Characterisation of novel lignocellulosic enzymes from the shipworm symbiont *Teredinibacter turnerae*

Claire Alexandra Fowler

PhD

University of York

Chemistry

September 2018

Abstract

Biomass is a renewable resource which can be broken down into bio-derived high value molecules using industrial processing, parts of which rely on enzymatic degradation. Nature provides a vast pool of highly efficient lignocellulosic enzymes capable of breaking down the multitude of natural polymers found in biomass. Identifying organisms that thrive within ecological niches based around the consumption of lignocellulosic material is likely to yield novel biomass degrading enzymes. Shipworms are marine organisms which use shell-like protrusions to burrow into submerged wooden substrates in oceanic environments. Endosymbiotic bacteria, *Teredinibacter turnerae*, found within the gills of the shipworm is thought to provide the host with a variety of carbohydrate active enzymes for use in the digestion. This work characterises carbohydrate active enzymes identified within the genome of *T. turnerae* including glycoside hydrolases (GH) from families 5, 8 and 12 as well as a lytic polysaccharide monooxygenase (LPMO) from subfamily AA10. Analysis of three GH5 proteins found activities on cellulose (*Tt*GH5_2 and *Tt*GH5_un) and xyloglucan (*Tt*GH5_4), with one protein (*Tt*GH5_un) likely representing a new GH5 subfamily. *Tt*GH8 was found to be highly active on xylans, with kinetics and structural studies showing a preference for longer xylooligosaccharides. The molecular structure of *Tt*GH12 was solved, but substrate specificity remains unknown, indicating a potentially new activity subclass within the GH12 family. *Tt*AA10 was active on crystalline cellulose through copper mediated oxidative attack at C1 or C4 either side of the glycosidic bond, and modelled structurally at high resolution. Like many other specialised organisms, the genome of *T. turnerae* is a treasure chest full of potentially useful lignocellulose degrading proteins. This work has assessed the activity and structure of several enzymes from *T. turnerae* and found multiple different substrate activities and potentially novel functions, paving the way for further experimentation and potential utilisation in industry.

List of Contents

List of Tables

List of Figures

Acknowledgements

This is the end of a long road, bound from a decision made in optimistic youth to one steeped in determination to succeed. This road may have become less well-travelled, narrowed and winding along the way, coming to a close, but the journey isn't over – there are only ever new roads to follow, but York will forever be a place to call home. Eight years spent here in the pursuit of knowledge and personal growth, it is a long time and I wouldn't have got through it without the help and support of so many wonderful people. Transferring from a world of atoms and bonds to the macro world of proteins was a well chosen decision, but one I could not have undertaken without the encouragement I received from my supervisors, Gideon Davies and Paul Walton, who gave me a wonderful scientific opportunity that I could not shy away from, no matter how nervous I was before I undertook this PhD. I would like to thank them for their input in this work, but also for giving me the space to be creative and reach for my own goals. I have gained an independence pre-York-Claire wouldn't have known to be possible. I would also like to thank Martin Fascione for his encouragement and kind words throughout my four years tackling this project.

A great thank you also goes to Glyn Hemsworth for taking on a nervous chemistry MChem project student and teaching me the ways (and bad habits!) of a real protein chemist. For all your help in those early years I will always be grateful. The York Structural Biology Lab (YSBL) may be essentially a scientific chemistry island, far away from the department, but it has been a wonderful place to work and I know how lucky I am to have been here. With the constant support of those in charge of keeping things running smoothly, I have learnt how a team should work together; Simon, Louise, Jules, Sam and Johan, Catherine and Jane, I give my thanks. YSBL means something different to the different people that flow through its corridors over time. My happiness in YSBL came from the friends I have made here, Becky, Imogen, Adam, Alexandra, Tamara, Paul, Bryony, Lukasz, Richard, Harriet, Lewis, Martin, Ali and Emma. You are the people that define my time in this department and your friendship and support means everything to me. Across the lake, so to speak, Luisa, you have been a rock for me the past couple of years and I am forever grateful for all the late-night ranting you put up with and for your unwavering support and encouragement. Lastly, to Ben, without whom I would never have had the courage for this endeavour.

My parents, Joan and Ken and my brother, Richard, I know you'll pretend to read my thesis and understand it, but nevertheless I know how proud you are and I'm so glad of this achievement and for being able to be so close to you all for the last eight years. I think we will all miss driving these roads. Although they won't really understand (as they can't read), my thanks go to Monty and Luna – two wonderful house bunnies and my little darlings, who have always been there no matter what, never judging, always providing companionship and love, even whilst I was trying to navigate the windiest of bends.

Finally, a thank you to my past self, for always staying optimistic and developing the independence and drive to push myself into such a promising direction. Here's to the future.

Author's Declaration

The work presented in this thesis has been carried out by the author, unless otherwise stated in the text. I declare that this thesis is a presentation of original work and I am the sole author. This work has not previously been presented for an award at this, or any other, University. All sources are acknowledged as References.

Work carried out by other persons;

- Protein ESI-MS was carried out in all cases by Dr. Andrew Leech, York Technology Facility.
- EPR experiments were carried out under the guidance of Dr. Luisa Ciano or Martin Steward. All EPR simulations were carried out by Dr. Luisa Ciano
- The SUMO-*Tt*AA10 construct was originally made by Dr. Glyn Hemsworth before the start of this work. Primers to make the pelB-SUMO-*Tt*AA10 construct and psf SUMO-like *Tt*AA10 constructs were designed by Dr. Glyn Hemsworth
- SEC-MALS analysis was carried out by Dr. Andrew Leech, York Technology Facility.
- HPAEC analysis was carried out under the guidance of Dr. Fiona Cuskin and Prof. Harry Gilbert, Newcastle University.
- Figure 3 was adapted from one kindly provided by Prof. Gideon Davies
- Figure 5 includes a diagram kindly provided by Prof. Paul Walton
- GH catalytic domain boundaries were chosen based on information provided by Prof. Bernard Henrissat
- The vector containing the maltose binding protein (MBP), pETFPP2 was given by the Prof. J Potts group (Biology, University of York) and originally made by the University of York Technology Facility
- Collaborative work to establish the substrate specificity of *Tt*GH12 was carried out by Gregory Arnal and sequence analysis by Dr. Alexander Holm Viborg, under the direction of Prof. Harry Brummer, The University of British Columbia.
- Collaborative work to further assess the activity of *Tt*GH5_2 and *Tt*GH5_un on mannans using PACE was carried out by Dr. Theodora Tryfona, under guidance from Prof. Paul Dupree.

1

An Introduction

1.1 Abstract

Demand for biofuels is increasing amid positive shifts in political and public opinion regarding the growing need for more sustainable fuel sources. In this Chapter the nature of complex carbohydrates, its effective recalcitrance towards degradation and ultimately the enzymatic tools for polysaccharide degradation will be discussed. Breakdown of the majority of different plant polysaccharides found in Nature is carried out by a class of enzymes known as glycoside hydrolases (GH), enzymes able to break the glycosidic bonds between the carbohydrate monomers of enzyme-specific substrates through hydrolytic mechanisms. These carbohydrate active enzymes can be classified into over 150 enzyme families based on their sequence homology with family groupings often showing activity through specific mechanisms or on specific types of polysaccharides. The portfolio of enzymes known to degrade complex polysaccharides was revolutionised by the recent discovery of lytic polysaccharide monooxygenases (LPMO); copper dependent enzymes capable of using oxidative chemistry to create chain breaks in otherwise GH inaccessible, insoluble, crystalline polysaccharides. The story of their discovery and applicability to polysaccharides degradation will be discussed. Considered by some as a common marine pest, shipworms (marine bivalve molluscs) are able to survive off a diet of wood alone despite possessing a gut mostly devoid of bacteria; this makes them an interesting target of novel enzyme discovery. Lignocellulose degradation may occur through synergistic action of host enzymes and enzymes produced by a community of endosymbiotic bacteria, *Teredinibacter turnerae*, housed within the gills of the shipworm. These fascinating animals and their symbiotic bacteria are the source organism for this entire work, whereby the goals of characterising novel carbohydrate active enzymes taken from the genome of *Teredinibacter turnerae* will be discussed.

1.2 Fuels of the future

A barrier towards the sustainable and efficient usage of plant biomass for fuel conversion lies in the complexity and recalcitrance of plant cell walls. The heterogeneous matrix of various carbohydrate compounds hinders the enzymatic breakdown of plant cell walls into energy rich carbohydrate monomers. Crystalline cellulose regions are interspersed with a web of more soluble polysaccharides, known as hemicelluloses. The plant cell wall matrix is strengthened by a hydrophobic and insoluble barrier, a mix of phenolic compounds known as lignin.

The effect of human industrial activity on the planet is very significant and is likely the cause of the greatest issue currently facing us as a whole; climate change. The burning of fossil fuels is coupled with the slow increase in the Earth's temperature due to a build up of CO_2 gas within the atmosphere. The greenhouse effect resulting from the build up of emissions had been predicted (and unfortunately put aside) as early as 1896 by Svante Arrhenius[1-2], and it is only now that the true effects of unnatural increases in the planet's temperature are being felt. Governments world-wide have been alerted to the serious threat of climate change and many have pledged to reduce carbon emissions in an effort to reduce the overall temperature increase. Reduction in emissions on a global level could come from removing our reliance on the burning of fossil fuels for energy. Fossil fuels are a non-renewable resource and one that is becoming scarcer as the decades progress. There are a wide selection of relatively untapped renewable sources of energy; solar, wind, tidal and hydro that are all valuable sources of potential energy. Indeed, when observing the countryside of Great Britain these days you will often come across enormous wind turbines, or looking out to sea you will be greeted by views of expansive wind farms. Steps are being made to use these cleaner sources of energy and another potential source is locked in the energy rich carbohydrates that make up plants and other biomass.

Biomass is a term for any substance of natural origin and such materials can be used in the production of biofuels. Industrial biorefineries aim to produce energy in the form of ethanol, fermented from glucose. In concept, a biorefinery can use any type of biomass material as a feedstock and produce biofuels or high value bio-based molecules, but in practise this process is limited by several issues. The major concern in the use of biomass feedstocks is the origin of the material. There is an ethical conundrum over whether land should be prioritised for the growth of crops for food or fuel. Annual production of biomass is in the realm of 10^{11} tons, made up of approximately 60 % terrestrial and 40 % aquatic sources.[3]

Only a tiny percentage of the annual biomass production is related to human cultivation (3 %)[3], a small number, but likely due to limitations such as land availability and environmental conditions. Overall, approximately 50 % of biomass is lignocellulosic in nature, however access to biomass varies by environmental conditions and differences in cultivation preferences of different countries.[4] Crops such wheat, corn and rapeseed, grown specifically for industrial conversion into fuels are known as first generation feedstocks. Sugar, starch, vegetable oils and even animal fats can all act as first generation feedstocks and be converted into useful products. Different products can be gained from different feedstocks; bioethanol can be made from sugarcane whilst biodiesel is made from crops such as rapeseed. Biomass from plants contains two types of raw material; assessible sugars naturally stored by the plant such as starch, and structural sugars bound in the lignocellulosic fraction of the plant. Starch is a polymer of glucose which is relatively easy to depolymerise, which can then be fermented into ethanol. Lignocellulose is recalcitrant towards depolymerisation due to the combination of polysaccharides which form the structural plant component. An alternative source of biomass is the waste products of food production, namely the lignocellulosic fraction; for instance, in a crop of corn, only the corn is taken for food use and the rest of the plant including the leaves, stem and corn stover are waste. Waste biomass is termed a second generation feedstock and its use bypasses the ethical issue of taking land away from food production, whilst maximising efficiency by recycling waste material. In second generation biorefineries the glucose (and other oligosaccharides) is removed from plant lignocellulosic materials through a mixture of chemical and enzymatic steps. Chemical refers to non-natural lignocellulosic degradation processes and these steps are commonly known as pre-treatments; pre-treatments aim to change the state of the plant material into something more amenable to enzymatic breakdown. For example, thermochemical pre-treatments use high temperatures and dilute acids or ammonia to solubilise hemicelluloses and break up the crystalline cellulose regions. Other treatments may involve different conditions which result in different changes to the accessibility of the polysaccharides.[5-7]

Political change is driving research towards finding solutions to reduce the cost of producing bio-based molecules, with several countries committed to reaching targets involving use of bio-derived fuel sources in the near future.[8] The success of the biorefinery concept is complex and relies on a multitude of issues which can limit the cost benefit of biomass processing. For example, transport of biomass from its source to the plant, coupled with storage of the material can be an inefficient and costly process. Many avenues of industrial

processes and political policies must combine together to efficiently take advantage of the abundant source of biomaterial available to us. Nature has already defined its own efficient process of lignocellulosic breakdown, through the synergistic action of a wide variety of different enzymes. A single organism can produce a consortium of enzymes capable of lignocellulosic degradation or can utilise the symbiotic behaviour of other smaller organisms such as bacteria. Enzyme cocktails are available to purchase from various biotechnology companies such as Novozymes, with the most recent incarnations (i.e. Cellic© CT3 HS) containing propriatory mixtures of β-glucosidases, hemicellulases and AA9 lytic polysaccharide monooxygenases. Industrial and academic research into enzymes capable of degrading the various natural polymers thus provides a route to improving industrial processing, by finding the most efficient enzymes for particular biomass materials. Improvements in the activity of industrial enzyme cocktails can shift biofuel production into a more economically vialble process, which can be used at large scale to produce cleaner and renewable fuels. For example, the more active a set of enzymes is in breaking down biomass, the less enzyme dosing is required, which in turn makes the process more cost effective. Biochemical research, such as the work presented in this thesis, has the potential to improve the knownledge basis about novel lignocellulosic enzymes. Thus, it is my hope, that this research may in the future be used to contribute to improving existing enzymatic cocktails or provide new catalytic routes in the production of renewable biofuels.

1.3 Complex Carbohydrates

Whilst there is an abundance of biomass produced globally every year, there is also a large degree of variation in the type of biopolymers made available by nature. Carbohydrate structures are highly variable, with their structure formation underpinned by the vast number of possible stereochemical assemblies. The number of isomers, both linear and branched, able to form from the set of atoms within a single reducing hexasaccharide is vast and nature has been able to use this diversity to its advantage, forming an enormous variety of structures from the most basic sugar building block.[9] The structure of glucose, arguably the most recognisable monosaccharide is shown in **Figure** 1. A free monosaccharide unit in solution is able to undergo reversible ring opening and closing, forming either α or β isomers; carbon 1 (C1) is the anomeric carbon within the ring structure and in open form is linked to an aldehyde functional group. Rotation of the anomeric carbon into different orientations followed by ring closing produces either the α-isomer, in which the OH group point downwards in an axial direction, or the β-isomer, where the OH group is in an

equatorial position. The same process can occur in sugar units at the end of a polysaccharide chain in which the anomeric carbon is not bonded to another molecule. Free monosaccharides and free chain ends are known as reducing sugars, and reducing chain ends respectively due to the reversible formation of the aldehyde at C1. Functional groups on each carbon position of the hexose ring, C1-5 can be either axial or equatorial and differences in positioning and type of functional group defines an individual monosaccharide, as shown in **Figure** 2. From formation of structure within plants, energy storage to biological signalling, carbohydrates play a role in most areas of life.

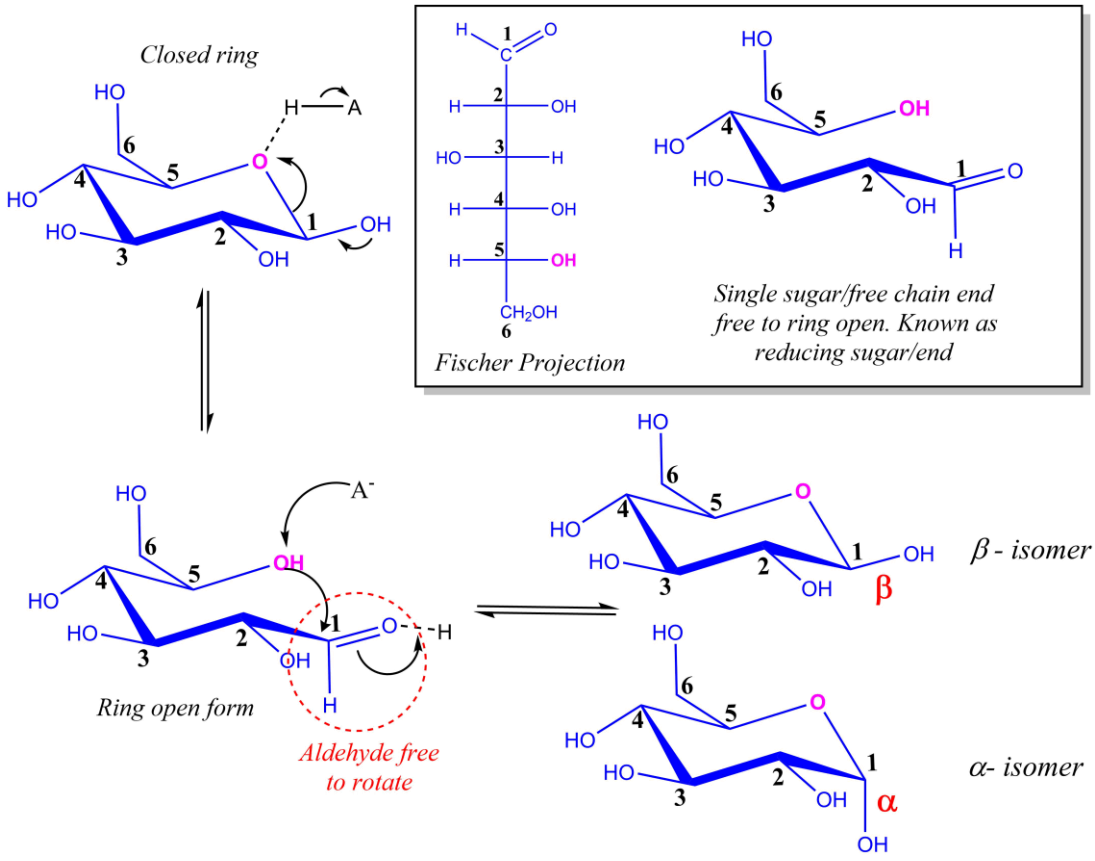


Figure 1 Diagram showing the ring opening mechanism of free single sugar units, or free chain ends of a polysaccharide. Ring closing involves the ring oxygen, (OH covalently linked to carbon 5 in the Fischer projection) interacting with the aldehyde functional group on carbon 1. Ring opening and closing within solution is able to produce either isomer, α (axial) or β (equatorial) around the anomeric carbon. A free sugar unit such as this is known as a reducing sugar or if on a polysaccharide, a reducing end.

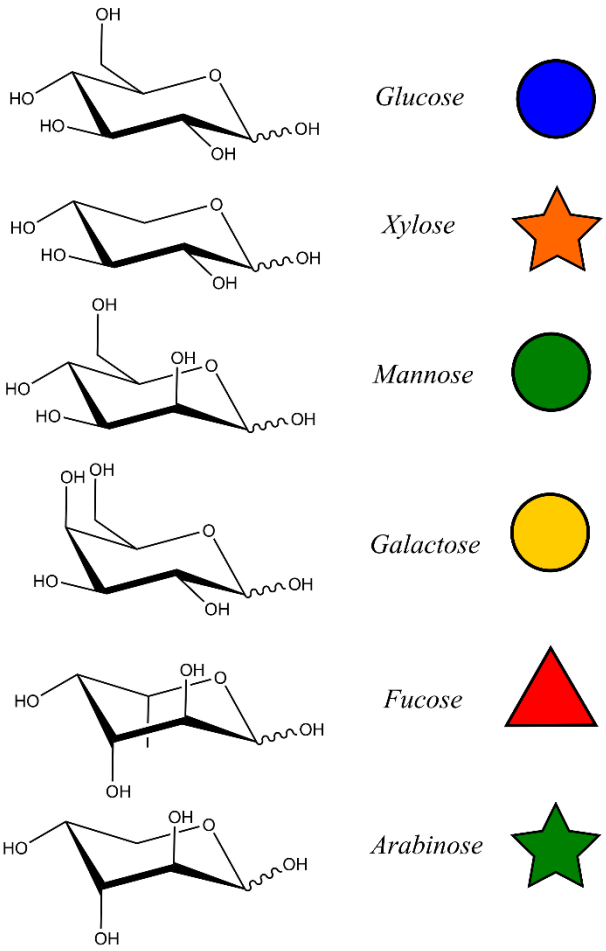


Figure 2 Examples of some common monosaccharides used to make up various different polysaccharides. The carbohydrate structures are shown without hydrogens labelled for ease of comparison of OH positioning. The OH group linked to the anomeric carbon is shown by a wiggly line, which indicates the OH group can take either axial or equatorial positions. The structures are annotated with their names and associated symbols as described by the Symbol Nomenclature for Glycans (SNFG).[10]

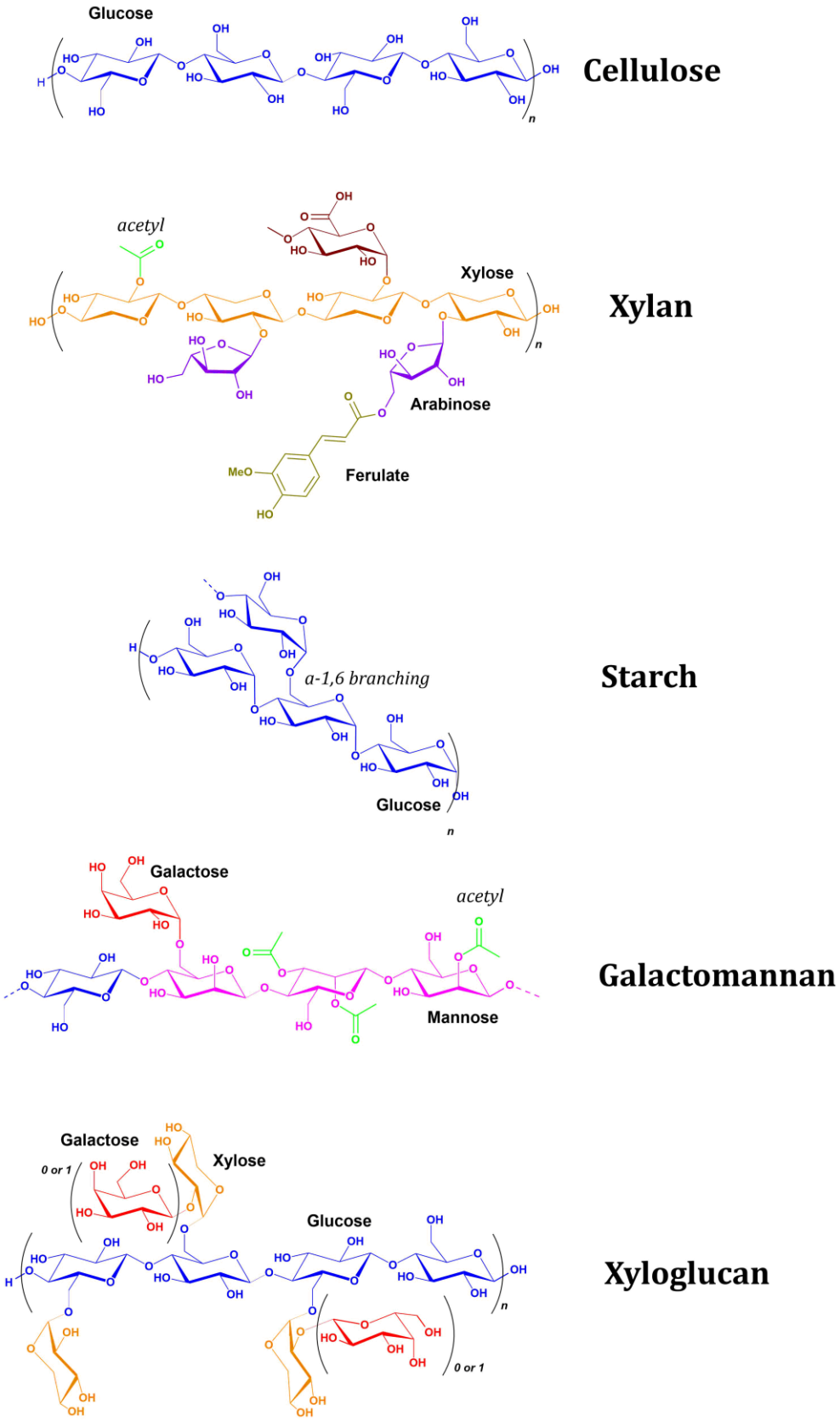


Figure 3 Diagram showing different types of common polysaccharides, formed by various different monomers. Figure adapted from one provided by G. Davies.

This work will focus on the enzymatic breakdown of structural plant polysaccharides, carbohydrate substrates which are built to withstand degradation, **Figure** 3. Surrounding a plant cell is the plant cell wall (PCW), a protective layer built up of networks of complex polysaccharides and often segmented into the primary and secondary cell wall.[11] The primary cell wall is firstly comprised of crystalline cellulose, simple linear chains of β (1-4) linked glucosyl units, which are embedded in a matrix of complex polysaccharides formed of either pectin, or hemicelluloses, **Error! Reference source not found.**. The cellulose chains naturally pack together due to extensive hydrogen bonding between chains. Hemicellulose is formed from the same simple glucan backbone as cellulose but contains branching structures and other modifications that prevent the formation of ordered microfibrils. Instead the single chains interact and wrap around the surface of the ordered cellulose fibres. Pectins are complex polysaccharides with a multitude of different functions containing a linear backbone of α (1-4) linked D-galacturonic acid moieties, often with substitutions branching off the main chain by moieties such as xylose, rhamnose, galactose and arabinose depending on the species origin. Pectins can form hydrated gels and glues able to adhere neighbouring cells together via the middle lamella.[11]

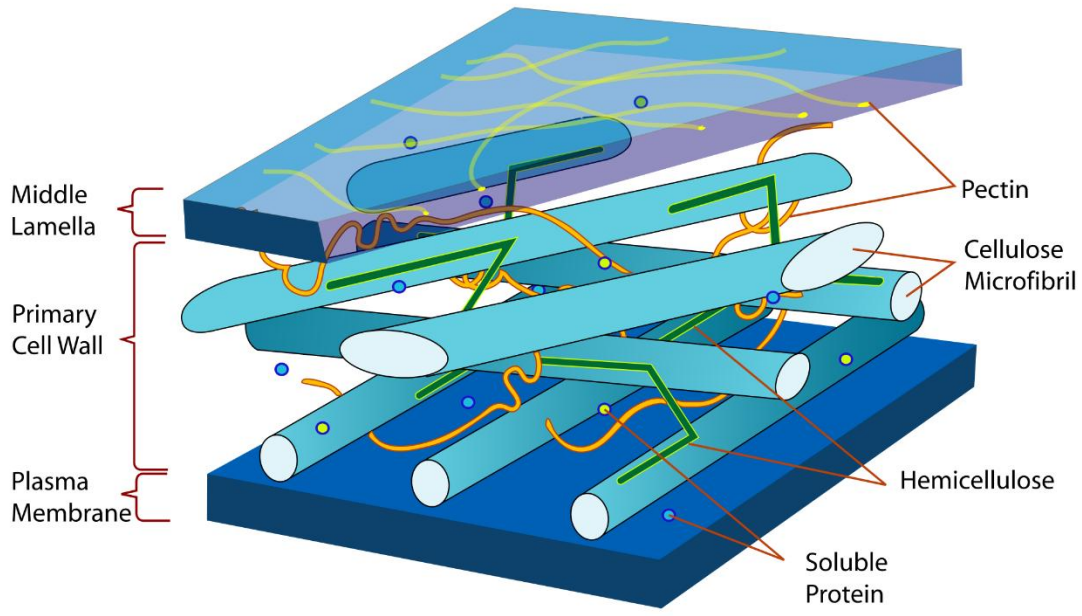


Figure 4 Diagram of a plant cell wall structure showing cellulose microfibrils linked with hemicellulose and pectin . Image free for public use from www.wikimediaorg.

The primary cell wall can be classified by type. The primary cell wall of dicots for example is made up of type I materials; cellulose microfibrils surrounded by an overlapping network of xyloglucans (a hemicellulose) which crosslink between separate microfibrils of cellulose and other complex polymers (i.e glucuronoarabinoxylans (GAXs), a pectin). Primary cell walls in monocots are normally formed of type II structures, where the cellulose fibres are mostly wrapped in networks of cross linking pectins and less hemicelluloses (i.e. xyloglucan) is present.[11]

The secondary cell wall can in most cases be attributed to plants cells which are required for strength and structure.[11] The strength and rigidity of the secondary cell wall is formed of organised cellulose microfibrils buried inside interacting hemicelluloses, which is further cross linked by aromatic compounds such as lignin.[12] Lignin is the second most abundant natural polymer after cellulose and is used in the cell wall for strength and stiffness and even to waterproof the cell from the environment. It is a heterogeneous mixture of aromatic polymers with basic structures formed from various phenylpropanoid monomers.[13]

The issue with effectively breaking down plant materials for energy gain, be it ourselves and our drive towards more sustainable fuels or simply the attack of a plant pathogen is that nature has caused plants to evolve structures that are recalcitrant to degradation; they are difficult to break down. In terms of the secondary cell wall, the lignin provides the first line of defence. Lignin is a major barrier towards the breakdown of plant polysaccharides in an industrial sense and enzymatic methods are not yet fully developed to deal with this outer cell wall layer.[8] If the lignin layer can be bypassed, the breakdown of hemicelluloses is in a relative sense, easier. The hemicelluloses are more freely assessable to enzymes due to their inability to naturally form ordered structures, and importantly can be solubilised. The main energy component of the plant cell wall in both secondary and primary structures is that of cellulose. Cellulose is the most abundant terrestrial polymer and stores a vast amount of potential energy if it can be broken down to glucose. The repeating and linear nature of the glucan chain in cellulose allows single chains to easily stack together through hydrogen bonding, Error! Reference source not found.; the relaxed chair conformation of glucose places the hydroxyl groups in equatorial positions, whilst the ring hydrogen atoms are axial. The edge hydroxyl groups of linear chains interact with those of another parallel glucan chain, forming a flat ribbon. The ribbons (or sheets) stack on top of each other with a slight stagger, again caused by the locked positioning of the hydrogen bonding pairs. Whilst strong hydrogen bonding holds chains together in microfibril groups, there are also hydrophobic interactions occurring between fibres and even weak C-H-O hydrogen bonding.[14] Stacking of

multiple ribbons produces an ordered microfibril which hinders breakdown of the cellulose, where the outer face of the fibres is hydrophobic and thus insoluble. Furthermore, the natural assembly of glucan chains in this manner leads to highly ordered structures which are essentially crystalline.[5] Regions of cellulose can become disordered, losing their crystalline nature somewhat and are described as amorphous. Whilst the ordered nature of cellulose provides an excellent structural polysaccharide for plants, it creates problems for organisms wishing to utilise it as an energy source.

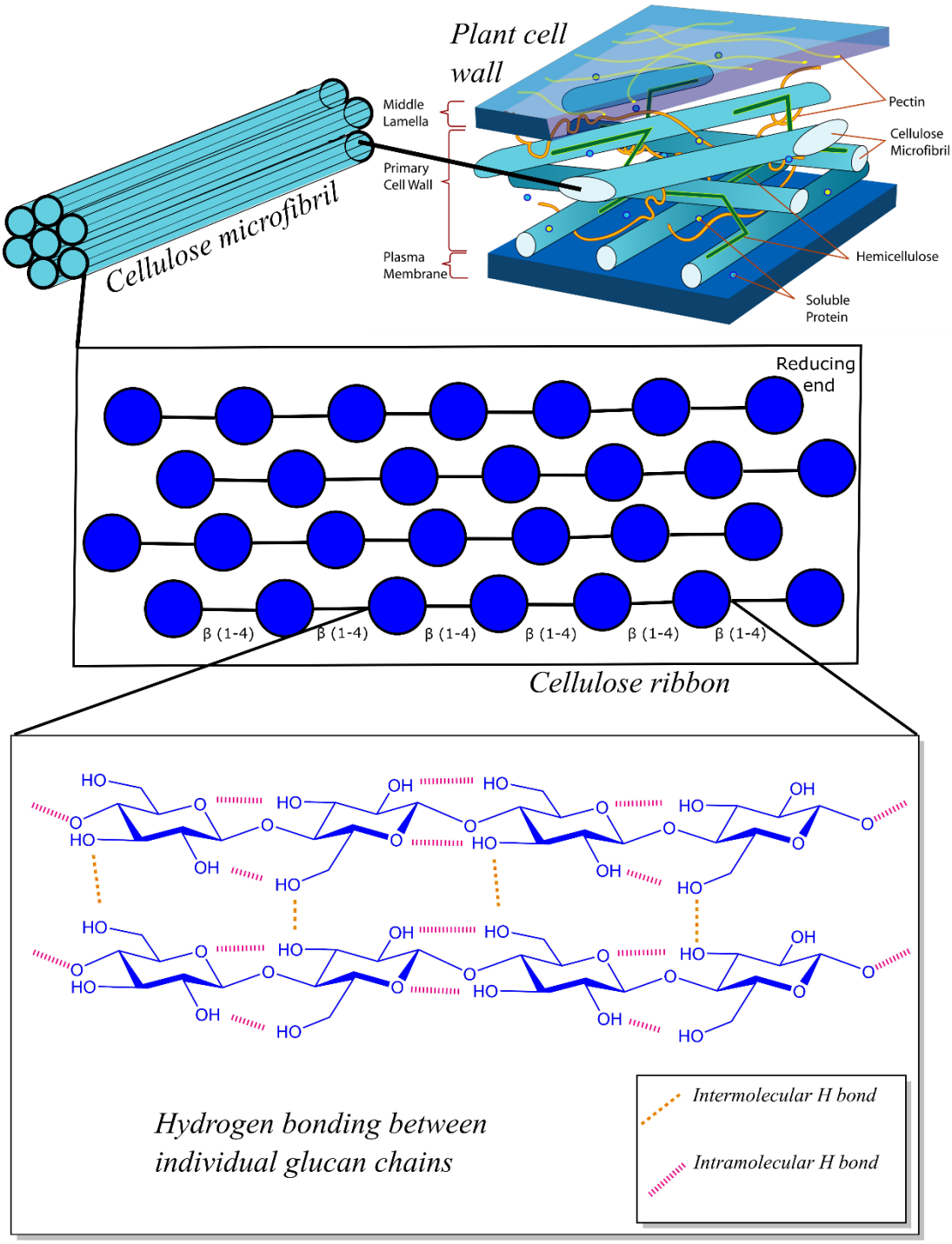


Figure 5 Structure of cellulose – microfibrils are made up of flat ribbons of stacking glucan chains. The glucan chains interact through hydrogen bonding, both within a single chain and between chains, locking the chains in position.

1.4 Glycoside Hydrolases

As Nature is able to build these impressive and complex carbohydrate structures, it must also possess a way of reversing the process and glycoside hydrolases (GH) are nature's answer. GHs are found across the tree of life, providing a large degree of roles to many different organisms which utilise them; the clearest example being the breakdown of the structural carbohydrates in plants during digestion. Organisms may make their own GHs, or play host to digestive bacterial communities which are predisposed for the production of large amounts of these essential enzymes. Most GHs work on the principle that two key catalytic residues, held in an active site, are able to break apart a glycosidic bond within a bound substrate through a specific hydrolytic enzyme mechanism. Enzymatic hydrolysis of a polysaccharide takes place at the individual bonds between the sugar units, specifically at the (aldose) glycosyl C1-O bond by general acid catalysis. The hydrolysis reaction can be thought of as a formal nucleophilic substitution, whereby the cleavage of the glycosidic bond yields two products, one of which contains a new reducing end unit. The leaving group alcohol is a poor (high p*K*a) leaving group and requires protonic (Brønsted acid) assistance for departure. An amino acid residues acting as a proton donor is required to carry out this task. Likewise, nucleophilic attack in a double displacement, or Brønsted base assistance to attack by water are also required and in GHs these acid/base/nucleophile roles are taken up by what are known as the catalytic residues.

With stereochemistry in mind, two major catalytic pathways are used to break apart a glycosidic bond; namely, net retention or inversion of the anomeric configuration after the scissile bond is broken. This leads to the enzymes being classified as either retaining or inverting enzymes. The residue acting as a proton donor in both retaining and inverting enzymes is held in an equivalent position, enabling a strong hydrogen bonding interaction with the glycosidic oxygen of the bound substrate moiety. The position of the second catalytic residues changes based on whether there is net retention or inversion of the anomeric centre. In a retaining enzyme, the nucleophilic catalytic base is close to the bound substrate (approximately 5.5 Å), whereas a larger distance from the catalytic base to the substrate is required (approximately 10 Å) for an inverting enzyme as this creates enough room for a water molecule to be positioned in between the substrate and the base.[15-16]

1.4.1 Inversion of stereochemistry

Hydrolysis with inversion of the anomeric configuration occurs over a single enzymatic step which is known to proceed through an oxocarbenium-ion like transition state; where double bond character between the ring oxygen and C1 is formed transiently.[17] An inverting reaction uses two catalytic residues, typically carboxylate resides such as glutamic acid and aspartic acid, which take on the roles of an acid and a base. A cascade of electron movement drives the formation of the transition state as shown in **Figure** 6.[18]

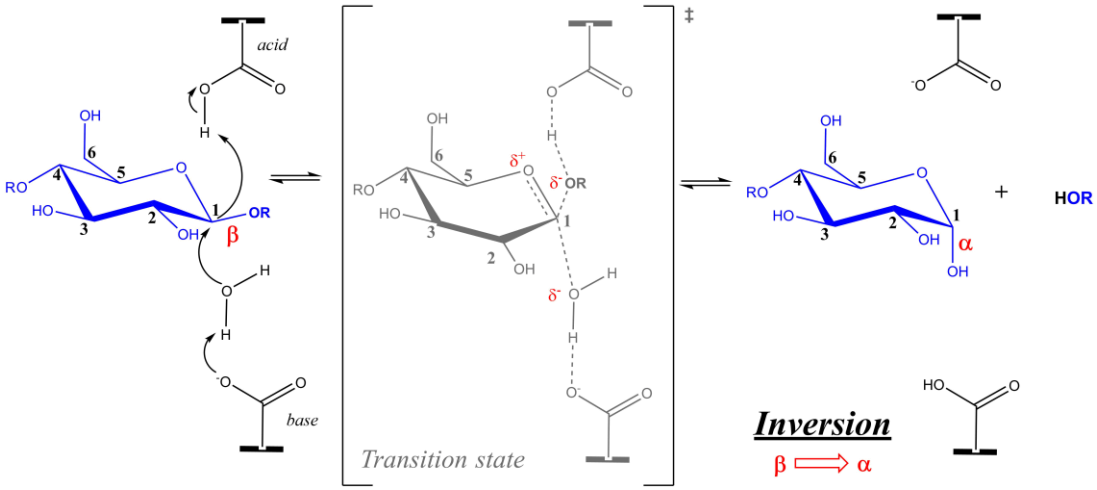


Figure 6 Diagram of glycoside hydrolysis via the inversion mechanism showing cleavage of a β-glycosidic bond in a single glucan chain (where R represents the continuation of the chain). The inversion mechanism proceeds through a single transition state which shows double bond character between the ring oxygen and anomeric carbon. The simultaneous addition of water inverts the stereochemistry at the anomeric carbon as the leaving product, OR departs and becomes protonated, leaving the product with inversion of its original stereochemistry.

1.4.2 Retention of stereochemistry

Hydrolysis with net retention of the configuration around the anomeric centre proceeds over two separate enzymatic steps involving the formation, and subsequent breakdown, of a covalent glycosyl-enzyme intermediate; the two steps often called glycosylation and deglycosylation, are shown in Figure 7.[18] Each individual step results in inversion of the anomeric configuration but overall leads to retention of the original stereochemistry and the reaction was coined as a double-displacement mechanism by Koshland in 1953.[17] In the glycosylation step, the bound substrate is attacked by the nucleophilic residue at the anomeric carbon, whilst the second catalytic residue initially acts as an acid, protonating the leaving group; this proceeds through an oxocarbenium ion-like transition state, where the displacement of the leaving group aglycon from either the α or β configuration results in the formation of a covalent glycosyl-enzyme intermediate with the opposite configuration around the anomeric centre. In the deglycosylation step, the second catalytic residue switches role to act as a base, whereby it deprotonates a nearby water molecule, which then attacks the anomeric carbon, resulting in a second transition state through which the covalent glycosyl-enzyme intermediate is broken down. This second step proceeds through the same transition state, creating a second inversion of stereochemistry, which results in an overall retention of the original conformation of the substrate.[17] The ability of the second catalytic residue to switch roles between acid and base during catalysis is known as pKa cycling, in which the pKa values oscillate to be appropriate for the changing role of the residue.[19]

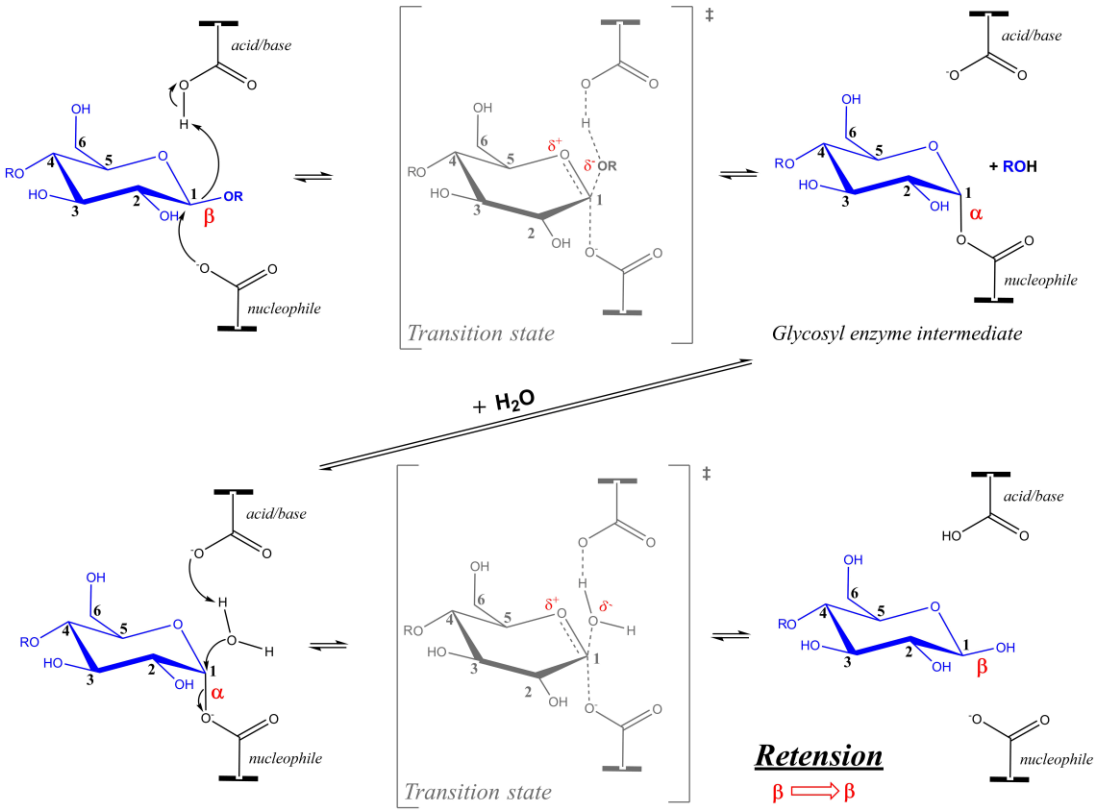


Figure 7 Diagram of glycoside hydrolysis via the retention mechanism showing cleavage of a β-glycosidic bond in a single glucan chain (where R represents the continuation of the chain). The retention mechanism proceeds through two transition states which show double bond character between the ring oxygen and anomeric carbon. A glycosyl enzyme intermediate exists between the two transition steps in which the anomeric carbon at the newly formed reducing end forms a covalent bond with the nucleophilic residue in the opposite conformation to the starting material, whilst the leaving group, RO becomes protonated and departs. The addition of water through a second transition state inverts the stereochemistry at the anomeric carbon again, leaving the product with a net retention of its original stereochemistry.

1.4.3 Binding and Conformations

Binding of a ligand within the active site of a glycoside hydrolase involves interactions beyond those carried out by catalytic residues. The available binding positions for individual sugar units are known as subsites, which are conventionally numbered *–n* to + *n,* where cleavage occurs between the sugar units taking up the -1 and +1 subsites, Figure 8. The catalytic residues are positioned between the two subsites, ensuring optimal positioning of the glycosidic bond to be broken. Common interactions which hold a ligand within an active site result from ring stacking interactions between the ligand and aromatic residues lining the active site space. Hydrogen bonding networks created by hydrophilic residues and water molecules also provide strong binding affinity to the ligand.[20]

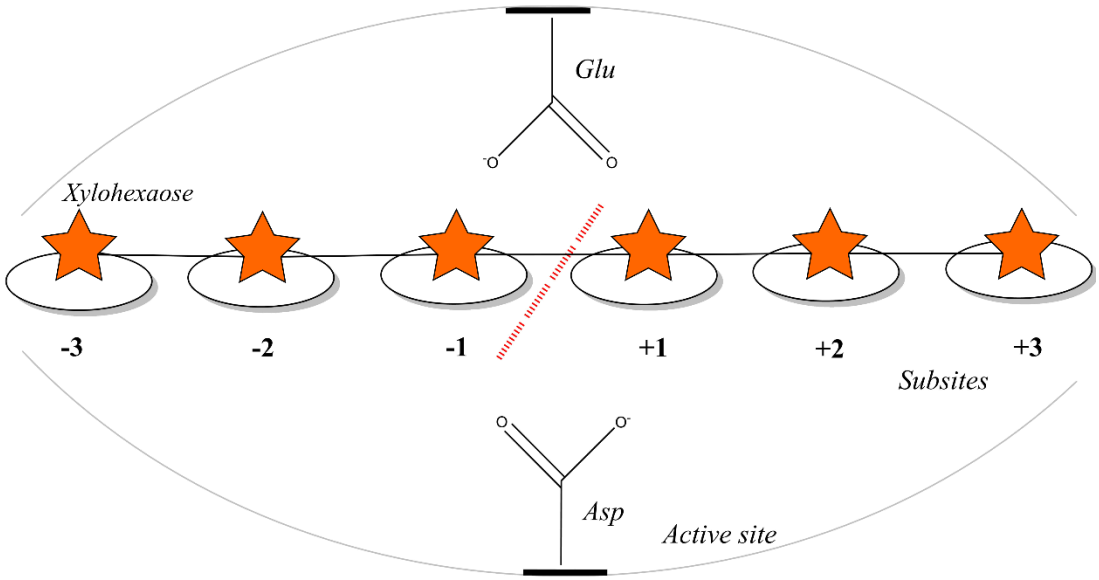


Figure 8 Representation of subsites with the active site of a generic xylanase. The xylose monomers of xylohexaose, depicted as orange stars each take up one of the subsites within the active site. The catalytic residues, Glu and Asp are shown in the middle with the bond being cleaved (indicated by a dashed line) between the -1 and +1 subsites.

The orientation of amino acid residues and the way they are able to interact with a substrate can assist with the hydrolysis reaction by inducing a conformational change on the sugar unit in the -1 position. The functional groups on a pyranose ring provide a large degree of variability in structure, a monocyclic ring can also undergo ring puckering which changes its conformational shape. In 1975 Cremer and Pople[21] described a coordinate system for a 4, 5 and 6 membered rings, in which the various conformations were mapped out using amplitudes and phases. For pyranoid rings, Jeffrey and Yates[22] showed how the many different conformations could be mapped out spherically and how this could relate to conformational nomenclature. The sphere is described as having a north and south pole, where at the north point is positioned the classic chair form of a sugar moiety, written as $^{4}C_{1}$. This simple nomenclature, $^{A}X_{B}$ describes the puckering of the ring. Natural puckering of the ring will leave 4 atom positions within the same plane, whereas the remaining two positions will sit out of this defined plane. In the nomenclature, (superscript) *A* describes atoms which lie above the plane of the ring and (subscript) *B* those ring positions which lie below plane, and X is shorthand for the type of puckering; in the $^{4}C_{1}$ example, C refers to the relaxed chair conformation, in which position 4 and 1 lie above and below the plane of the ring respectively, **Figure** 9.[22] At the opposite pole, is an inverted version of this chair form, written as $^{1}C_{4}$. As the conformations are mapped spherically based on the set of parameters calculated by Cremer and Pople, movement around the sphere based on latitude describes the degree of distortion away from the perfect chair conformation, whilst the longitude

movement describes the type of distortion; the most recognisable conformations being half chair, skew boat, boat and half boat, Figure 9.[22] The nature of the puckering is known as conformational analysis and each individual conformation can be given a relative energy. Different sugar moieties will have access to different conformations based on their structure, as the changes in functional groups on each ring positions can help or hinder formation of certain conformations. As described previously, GHs can catalyse hydrolysis through either inversion or retention of the anomeric carbon, and both mechanisms go through transition states that possess oxocarbenium double bond character; as the ring interacts with the catalytic residues and/or nucleophilic water the transition states has a dissociative nature, whereby there is a large build-up of positive charge on the ring. To counter this build-up of charge and stabilise the structure, the lone pair of electrons from the ring oxygen dissociates and forms the double bond character observed between the ring oxygen and anomeric carbon. To carry out this delocalisation, the conformation must change to something which allows for significant orbital overlap (to allow the sharing of electrons).[23] In Figure 6 and Figure 7 the transition states are shown distorted away from the chair conformation. Different combinations of substrate and enzyme will follow different conformational pathways during hydrolysis which result from several aspects of the reaction; the structure of the substrate and the non-covalent interactions within the binding site are key factors. Binding of a substrate in the -1 position can often induce a conformational change which is favourable in the formation of the transition state, aiding the reaction by lowering the energy barrier. This concept is used widely in biochemistry, for example to inhibit an enzyme efficiently one may often choose to design a transition state mimic, which displays a similar conformation to the reaction pathway at a higher binding affinity.

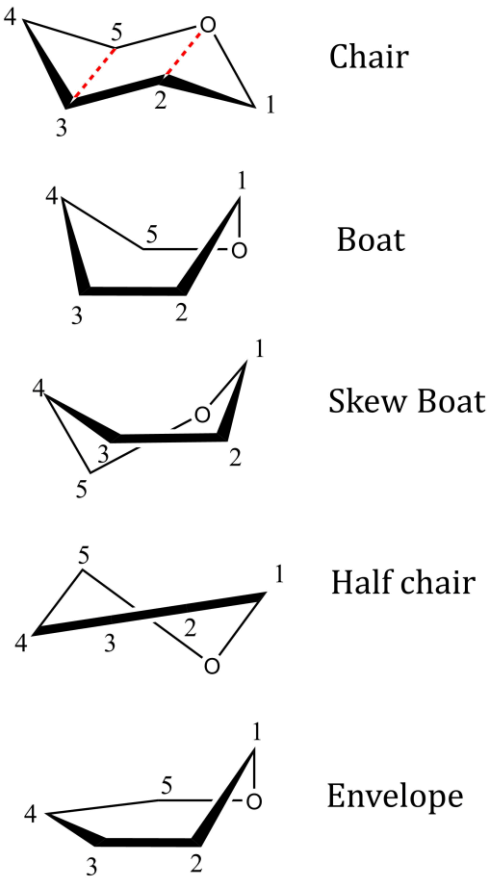


Figure 9 Conformers of a 6-membered ring. Dashed red lines give an example of four atoms lying in a plane in the relaxed chair form.

1.4.4 Topology of attack

Further complexity is introduced in whether the GH is *endo*-acting, meaning within a chain, or *exo*-acting, meaning action upon the end of a sugar chain (often, but not exclusively the reducing end, see Figure 1) as shown by the cartoon in Figure 10. The tertiary structure of a GH defines its mode of action on a polysaccharide chain, and the active site architecture of GHs can be classified into three topological types; pocket, groove and tunnel. Examples of each topology type are shown in Figure 11 using examples of GHs with different activities. Pocket active sites are typically specific to exo attack on the non-reducing chain end, producing monosaccharide products. Open active sites in the form of grooves (or clefts) are tailored for endo action, whereby the enzyme binds to a 'random' region of the polysaccharide chain, with active sites often offering binding sites to multiple monomers of the substrate. Finally, tunnels are simply 'closed over' groves, whereby long loop structures have evolved to cover the substrate groove. Generally, this architecture is found in cellobiohydrolases, and these types of enzyme are able to release product whilst remaining bound to the substrate. After product release, the enzymes are able to move along (process) the substrate and commence another hydrolysis reaction.[15] The use of loops can sometimes assist in substrate binding, or prevent unwarranted binding. An example is the 'thumb' region of GH11 enzymes, a long loop structure that crosses the active site and is thought to be involved in substrate recognition, as discussed further in Chapter 4. GHs differ on which mechanism they use, their preference for endo/exo activity and most importantly on their structure and function.[24] GHs are characterised and split up based on sequence and function into 135 families as described by the CAZyme classification system.[25]

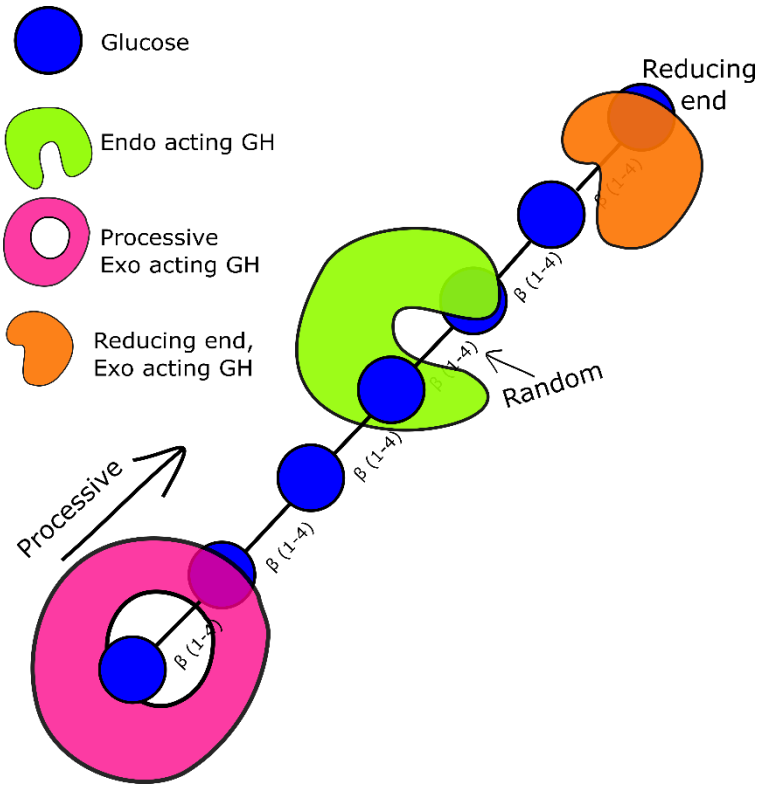


Figure 10 Diagram depicting generic GHs attacking a single glucan chain . Endo acting GHs will attack randomly along the chain, whereas exo acting GHs will start at a chain end and work processively along the chain.

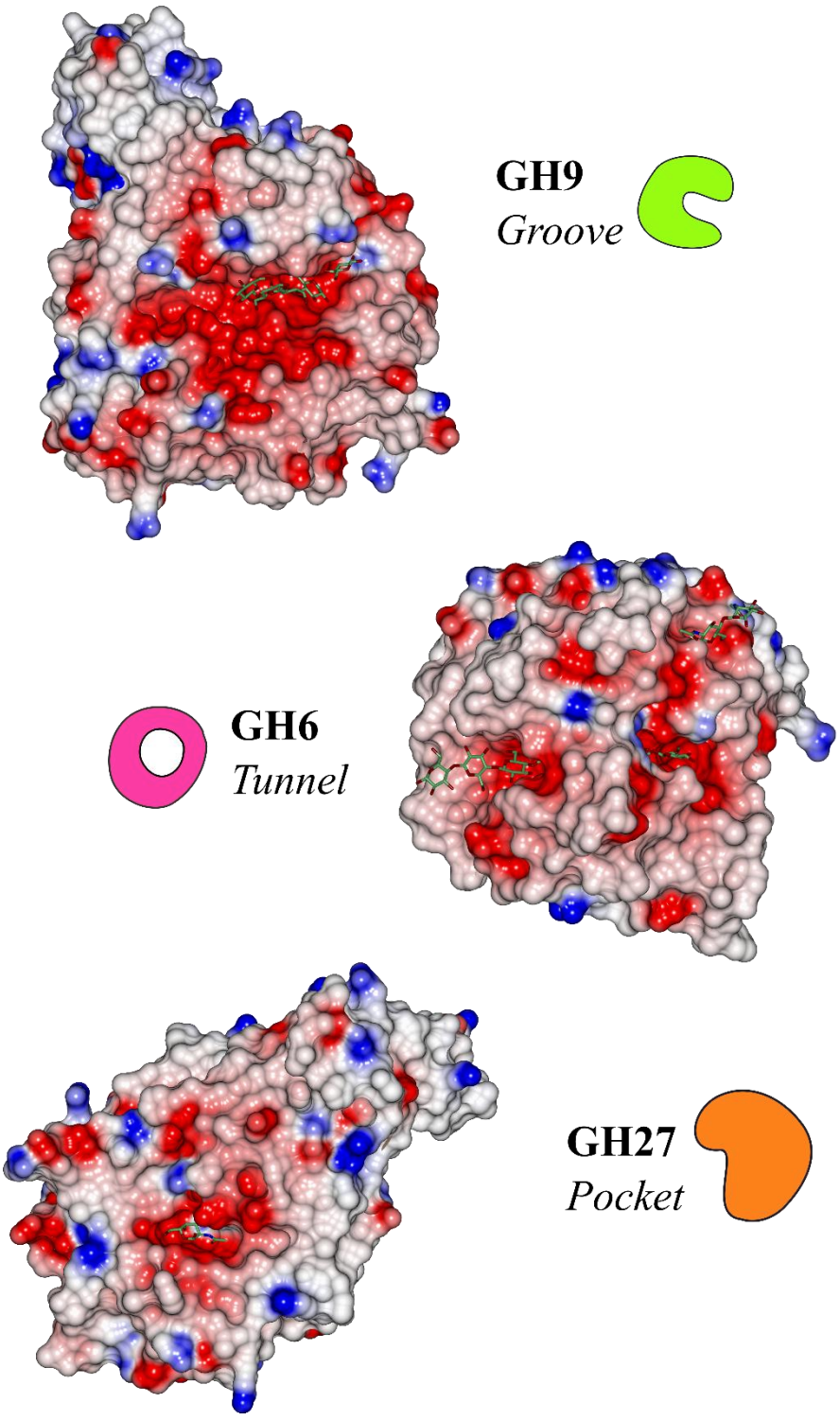


Figure 11 General active site topology of GHs as shown by three examples. GH9 (PDB code; 3H3K), cellulase CelA from *Alicyclobacillus acidocaldarius* in complex with cellotetraose; GH6 (PDB code; 4A05), cellobiohydrolase Cel6A, from *Chaetomium thermophilum*, in complex with cellotetraose and cellobiose; GH27 (PDB code; 1KTC), alpha-N-acetylgalactosaminidase from *Gallus gallus* in complex with N-acetyl-D-galactosamine.

1.4.5 Classifying Carbohydrate Active Enzymes

Before the genomic data boom, there was a time when enzymes of different origins displayed little homogeneity, and similarities in sequences were hard to detect. A major development in the analysis of protein sequences and the relationship with structural folds came about in the late 1980s, where a technique called hydrophobic cluster analysis (HCA) [26] was used to bypass the issues related to low sequence homology (quoted as low as 10%).[27] The power of HCA was shown in the classification of a small set of cellulolytic enzymes from different organisms into 6 families based on their amino acid sequence.[27] Even though the sequences came from different organisms, HCA was able to pinpoint conserved regions within sequences based on the clustering of hydrophobic residues. The theory of HCA uses the understanding that during production of an amino acid chain from the ribosome, the polypeptide exists as an unfolded fluctuating α-helix, consisting of 3.6 residues per turn.[28] The amino acid sequence is drawn out as a perfect α-helical cylinder which uses the residues per turn as scale. At some point along the cylinder, the position of the first residue is parallel with another, and the imaginary cylinder is cut along this parallel line producing a 2D 'α-helical net'. The flattening of the cylinder causes separation in 2D space of residues that were part of a helical turn and to mitigate for this the sequence is duplicated and stacked so that analysis of separated residues can be carried out.[28] Certain amino acids are naturally hydrophobic and others hydrophilic. Sequences drawn out in such a way highlight hydrophobic residues which cluster together in different shapes, and such clusters were known to form secondary structure elements within a protein. The method is shown in Figure 12, taken from the original explanation by Gaboriaud *et al.*[28] HCA compares the extent to which two (or more) sequences show a conservation of hydrophobic clustering, which indicates a conservation of structural fold. Interestingly, non conserved clustering of hydrophilic residues does not necessarily reduce fold homogeneity, as differences in hydrophilic regions often corresponds to loops, which are protein features that normally do not confer importance to the overall protein fold.[27]

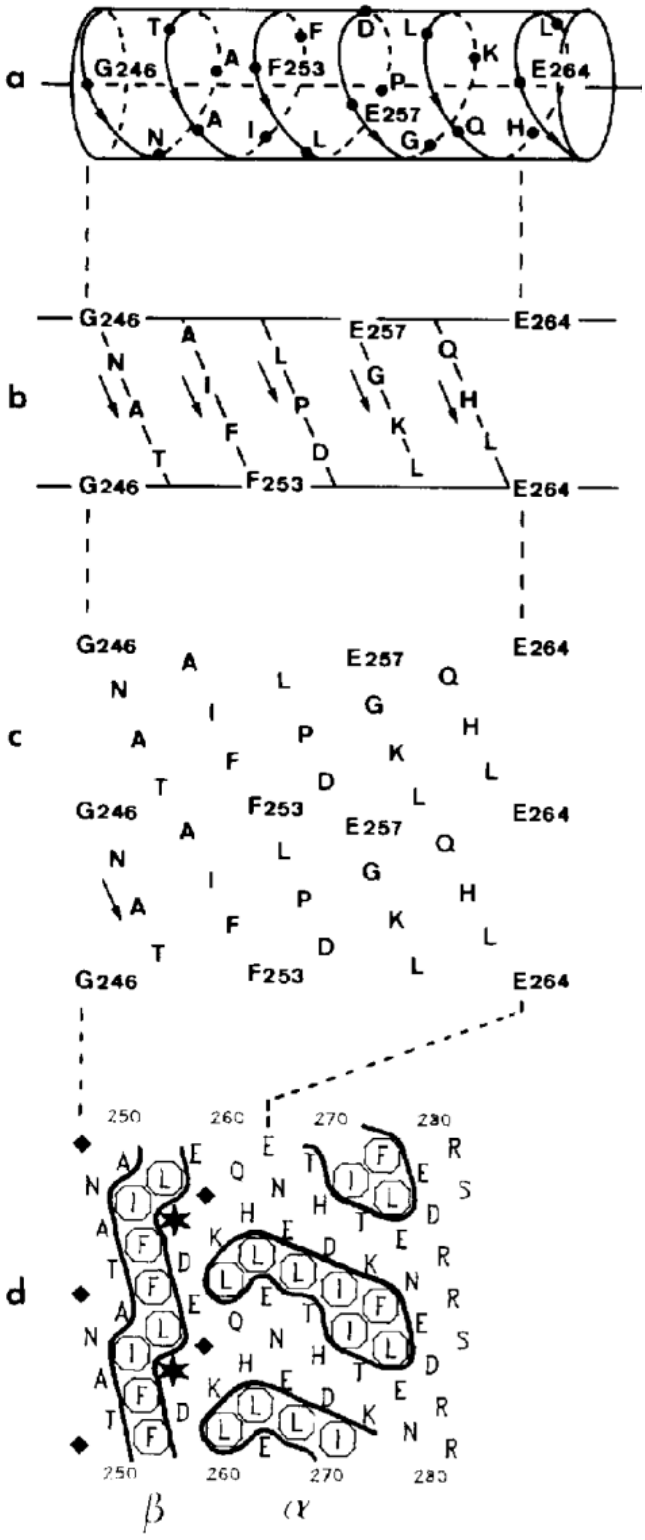


Figure 12 Image taken from Gaboriaud et al, 1987, which gives an example of how HCA works , by firstly representing an amino acid sequence as an α-helix. Secondly cylinder is cut on a parallel line between two residues and unravelled as a 2D net. The net pattern is duplicated one under another to allow spatial representation of residues on the turn as 2D. Hydrophobic residues near each other are circled.

HCA began a new classification system that did not rely on functional studies (which are sometimes variable and as such, not comparable). The first work by Henrissat classified 6 'families' of sequences based on the similarity of their, sequence defined, structural fold (based on a HCA scoring system [28]) and found there to be no similarity between sequence and protein folds of the 6 families. The lack of fold similarity between families suggested that cellulases may adopt many different fold structures and these folds are conserved between similar proteins. The concept was duly extended to the prediction of active site residues and non-catalytic invariant residues, a feature that would be expected to be conserved within a family of similar proteins.[27] Two years later, Henrissat analysed a further 301 sequences thought to be GHs, increasing the number of potential sequence families to 35. The classified families were found to be either monospecific in function or contain multiple known functionalities (polyspecific) as shown by a protein's associated EC number (EC numbers; refer to enzyme type, specific enzyme reaction mechanisms acting on specific substrates).[29]

The occurrence of the same functionality in different fold families reflects the convergent nature of evolution, whereby proteins exhibiting different folds are able to carry out the same function. Likewise the reserve was found, whereby single fold families were found to contain multiple activities suggesting a specific fold is not limited to a single substrate activity. Indeed simple analysis comparing the activity of one enzyme to another could not entertain this extra complexity and HCA analysis blossomed into what we now know as the Carbohydrate Active Enzyme (CAZy) Database. The genomic data boom advanced the classification system, which originally focused on GHs, but now encompasses glycosyl transferases, [30] polysaccharide lyases, carbohydrate esterases, carbohydrate binding modules [31] and auxiliary activities [32] (the latter will be discussed later in this Chapter). Although structural analysis in the form of x-ray crystallography was slow to catch up, the steady stream of 3D structures of GHs emphasised the power of the HCA classification technique. Families were found to contain similar folds as may be expected from the high conservation of sequence observed in HCA. However, it should be noted that HCA is not able to predict the fold of a protein, merely identify sequences which share particular conserved features. Conservation of residues and their relevance to protein function would be a key GH research focus over the years, helped by the family system; if one protein within a family had had its active site residues successfully identified experimentally, the invariant residues could be clearly identified in proteins of the same family.

Successful classification of a protein sequence into a GH fold family did not always occur. In the early study by Henrissat, 10 % of the sequences analysed were found not to have sufficient overlap in hydrophobic clustering, meaning they could not be assigned to a family; the definition of a 'family' requires more than one member. However, sequences which display little sequence similarity to other proteins of known families more likely indicates that sequences identified in the future may align and produce different families with potentially different structural folds and new activities. This concept was shown in the later set of classifications by Henrissat, which added an additional 181 sequences thought to be GHs or related proteins to the previously classified set of 301. Indeed, of the 181 sequences many were classified into the previously defined 35 families, whilst a further 10 new GH families were identified.[33] Cellulase families were historically described with a letter, i.e. Cellulase family A, whereas the current CAZy classification describes families as the enzyme type, abbreviated, followed by a number – for example, glycoside hydrolase family 5 corresponds to GH5.[34]

The expansion of the CAZy database continues to this day, as evermore genomic data becomes available.[31] This has led to a large collection of publicly available data, which is continually curated as an online database.[34-35] As the number of families grew, as predicted, it became clear that unrelated sequence families often shared a type of protein fold. As sequences are classified into families, the families can themselves be grouped into clans; where a clan is a collection of two or more CAZy families that are related in protein fold, but not in amino acid sequence, as well as display the same catalytic machinery and mechanism. These shared features may indicate families whose members share a common evolutionary ancestor.[36] Protein fold is often better conserved than the sequences that form it and several families unrelated by sequence still adopt a specific type of structural fold, as shown in **Table** 1.

Table 1 GH Clans A-Q, their structural fold and containing GH families. Information taken from www.cazy.org accessed August 2018.[34]

Clan	Structural Fold	GH Family
GH-A	$(\beta/\alpha)_8$	1, 2, 5, 10, 17, 26, 30, 25, 29, 42, 50, 51, 53, 59, 72, 79, 89, 113, 128, 147, 148
GH-B	β-jelly roll	7, 16
GH-C	β-jelly roll	11, 12
GH-D	$(\beta/\alpha)_8$	27, 31, 36
GH-E	6-fold β-propeller	33, 34, 83, 93
GH-F	5-fold β-propeller	43, 62
GH-G	$(\alpha/\alpha)_6$	37, 63, 100, 125
GH-H	$(\beta/\alpha)_8$	13, 70, 77
GH-I	α+β	24, 80
GH-J	5-fold β-propeller	32, 68
GH-K	$(\beta/\alpha)_8$	18, 20, 85
GH-L	$(\alpha/\alpha)_6$	15, 65
GH-M	$(\alpha/\alpha)_6$	8, 48
GH-N	β-helix	28, 49
GH-O	$(\alpha/\alpha)_6$	52, 116
GH-P	$(\alpha/\alpha)_6$	127, 146
GH-Q	$(\alpha/\alpha)_6$	94, 149

1.5 Lytic Polysaccharide Monooxygenases

Lytic polysaccharide monooxygenases (LPMO) are a recently discovered class of enzymes capable of using copper dependent oxidative chemistry to break glycosidic bonds within typically recalcitrant polysaccharide substrates such as crystalline cellulose and chitin. Unlike GHs which rely on a nucleophilic substitution process to break glycosidic bonds, LPMOs are able to form powerful copper-oxygen intermediates able to abstract a hydrogen atom from the strong C-H bond (bond dissociation enthalpy ~95–103 kcal mol^{-1}) associated with the anomeric carbon atom. Hydrogen atom abstraction (or hydrogen atom transfer, HAT) is followed by breakage of the C-O glycosidic bond.[37] LPMO families contain individual clades which group those enzymes together with similar substrate preference (i.e. cellulose or chitin, or both) and sequence similarity, which is further subdivided into type based on regioselectivity of cleavage (C1, C4 or either). Towards the end of this chapter **Table** 2 shows how AA9 and AA10 LPMOs are divided based on their clade and type. Enzyme activity is commonly experimentally determined with matrix assisted laser desorption-time of flight mass spectrometry (MALDI-TOF-MS), whereby LPMO attack of the C1 or C4 carbon of the scissile glycosidic bond produces classic and easily identifiable oxidised products, aldonic acids (C1 attack) and ketoaldoses (C4 attack) as shown in **Figure** 13.[38] The active site of LPMOs is strictly conserved in its basic form, the histidine brace; a copper ion bound in T-shaped geometry to 3 nitrogen atoms from a pair of invariant histidine residues. There are currently 6 classes of LPMO (AA9, AA10. AA11, AA13, AA14 and AA15), as shall be described in the forthcoming section, all of which contain this key metal binding site capable of using oxidative chemistry to break apart recalcitrant polysaccharides.[39]

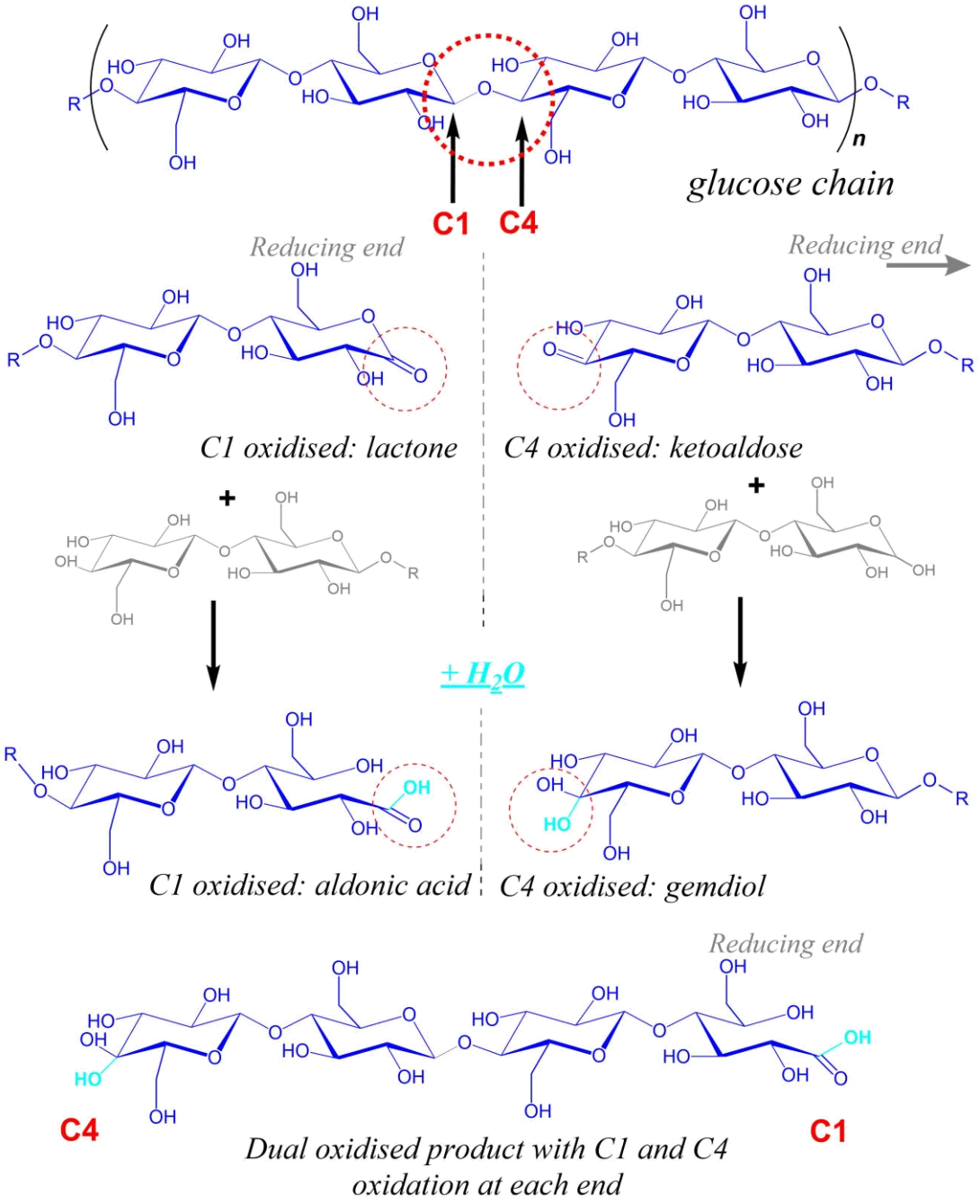


Figure 13 Schematic showing the carbon positions upon which LPMO oxidative attack has been shown to occur. Oxidation at C1 leads to production of a lactone, which readily undergoes water addition to the aldonic acid form. C4 oxidation produces a ketoaldose, which can also undergo water addition to form a gemdiol product. The reducing end of the chain is highlighted throughout. Different LPMO display preferences for C1, C4 or either C1 and C4 oxidation and as such, products containing both aldonic acid and gemdiol can be formed where oxidation has occurred at both ends of a chain.

1.5.1 An important development towards efficient utilisation of biomass

The ability of these enzymes to work on substrates typically out of the realms of activity of traditional GHs has caused an explosion in research into their activity, function and structure since 2010. Industrial processes which work on the breakdown of lignocellulosic material adopted the use of these enzymes very early on in their discovery, following on from reports of their ability to increase efficiency of traditional cellulase enzyme mixtures by two-fold.[40] The cost of enzyme mixtures required during industrial processing of biomaterials is high and limited stocks are available due to the low numbers of companies producing enzymes for saccharification (higher costs, due to lower competition). Despite this, mixtures using thermostable enzymes capable of withstanding the harsh conditions of an industrial plant are in wide usage and improve cost efficiency of biofuel production. LPMOs are another avenue taken by the major enzyme production companies as they are known to improve the efficiency of the enzyme cocktails used during production; they are effective at reducing the recalcitrance of the feedstock, which is the major issue in producing fermentable sugars. Predictions suggest that a boom in development of industrial plants working on producing bio-ethanol from the enzymatic breakdown of waste lignocellulosic feed stocks will occur over the next several years.[41]

Knowledge of LPMOs, and debate on their mechanisms and functionality is ever expanding and has been extensively reviewed over the past several years in order to keep up with the fast paced nature of this field of research: 144 papers are found under the search term 'lytic polysaccharide monooxygenase' from Web of Science's core collection, August 2018, within the 8 years since their initial characterisation. The CAZy database has over 3500 sequences designated as LPMOs as of August 2018.[34, 39] Some reviews which have tracked the progression of research on LPMOs can be found by Horn *et al* (early review of the emerging LPMO field, 2012),[42] Beeson *et al* (overview of LPMOs with a focus on diversity, 2015),[37] Frandsen *et al* (a crystallography perspective, 2016),[43] Meier *et al* (focusing on oxidative cleavage of various polysaccharides by LPMOs, 2018),[44] Frommhagen *et al* (substrate specificities and electron-donating systems of fungal LPMOs are summarized, 2018)[12] and Ciano *et al* (focusing on the histidine brace, 2018).[39]

1.5.2 Discovery of LPMOs – from non-catalytic to oxidative powerhouses!

The concept of breaking a glycosidic bond is synonymous with the idea of a glycoside hydrolase, a protein class efficient in using hydrolysis to degrade plant materials (as discussed in the previous section). The diverse number of different types of GHs and their wide ranging abilities to break down a variety of different polysaccharides and oligosaccharides, working in nature as a coherent whole is an elegant concept. Yet there were flaws in our understanding of nature's methods to breakdown those energy rich substrates that only came to the surface of research in the early 2010s. Importantly, whilst GHs are very efficient in their role as hydrolytic enzymes, they require their specific substrates to follow certain protein binding interactions, namely to dock to protein active site surfaces formed as grooves, clefts or tunnels.

In 1950, Reese and co-workers had envisaged a system in which GHs worked in tandem with an unknown enzymatic entity, to proceed with efficient breakdown of polysaccharides. Reese suggested this unknown factor, termed C_1, acted as a prelude to the degradation carried out by the GHs, or as he wrote the C_X component.[45] In order for GHs to work on substrates that are insoluble, and recalcitrant, there must first be a method of disrupting the ordered, crystalline nature of the substrate. Work carried out towards the end of the last century held hints about the enigmatic C_1 component, which had mostly been put to one side due to conformations by authors such as Wood and McCrae that the C_1-C_X theory was manifested in the synergistic nature of *endo* and *exo* GHs. Authors of note, were those who were investigating fungal cellulose degradation systems such as Raguz *et al* (1992),Schrempf (1995) and Saloheimo *et al* (1997), all of whom encountered proteins linked to cellulose degradation which did not conform to known cellulases of the time, yet were found to be involved in cellulose degradation.[46-48] The notion of an enzyme capable of oxidising cellulose was touched upon in studies by Eriksson *et al* in 1974, whereby it was shown that a mixture of enzymes taken from a cell culture were able to breakdown a cellulose substrate more efficiently in the presence of oxygen. Indeed, the authors discussed the C_1-C_X theory and stated that *"the discovery of the oxidative enzyme reported here makes it clear that it is not only hydrolytic enzymes that are involved in cellulose degradation"*. Despite this assertion, the enzyme studied was not assigned a specific function.[49] Nearly 23 years later, work by Saloheimo *et al* formed the basis of the now re-classified GH family 61, into which LPMOs active on cellulose were initially placed. Their work on *Trichoderma reesei* discovered a new

enzyme capable of 'endoglucanase' activity on cellulose and found it homologous to a protein from *Agaricus bisporus* of unknown function, thus creating a new CAZy family, GH61.[48] Similarly, work carried out around the same time by Schnellmann *et al* was likely the first activity identification of what is *now* known to be a chitin active LPMO, CHB1.[50] Reclassification of the GH61 family into what we now know as auxiliary activity family 9 (AA9) would not take place for several years[51] and would be sparked by analysis of a mysterious enzyme able to break down chitin.

The structure of what is now known as an LPMO was first solved in 2005 by Vaaje-Kolstad *et al*, and described as a chitin binding domain, CBP21. Importantly the chitin binding domain, whilst found to be essential to the function of *Serratia marcescens* in breaking down chitin, was determined to be 'non-catalytic', yet able to bind "*to the insoluble crystalline substrate, leading to structural changes in the substrate and increased substrate accessibility"*.[52] Several proteins with similar overall structures had also been identified and classified as chitin binding domain family 33 (CBM33), and sequence alignments were used to try to establish which residues may be important for chitin binding. A group of tryptophan residues within the protein sequence were hypothesised to cause chitin binding through aromatic interactions, but this was clearly disproved during structural analysis, which found them buried within the core of the structure.[53] Instead, Vaaje-Kolstad *et al* noted a patch of hydrophilic residues on a flat surface of the protein, Figure 14, which when mutated, reduced overall activity; mutation of residues Try54, Glu55, Glu60, His114 and Asp182 to alanine reduced the amount of products (chitobiose) formed during a combined assay with a chitinase, ChiC, to a similar level as if no CBP21 had been added.[52] Despite the observation of a conserved patch of residues (which included the histidine brace), the second feature that was found lacking was an active site. Naturally, the structure of CBP21 was being compared against canonical GHs and found not to conform to the structural expectations of that class of proteins: no groove, cleft or tunnel to accommodate a polysaccharide chain, no conserved Asp and Glu potential catalytic GH residues. As such it had been identified as more similar to non-catalytic carbohydrate binding domains (CBMs). The authors of this first structural insight into LPMOs therefore defined the protein as non-catalytic as it lacked any surface structure which could accommodate a chitin substrate in a typical hydrolytic fashion, and indeed the authors mention the lack of co-crystallisation with any soluble chitin oligosaccharides as evidence.[52-53] With hindsight, one can look back at these now superseded papers and easily identify the iconic N-terminal Histidine residue that forms the basis of an LPMO active site in the sequence alignments shown of homologous proteins in

the designated chitin CBM33 family.[53] Despite classification as non-catalytic it was becoming apparent, that this was a newly forming family of proteins able to disrupt the structure of insoluble, recalcitrant chitin, an idea that satisfies the requirements set by Reese *et al* for his efficient polysaccharide degradation system.[45] Further evidence for the assistive power of CBM33 proteins came when Moser *et al* in 2007 described two proteins from *Thermobifida fusca,* E7 and E8 as secreted when the source organism was grown on polysaccharide substrates and able to improve degradation of substrates by GHs in activity tests. Moser found E7 bound strongly to β-chitin, whilst E8 showed a higher binding preference for crystalline cellulose. Both proteins were shown to improve the activity of *T. fusca* cellulases on degradation of filter paper and on the activity of Chitinase C (*Serratia marscescens*) on β-chitin (evidence which supported the activity shown by Vaaje-Kolstad *et al* by CBP21 and its hydrophilic surface mutants in 2005).[54] Whilst the activity enhancements were becoming apparent it wasn't until further work in 2010 carried out by the same group, did it become known that the assistive degradation was in fact catalytic, and importantly relied on an oxidative mechanism.[38]

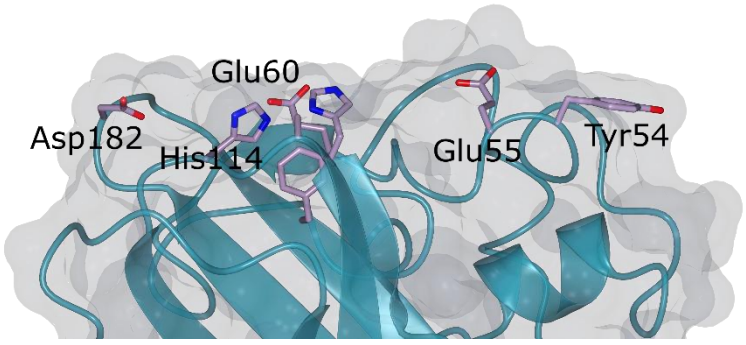


Figure 14 The conserved surface residues observed on CBP21 , which were mutated to Ala in the study by Vaaje-Kolstad *et al*, and found to significantly affect the function of chitin degrading ChiC used in the synergism assay.[53] Image produced in CCP4mg.[55]

Whilst there was steadily mounting evidence to suggest that the CBM33 proteins were able to improve polysaccharide degradation by canonical GHs, similar work was being carried out with proteins classified to family GH61. Before conformation of LPMO oxidative activity by Vaaje-Kolstad *et al*, two key studies were able to identify a patch on the flat planar surface of two cellulose active enzymes from family GH61, which were most importantly conserved and found to accommodate a metal ion. Firstly, in 2008 Karkehabadi *et al* produced the first crystal structure of a protein from the enigmatic GH61 family, where only weak endoglucanase activity had ever been detected.[48, 56] The structure of Cel61B from *Hypocrea jecorina* was solved in a condition rich in nickel ions, and as such the abundance of nickel in

the structure was used as the source of anomalous scatter, enabling structure solution without the need for a known model. The structure was shown to have a core consisting of a twisted β-sandwich, with a distinct lack of a typical polysaccharide binding site. The authors describe for the first time, the conservation of a surface region of the protein containing polar amino acids, found in the sequences of all GH61 members and forming the basis for a nickel ion coordination site, as observed in their structure solution. Most importantly, the authors draw attention to the fact that the closest structural homology of Cel61B was that of CPB21, which shared the highly conserved polar surface region (the histidine brace) as well are propensity for conserved hydrophilic surface resides [56], which in CBP21 were found to interact with chitin as described previously, **Figure** 15.[52-53] The authors note the unusual occurrence of the bound nickel ion, and do recognise the notion that this may not be the natively bound ion within the coordination site, whilst holding back on their speculations that the metal site may be involved in some sort of catalytic function.[56] Secondly, in 2010, Harris *et al* undertook a study looking into the boosting effect of several GH61 proteins in the presence of different metal cations (but not copper) towards the degradation of cellulose. The metals probed in this study were found to have an appreciable effect on the degree to which cellulose assistive degradation was achieved, but the major highlight of this study comes from their structural mutant analysis. Mutation of the N-terminal His to Asn or the mid-chain histidine to Ala of one GH61 protein was found to abolish catalytic function in both mutants, further asserting the hypothesis that this conserved surface patch, containing a metal ion was imperative to the function of the protein.[40]

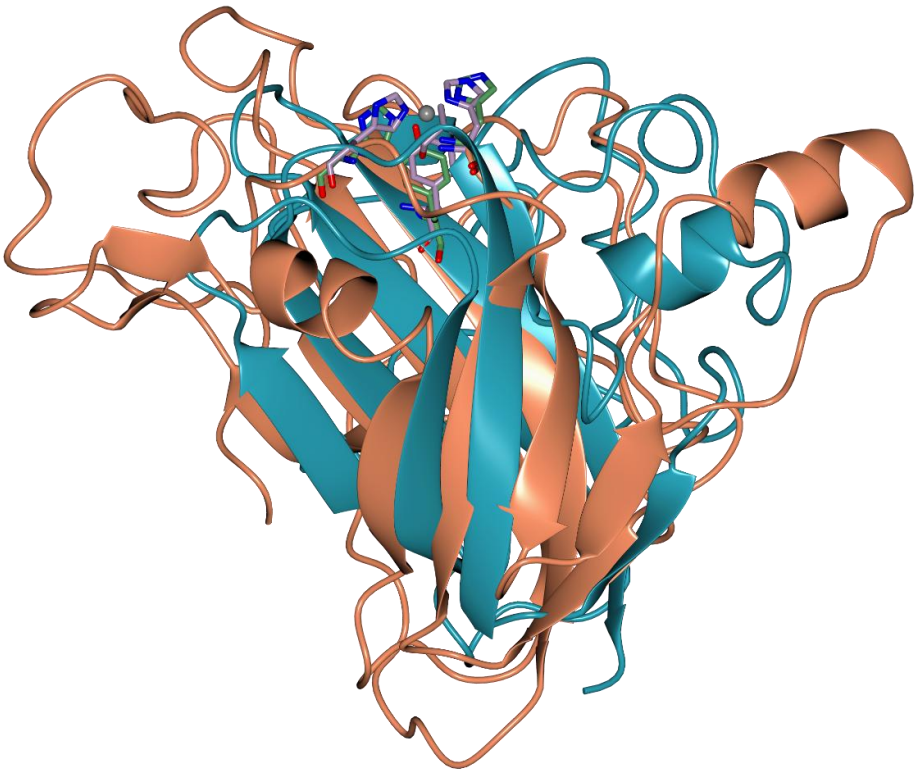


Figure 15 Structural comparison of CBP21 (dark cyan ribbon, dark green cylinders, PDB code; 2BEM) and Cel61B (coral ribbon, lilac cylinders, PDB code; 2VTC) , with the histidine brace shown as cylinders. A Ni atom was modelled into the histidine brace of Cel61B, as shown in grey. Image produced in CCP4mg.[55]

Shortly after the study by Harris et al, the oxidative function of CBP21 was finally realised by Vaaje-Kolstad *et al*.[38] Mass spectrometry was used to analyse the products of chitin degradation this time using a method in which a chitin deacetylase was used to solubilise the longer oligomers produced during degradation reactions. It was observed the CPB21 did not act like a random *endo*-GH, but produced even numbered products that one might associated with specific binding to one side of the chitin substrate, specifically a crystalline region. The use of water labelled with ^{18}O identified that in the aldonic acid products of chitin degradation, only one oxygen was due to the presence of water. Further tests in which the air in the assays was replaced with nitrogen significantly reduced enzyme function, indicating that the second oxygen was due to reaction with molecular oxygen. As such, Vaaje-Kolstad *et al* suggested that CBP21 was carrying out both a hydrolytic step and an oxidative step.[38] One can now easily rationalise the incorporation of the two oxygen atoms from the different sources, where the oxidative step uses molecular oxygen during catalysis with the copper ion in the active site and the hydrolytic step is simply the change

from the initial lactone product to the stable aldonic acid through the interaction with water, Figure 16.

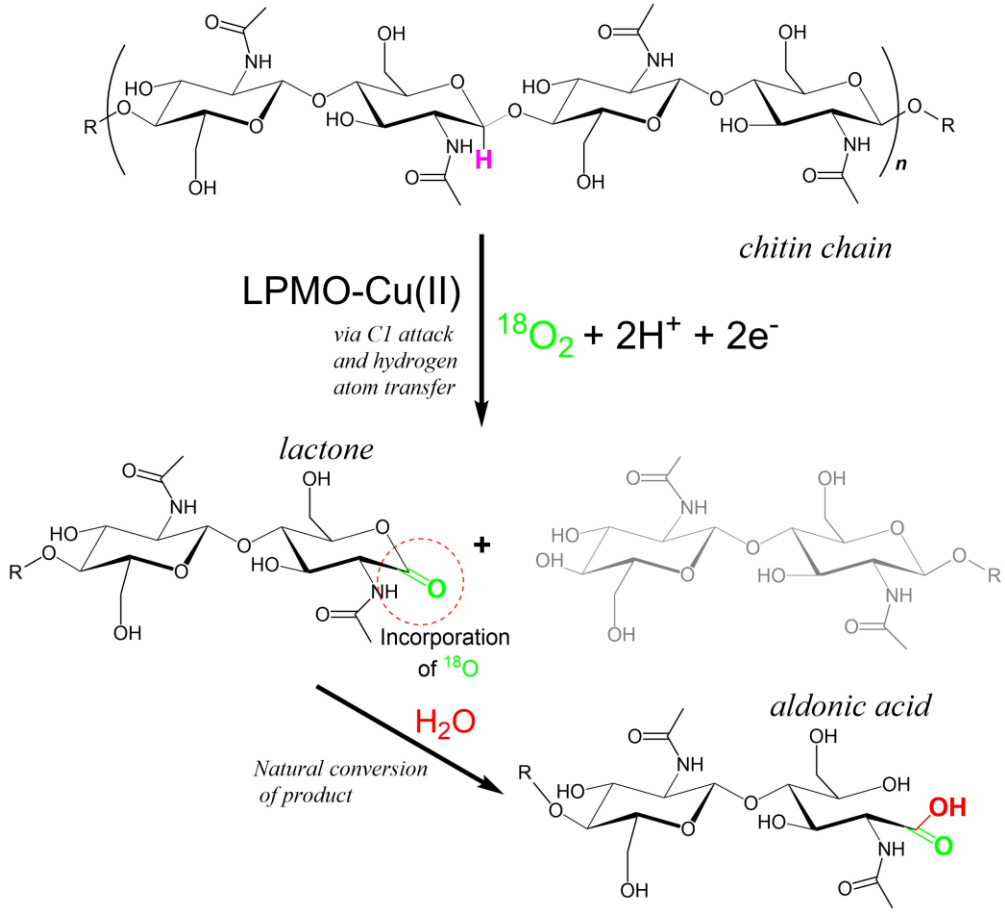


Figure 16 Diagram of LPMO mediated incorporation of ^{18}O from labelled molecular oxygen gas into the aldonic acid product of CBP21 degraded chitin, as observed in experiments by Vaaje-Kolstad et al.[38]

The ambiguity about the bound metal ion was cleared up in foundational work carried out by Quinlan *et al*.[57] This study looked in great detail at the active site of a GH61 from *Thermoascus aurantiacus* and was first to clearly define the catalytic properties of a GH61 family protein; whereby the proteins were described at lytic polysaccharide monooxygenases (LPMO), able to break the glycosidic bonds of crystalline cellulose using an oxidative mechanism based on copper. Firstly, they perform an analysis of the potential sources of electrons from PSC. Vaaje-Kolstad had remained ambiguous regarding the need for a co-factor to support oxidative function of CBP21, whilst Harris suggested that the heterogeneous mixture of pre-treated corn stover (PCS) used during his study on GH61 proteins provided a source of electrons.[38, 40] Quinlan looked at a range of soluble products within the PCS mixture and found that gallate (a reducing agent) could be utilised by the *Ta*GH61 as a redox partner during breakage of glycosidic bonds, whilst not affecting the cellulose itself. Study of fungal LPMOs, as they were becoming known, by Phillips *et al* found that it wasn't just small molecule reductants that were able to act as reducing partners; a GH61 fungal protein was found to be able to accept electrons from cellobiose dehydrogenase.[58] However, no protein electron partners have been identified for historically CBM33 LPMOs as yet. Future studies would also identify lignin as a potential source of electrons, with LPMO activity on cellulose linked to the removal of unpaired electrons from lignin [59] and with assays, regarding LPMOs active on hemicelluloses, carried out on complete wood samples were found to continue to offer a boosting effect even without the addition of a reducing agent or protein electron donor.[60]

Secondly, Quinlan was able to clarify the debate regarding coordination of a metal ion within the conserved patch of surface polar residues. Where previous studies had assigned the coordinated metal ion based on environmental factors such as crystallisation conditions, or saturation of the protein with different divalent cations, this study analysed *Ta*GH61 using isothermal titration calorimetry (ITC) to assess the binding affinity of various metal ions. Despite experiments and crystal structures indicating otherwise, the binding of divalent metals, Mg^{2+}, Ca^{2+}, Mn^{2+}, Co^{2+}, Ni^{2+} and Zn^{2+} to the conserved metal site was not detected. Instead a clear 1:1 binding ratio of *Ta*GH61 with Cu^{2+} was observed, with binding so tight that calculating the dissociation constant was beyond the realms of experimental determination by classical ITC. The authors go on to suggest a practical problem noticed by all who now work with LPMO enzymes, they are so adept at binding copper that the proteins are able to scavenge it from their environment. Even introduction of EDTA (a strongly metal coordinating ligand) lead to the slow removal of copper, and it is likely that all previous

studies carried out on enzymes such as CBP21 and CelB61 already had copper bound tightly in their active sites, despite 'saturation' with other metal cations.[57] NMR studies went on to support the assignment of copper in the binding site of LPMOs, with substrate binding studies showing that the planar surface, including the copper coordination site was used in substrate binding.[61]

1.5.3 The Histidine Brace

The strong binding of copper in *Ta*GH61 was further assessed by Quinlan *et al* using a technique little utilised by biochemical research; electron paramagnetic resonance (EPR). EPR is a technique which uses an atom's unpaired electrons and how they relax after the influence of microwave radiation within a strong magnetic field to produce a spectroscopic interpretation of the atom's coordination environment. Copper(II) is a prime candidate for EPR analysis due to its single unpaired electron. As such, the coordination of the copper ion within the conserved site of *Ta*GH61, coined by the authors at the 'Histidine brace' was interpreted in great detail and found to be similar to particulate methane mono-oxygenases (MMO) (although at the time, MMOs were thought to have two Cu ions in their active site, a theory that has now been revised) [62-63].

Crystallography had already managed to observe a metal ion forming a coordination site with N-terminal histidine and a mid-chain histidine. The advantage of using EPR is clear in the analysis of LPMOs, as it captures the coordination site of the copper in pure spectroscopic detail, supporting crystallographic analysis with chemical interpretations, as shown in **Figure** 17, taken from Quinlan *at el*.[57] EPR analysis of *Ta*GH61 which had been fully saturated with copper produced spectra that represented a copper ion in a tetragonal coordination geometry. The copper species bound within the proteins conserved active site was found to produce spectra similar to those of other copper oxidases, which contain a classical type 2 copper species.[64] This result was supported by crystallographic analysis, where a copper ion was modelled into what the authors coined the 'histidine brace'; the copper was shown coordinating to three nitrogen atoms, two from the N-terminal histidine (one ring nitrogen and the nitrogen of the amino terminus) and a third from the ring of the mid-chain histidine.[57] It is this structural conformation of two histidine residues forming a coordination sphere for the copper ion that is present in all LPMOs. The EPR analysis showed that the copper was a type 2 species which supported the idea that oxidative chemistry [38] would occur, as other enzymes with similarly coordinated copper ions were known to be able to use molecular oxygen and insert it into substrates.[64]

The similarities between enzymes placed in the original families GH61 and CBM33 were becoming more apparent especially in terms of structure and function; whilst the primary scaffold of the histidine brace was consistently present, changes in the other coordinating ligands were found to be indicative of the type of LPMO being studied. Quinlan *et al* had observed T-shaped geometry (distorted square planar) around the copper site of *Ta*GH61, defining the copper species as type II. Analysis of a CBM33 protein from *Bacillus amyloliquefaciens* followed shortly afterwards by the same group, and EPR studies showed significant differences within the copper coordination. They describe the copper coordination in *Ba*CBM33 as *"significantly distorted from axial, such that it lies somewhat between the usual Peisach–Blumberg type 1 and type 2 classifications".* The positioning of two non-coordinating conserved residues in the active site of *Ba*CMB33, an alanine and phenylalanine in the apical ligand positions allows the copper ion to only coordinate with the ligands in the equatorial positions.[65] The differences in coordination sphere between the two types of LPMO are clear, and likely arise because of the differences in steric/electronic environments in the secondary coordination sphere of the histidine brace.

Lastly, an obvious difference was first observed on the N-terminal histidine of *Ta*GH61, it was methylated (see **Figure** 17, A and C).[57] This post-translation modification is found in all LPMOs of fungal origin and the instance of such a modification remains mysterious; fungal LPMO proteins produced using *E. coli* lack this modification, yet still function.

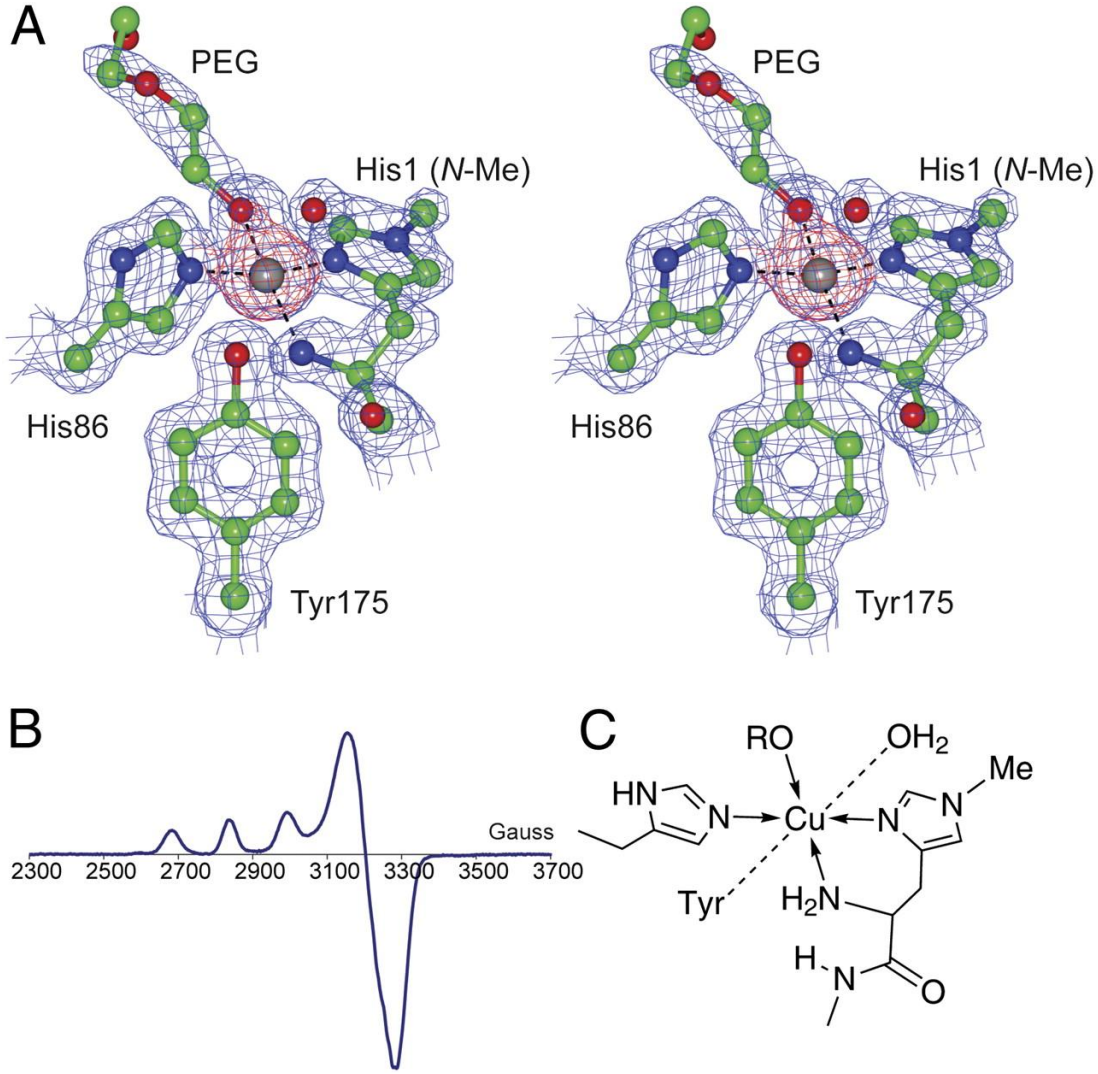


Figure 17 Image taken from Quinlan *et al* [57] showing the clear evidence coming from EPR and crystallography to support the binding of copper into the *Ta*GH61 histidine brace. A) shows the copper ion modelled into the electron density of copper-saturated *Ta*GH61. B) shows the EPR spectra obtained from TaGH61, with clear peaks that are similar to other type II copper species and C) which shows a representation of the coordination sphere of the copper ion in the histidine brace, with main tetragonal coordination coming from the three nitrogen atoms and a PEG molecule. A water molecule and a tyrosine residue are in the apical binding positions.

1.5.4 LPMO Reclassification and Expansion

As researchers began to focus more on these intriguing enzymes it became clear that the previous CAZy classifications were incorrect; proteins established as belonging to GH61 were definitely not glycoside hydrolases, and those proteins in family CBM33 were not just chitin binding modules. Thus a new family was created in the CAZy database to house LPMOs (among other redox active enzymes) called auxiliary activities. GH61 was renamed as AA9 whilst CBM33 was renamed AA10.[32] The reclassification came at a convenient time, for other LPMO families were and are still being discovered; namely AA11, AA13, AA14 and

AA15 which shall be discussed briefly below. All currently known LPMO families are summarised in Table 2, at the end of this section.

1.5.4.1 AA11

The lack of sequence homology between proteins found in families AA9 and AA10 means that when one conducts a sequence search, a protein from AA9 will not yield a result from family AA10. Clever methods needed to be employed to search for non-sequence homologous LPMOs. One such method, known as 'module walking' was used by Hemsworth *et al* to discover proteins that would become known as LPMO AA11's.[66] The technique involved identifying domains that were found appended to certain LPMOs and searching for these domains using NCBI BLAST. LPMOs are sometimes appended with CBMs and other domains of unknown function, termed X-domains. Hemsworth *et al* found binding of a specific X-domain to a GH18 protein, which itself was attached to a domain of unknown function – whose sequence started with a signal peptide followed by an N-terminal histidine residue. This AA11 from *Aspergillus oryzae* was found to be active on squid pen chitin. The structure of *Ao*AA11 shares the common LPMO fold, with a core β-sandwich and flat binding surface, Figure 18. Notably, as this fungal LPMO gene was produced in *E. coli* it lacked the post-translation N-terminal histidine methylation (*E.coli* is not capable of carrying out this modification), it is not yet determined what the effect of the methylation of the N-terminal histidine is.[66] Analysis of the copper coordination geometry proved surprising, *Ao*AA11 was somewhere in-between the geometry posed by AA9 and AA10 enzymes. Like in AA9s, a non-coordinating tyrosine residue was found in the lower apical coordination site. However, like in AA10s, an alanine residue was found in the top apical position, preventing coordination of water as in AA9s.

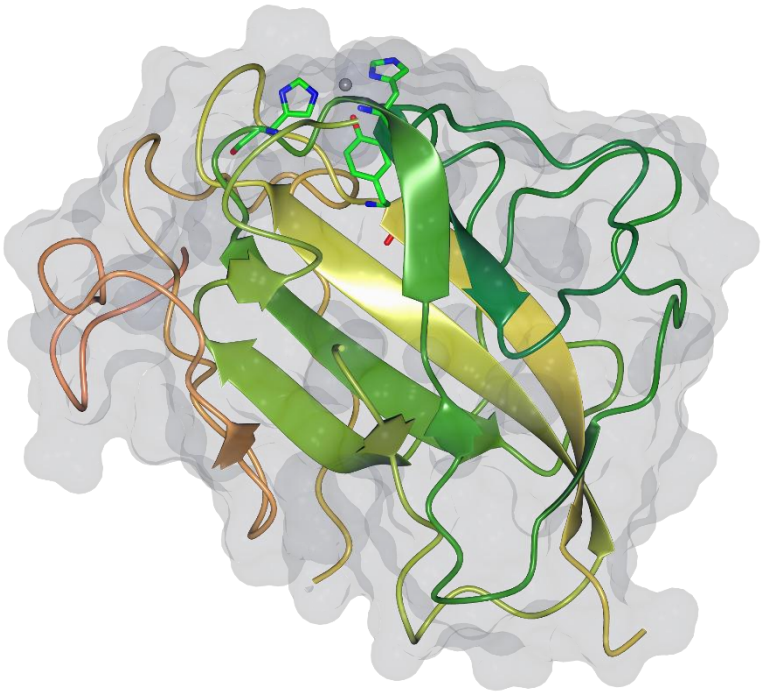


Figure 18 Structure of *Ao*AA11 (PDB code, 4MAI) . Image produced in CCP4mg.[55]

1.5.4.2 AA13

The AA13 family was found using a different bioinformatics method, which relied on the knowledge that all known fungal LPMOs were secreted. Vu *et al* looked at secreted proteins in the genome of *Neurospora crassa*, identifying those with a signal peptide followed by a strictly conserved N-terminal histidine (when compared with homologues). Other defining features, such as the second copper coordinating histidine and a known sequence motif were used to further identify the proteins as LPMOs. One of the newly identified LPMO sequences, was appended to a CBM20 domain, which is known to bind starch. Further characterisation proved that the LPMO was able to break down starch using the same oxidative chemistry as required for cellulose and chitin degradation by AA9, AA10 and AA11 LPMOs.[67] Shortly after and using a similar method, Lo Leggio *et al* identified another starch degrading LPMO in the fungi *Aspergilllus nidulans.*[68] The activity of AA13 enzymes on starch shows they are capable of breaking α-1-4 glycosidic bonds, a feat not observed before within LPMO activity, where chitin and cellulose are made up of β-1-4 glycosidic bonds. However, an AA13 found in *Aspergillus oryzae* was found not to degrade starch, which was

hypothesised to be due to the lack of the C-terminal CBM20 domain. [68] Indeed, a large majority of AA13 enzymes possess a C-terminal CBM20 domain, which may be required for activity of the LPMO on starch.[69] One of the most interesting aspects of AA13 is those with corresponding 3D structures show an alternation to the normally planar LPMO surface. Whilst some LPMOs have been observed to have protrusions, or binding surfaces with concave features, AA13 LPMOs have been shown to have a shallow groove. The authors suggest that this interesting adaption may be in response to the helical nature of the substrate amylopectin. However, it would be interesting to see how many AA13 proteins which are not associated with a starch binding domain contain this adaption.[68]

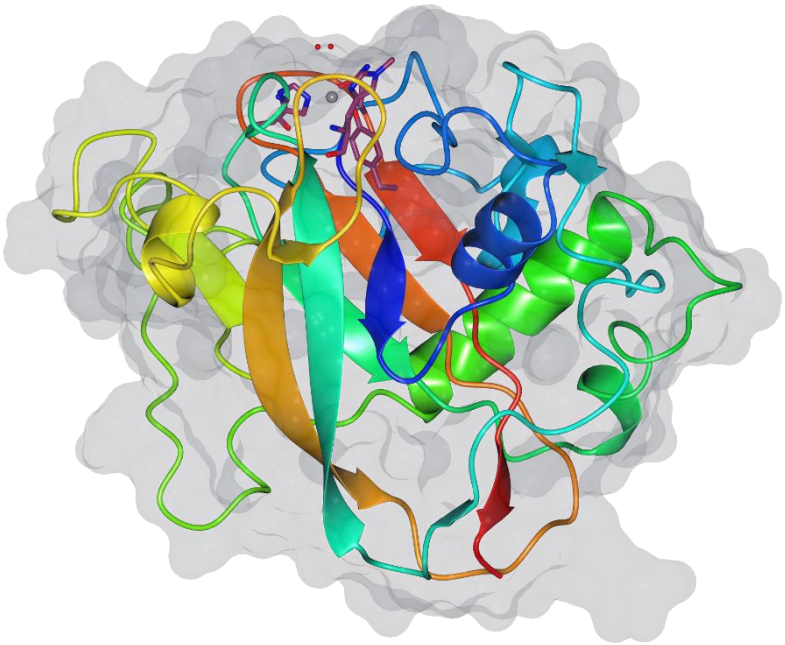


Figure 19 Structure of AoAA13 (PDB, 4OPB). Image made in CCP4mg.[55]

1.5.4.3 AA14

LPMOs do not always have C-terminal domains, so module walking approaches do not apply in all cases, as shown by the discovery of the AA14 family of xylan degrading LPMOs, **Figure** 20. The degradation of woody biomass is carried out in nature effectively by white and brown rot fungi. Couturier *et al* were investing the wood degrading ability of *Pycnoporus coccineus*, a white rot fungus, by monitoring the up-regulation of genes in response to different woody substrates.[60] A protein secreted during growth on pine and poplar was compared to homologous proteins using BLAST and all were found to display the classical N-terminal histidine residue. The authors showed a phylogenetic analysis that identified the

homologous proteins as clustering together into a family, made up of sequences from both white and brown rot fungi and was thus classified as AA14. Structural analysis by EPR showed that *Pc*AA14A and *Pc*AA14B contained strictly type 2 copper species, as seen in AA9 LPMOs, but both were found unable to degrade cellulose. Their activity was instead focused on xylans, and importantly those xylan chains which were interacting with cellulose, the so-called hemicellulose fraction of woody biomass. Transmission electron microscopy and NMR analysis observed an enzymatic effect only on the xylans bound to the underlying cellulose structure. Interestingly, the AA14 enzymes were not active on soluble xylans, to which the authors suggest is due to the three-fold helical screw formed by the solubilised substrate. Instead it is suggested that the binding of the xylan chain in a parallel fashion orientates the xylan in such a way that it is susceptible to oxidative cleavage in the same way as cellulose.[60]

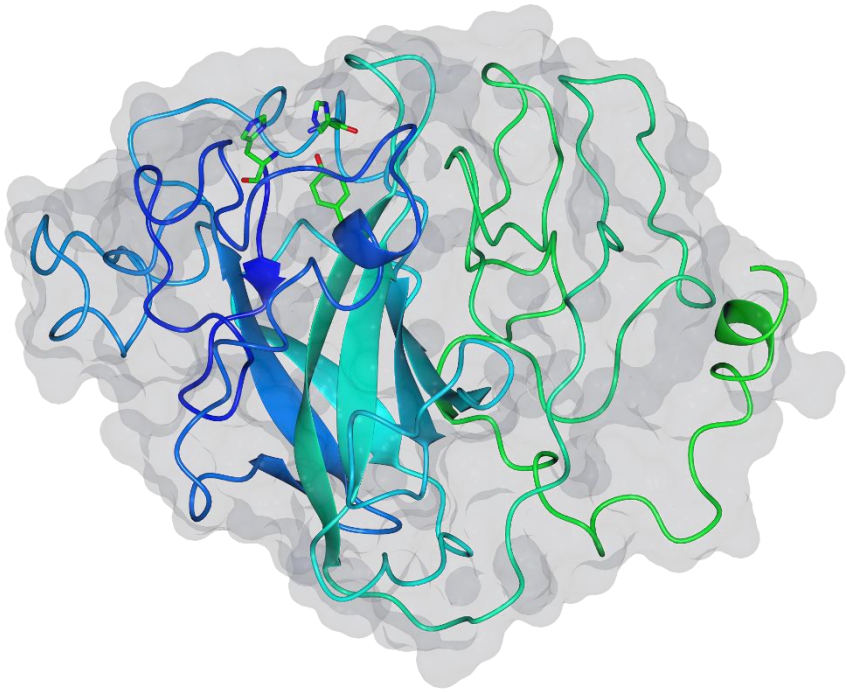


Figure 20 Structure of *Pc*AA14 (PDB code, 5NO7) in apo form, meaning the protein structure lacks copper in the active site. Image made in CCP4mg.[55]

1.5.4.4 AA15

The most recent addition to the LPMO family is AA15, which contains a selection of proteins found through analysis of the genome of the *firebrat* insect, *Thermobia domestica,* Figure 21.[70] This primitive and ancient insect is able to degrade crystalline cellulose without the need for a digestive microbial community, much like the shipworm as will be discussed shortly. Identification was carried out by analysing the gut contents of animals feed on Avicel, and of the identified CAZymes, 20.2% were attributed to being putative LPMOs. Comparison of the sequences identified using transcriptomics, BLAST searching lead to the identification of a multitude of homologues from a wide array of organisms – marine and terrestrial insects, algae and oomycetes. Uncommonly, a sequence homology link was observed to other LPMOs and the group of proteins was classified as AA15, LPMOs active on cellulose.[70]

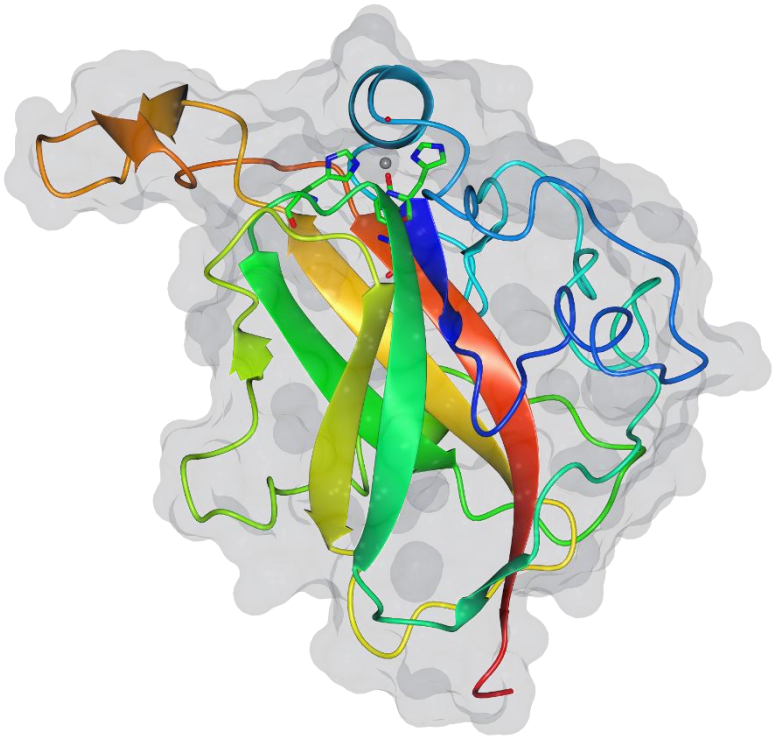


Figure 21 Structure of *Td*AA15 (PDB code 5MSZ) , showing a change from the normal rounded globular structure of other LPMOs. Image made in CCP4mg.[55]

Table 2 Overview of the 6 LPMO classes currently known. A more detailed analysis can be found in Ciano *et al*, supplementary spreadsheet.[39]

AA Family	Bacterial /Fungal	Structures of different species	Clade	Sub group*	Cleavage Site	Substrate	Cu Inner Sphere (other than Histidines)	Copper Type	Points of Interest*
AA9	Fungal	13	-	Type 1	C1	Cellulose	Apical: Tyr, H_20	I	N-terminal methylation , elongated square bipyramid geometry
			-	Type 2	C4	Cellulose			
			-	Type 3	C1/C4	Cellulose			
AA10	Bacterial	12	I	Type 1A	C1	Chitin	Apical: Phe, Ala	II	Trigonal bipyramid geometry
	Viral	3	I	Type 1B	C1	Chitin			Distorted square pyramidal geometry
			II	Type 2	C1	Cellulose			See-saw geometry
			II	Type 3	C1/C4	Cellulose	Apical: Tyr, Ala		Elongated square bipyramid geometry
AA11	Fungal	1	-	-	C1	Chitin	Apical: Tyr, Ala	Between I and II	Active site a mixture of AA9/AA10 elongated square bipyramid geometry
AA13	Fungal	1	-	-	C1	Starch	Apical: Tyr, Gly	II	Grooved binding site, elongated square bipyramid geometry Recombinant protein, no N-terminal methylation
AA14	Fungal	1	-	-	C1	Xylan attached to cellulose	Apical: Tyr, H_20	II	Recombinant protein, no N-terminal methylation
AA15	Invertebrates, alage, oomcytes, etc.	1	-	-	C1	Cellulose Chitin	Apical: Tyr, ?	II (distorted)	No N-terminal methylation

*Subgroup information and geometry taken from Vu *et al*.[71]

1.5.5 LPMO – A fast moving research focus

Since their discovery, LPMOs have had a major impact on the field of lignocellulosic breakdown, and research into these oxidative enzymes is fast paced and still gaining momentum as more and more research groups study every aspect of their behaviour and function. A key area of this research is how the oxidative mechanism actually works; how does the enzyme use copper to catalyse the breakdown of a glycosidic bond. There are other enzymes in nature that use metal catalysed oxidative chemistry. In what seemingly cannot just be a fascinating coincidence, particulate methane monooxygenases also play host to a copper ion, housed within a histidine brace capable of oxidative chemistry.[62-63, 72-73] As such, one may assume that nature has produced a copper coordination complex that is powerful in producing different activities in different scenarios. The mechanism of action is a widely debated topic and is reviewed well in Walton and Davies (2016)[74], Vu and Ngo (2018)[71] and Ciano *et al* (2018).[39]

In an early study, Beeson *et al* used isotopic labelling to determine that an AA9, from *Neurospora crassa*, was able to facilitate hydrogen atom abstraction at either the C1 or C4 carbon atom positions of the glycosidic bond, which was then followed by cleavage of the joining glycosidic C-O bond (carbon at either position).[37] As has been noted in many crystal structures, the copper ion is almost always in the Cu(I) state, most likely due to reduction of any Cu(II) form to Cu(I) during x-ray exposure. The Cu(I) state is thought to be more stable than Cu(II), an idea supported by electrochemical studies of the reduction potential of the enzymes in solution. As such, most catalytic mechanisms that have been suggested (of which there are many) centre on the copper being the (I) state at rest.[74] Combination of experimental observations, orbital theory and density functional theory (DFT) have produced several possible mechanisms for the copper oxidative attack on polysaccharides. LPMOs are a remarkable example of how different fields of study can come together to support each other in the understanding of enzyme action. For example, in studies by Kim *et al* [75] and Kjaergaard *et al* [76] both were able to produce DFT models of LPMO active sites which matched the copper active site geometry and observed oxidation states seen in crystal structures.[74] There are several proposed mechanisms of action, or the LPMO 'catalytic cycles', each one with their own merit and potential issues. Walton and Davies, review these potential cycles, which can be split into two routes; catalysis in which molecular oxygen binds before the substrate, and catalysis in which molecular oxygen binds after the substrate. A multitude of copper-oxygen intermediates (such as copper-superoxide

and copper-oxyl species) are suggested amid various different catalytic routes. Unresolved issues remain that confound and complicate these proposed mechanisms, as described in Walton and Davies[74], including but not limited to; the variety of substrate specificity and differences between substrates (such as the strength of the C-H bond from which abstraction takes place) and how this may affect mechanism, whether it is possible to spectroscopically observe a copper intermediate state (which would undoubtedly assist in the conformation of certain aspects of mechanism), the lack of available kinetic measurements (methods are emerging which can measure LPMO rates even when the substrate is insoluble), understanding the binding nature of LPMOs to the substrate (knowing whether the enzyme is processive, remaining bound to the substrate is a key factor in some proposed mechanisms, and what effect a CBM has on LPMO action), and the effect of the reducing agent (fungal AA9s are able to gain electrons from CDH, whereas there is no known protein based equivalent for other classes of LPMO).[74] Elucidating the mechanism of LPMOs will prove a fundamental step in further utilising their biomass degrading power, be it in the form of engineered artificial enzymes to tailor substrate specificity, or make a one-substrate-fits-all approach, to the creation of small molecule mimics which could potentially be made in bulk, cutting out the need for high cost enzyme production.

1.6 Mimicking Nature

The ability of an organism to successfully breakdown a variety of polysaccharide substrates relies on the combined usage of many different CAZymes. The boosting effect provided when an LPMO is used in addition to traditional GHs on the breakdown of polysaccharides is well known[42], and is a mixture already employed in biorefineries worldwide. Depending on the source biomass material, there is likely to be a difference in structural composition. As such, to mimic Nature, we must also use a variety of enzymes to efficiently break down biomass materials. At present, the 'gold standard' in producing industrial enzymes for bio-processing is *Trichoderma reesei*; a fungus identified during World War II, found degrading the cellulose fibres that made up US army tents on the Soloman Islands, reviewed by Bischof *et al*.[77] Over time, this model organism has undergone many changes to become what is known as a 'hyperproducer', and those strains used to produce proteins in industry are capable of making over 100 g L^{-1} of protein. Indeed, Bischof quoted in 2016 that of the 480.5 million liters per year of ethanol produced from cellulosic biofuel production, 80 % of this was due to enzymes produced by *T. reesei*.[77] Despite sounding as if the industrial process of

lignocellulose breakdown into usable products is firmly taken care of by *T. reesei*, there are still many improvements to be made. Enzymes need to work in a synergistic manor, have high activities, be stable over the course of their usage, be resistant to inhibition by end products and be able to operate under the same conditions as one another. There is also the fact that lignocellulose is a very heterogeneous mixture and requires a large number of enzymes with different substrate specificities to effectively breakdown a type of biomass. As such, there are always more enzymes to be found, a practise known as 'bioprospecting', whereby new and superior enzymes which could potentially improve these industrial 'enzyme cocktails' are searched for.[78] As many organisms have evolved to use specific substrates as a food source, their genomes provide a treasure chest of possibilities that could potentially be used in an industrial process. Organisms that survive in different ecological niches can be a source of new protein-industrial advances in terms of improving the biomass breakdown; be it enzymes capable of tolerating higher temperatures, those with high levels of stability or longevity, or those with usual substrate specificity that allows for the degradation of particularly difficult polysaccharides. With this thought in mind, the work described herein discusses an organism which has evolved to survive in a very specific niche. The organism of interest, the shipworm, was chosen due to its combined choice of habitat and food source; the marine animals burrow into submerged wooden substrates and use the material extruded during burrowing as a food source.[79] Success within this very specific ecological niche would suggest that the animal has a set of CAZymes capable of efficiently providing enough nutrients for survival from a substrate displaying high levels of recalcitrance. This work uses the genome of a symbiotic bacterium, *Teredinibacter turnerae,* found within the gill region of shipworms, as a potential source for new and novel CAZymes.

1.7 A question of symbiosis

Throughout the 'Tree of Life' there exists a multitude of symbiotic associations in which host and symbiont may rely on one another to provide them with key elements of survival.[80] The variations in the type of symbiotic relationships are vast and range from specific adaptations (such as the algae growth on the fur of sloths to provide camouflage) to more general harmonious communities of bacteria within a host digestive system. Ruminants are an excellent example of this, where the digestion of nutrient poor grass is assisted by a vast array of bacteria contained within the digestive system of the animals, enabling access to a wider source of nutrients such as carbon sources from the diet and nitrogen from bacterial fixation. The development of bacterial relationships within terrestrial animals is thought to

provide the symbionts with a relatively stable and importantly, 'aquatic' environment in which they may exist. Without habitation within terrestrial organisms a large proportion of bacteria would be limited to niche areas such as lakes or rivers. Thus, co-habitation may represent a necessary shift in bacterial evolution as aquatic animals moved towards terrestrial environments.[81] In aquatic environments one may then ask the question, why is there a need for symbiosis between bacteria and aquatic animals? Symbiosis may be considered a choice of the bacteria, as water is available both inside and outside of the aquatic host. Yet, in a wide array of marine species, there exists dense populations of bacterial symbionts. Bacterial communities present in several forms, transient, opportunistic and resident, where the latter is more likely associated with bacteria containing some adaption which allows them to maintain survival within the host; for example, adhesion to the lining of a host stomach to avoid extrusion along with the remains of digested food.[81] The type of association often relates to the diet of the host animal, with animals ingesting mostly cellulose substrates found to have digestive systems which support a diverse microbiome (genes provided by a community of microbes), adept at cellulose degradation.

Figure 22 Big ear shipworms, photographed within their burrows. Image by David Fenwick and taken with permission from www.aphotomarine.com

1.8 The Humble Shipworm

This work focuses on novel enzyme characterisation from bacteria living in a symbiotic relationship with marine bivalve molluscs. This specific group of bivalve molluscs of the family *Teredinidae* are found across the globe, and are commonly referred to as shipworms or teredo. The creatures do indeed resemble worms, **Figure** 22, having evolved an extended

translucent tubular body at the head of which sits two shells, or valves, covered in a row a tiny teeth like structures which the animal uses to mechanically bore into wooden substrates. Indeed, the way the animals bore through the wood without collapse of the resulting tunnel is said to have inspired the great engineer Sir Marc Isambard Brunel to invent the tunnelling shield, a machine capable of supporting the leading face of a tunnel during deep excavations, and most famously used to build the Thames tunnel in the early 19th century.[82] The animals live within the submerged wooden material, burrowing along the grain of the substrate as they grow, ingesting and processing the extruded wood as a food source as they go along, **Figure** 22. Food particles are known to collect in a region of the digestive system called the cecum (see **Figure** 23), whereby agitation of the substrate and absorption of the solubilised degradation products such as glucose takes place.[70] They are able to extrude a protective layer of calcium carbonate around the surface of the burrow, which forms a protective yet brittle layer which may strengthen the structure, provide a water-tight layer and protect against predators. Animals have been known to create walls of calcium carbonate to seal off sections of their burrow, as they change direction with the grain of the wood.[83] Two siphons from the tail-end of the animal, one for the controlled influx of water containing essential oxygen and phytoplankton and the other for excretion often protrude out of the small initial bore hole, where the juvenile shipworm entered the wood substrate. Once the siphons are retracted, the animal is able to seal the entry hole by pushing forward its pallets, paddled shaped calcareous structures that flank the siphons.[83-84] This is an important adaption if the wooden substrate enters or is sustained by a habitat containing tidal surges, as sealing the burrow effectively allows the shipworm to remain submerged in water even if the outside of the burrow is exposed to the air. The siphons are known to pump 'fresh' sea water over the gills of the animals which likely contains microorganisms such as plankton, which can be filtered into the digestive system through a connective structure to the mouth, **Figure** 23.[85] Indeed, studies have noted the incorporation of ^{14}C within the tissue of adult animals from labelled phytoplankton, indicating the shipworm's ability to filter particulates from seawater and utilise them as a source of nutrition.[86] Reproduction in shipworm results in larvae release into the water column, where the only practical food source seems to be abundant microscopic plankton, before the animals settle on a burrow location.[83] However, larvae were found not to contain tissues labelled with ^{14}C from plankton. Hence filter feeding of plankton (mainly in the summer months) may be opportunistic once the animals become sedentary within a substrate. Some studies have shown that shipworms can survive in sea water alone, utilising

plankton as a food source, whilst others saw decline of the animals upon removal of a wooden food substrate. Plankton may provide a source of many essential amino acids to the animals, but its abundance seems to affect certain species of shipworms differently.[84-85]

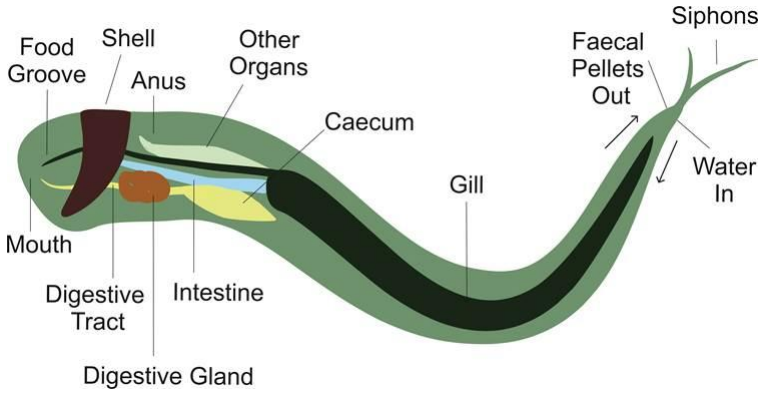


Figure 23 Basic diagram of a shipworm (image not to scale) showing the general layout of the main organs within the animal. A connective structure runs from the gill region to the mouth of the animal.

1.8.1 A historic animal

Whilst interesting to some, shipworms are generally considered pests by the maritime profession, inflicting damage onto wooden structures such as piers and ship hulls by burrowing into the material and destabilising it.[87] Shipworms have been endemic in oceanic environments since antiquity, one event of historical note was the great shipworm 'plague' of the 1730's, where changes in weather caused storm surges around the coast of the Netherlands, creating a prime environment for a shipworm epidemic which threatened the structural integrity of the dikes. Public outcry was high, with many people frightened by the greatly enlarged images of animals threatening the coastline.[88] The abundance of shipworms within the ocean is high, illustrated by common experiments carried out by researchers, whereby blocks of wood left submerged for a period of time will become entrenched with shipworms, allowing collection and study of the animals. The natural environment of shipworms has become intrinsically mixed with human settlement, where terrestrial wood unlikely to interact with oceanic organisms finds its way through to the ocean in the form of man-made structures such as boats, flood defence mechanisms or coastal structures. It is therefore unsurprising that the shipworm is found spread across the globe, most likely having been easily transported around as shipping routes expanded. Studies have also shown how natural disasters such as tsunamis can aid the dispersion of local shipworm species further afield, with one study tracking the movement of Japanese localised species all the way to America.[83] Animals are known to exist in various environments made up of

different types of wooden substrates – these substrates may be sedentary such as man-made docks or mangrove forests, or be pelagic, driftwood or other wood-like substrates such nuts and seeds held aloft in the open ocean.[83]

Table 3 List of Shipworms mentioned in the main text, their geographical location, key research finding and associated reference.

Shipworms	Origin	Key Research finding	Analysis
Teredo navalis	Worldwide distribution, transported by ships worldwide, dubbed 'The naval shipworm'	Reduction in lignocellulose content by passage through shipworm digestive system	Dore, 1923.
Bankia setacea	Pacific Coast	Reducing sugars found after digestion of wood	Boynton, 1926
Bankia australis	Western Pacific Ocean, Australian coastline, predominantly intertidal, mangrove forests.	Electron microscopy study observed bacterial associations in gills	Popham, 1973
Teredora malleolus	North Atlantic Ocean, intertidal, British Coastline	Nitrogen fixation study	Carpenter, 1975
Lyrodus pedicellatus	Northern Altantic Ocean, English Chanel, coastline	Survival off of wooden substrate alone found possible	Gallager, 1981
Bankia gouldi	Western Atlantic Ocean	Isolation of gill bacteria – Teredinibacter Turnerae	Waterbury, 1983
Kuphus polythalamia	Western Pacific Ocean, western and eastern Indian Ocean and the Indo-Malaysian area	Giant shipworm, sulfur oxidising enzymes	Distel, 2018

1.8.2 Bacterial Associations

In 1975, Carpenter and Culliney studied four shipworm species and found a high rate of nitrogen fixation, suggested to support the growth of the animals. They identified a bacteria able to fix nitrogen under anaerobic conditions in one of the species, which was interestingly also found capable of 'liquefying cellulose', isolated from the cecum.[89] Nitrogen fixation is commonly carried out by prokaryotes and the products used by other organisms. The process involves the conversion of atmospheric nitrogen (N_2) into ammonia (NH_3) which can then be used in protein synthesis.[90] In the late 1970s, studies by Gallager observed that some shipworm species could survive without filtration of particulate matter and could instead survive solely off of woody substrates. It was also noted that despite the wood diet, a nutritionally poor source of nitrogen with a high C:N ratio, the shipworm seemed to have an unusually high nitrogen conservation level.[91]

During burrowing, wood which has been ingested after being mechanically destabilised by the rasping motion of the shells, is processed within the digestive system and the remains extruded as faecal pellets. Unlike most other organisms, the digestive system of the shipworm lacks any significant microbial communities.[87] Although Carpenter's study describes isolation of a bacterium from the cecum of a shipworm, colonisation of the digestive system is not commonly observed. Larger organisms existing on a diet containing lignocellulose will utilise bacterial symbiosis to supplement host digestion. In 1923 Dore observed the disappearance of 80% of cellulose and 15-56% of hemicellulose from Douglas fir piling during its transit through the digestive tract of *Teredo navalis,* which was later confirmed to occur in *Bankia setacea* by Boynton.[92-93] Without bacterial assistance, this would be an amazing feat of eukaryotic lignocellulose degradation.

The first observation of bacteria associations was in an electron microscopy study in 1973 by Popham and Dickson. They were studying the so called 'gland of Deshayes' a region of the gills of shipworms that had been studied and named in 1848 by Deshayes. The glandular nature of this region of the gills was disproved by Popham and Dickson's use of electron microscopy, which instead observed a mass of rod shaped gram negative bacteria, Figure 24. In this short paper the authors show that the shipworm does indeed have some sort of bacterial association, localised in the gill region, which they suggest may be symbiotic.[94] Carpenter mentions this study and how they were unable to isolate bacteria from the gill region.[89] 10 years later, in seminal work by Waterbury *et al* the gill bacteria were physically isolated, and grown as a pure culture. From this work, their existence was experimentally

described as symbiotic, specifically endosymbiotic, supporting the original instinctual suggestion by Carpenter and Culliney.[79]

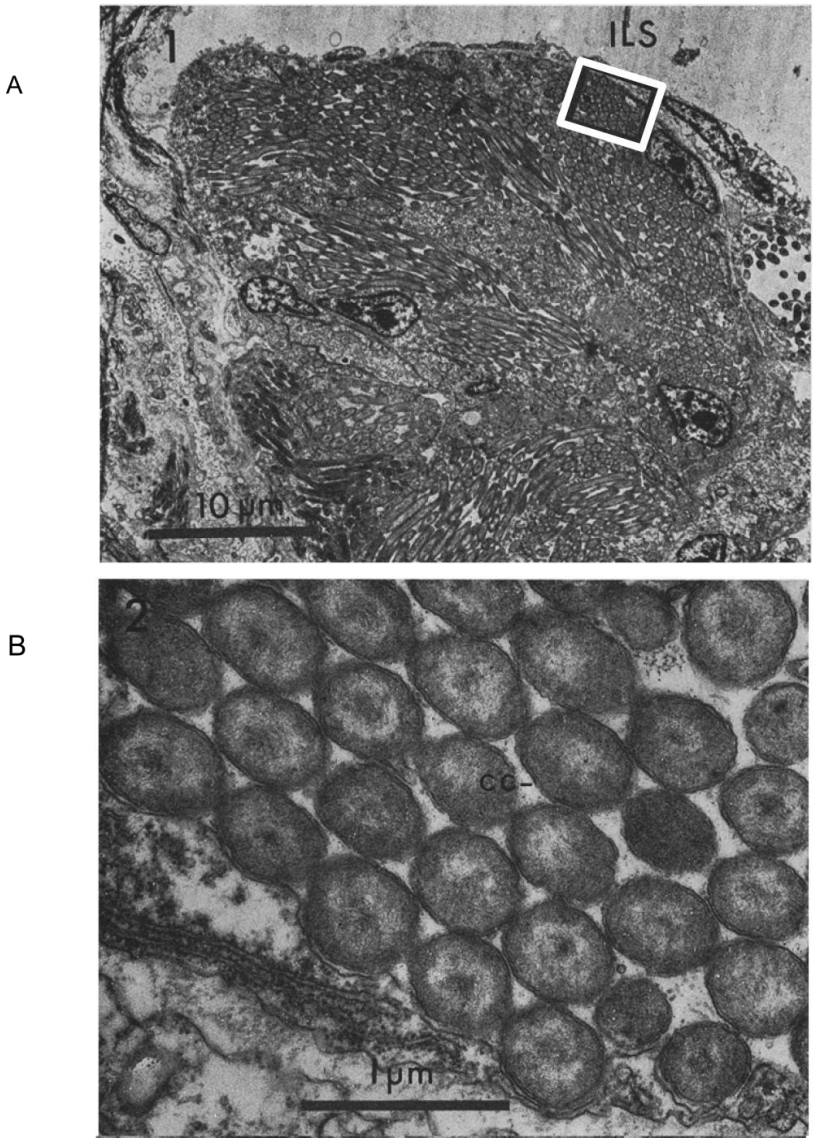


Figure 24 Electron micrographs taken from Popham and Dickson, 1973 .[94] A) View of the 'gland of Deshayes' shown instead to contain a mass of rod shaped bacteria. White box is magnified in B) showing the ordered hexagonal arrangement of bacterial cells.

Dense populations of the bacteria were observed by Waterbury being held within specialised cells called bacteriocytes inside the same internal region of the gills (now without the name, gland of Deshayes). The bacteria were isolated from the shipworm *Bankia gouldi* collected from Beaufort, North Carolina in 1979 and named *Teredinibacter turnerae* (strain T7901). Waterbury suggested that as well as the provision of nitrogen to the host, *T. turnerae* also provided some cellulolytic function as the bacteria were able to grow as a pure culture on cellulosic substrates.[79] Griffin *et al* examined this hypothesis and did find evidence of secreted endoglucanase enzymes from *T. turnerae* that were able to break down long

chain soluble cellooligosaccharides.[95] In 2009, Distel *et al* showed evidence that shipworm gill regions contain small consortiums of different endosymbionts, with *T. turnerae* being the most prevalent.[96] A breakthrough in the understanding of what role the symbiosis may provide the host came about after the genome of *T. turnerae* was sequenced. Yang *et al* analysed the *T. turnerae* genome and compared the amount of potentially carbohydrate active enzymes to other closely related bacteria, *Saccharophagus degradans* and *Cellvibrio japonicas*. Of the potential GH domains found within the genome of *T. turnerae*, 53.5 % were thought to be associated with cellulose or xylan, almost double the number of similar observed genes in *S. degradans* and *C. japonicus*. Furthermore, the GH domains specific for cellulose and xylan were thought to be adapted to the degradation of terrestrial wood. Interestingly, 7 genes were found to consist of multiple catalytic domains, which unusually were predicted to have different functions, which could possibly be involved in the degradation of hemicellulose. The authors also investigated the high number of CBMs present in the genome; 117 domains were found. Approximately 50 % of the domains belong to family CBM2 and CBM10, which are known to be associated in binding crystalline cellulose.[97-98] A similar finding can be observed in the genomes of *S. degradans* and *C. japonicus.* However, *T. turnerae* was also found to contain a high proportion of CBMs thought to be involved in binding to xylan (CBM22) compared with *S. degradans*. The analysis of the genome suggested that *T. turnerae* lacks the typical carbohydrate active enzymes required for degradation of typical marine polysaccharides such as agar and chitin. Instead, the plethora of CAZymes, which represent 38 GH families appear to specialise in the degradation of substrates rich in cellulose, xylan and mannans, which are typically found in woody plant materials.[99]

1.8.3 Probing the shipworm symbiotic relationship

The question remains as to why a community of bacteria containing such a large proportion of potentially wood-degrading genes are held localised in the gills, and not within the digestive system of the animal. *T. tunerae* was known to fix nitrogen in pure culture and animals were known to have a higher nitrogen content than would otherwise be expected from their diet. The nitrogen fixation by *T. turnerae* in situ was elegantly shown by Lechene *et al*, who in 2007 used a technique called multi-isotope mass spectrometry (MIMS) to measure the abundance of ^{15}N in the gill tissue of shipworm *L. pedicellatus*. The animals had been grown in tanks enriched with isotopically-labelled nitrogen gas, and stable ^{15}N isotope incorporation into the gill region was mapped by MIMS coupled with transmission electron

microscopy (TEM). The visual map provided evidence that newly fixed nitrogen was highly abundant within the symbiotic bacteria and lower amounts were observed in nearby host cells, indicating that transfer of the newly fixed isotopically labelled nitrogen to the host cells was occurring.[100] Further mapping of host-symbiont interactions was carried out by O'Connor *et al.*[101] Probes specific for 16S rRNA endosymbiont gene fragments were detected using fluorescence in situ hybridisation (FISH). The probes were highlighted in high abundance within the gill region where the bacteriocytes were located. Interestingly, the probes were also identified within the cecum of the animals, which are known to be free of bacteria. This identification of endosymbiont gene fragments within the digestive system of the host strongly indicated that bacterial proteins active on carbohydrates were being translocated from bacteriocytes in to the gills to the gut of the animal.[101]

The majority of *T. turnerae* CAZymes are secreted, with experiments showing accumulation of GH proteins within spent growth media was higher when the culture conditions included cellulose as a substrate.[99] A major gap in the understanding of the shipworm-symbiont relationship is how the bacteria know to produce the vast array of lignocellulosic proteins when the substrate is distant to their localisation. O'Connor *et al* studied the movement of bacteria proteins around anatomy of the shipworm and several intriguing aspects were uncovered. Firstly, much like the earlier study (carried out by the same group), a selection of bacterial (phylogenetically similar to *T. turnerae*) enzymes were found within the gut of the shipworm, *Bankia setacea*. Importantly, this study highlights the fact that only a small number of bacterial enzymes were translocated to the gut, and all were predicted to be 'wood-degrading'.[101] Tissue extracts containing bacteriocytes were analysed for their protein content, of which 11 % was suggested to be due to expression of functional CAZymes with potential activity or binding to cellulose and hemicelluloses. As if confirming previous studies, the cecum contents of the animals were analysed for the endosymbiont protein content. Approximately all (~98 %) of the bacterial proteins detected within the cecum were predicted to have some involvement in cellulose or hemicelluloses binding or catalytic activity. All the cecum located proteins were also identified within the bacteriocytes themselves. Interestingly, a selection of bacterial proteins identified in the cecum were CBM domains linked to domains of unknown function, likely to represent uncharacterised activities.[101]

Several studies have focused on the bacterial association, whilst little research into shipworm endogenous proteins has been carried out. Sabbadin *et al* undertook a study which focused on assessing the role of native shipworm proteins in lignocellulose digestion.

They identified the 'digestive gland' as the major source of host lignocellulosic protein transcription in *L. pedicellatus*. Like with the bacterial proteins, the majority of host CAZymes are also secreted GHs, the majority of which are predicted cellulases. Sabbadin and colleagues highlight a potential flaw in the analysis of the cecum proteome of *B. setacea* by O'Connor in 2014, which focused entirely on analysing the bacterial content. This recent study shows that of the CAZymes found within the cecum of *L. pedicellatus*, only 15 % are of bacterial original, with the remaining 85 % thought to be produced by the shipworm itself; they are also able to link abundant cecum GHs to those highly transcribed within the digestive gland.[70] It seems that there exists both a complex translocation system of both bacterial and endogenous proteins to the site of digestion.

The relationship between shipworm and their endosymbiotic bacteria appears complex, despite several years' worth of research on this interesting relationship. The abundance of bacteria containing so many potential carbohydrate active genes in a region far distinct from the digestive system of the animal is puzzling. The way the bacteria are housed within specialised gill cells also seems strange, as if their existence is somehow being controlled by the host animal. The gill region is connected by a long tubular structure, known as the food groove which links the gill to the mouth (and thereafter the digestive system). This structure may be involved in filtering out particulates taken in by the gills, such as plankton. However, this also represents a prime route for bacterial enzyme translocation. The nitrogen fixation provided by *T. turnerae* is of great importance to the shipworm, as it supplements the low nitrogen content of their wooden food source. The utilisation of CAZymes could have been a secondary outcome as the bacterial population is maintained at a steady level by the host, and only found in a specific area of the animal. Some hypothesise that the evolutional increase in gill size may have induced growth of bacterial populations rather than increase the amount of filter feeding; a higher surface area of the gill can accommodate more bacteriocytes and indeed this idea holds true based on the studies of shipworms grown in laboratory conditions where the only food source was their wooden burrows.[85, 91] A recent advance in 2017 saw the discovery of a new species of shipworm, a greatly enlarged version found living within mud burrows instead of wooden substrates; Distel *et al*, used video evidence in their study, showing the animal being removed from its protective calcium carbonate layer, and is well worth seeing as a visual representation of the size of the animal. The metre long animals, *Kuphus polythalamia* are an interesting development in the evolutional study of the shipworm species, who have a greatly reduced digestive system.

Instead of cellulytic bacteria, the gill regions of the giant shipworm play host to bacteria capable of oxidising sulfur, a major component within the buried marine sediments.[102]

1.9 This Work

To summarize, there is still mystery surrounding the relationship between shipworm and bacterial symbionts such as *T. turnerae*. The shipworms have been shown to produce their own digestive enzymes, yet they are thought to also utilise the enzymes produced in the gills by their symbiotic bacteria for both nitrogen fixation and for the digestion of woody substrates.[87, 96, 99, 101, 103] Experiments have shown that certain enzymes are secreted when *T. turnerae* is grown on cellulose substrates, but little characterisation of these enzymes has been carried out.[104-105] Shipworms and their bacterial symbionts represent one of many untapped enzymatic resources for the discovery of novel lignocellulosic enzymes. Biochemical characterisation of the lignocellulosic enzymes used by shipworms may prove useful in future research focusing on improvements to industrial systems creating cleaner energy sources from waste biomaterials.

1.9.1 General Aims

The aims of this project were to investigate the activity and structural properties of several likely carbohydrate active enzymes taken from the genome of *T.turnerae.* This work has focused on the characterisation of several novel glycoside hydrolases and a single LPMO. Several experimental enzyme characterisation techniques have been used throughout this work and for ease these will be described in Appendix 2, in the hope that this provides the reader with enough background knowledge to underpin the subsequent results chapters.

In the initial 1-2 years of this 4-year project a large proportion of time was dedicated to the analysis of a several bacterial homologues of a single gene found within the cecum of the shipworm *L. pedicellatus*, thought to be an endogenous protein similar to peptidylglycine alpha-hydroxylating monooxygenase (PHM) – another example of a copper oxidative enzyme.[106] This work is shown in Appendix 1.

Thereafter, work focused on the shipworm symbiont *T. turnerae*. Chapter 2 describes the strategies used to produce recombinant versions of several *T. turnerae* CAZYmes. Characterisation of novel enzymes from the bacterial symbiont genome is organised based

on enzyme family and Chapters 3, 4, 5 and 6 offer insights into *T. turnerae* enzymes from families GH5, GH12, GH8 and LPMO AA10 respectively. Chapter 7 will summarise the results of this novel enzyme characterisation endeavour and provide the reader with perspective and describe potential opportunities for future work in this exciting marine area of discovery.

2

Protein Production

2.1 Abstract

A marine bivalve mollusc, commonly known as a shipworm is adept at surviving on a diet rich in recalcitrant carbohydrates, as its main food source is woody material found in oceanic environments. The shipworm is thought to utilise carbohydrate-active enzymes produced from a community of symbiotic bacteria housed within the gills of the animal. A set of 245 carbohydrate active enzymes are found within the genome of shipworm endosymbiont, *Teredinibacter turnerae,* over half of which are thought to be involved in the breakdown of polysaccharides to useful products. 15 *T. turnerae* genes were selected for protein expression trials on the basis of their glycoside hydrolase (GH) family, whether they were the sole representative of a family and the expected or unknown activity of the target protein. As such the targets covered a broad range of GH families as well as a single LPMO family, and potentially display degradative ability on a wide variety of lignocellulosic substrates. After several purification trials, pure protein for 5 GH enzymes from family 12 (ACR14297.1), 8 (ACR14722.1) and 5 (ACR12145.1, ACR12247.1, ACR11279.1), were obtained through cytoplasmic expression by *E. coli* of the enzyme catalytic domains, tagged with a removable N-terminal hexahistidine purification tag. Purification of an LPMO, from subfamily AA10, *Tt*AA10 (ACR14100.1), the only LPMO in the genome of *T. turnerae* was carried out by expression with a pelB signal peptide, allowing soluble protein to accumulate in the periplasm of *E. coli* and purified with the aid of a C-terminal Strep tag. This chapter shows the purification strategies employed leading to the development of suitable, and importantly soluble, target expression routes for 6 carbohydrate active enzymes from *T. turnerae*; 3 proteins from family GH5, a GH12, GH8 and an AA10 LPMO.

2.2 Introduction

2.2.1 Carbohydrate Active Enzymes in T. turnerae

As discussed in **Chapter 1**, the genome of *T. turnerae* contains 245 open reading frames that have been classified as carbohydrate active enzymes.[99] A large proportion of these are further expected to be glycoside hydrolases and the sequences have been classified into a variety of GH families, as shown in the CAZy database.[34] The mechanism of movement of the carbohydrate active enzymes produced by *T. turnerae* from their initial localisation, held in the specialised bacteriocytes within the gills of the host shipworm has been discussed by O'Connor *et al* and more recently by Sabbadin *et al* and the final destination of the extracted bacterial proteins is suggested to be in the digestive region of the shipworm's anatomy.[70, 101] Perhaps unsurprisingly, large numbers of the CAZymes produced by *T. turnerae* are secreted, with peptide leader sequences found at the beginning of the each GH sequence and LPMO, indicating that proteins are secreted and held within the host bacteriocytes until they are mechanically extracted; bacteriocytes are not found within the gut of the animal, yet bacterial proteins are, indicating there to be some host driven process of protein removal from bacteriocytes Questions still remain as to how the relationship between host and endosymbiont is maintained; especially with regards to access to a food source by the bacteria.

The large number of carbohydrate-active enzymes (or binding domains) [34, 104, 107-108] within the genome of *T. turnerae* are classified into many different GH families, which suggests that these proteins will have very specific activities, and cover a broad range of substrates. Whilst the host organism does encode its own carbohydrate active enzymes, combination with the wide variety of GHs produced by *T. turnerae* would provide the host with the large enzymatic toolbox needed to efficiently breakdown the environmentally localised polysaccharides. The type of polysaccharides found within the shipworm food source depends on the localisation of the animal in the world and source of woody material used for the animals burrow; for example, animals living in burrows made in the bottom of boat hulls would likely require a slightly different set of enzymes to the animals living in natural submerged mangrove forests, due to the difference in wood type. It remains unknown whether *T. turnerae* produces all its CAZymes simultaneously, or whether there is some feedback system which provides the bacteria with information on which enzymatic functions are required depending on the wood substrate chosen for inhabitation by the

host shipworm. It may be that *T. turnerae* possess a wide variety of enzymes from different GH families to combat this lack of choice when it comes to the food source, or that a selection of key enzymes remain encoded and expressed based on evolution of the symbiosis over time; however a 2018 study into the mutation rate of *T. turnerae* found that the symbiont was resistant to the typical erosion of the genome normally associated with bacteria living as endosymbionts, which could explain its ability to survive as a free-living culture in laboratory studies [109] and possibly allowing it to retain a wider variety of lignocellulosic proteins.

As discussed in **Chapter** 1, whilst *T. turnerae* is the main characterised bacterial strain found within certain species of shipworm[110], it is clear that other related bacterial species are prevalent in both single animals and in animals found in different regions.[96, 108, 110-111] Sequence analysis shows a high similarity between *T. turnerae* proteins and those from other bacterial clades such as Bs12/Bs08 [110]. However, it is likely that differences in the type of GHs produced by bacteria from different shipworm species occurs and may be a response to the type of substrate chosen by the animal. Despite the circumstances of expression of GH proteins remaining unknown, the genome of *T. turnerae*, and by extension other related bacterial clades provides an excellent pool of targets for which lignocelluloses degradation can be probed.

2.2.2 The Targets

Of the 245 carbohydrate active proteins and binding domains found within the genome of *T. turnerae*, 101 are suggested to be GHs.[34, 101] Work by Yang *et al* suggested that of these 101 GHs, just over half were expected to be involved in the degradation of cellulose or xylan, about a quarter devoted to other polysaccharides such as chitin, agarose, laminarin, and pectin, whilst the last quarter were defined as 'non-specific'. [99] It is likely that the non-specific portion of the GHs are labelled as such due to insufficient similarity to existing characterised enzymes. As such, this set of marine CAZymes provides a vast pool of sequences from which novel proteins with existing and possibly new function may be derived. Some of the work needed to begin characterising these likely polysaccharide degrading enzymes has already been started by their incorporation in the CAZy database. Sequence analysis, involving HCA is used to match the proteins to their likely structural fold and to members of known GH families.[15, 25, 27, 34, 36] Indeed, there is an extensive list of sequences from the *T. turnerae* genome shown as classified into predicted GH families, from which one can narrow down possible functions.

Whilst the scope of this project would not be able to cover all 101 GH proteins from *T. turnerae*, a selection of 15 sequences were chosen which covered several GH families and a single LPMO from auxiliary activity family AA10. Table 4 shows the sequences chosen with their associated CAZy codes, expected family and subfamily, as well as the potential functions as determined by sequence analysis with BLAST [112-113]. A couple of sequences were chosen based on the fact that they were the sole representative of their associated GH family, namely the GH8 and GH12. A third, GH5_un, was chosen as the sequence has not yet been characterised into a GH5 subfamily (see Chapter 3 for a discussion of GH5 subfamily assignments).

Table 4 Carbohydrate active genes chosen from *T. turnerae* for characterisation, with potential function as suggested by sequence homology searches.

Name	ORF	Genbank Code	Family _subfamily	Potential Function
AA10	TERTU_0046	*ACR14100.1*	**AA10**	Chitin binding protein/likely LPMO
GH5-2895	TERTU_2895	*ACR12145.1*	**GH5_2**	B-1,4-glycan cleaving enzyme (extracellular)
GH5-3565	TERTU_3565	*ACR11017.1*	**GH5_1**	B-1,4-glycan cleaving enzyme
GH5-0183	TERTU_0183	*ACR12128.1*	**GH5_26**	Endo β-1,4-glycanase
GH5-3751	TERTU_3751	*ACR11279.1*	**GH5_un**	Unknown subfamily/activity
GH5-3361	TERTU_3361	*ACR12247.1*	**GH5_4**	Endo-β-1,4 glucanase-specific for xyloglucan
GH5-0428	TERTU_0428	*ACR12792.1*	**GH5_53**	Cellodextrinase/β-glycanase
GH5-0427	TERTU_0427	*ACR13327.1*	**GH5_53**	Cellodextrinase/β-glycanase
GH6-2898	TERTU_2898	*ACR12723.1*	**GH6**	Cellobiohydrolase
GH6-3996	TERTU_3996	*ACR14000.1*	**GH6**	Cellobiohydrolase
GH6-2895	TERTU_2895	*ACR12145.1*	**GH6**	Cellobiohydrolase
GH8-4506	TERTU_4506	*ACR14722.1*	**GH8**	Endo β-1,4-xylanase
GH9-0607	TERTU_0607	*ACR11786.1*	**GH9**	Cellulase/endo-β-1,4-D-glucanase
GH9-0645	TERTU_0645	*ACR14629.1*	**GH9**	Endoglucanase
GH12-0353	TERTU_0353	*ACR14297.1*	**GH12**	Endoglucanase, hydrogenase/urease accessory protein hupe prior to GH12
GH45-3400	TERTU_3400	*ACR13005.1*	**GH45_A**	Endoglucanase

2.2.3 The Purification Key: Searching for Solubility

Successful production of target proteins is often underpinned by inherent or conditional solubility of the translated protein. Incompatible conditions can cause a particular protein construct to be produced in an insoluble form. This can be a serious road block on the journey towards production of pure protein and often needs a significant amount of time devoted to the de-convolution of the appropriate 'soluble' experimental expression condition. Insolubility can arise due to issues with the protein structure, for example a high degree of hydrophobic surface residues may cause the formation of inclusion bodies which minimise solvent-protein surface interactions. Inclusion bodies can sometimes be used as

an advantage in some cases, where the protein is actually in a sense protected within the aggregate from outside interference such as proteases or reactions to changing conditions; the protein can be reclaimed in a soluble form if the protein aggregate is appropriately denatured and refolded on return to more native conditions. The production of inclusion bodies can sometimes in itself act as an *in vivo* purification step. Successful reclamation of protein from inclusion bodies is however, not always possible. Production of the protein may not be optimised to the system, leading to misfolding events which can cause insolubility. Refolding of proteins with unknown functions may also result in a folded final protein product, but there may be changes in structure caused by the laboratory conditions that do not reflect the native structure. As such it is important to optimise expression conditions to suit an individual target and this is often driven towards producing soluble forms of a protein as this is most often easier to work with. Lines of optimisation can include the following; construct design, *E. coli* cell lines, temperature, additives, growth media, amount of oxygen during culture growth, construct toxicity to the cell strain, inducer concentration, shaking speed, growth time, protease inhibitors and extraction buffer.

2.2.4 Ensuring a good start; constructs with a Native N-terminal Histidine

The *T. turnerae* LPMO target protein, *Tt*AA10, was first produced by Dr. G Hemsworth using SUMO tagged expression construct, SUMO-*Tt*AA10, albeit in a very small amount. At the start of this work, very little protein was extracted using this construct, and it did not yield enough protein for further characterisation. As an LPMO, the N-terminal histidine residue forms an integral part of the active site of the protein. Therefore, the construct designs need to ensure that the native N-terminal Histidine residue remains unencumbered. A SUMO tag, which stands for Small Ubiquitin-like Modifier, is eukaryotic in origin and forms partners with other proteins as a post translational modification, which is thought to play a role in cell processes. The SUMO protein is now a well-used biotechnological system, in that its attachment to a recombinant protein has been found improve the solubility and stability of the target. On top of this, the SUMO tag contains its own N-terminal hexahistidine tag, making it both ideal for improving solubility and enabling typical affinity chromatography for purification. The SUMO protein is easily removed from the recombinant protein assembly using a specific SUMO protease, which recognises the C-terminal region of the tag, cutting after the last two glycine residues, leaving no left over

residues on the partner protein, ***Figure 1*** [114-118]. The production of a native N-terminal residue on the released partner protein makes it an ideal system to use with LPMOs, which require their N-terminal Histidine for activity, and has been demonstrated by Gregory *et al* to improve LPMO yield in an *E. coli* expression strategy.[119]

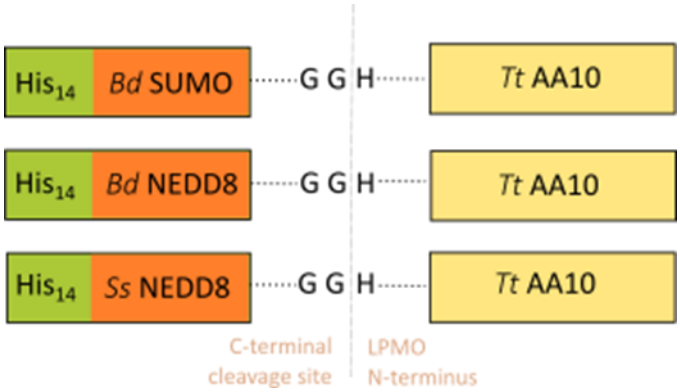


Figure 1 Schematic showing the SUMO and SUMO-like NEDD8 solubility tags (preceded by a 14 residue long histidine tag) and LPMO sequence. Two c-terminal glycine residues, positioned directly before the N-terminal Histidine of the LPMO are the recognition site for cleavage by the corresponding SUMO protease. (19)

2.2.5 Purification aims

The goal of this section of work was to produce soluble forms of as many target proteins out of the chosen set of 15 sequences from *T. turnerae* described previously as possible. In the drive to produce the GH targets, several different expression testing conditions were employed which enabled identification of the most promising targets for soluble and large scale protein expression. This ultimately led to the successful large scale production of 5 GH proteins, *Tt*GH12, *Tt*GH8, *Tt*GH5_2, *Tt*GH5_4 and *Tt*GH5_un, and their purification route is shown herein. Production of *Tt*AA10 proved more difficult and strategies to improve the expression yield of *Tt*AA10 included specialised induction conditions such as autoinduction and the effects of stress on chaperone co-expression. A large number of constructs were designed in the hope of improving the solubility and yield of *Tt*AA10, whilst maintaining an uncompromised free N-terminus of the protein. The protein was successfully produced using a C-terminal Strep tag and periplasmic expression (similar to the native production of the protein), using a construct designed by Luisa Elias, (CNAP, and University of York), and its purification will be shown in the forthcoming **Chapter**.

2.3 Methods:

2.3.1 Competant Cells

Six competent cell lines were using during cloning and are described in Table 5

Table 5 List of *E. Coli* expression strains used for expression testing of *T. turnerae* GH and LPMO targets, with key characteristics described (as noted in manufacturer product details).

Expression Strain	Supplier	Characteristic
NEB 5-Alpha	New England Biolabs	T1 phage resistant and *end*A deficient
BL21(DE3)	New England Biolabs	Lacks two proteases normally found within the cytoplasm (Lon) and outer membrane (OmpT)
One Shot BL21 Star (DE3)	New England Biolabs	Greater mRNA stability
SHuffle T7	New England Biolabs	Promotes disulfide bond formation within cytoplasm, by expression of disulfide bond isomerise (DsbC)
pGro7	Takara	Co-expression of chaperone groES-groEL by induction of the araB promoter with L-Arabinose.
pTf16	Takara	Co-expression of chaperone tig by induction of the araB promoter with L-Arabinose.

2.3.2 Vectors, Primers and DNA

The vector pNT-TrxT containing the thioredoxin tag was purchased from Addgene; the pNT-TrxT vector was received as a cell stock and was plated onto agar (kanamycin, 30 µg/ml) and stock plasmid produced using the QIAprep Spin Miniprep kit (QIAGEN). The vector containing the maltose binding protein (MBP), pETFPP2 was given by the Professor. J Potts group (Department of Biology, University of York) and originally made by the University of York Technology Facility. All primers were bought from Eurofins-GATC and reconstituted on arrival according to the manufacturer instructions. DNA sequences for all *T. turnerae* GHs were produced, and constructs cloned into pET28a vectors, by Genscript.

2.3.3 CAZyme Mining

The *T. turnerae* (Taxid; 377629) genome is publicly available and the 245 potential carbohydrate active enzymes are displayed in the CAZy database. A selection of glycoside hydrolase families were chosen to provide a spread of coverage over a range of possible

substrate specificities, as described in Table 4. A single LPMO is found within the *T. turnerae* genome, henceforth known as *Tt*AA10, and was also proposed as a target for characterisation. 15 sequences were taken from NCBI and analysed using NCBI BLAST for the presence of different domains. The genes contained mixtures of domains; signal peptide, linker regions, CBM domains and the catalytic domains. The catalytic domain boundaries were identified based on those used in the CAZY modular classification, and the information used to design constructs suitable for cytoplasmic expression.

2.3.4 GH Construct Design

All GH genes used in this work were designed as follows, from N to C-terminus:

- Two CC bases – to move insert into the correct reading frame in PET28a,
- Methionine codon (ATG)
- N-terminal hexahistidine affinity tag,
- 3C protease cleavage site
- GH target sequence
- Stop codon (TAA)

Sequences were codon optimised for expression by *E. coli* (as collated in Appendix 3) and cloned into plasmids (pET28a) containing antibiotic resistance (kanamycin) by GenScript. The plasmids were individually transformed by *E. coli* competent cells. Transformation was carried out by adding 1 µl plasmid to 25 µl competent cells and left on ice for 30 mins, before heat shock at 42 °C for 30 secs and cooling on ice for 2 mins. Cells were added to 300 µl SOC media (super optimal broth with catabolite repression) and incubated at 37 °C for 1 hr with shaking (800 rpm). The cells were then spread on agar plates containing antibiotic (kanamycin at 30 µg ml^{-1} for BL21 (DE3) and SHuffle T7, and kanamycin/chloramphenicol 30 µg ml^{-1} and 20 µg ml^{-1} respectively for chaperone co-expression cells pGro7 and pTf16 and incubated overnight at 37 °C. Single colonies were picked and small cultures grown in LB containing appropriate antibiotic to an OD of 0.6 and mixed in a 1:1 ratio with glycerol (50 % v/v) to create cell stocks (25%) at -80 °C.

2.3.5 Cloning of LPMO

A plasmid containing the gene SUMO-*Tt*AA10 within the vector Champion SUMO (Appendix 3) made previously by Dr G. Hemsworth prior to the start of this project was used as a starting point to produce all further *Tt*AA10 constructs.

2.3.5.1 Cloning: Vector and Insert Preparation

Polymerase chain reaction (PCR, see Appendix 2 for method theory) was carried out using the following mixture to amplify DNA;

- Template DNA (1 µl)
- Q5 buffer (10 µl) (New England Biolabs)
- dNTPs (10 mM) (1 µl) (Thermo Fisher Scientific)
- Forward primer (1.5 µl)
- Reverse primer (1.5 µl)
- Water (34 µl)
- Q5 DNA polymerase (1 µl) (New England Biolabs)

PCR reaction times were specific to each primer pair (forward and reverse) annealing temperature – see Table 6 and Table 7. Primers for the vectors produced blunt ended double stranded vector DNA. One or more points within the vector sequence were chosen as the annealing positions for the forward and reverse primers, Table 6. PCR initiates DNA replication forming a double stranded DNA sequence containing only the required vector sequence, removing any unwanted vector sequence which was originally between the two primer annealing positions. To make the inserts, primers were designed with two regions; once region binds to the DNA of the target sequence, and a second region which acts as an overhang. The overhang sequence is complementary to a section of DNA on the vector sequence. After PCR, additional nucleotides would have been built along the single stranded overhang region by the DNA polymerase. Before ligation of the insert and vector DNA fragments can occur, the double stranded overhang on the insert must be 'chewed' back enzymatically.

Vector and insert were combined using the HiFi DNA assembly cloning kit (New England Biolabs). In the reaction, a 5'-3' exonuclease creates single stranded 3' overhangs – allowing the complementary regions on the 3' end of the vector and insert to bind together. The DNA polymerase then carries out extention of the newly forming complimentary strand using the available nucleosides within the ligation reaction. A DNA ligase is then able to seal the nicks in the sequence. Ligation reactions were carried out by mixing vector and insert fragments in a 1:3 molar ratio. Quantities required for ligation were calculated using the free online tool NEBioCalculator (Clontech, Takara) (http://NEBiocalculator.neb.com/#!/ligation) based on the size of the DNA sequence. The ligation reaction (total reaction volume, 20 µl) was heated at 50 °C for 1 hour and used the HiFi DNA Assembly Master Mix. Reactions were then transformed into competent cells, NEB 5-alpha (New England Biolabs) and plasmid extracted

using the QIAprep Spin Miniprep kit (QIAGEN). Plasmid stocks were then transformed into different expression cell strains.

2.3.5.2 DNA and Plasmid Analysis

PCR reactions were analysis for the correctly sized fragments by agarose gel electrophoresis. Agarose gels (6 %) for DNA fragment separation were made using the following materials;

- TAE buffer (Tris-acetate-EDTA), 60 mL – microwaved for 1 minute
- 0.6 g agarose
- 1 μL SYBR safe DNA gel stain (Thermo Fisher)

Agarose gels were run for 1 hr at 100 V, using TBE (Tris-Borate-EDTA) running buffer. DNA samples were pre-mixed with DNA gel loading dye (6X, Thermo Fisher). A 1 kb DNA ladder was purchased from New England Biolabs.

Plasmid uptake into cloning cells (OneShot Top10) was tested using colony PCR. Single colonies were picked from agar plates and analysed to check if the correct target insert had been ligated successfully into the plasmid. The single colony was mixed into a pre-made PCR reaction solution containing the following:

- Taq buffer (10 μl) (New England Biolabs)
- dNTPs (10 mM) (1 μl) (Thermo Fisher Scientific)
- Forward primer (1.5 μl) (specific for gene insert)
- Reverse primer (1.5 μl) (specific for gene insert)
- Water (34 μl)
- Taq polymerase (1 μl) (New England Biolabs)

The PCR reaction was run using the following thermal cycling conditions:

- Denaturation – 94 °C – 30 sec
- Annealing – *temperature dependent on primers* -30 sec
- Extension – 72 °C - *time dependent on DNA fragment length*
  Three stage process cycled 30 times.

Table 6 Forward and reverse primers were designed to anneal to specific points on the vector sequence, which when amplified using PCR would create linear double stranded vector DNA.

Vector	Forward Primer	Reverse Primer	Heating conditions	Purpose
pET11a pelB-SUMO-*Tt*AA10*	CGCCATCGC CGGTTGGGC	TAAGGATC CGGCTGCT AACAAAGC	94 °C (30sec) [94 °C (30sec), 69 °C (30sec) 72 °C (3 mins)] **(35 cycles)** 72 °C(5min)	Retain pelB site but remove SUMO-TtAA10
pNT-TrxT	AGTAAAGGT GGATACGGA TCC	CATGGATT GGAAGTA CAAGTTCT C	94 °C (30sec) [94 °C (30sec),64 °C (30sec) 72 °C (2 mins, 10 sec)] **(35 cycles)** 72 °C(5min)	Vector kept intact, only made linear
pETFPP2 His_6-MBP-3C	TGAGATCCG GCTGCTAAC AAAGCCCG	TGCTGGTC CCTGGAAC AGAACTTC CAG	94 °C (30sec) [94 °C (30sec),63 °C (30sec) 72 °C (2 mins, 20 sec)] **(35 cycles)** 72 °C(5min)	Anneals at 3C site and at the stop codon after the C-terminal His_6 tag
pET28 YSBL LIC	CACCACCAC CACCACCAC TGAGATCCG GC	CATGGTAT ATCTCCTT CTTAAAG	94 °C (30sec) [94 °C (30sec),58 °C (30sec) 72 °C (2 mins, 20 sec)] **(35 cycles)** 72 °C(5min)	Removal of N-terminal His and YSBL LIC section, but keeps C-terminal His tag
pSF 1477 *Bd*SUMO-MBP	TAAGGATCT CATCACCAT CACCATCA	GCCACCAC GCAGGGC CAGA	94 °C (30sec) [94 °C (30sec),67 °C (30sec) 72 °C (2 mins 20 sec)] **(35 cycles)** 72 °C(5min)	Remove MBP sequence
pSF 1478 *Bd*NEDD8-AGT	*As above*	*As above*	94 °C (30sec) [94 °C (30sec),67 °C (30sec) 72 °C (2 mins 20 sec)] **(35 cycles)** 72 °C(5min)	Remove AGT sequence
pSF 1479 SsNEDD8-AGT	*As above*	*As above*	94 °C (30sec) [94 °C (30sec),67 °C (30sec) 72 °C (2 mins 20 sec)] **(35 cycles)** 72 °C(5min)	Remove AGT sequence
pelB pSF 1477 *Bd*SUMO-MBP	*As above*	*As above*	94 °C (30sec) [94 °C (30sec),67 °C (30sec) 72 °C (2 mins 20 sec)] **(35 cycles)** 72 °C(5min)	Remove MBP sequence

Table 7 Forward and reverse primers used to produce *Tt*AA10 inserts for the intended vectors as shown in **Table** 6. Primers contain a complementary region to the target gene sequence and overhangs (small case) which are complementary to each end of the linear vector. All inserts were created from a complete plasmid made by Dr G. Hemsworth prior to the start of this project, in which sequence was available for SUMO-*Tt*AA10, within the Champion SUMO vector.

Insert Sequence	Forward Primer	Reverse Primer	PCR Conditions	Vector	Final expression cassette
SUMO-*Tt*AA10	gagaacagattggtggt CATGGCTATATTG AATCGC	gcagccggatcctta TCAACCAAAGTC CACGTC	94 °C (30sec) [94 °C (30 sec),69 °C(25sec) 72 °C(15 sec)] **(30 cycles)** 72 °C(5min)	pET11a	**pelB-SUMO-*Tt*AA10**
*Tt*AA10	caaccggcgatggcgC ATGGCTATATTGA ATCG	gcagccatccttaTC AACCAAAGTCCA CGTC	94 °C (30sec) [94 °C (30sec),69 °C (15sec) 72 °C (15 sec)] **(30 cycles)** 72 °C(5min)	pET11a	**pelB-*Tt*AA10**
*Tt*AA10	gaaggagatataccat gCATGGCTATATT GAATCGC	tcagtggtggtggtg gtggtgTCAACCA AAGTCCACGTC	94 °C (30sec) [94 °C (30sec),58 °C(30sec)72 °C(15 sec)] **(30 cycles)** 72 °C(5min)	pET28 YSBL LIC	***Tt*AA10-His**
pelB-SUMO-*Tt*AA10	gaaggagatataccat gATGAAATACCTG CTGCCG	tcagtggtggtggtg gtggtgTCAACCA AAGTCCACGTC	94 °C (30sec) [94 °C(30sec),62 °C(30sec) 72 °C (25 sec)] **(30 cycles)** 72 °C(5min)	pET28 YSBL LIC	**pelB-SUMO-*Tt*AA10-His**
SUMO-*Tt*AA10	ccagggaccagcaAT GGGCAGCAGCC	gccggatctcaACC AAAGTCCACGTC CAC	94 °C (30sec) [94 °C (30sec),69 °C (30sec) 72 °C (15 sec)] **(30 cycles)** 72 °C(5min)	pETFPP2	**His-MBP-3C-SUMO-*Tt*AA10**
SUMO-*Tt*AA10	tgtacttccaatccatg ATGTCGGACTCAG AAGTC	ccgtatccacctttac tTCAACCAAAGT CCACGTC	94 °C (30sec) [94 °C (30sec),69 °C (30sec) 72 °C (15 sec)] **(30 cycles)** 72 °C(5min)	pETFPP2	**Thioredoxin-TEV-SUMO-*Tt*AA10**
*Tt*AA10	caccagactggtggcC ATGGCTATATTGA ATCGC	tggtgatgagatcctt aTCAACCAAAGT CCACGTC	94 °C (30sec) [94 °C (30sec),69 °C (30sec) 72 °C (15 sec)] **(30 cycles)** 72 °C(5min)	*Bd*SUMO-MBP	**B*d*SUMO-*Tt*AA10**

Table 5 continued...

Insert Sequence	Forward Primer	Reverse Primer	PCR Conditions	Vector	Final expression cassette
*Tt*AA10	gccctgcgtggt ggcCATGGC TATATTGAA TCGC	tggtgatgagat ccttaTCAAC CAAAGTCCA CGTC	94 °C (30sec) [94 °C (30sec),69 °C (15sec) 72 °C (15 sec)] (30 cycles) 72 °C(5min)	B*d*NEDD8-AGT	**B*d*NEDD8-*Tt*AA10**
*Tt*AA10	gccctgcgtggt ggcCATGGC TATATTGAA TCGC	tggtgatgagat ccttaTCAAC CAAAGTCCA CGTC	94 °C (30sec) [94 °C (30sec),69 °C (15sec) 72 °C (15 sec)] (30 cycles) 72 °C(5min)	*Ss*NEDD8-AGT	***Ss*NEDD8-*Tt*AA10**
*Tt*AA10	caccagactgg tggcCATGGC TATATTGAA TCGC	tggtgatgagat ccttaTCAAC CAAAGTCCA CGTC	94 °C (30sec) [94 °C (30sec),69 °C (15sec) 72 °C (15 sec)] (30 cycles) 72 °C(5min)	pelB *Bd*SUMO-MBP	**pelB *Bd*SUMO-*Tt*AA10**
*Tt*AA10	gccctgcgtggt ggcCATGGC TATATTGAA TCGC	tggtgatgagat ccttaTCAAC CAAAGTCCA CGTC	94 °C (30sec) [94 °C (30sec),69 °C (15sec) 72 °C (15 sec)] (30 cycles) 72 °C(5min)	pelB *Bd*NEDD8-AGT	**pelB *Bd*NEDD8-*Tt*AA10**
*Tt*AA10	gccctgcgtggt ggcCATGGC TATATTGAA TCGC	tggtgatgagat ccttaTCAAC CAAAGTCCA CGTC	94 °C (30sec) [94 °C (30sec),69 °C (15sec) 72 °C (15 sec)] (30 cycles) 72 °C(5min)	pelB *Ss*NEDD8-AGT	**pelB *Ss*NEDD8-*Tt*AA10**

2.3.6 Soluble Protein Expression Testing

2.3.6.1 Visualisation by Sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS PAGE)

SDS PAGE gels were used to analyse expression levels of target proteins. SDS PAGE gels were made at 12% using the following amounts of materials:

:

Resolving gel -

- 2.5 ml resolving gel buffer (1.5 M TRIS pH 8.8, 0.4% SDS)
- 4.2 ml acrylamide
- 3.2 ml water
- 16µl Tetramethylethylenediamine
- 100 µl 10% ammonium persulfate

Stacking gel -

- 1.3 mL stacking gel buffer (0.5 M TRIS pH 6.8, 0.4% SDS)
- 0.5 mL acrylamide
- 3.2 mL water
- 8 µl Tetramethylethylenediamine
- 100 µl 10% ammonium persulfate
- 8 µl 1% Bromophenol blue

SDS PAGE gels were run at 200 V for 50 minutes using 10 % SDS as running buffer. Gels were stained using Magic Dye (150mg Coomassie Brilliant Blue (Thermo Fisher), 2.49 L water, 8.6mL HCl).

Samples were prepared using SDS PAGE loading dye (2X) by mixing an equal volume of protein samlpe to loading dye. Samples were heated at 95 °C for 5 minutes prior to loading onto the gel. Low and broad range molecular weight SDS PAGE ladders were purchased from Biorad.

The loading dye was made up of the following material:

- 4.8mL Milli-Q water
- 1.2mL 0.5M TRIS HCl pH6.8
- 1.0mL Glycerol
- 2.0mL 10% SDS
- 0.5mL 1% Bromophenol blue
- 0.5ml B-mercaptoethanol

2.3.6.2 Expression testing

Expression of protein targets was tested in several different *E. coli* expression strains, described in Table 5. Small test cell cultures (LB media with either kanamcyin (30 µg mL^{-1}) or ampicillin (100 µg mL^{-1}) and/or chloramphenicol (30 µg mL^{-1}) depending on construct/expression system) were analysed for target expression using Bugbuster Protein Extraction Reagent (using $1/5^{th}$ culture volume for cell lysis, Sigma Aldrich) and SDS PAGE visualisation (12%, 50 mins, 200 mV). Cultures were inoculated using glycerol stocks of Cultures were grown at 37 °C with shaking until an OD of 0.6 or above was reached. Target expression was induced with IPTG (1 mM final concentration) followed by incubation at 16 °C overnight with shaking. Further expression testing was carried out on different targets using different *E. coli* cell strains and specific expression conditions. GHs not produced in a soluble form in the BL21 strain were also screened with SHuffle T7 (New England Biolabs) (incubation with IPTG was carried out at 37°C as well as 16°C overnight) and with a set of chaperone co-expression strains, pGro7 cells (Takara), where co-expression was controlled by the addition of L-arabinose at the start of culture growth. *Tt*AA10 constructs were all initially screened for soluble protein expression using *E. coli* cells BL21*, followed by SHuffle T7 and BL21 (DE3). The construct *Ss*NEDD8 *Tt*AA10 was tested for soluble protein expression using cell strain pGro7.

Table 8 *Tt*AA10 constructs and the type of E.coli expression strain that was tested. A dash indicates unsuccessful transformation into expression cells.

*Tt*AA10 Construct	*E.coli* Expression Strain
pelB SUMO *Tt*AA10	BL21*, SHuffle T7
*Bd*SUMO *Tt*AA10	BL21*, SHuffle T7
*Bd*NEDD8 *Tt*AA10	-
*Ss*NEDD8 *Tt*AA10	BL21*, SHuffle T7, Takara Chaperone plasmid set
pelB *Bd*SUMO *Tt*AA10	BL21*, SHuffle T7
pelB *Bd*NEDD8 *Tt*AA10	-
pelB *Ss*NEDD8 *Tt*AA10	-
pelB *Tt*AA10	BL21*,BL21 (DE3), SHuffle T7
SUMO *Tt*AA10	BL21 (DE3), SHuffle T7
PNT Trx SUMO *Tt*AA10	BL21 (DE3), SHuffle T7
His MBP SUMO *Tt*AA10	BL21 (DE3)
*Tt*AA10-His	BL21 (DE3)

2.3.6.3 Expression testing of SUMO-*Tt*AA10 under Stress Conditions

SUMO-*Tt*AA10 was tested for expression under stress conditions. Conditions included Lysogeny broth media containing an additional 500 mM NaCl or 5 % glycerol and both 500 mM NaCl and 5 % glycerol. One culture containing both additives was subjected to heat shock: the culture tube was heated in a water bath (approx. 47 °C) for 10 minutes before induction with IPTG (1 mM). Expression was carried out overnight at 16 °C, samples harvested by treatment with Bugbuster and fractions analysed on a 12 % SDS PAGE. The remaining culture (9 ml) was lysed by sonication in a buffer containing 20 mM Tris, 250 mM NaCl and 30 mM imidazole and both unlysed and lysed soluble/insoluble fractions analysed by gel electrophoresis.

2.3.7 Optimised Culture Growth of *Tt*AA10-Strep

Final expression conditions (expression strain, growth temperature, IPTG concentration) were chosen based on the amount of soluble protein produced as visualised by SDS PAGE. Large cultures (6 x 500 ml LB, kanamycin 30 µg/ml) were inoculated with 500 µl starter culture (grown overnight) and grown at 37 °C, 200 rpm, until an OD of 0.6 was obtained. IPTG (1 mM final concentration) was then added and cultures cooled to 16 °C and left overnight.

*Tt*AA10-Strep was initially made on a large scale by Luisa Elias (University of York) using Rosseta *E. coli* competent cells. Subsequent batches made by myself were expressed using Tig Chaperone cells (Pgro7 Chaperone set TAKARA). Culture media, (LB, Ampicillin (100 ug/ml), chloramphenicol (35 ug/ml), was inoculated with starter culture and L-arabinose (0.5 g L^{-1} final concentration) added to induce expression of the Tig chaperone. Cultures were grown at 37 °C until the OD was approximately 0.6. The cultures were left to cool slightly before IPTG was added to a final concentration of 1 mM. Cultures were incubated overnight at 16 °C with shaking.

2.3.8 Protein purification of GHs

Cultures were harvested by centrifugation and the pellet resuspended in buffer A (50 mM HEPES, 250 mM NaCl, 30 mM imidazole, pH 7). The sampled was then sonicated (30 s on, 30 s off x5, in an ice bath) and the lysed cell mixture centrifuged at 15 *g* for 30 minutes. The supernatant was collected and loaded onto a pre-equilibrated Ni HiTrap Crude 5ml affinity column (GE Healthcare Life Sciences). Ni affinity chromatography was carried out using the ÄKTA Start (GE Healthcare Life Sciences), with an elution gradient of buffer B (50 mM HEPES, 250 mM NaCl, 300 mM Imidazole, pH 7), 30 -300 mM imidazole over 25 column volumes. Peak fractions containing the protein were pooled and concentrated down for concentration measurement (Nanodrop). Protein samples were diluted to approximately 1mg ml^{-1} to dilute the imidazole concentration and GH targets treated with 3C protease (1:100 ratio 3C protease:target protein), DTT (5 mM) and left at 4°C overnight to remove the hexahistidine tag. The cleavage products were loaded onto a pre-equilibrated (buffer A) Ni HiTrap Crude 5ml affinity column and the flow through and wash collected. The collected sample was concentrated to approximately to 1 ml or less and loaded on a Superdex gel filtration column (S75 or S200 depending on protein size, GE Healthcare Life Sciences) and run using an ÄKTA Start with buffer C (50 mM HEPES, 250 mM NaCl). Peak fractions were combined,

concentrated down and exchanged into a 10 mM HEPES storage buffer. A wide variety of protein yields were obtained depending on the target protein being purified.

2.3.9 LPMO- Streptavidin chromatography

Cultures were harvested by centrifugation, 5000 *g* for 30 min. For each 100 ml of original culture the pellet was gently resuspended in 5 ml of ice cold buffer A (50 mM Tris, 20% v/v sucrose, pH 8) and left on ice for 30 minutes with occasional mixing. The cell suspension was centrifuged again at 8000 rpm for 10 minutes, and the cells subjected to osmotic shock; pellet was resuspended in ice cold buffer B (1 mM MgSO4 + protease inhibitor, 5 ml per 100 ml of initial culture) and left on ice for 30 minutes with occasional mixing. The suspension was centrifuged again and the supernatant collected, filtered and diluted to make up a 1X PBS solution (using stock 10X PBS, pH 7.4).

The sample was loaded onto a strep column (GE Healthcare) pre-equilibrated in buffer D (1X PBS, pH 7.4). The column was washed with 1X PBS buffer for 5 column volumes. The protein was eluted from the column using 5 column volumes of buffer D (1X PBS, 2.5 mM desthiobiotin, pH 7.4) and protein collected.

The dilute *Tt*AA10-strep protein sample was copper loaded by incubation with $CuSo_4$, (1 mM final concentration equivalent to 10X protein concentration) at 4 °C overnight. The protein sample was filtered and concentrated down to less than 1 ml. The concentrated protein sample was loaded onto an S75 gel column and run using buffer A to remove the excess copper.

2.3.10 Protein Quality Analysis

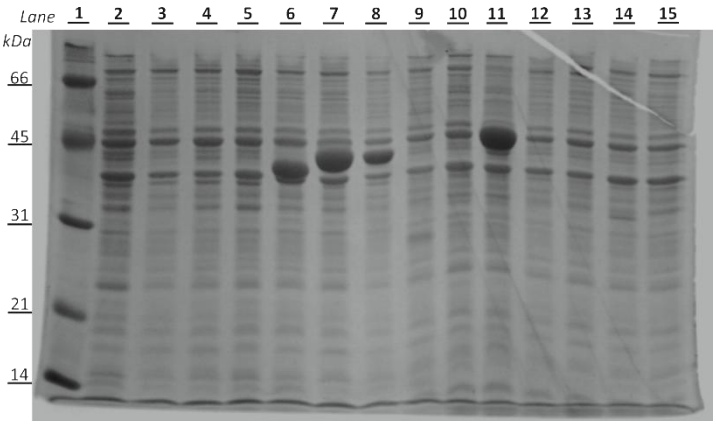
During all protein preparations SDS PAGE (12%, 50 mins, 200mV) were run using aliquots from each step of purification, including samples from cell lysates, column loads, column flow through, cleavage reactions and peak fractions. Final sample purity was assessed with mass spectrometry (ESI/TOF) and in one instance SECMALS (Andrew Leech, Biology TF). For mass spectrometry analysis, protein samples were buffer exchanged into 2 mM Tris pH 8.

2.4 Results: Production of Tt GH targets

2.4.1 Expression Testing of GH Targets

All 14 GH targets were initially tested for soluble protein expression using cell line BL21 (DE3), culture growth at 37 °C, followed by addition of IPTG before overnight incubation at 16 °C. Culture samples were analysed for soluble protein expression by comparison of soluble material from cultures induced with IPTG and a negative control (no IPTG), using SDS PAGE. Figure 25 shows the results of this first expression strategy, where 4 targets are found to be successfully over expressed as the desired soluble protein. The four proteins, shown in lanes 6, 7, 8 and 11 of Figure 25, *Tt*GH5_un (3751) (40.4 kDa), *Tt*GH5_4 (3361) (40.6 kDa), *Tt*GH5_53 (0427) (53.7 kDa) and *Tt*GH8(4506) (45.1 kDa) respectively, were observed at their expected molecular weight.

The soluble expression of the 10 remaining GH targets was tested with SHuffle T7 cells. After induction of the cultures with IPTG, incubation was tested at either at 37 °C for 2 hours (Figure 26A) or 16 °C overnight (Figure 26B). In the higher temperature test, Figure 26A, only lane 11 shows a small amount of soluble protein compared with the control; *Tt*GH12 (0353) (35.8 kDa). Incubation of the test cultures for a longer period at 16 °C resulted in a high yield of soluble expression of *Tt*GH5_2 (2895), lane 12 Figure 26B. A slight increase in yield was also seen for *Tt*GH12 (0353), lane 5, Figure 26B. The test cultures were also analysed for production of insoluble protein, by analysis of the solid pellets and it was found that *Tt*GH9 (0607), *Tt*GH9 (0645), *Tt*GH5_26 (0183) produced large amounts of insoluble protein (data not shown).



Samples:
1)Low range molecular marker
2)Uninduced control sample
3TtGH5_2 (2895)
4)TtGH5_1 (3565)
5)TtGH5_26 (0183)
6)TtGH5_un (3751)
7)TtGH5_4 (3361)
8)TtGH5_53 (0427)
9)TtGH6 (2898)
10)TtGH6 (2895)
11)TtGH8 (4506)
12)TtGH9 (0607)
13)TtGH9 (0645)
14)TtGH12 (0353)
15) TtGH45 (3400)

Figure 25 Initial expression test of the Tt glycoside hydrolases visualised using SDS PAGE . Lanes contain the samples described to the right of the image. Target proteins were expressed by BL21 *E. coli*. Cultures were grown at 37 °C until an OD above 0.6 was achieved, whereby the cultures were induced with IPTG (1 mM final concentration) and incubated overnight at 16 °C. Lane 2 contains a culture sample in which no IPTG was added and therefore displays no target proteins. This control sample is used as comparison to pick out over expression of target proteins, and bands compared to the molecular marker (Biorad). *Tt*GH5_un (3751) (40.4 kDa), *Tt*GH5_4 (3361) (40.6 kDa), *Tt*GH5_53 (0427) (53.7 kDa) and *Tt*GH8 (4506) (45.1 kDa) all display soluble over expression.

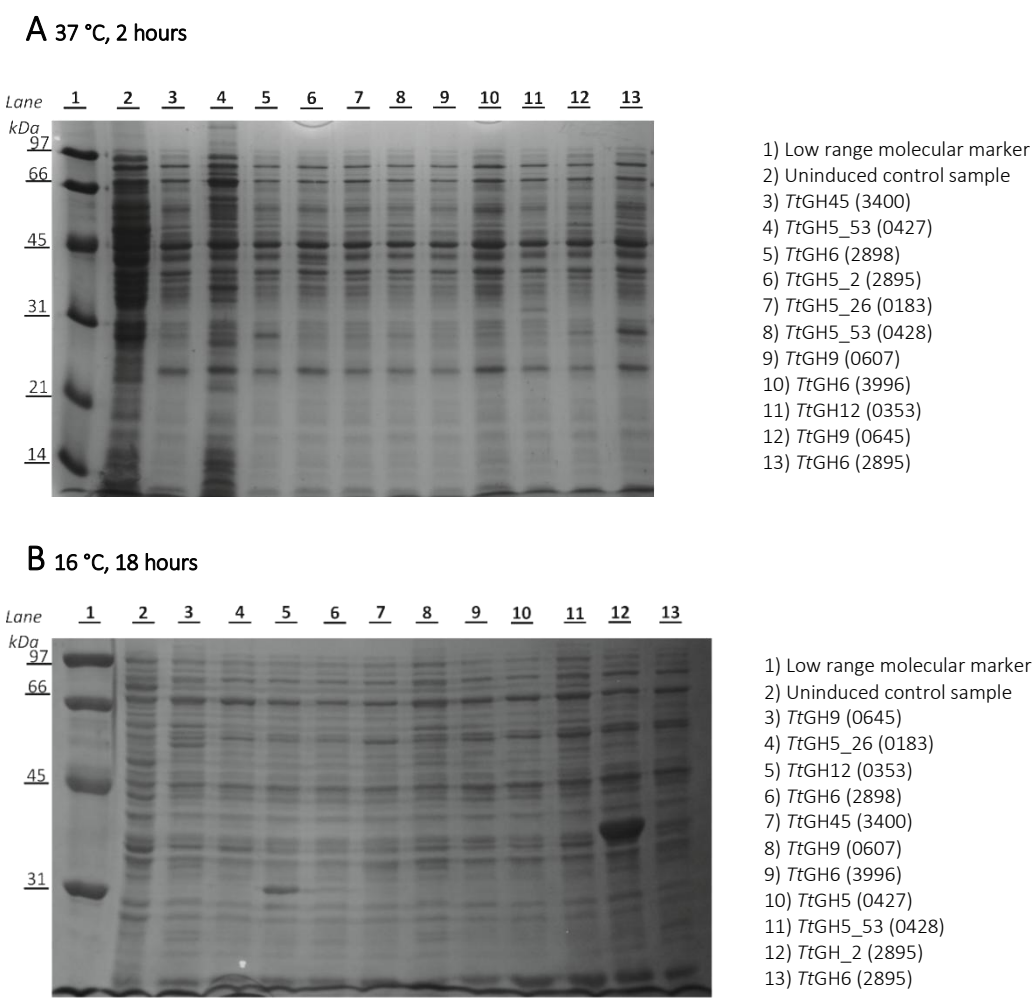
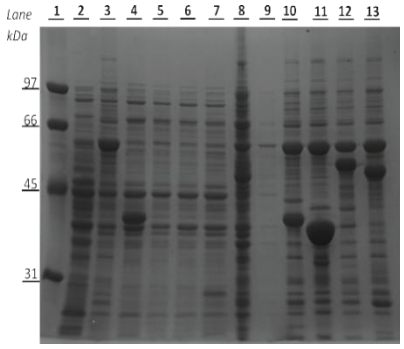


Figure 26 Expression test of the Tt glycoside hydrolases analysed by SDS PAGE. Lanes contain the samples described to the right of the image. Expression was trialled using T7 Shuffle E. coli. Cultures were grown at 37 °C until an OD above 0.6 was achieved, whereby the cultures were induced with IPTG (1 mM) incubated either at 37 °C (A) or 16 °C (B) overnight. Lane 2 (A,B) contains a culture sample in which no IPTG was added and therefore displays no target proteins. This control sample is used as comparison to pick out over expression of target proteins. Lanes 5 and 12 show over-expression bands compared with the controls and the molecular marker (Biorad); TtGH12 (0353) and TtGH5_2 (2895) are 31.3 and 35.8 kDa respectively.

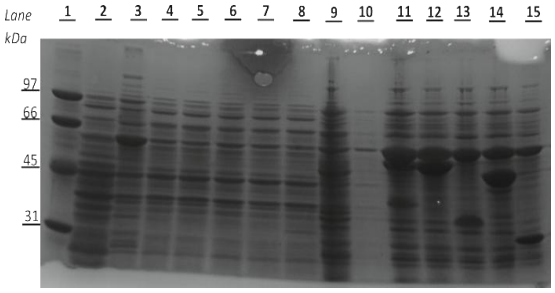
The GH targets were then tested for soluble protein expression using a cell strain, pGro7 from the plasmid chaperone set from TAKARA. Co-expression of a recombinant protein with a chaperone team such as those available in the plasmid chaperone set have been known to improve protein solubility and prevent aggregation [120-121]. Protein expression was tested using co-expression of the chaperone team grosEs-groEL (pGro7), which was induced with L-arabinose. Controls during the experiment were tested based on samples which had no induction and chaperone induction only. It was found that use of IPTG to induce target expression caused the chaperone to become insoluble, **Figure** 27. This may have decreased the likelihood of the targets being successfully expressed in a soluble form. **Figure** 27A-D shows a large number of targets found in the insoluble fraction of the prepared samples. Despite this, *Tt*GH5-2895 was once again found to be highly soluble, lane 4 and 5, **Figure** 27A and 3D respectively. To assess the effect of temperature on soluble protein expression using this system, expression was also tested at a lower temperature and cultures incubated overnight, **Figure** 28. Soluble expression was improved for *Tt*GH12 (0353), but the already large yield of *Tt*GH5_2 (2895) was not affected.

A 37 °C, 2.5 hours



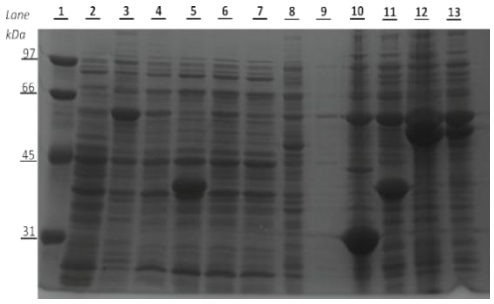
Samples:
1) Low range molecular marker
2) Uninduced control
3) Chaperone only
4) *Tt*GH5_2 (2895)
5) *Tt*GH5_26 (0183)
6) *Tt*GH6 (3996)
7) *Tt*GH6 (2895)
8) Insoluble uninduced control
9) Insoluble chaperone only
10) Insoluble *Tt*GH5_2 (2895)
11) Insoluble *Tt*GH5_26 (0183)
12) Insoluble *Tt*GH6 (3996)
13) Insoluble *Tt*GH6 (2895)

B 37 °C, 2.5 hours



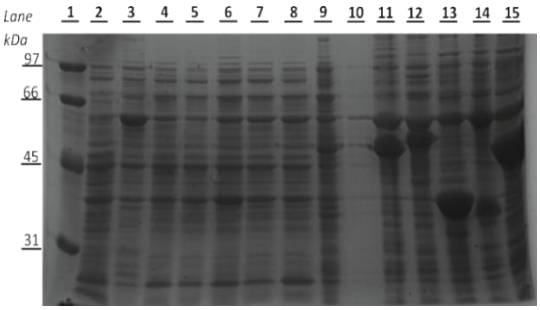
Samples
1) Low range molecular marker
2) Uninduced control
3) Chaperone only
4) *Tt*GH5_53 (0428)
5) *Tt*GH9 (0645)
6) *Tt*GH45 (3400)
7) *Tt*GH9 (0607)
8) *Tt*GH12 (0353)
9) Insoluble uninduced control
10) Insoluble chaperone only
11) Insoluble *Tt*GH5_53 (0428)
12) Insoluble *Tt*GH9 (0645)
13) Insoluble *Tt*GH45 (3400)
14) Insoluble *Tt*GH9 (0607)
15) Insoluble *Tt*GH12 (0353)

C 37 °C, 18 hours



Samples
1) Low range molecular marker
2) Uninduced control
3) Chaperone only
4) *Tt*GH12 (0353)
5) *Tt*GH5_2 (2895)
6) *Tt*GH9 (0465)
7) *Tt*GH5_53 (0428)
8) Insoluble uninduced control
9) Insoluble chaperone only
10) Insoluble *Tt*GH12 (0353)
11) Insoluble *Tt*GH5_2 (2895)
12) Insoluble *Tt*GH9 (0465)
13) Insoluble *Tt*GH5_53 (0428)

D 37 °C, 18 hours



Samples
1) Low range molecular marker
2) Uninduced control
3) Chaperone only
4) *Tt*GH6 (2895)
5) *Tt*GH6 (3996)
6) *Tt*GH5 _26 (0183)
7) *Tt*GH45 (3400)
8) *Tt*GH9 (0607)
9) Insoluble uninduced control
10) Insoluble chaperone only
11) Insoluble *Tt*GH6 (2895)
12) Insoluble *Tt*GH6 (3996)
13) Insoluble *Tt*GH5_26 (0183)
14) Insoluble *Tt*GH45 (3400)
15) Insoluble *Tt*GH9 (0607)

Figure 27 Chaperone co-expression test of the Tt glycoside hydrolases using pGro7 cells (Takara), visualised by SDS PAGE. Chaperone (grosEs-groEL) was induced with L-arabinose at the start of culture growth. Expression of target proteins was achieved with IPTG (1 mM final concentration) at an O_D above 0.6. Cultures were either induced with IPTG *(continued…)*

(…continued) and incubated at 37 °C for a further 2.5 hours (A,B) or 18 hours (C,D). Gels A and C both show over expression of soluble TtGH5-2895 (36 kDa). Gel D shows possible soluble expression of TtGH5-0183 (37 kDa), albeit in a low amount.

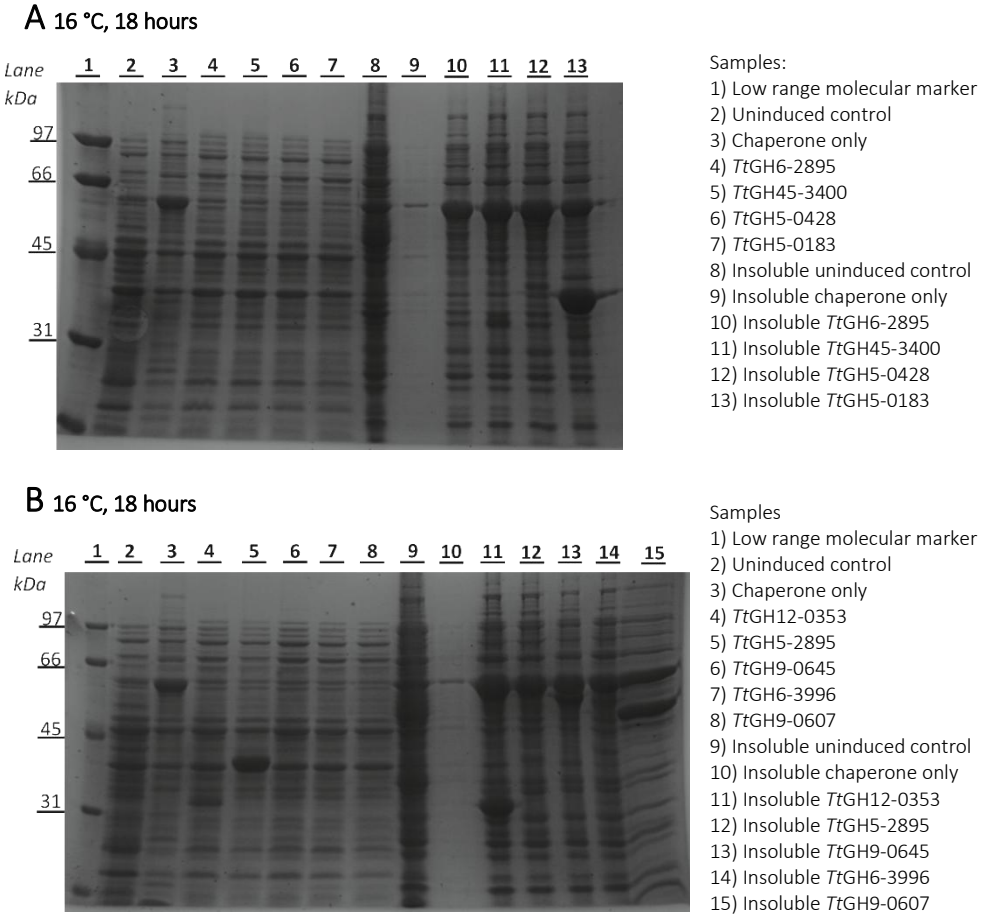


Figure 28 Lower temperature chaperone co-expression test of the Tt glycoside hydrolases using Prog7 cells (Takara), visualised by SDS PAGE. Chaperone (grosEs-groEL) was induced with L-arabinose at the start of culture growth. Expression of target proteins was achieved with IPTG (1 mM final concentration) at an O_D above 0.6. A,B) Cultures were induced with IPTG and incubated at 16 °C for a further 18 hours. Samples are described below each image. Two controls were used, uninduced (no L-arabinose or IPTG) and chaperone only expression (L-arabinose only). Gel B shows over expression of TtGH5-2895 (35.8 kDa). Soluble expression of TtGH12-0353 (31 kDa) is also seen in gel B, where the protein is found in both soluble and insoluble fractions. Comparison of the bands to the molecular marker (Biorad) suggests the proteins are the expected size.

2.4.2 GH Protein Production

Protein expression testing indicated that several *T. turnerae* GH targets could be produced in soluble forms on a test scale (approx. 10 ml), as described in Table 9.

Table 9 Summary of best expression conditions to yield soluble protein

Protein	Expression Cell Strain	Conditions
*Tt*GH8	BL21	37 °C culture growth to OD 0.6, then IPTG induction, 16 °C overnight
*Tt*GH12	pTf16	37 °C culture growth containing L-Arabinose for chaperone expression to OD 0.6, then IPTG induction, 16 °C overnight
TtGH5_2	pGro7	37 °C culture growth containing L-Arabinose for chaperone expression to OD 0.6, then IPTG induction, 16 °C overnight
*Tt*GH5_4	BL21	37 °C culture growth to OD 0.6, then IPTG induction, 16 °C overnight
*Tt*GH5_un	BL21	37 °C culture growth to OD 0.6, then IPTG induction, 16 °C overnight

Large scale (3 L) batch production and purification of the 5 soluble GH targets is shown in the following sections, carried out using the same conditions as described above. It should be noted that soluble expression and large scale production of pure protein was obtained for *Tt*GH5_53 (0428) but little subsequent characterisation was carried out, so this has been omitted from the following work. The targets will henceforth be named by their GH family and subfamily only, as the CAZy identifiers are no longer required to distinguish between targets from the same family; *Tt*GH12, *Tt*GH8, *Tt*GH5_2 *Tt*GH5_4 and *Tt*GH5_un. The *Tt* GH constructs were designed to be used with Ni affinity chromatography, where the incorporated N-terminal hexahistidine region could be used as an Ni column purification affinity tag to separate out the target protein from the soluble fraction of the cell lysate. In the interest of protein characterisation, it is preferable to remove any tag from a target protein to ensure that future experiments are not inhibited or altered by the presence of the solubility or affinity tag. The hexahistidine tag was engineered to be removable by incubation of the tagged protein with 3C protease, which is able to recognise the incorporated cleavage site after the tag and prior to the start of the target sequence. Cleavage by the protease is efficient, but does leave two non-native residues left over at the start of the target sequence. The overall purification strategy used to produce the soluble *Tt* GH targets on a larger scale is shown in Figure 29 where the chromatograms from three different chromatography steps correspond to the purification of *Tt*GH12.

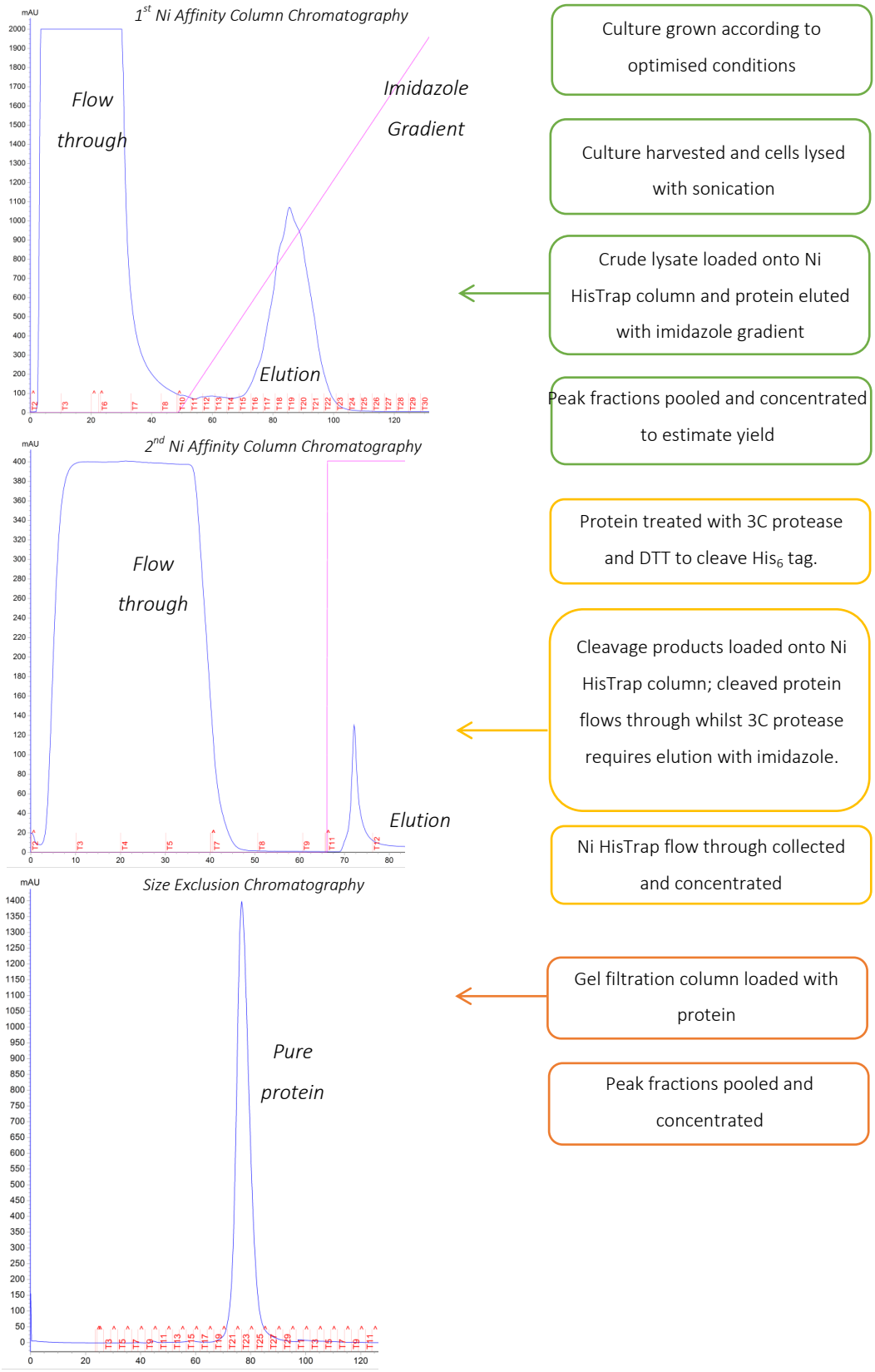


Figure 29 The purification strategy used during the production of all *Tt*GH12 , which was used during production of all *Tt* GH target. The displayed chromatograms show the purification by Ni affinity and size exclusion chromatography, with response of protein absorbance *(...continued)*

(continued...) at A_{280} monitored over time. On each chromatogram, blue lines indicate the absorbance at A_{280} in mAU (flat lines occur as the detector limit is 2000 mAU on the AKTA Start), whereas the pink line indicated the concentration of imidazole (sloping line is gradient elution, whereas upwards line is using 100% imidazole buffer).

2.4.2.1 *Tt*GH12

Large scale expression of *Tt*GH12 was carried out using the T7 Shuffle cell strain. The size exclusion column chromatogram in **Figure** 29 shows a sharp peak, indicating that the protein eluted from the column was homogeneous. Subsequent analysis of the final pure sample of *Tt*GH12 by ESI mass spectrometry confirmed that the protein was indeed pure and at the correct molecular weight, **Figure** 30. The yield was approximately 12 mg purified protein per litre culture media (12 mg/L).

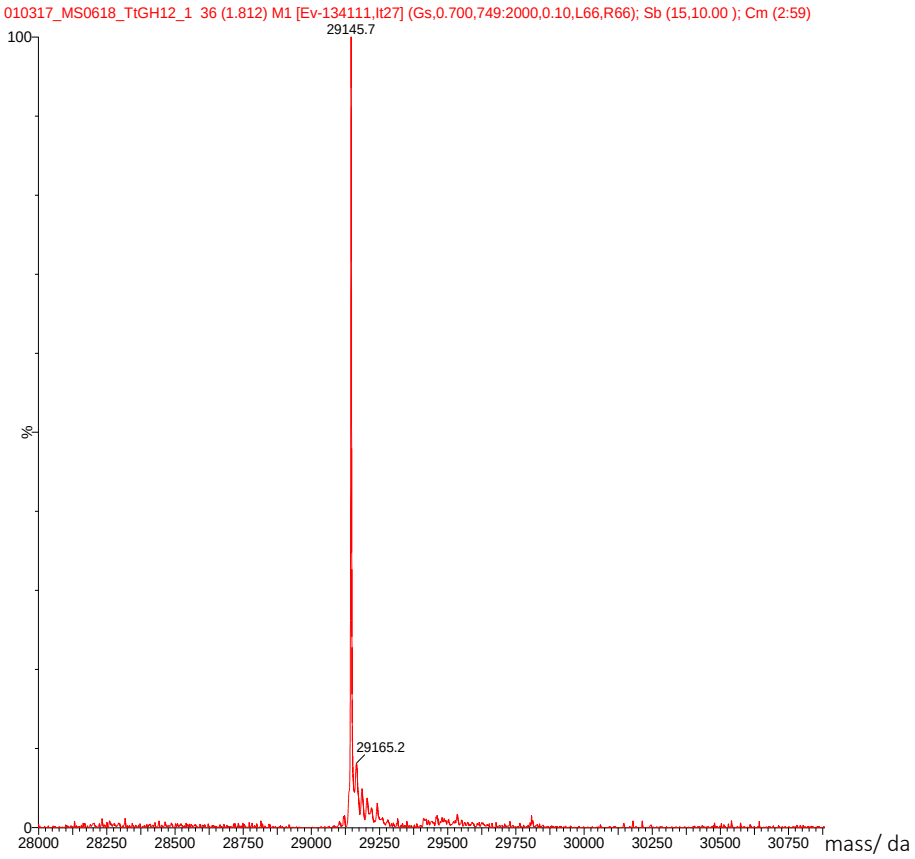


Figure 30 *Tt*GH12 protein molecular weights and purity analysed by mass spectrometry (TOF Ms ESI). A) *Tt*GH12 peak observed at 29145.7 Da, corrected to 29150.5 Da by reference to an external standard (myoglobin). *Tt*GH12 was found to be pure and no other protein peaks observed.

2.4.2.2 *Tt*GH8

*Tt*GH8 was expressed using BL21 *E. coli* strain and **Figure 31** shows the first and third chromatography steps carried out during purification. Samples taken from the various steps of the purification are shown in **Figure 32** by SDS PAGE visualisation. Due to the large amount of protein being purified, the size exclusion chromatogram shows a very broad elution peak, with a shoulder peak beforehand. The broadness of the peak may have hidden some impurities, and as such a few small bands can be seen below the main *Tt*GH8 band in the SDS PAGE shown in lane 8, **Figure 32**. The smaller protein bands may be degradation products, but their concentration compared with the amount of *Tt*GH8 protein present suggested their presence was not significant enough to disrupt further analysis. As such, the protein was tested for purity by ESI mass spectrometry and found to be relatively pure with only a small secondary peak which was suggested to be a salt adduct, **Figure 33.** The overall yield from the purification process was approximately 6 mg/L.

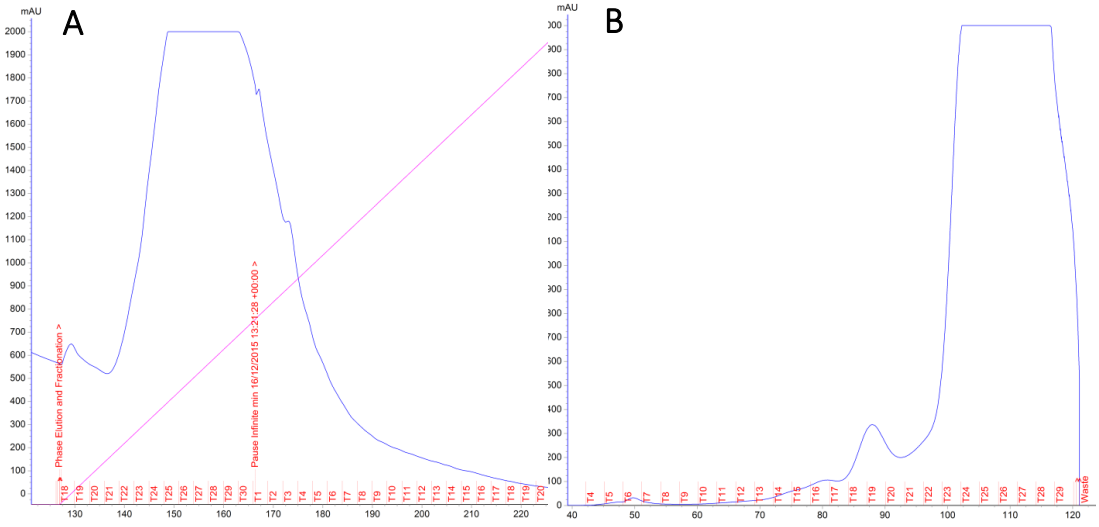


Figure 31 *Tt*GH8 chromatography purification measured as protein absorbance at A_{280} using A) Nickel affinity chromatography on the crude lysate to collect his_6 tagged protein and B) Size exclusion chromatography to collect pure protein.

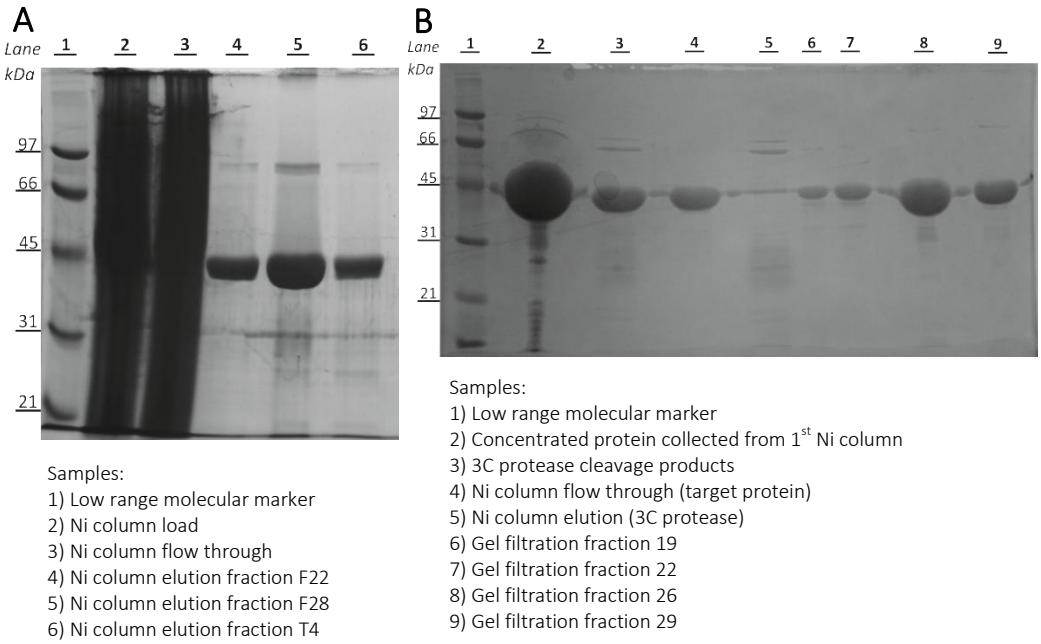


Figure 32 Purification of *Tt*GH8 as shown by SDS PAGE . Lanes contain the samples described below each image. Gel A shows the first step in protein purification, whereby harvested cultures are loaded onto a Ni affinity column and bound protein eluted with an increasing gradient of imidazole. Fractions were taken from the start, middle and end of the protein elution peak obtained during chromatography. Gel B shows the second and third steps in purification; passing the 3C cleavage products through a Ni affinity column and gel filtration. Comparison of the bands to the molecular marker (Biorad) suggests the *Tt*GH8 is the expected size.

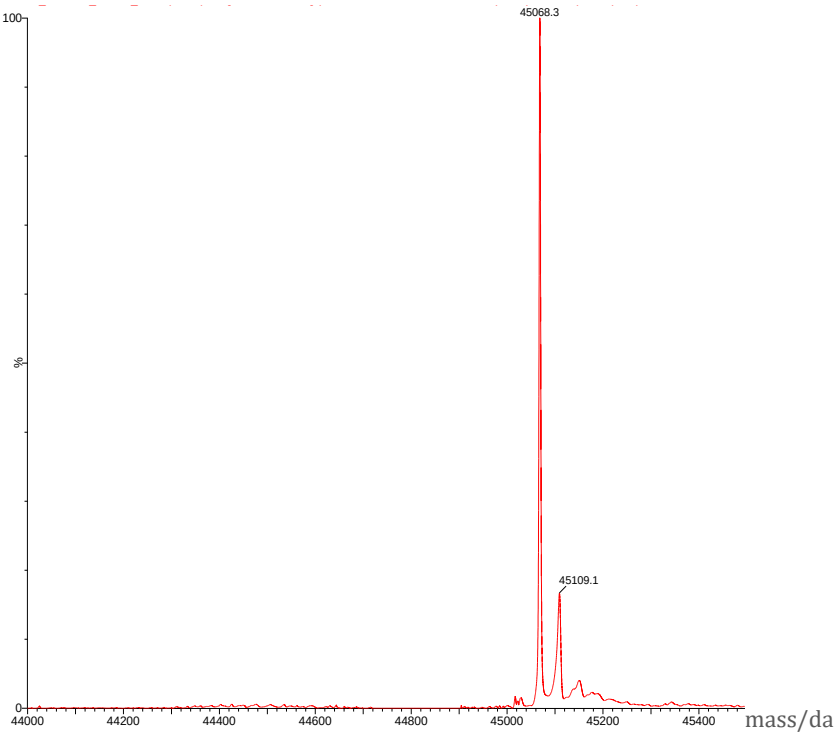
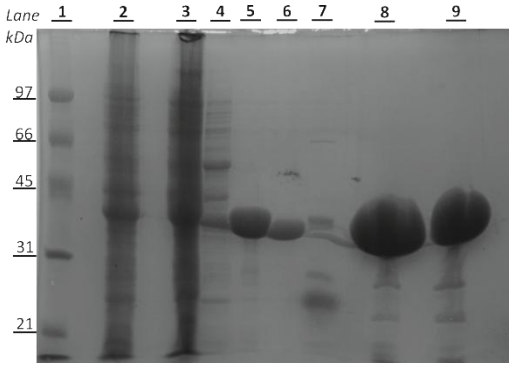


Figure 33 *Tt*GH8 Protein molecular weight and purity analysed by ESI MS . TtGH8 main peak observed at 45068.3 Da, corrected to 45070.9 Da by reference to an external standard (myoglobin). The smaller peak at 45109.1 Da, corrected to 45111.7 Da was suggested to be a salt adduct. TtGH8 was found to be pure and no other protein peaks observed.

2.4.2.3 *Tt*GH5_2

*Tt*GH5_2 was produced by co-expression with the chaperone grosEs-groEL chaperone using the pGro7 *E. coli* cell strain. Samples taken from various stages of purification are shown in Figure 34, where the last lane indicates there to be some smaller distinct bands in the final pure sample. However, the protein was shown to be relatively homogenous during elution from the size exclusion column and as such the bands are possibly degradation products of the protein, Figure 35C. An approximate yield of 4 mg L^{-1} of protein was obtained. *Tt*GH5_2 was tested by ESI mass spectrometry and found to be relatively pure and of the correct molecular weight, Figure 36.



Samples:
1) Low range molecular marker
2) Lysate
3) Ni column load
4) Ni column flow through
5) Ni column elution
6) 3C protease cleavage product
7) Uncleaved protein and 3C protease elution
8) Gel filtration column load
9) Final sample

Figure 34 Purification of *Tt*GH5_2 as shown by SDS PAGE . Lanes contain the samples described to the right of the image. Harvested culture lysate was loaded onto a Ni affinity column and bound protein eluted with an increasing gradient of imidazole. Further purification was carried out by treatment with 3C protease, a second Ni affinity chromatography step to collect unbound target protein and a final polishing step of gel filtration. Comparison of the bands to the molecular marker (Biorad) suggests the *Tt*GH5_2 is the expected size.

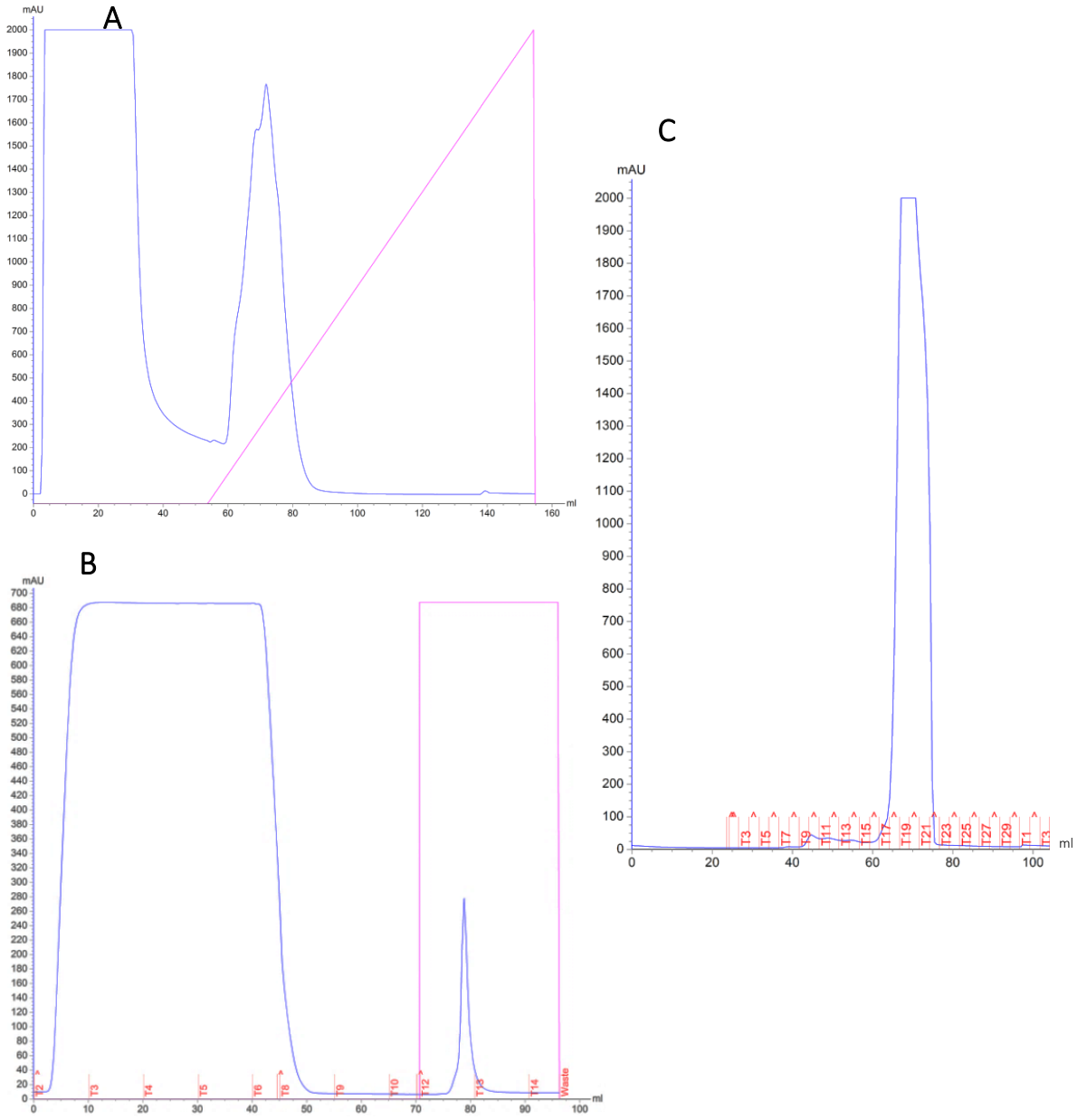


Figure 35 *Tt*GH5_2 chromatography purification steps measured as protein absorbance at A_{280} over time ; A) Crude lysate loaded onto Ni affinity column and protein eluted with imidazole gradient (pink line). B) 3C protease treated protein loaded onto Ni affinity column and flow through collected. C) Size exclusion chromatography.

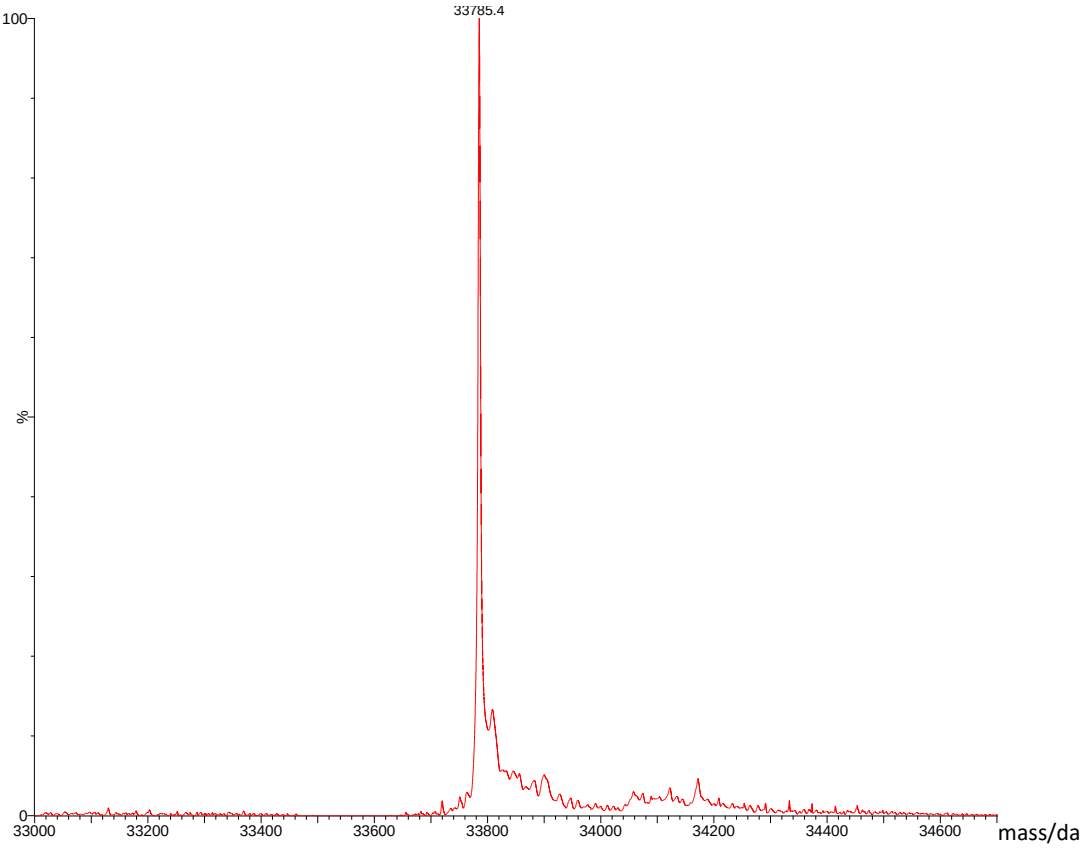
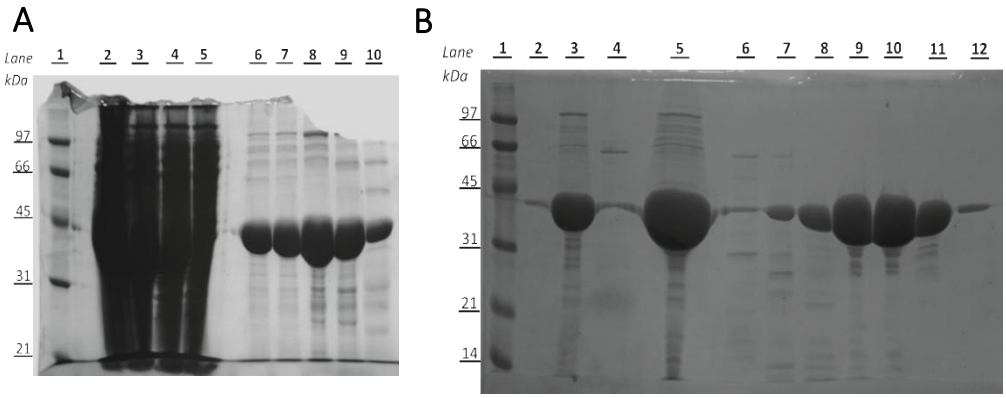


Figure 36 *Tt*GH5_2 protein molecular weight and purity analysed by ESI MS . *Tt*GH5_2 main peak observed at 33785.4 Da, corrected to 33792.1 Da by reference to an external standard (myoglobin). *Tt*GH5_2 was found to be pure and no other protein peaks observed.

2.4.2.4 *Tt*GH5_4

*Tt*GH5_4 was expressed using the BL21 *E. coli* strain. Samples taken from the different stages of protein purification are shown in **Figure 37**, where SDS PAGE analysis summarises the state of the protein in the different steps.**Figure 38Figure** 31 shows the first and third chromatography steps carried out during purification. The protein was tested for purity by ESI mass spectrometry and despite the peaks observed in lanes 7-11, **Figure 37B**, *Tt*GH5_4 was shown to be relatively pure and of the correct molecular weight, **Figure 39**. An approximate protein yield of 6 mg L^{-1} was achieved.



Samples:
1) Low range molecular marker
2) Crude sample
3) Lysate
4) Ni column load
5) Ni column flow through
6) Ni column elution fraction 14
7) Ni column elution fraction 16
8) Ni column elution fraction 20
9) Ni column elution fraction 24
10) Ni column elution fraction 28

Samples:
1) Low range molecular marker
2) Ni column load (3C protease cleavage products)
3) Ni column flow through (target protein)
4) Ni column elution (3C protease)
5) Gel filtration load
6) Gel filtration fraction 15
7) Gel filtration fraction 16
8) Gel filtration fraction 17
9) Gel filtration fraction 18
10) Gel filtration fraction 19
11) Gel filtration fraction 20
12) Gel filtration fraction 21

Figure 37 Purification of *Tt*GH5_4 as shown by SDS PAGE. Lanes contain the samples described below each image. Gel A shows the first step in protein purification, whereby harvested cultures are loaded onto a Ni affinity column and bound protein eluted with an increasing gradient of imidazole. Fractions 14-28 were taken from the start, middle and end of the protein elution peak obtained during chromatography. Gel B shows the second and third steps in purification; passing the 3C cleavage products through a Ni affinity column and gel filtration fractions 15-21. Bands observed as lower molecular weight in the gel filtration fractions are most likely degradation products. Comparison of the bands to the molecular marker (Biorad) suggests the *Tt*GH5_4 is the expected size.

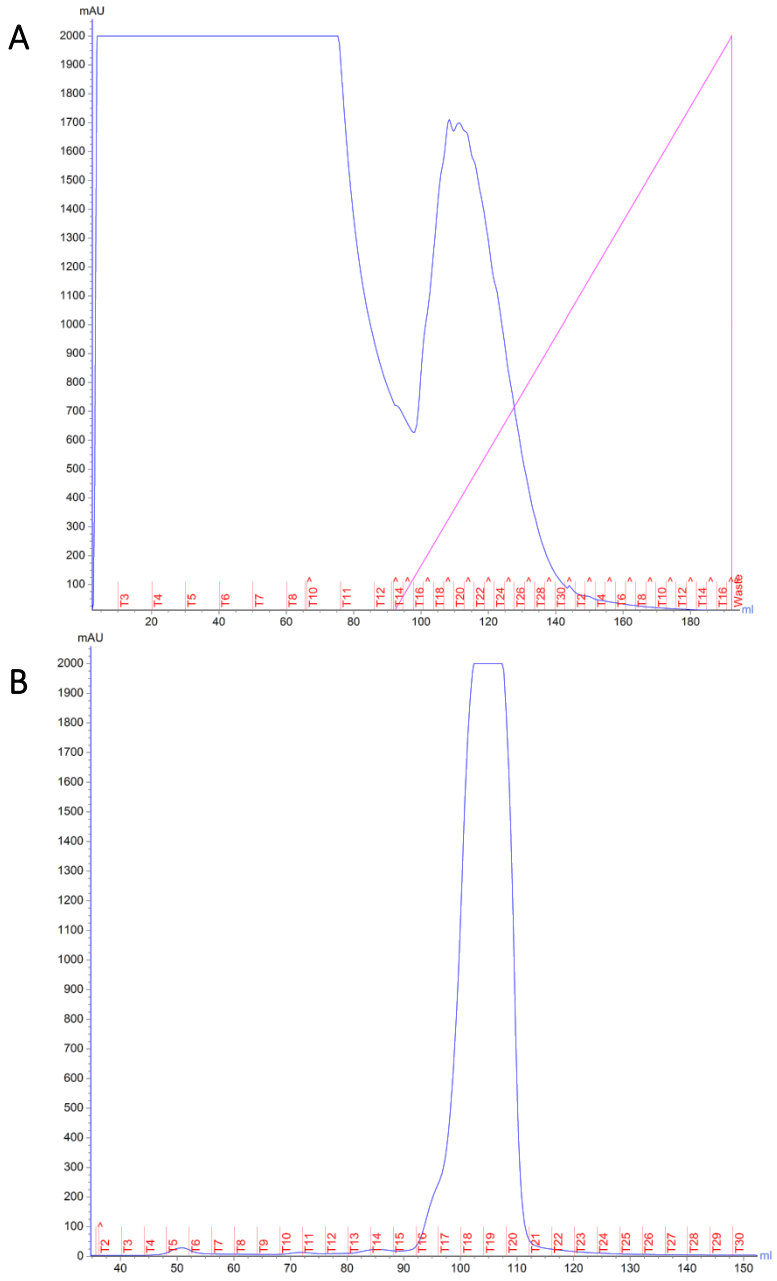


Figure 38 *Tt*GH5_4 chromatography purification steps measured as protein absorbance at A_{280} over time; A) crude lysate run on a Ni affinity column and protein eluted with an imidazole gradient (pink line) and B) size exclusion chromatography.

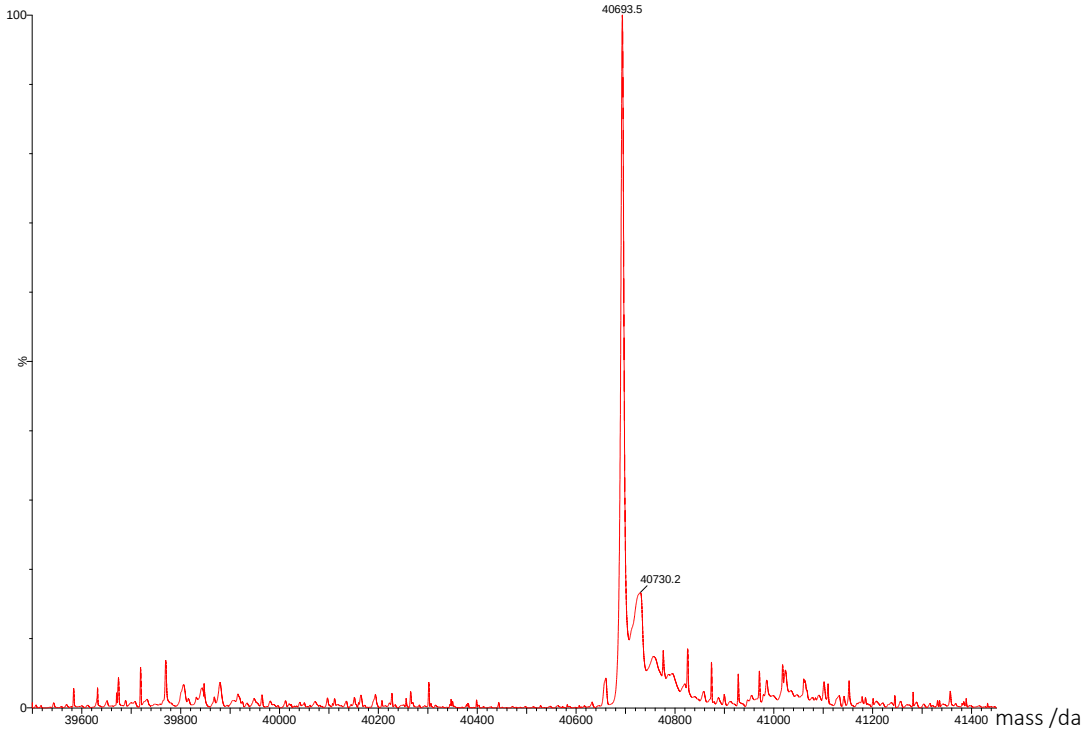
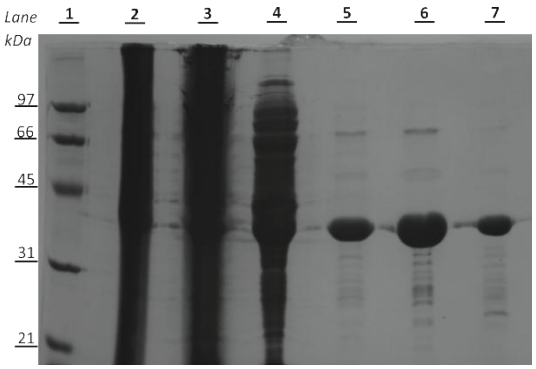


Figure 39 *Tt*GH5_4 protein molecular weight and purity analysed by ESI MS. . TtGH5_4 peak observed at 40693.5 Da, corrected to 40696.6 Da by reference to an external standard (myoglobin). TtGH5_4 was found to be pure and no other protein peaks observed.

2.4.2.5 *Tt*GH5_un

*Tt*GH5_un expression was carried out with the BL21 *E. coli* cell strain. Samples taken from the different steps during the purification are shown visualised by SDS PAGE in Figure 42. The affinity chromatograms from one of the batch purifications are shown in Figure 41. Analysis of the protein by SEC-MALS and ESI mass spectrometry showed that whilst the protein is pure it is not at the expected molecular weight of 40.4 kDa, Figure 42. Instead the protein molecular weight was characterised from two separate protein preparations to be 33.4 kDa, which indicated a loss of 64 amino acids from the protein sequence. A high protein yield of 19 mg L^{-1} was achieved.



Samples:
1) Low range molecular marker
2) Culture lysate
3) Ni column load
4) Ni column flow through
5) Ni column elution fraction T18
6) Ni column elution fraction T22
7) Ni column elution fraction T28

Figure 40 Purification of *Tt*GH5_un- as shown by SDS PAGE. Lanes contain the samples described to the right of the image. This gel shows the first step in protein purification, whereby harvested cultures (lysate) are loaded onto a Ni affinity column and bound protein eluted with an increasing gradient of imidazole. Fractions were taken from the start, middle and end of the protein elution peak obtained during chromatography. Further purification was carried out by treatment with 3C protease, a second Ni affinity chromatography step to collect unbound protein and a final polishing step of gel filtration (not shown). Comparison of the bands to the molecular marker (Biorad) suggests the *Tt*GH5_un around the expected size.

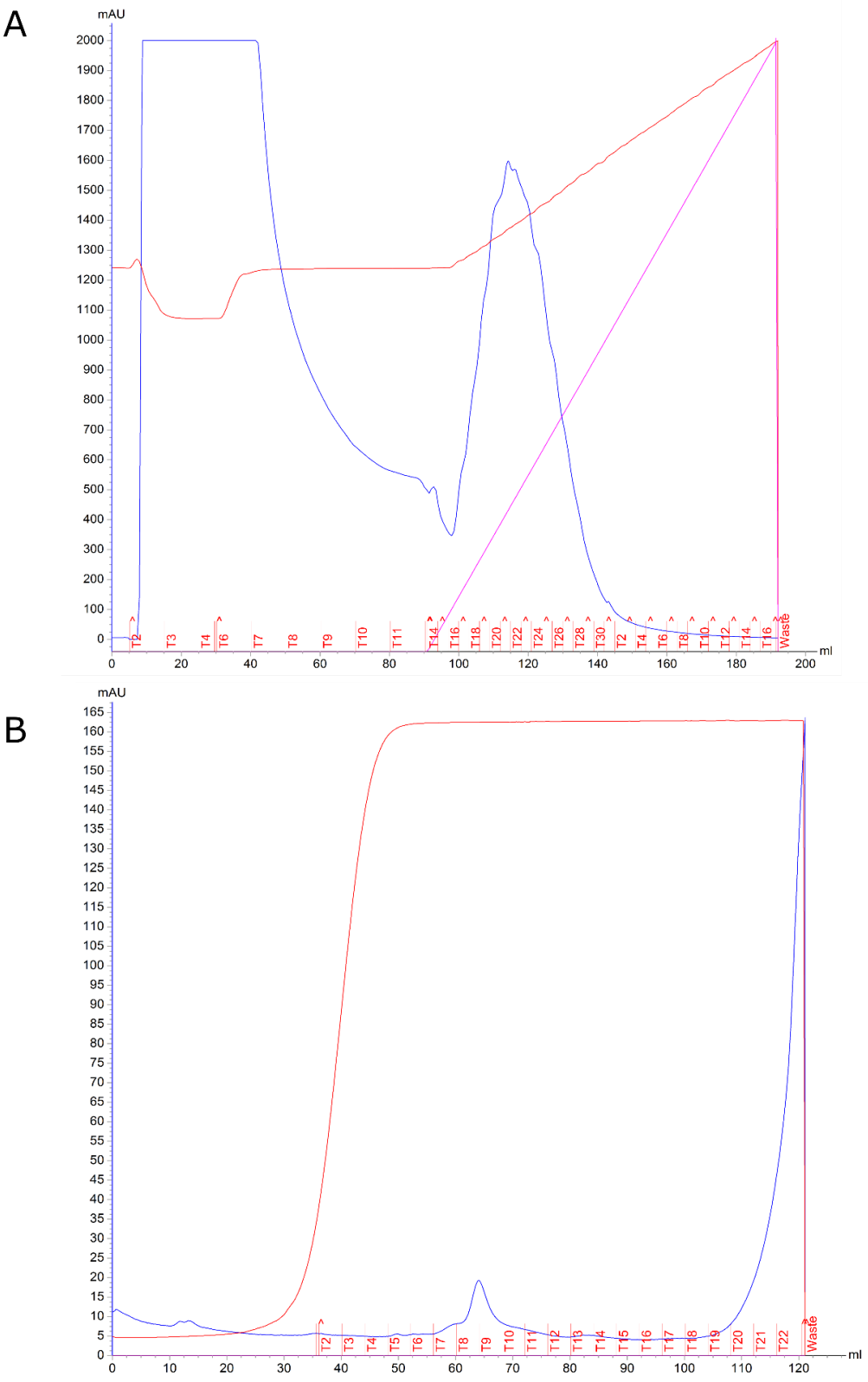


Figure 41 Purification of *Tt*GH5_un , A) Ni affinity chromatography and B) Size exclusion chromatography, where protein eluted at the end of the column, but was collected manually after the set programme ended. Red lines show conductivity, pink buffer gradient and blue is the response measured at A_{280}.

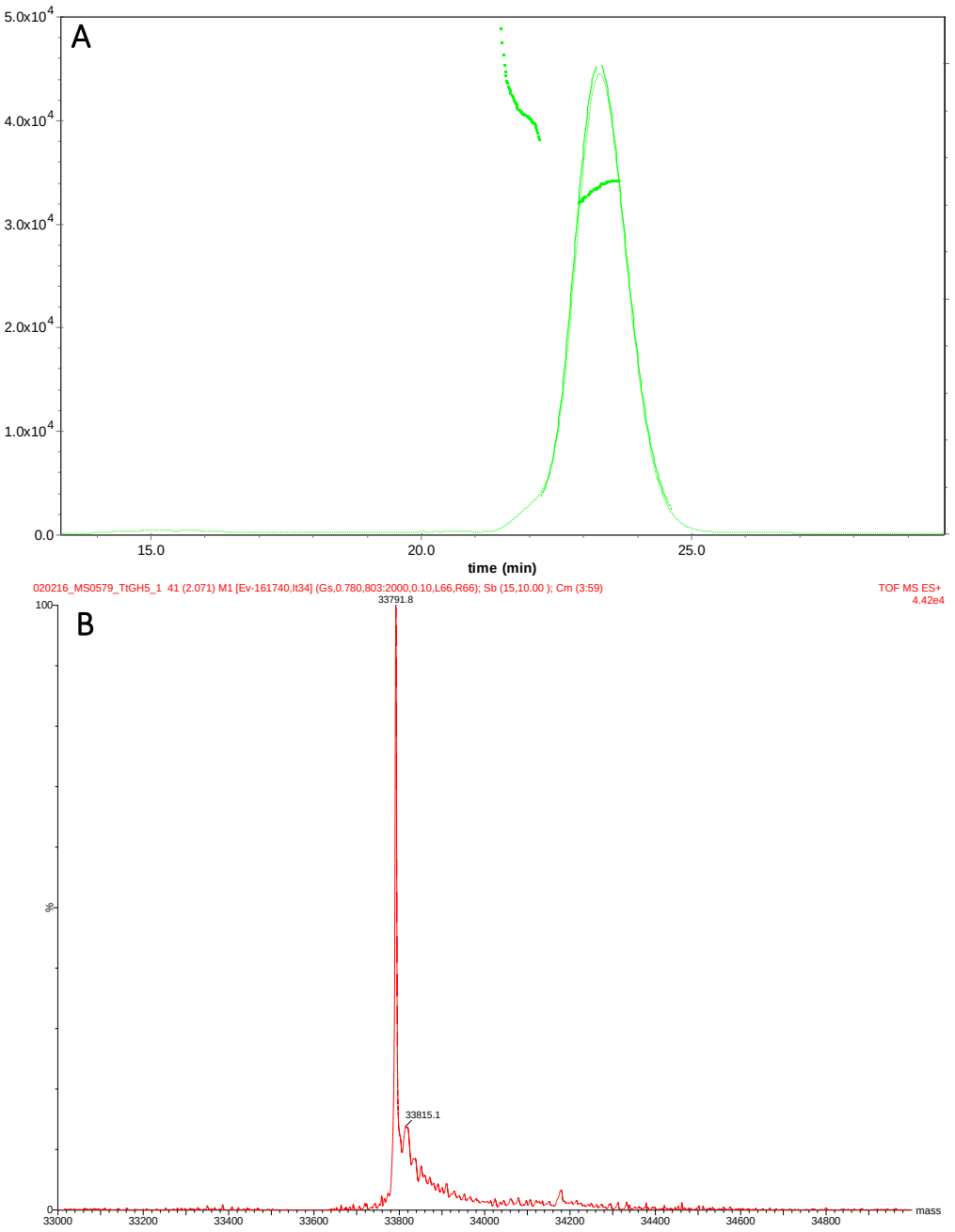


Figure 42 *Tt*GH5_un SEC MALS plotted as molar mass (g/mol) against time. The solid green line is the refractive index, the dashed is light scattering. The line within the central peak is an estimate of the 'weight averaged molecular weight' which equates to 3.32 x10^4 g/mol. The second green line to the left of the main peak indicates the 'weight averaged molecular weight' of the small shoulder peak, at 3.34 x10^4 g/mol.

2.5 Results: LPMO Production

2.5.1 Construct Creation

To make the various constructs, primers were designed to enable extraction of the various sequence building blocks, be it the *Tt*AA10 sequence from the original stock plasmid containing pelB-SUMO-*Tt*AA10 or the solubility tags from different vectors. The 14 different designs are shown in

where the N-terminal region of the construct is built up with the different sequence blocks taken from various vectors, conferring either a signal peptide for target movement to the periplasm and/or a solubility tag. Complete removal of either the signal peptide (by natural cell processing), the SUMO tag, or SUMO-like tag (NEDD8, both removable by specific protease cleavage) would be possible and produce *Tt*AA10 with the native N-terminal Histidine intact and not encumbered with left over residues from a cleavage site. The original SUMO tag was used (taken from the Champion SUMO vector by Dr G.Hemsworth whilst making the stock pelB-SUMO-*Tt*AA10 plasmid used in this work) along with two other similar versions taken from vectors gifted for use in this work by the Frey group, *Bd*SUMO and the SUMO-like *Ss*NEDD8 tags from two different organisms, *Brachypodium distachyon* (*Bd*) and *Salmo salar*(*Ss*).[122] Three constructs were designed to have non removable C-terminal tags, two contained a hexahistidine tag, and the third (designed and made by Luisa Elias) included a Strep tag. The C-terminal position of these tags was thought to be unlikely to affect LPMO activity.

Table 10 Construct designs to try improve soluble expression level of *Tt*AA10. Construct 14 was designed and made by Luisa Elias, CNAP, University of York.

	N-terminal Tag				Target Protein	C-terminal Tag	Expected weight (KDa)
1		pelB	SUMO		*Tt*AA10		38.5
2			*Bd*SUMO		*Tt*AA10		36.3
3			*Bd*NEDD8		*Tt*AA10		34.3
4			*Ss*NEDD8		*Tt*AA10		34.4
5		pelB	*Bd*SUMO		*Tt*AA10		38.4
6		pelB	*Bd*NEDD8		*Tt*AA10		36.4
7		pelB	*Ss*NEDD8		*Tt*AA10		36.5
8			SUMO		*Tt*AA10		36.4
9			pelB		*Tt*AA10		25.1
10		pelB	SUMO		*TtAA10*	His	39.32
11					*Tt*AA10	His	23.75
12	His	MBP	*Linker (3C Cleavage site)*	SUMO	*Tt*AA10		78.79
13		Thioredoxin	*Linker (TEV cleavage)*	SUMO	*Tt*AA10		48.29
14				pelB	*TtAA10*	Strep	24.0

- Several different vectors were required to make the 14 *Tt*AA10 constructs. An example agarose gel is shown in **Figure 43**, where three vectors have been linearised – PCR reactions to create linear double stranded DNA using specific primers as described earlier in

Table 6. The PCR of pET28 YSBL LIC 3C vector removed the YSBL LIC region, removing both the N-terminal hexahistidine tag and the 3C cleavage site. This left only the C-terminal hexahistidine tag ready for incorporation of the *Tt*AA10 insert, by ligation. The PCR of pETFPP2 His-MBP-3C amplified vector sequence without MBP-3C and the C-terminal hexahistidine tag. The PCR pf pnT-TRxt vector sequence was amplified after the TEV cleavage site (forward direction) and after the LIC cleavage site (reverse) leaving the thioredoxin tag, as the soluble tag was intended to be positioned prior to the SUMO tag of *Tt*AA10.

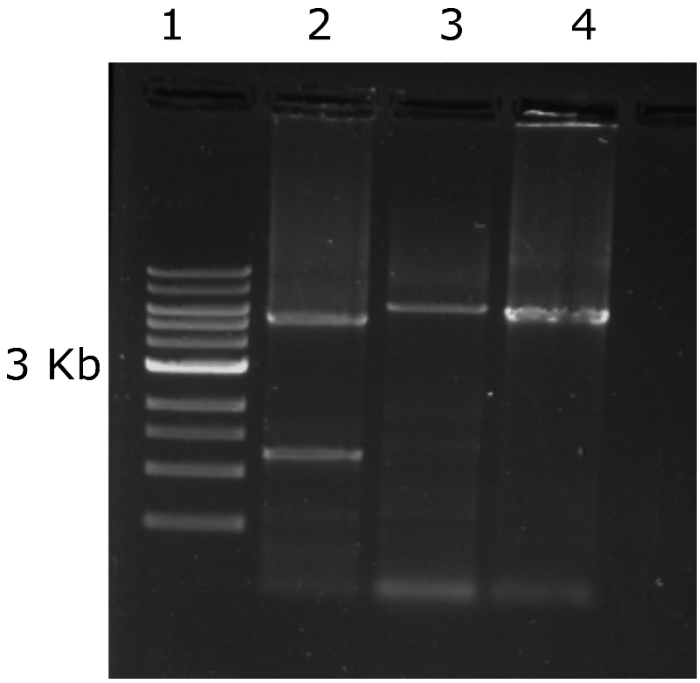


Figure 43 Agarose gel of the linear vectors required to build some of the *Tt*AA10 constructs. Lanes are the following, 1) DNA ladder, 2) pET28 YSBL LIC 3C, 3) pETFPP2 His-MBP-3C and 4) pNT-Trxt.

Inserts were made by designing different primers to anneal to specific sections of the target gene sequence, as described in Table 7, with sequence overhangs which were complementary base pairs on the intended vector. The amplified double stranded DNA fragments were ligated to the vectors via the complementary base pair overhangs (sticky ends). Figure 44 shows an example agarose gel of four inserts prepared from the stock plasmid, each containing different overhangs and tag 'building blocks'.

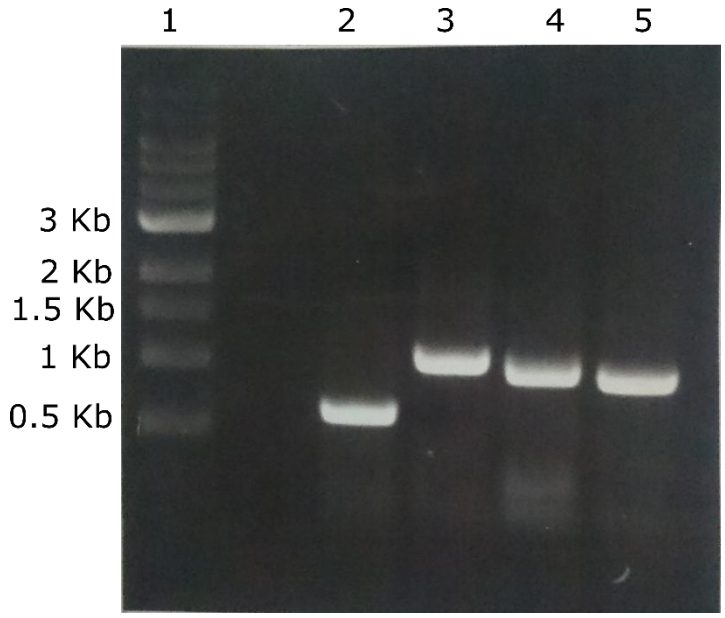
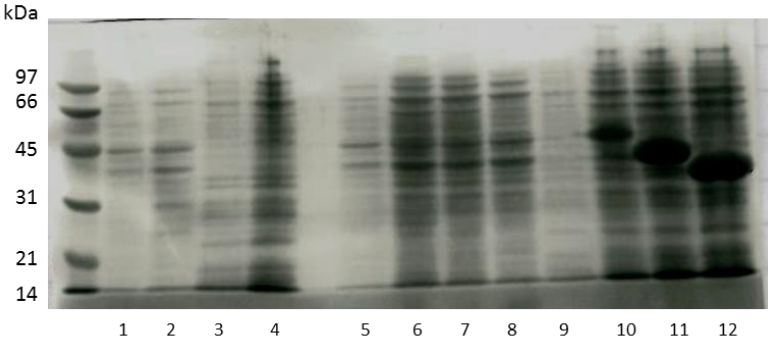


Figure 44 Agarose gel showing four TtAA10 inserts with different overhangs ready for incorporation into the various linear vectors. All inserts were made from a stock plasmid containing pelB-SUMO-*Tt*AA10. Lanes 1) DNA ladder, 2) *Tt*AA10 to make the *Tt*AA10-His construct, 3) pelB-SUMO-*Tt*AA10, 4) SUMO-*Tt*AA10 with overhang for the MBP construct and 5) SUMO *Tt*AA10 with overhang for the thioredoxin vector.

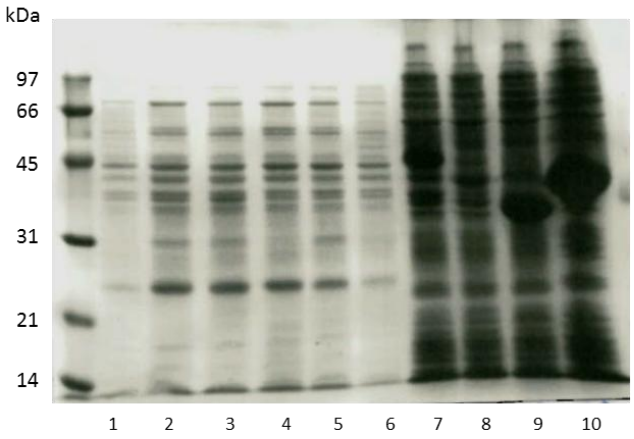
2.5.2 Expression Testing using different cell strains

Constructs 1-7 (Table 10) were tested for soluble protein over expression at different times using BL21* and SHuffle T7 competent cells. Figure 44 shows constructs 1-4 being tested for the presence of soluble or insoluble target protein. It should be noted that the SDS PAGE also shows several constructs of *Sd*AA10 which were being tested alongside *Tt*AA10 during the first year of this work but were subsequently not taken further than expression testing. None of the constructs shown in Figure 45 were produced in a soluble form. A large amount of insoluble protein was observed for *Bd*SUMO-*Tt*AA10 and *Ss*NEDD8-*Tt*AA10. A further expression test was carried out on the SUMO constructs 5-7 which contained an N-terminal pelB signal peptide, but no expression was observed (data not shown). It was noted that during expression testing, no over expression of any construct containing a pelB signal peptide was observed, not even as an insoluble aggregate. Constructs 11-13 were tested for soluble protein production using BL21* and SHuffle T7 *E. coli* cell strains, Figure 46. Despite the incorporation of two solubility tags (MPB or thioredoxin as well as the SUMO), the proteins were found in the insoluble cell extract, and expression was found to not be under the control of IPTG induction. A small band at the approximate molecular weight for the Trxt-SUMO-*Tt*AA10 protein was observed in the soluble fraction, lane 7, Figure 46. No expression was observed for *Tt*AA10-His (data not shown).



A BL21*

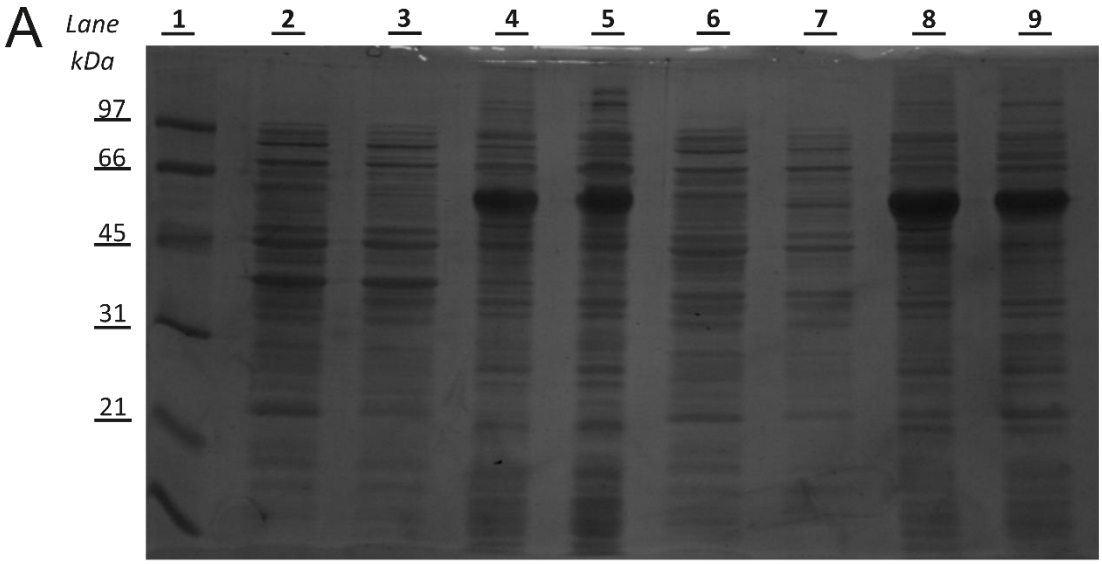
1) Uninduced pelB-SUMO-*Tt*AA10 Soluble
2) pelB-SUMO-*Tt*AA10 Soluble
3) uninduced pelB-*Tt*AA10 Insoluble
4) pelB-SUMO-*Tt*AA10 Insoluble
5) Uninduced SsNEDD8 *Tt*AA10 Soluble
6) *Bd*SUMO *Tt*AA10 Soluble
7) *Ss*NEDD8 *Tt*AA10 Soluble
8) *Ss*NEDD8 *Sd*AA10 Soluble
9) Uninduced *Ss*NEDD8 *Tt*AA10 Insoluble
10) *Bd*SUMO *Tt*AA10 Insoluble
11) *Ss*NEDD8 *Tt*AA10 Insoluble
12) *Ss*NEDD8 *Sd*AA10 Insoluble



B SHuffle T7

1) Uninduced *Bd*SUMO-*Tt*AA10 Soluble
2) *Bd*SUMO-*Tt*AA10 Soluble
3) *Bd*SUMO-*Sd*AA10 Soluble
4) *Bd*NEDD8-*Tt*AA10 Soluble
5) *Ss*NEDD8-*Tt*AA10 Soluble
6) Uninduced *Bd*SUMO-*Tt*AA10 Insoluble
7) *Bd*SUMO-*Tt*AA10 Insoluble
8) *Bd*SUMO-*Sd*AA10 Insoluble
9) *Bd*NEDD8-*Tt*AA10 Insoluble
10) *Ss*NEDD8-*Tt*AA10 Soluble

Figure 45 SDS PAGE results for expression tests of *Tt*AA10 and *Sd*AA10 in BL21* and T7 SHuflle. *Bd*SUMO-*Tt*AA10, *Ss*NEDD8-TtAA10 and *Ss*NEDD8-*Sd*AA10 expressed in the insoluble fraction in BL21*. *Bd*SUMO-*Tt*AA10, *Bd*NEDD8-*Sd*AA10 and *Ss*NEDD8-*Tt*AA10 also expressed insolubly in SHuffle T7.



1)Low molecular weight marker

2)Uniduced BL21 Trxt-SUMO-*Tt*AA10 Soluble

3)BL21 Trxt-SUMO-*Tt*AA10 Soluble

4) Uniduced BL21 Trxt-SUMO-*Tt*AA10 Insoluble

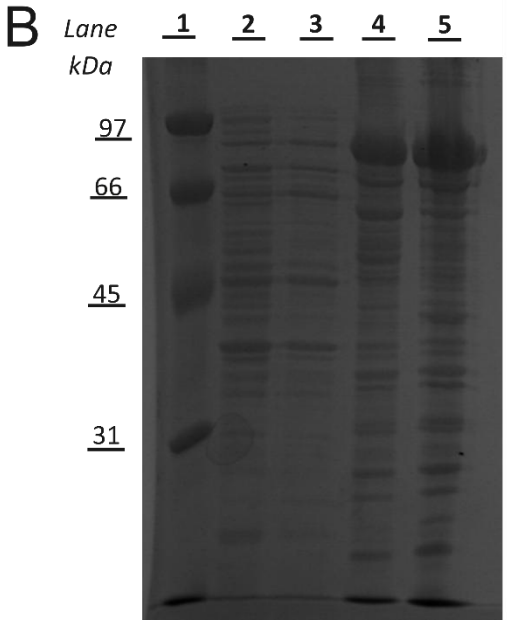
5)BL21 Trxt-SUMO-*Tt*AA10 Insoluble

6)Uninduced SHuffle T7 Trxt-SUMO-*Tt*AA10 Soluble

7) SHuffle T7 Trxt-SUMO-*Tt*AA10 Soluble

8) Uninduced SHuffle T7 Trxt-SUMO-*Tt*AA10 Insoluble

9) SHuffle T7 Trxt-SUMO-*Tt*AA10 Insoluble



1) Low molecular weight marker

2) Uniduced His-MBP-SUMO-*Tt*AA10 Soluble

3) His-MBP-SUMO-*Tt*AA10 Soluble

4) Uniduced His-MBP-SUMO-*Tt*AA10 Insoluble

5) His-MBP-SUMO-*Tt*AA10 Insoluble

Figure 46 Expression testing of the constructs TrxT-SUMO-*Tt*AA10 (BL21 and SHuffle T7) and His-MBP-*Tt*AA10 (BL21) under typical growth conditions , 37°C, induction with IPTG (1 mM) once cultures reached an OD of 0.6, followed by incubation at 16°C overnight. The expression of both constructs appears to even within induction with IPTG.

2.5.3 Increasing the availability of copper

Constructs, *Bd*SUMO-*Tt*AA10, pelB-*Bd*SUMO-*Tt*AA10 and pelB-SUMO-*Tt*AA10 were expressed in LB media containing additional copper. It was thought that an excess of copper ions in the growth medium may stabilise *Tt*AA10, as it could be incorporated into the protein active site. The culture growth was not inhibited by the addition of $CuSo_4$ (7 µM or 100 µM) but there was no improvement in protein solubility or expression. Failure to improve the stability of the protein may be due to the solubility tag preventing the correct formation of the active site, holding the N-terminal residue in a position that prevents the correct coordination of the copper ion.

2.5.4 Inducing *E. coli* Chaperone production with Stress

The actions of native *E. coli* proteins during expression of a target protein can often interfere with protein production. A key example lies in the proteolytic breakdown of a recombinant protein during culture growth. One way to take advantage of using a living organism to produce a target protein is to encourage the production of native chaperones. In *E. coli* this includes DnaK-DnaJ-GrpE, ClpB, GroES-GroEL, and IbpA/IbpB. These proteins will often be produced during conditions of cellular stress, to fight against stress induced alterations to native *E. coli* proteins. Chaperones are known be able to encourage proper folding of aggregation prone recombinant proteins. As such, they are sometimes co-expressed with a target protein to facilitate this.[123-124] By exposing the *E.coli* cells expressing SUMO-*Tt*AA10 to conditions such as high salt molarity, glucose content and heat shock before expression, they may produce helpful chaperones. Once induced, these chaperones will still be within the cell and may be able to assist in the production of soluble and correctly folded target protein. [52]

*Ss*NEDD8-*Tt*AA10 had been observed to be produced as insoluble protein in previous expression tests and so was deemed an appropriate target to undergo expression under different 'stress' conditions, Figure 47. Unfortunately, the gene expressed by *E. coli* was still producing insoluble protein in all cases; however some small bands could be attributed to soluble protein in gel B, lanes 4-6, but it is unclear if this is directly related to the conditions imposed on the cells during growth and expression.

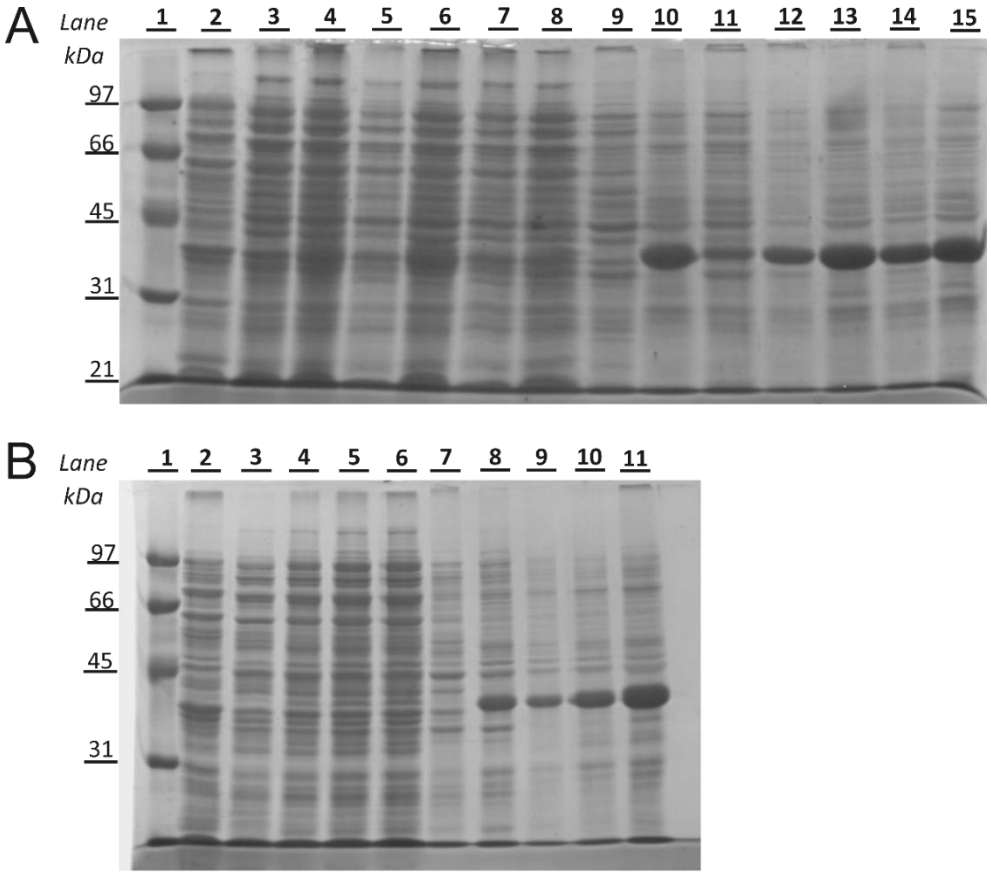


Figure 47 Expression testing of Ss NEDD8-*Tt*AA10 using various conditions shown over two gels. Gel A Soluble fractions: 1) Low weight molecular marker, 2) uninduced *Ss*NEDD8-*Tt*AA10, 3) 1 mM IPTG, 16°C overnight, 4) 0.1 mM IPTG, 16°C overnight, 5) Cold shock before induction, by cooling culture for 15 mins in an ice bath, 1 mM IPTG, 37°C for 1 hour, 6) Cold shock before induction, by cooling culture for 15 mins in an ice bath, 1 mM IPTG, 16°C for overnight, 7) 1 mM IPTG, 37°C overnight, 8) Repeat, 1 mM IPTG, 16°C overnight,
Insoluble Samples of above (2-8) are in lanes 9-15.
Gel B Soluble fractions: 1) Low weight molecular marker, 2) uninduced *Ss*NEDD8-*Tt*AA10,
3) Media changed before induction and replaced with media containing an extra 300 mM NaCl, the 1 mM IPTG, 37°C for 16 hours overnight, 4) 1 mM IPTG, 16°C overnight with containing covered with loose foam bung rather than screw cap, 5) 2 mM IPTG, 16°C overnight, 6) Repeat of number 4, cold shock before induction, by cooling culture for 15 mins in an ice bath, 1 mM IPTG, 16°C for overnight. *Insoluble samples of above (2-6) are in lanes 7-11*

2.5.5 Successful Protein Production

The pelB *Tt*AA10-Strep construct was made in the first instance by Lusia Elias, transformed by *E. coli* Rosetta competent cells and co-expressed with a chaperone. Further production of the target protein was carried out with co-expression of the chaperone Tig (trigger factor, pTf16, using cells from the Prog7 chaperone set from TAKARA), which was induced with L-arabinose. Expression of this construct required extraction of the soluble protein from the periplasm of *E. coli*. This was carried out using the osmotic shock method, whereby the cells were soaked in a solution of sucrose before centrifugation and resuspension in a water-$MgSO_4$ solution. Resuspension in water was sufficient to burst only the outer membrane of the cells, releasing the target protein into the supernatant which was then purified using a streptavidin column, Figure 48. Displacement of the bound protein, with high resolution, was achieved by washing the column with a buffer containing desthiobiotin, which binds strongly to the streptavidin. An approximate yield of 1.6 mg L^{-1} of protein was obtained. The protein samples were produced in several batches and each batch was individually treated with a solution of $CuSO_4$ to ensure that the active site of each protein in the sample was fully saturated. The excess copper was removed from the protein by gel filtration.

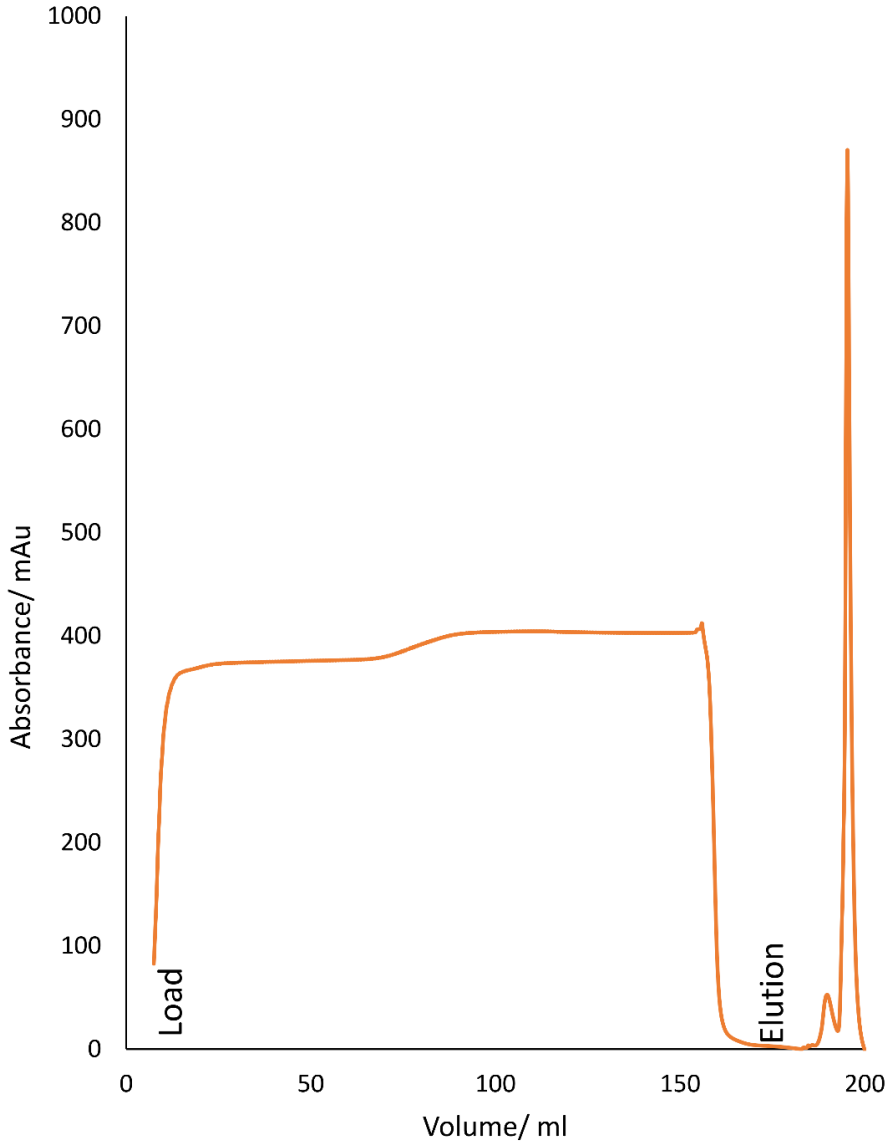


Figure 48 Strep column purification of *Tt*AA10-Strep. The column is loaded with approximately 150 ml of dilute sample, before a short wash of the column with buffer. Elution is carried out with buffer containing desthiobiotin, which displaces the bound protein from the column, causing a sharp elution of the protein.

2.5.5.1 Analysing the purity of *Tt*AA10-Strep

After purification, *Tt*AA10-Strep was analysed by SDS PAGE, Figure 49 and ESI-TOF mass spectrometry to assess its purity, Figure 50. The protein was found to be pure, but was further tested by SEC-MALS to determine if the Strep tag might initiate dimer formation in solution. Figure 51 shows that only a single peak was obtained during the SEC-MALS experiment, which indicates that the protein exists as a monomer in solution. The molecular weight measured by mass spectrometry was also supported, with SEC-MALS suggesting an approximately equivalent weight of 2.442 x10^4 g/mol.

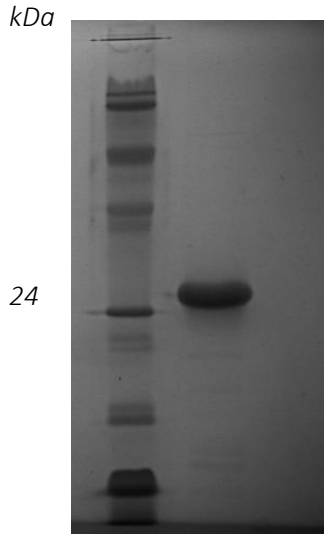


Figure 49 SDS PAGE of *Tt*AA10-Strep during and after purification *Tt*AA10 molecular weight of 24 kDa is indicated.

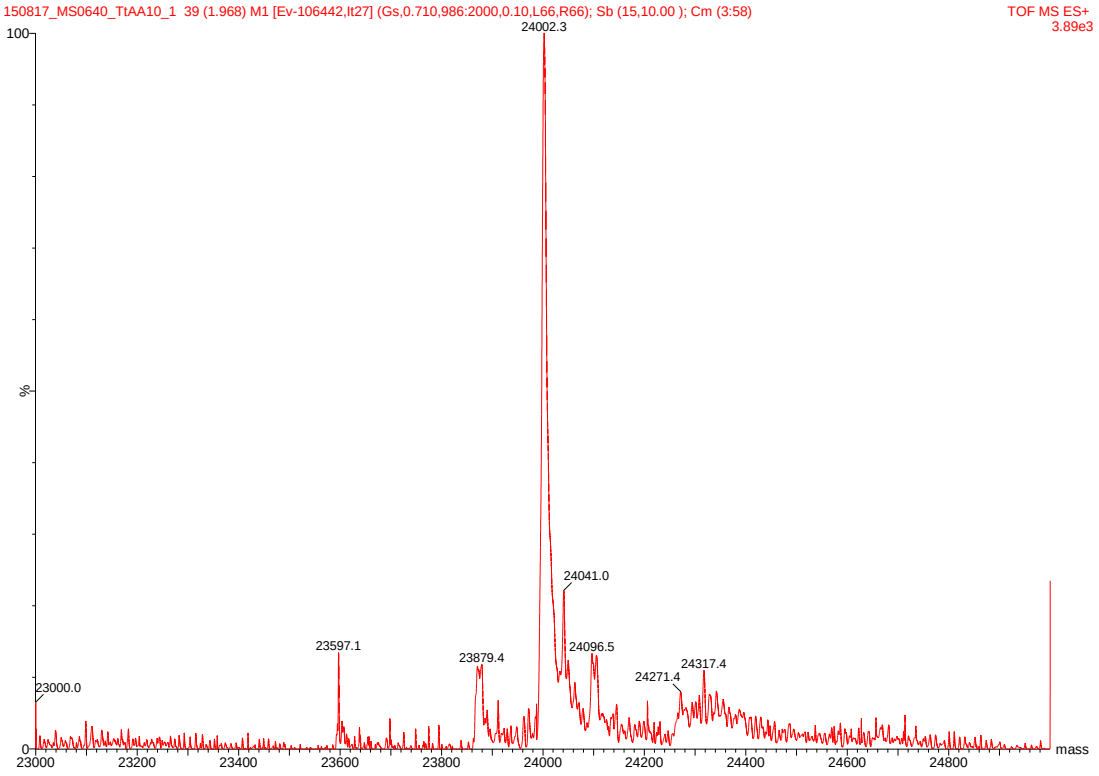


Figure 50 ESI mass spectrometry analysis of *Tt*AA10-Strep, where a peak at 24 kDa was observed.

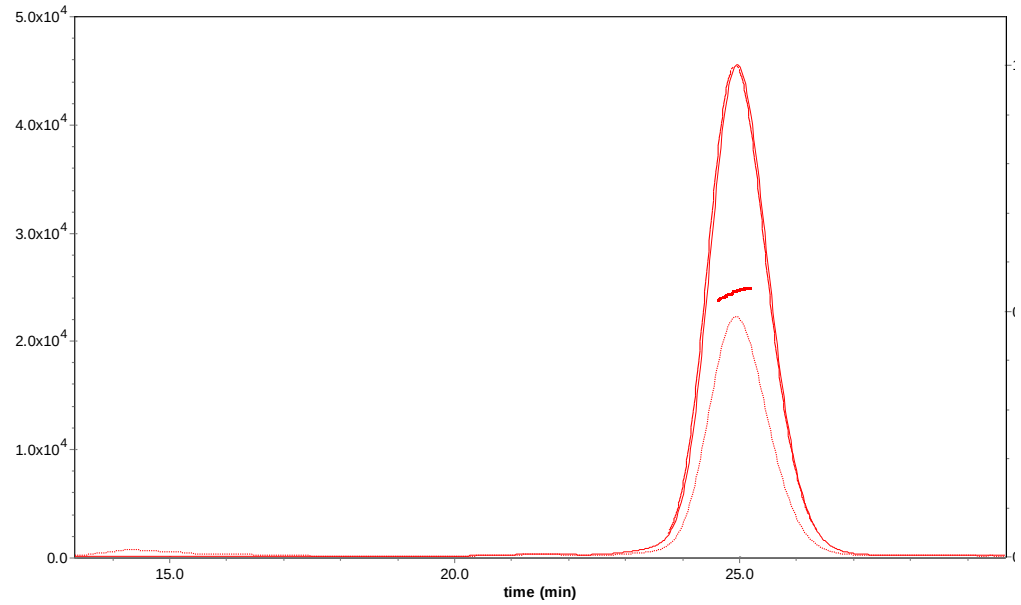


Figure 51 SEC-MALS analysis of TtAA10-Strep, where the solid red line indicates the refractive index, the dashed line is the light scattering, dotted line the UV response at 280 nm and the central line inside the peak representative of the molecular weight (2.442 x104 g mol^{-1}).

2.6 Discussion

2.6.1 GH Targets

Extensive expression testing using different *E. coli* cell strains and temperature conditions lead to the soluble protein expression of 6 GH targets, as summarised in Table 11. Large scale production of these proteins was carried out in *E. coli* using batch cultures. Production of *Tt*GH5_53 was achieved but little characterisation was carried out on this protein, so its purification was intentionally not included in this work. The proteins were purified using a simple method of Ni affinity chromatography, whereby the incorporated hexahistidine tag binds the protein to the column, allowing it to be collected from the cell lysate. The proteins were produced in good yields, and all final protein samples deemed to be of high enough purity to be used in the forthcoming protein characterisation experiments (subsequent Chapters). *Tt*GH5_un was found to have been cleaved at some stage during preparation, possibly due to protease activity. This was found to be the case during two separate preparations of the protein. There were 14 GH targets tested for protein expression and whilst 5 were successfully produced, 9 were not. Several proteins were found to be expressed in an insoluble form. Future work on these GH targets from *T. turnerae* could focus on trying to purify these particular proteins from the inclusion bodies, alter the constructs to introduce a solubility tag or try localisation to the periplasm during expression.

Table 11 Overview of the expression and purification of the 6 T.turnerae GHs produced in this work

Protein	Expression System	Expression Conditions	Purification – Ni Affinity Chromatography	Average yield
***Tt*GH8**	BL21	37 °C, IPTG 1mM, 16 °C overnight	**Binding buffer:** 50 mM HEPES, 250 mM NaCl, 30 mM imidazole, pH 7 **Elution Buffer:** 50 mM HEPES, 250 mM NaCl, 300 mM imidazole, pH 7 **His Tag removal:** 3C protease 1:100, DTT 5 mM, 4 °C overnight. **Final buffer:** 50 mM HEPES, 250 mM NaCl	6 mg/L
***Tt*GH12**	T7 Shuffle			12 mg/L
***Tt*GH5_4**	BL21			6 mg/L
***Tt*GH5_2**	BL21			4 mg/L
***Tt*GH_un**	BL21			19 mg/L

2.6.2 LPMO

The production of *Tt*AA10 proved difficult. The original plasmid containing the gene SUMO-*Tt*AA10 was used to produce small amounts of protein by Dr G. Hemsworth prior to the start of this work. The amount of protein produced from a normal 3 L culture using this construct was very low which was subsequently lost during cell preparation and the different chromatography steps. An attempt was made to produce a set of different constructs with a set of sequence 'building blocks', notably different solubility, affinity tags and the pelB signal peptide. It was hoped that one of the 14 constructs designed may improve the expression yield of *Tt*AA10, but unfortunately the majority of constructs were unsuccessful in this endeavour. A set of plasmids were obtained from the Frey group, which contained a variation of the typical SUMO tag found in commercial plasmids (the SUMO tag used in this work was taken from the Champion SUMO vector). These constructs, *Bd*SUMO-*Tt*AA10, *Bd*NEDD8-*Tt*AA10 and *Ss*NEDD8-*Tt*AA10 were able to produce a good yield of the target protein, but it was unfortunately located in the insoluble fraction of the cell preparations. Versions of the later constructs containing the pelB signal peptide were not found to express. An attempt was made to produce a construct containing multiple solubility tags with the desire that there would be enough 'folding power' from the tags to pull the target protein into a soluble form. Two constructs, one containing MBP and the other thioredoxin were truncated to the N-terminus of the SUMO-TtAA10 sequence. Unfortunately, both constructs suffered from leaky expression, where the protein was found in the insoluble cell lysate fraction in both the IPTG induced and uninduced samples. The efforts to design a construct capable of producing larger amounts of TtAA10 had failed. *Tt*AA10 was finally made in several batches using a construct designed by Luisa Elias, whereby the protein was collected from the periplasm and purified using a non-removable C-terminal Strep tag (a feature which actually proved fruitful during crystallisation and structure solution, as shall be described in Chapter 6). Whilst the yield was still very low, approximately 1.6 mg/L, repetitions of gene expression and protein purification were sufficient to produce enough protein sample for characterisation and analysis.

2.7 Conclusions

In this section of work, the design, expression trialling and purification of 6 target proteins from *T. turnerae* has been described. 6 out of 14 recombinant GH targets were successfully synthesised in soluble forms by *E. coli*, with successful large scale purification carried out using a simple hexahistidine affinity tag of 5 proteins, *Tt*GH8, *Tt*GH12, *Tt*GH5_2, *Tt*GH5_4 and *Tt*GH5_un shown. Several other GH targets were produced as insoluble aggregates. The single LPMO found in the genome of *T. turnerae, Tt*AA10 was also successfully made, but attempts to improve the yield of the protein did not prove fruitful. In conclusion, this work was able to produce nearly half of the protein targets selected for characterisation in forms suitable for further study. In the forthcoming Chapters, characterisation of the 6 successfully produced proteins will be described in detail and divided based on CAZy family; experiments pertaining to substrate specificity, enzyme function, kinetic analysis and in several cases, structural insight will be presented and discussed.

3

Characterisation of three *Tt*GH5s

3.1 Abstract

T. turnerae possesses a wide variety of glycoside hydrolases from many different families, possibly providing it with a large protein 'reservoir' which can be used to tackle the various substrates that the host Shipworm ingests. Three proteins from *T. turnerae*, classified as GH family 5 proteins under the CAZy database, *Tt*GH5_2 (ACR12145.1), *Tt*GH5_4 (ACR12247.1) and *Tt*GH5_un (ACR11279.1) had been successfully purified and have now been characterised by activity analysis using thermal shift assays, thin layer chromatography and mass spectrometry. *Tt*GH5_2 and *Tt*GH5_un were both found to be similar in their substrate specificity, and have been shown by this work to be β-glucanases. Both proteins were active on konjac glucomannan (kGM) and cellulose based substrates, but not ivory nut mannan (inM), indicating preferential cleavage at glucose residues. *Tt*GH5_un is as yet unassigned to a GH5 subfamily due to the absence of related sequences from the available genomic data. This study provides a base line characterisation for any subsequent similar sequences discovered in the future and indicates the likelihood that this protein will form a new GH5 subfamily. *Tt*GH5_4 was found to be specific for xyloglucan, adding to the small number of xyloglucanases observed within this subfamily and preliminary kinetics were carried out. Crystallisation was attempted for all three GH5s, with *Tt*GH5_un producing crystals that diffracted to 1.8 Å, but due to the lack of available reference models of sufficient sequence homology, the structure was not solved. This study has identified the activities of three new GH5 enzymes from *T. turnerae*, two of which possess preferential activity for a glucose-based polymer and a third which is able to accommodate the complex branching structure of xyloglucan. It is suggested that *Tt*GH5_un will form a new GH5 subfamily in the future and

its established activity will assist characterisation of similar novel enzymes when they become available through the ever-increasing influx of genomic data.

3.2 Introduction

3.2.1 The GH5 Family

One of the largest GH families is that of GH5, containing 11640 individual entries in the CAZy database at the time of writing.[34] Historically known as 'cellulase family A' it was the first cellulase family to be described. The GH5 family belongs to clan GH-A, a group of 19 GH families linked together by a distant evolutionary ancestor, where all members of the clan share certain features such as protein fold and catalytic machinery.[31, 36, 125] The GH5 family contains a variety of different functionalities, with a significant portion active on breaking down polysaccharides with a glucose backbone. Whilst the general structure and catalytic ability of GH5s remains similar across the family, the low sequence identity between different clades within the family has created a vast pool of enzymes capable of breaking down a variety of different substrates. For example, cellulases typically exhibit a tight binding groove in which the polysaccharide chain fits snugly, whereas some xyloglucanases have evolved larger binding clefts in order to accommodate the branched nature of the substrate.[126]

3.2.2 Subfamilies of GH5s

During the 1990s, 'cellulose family A' was split into subfamilies as new emerging experimental data was produced. The subfamilies, known as A1-A10 covered a range of enzymatic specificity and origin. As is the same for all GH families, but more importantly for older families, the ever-increasing amount of sequencing data available brings into question whether the older classifications meet the needs of today's data. Members of the GH5 family cover a very wide diversity of activities and are found in a multitude of ecological niches. Current estimates suggest over 20 experimentally determined enzymatic activities form a broad degree of functionality within the family, yet this is most likely set to increase due to recent work by Aspeborg *et al* in 2012, who undertook a large phylogenetic based reclassification of the family. The work by Aspeborg and colleagues was induced, in some respect, by a rearrangement of several GH5 sequences into family GH30. This came about due to new structural insights that linked certain conserved structural features of the catalytic machinery not to the GH5 family, but to GH30. Aspeborg states that the steady, but

enormous increase in genomic data has led to a large proportion of GH5 sequences sitting in the family, unassigned to subfamilies, and with often unknown functions. Aspeborg's new classification system of GH5 throws away the old A1-A10 subfamilies and presents 51 new subfamilies, Table 12. In 2012, their bioinformatics analysis split the publicly available set of 2300 GH5 catalytic modules based on the production of a phylogenetic tree, of which the distinct branches (clades) represent an individual subfamily. Some of the historical subfamily classifications held true, whilst some were combined. As such, GH5_3 and GH5_6 no longer exist. The naming system of each subfamily is simply GH5 followed by an underscore and a number. As would be expected, those subfamilies with more historical entries contain a much larger selection of sequences. GH5_2 for example, historically the A2 subfamily, now consists of 245 sequences. Similarly, GH5_4 gained a combined data set coming from historical subfamilies A3 and A4, giving a total of 160 entries. A baseline definition of five publicly available sequences was set as the requirement to create a new subfamily; ten subfamilies contain only five sequences. The classification success rate is deemed to be 80%, meaning 20% of the publicly available sequences did not meet these criteria. This hints at the significant number of sequences yet to be assigned to a subfamily due to lack of genomic data, suggesting that there will be new subfamilies uncovered as time goes on. The spread of diversity over the new subfamily classification also highlights taxonomical data; several subfamilies contain sequences that are limited to belonging to a specific taxonomical origin, for example to fungal, or a certain species. Aspeborg also draws attention to the prevalence of non-catalytic subfamilies, for example GH5_30 in which all sequences are deemed unable to carry out canonical GH activity, due to a loss of essential elements that traditionally yield catalytic ability. Other losses of catalytic prowess are seen in a couple of other families, where certain members display a lack of catalytic activity and are appended to other modules.[125] It can be suggested then that 20% of publicly available sequences will in the future be characterised completely, whether into existing GH5 subfamilies or new ones with catalytic or novel non-catalytic functions.

Table 12 Summary of the GH5 subfamily classification as proposed by Aspeborg *et al.* Subfamilies where evidence exists for the characterisation of members, shown as taxonomical origin and general activities with their associated EC numbers. Several sub families have no activities determined.

Subfamily	Origin	Activity
GH5_1	Archaea, bacterial, uncultured symbionts	Endo-β-1,4-glucanase activity (EC 3.2.1.4)
GH5_2	Extracellular	Endo-β-1,4-glucanase Activity (ec 3.2.1.4)
GH5_4	Majority extracellular bacterial	Endo-β-1,4-glucanases(EC 3.2.1.4) , Xyloglucan-specific Endo-β-1,4-glucanase (EC 3.2.1.151), licheninases (EC
GH5_5		Endo-β- 1,4- glucananase (EC 3.2.1.4)
GH5_7	Eukaryotic, archaeal, bacterial	B-1,4-mannan-cleaving enzymes (EC 3.2.1.78 and EC 3.2.1.25)
GH5_8		Endo-β-1,4-mannanase (EC 3.2.1.78)
GH5_9	Fungal, cell wall modifying	Exo-β-1,3-glucanase (EC 3.2.1.58), endo-β-1,6-glucanase (EC 3.2.1.75), β-glucosidases (EC 3.2.1.21)
GH5_10		Endo-β-1,4-mannanase (EC 3.2.1.78)
GH5_11		*Not experimentally determined*
GH5_12	Fungal	B-glucosylceramidases (EC 3.2.1.45)and (flavonoid) β-glucosidase (EC 3.2.1.21)
GH5_13	-	*Not experimentally determined*
GH5_14	Plant	Exo-β-1,3-glucosidase (EC 3.2.1.58)
GH5_15	Fungal	*Not experimentally determined*
GH5_16	Fungal	Endo-β-1,6-galactanase (EC 3.2.1.164)
GH5_17		Endo-β-1,4-mannanase (EC 3.2.1.78)
GH5_18	-	*Not experimentally determined*
GH5_19	-	*Not experimentally determined*
GH5_20	-	*Not experimentally determined*
GH5_21	Bacteroidetes	Endo-β-1,4-xylanase (EC 3.2.1.8)
GH5_22		Endo-β-1,4-glucanase activity(EC 3.2.1.4)
GH5_23	Fungal, secreted	Fungal β-diglycosidases (EC 3.2.1.149 and EC 3.2.1.168)
GH5_24	-	*Not experimentally determined*
GH5_25	Thermophilic	Endo-β-1,4-glycanases (EC 3.2.1.4 and EC 3.2.1.78)

GH5_26	Uncultured microorganisms	Endo-β-1,4-glycanases (EC 3.2.1.4 and EC 3.2.1.73)
GH5_27	Eukaryotic, extracellular	Endo-glycosylceramidases (EC 3.2.1.123)
GH5_28	*Actinobacteria*, bar one	Endo-glycosylceramidases (ec 3.2.1.123)
GH5_29	*Rhodococcus* bacterial	Endo-glycosylceramidases
GH5_30	-	*Not experimentally determined*
GH5_31	Fungal, secreted	B-1,3-(gluco)mannanase activity (EC 3.2.1.-)
GH5_32	-	*Not experimentally determined*
GH5_33	-	*Not experimentally determined*
GH5_34	Bacterial, extracellular	Arabinoxylanase activity (EC 3.2.1.-)
GH5_35	-	*Not experimentally determined*
GH5_36	Bacterial	Endo-b-1,4-glycanases (EC 3.2.1.73 and EC 3.2.1.78), endo-
GH5_37	Bacterial	Endo-β-1,3/4-glycanases (EC 3.2.1.4 and EC 3.2.1.73),
GH5_38	Bacterial	Activity on model plant cell wall (PCW) compounds
GH5_39		Endo-β-1,4-glucanase activity (EC 3.2.1.4)
GH5_40	-	*Not experimentally determined*
GH5_41	-	*Not experimentally determined*
GH5_42	-	*Not experimentally determined*
GH5_43	-	*Not experimentally determined*
GH5_44	-	*Not experimentally determined*
GH5_45	-	*Not experimentally determined*
GH5_46	Bacterial	Activity on model plant cell wall (PCW) compounds
GH5_47	-	*Not experimentally determined*
GH5_48	Bacterial	Activity on chitin and chitosan derivatives
GH5_49	-	*Not experimentally determined*
GH5_50	-	*Not experimentally determined*
GH5_51	-	*Not experimentally determined*
GH5_52	Bacterial	Cellodextrinases (EC 3.2.1.74)
GH5_53	Extracellular	Cellodextrinase (EC 3.2.1.74)

3.2.3 Structural Overview of GH5s

Despite the separation into subfamilies and differences in activity, members of the GH5 family all share some common features, including protein fold, catalytic machinery and reaction coordinate itinerary. GH5 sequences vary little in their overall tertiary structure, however, there are only a handful of residues that are invariant across the entire family.[127] Structural data are available for 75 separate "apo" proteins from subfamilies, 1, 2, 4, 5, 7, 8, 9, 10, 26, 27, 28, 34, 36 and 37. Some of studies have also expanded further and produced structures of protein-substrate complexes to map out the binding site and assess catalytic activity. The overall structure of all GH5 proteins is conserved and can be described as having a $(\beta/\alpha)_8$ fold, as first identified in the structural characterisation of *endoglucanase CelC* in the late 1990s, **Figure** 52.[128-129] This fold is also known as a TIM (named after Triose-phosphate isomerise) barrel structure and consists of an alternating backbone configuration of eight parallel β-sheets and eight α-helices. This common 'doughnut' shape fold is created by a circular repeating pattern of 'β-loop-α-loop' secondary structures. The inner surface is primarily made up of the parallel β-sheets, inside which there is a hydrophobic core. The α-helices make up the protein exterior, meaning there is a secondary hydrophobic 'interface' positioned between the β-sheets and α-helices which run like a channel between the C- and N-terminus.[130-131] The amino acid side chains on the β-sheets are described as forming two layers; sides chains on strands 1,3,5 and 7 point inwards towards the central hydrophobic core, whilst strands 2,4,6 and 8 side chains point outwards towards the α-helices. As such, the barrel displays a four-fold symmetry instead of eight-fold.[130] Silverman *et al*, carried out a study focused on understanding the structural importance of components creating the $(\beta/\alpha)_8$ protein fold in the namesake protein, triose-phosphate isomerise. They mutated residues in different regions of the fold to assess the effect on structure and protein functionality. It was suggested that the hydrophobic interface between the β-sheets and α-helices was relatively flexible compared with the internal hydrophobic core, where mutations in the internal core resulted in a severe loss of function.[131] Being the most common protein fold suggests it has existed for a relatively long time on the evolutionary timescale. The overall barrel structure has recently been proposed to have evolved from fusion events of 'half barrels' ($(\beta/\alpha)_4$), indicating the possibility of hidden or lost sub-domains that could assist in sequence analysis. [130, 132-133]

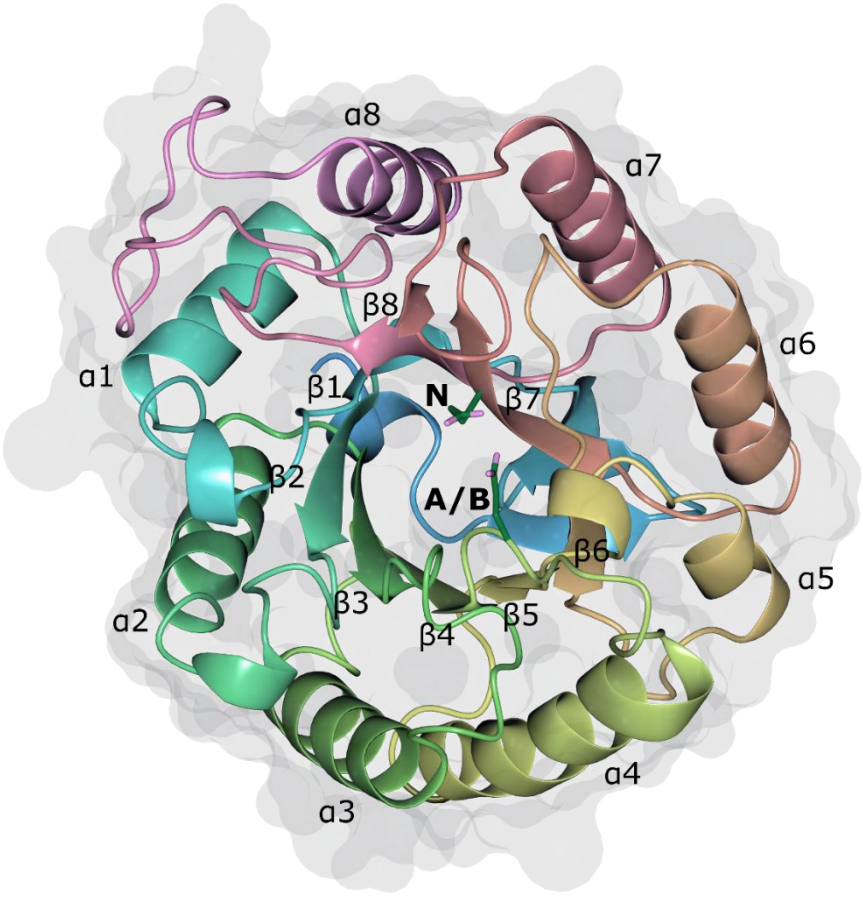


Figure 52 Structural overview of a GH5 enzyme , Cel5A, showing the classic $(\beta/\alpha)_8$ protein fold. The helices and sheets are labelled and pack in a doughnut shape. The inner hydrophobic core and interface region between the sheets and helices can be easily seen. The nucleophilic glutamate is labelled N and is positioned on strand β7, whilst the glutamate that acts as the catalytic acid/base, labelled A/B is on the C-terminal loop of strand β4. PDB code: Q102

3.2.4 Catalytic Activity in GH5s

The catalytic breakage of glycosidic bonds is carried out through a double displacement mechanism, in which the configuration around the anomeric carbon is 'retained' (see **Chapter** 1).[31] Two carboxylic acid side chains, which have been experimentally determined as glutamates, act as the catalytic residues and sit on one of the loops between a β-sheet and α-helix. In a sustained effort, Davies *et al* were able to analyse the catalytic reaction itinerary of an endoglucanase known as Cel5A from *Bacillus agaradhaerens*. This characterisation created a representative catalytic reaction mechanism of glycosyl catalytic activity in GH5. Crystallography of Cel5A with chemical mimics capable of producing complexes where the substrate is effectively trapped, **Figure** 53 (known as Wither's reagents, covalent inhibitors of glycoside hydrolases that use sugar units containing fluorine[134]), allowed the authors to see snapshots of the reaction itinerary within the electron density. Such extensive crystallographic work on Cel5A enabled the reaction conformational itinerary to be mapped.

The cleavage reaction was found to proceed through initial substrate binding in a distorted $^{1}S_{3}$ conformation, followed by formation of a $^{4}C_{1}$ glycosyl enzyme intermediate, where the configuration around the anomeric carbon is inverted. The transition state between the two conformations was suggested to continue through a $^{4}H_{3}$ half chair conformation. Finally, the binding of the product complex was found to be unfavourable, leading to dissociation and opening up the binding site for further and higher affinity substrate binding.[135]

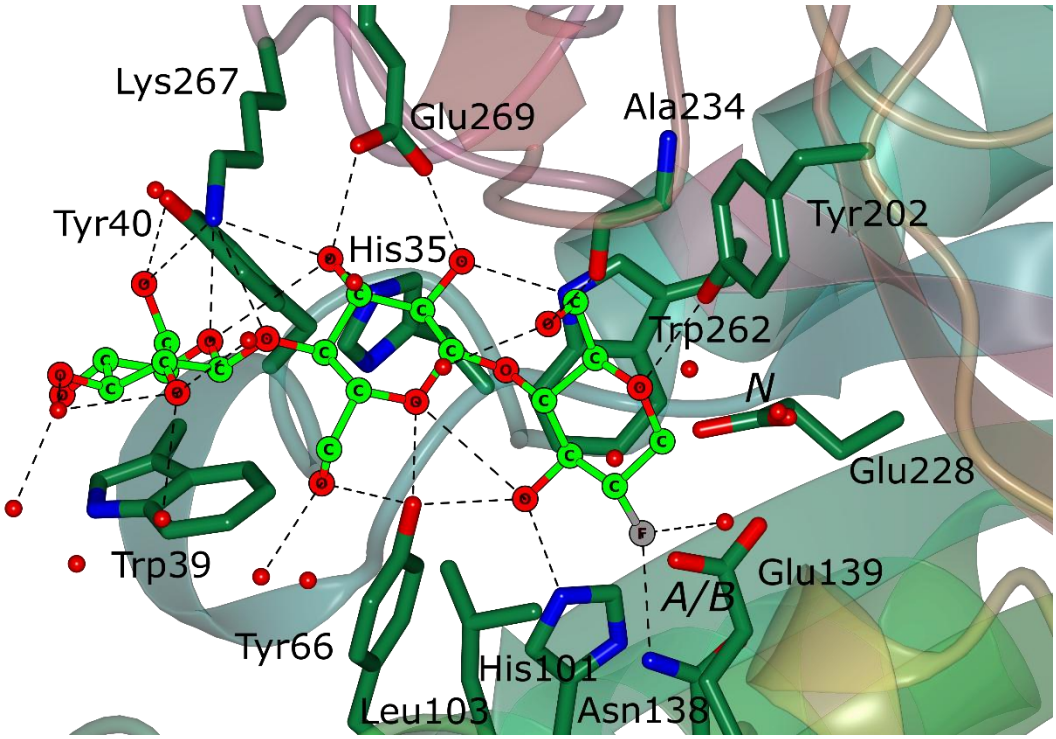


Figure 53 Close up view of Cel5A bound with a Wither's trapping reagent ; β-D-Glcp-(1-4)-β-D-Glcp-(1-4)-α-D-Glcp2F-(1-4)-GLU. The catalytic nucleophile (N) and acid/base (A/B) are labelled, and correspond to Glu228 and Glu139 respectively. The sugar is bound in the -1 to -3 subsites, with hydrogen bonding shown to the nearest neighbours (within 4 Å). Several residues remain conserved throughout the GH5 family, for example, Trp39 and Trp262 that provide ring stacking interactions with the bound substrate. PDB: 1QI0.

3.2.5 Conservation in GH5s

The positions of the catalytic residues are conserved throughout the family and general protein fold bar a few inactive subfamilies, as previously mentioned. The nucleophilic glutamate is positioned three quarters of the way along β-strand 7, whilst the glutamate acting as the catalytic acid/base is found on the C-terminal loop end of β-strand 4.[127, 129, 136] This suggests that the loop regions create the active site environment, possibly endowing substrate specificity due to the large degree of promiscuity.[130] Sequence analysis can determine the identity of those residues that are conserved throughout the entire GH5 family, as well as those specific for certain subfamilies. Several studies have gone on to

access the function of these invariant residues to ascertain the role they play within; structural, catalytic, or substrate specificity. Indeed, including the catalytic residues (written in italics), only a small number of amino acids are conserved across family GH5; *Glu303*, His249, *Glu172*, Tyr251, Trp339, Arg67, His111 and Asn171 (numbering from a xyloglucanase, assigned to GH5_4, AC2aCel5A).[126] For example, substrate specificity is provided by an invariant cis-peptide, of trp339 (Cel5A). It acts as a base for the -1 subsite, allowing the formation of a hydrogen bond between the residue and the O-2 hydroxyl group of the moiety bound in the adjacent -2 subsite.[127]

3.3 *T. turnerae* GH5s

The bacterium *Teredinibacter turnerae* contains 107 glycoside hydrolases as quoted in the CAZy database, of which 14 have been classified as belonging to family GH5. As discussed in Chapter 2, a selection of *T. turnerae* GH targets from a range of different families were chosen for purification trials using an *E. coli* expression strategy. The 14 targets had the potential to cover a large range of activities and substrate specificities due to the diverse GH family classifications. The work described in this Chapter builds on the successful purification of 3 recombinant *T. turnerae* GH5 proteins, as described in Chapter 2. Two of the successfully produced GH5 proteins are classified as belonging to subfamilies 2 and 4, which shall be described in more detail in the forthcoming section. The third GH5 protein produced in this work has not been classified into a subfamily as of yet and the implications of sequences lacking significant homology to other members within the GH5 family shall be discussed further below.

3.3.1 *Tt*GH5_2

As mentioned above, *Tt*GH5_2 is the largest subfamily, and at the time of writing (August 2018), CAZy quoted 119 sequences as characterised, with 14 structures available, out of a potential pool of 986.[34] The major activity for this subfamily is β-1,4-glucanase, where the proteins are often multimodular. Some contain CBM binding domains and others are linked to dockerin-like domains. The sequences mostly come from bacteria, but representatives from archaea and uncultured protists are also included.[137] Even within a subfamily there exists a disparity in the sequence, yet there remain several conserved motifs that endow structural and functional similarity. The most structurally characterised member of GH5_2 is that of Cel5A from *Salipaludibacillus agaradhaerens*, where 22 separate structures exist; 2 of

which are apo and the remaining 20, substrate complexes. Such high level of investigation allowed the substrate binding site to be mapped. For example, the study of a cellobiose complex showed the interaction of tyrosine, common to the subfamily and located on a turn between strand 5 and helix 5, acting as a substrate binding plane at the -1 subsite; the aromatic side chain and the pyranose ring of the sugar moiety of the substrate are able to interact through ring stacking.[127]

3.3.2 *Tt*GH5_4

At the time of writing, the GH5_4 subfamily contained 578 members, of which, 98 have been characterised by biochemical methods and structures of 17 deposited in the PDB.[34] In GH5_4, 79 entries out of the 98 characterisations specifically quote endo β-1,4-glucanase (EC 3.2.1.4) as the enzyme activity. A further selection of entries quote β-glucanase activity without definition of what form this activity takes, i.e. endo or exo, whilst several entries quote multiple activities, including that of β-1,4-glucanase. Of the 98 characterised entries, only 6 bacterial sequences exhibit known xyloglucanase function, with a further 4 of these structurally characterised.[34] GH5_4 is the only subfamily containing enzymes active on xyloglucan and these activities appear to have developed at different times; the xyloglucanases form part of two different clades within the overall GH5_4 subfamily phylogenetic tree.[137] Xyloglucanase is composed of a backbone of glucose residues, at which branching occurs at the C6 hydroxyl position with xylose residues. Further branching of the xylose residues with other moieties, galactose, fucose and arabinose, create a spatially large and complex polysaccharide. Due to the complexity, the substrate is often described in 'modules' and labelled as follows; G = Glc*p*, X = [Xyl*p*(_1,6)]Glc*p*, and L = [Gal*p*(_1,2)Xyl*p*(_1,6)]Glc*p*.[138] **Figure** 54 describes the various different sides chains found in the xyloglucan of different species of plants, using common glycan symbols and showing the typical linkages.[138-139] Alterations in the active site cleft enable xyloglucanases to accommodate a more complex substrate, where a broadening of the active site has been observed crystallographically compared with GH5 enzymes active on simpler substrates, such as cellulose. For example, in **Figure** 55, a common cleavage product of xyloglucan degradation is shown in the binding site of a xyloglucanase from *Paenibacillus pabuli*, written as GXLG.

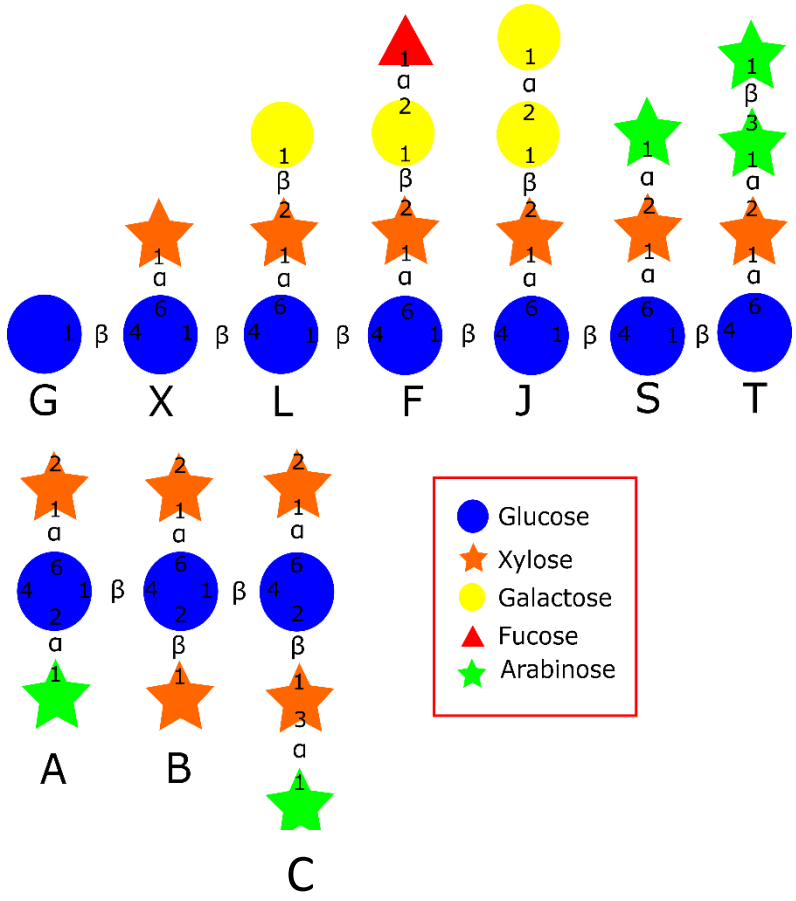


Figure 54 Possible branching chains found, where arrangement of the various 'modules' can fully describe a specific polysaccharide pattern.

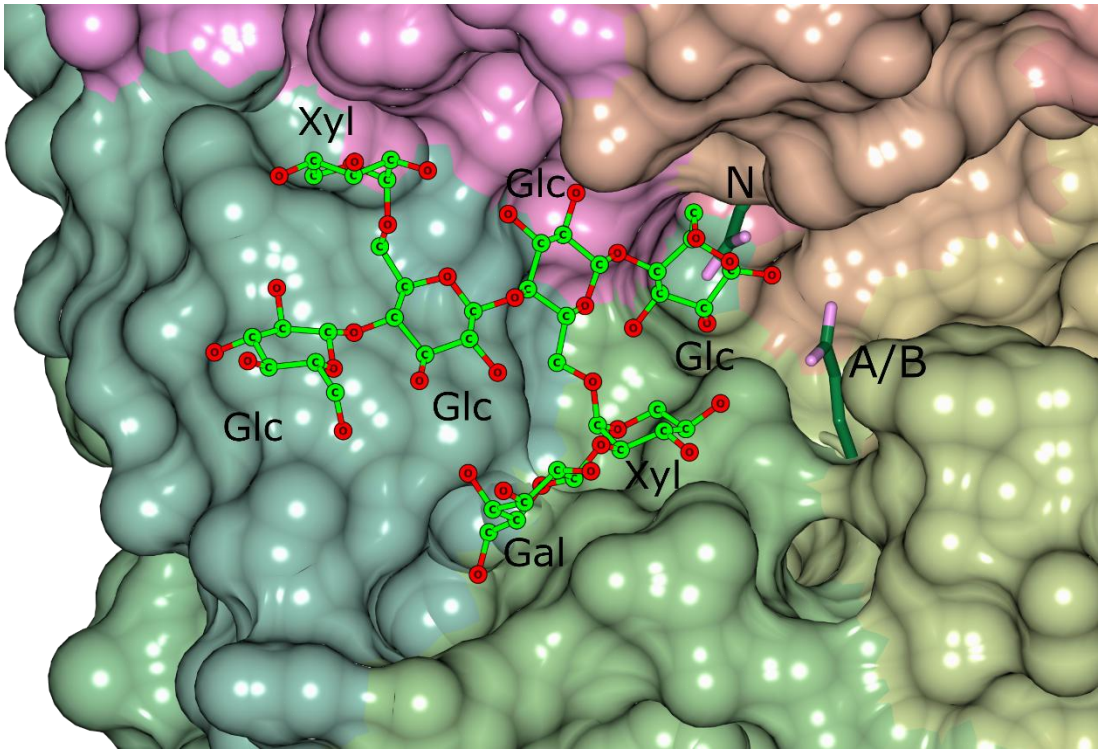


Figure 55 Close up structural view of a GH5_4 enzyme active on xyloglucan , where a xyloglucan oligosaccharide (GXLG) is shown bound in the active site. The glutamate nucleophile and catalytic acid base are labelled as N and A/B, respectively.

3.3.3 *Tt*GH5_un

As mentioned previously, a large proportion of glycoside hydrolases in GH5 have not been assigned to a subfamily. This is due to the cut-off point imposed by Aspeborg and colleagues that defined their most recent GH5 classification system; 5 similar sequences are needed to define a new subfamily.[137] As such, GH5_un is representative of any sequence that has been classified overall as a GH5, but it does not have enough similarity to other sequences to be placed into a subfamily. However, the ever expanding CAZy database, which was earlier proposed as an issue to GH overall classification, will most likely result in providing more genomic data and as such, more matching sequences to those 'unknown' sequences already in existence. As such, it can be suggested that once these unassigned sequences are characterised, they will inform new activities/subfamilies in the future. A major issue regarding the characterisation of GH5_un sequences is that their sequence similarity with other characterised sub-families enzymes is very low. Sequence analysis determines a protein's position within the CAZy classification system, but a sequence defined as a GH5 still has a large degree of possible activities.[25, 27, 129] Sequence alignments with online tools, such as NCBI BLAST, can often assist in forming a proposal of a protein's activity based on the activities of the highest sequence match.[112-113] When a sequence is known to be relatively

unique and has little similarity with other proteins, the GH5_un protein in question can be difficult to hypothesise a function for. This can cause issues with designing experiments to test the substrate specificities of the enzymes, as unless one takes a broader view on potential activities; it will be a stab in the dark.

3.4 This Work

The aims of this section of work were to analyse *Tt*GH5_2, *Tt*GH5_4 and *Tt*GH5_un to assess their substrate specificity, activity and structure. A broad spectrum of characterisation methods, has been applied to successfully define the activity and function of all 3 enzymes. Kinetic analysis of all 3 enzymes was also carried out to assess the efficiency of each enzyme for their proposed substrate.

3.5 Methods

3.5.1 Materials

Substrates used in this study were tamarind xyloglucan (tXyG), mixed linkage Barley beta glucan (bMLG), Avicel (Sigma), Phosphoric acid swollen cellulose (PASC) (Sigma), ivory nut mannan (inM), carob galactomannan (cGM), konjac glucomannan (kGM), mannohexaose (Sigma), birchwood xylan (bX), wheat arabinoxylan (wAX), rye arabinoxylan (rAX), corn arabinoxylan (cAX), potato starch, rice starch, squid pen chitin, crab chitin (Sigma), guar gum, locust bean gum, apple pectin (Silverspoon). Orcinol monohydrate used for staining thin layer chromatography plates was purchased from Sigma. Substrates for kinetics and dyed insoluble polysaccharides were purchased from Megazyme (Ireland); 4-Methylumbelliferyl-β-cellobioside (4-MU-C2), 4-Methylumbelliferyl-β-cellotrioside, 4-Methylumbelliferyl-β-laminaribioside (4-MU-B1/3-C2), AZCL-HE galactomannan (Carob) and AZCL-HE cellulose (fine). Pure recombinant *Tt*GH5_2 (ACR12145.1), *Tt*GH5_4 (ACR12247.1) and *Tt*GH5_un (ACR11279.1) were produced as described in Chapter 2.

3.5.2 Thermal shift analysis

Samples containing SYPRO-orange dye (15 µL) and enzyme (final concentration of 1 mg mL^{-1}) with either buffer or substrate (30 µL total) were run using a Stratgene qPCR. Small amounts of solid polysaccharides were added directly to the sample tubes, to a point where a quarter of the 30 µL volume of the sample was taken up by the solid. Soluble substrates were used at concentrations of either 1 mg mL^{-1} or 10 mM depending on substrate type. Samples were heated from 20 °C to 91 °C in increments of +1 °C over 71 cycles. The fluorescence of SYPRO orange was monitored throughout and the data used to calculate the protein melting temperature (see Appendix 2 for further information about this method) Curves were fitted using a free online tool developed by Paul Bond at the University of York and is available at; http://paulsbond.co.uk/jtsa.

3.5.3 Activity Assays

Reactions were carried out in 1.5 mL Eppendorf tubes, held horizontally during shaking to ensure thorough mixing of insoluble substrates during incubation. Enzyme concentrations were used at 1 mg mL^{-1} and substrates at 1 mg mL^{-1} unless otherwise stated. Assays were carried out either at 37 °C (*Tt*GH5_2, *Tt*GH5_un) or room temperature (22 °C) (*Tt*GH5_4)

and left overnight before further analysis. Samples were centrifuged to pellet the remaining solids and to ensure only the soluble products were taken forward for analysis. A second quick method for activity analysis involved the use of dyed insoluble polysaccharides – AZCL-HE galactomannan (Carob) and AZCL-HE cellulose (fine) (Megazyme), dissolved to 1 mg mL^{-1} in buffer. Addition of an enzyme capable of breaking down the substrates releases the dye, measurable at 590 nm.

3.5.4 Thin Layer Chromatography (TLC)

Soluble products (4 µL total) were spotted on a TLC plate 1 µL at a time and left to dry between applications. Appropriate oligosaccharide standards were run alongside reaction samples to provide a weight ladder for comparison. The TLC plate was placed in a pre-equilibrated chromatography tank containing the running buffer (50% v/v n-butanol, 25% v/v acetic acid, 25% v/v water). Plates were run once, dried and then re-run to improve the separation of sugars. The plates were visualised using a staining solution (3% v/v sulphuric acid, 75% v/v ethanol, 0.1% w/v orcinol monohydrate), dried and then heated to approximately 100 °C using a hot plate. TLC plates were photographed immediately after visualisation as the stain fades quickly.

3.5.5 High performance anion exchange chromatography with pulsed amperometric detection (HPAEC-PAD)

HPAEC-PAD was carried out at Newcastle University under guidance of the Gilbert group, using a Dionex system. *Tt*GH5_4 overnight hydrolysis reactions were run at 37 °C on tXyG. The soluble samples were mixed with a fucose internal standard and run on an anion exchange column (CARBOPAC) using a sodium acetate gradient. The peak areas observed in the HPAEC-PAD traces were normalised against a fucose internal standard.

3.5.6 Matrix assisted laser desorption ionisation – time of flight- mass spectrometry (MALDI-TOF)

Soluble fractions activity reactions were analysed using MALDI-TOF (ultraflex III) mass spectrometry. Samples were mixed with the 2,5-dihydroxybenzoic acid matrix (20 mg mL^{-1} concentration, in 50% acetonitrile, 0.1% trifluoroacetic acid) in either a 1:1 or 1:5 sample to matrix ratio. Sample-matrix mixtures were spotted (1 µL) onto the sample plate and left to dry. A pre-made calibration solution (containing a set number of protein standards) was

mixed with the matrix and spotted onto the plate, making sure that all samples were within 1 spot of the calibration spot. The instrument was controlled using the Bruker FlexControl software. Using the spot imaging, the laser (100 Hz laser frequency, laser power of 40 mW) was manually moved around the spot to prevent damage to the sample. Depending on the sample concentration, data collection (standard run set at 800 shots) was repeated (sometimes at different laser intensities) and spectra added together in a cumulative manner to improve signal definition and to observe the growth of product peaks relative to noise. Spectra were processed and analysed using the related Bruker flexAnalysis software.

3.5.7 Polysaccharide Analysis Using Carbohydrate Gel Electrophoresis (PACE)

*Tt*GH5_2 and *Tt*GH5_un were analysed by PACE[140], carried out by Dr. Theodora Tryfona, University of Cambridge. 100µl reaction volume using reaction buffer (20 mM HEPEs, 200 mM NaCl, pH 7) were mixed with 3 µg enzyme (*Tt*GH5_2 or *Tt*GH5_un) and 250µg substrate (bMLG, inM, kGM, GM substrates were boiled for 10min to solubilise). For deacetylation the samples were treated with 20µl 4M NaOH for 1h at room temperature followed by desalting on PD-10 column. Reactions were incubated at 37 °C overnight. The undigested oligosaccharides were precipitated out with 80% ethanol followed by collection of the supernatant. Sequential digestion analysis of the products of *Tt*GH5_un were carried out using 1µl of β-glucosidase and/or β-mannosidase and samples incubated for 4h at room temperature. Enzymes were thermally deactivated (100 °C for 30min) before the next enzyme was added. Finally, samples were labelled with ANTS (8-aminonaphthalene-1,3,6-trisulfonic acid) and run on the gel.

3.5.8 Kinetics

The activity of *Tt*GH5_2 and *Tt*GH5_un was tested on 4-Methylumbelliferyl-β-cellobioside (4-MU-C2), 4-Methylumbelliferyl-β-cellotrioside and 4-Methylumbelliferyl-β-laminaribioside (4-MU-B1/3-C2). Enzyme scoping experiments were used to find a suitable enzyme concentration for the assay (0.06-1 µM) using substrate at 1 mM. A standard curve of fluorescent product 4-MU over a concentration range, 15.6 - 1000 µM was carried out and used for analysis. Substrate was distributed in 48 well black plates and enzyme pipetted into each condition simultaneously and the assay started in a Clariostar plate reader. The assay was run using 0.5 µM enzyme and a substrate gradient 1-0.015 mM and data recorded every

5 seconds for 400 s. The experimental gain had to be decreased for reactions involving 4-MU-C3 as fluorescence was high. The gradient from the separate plots of each different substrate concentration were plotted against their corresponding assay substrate concentration. The gradient of this graph, if linear as in the case of 4-MU-C2, was used to calculate k_{cat}/K_m by dividing by the enzyme concentration. The 4-MU-C3 substrate produced curved profiles with both *Tt*GH5_2 and *Tt*GH5_un, which was analysed using the Michaelis-Menton kinetics option available in Origin, to calculate values for K_m, k_{cat} and k_{cat}/K_m.

The activity of *Tt*GH5_4 was analysed by 3,5-Dinitrosalicylic acid (DNSA) reducing sugar assay (see Appendix 2 for more information) on tamarind xyloglucan (tXyG). Five different substrate concentrations were used, 0.25-2 mg mL^{-1} and incubated with enzyme (100 nM) at 37 °C and aliquots removed over time and enzyme deactivated. The rates for each different substrate concentration were plotted against substrate concentration, to give a relatively linear plot. Division of the gradient by the enzyme concentration gave a value for k_{cat}/K_m.

3.5.9 Crystallisation and data collection

Initial crystallisation screening was performed on all 3 *Tt*GH5 proteins using commercial screens; PACT, Index and Crystal HT. Wells were set up using a Mosquito crystal robot to produce sitting drops in a 1:1 ratio of protein to mother liquor; total volume of the drop was 300 nL.

3.6 Results

3.6.1 Thermal Shift Analysis

3.6.1.1 *Tt*GH5_un

Thermal shift analysis was used as a tool to probe any potential substrates of *Tt*GH5_un that could affect the protein melting temperature, measured at 65 °C, Figure 56. The bar chart clearly shows two increases in melting temperature; kGM and bMLG. Figure 57 shows the fluorescence curve measured for the apo protein and protein mixed with kGM. The extremely large shift on addition of solid kGM was investigated further and it was found that incubation with a lower concentration of substrate (1 mg mL^{-1}) affected the degree of positive shift in the melting temperature of *Tt*GH5_un. A higher amount of substrate may have allowed more of the protein access to the substrate and as such produce a larger shift in melting temperature. The kGM stock suspension was made at 1 mg mL^{-1}, however, removing small aliquots from the stock solution of insoluble compounds can often lead to errors in the expected substrate concentration of the assay solution (due to pipette clogging).

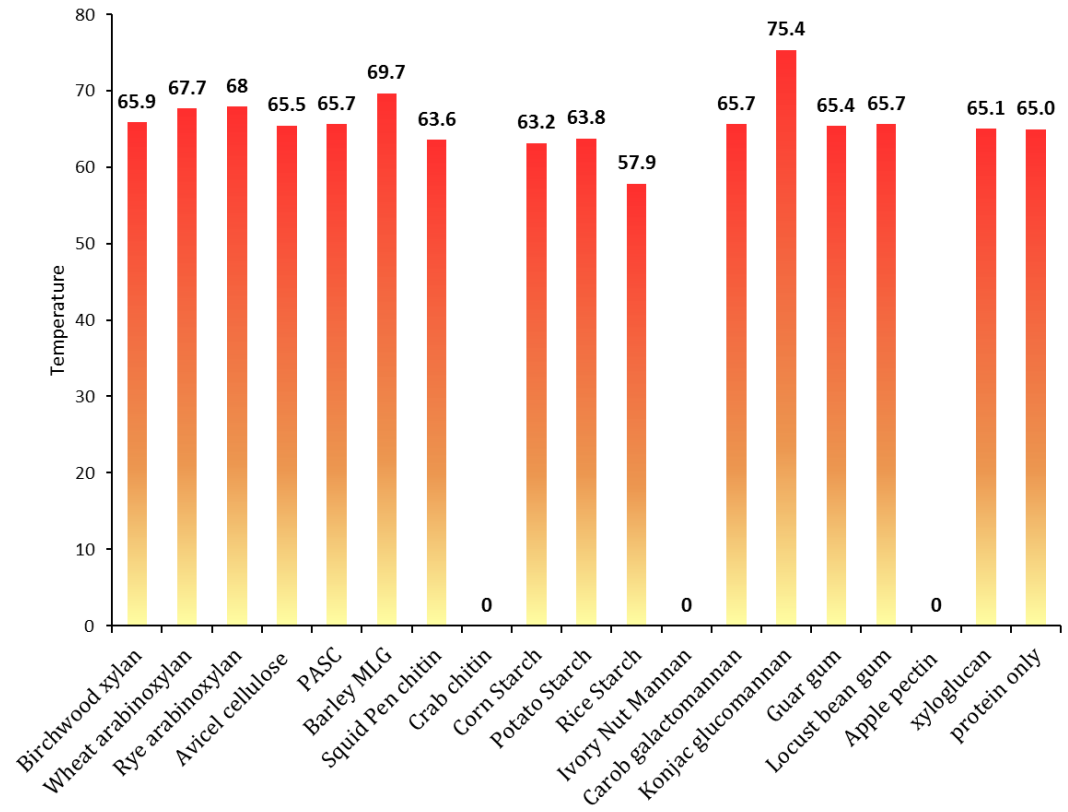


Figure 56 Thermal shift assay of *Tt*GH5_un against a range of different polysaccharides, with melting temperatures displayed as the average of three experiments. The protein only melting temperature is displayed on the far right at 65.0 °C, and the largest differences in melting temperature observed are with Barley mixed linkage glucan and Konjac glucomannan.

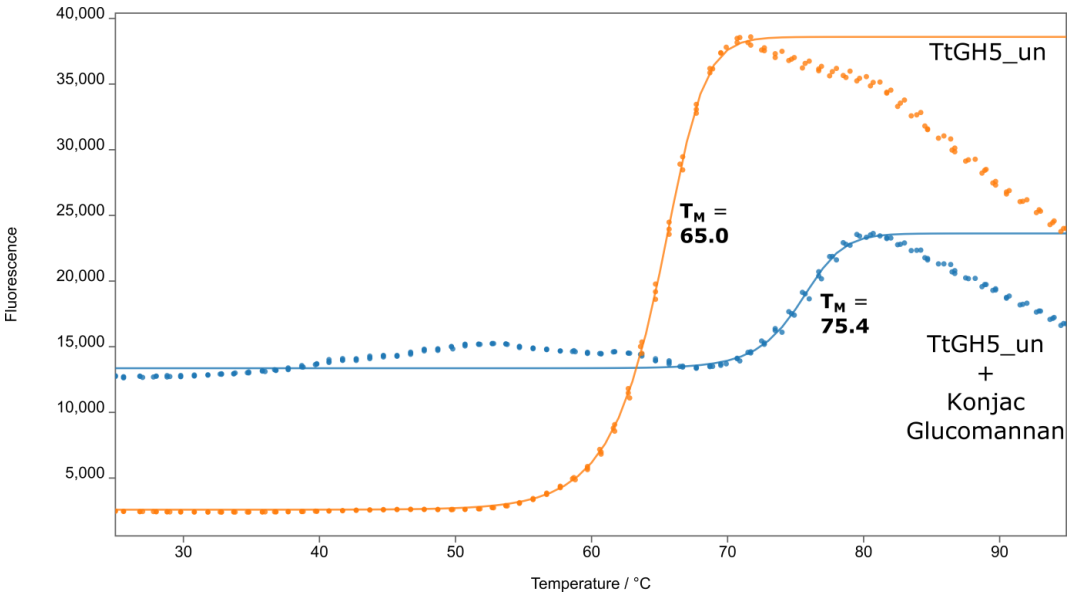


Figure 57 Thermal shift assay of *Tt*GH5_un with and without addition of solid Konjac glucomannan. Samples were mixed with SYPRO orange dye and temperature increased in 1 °C cycles up to 96 °C. Melting temperatures are displayed to the side of each curve, where a shift of approximately +10 °C was observed upon mix *Tt*GH5_un with the substrate.

3.6.1.2 *Tt*GH5_2

Small shifts were observed when *Tt*GH5_2 was mixed with mannose-based substrates. Figure 58 looks at the melting curves for GM, inM and kGM in detail compared with apo protein. Interestingly, a large destabilising effect was seen upon addition of mannohexaose, where a negative change in protein melting temperature of approximately 5 °C was observed.

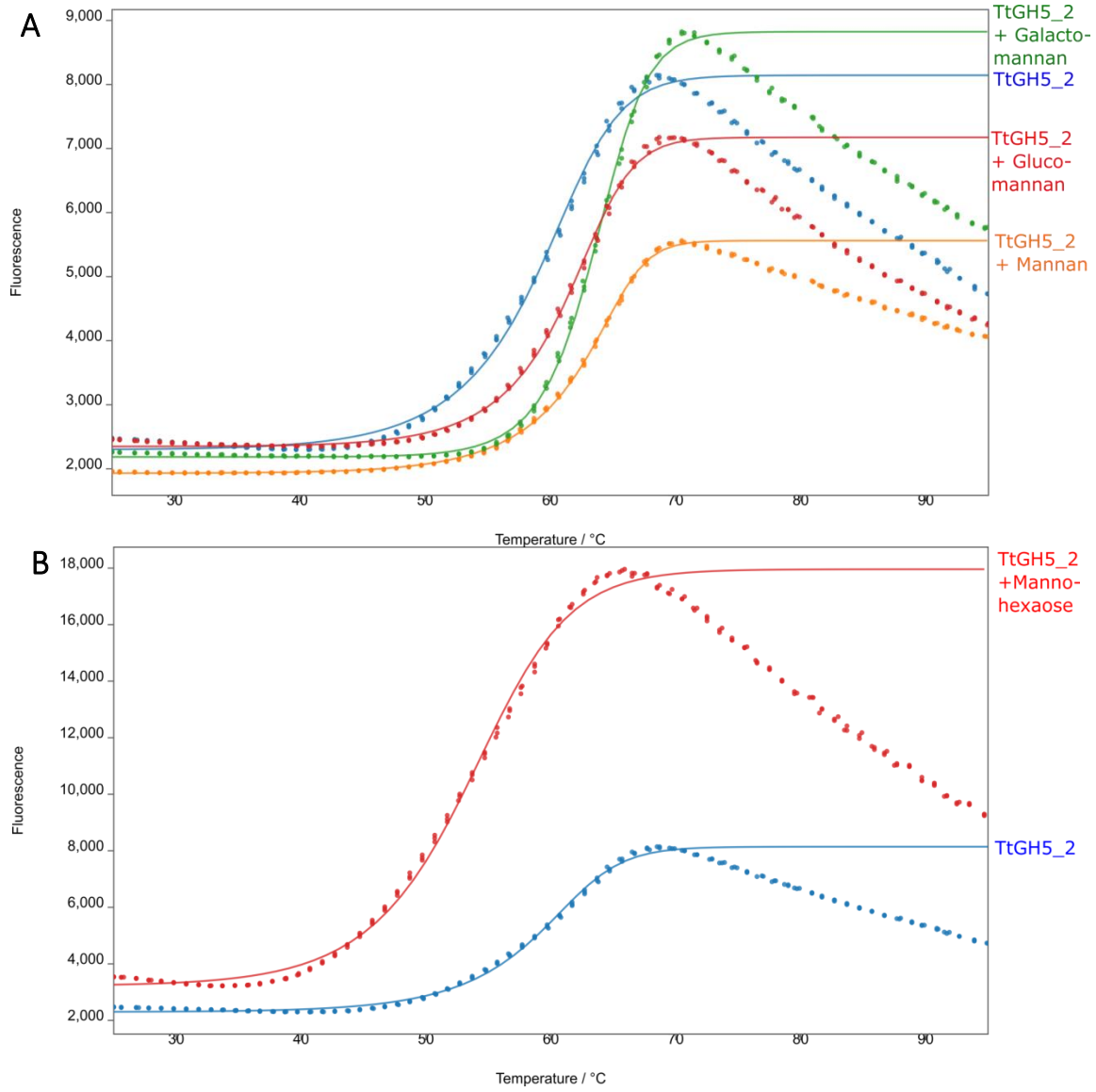


Figure 58 Thermal shift assay of *Tt*GH5_2 with and without addition of A) Ivory nut mannan, Carob galactomannan, Konjac glucomannan, and B) mannohexaose. Samples were run using SYPRO orange dye, a protein concentration of approximately 1 mg mL^{-1} and addition of approximately half the sample volume of substrate suspensions (1 mg mL^{-1}). Melting temperatures were calculated as the average of three experimentss and are as follows; Apo – 59.1°C, Ivory nut mannan – 62.8 °C, Konjac glucomannan – 63.6 °C, Carob glactomannan – 61.2 °C, mannohexaose – 54.2 °C.

3.6.1.3 *Tt*GH5_4

A positive shift in melting temperature of 3.2 °C is observed between the apo protein and the protein mixed with rAX, as shown in **Figure** 59. The proposed substrate, xyloglucan (tXyG) for this enzyme based on sequence analysis did not show a significant positive shift in melting temperature.

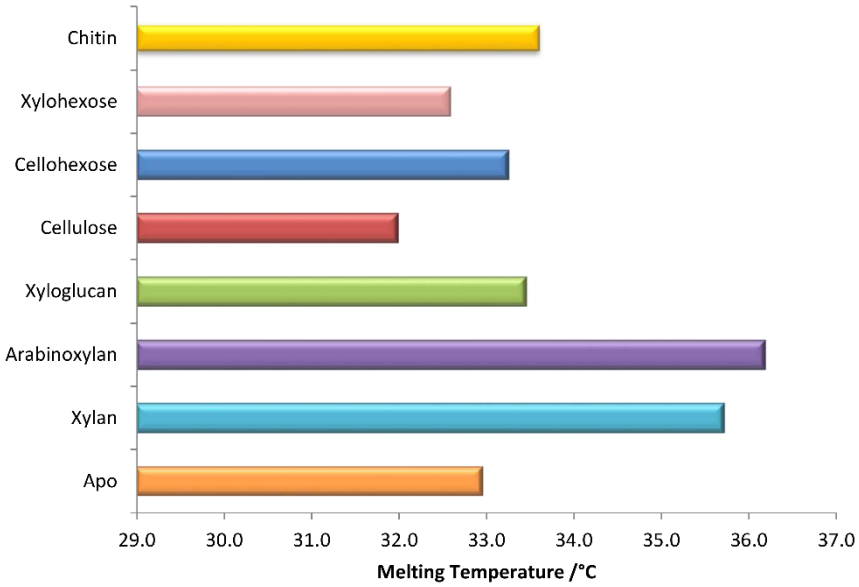


Figure 59 Protein melting temperatures of *Tt*GH5_4 when mixed with various polysaccharides as determined by TSA.

3.6.2 Activity analysis with insoluble colourmetric substrates

*Tt*GH5_2 was tested on AZCL-HE galactomannan and AZCL-HE cellulose (1 mg mL^{-1}). Breakdown of the dyed substrate is synonymous with the release of the dye molecules held within the cross linked insoluble substrate matrix, **Figure** 60. Once in solution, the dye molecules are soluble providing a quick visual test for enzyme activity whilst also being measurable by UV/Vis spectrometry at 590 nm. Addition of *Tt*GH5_2 and incubation overnight saw a small amount of dye released from the GM substrate compared with the control (0.115 Abs with enzyme, 0.04 Abs control). Whilst this small amount of dye release indicated that there was some activity, GM is unlikely to be the real substrate for *Tt*GH5_2, as the majority of the dye remained trapped in the insoluble substrate. In contrast, the breakdown of AZCL-HE cellulose with *Tt*GH5_2 was visible almost instantly, with the soluble fraction of the reaction mix turning blue within minutes of enzyme addition. **Figure** 60 shows

the dye release after an hour of incubation with *Tt*GH5_2 and the difference between the two types of substrates is very clear.

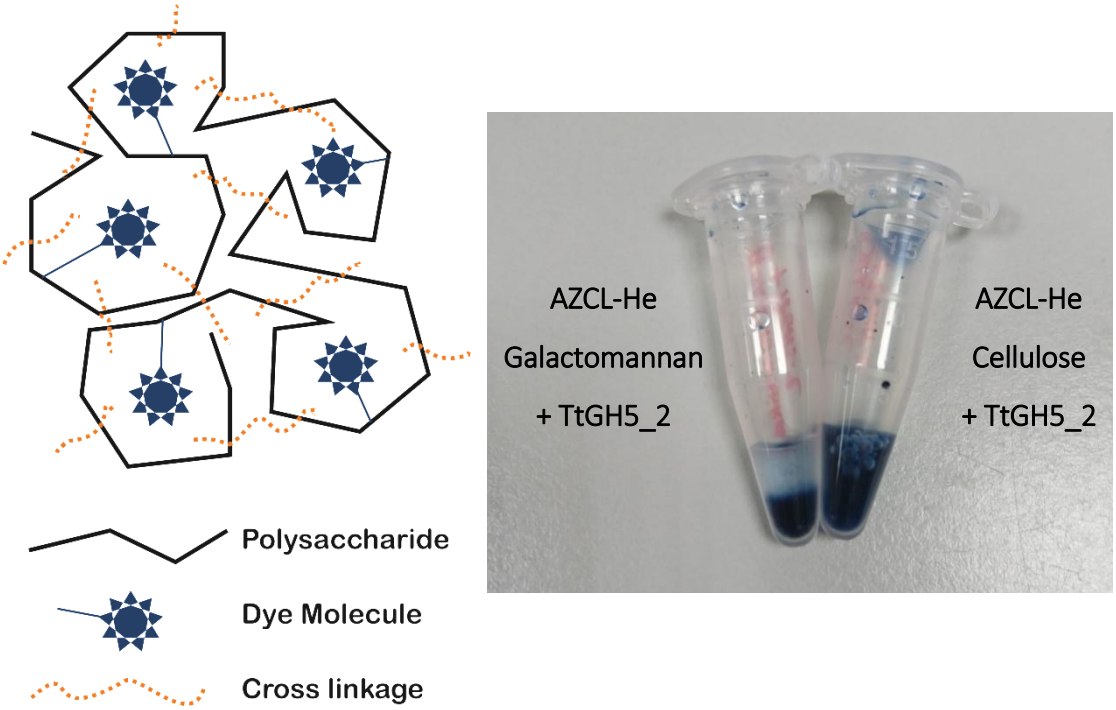


Figure 60 Insoluble polysaccharide colourmetric assay. Left; general diagram of the dye substrate, where dye molecules are held with the insoluble substrate matrix. Once substrate is degraded, dye molecules are released into the solution. Right; Incubation of AZCL-HE Galactomannan and AZCL-HE Cellulose (100 uL, 0.2% w/v solution) with *Tt*GH5_2 (5 uL, 45 mg mL^{-1} stock) for 1 hour at 25 °C shows visible dye release. A small amount of dye is released into solution in the galactomannan reaction but is not significant compared to the amount released from the breakdown of the cellulose substrate.

3.6.3 Hydrolytic Activity Analysis using TLC

3.6.3.1 *Tt*GH5_2

Based on the thermal shift assay indications, the activity of *Tt*GH5_2 on mannans was tested further with TLC, **Figure** 61. No degradation of GM was observed, which supports the lack of efficient enzyme activity during incubation with the dyed AZLC-He galactomannan. No degradation of inM or manno-oligosaccharides (2-6) was observed compared with the controls. *Tt*GH5_2 did show activity on kGM with discrete bands at various weights being observed on the TLC plate.

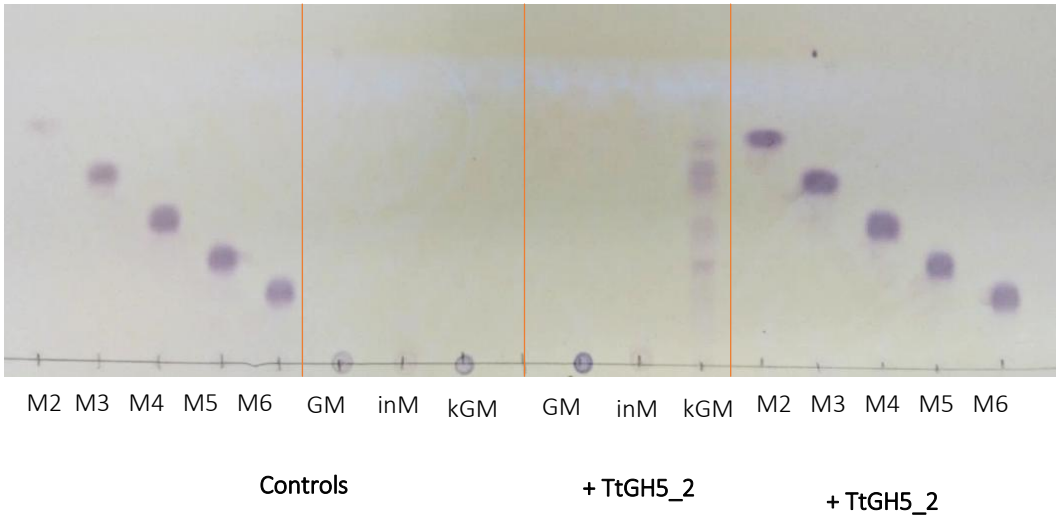


Figure 61 TLC analysis of TtGH5_2 activity on glucomannan (kGM), galactomannan (GM) and ivory nut mannan (inM). Standards shown on the left are mannobiose - mannohexaose. No activity is observed for galactomannan or ivory nut mannan when compared with the control samples. Glucomannan degradation products are observed at various weights as compared with the oligosaccharide standards. No activity is observed on the manno-oligosaccharides.

3.6.3.2 *Tt*GH5_un

The thermal shift assay indicated that *Tt*GH5_un showed significant protein-substrate interactions with kGM and bMLG. TtGH5_un was tested for hydrolytic activity on mannose-based substrates (kGM, GM and iNM) and bMLG by TLC, **Figure** 62. Hydrolysis reactions were run overnight at 37 °C using 1 mg ml^{-1} protein and 1 mg ml^{-1} substrate concentration. The soluble fraction of each reaction was analysed alongside control reactions that contained no protein. **Figure** 62 shows hydrolysis products of kGM, but the protein showed no hydrolytic effect on either GM or iNM as compared with the control samples. *Tt*GH5_un also displays very specific activity on bMLG.

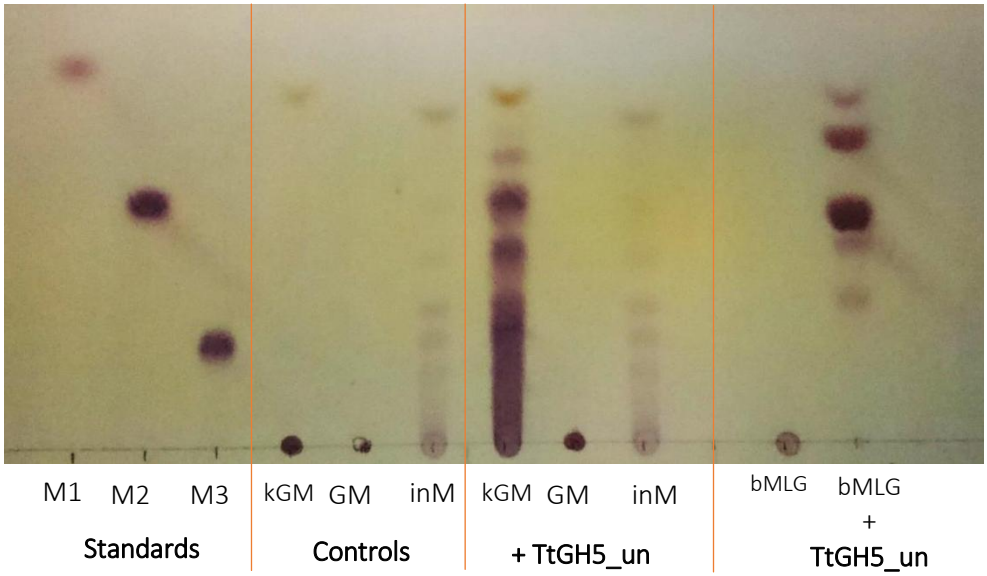


Figure 62 TLC analysis of *Tt*GH5_un activityon glucomannan (KGM), galactomannan (GM), ivory nut mannan (INM) and mixed linkage glucan (BMLG). Standards shown on the left are mannose, mannobiose and mannotriose. No activity is observed for galactomannan or ivory nut mannan when compared with the control samples. Glucomannan and mixed linkage glucan are broken down by TtGH5_un into discrete products of various sizes. bMLG is broken down in small products equivalent in size of mannobiose and smaller, whereas, kGM is broken down into a high density of different sized products.

3.6.3.3 *Tt*GH5_4

Hydrolytic activity of *Tt*GH5_4 was tested on a variety of xylans and tXyG, following on from the indications of possible interactions in the thermal shift assay. TLC analysis of the reaction products show that *Tt*GH5_4 is only catalytically active on tXyG, **Figure** 63. Two strong and discrete bands are shown on the TLC plate, and by comparison with the ladder, at molecular weights larger than X6 (810 g mol^{-1}). Several smaller molecular weight bands are seen but appear significantly fainter that the major reaction products. The lack of substrate remaining at the baseline, compared with the other substrate spots indicates that all the substrate has been solubilised and broken down.

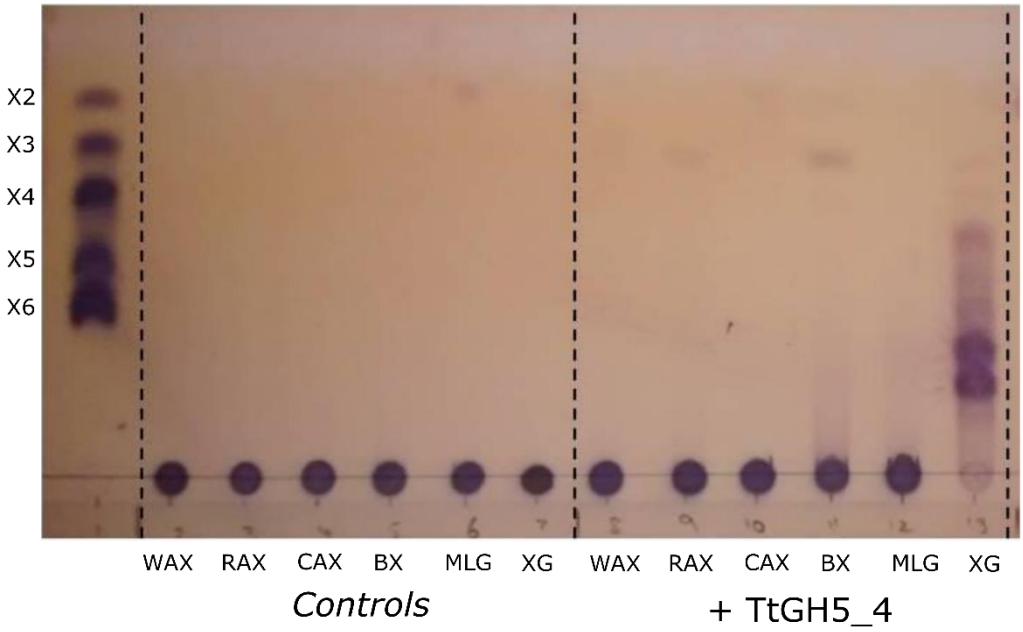


Figure 63 TLC of *Tt*GH5_4 on wAX, rAX, cAX, bX, bMLG and tXyG. The tXyG (written as XG above) was the only substrate specifically degraded by *Tt*GH5_4, showing two main discrete bands at a larger molecular weight than the standard X6 on the left-hand side of the plate. Fainter bands can be seen at lower molecular weights but are hard to distinguish. The general smearing of the band could be related to the complexity of the substrate, with a multitude of products possible. There are also faint bands around a molecular weight equivalent to X3 for rAX and bX.

3.6.4 Hydrolytic Activity Analysis using HPAEC-PAD of *Tt*GH5_4

The soluble fraction of the reaction of *Tt*GH5_4 on tXyG was then analysed by HPAEC-PAD, Figure 64. Comparison with the control shows a large release of soluble products upon incubation with the enzyme. Unfortunately, it was not possible to assign these peaks, which have instead been used, in this instance, to highlight the overall activity of TtGH5_4 on the substrate. Xyloglucan is primarily formed of a linear backbone of β (1-4) linked glucosyl residues but it is increased in its complexity through the addition of xylose moieties bound in an α (1-6) fashion with extended regions of galactosyl ,fuscosyl and/or arabinosyl resiudes as discussed in earlier in this chapter, Figure 54.[141-142] As such, the degradation pattern observed in Figure 64 reflects the complex nature of this substrate, and whilst it is not beneficial in determining an average 'output' product of *Tt*GH5_4, it highlights the enzyme's ability to navigate a complex polysaccharide.

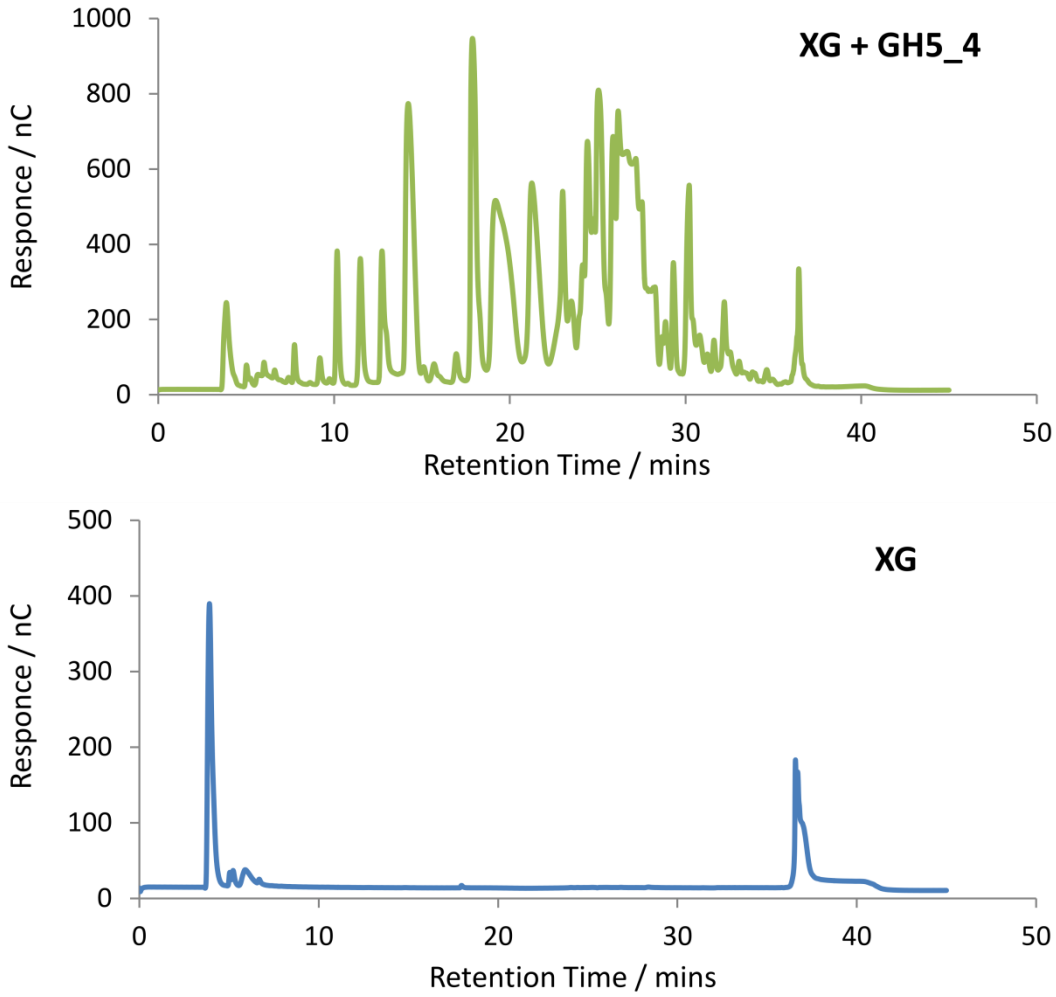


Figure 64 HPAEC-PAD spectra of *Tt*GH5_4 activity on tXyG. Hydrolysis reactions were carried out overnight at 37 °C using approximately 1 µM enzyme and 1 mg mL^{-1} substrate. The soluble fraction of the reaction and control was run on the HPAEC-PAD Dionex system using a sodium acetate gradient. Breakdown of tXyG, producing soluble products is observed on the addition of enzyme.

3.6.5 Hydrolytic Activity Analysis using MALDI-TOF-MS

3.6.5.1 *Tt*GH5_4

To expand on the results obtained from HPAEC-PAD analysis, where peaks were not characterised, the soluble reaction products of *Tt*GH5_4 incubated with tXyG was examined using MALDI-TOF, **Figure** 65. In the control reaction with no enzyme, there were no identifiable oligosaccharide peaks. A very high product response was observed in the hydrolysis reaction sample, indicating a large degree of substrate degradation by the enzyme. Major peaks were observed at 1409.5, 1247.4, 1115.4, 923.3, and 659.1 m/z and all have been assigned as oligosaccharide products of xyloglucan hydrolysis. The observed products correspond with those typically formed during the enzymatic breakdown of *tXyG* as discussed previously in this chapter.[138, 142-143]

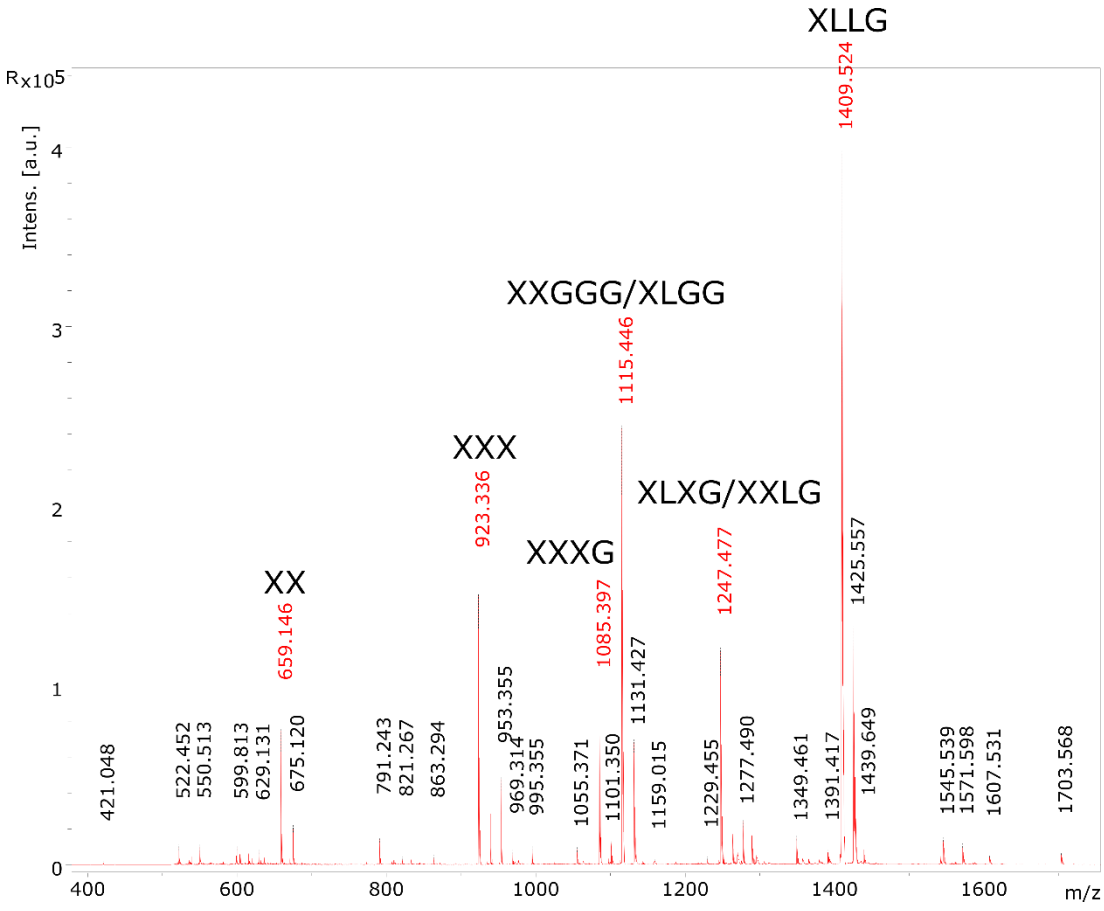


Figure 65 MALDI-TOF MS spectra of *Tt*GH5_4 (2.7 mg mL^{-1}) activity on xyloglucan (1 mg mL^{-1}). The m/z values of the major peaks have been labelled in red and annotated with the most likely oligosaccharide products; 1409.5 m/z (XLLG), 1247.4 m/z (XLXG/XXLG), 1115.4 m/z (XXGGG/XLGG), 923.3 m/z (XXX), 659.1 m/z (XX).

3.6.5.2 *Tt*GH5_2

The soluble products resulting from the hydrolysis reactions of *Tt*GH5_2 with mannose- and glucose-based substrates were analysed by MALDI-TOF mass spectrometry. As expected, soluble hydrolysis products of inM and GM were not observed. Hydrolysis of kGM saw the production of native oligosaccharides 4-17 degrees of polymerisation, Figure 66. Smaller peaks separated by 42 m/z units after the main oligosaccharide peaks are likely to be caused by an acetyl group found somewhere within the oligosaccharide chain. The soluble degradation products of PASC (phosphoric acid swollen cellulose) and Avicel also showed a clear production of oligosaccharides; PD 4-11 for PASC and PD 4-13 for Avicel. The mass spectra are shown as a mass range 825-2100 m/z, due to the presence of a highly intense peak in both samples at 803 m/z (also presentment in Figure 67, but to a smaller extent). This weight does not correspond to an expected degradation peak. Mass spectrometry of *Tt*GH5_2 activity on bMLG saw only one peak corresponding to that of a potential sugar product, cellotetraose. The mixed linked nature of the substrate may prevent production of larger products, allowing the enzyme to only cleave at certain points only the chain, avoiding the β (1-3) linkages.

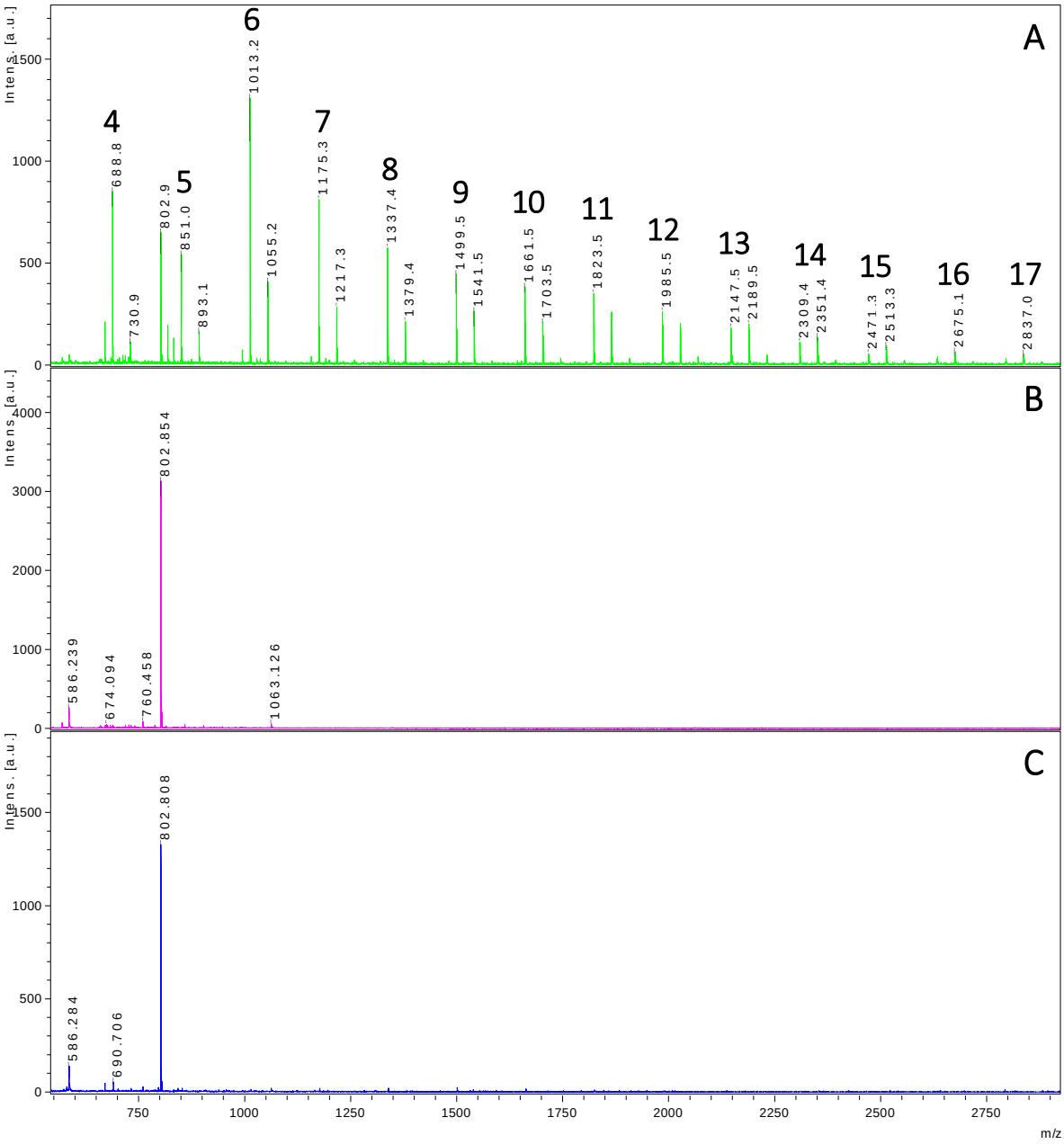


Figure 66 MALDI-TOF MS of the soluble hydrolysis products of TtGH5_2 activity with A) kGM, B) GM and C) inM. Clear degradation products are observed for KGM; oligosaccharides which are shown peak picked and labelled 4-17. Oligosaccharides with acetylation are separated by +42 m/z after the main peaks.

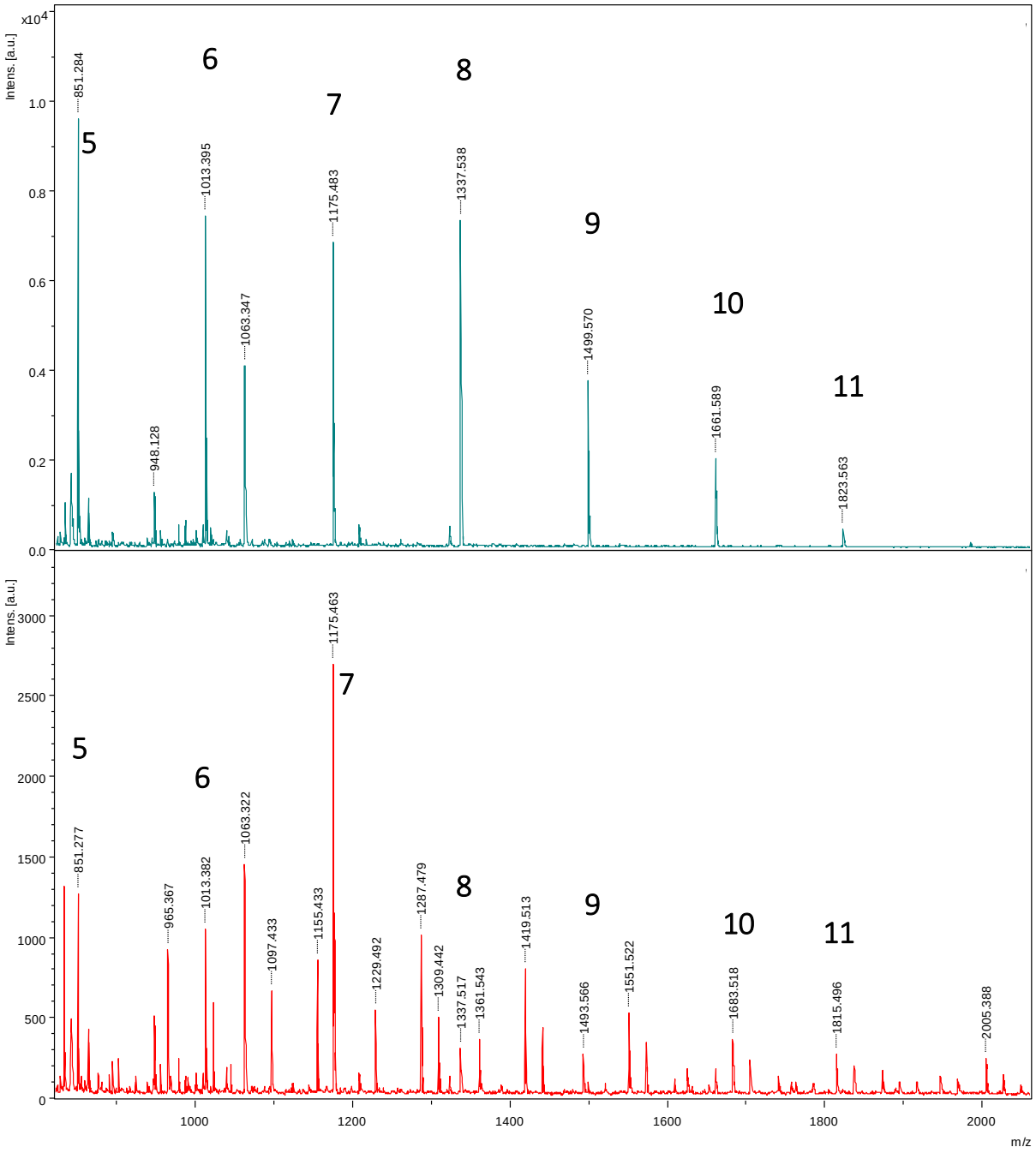


Figure 67 MALDI-TOF MS of *Tt*GH5_2 activity on PASC (top) and Avicel (bottom). Spectrum centred on 825-2100 m/z due to the presence of an intense peak at 803 m/z to allow analysis of the products. Smaller hydrolysis peaks were also observed at 689 m/z indicating for both substrates that products of a PD of 4-11 are formed.

3.6.5.3 *Tt*GH5_un

MALDI-TOF analysis of the soluble degradation products after incubating *Tt*GH5_un with kGM shows species for DP 4-11 (in the mass range tested, 500-2200 m/z), Figure 68. As seen on the TLC analysis, no product preference for TtGH5_un is observed, suggesting that the enzyme is endo acting. TLC analysis of TtGH5_un activity on bMLG emphasised the formation of products of a unit length of 4 and perhaps 1 or 2, with the latter allocation being difficult to assign. However, in Figure 69, the MALDI mass spectrum of the hydrolysis products show a clear product pattern; DP 4,7,9,12,15,18 and 21 display significantly larger intensities than any of the other observed products.

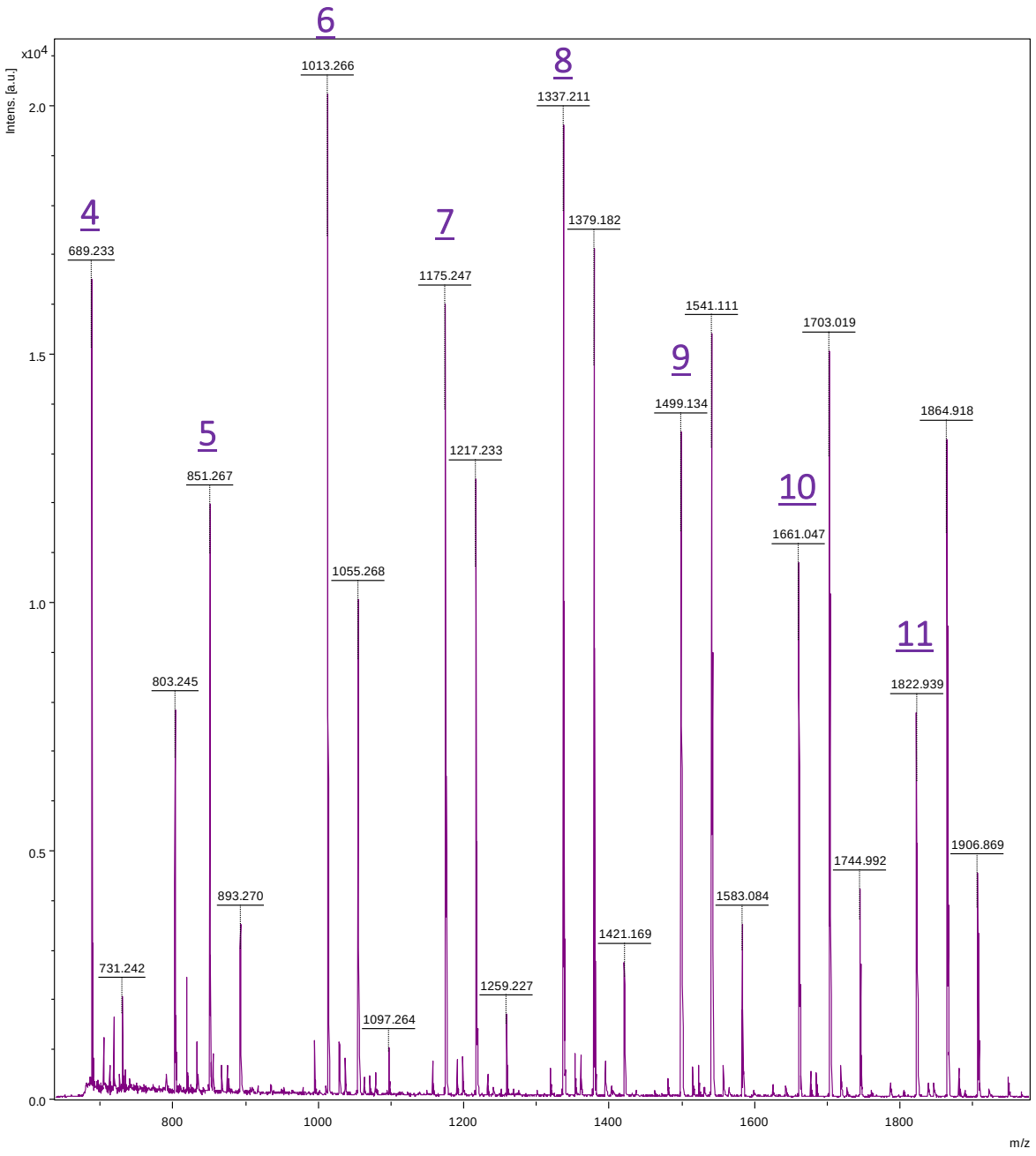


Figure 68 MALDI-TOF MS of the soluble reaction products of *Tt*GH5_un (0.1 mg mL^{-1}) incubated, shaking sideways, at 37 °C overnight with Konjac glucomannan. Native oligosaccharides are labelled above the corresponding peak, from 4-11 degrees of polymerisation.

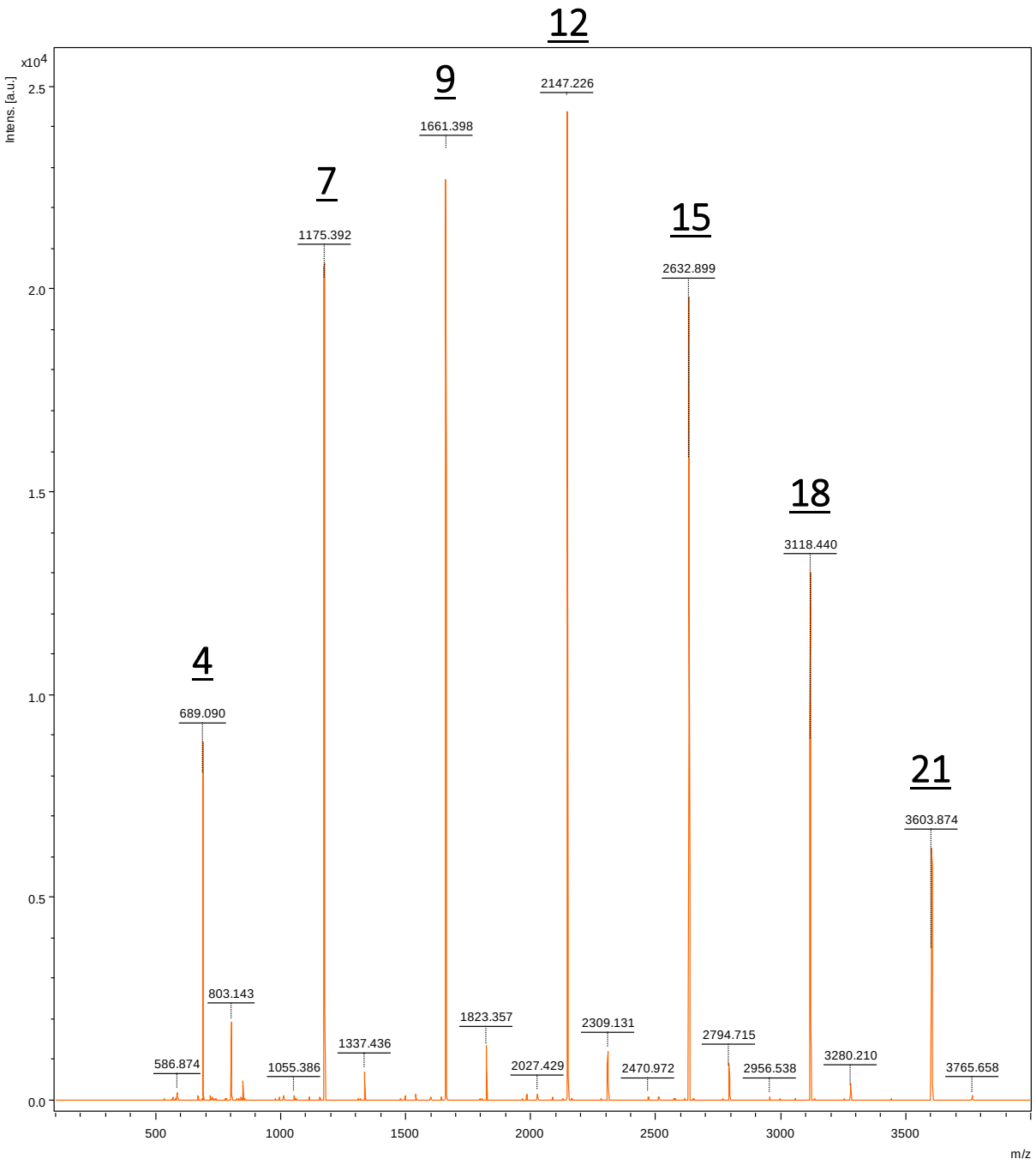


Figure 69 MALDI-TOF MS of the soluble reaction products of *Tt*GH5_un (0.1 mg mL-1 incubated, shaking sideways at 37 °C overnight with mixed linkage Barley glucan. The oligosaccharide product peaks are labelled in black as the DP size, with the spectrum showing a clear preference for products of certain sizes.

3.6.6 PACE

Both *Tt*GH5_2 and *Tt*GH5_un were analysed for activity by Dr. Theodora Tryfona, University of Cambridge using PACE.[140] The enzymes were analysed by PACE and confirmed to act strongly on MLG, weakly on iNM and relatively strongly on kGM, Figure 70.

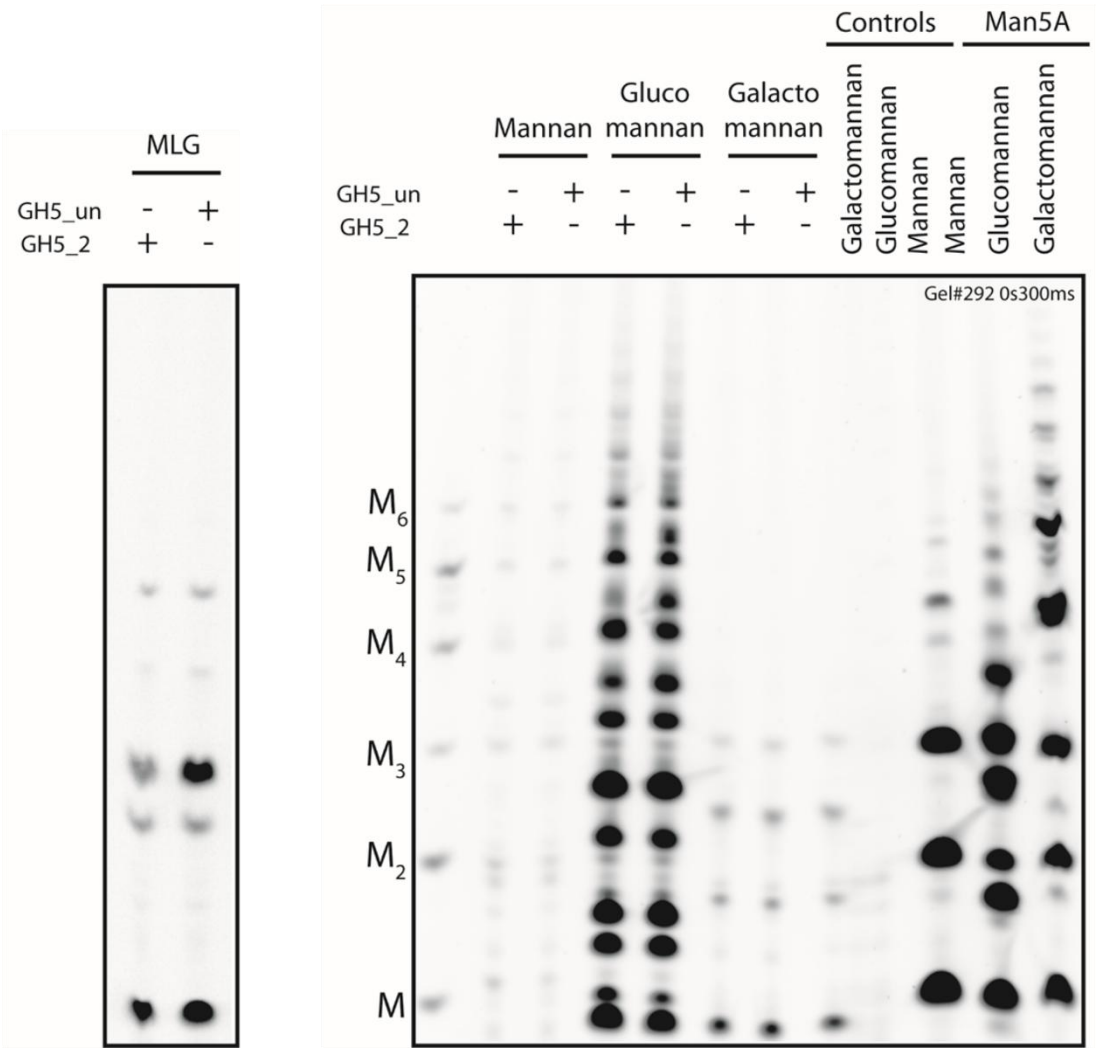


Figure 70 PACE showing activity of *Tt*GH5_2 and *Tt*GH5_un on MLG, iNM, kGM and GM, alongside various controls and a manno-oligosaccharide ladder. Experiment, analysis and figure performed and prepared by Dr. Theodora Tryfona, University of Cambridge.

*Tt*GH5_2 and *Tt*GH5_un were both found to act on deacetylated and acetylated kGM. The lack of co-migration of the products of both enzymes with those products resulting from control reactions with Man5A suggests that the GH5 products do not have mannose moieties at the reducing end. The kGM products of *Tt*GH5_un were analysed further by sequential addition of β-glucosidase and β-mannosidase which show a propensity for products of the form Man_xGlc with a small proportion of them also being $GlcMan_xGlc$.

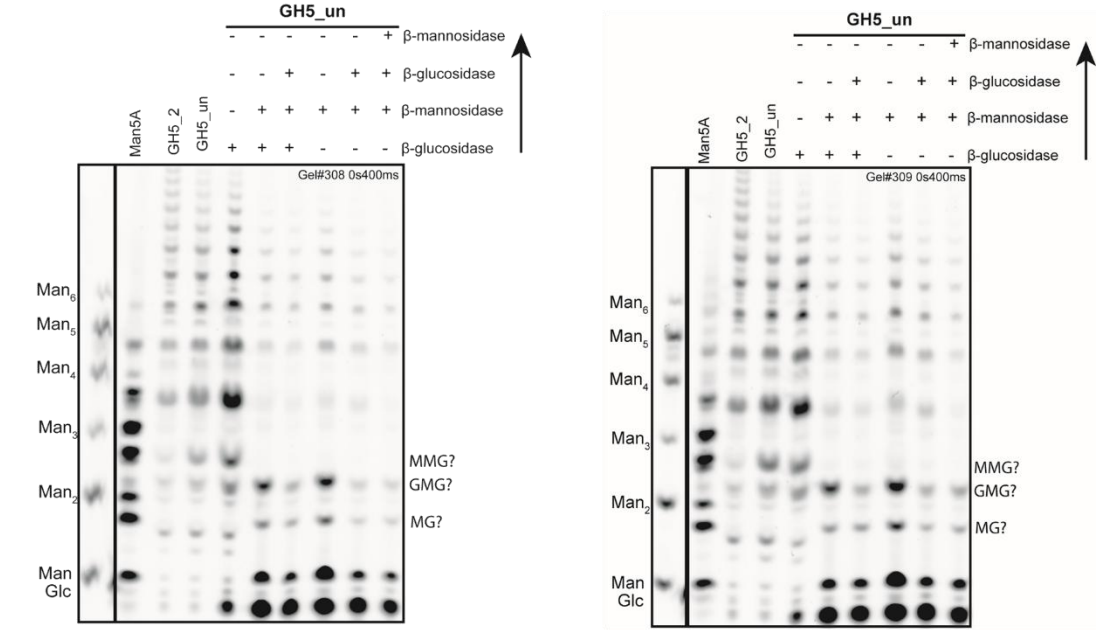


Figure 71 PACE of TtGH5_un on kGM, with sequential addition of either β-glucosidase or β-mannosidase to further break down the products. Left diagram refers to deacetylated kGM, and the right diagram shows acetylated kGM. Products are labelled as MMG, GMG and MG as potential product forms (where M is mannose, and G is glucose). Experiment, analysis and **Figure** prepared by Dr. Theodora Tryfona, University of Cambridge.

3.6.7 Kinetic Analysis

All 3 *Tt*GH5 enzymes have been subjected to activity analysis with certain substrates, and k_{cat}/K_M derived. Based on the reactivity of *Tt*GH5_2 and *Tt*GH5_un on both bMLG and cellulose substrates, it was deemed sensible to test both enzymes on 4-Methylumbelliferyl-β-cellobioside (4-MU-C2), 4-Methylumbelliferyl-β-cellotrioside and 4-Methylumbelliferyl-β-laminaribioside (4-MU-B1/3-C2). *Tt*GH5_4 has been shown to be active on xyloglucan and was instead analysed using the 3,5-dinitrosalicylic acid (DNSA) reducing sugar assay.

3.6.7.1 *Tt*GH5_2

An appropriate enzyme concentration was determined for *Tt*GH5_2 by testing between 1, 0.5 and 0.125 µM enzyme with a constant substrate concentration of 1 mM 4-MU-C2, Figure 72. This scoping experiment also determined whether the enzyme was active on the fluormetric substrates. Figure 72 shows the change in fluorescence over time for the reaction of TtGH5_2 with 4-MU-C2. The rate of change over the experimental timescale indicates that the substrate is being cleaved by the enzyme. The same experiment repeated with 4-MU-B1/3-C2 however did not show any change in fluorescence, indicating that the enzyme is incapable of breaking down this substrate.

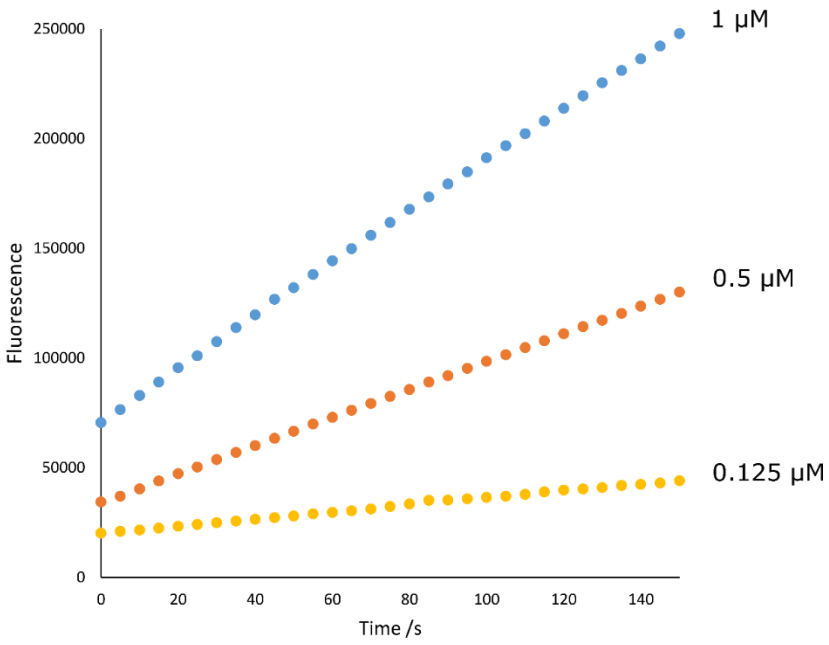


Figure 72 Enzyme scoping experiment using 4-MU-C2 (1mM) and TtGH5_2 (1, 0.5 and 0.125 µM).

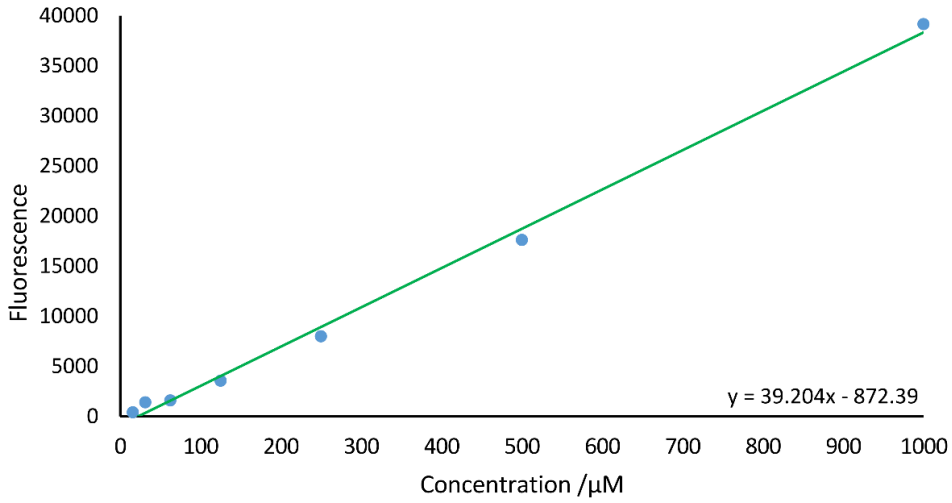


Figure 73 Standard curve of fluorescent product 4-MU over a concentration range, 15.6 - 1000 µM.

In order to assess the amount of product formation during the enzymatic reaction, a standard curve with known amounts of the fluorescent product 4-MU was produced as shown in **Figure** 73. A concentration of 0.5 µM enzyme was chosen based on the concentration scoping experiment and *Tt*GH5_2 was tested on different concentrations of 4-MU-C2 and 4-MU-C3 over a short timescale. A significant difference can be seen in the rate of reaction of *Tt*GH5_2 against the two substrates, with more activity observed for the longer cellotriose based substrate. As such during collection of the data, the gain of the

instrument needed to be reduced for the reaction of TtGH5_2 on 4-MU-C3 as the fluorescence was out of the measurable range (indicating high reactivity). **Figure** 74 shows the activity of TtGH5_2 on 4-MU-C2, where the rates of reaction are plotted against the substrate concentration. **Figure** 75 shows the product formation curve of TtGH5_2 on 4-MU-C3, followed by **Figure** 76, where the rates are plotted against substrate concentration. In the latter **Figure**, the plot observed true Michaelis-Menten kinetics, and as such the data was fitted using the 'enzyme kinetics' mode within Origin to deduce the values for V_{max} and K_M. **Table** 13 draws together the kinetic values determined for *Tt*GH5_2.

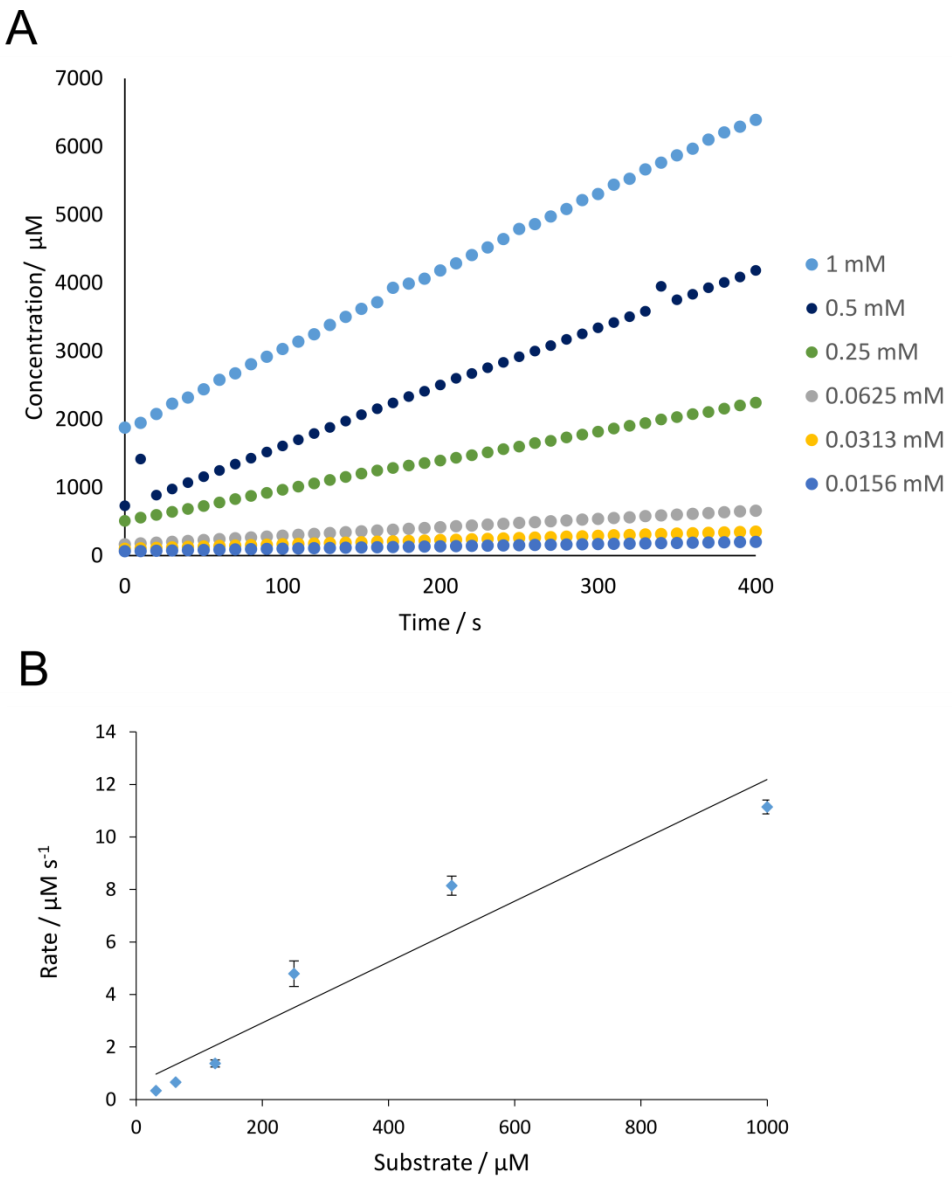


Figure 74 Kinetic analysis of *Tt*GH5_2 on 4-MU-C2. A) Plot of the formation of 4-MU product over time. B) rate of reaction plotted against the corresponding substrate concentration.

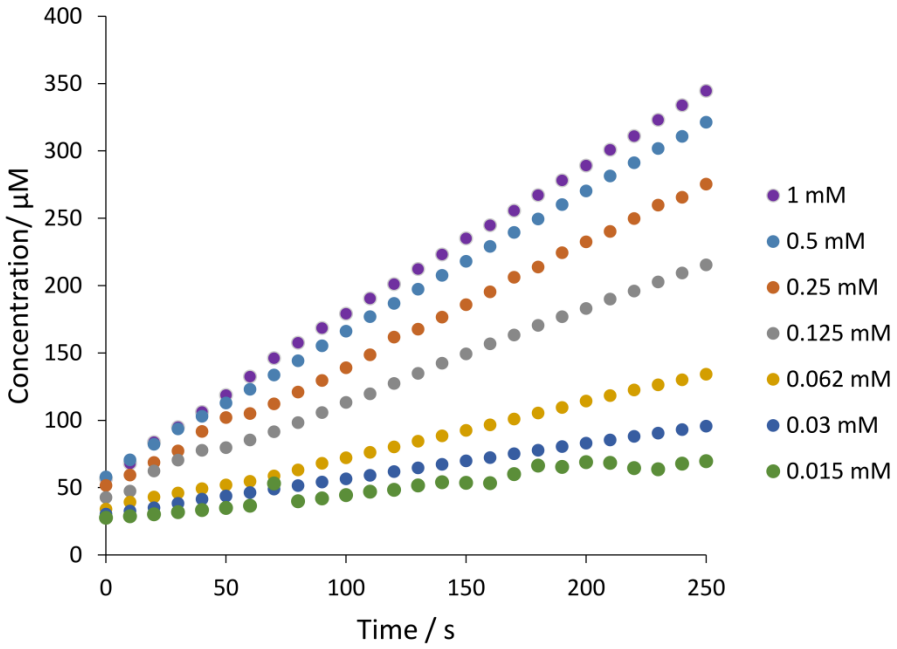


Figure 75 Product formation curves of *Tt*GH5_2 on 4-MU-C3 over time.

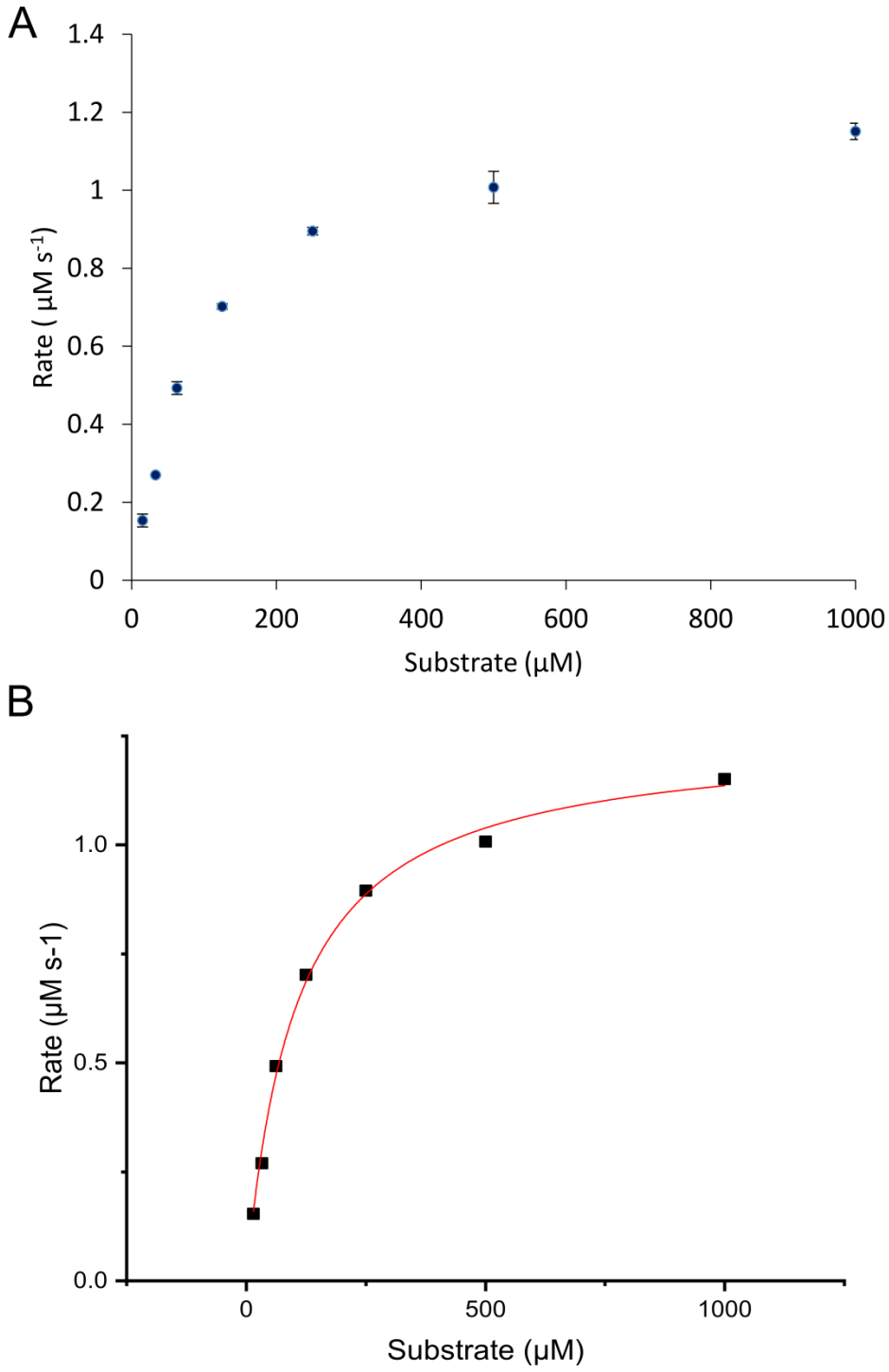


Figure 76 Kinetic analysis of *Tt*GH5_2 on 4-MU-C3 . A) Plot of reaction rate against substrate concentration, showing error bars determined by calculating the standard error of the mean of the data. B) the same plot fitted to Michaelis-Menten kinetic analysis using Origin.

Table 13 Kinetic parameters determined for TtGH5_2 on three substrates, 4-MU-C2, 4-MU-C3 and 4-MU-β1/3-C2.

Substrate	V_{max} (μM s^{-1})	K_M (μM)	k_{cat}/K_M (μM^{-1} s^{-1})
4-MU-C2	Not determined	Not determined	2.32 x10^{-2} ± 1.05 x10^{-4}
4-MU-C3	1.25311 ± 0.0268	103.12 ±7.29	2.43 x10^{-2} ± 7.2 x10^{-3}
4-MU-β1/3-C2	*No reaction*	*No reaction*	*No reaction*

3.6.7.2 *Tt*GH5_un

*Tt*GH5_un was treated in the same way as described for *Tt*GH5_2 above. The enzyme scoping experiment suggested an ideal concentration of 0.5 μM and suggested that the protein is active on 4-MU-C2 albeit at a lower rate than observed for *Tt*GH5_2, **Figure** 77. The timescale of the experiment was significantly longer than that used for TtGH5_2 and produced a smaller overall fluorescence change. *Tt*GH5_un proved similar to *Tt*GH5_2 in that it was also unable to react with 4-MU-B1/3-C2, as no change in fluorescence was observed over the timescale of the reaction.

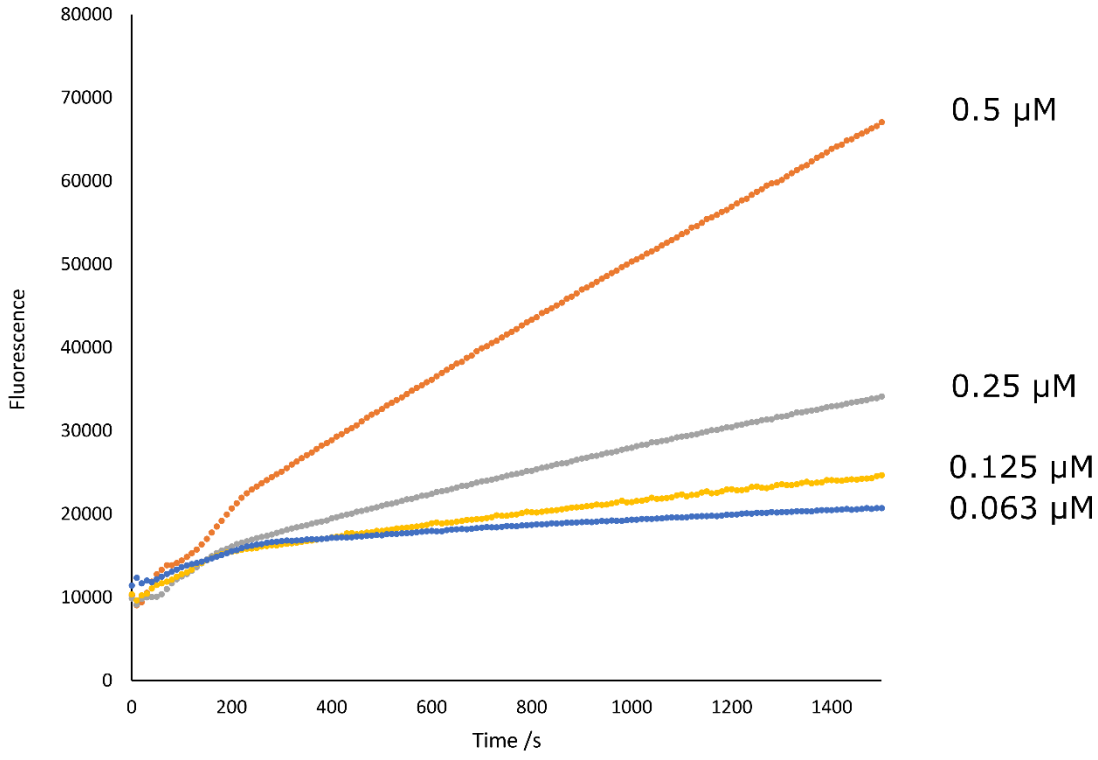


Figure 77 Enzyme scoping reaction with 4-MU-C2 (1 mM) and *Tt*GH5_un (0.5, 0.25, 0.125 and 0.063 μM).

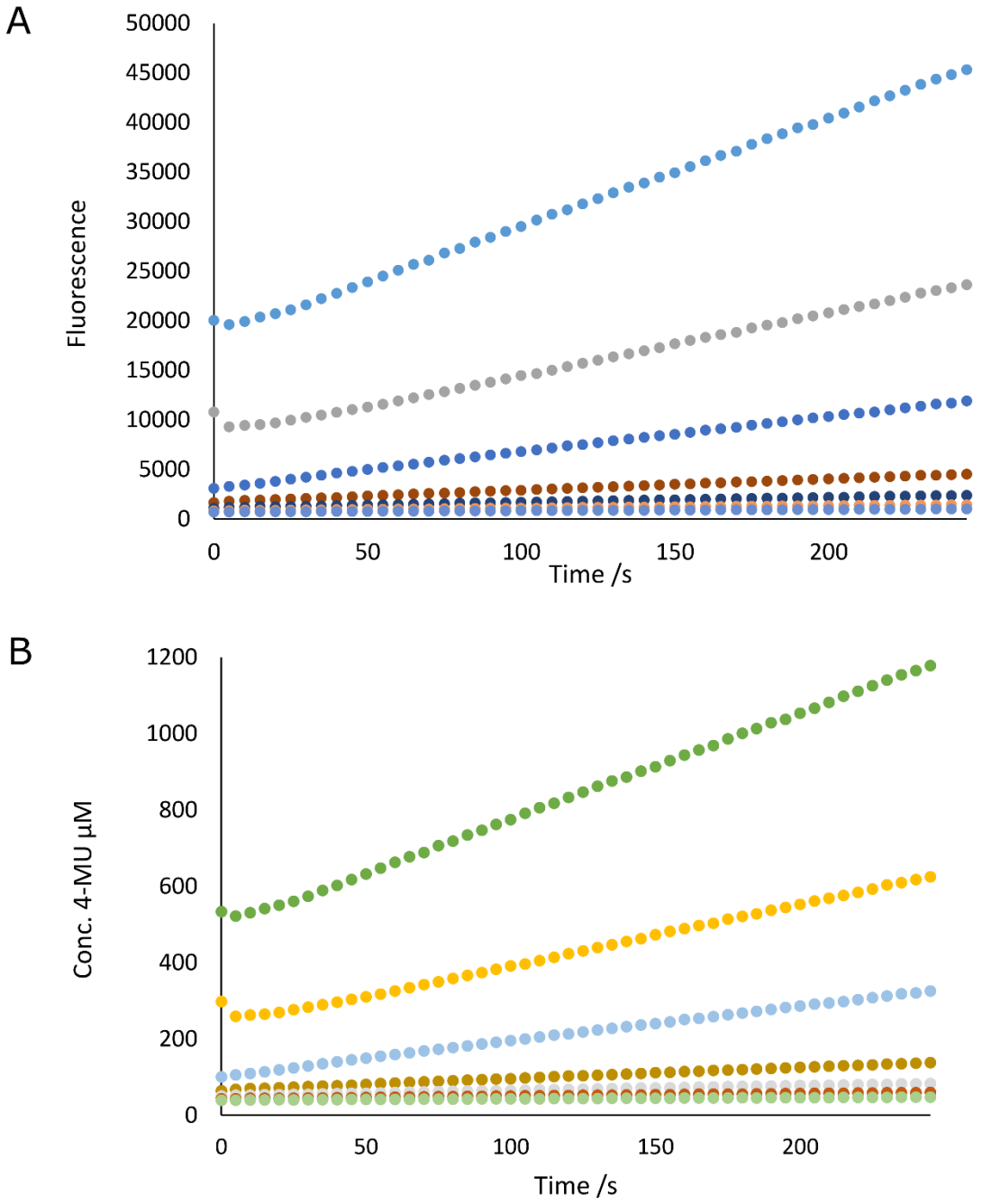


Figure 78 Analysis of multiple curves produced in *Tt*GH5_un assay. A) The effect of changing the substrate concentration on the reaction of *Tt*GH5_un with 4-MU-C2. B) reaction shown as the increase in production formation over time, as calculated using the fluorescence data in A and a standard curve of the product complex (not shown).

Figure 78 shows the experiment in which the change in fluorescence from reactions of different substrate concentrations with a set amount of enzyme, can be monitored. The recorded fluorescence values can be converted into a measure of product formation using the specific relationship *y=mx+c* that describes the standard curve shown in **Figure** 73 Standard curve of fluorescent product 4-MU over a concentration range, 15.6 - 1000 µM.

The gradients for each individual plot of concentration of 4-MU vs. time are themselves plotted against their corresponding substrate concentration, **Figure** 79. The change in product formation over time is essentially the rate of reaction, and as such the rates plotted against substrate concentration yield a final overall rate constant, in this case equal to 0.0026 s^{-1}. Division of this value by the enzyme concentration (0.5 µM) produces a value for the rate constant k_{cat}/K_M; 5.2 $x10^{-3}$ ± 1.05 $x10^{-4}$ μM^{-1} s^{-1}.

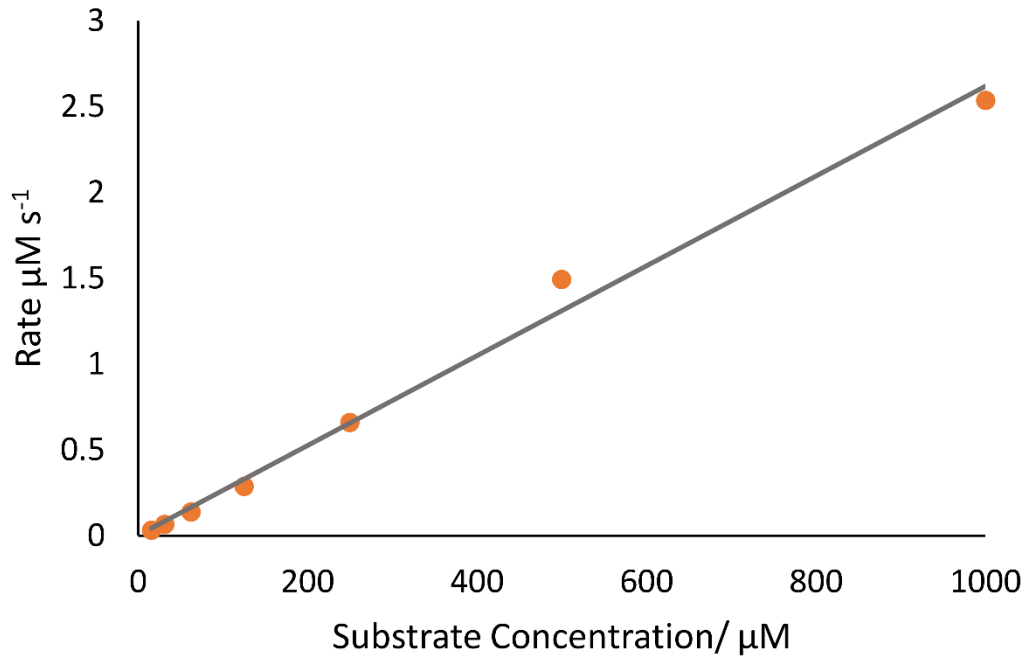


Figure 79 Final graphical analysis of the activity of *Tt*GH5_un on 4-MU-C2, showing rate of reaction against substrate concentration. The gradient of this plot is used to work out a value for k_{cat}/k_M.

*Tt*GH5_un was then tested on 4-MU-C3 and found to display Michaelis-Menten kinetics on this substrate, **Figure** 80. Origin was used to analysis plot B in **Figure** 80, from which the values for V_{max} and K_M have been determined. The third kinetic parameter, k_{cat}/K_M was also calculated and all parameters shown in **Table** 14.

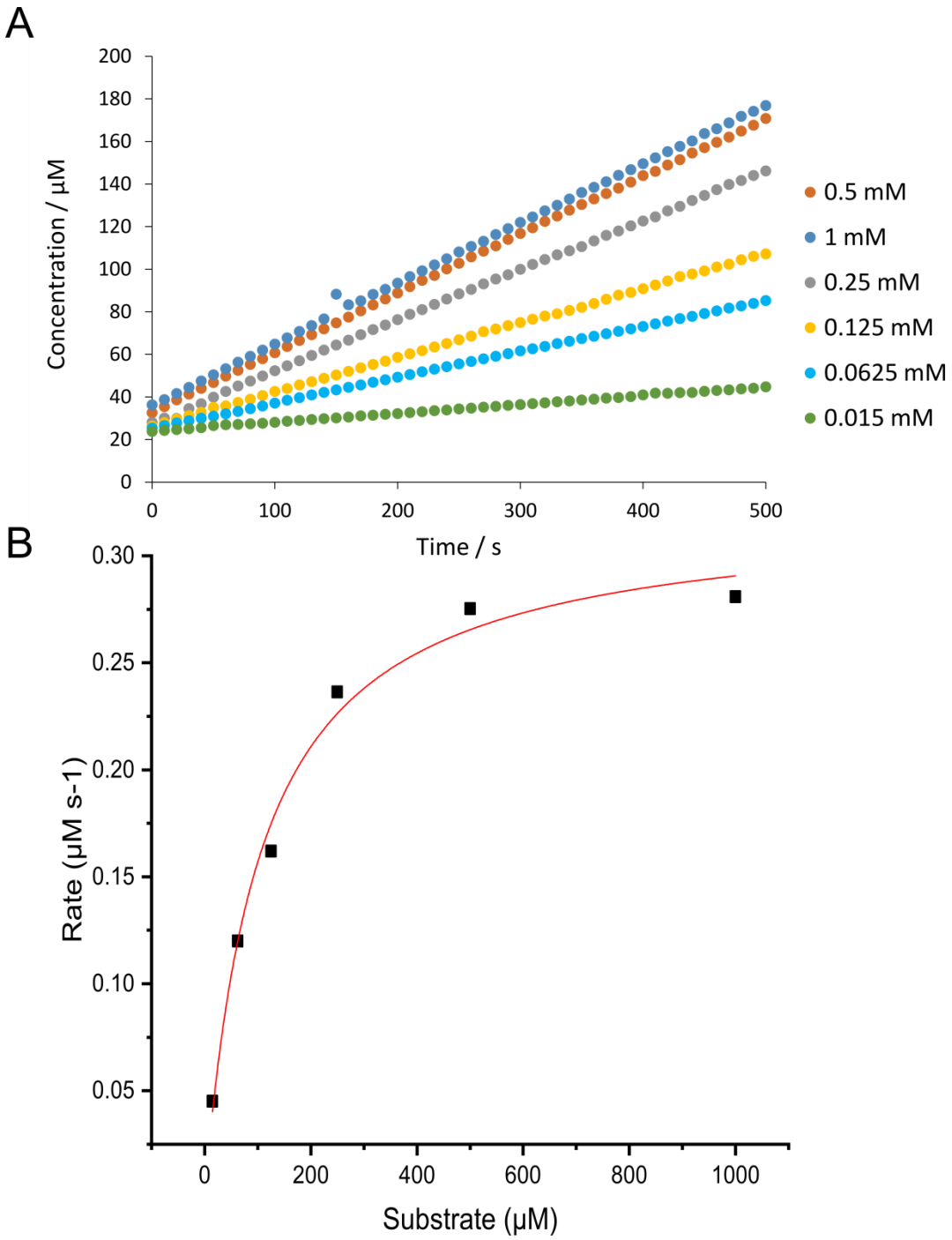


Figure 80 Kinetic analysis of *Tt*GH5_un on 4-MU-C3. A) Plot of product formation over time. B) Michaelis-Menten analysis of the rates of reaction plotted against substrate concentration.

Table 14 Kinetic parameters determined for TtGH5_un on three substrates, 4-MU-C2, 4-MU-C3 and 4-MU-β1/3-C2.

Substrate	V_{max} (μM s^{-1})	K_M (μM)	k_{cat}/K_M (μM^{-1} s^{-1})
4-MU-C2	Not determined	Not determined	5.2 x10^{-3} ± 1.05 x10^{-4}
4-MU-C3	0.321 ± 0.0124	104.61 ±14.18	6.00 x10^{-4} ± 1.75 x10^{-3}
4-MU-β1/3-C2	*No reaction*	*No reaction*	*No reaction*

3.6.7.3 *Tt*GH5_4

Initial kinetic analysis of *Tt*GH5_4 on tXyG was carried out using the 3, 5-dinitrosalicyclic acid (DNSA) reducing sugar assay (Appendix 2 for information on method). The enzyme was tested using 5 different substrate concentrations, carried out at 37 °C, and rates taken and plotted against substrate concentration. Figure 81 shows the plots required to produce a value of k_{cat}/K_m, whereby the gradient of the plot of rate against substrate concentration is divided by the concentration of protein used to yield a value of 7.71 x10^5 mg^{-1} ml min^{-1}.

3.6.8 Crystallisation and Data Collection

A crystal hit was obtained for *Tt*GH5_un, where crystals grew in a condition containing 0.1 M HEPES, 25% w/v peg 3350, 0.2M NaCl, at pH 7.4. A 24-well optimisation screen containing variations on the mother liquor was set up using protein at a very high concentration (300 mg mL^{-1}). Crystals were fished and found to diffract to 1.8 Å. Crystal data sets were collected at the Diamond synchrotron (Oxford) by remote collection. Due to lack of sequence homology with other available proteins in the PDB, a structure solution was not determined by molecular replacement.

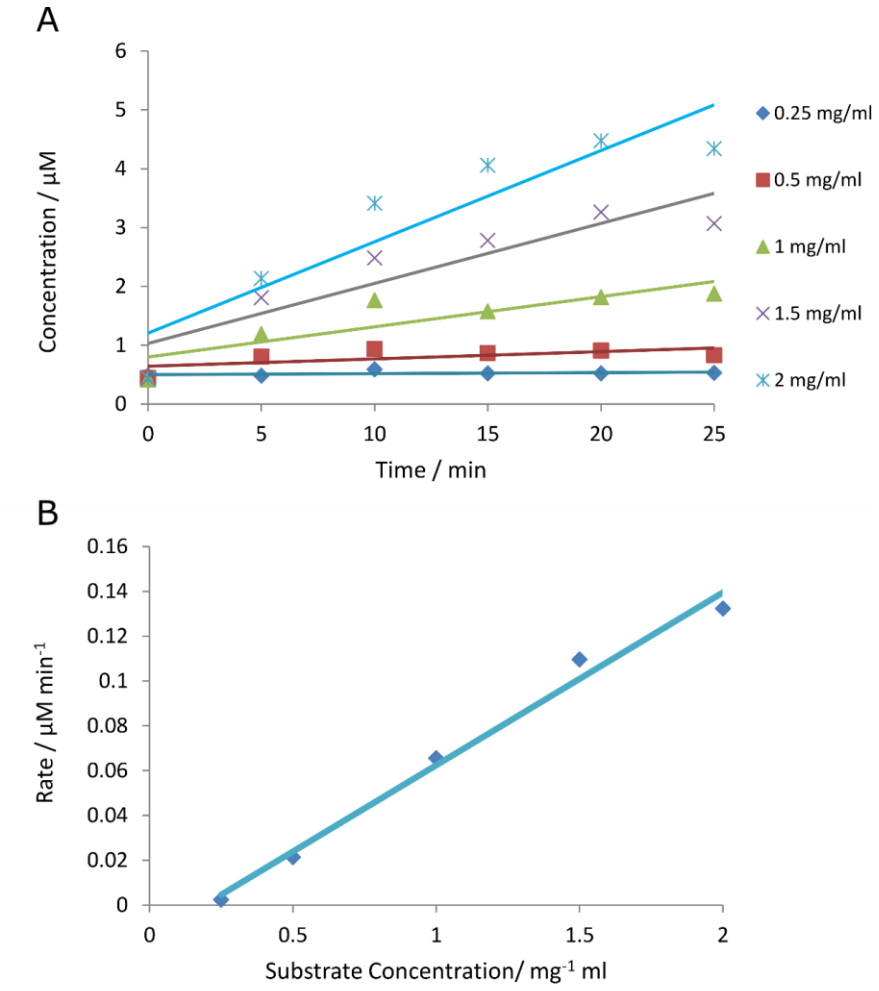


Figure 81 DNSA analysis of *Tt*GH5_4(100 nM) on tXyG. A) Concentration of reducing sugars produced by action of TtGH5_4 against time for 5 different substrate starting concentrations of tXyG. B) Rate of tXyG hydrolysis by *Tt*GH5_4 against substrate concentration.

3.7 Discussion

3.7.1 *Tt*GH5_2

During the analysis of this work it came to light that *Tt*GH5_2 had been previously studied by Ekborg *et al* as part of a multidomain protein called CelAB, which found both cellobiohydrolase and endo- glucanase activity.[105] CelAB contains both GH5 and GH6 domains as was noted in Table 4 (Chapter 2). Due to the difference in name, and accession code at the start of this project, this study was not refered to during the experimental work. Subsequent updates the the CAZy database (including a change in the accession code to be a combined code) brought this study to light and it has proved a useful comparison regarding the functionalty of the GH5_2 domain when separated from its GH6 partner domain.

The combined activity experiments testing the degradation of kGM, inM and AZCL-He cellulose suggest that *Tt*GH5_2, whilst it may interact with mannose-based substrates (based on small positive shifts in melting temperature) the interaction is likely to be non-specific and does not yield hydrolytic activity against mannose glycosidic bonds. The ability of the protein to break down kGM but not inM, where the only difference is the dispersion of glucose moieties within the polysaccharide chain, suggests that enzymatic cleavage is limited to glycosidic linkages at glucose moieties. TLC analysis showed a wide range of weights of hydrolysis products on incubation of *Tt*GH5_2 with kGM. This result was confirmed by MALDI mass spectrometry, where native oligosaccharides products in the range 4-17 degrees of polymerisation where observed. Smaller peaks at +42 m/z to the oligosaccharides are suggested to be due to acetylation of the mannose residues, a feature known to be present within the kGM substrate. Although defining the position of the acetyl groups would not be possible by this method of mass spectrometry, fragment analysis may provide a better view. Glucomannan is made up of β 1-4 linkages of glucose and mannose in a 1:1.6 ratio respectively.[144] However it is likely that the activity of *Tt*GH5_2 on kGM is an opportunistic hydrolytic reaction, and kGM is not the natural substrate of the enzyme. Indeed, analysis using PACE showed that the enzyme was able to degrade both MLG and kGM, but the latter products were not likely to have mannose reducing ends, which signifies that the enzyme only cuts at a glucose moiety.

The high rate of dye release upon addition of *Tt*GH5_2 to AZCL-He cellulose lead to the hypothesis that glucose-based polysaccharides are the major substrate for this enzyme. Dye release was observable by eye as soon as the enzyme was added to the polysaccharide

solution; the reaction could be seen occurring as a 'cloud' of blue colouring emerging around the 1 µL of enzyme, before the sample was mixed thoroughly. Conformation of activity on normal cellulose substrates, PASC and Avicel was carried out by MALDI mass spectrometry. Similarly, to the activity observed on kGM, *Tt*GH5_2 produced cleavage products in a wide range of molecular weights; PD 4-11 for PASC and PD 4-13 for Avicel. *Tt*GH5_2 was found to only produce one peak relating to an oligosaccharide (likely cellotretraose) with mass spectrometry when incubated with bMLG. It may be that the enzyme is unable to accommodate parts of the substrate into the binding pocket due to the addition of β-(1,3) bonds. Another possible explanation may be due to substrate solubility, bMLG is easily solubilised whilst the other cellulosic substrates PASC and Avicel remain insoluble in solution. The enzyme may prefer to hydrolyse more crystalline regions of cellulose, rather than the easily accessible substrates like bMLG. The hydrolysis pattern of the cellulose substrates suggests that *Tt*GH5_2 acts as an endoglucanase, working at random points within the polysaccharide chain, to produce a wide range of products. This is an interesting finding when compared with the Ekborg study on CelAB, where it was found that the multidomain protein displayed a higher specific activity on bMLG than on CMC, but no activity was displayed on PASC.[105]

*Tt*GH5_2 displayed Michaelis-Menten kinetics on the cellotriose based fluorescent substrate, 4-MU-C3, and not on the cellobiose substrate. There was also no activity with the substrate 4-MU- β1/3-C2 which suggested that the enzyme is not able to fit β (1-3) linkages within its active site. The higher activity of the enzyme on 4-MU-C3 allowed the kinetic parameters V_{max} and $K_{M,}$ to be determined as 1.25311 ± 0.0268 µM s^{-1} and 103.12 ±7.29 µM respectively. The activity of *Tt*GH5_2 is comparable to other GH cellulases, with similar K_M in the µM range, indicating that the enzyme has a high binding efficieny for 4-Mu-C3.[127, 145] The enzyme is suggested to be more proficient at interacting with a substrate of three units or larger.

3.7.2 *Tt*GH5_4

*Tt*GH5_4 was found to be active on xyloglucan. Only six GH5 enzymes are active on xyloglucan and are subsequently only found in subfamily *Tt*GH5_4. Whilst TSA of the protein with different substrates indicated a positive shift in melting temperature for rAX, through TLC the activity was defined as specific only for tXyG. The degradation pattern of tXyG by *Tt*GH5_4 observed by TLC showed two discrete bands at a slightly higher molecular weight than the largest standard on the xylose oligosaccharide ladder, X6 (810 g/mol) as well as

smaller more indistinct bands. Due to the branching structure of tXyG, this most likely does not indicate a product with a backbone length larger than that of X6. *Tt*GH5_4 family members active on xyloglucan are thought to cleave primarily between the unbranched glycosyl units. Cleavage at unbranched glycosyl units creates products with variable side chains; 19 possible side chain structures have been identified, albeit not all at once but on a variety of different xyloglucans from different plant species.[146] Whilst there is a good variety of possible products, a set of four xyloglucan oligosaccharides are canonically observed in the degradation of tXyG by other members of the GH5_4 subfamily; XXXG, XLXG, XXLG and XLLG, whereby G equates to Glc*p*, X to [Xyl*p*(_1,6)]Glc*p*, and L to [Gal*p*(_1,2)Xyl*p*(_1,6)]Glc*p*.[138, 142] These canonical GH5_4 degradation products are made up of 7, 8, 8 and 9 sugar units, respectively. The TLC indicated two major bands were a higher molecular weight than was observed for the X6 standard that contains 6 sugar units. Therefore, it could be suggested that XXXG as well as XLXG/XXLG are being formed as the major products. The *Tt*GH5_4-xyloglucan reaction was analysed by HPAEC-PAD but due to lack of available standards at the time, the individual product components were not characterised. The HPAEC-PAD spectrum for the breakdown of tXyG by *Tt*GH5_4 was a complex picture, possibly indicating the complex nature of the polysaccharide and the variety of products produced by the enzyme. On the other hand, comparing the observed *Tt*GH5_4 HPAEC-PAD data with similar enzymes in the literature may suggest that during this analysis there was incomplete hydrolysis of the substrate. Whilst it is very likely that there are products in the format of XXXG, XXLG, XLXG and XLLG (around a retention time of 10 minutes), the peaks at longer retention times indicate larger but soluble oligosaccharides; these could be in the form of repeat units of the common tXyG breakdown products that had not yet been cleaved by the enzyme – i.e. $[XXXG]_2$, $[XXXG]_4$, $[XXXG]_6$, $[XXXG]_8$. There could also be differences in chain extensions to which *Tt*GH5_4 is unable to accommodate, yielding a slightly different pattern of products to presented in the literature. To further elucidate these findings, the soluble reaction products were also analysed by MALDI-TOF. Four major peaks were assigned as XLLG (1409.5 m/z), XLXG/XXLG (1247.4 m/z), XXGGG/XLGG (1115.4 m/z), XXX (923.3 m/z), XX (659.1 m/z). Despite accurate molecular weights, there remains some ambiguity in the assignment due to several species having the same molecular weights. However, the major peaks fit well with the expected degradation pattern on tXyG by a GH5_4 enzyme; where cleavage occurs at the unbranched glucose residue at position *G.* Preliminary kinetics analysis using the DNSA reducing sugar assay was carried out on this enzyme, albeit at a temperature that is slightly above the protein melting

temperature. Despite this, the enzyme proved efficient at degrading tXyG with a k_{cat}/k_M of 7.71 x10^5 mg^{-1} ml Min^{-1}. It can therefore be expected that further kinetic analysis on this protein at a more suitable temperate would result in a much higher value of k_{cat}/k_M. Other GH5_4 enzymes show similar kinetic parameters, with examples being *Bo*GH5 (kcat 2.61 x10^4 min^{-1}, Km 0.82 mM) and *Pp*XG5 ($V_0/[E]_t$= 8700 min^{-1} at 0.5 mg ml^{-1} substrate) which, as seen in the analysis of *Tt*GH5_4, show enzyme specificity for xyloglucan only. [147-148]

3.7.3 *Tt*GH5-un

As *Tt*GH5_un remains unassigned to a GH5 subfamily, there are many possible functional roles for this enzyme. Sequence analysis of *Tt*GH5_un using NCBI BLAST did not lead to any definitive possibilities in enzyme activity due to the very low sequence identity the protein shares with other members of GH5. Purification of *Tt*GH5_un was extremely successful in terms of yield, but cleavage of the protein had occurred during its preparation and as such, a loss of 64 amino acids was observed by ESI MS. Despite this, the protein was found to be stable, with a melting temperature of 65.0 °C. A broad range of substrates were initially tested by TSA. Large shifts in protein melting temperature were observed upon mixing with solid kGM and bMLG, although little increase was observed when the protein was mixed with a 1 mg mL^{-1} suspension of kGM. The latter result was most likely due to the insoluble nature of the substrate, resulting in little substrate reaching the test sample after pipette transfer. Test reactions were set up for *Tt*GH5_un on both substrates, as well as inM and GM and products visualised using TLC. TLC showed no observable bands for samples from the inM or GM reactions, an observation that was backed up by MALDI-TOF analysis; no oligosaccharide peaks were observed in the soluble fractions taken for mass spectrometry analysis and the samples matched with the soluble fractions of the no-enzyme control reactions. *Tt*GH5_un hydrolysis of kGM was evident on the TLC plate, with a large number of products separating poorly on the plate. The product profile was also observed as a range of oligosaccharide products of PD 4-11 in MALDI-TOF analysis. The enzyme, whilst unable to cleave inM was able to act on kGM, indicating that the enzyme works on the cleavage of glycosyl and not manosyl residues. Indeed, PACE analysis suggested that degradation of kGM produced products with glucose at the reducing end, and not mannose. TLC saw discrete bands for activity of *Tt*GH5_un on bMLG, with the predominant product being a PD of 4. MALDI-TOF analysis of the breakdown of bMLG also observed a discrete degradation pattern, where beginning at PD 4, predominant product peaks were 3 sugar units apart. The ratio of products could suggest that the enzyme works in a processive manner and requires

movement of 3 units along the chain before another cleavage event takes place. However, this theory does not consider the mixed linkage nature of the substrate. The mixed linkage glucan is typical formed of repeating 3-5 units of β 1-4 linked glucose, separated by a single β 1-3 linked moiety.[149-150] The inability of the enzyme to produce oligomers at certain degrees of polymerisation may reflect the repeating nature of the β (1-4), β (1-3) bonding pattern. During kinetics analysis using a fluorescent assay with 4-MU-C2, 4-MU-C3 and 4-MU-B1/3-C2, *Tt*GH5_un was found to cleave only the compounds containing β 1-4 linked glucose. This, coupled with the mass spectrometry profile of *Tt*GH5_un on mixed linkage glucan suggests that the enzyme is not able to accommodate β 1-3 linked moieties within the active site; as such the enzyme is active only on sections of the substrate that do not containing β 1-3, which resulted in a strong preference for oligosaccharides of DP 3 units apart. Drawing on the kinetic analysis, TtGH5_un, like that observed for TtGH5_2 displayed Michaelis-Menten kinetics on the cellotriose based fluorescent substrate, and not on the cellobiose substrate. It can therefore be suggested, based on the values of V_{max} and $K_{M,}$ 0.321 ± 0.0124 μM s^{-1} and 104.61 ±14.18 μM respectively that the enzyme is more proficient at interacting with a substrate of three units or larger.

*Tt*GH5_un was successfully crystallised in apo form and a data set obtained at the Diamond synchrotron. The data set obtained was of good quality, resolution at 1.8 Å, however, molecular replacement consistently failed during structural analysis due to the low sequence similarity of closest related structures. The majority of structures in the NCBI BLAST PDB search produced results of less than 30% identity over an incomplete range of the sequence. As such, molecular replacement pipelines failed to produce a structural model. Heavy atom soaking strategies were employed; crystals were dip-soaked in a 0.5 M KBr solution (made up in mother liquor) and data sets analysed using Br as a heavy atom for SAD phasing. This technique as well as using the natural sulphur atoms contained in the structure did not have enough anomalous signal of the heavy atoms to produce a model. One train of thought was related to the issue in sequence length of the purified protein. It was noted in **Chapter** 2 that *Tt*GH5_un was not at the correct molecular weight when the protein was analysed by ESI mass spectrometry. It was suggested that cleavage of the protein had occurred at some point during purification (this was noted in two separate purifications) with a loss of 64 amino acids. Despite this the protein was still active, as evidenced by the previous discussion. In an attempt to alleviate any issues caused by inputting a protein sequence that was too long, the actual molecular weight was used instead. This method of molecular replacement still failed to produce a model.

3.8 Conclusion

This section of work has led to the substrate specific characterisation of three GH5 proteins from *T. turnerae.* Two proteins, *Tt*GH5_2 and *Tt*GH5_un were determined to cleave glycosyl moieties within insoluble cellulose-based polysaccharides and soluble mixed linkage glucans. The third protein, *Tt*GH5_4 was active on complex xyloglucan and able to navigate the highly branched structure, cleaving at unbranched glucose residues. Activity was determined using various analytical techniques to enable assignment of hydrolysis products. *Tt*GH5_un is a completely novel protein, unassigned to a GH5 subfamily, shown here to favour cellulose based substrates. This novel characterisation will hopefully pave the way for characterisation of new related GH5_un sequences once they become available from genomics.

4

Characterisation of *Tt*GH12

4.1 Abstract

The ability of the shipworm to use a variety of wooden substrates as a food source has been explored by focusing on characterisation of novel lignocellulosic enzymes provided by the animal's endosymbiotic bacteria, *T. turnerae*. Despite localisation of the bacteria within specialised gills cells, it is thought that the animals are able to use lignocellulosic bacterial enzymes within their own digestive system, which is otherwise void of any major microbial communities normally involved in polysaccharide degradation. This work features an enzyme belonging to the glycoside hydrolase 12 family, members of which are known to breakdown cellulose and xyloglucans. The structure of *Tt*GH12 (ACR14297.1) was solved at 1.6 Å using molecular replacement, and shows the classic 'jelly-roll' fold and long substrate binding grove found in other GH12 structures. A protein-inhibitor complex was also solved at 1.9 Å, where the compound, Glc β1,4 noeuromycin was found to bind in the -2 to -1 subsites, with the glucose moiety ring stacking with nearby tryptophan residues, two of which are rotated compared to other GH12 enzymes. *Tt*GH12 was found to lack the typical aspartate residue thought to form a catalytic triad in the centre of the active site and instead exhibits a threonine in this position. An arginine residue in close proximity to the -1 subsite was found to provide hydrogen bonding interactions with the Glc β1,4 noeuromycin and may be involved in substrate binding or catalysis. Sequence and phylogenetic analysis suggest *Tt*GH12 contains an alteration in the normally conserved catalytic triad, and sits with two other sequences as a single branch within the GH12 phylogentic tree. Thermal shift assays showed *Tt*GH12 to have a biphasic thermal profile, with T_M at 48 and 61 °C. Interaction with substrates xylan, mixed linkage glucan, xyloglucan and mixed linkage xylan produced only single melting curves at 62, 66, 54 and 62 °C respectively. *Tt*GH12 was found

to have some background activity on xylan and mixed linkage glucan over a long reaction time but no typical GH12 cellulase activity was observed. Based on lack of activity on substrates typically acted on by GH12 enzymes and by significant changes within the active site of the protein, *Tt*GH12 remains ambiguous but likely has a new, as yet unknown, substrate range within the GH12 family. This work provides a platform for further analysis of *Tt*GH12, which should elucidate the function and ultimately its place among other GH12 enzymes.

4.2 Introduction

4.2.1 The 12th Family

Glycoside hydrolase family 12 is one of the smaller families within the CAZy database, and as of 2018, is comprised of a mixture 828 bacterial and eukaryotic sequence entries. Of these, 69 sequences have been characterised, and enzyme function covers endoglucanases, xyloglucanases and β-1,3-1,4-glucanases.[34] The GH12 family, historically known as cellulose family H, was identified based on the similarity to known enzymes found within the GH11 family, which primarily contains xylanases. Despite this, no primary xylanase activity has been observed in the GH12 family. The GH11 and GH12 families are grouped together into clan GH-C; both of which share a similar protein fold and break β-1,4 glycosidic bonds through a retaining mechanism (for a discussion on hydrolytic cleavage mechanism, please refer to Chapter 1).[151]

Clan GH-C exhibits a β-jelly roll structure; whereby two twisted anti-parallel β-sheets, linked by flexible loops, stack together. Stacking of the sheets creates a 'sandwich' structure in which the aliphatic residues are exposed on the sheet outer surfaces, whilst hydrophobic residues are held as the 'filling' between the two stacked β-sheets.[151] A small α-helix is positioned packing against the outer face of the structure. Hydrophobic cluster analysis identified some major structural similarities between the two clans, an observation that was analysed crystallographically by Sulzenbacher *et al* from 1997-1990. [24, 151-152] Sulzenbacher looked at a GH12 protein, CelB2 from *Steptomyces lividans,* producing models which were extremely informative in providing a comparison between 'apo' and glycosyl-enzyme intermediate structures, as well as distinguishing the differences between GH12 and previously characterised GH11 enzymes. [24, 151] CelB2 consists of two separate but stacking β-sheets, one of which is made up of fewer individual β-strands than the other. The stacking of the two slightly differently sized sheets creates a curvature, whereby on the concave face,

a long substrate groove is observed. [151] The perimeter of the binding groove is defined by the loop regions, which pass through the two β-sheets. One inserted loop is significantly longer, and its positioning is thought to close off the reducing end of the substrate binding groove; a major similarity with GH11 enzymes, and commonly referred to as 'the cord'. A 12 residue long loop representing this 'cord' structure was shown to have well defined electron density within CelB, indicating a possible structural function. [151, 153] The cord has been suggested to undergo a conformation change upon substrate binding [153], but this has yet to be observed in GH12 or GH11 substrate complexes. The apo structure of CelB2 was conformation of the expected similarity of GH12 and GH11 proteins, whereby the core structure is common to both families, and differences are observed in loop regions. For example, it was noted that in CelB2, a loop near subsite +2 was longer compared with the corresponding loop in GH11 structures, allowing the binding groove of CelB2 at this position to be significantly deeper.[151] The shape of GH11 proteins has been previously described as analogous to a hand, where the β-sheets represent the 'fingers', the concave face as the 'palm' and a third region of β-sheets protruding in front of the concave face as the 'thumb'.[153] A major difference between GH12 and GH11 proteins was found to be that GH12 proteins lack the 'thumb' region; without the β-strand region crossing over the binding cleft, GH12 proteins can adopt a more open active site, **Figure** 82.[24]

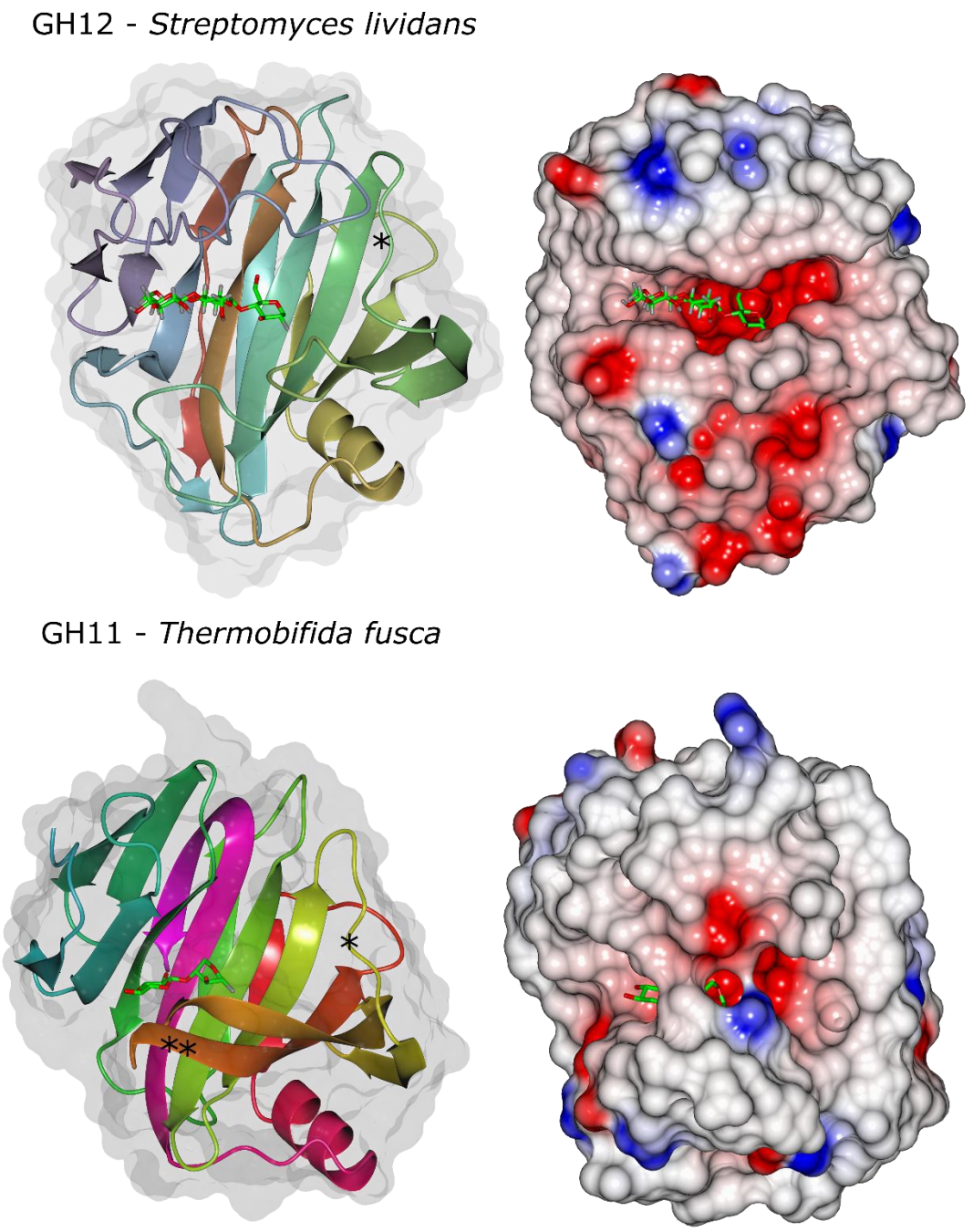


Figure 82 Structural comparison of a GH12 (CelB2, PDB; 2NLR) and GH11 (PDB; 3ZSE) protein , shown as ribbons and as surfaces. A single asterisk marks the 'cord' on both GH12 and GH11 ribbon structures, a structured loop conserved at the reducing end of the active site cleft in both GH families. A double asterisk on the GH11 ribbon structure highlights the 'thumb' region, which is not present in GH12 proteins, and protrudes around the front of the active site. The effect of this 'thumb' region can be easily seen on the surface views, where the GH11 active site is partially covered by this region, whilst the GH12 active site is completely open. Both proteins are shown with enzymatic trapping agents bound within the active sites. Images produced in CCP4mg.[55]

Catalytic residues are conserved in clan GH-C, and are known to be two glutamic acid residues, located closely to each other in order to facilitate the retaining mechanism of glycosidic bond cleavage. The similarity between GH11 and GH12 families meant that the catalytic residues could be predicted by sequence homology modelling. Experiments which were able to define the catalytic residues of GH11 xylanases firstly involved the use of enzymatic trapping agents,[154] (inhibitors displaying poor leaving groups, essentially trapping the covalent intermediate within the active site) the interpretations from which could be applied to CelB2, allowing the catalytic nucleophile to be automatically assigned. [151] Secondly, NMR analysis was able to identify the general acid/base residue in a GH11 xylanase from *Bacillus circulans,* as a second glutamic acid, by comparing the residue's pKa values during enzymatic titrations; changes in pKa in the general acid/base as it alters role during the reaction itinerary is described as pKa cycling.[19] Site directed mutagenesis was also employed to test whether altering the proposed catalytic glutamic acid residues would form inactive mutants of the endoglucanase Cel12A from *Trichoderma reesei*, whereby the authors reported a 98% loss of function upon mutation of both catalytic residues to glutamine.[155] Further use of enzymatic trapping agents, this time with a longer substrate allowed Sulzenbachen to map out the interactions occurring within the binding cleft. Crystallography of CelB in complex as an enzyme-intermediate and product complex was used to defined the hydrogen bonding network surrounding the bound ligands, confirming the occurrence of 5 subsites; -3 to +2.

Despite the conservation of catalytic machinery, the two families display different activities, with GH12 showing very limited activity on xylan. Sulzenbacher suggested this to be due to a slight change in the topology around the -1 subsite, where the directionality of a Tyr adjacent in space to the proton donor, forms a cavity which is likely to induce preference for the exocyclic hydroxy-methyl substituent of glucose moieties, over xylan which lacks this group.[151] Indeed, as mentioned above, the increased depth of the active site compared with GH11 proteins may drive substrate preference towards glucan based polysaccharides which are able to fill the cleft and form more stabilising interactions than a slightly simpler xylan backbone polysaccharide.

As mentioned previously, GH12 enzymes lack the 'thumb' region conserved in GH11 xylanases [24]. Lack of this cross over loop opens up the binding cleft to more substituted substrates and has resulted in some GH12 enzymes accommodating complex xyloglucans into their active site (see Chapter 3 for a discussion on xyloglucan structure). Indeed, studies

of a xyloglucanase XG12 from *Bacillus licheniformis* by Gloster *et al* highlighted the ability of the enzyme to tolerate substitutions on the glucan backbone in the -3,-2, +1 and +2 positions. This allows for activity to occur on both branched and unbranched glucose based polysaccharides.[148] As such, GH12 enzymes are one of the few families able to break down complex xyloglucans.

4.3 This Work

Teredinibacter turnerae exhibits only one GH12 sequence in its vast array of glycoside hydrolases and as such, the study of this enzyme, *Tt*GH12, was deemed an interesting target. The goal of the work presented herein was to establish the functionality of *Tt*GH12, assess its hydrolytic capability and use x-ray crystallography to investigate the 3D structure of the protein.

4.4 Methods

4.4.1 Materials

Substrates used in this study were tamarind xyloglucan (tXyG), mixed linkage barley beta glucan (bMLG), Avicel (Sigma), Phosphoric acid swollen cellulose (PASC) (Sigma), birchwood xylan (bX), mixed linkage β 1-3, β 1-4 xylan(MLX) (Elicityl/Oligotech). Crystallisation screens were obtained from Molecular Dimensions (PACT) and Hampton Research (PEG/ION and Crystal Screen HT). Cellohexaose (Megazyme) and Glc β 1,4 noeuromycin (a gift for the Davies Lab for a previous project) were obtained for crystal soaking experiments. Pure recombinant *Tt*GH12 (ACR14297.1) was prepared as described in Chapter 2.

4.4.2 Thermal shift analysis

The TSA assay was carried out using a Stratagene rtPCR machine, SYPRO orange dye (1000X stock, 15 µL) using 1 mg ml^{-1} final concentration of *Tt*GH12 protein, with small amounts of the solid polysaccharides mixed into the final sample. Samples were heated from 20°C to 91° in increments of 1 °C over 71 cycles. The fluorescence of SYPRO orange was monitored throughout and the data used to calculate the protein melting temperature. Curves were fitted using a free online tool developed by Paul Bond at the University of York and is available at; http://paulsbond.co.uk/jtsa.

4.4.3 Activity Assays

Reactions were carried out in 1.5 mL Eppendorf tubes, held horizontally during shaking to ensure thorough mixing of insoluble substrates during incubation, which was carried out at 37 and 45 °C. Enzyme concentrations were used at 1 mg mL^{-1} and substrates at 1 mg mL^{-1}. The products were analysed by TLC, as described in Chapter 3

4.4.4 Crystallisation

*Tt*GH12 (38 mg ml^{-1}) was tested for crystallisation hits using three commercial screens, PACT, Crystal Screen HT and Peg Ion HT. A 24 well hanging drop plate was set up using magnesium chloride hexahydrate (0.2 M) at pH 5 and pH 6 with the PEG 3350 concentration increasing from 15-25% w/v. Protein and buffer were mixed in a 1:1 ratio. Large long plate crystals grew overnight and showed a preference for pH 6 and 15% w/v PEG 3350. Crystals

were fished using cryo loops and crystals dipped in a drop containing 50% v/v mother liquor (magnesium chloride hexahydrate 0.2 M, pH 6, 15% w/v PEG 3350 and 25% v/v glycerol as a cryoprotectant before being flash frozen in liquid nitrogen. Ligand soaks with cellohexaose (1 mg ml^{-1}) glc β 1,4 noeuromycin and glucose β-1,3-isofagamine (powder soak) were also carried out on *Tt*GH12 crystals. The crystals were tested for diffraction in house, before being sent to the Diamond synchrotron and processed using the beamline I04.

4.4.5 Structure Solution

The diffraction data for *Tt*GH12 was collected to a resolution of 1.6 Å. The data was analysed using the CCP4I2 pipeline[156]. Data reduction was carried out using AIMLESS, with data cut to 1.9 Å. Molecular replacement was run using PHASER (Expert MR)[157]. The model chosen for molecular replacement was a truncated version (CHAINSAW) of 2JEM which has a 30% sequence identity with *Tt*GH12. The structure solution was used in BUCCANEER [158], where the model was only used to place and name chains, before iterations of seed chain growth were carried out. Scripted model building in COOT [159] was used to automatically build side chains into the remaining density, before several iterations of REFMAC [160] and manual model building in COOT were carried out. Data sets for crystals soaked in ligands were treated in the same manor except the model used for molecular replacement was the 'apo' *Tt*GH12 structure solution. The active sites were checked for extra electron density that could be attributed to a ligand molecule using COOT. Glc β1,4 noeuromycin was built into one model using JLigand [161] before being subjected to refinement cycles in REFMAC.

4.5 Results

4.5.1 TSA

The melting temperature of apo *Tt*GH12 was analysed by TSA assay and the protein was found to exhibit two distinct melting curves. The first melting event was at approximately 48 °C with the second event occurring at approximately 60 °C. To identify any possible polysaccharide interactions, TSA assays were conducted on *Tt*GH12 in conditions containing protein incubated with Avicel, PASC, bX, bMLG, tXyG and MLX. Incubation with bX, MLX, bMLG and tXyG removed the biphasic curve seen for the apo protein sample, and only a single melting event was observed. Incubation with crystalline cellulose Avicel and PASC saw no change in the dual melting event with similar melting temperatures calculated.

Table 15 summarizes the observed melting temperatures under the different substrate conditions, whilst Figure 83 shows some of the melting curves.

Table 15 *Tt*GH12 protein melting temperatures as analysed using the TSA assay with different substrates. Some conditions yielded dual melting events, whilst others saw only one protein melting event (highlighted in bold).

Substrate	Average Melting Temperature (°C)
Protein only	48,61
Cellulose (Avicel)	49, 60.1
Cellulose (PASC)	49, 60
Barley β glucan (bMLG)	**66.0**
Birchwood Xylan (bX)	**61.8**
Xyloglucan (tXyG)	**53.5**
Mixed linkage xylan (MLX)	**62.0**

The biphasic melting temperature of *Tt*GH12 was investigated further by first heating 10 individual protein samples at temperatures ranging between 41 and 66 °C for 2 hours, before analysing the samples by TSA. Samples heated at 41 °C showed two melting curves, albeit with only a small first melting curve. Whereas 41.5 and 42.7 °C exhibited only one melting curve, with a T_M of 55 and 59 °C respectively. All protein samples heated above 42.7 °C exhibited no melting curves indicating that protein degradation had already occurred.

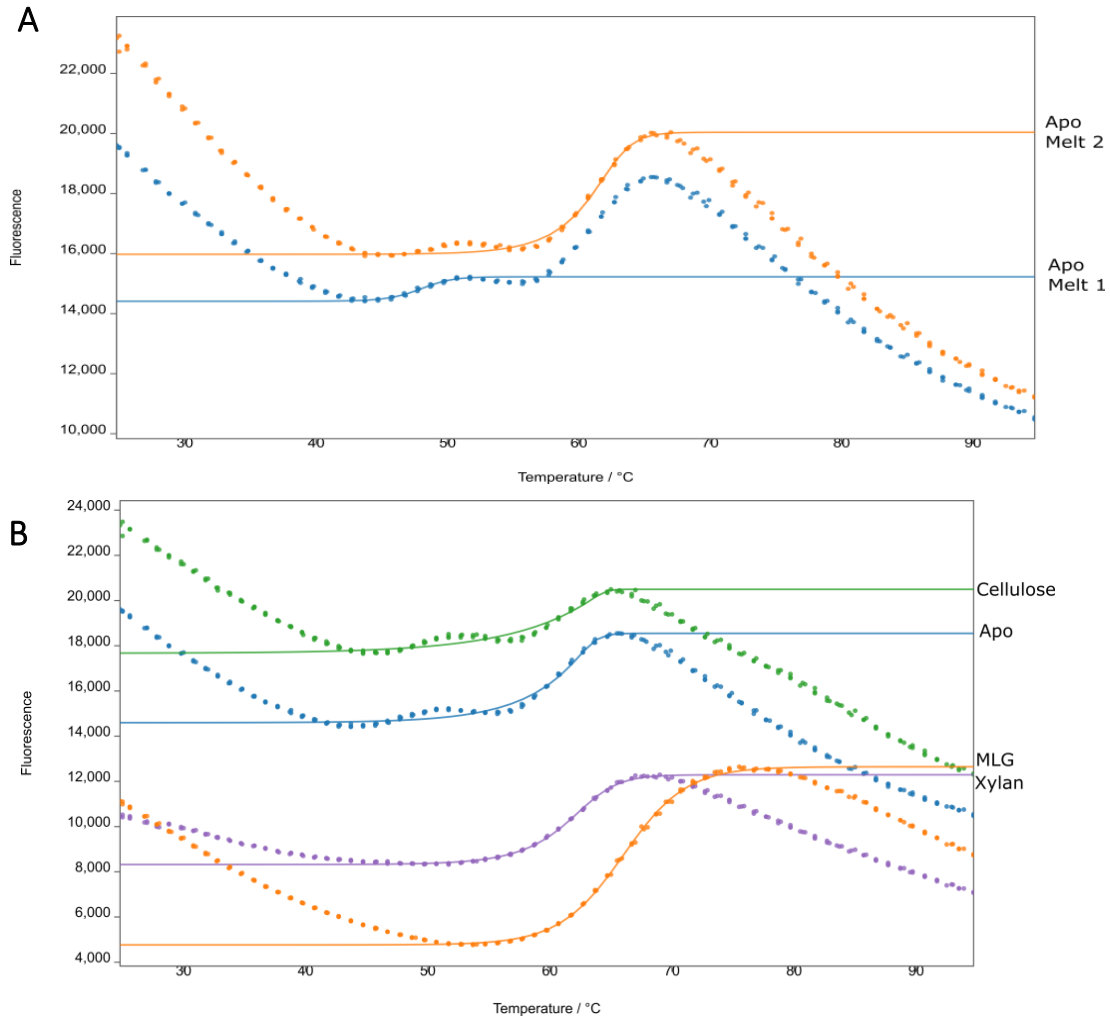


Figure 83 TSA of *Tt*GH12 . A) *Tt*GH12 apo melting curve, two runs displays which showed the two melting events and the subsequently fitted sigmoidal curves. B) *Tt*GH12 apo melting curve compared with *Tt*GH12 with cellulose (PASC), mixed linkage glucan (bMLG) and birchwood xlyan (bX).

4.5.2 TLC Analysis

*Tt*GH12 was incubated at 37°C overnight, at pH 5, 6 and 8, with tXyG, bX, bMLG and MLX. The soluble fractions of the reactions were analysed by TLC. No activity was observed for tXyG whilst very little activity was observed with bX (some smearing). Discrete bands were seen for reactions of *Tt*GH12 with bMLG and MLX, although the amount of smearing and substrate still at the initial sample spot position on the plate indicates incomplete degradation of the substrate. Analysis of the TLC plate suggested the smallest products were equal to or larger than 3 glucose units, by comparison with the xylooligosaccharide standards. There was a preference for lower pH values, with more activity seen in tests run at pH 5. The experiment was repeated at higher temperature; incubation overnight at 45 °C significantly reduced the amount of activity observed. Incubation at 50 °C destroyed all protein function observed at lower temperature, with a white precipitate being produced.

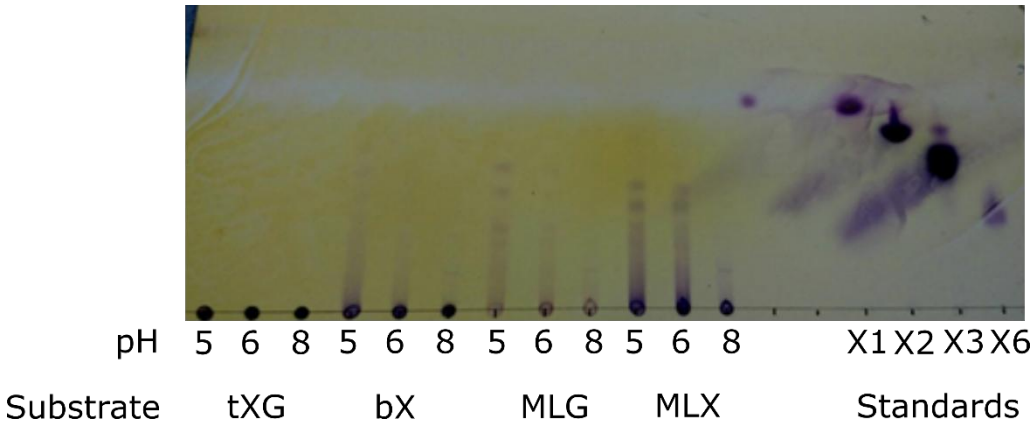
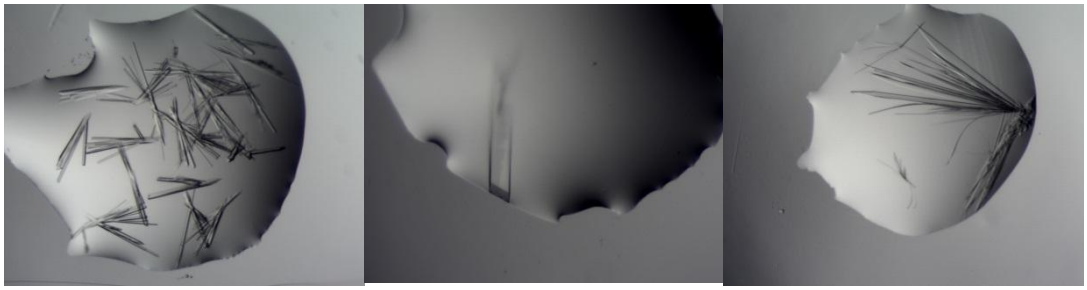


Figure 84 TLC plate showing the soluble reaction products from incubation of TtGH12, at three different pHs on tXyG, bX, MLG, MLX. Xylooligosaccharide standards, X1,X2, X3 and X6 are shown on the right.

4.5.3 Structure Solution of *Tt*GH12

4.5.3.1 Crystallisation

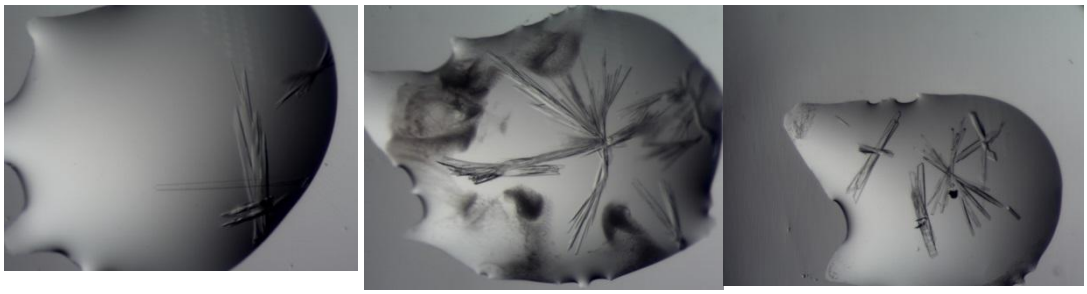
*Tt*GH12 was tested for crystallisation hits using three commercial screens, PACT, Crystal Screen HT and Peg Ion HT. *Tt*GH12 was quick to crystallise in various different crystallisation conditions, mostly forming long thin plates as shown in Figure 85. Condition A in Figure 85 was chosen for optimisation based on the size and separation of individual crystal clusters. A 24 well hanging drop plate was set up using magnesium chloride hexahydrate at pH 5 and pH 6 with the PEG 3350 concentration increasing from 15-25% w/v. Protein and buffer were mixed in a 1:1 ratio. Large long plate crystals grew overnight and showed a preference for pH 6 and 15% w/v PEG 3350.



A) PEG/Ion A5: 0.2 M Magnesium chloride hexahydrate, 20% w/v Polyethylene glycol 3,350, pH 5.9

B) PACT H11: 0.2 M Sodium citrate tribasic dehydrate, 0.1 M Bis-Tris propane, pH 8.5, 20 % w/v PEG 3350

C) PEG/Ion F1: 4% v/v Tacsimate, pH 6.0, 12% w/v Polyethylene glycol 3,350



D) PEG/Ion C9: 0.2 M Sodium sulfate decahydrate, 20% w/v Polyethylene glycol 3,350, pH 6.7

E) Crystal Screen HT B10: 0.2 M Sodium acetate trihydrate, 0.1 M TRIS hydrochloride pH 8.5, 30% w/v PEG 4,000

F) PACT C10: 0.2M Magnesium chloride hexahydrate, 0.1 M HEPES, pH 7.0, 20 % w/v PEG 6000

Figure 85 *Tt*GH12 crystallization screening produced hits in many different mother liquor compositions.

4.5.3.2 Data collection and structure solution

*Tt*GH12 crystals were initially tested in house to access the quality of x-ray diffraction. Figure 86 shows two images, how a single crystal is positioned inside small loops containing mother liquor and cryo-protectant during the diffraction experiment, and finally a diffraction image. *Tt*GH12 crystals were strongly diffracting and showed clear diffraction spots (Figure 86) during in-house analysis. Complete data sets used during structure solution of *Tt*GH12 were collected at the Diamond Light Source, using beamline I04. Crystals dip soaked in Glc β 1,4 noeuromycin had data sets collected separately.

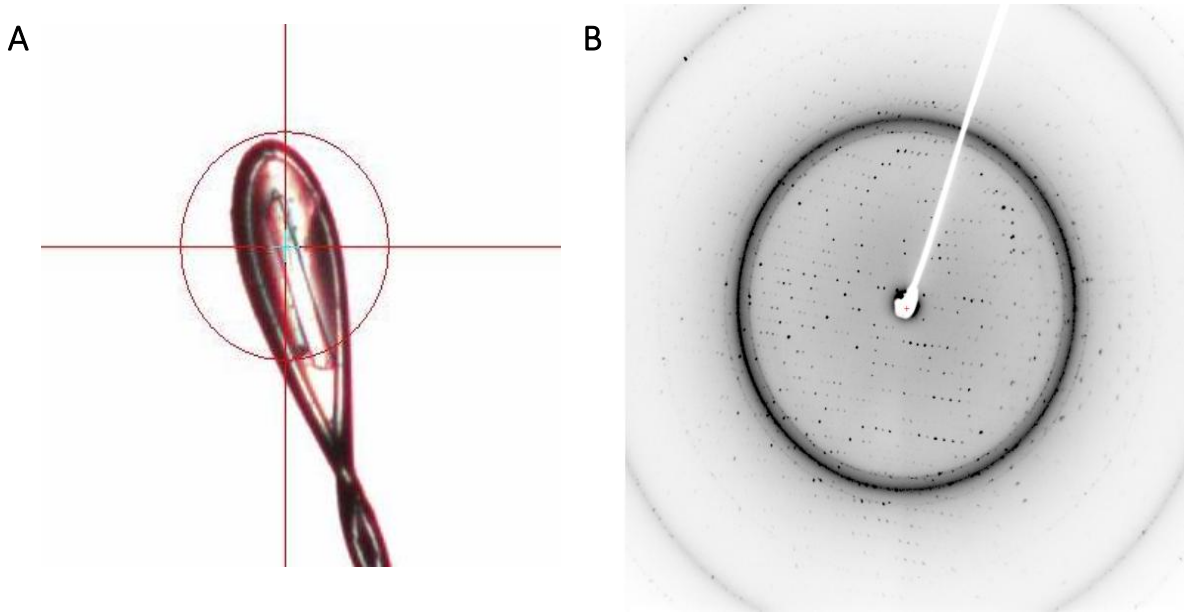


Figure 86 A) Crystal held in loop and photographed after in house testing. B) Test *Tt*GH12 crystal x-ray diffraction pattern, showing strong spots and an ice ring.

Table 16 Data collection and refinement statistics for *Tt*GH12 apo and Glc β 1,4 noeuromycin complex structures. Values given in brackets corresponds to the outer shell.

		*Tt*GH12 Apo	*Tt*GH12 Glc β 1,4 noeuromycin
Data collection	Space group	P 2_1 2 2_1	P 2 2_1 2_1
	a, *b*, *c* (Å)	51.5 81.2 139.1	52.43 72.99 141.09
	(°)	90.0 90.0 90.0	90.0 90.0 90.0
	Resolution (Å)	52.8 (1.6)	70.5 (1.9)
	R_{merge}	0.094 (1.46)	0.229 (2.65)
	R_{pim}	0.037 (0.576)	0.144 (1.20)
	CC(1/2)	0.995 (0.730)	0.995 (0.411)
	$I / \sigma(I)$	9.0 (1.4)	5.9 (1.0)
	Completeness (%)	99.9 (99.3)	99.9 (100.0)
	Redundancy	8.1 (8.3)	6.6 (6.8)
Refinement	Resolution (Å)	1.6	1.9
	No. reflections	77728	43507
	R_{work} / R_{free}	0.24/0.27	0.24/0.31
No. atoms	Protein	4021	3975
	Ligand/ion	30	56
	Water	194	299
B-factors (Å^2)	Protein	26.0	26.1
	Ligand/ion	39.9	36.2
	Water	33.0	32.6
R.m.s. deviations	Bond lengths (Å)	0.013	0.010
	Bond angles (°)	1.6	1.4
Ramachandran plot	Most favoured (%)	97.7	96.4
	Allowed (%)	1.8	3.4

The diffraction data for *Tt*GH12 was collected to a resolution of 1.6 Å. Subsequent data reduction using AIMLESS found the space group to be P 2_1 2_1 2_1 with a solution probability of 85 %. However, this was complicated by the occurrence of translational non-crystallographic symmetry within the unit cell (at a coordinate (0.5 0.5 0.458), chance probability 0.0177, height 0.435 of origin, as suggested during data reduction). Molecular replacement using the space group suggested by AIMLESS did not produce a structure solution, as indicated by r factors above 0.5. MOLREP was used to test the space groups, P 2 2 2, P 2 2 2_1, P 2_1 2 2 and P 2_1 2 2_1, which indicated the likely space group as P 2_1 2 2_1. A search model was found using NCBI BLAST (PDB search, [113]) by similarity of the *Tt*GH12 sequence to structures deposited in the PDB [113]. The model chosen was 2JEM, a GH12 protein from *Bacillus licheniformis* (*Bl*GH12), known to be active on xyloglucan, which shared a 30% sequence identity with *Tt*GH12. The search model was truncated using CHAINSAW based on sequence alignment (CLUSTALW) with *Tt*GH12. Molecular replacement was carried out in PHASER (Expert MR)[157], using the truncated search model and manual definition of the space group as P 2 2_1 2_1. A unique solution was found, with two molecules in the asymmetric unit, but with R_{cryst} and R_{free} as 0.49 and 0.52 respectively. The solution was used in BUCCANEER[158], where the model was only used to place and name chains, before iterations of seed chain growth were carried out. This significantly improved the model, with R_{factor} at 0.26 and a completeness of 100%. Scripted model building in COOT[159] was used to automatically build side chains into the remaining density, before several iterations of REFMAC[160, 162] and manual model building in COOT were carried out. The model was found to be acceptable, with final R_{factor} and R_{free} as 0.22/0.25 respectively. Aspects of the structure were validated using the various tools within the COOT manual interface. The structure of *Tt*GH12 is shown in **Figure** 87, created using CCP4mg [55], as front and side-on views, with and without a surface (coloured by electrostatic potential).

*Tt*GH12 crystals were soaked with three ligands; cellohexaose and two glyosidase inhibitors, Glc β 1,4 noeuromycin and glucose β-1,3-isofagamine. The latter two soaking experiments were carried out by powder diffusion into the crystal drop. Data were collected at the Diamond Light source, and processed in a similar manner to that of the apo structure. However, molecular replacement was carried out using the solved apo structure of *Tt*GH12 instead of the original search model 2JEM, and free R sets maintained. Only crystals soaked in Glc β 1,4 noeuromycin displayed extra density within the active site, **Figure** 88. The Glc β 1,4 noeuromycin ligand was manually incorporated into the model using the COOT extension JLigand[161], and refined with REFMAC. The final model for *Tt*GH12 in complex with

Glc β 1,4 noeuromycin displayed the statistics R_{crst} and R_{free} as 0.24 and 0.31 respectively. The ligand is observed binding in the centre of the active site cleft, and has a likely defined set of interactions from nearest neighbour residues (CCP4mg[55] was used to show residues/water molecules up to 4 Å away from ligand atoms) as shown in **Figure** 89.

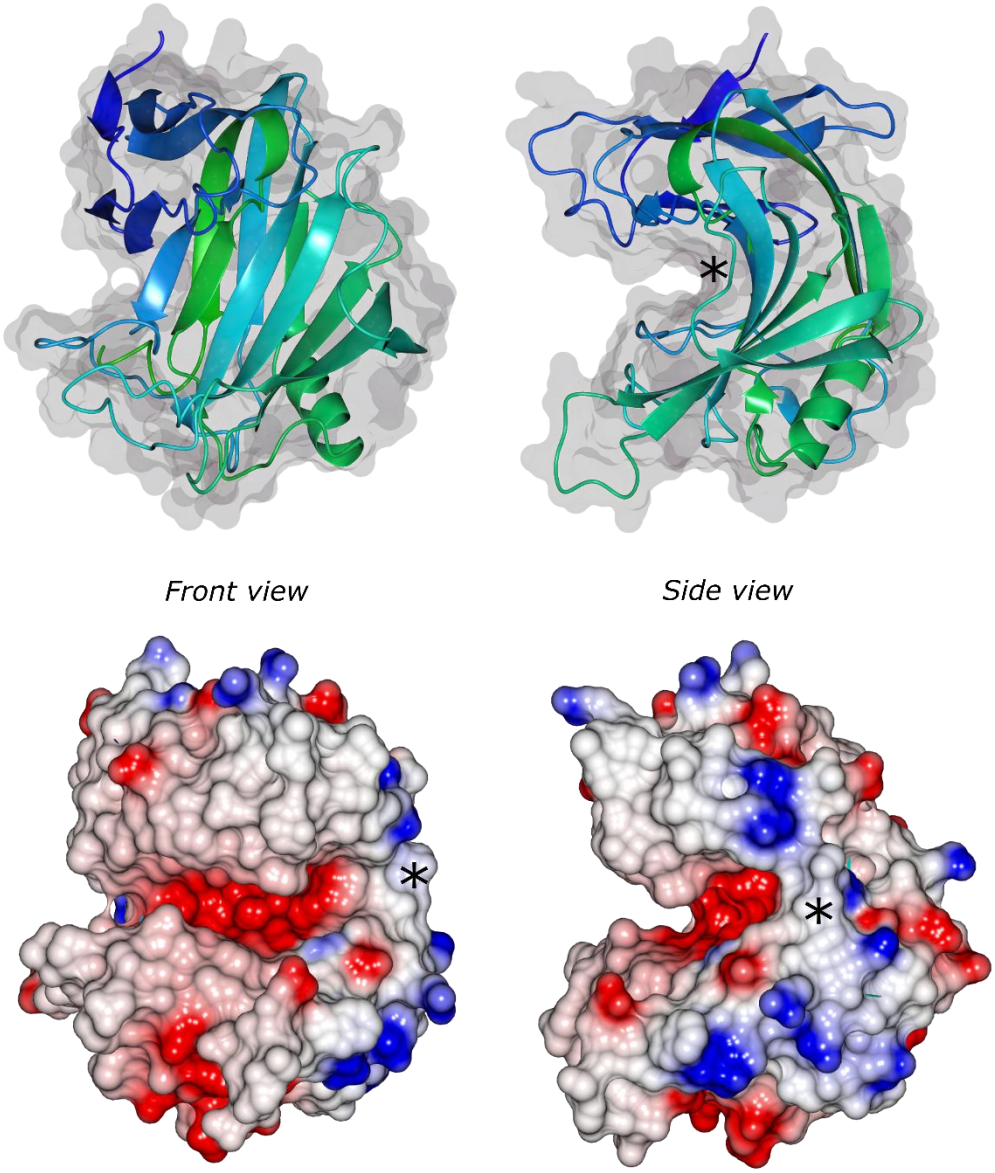


Figure 87 *Tt*GH12 apo structure, shown as different perspectives , 'front' and 'side' (coloured dark blue to light green, N-C terminus), with and without surface (coloured by electrostatic potential). An asterisk highlights the 'cord', a structured loop which crosses the reducing end of the active site. Images produced in CCP4mg[55].

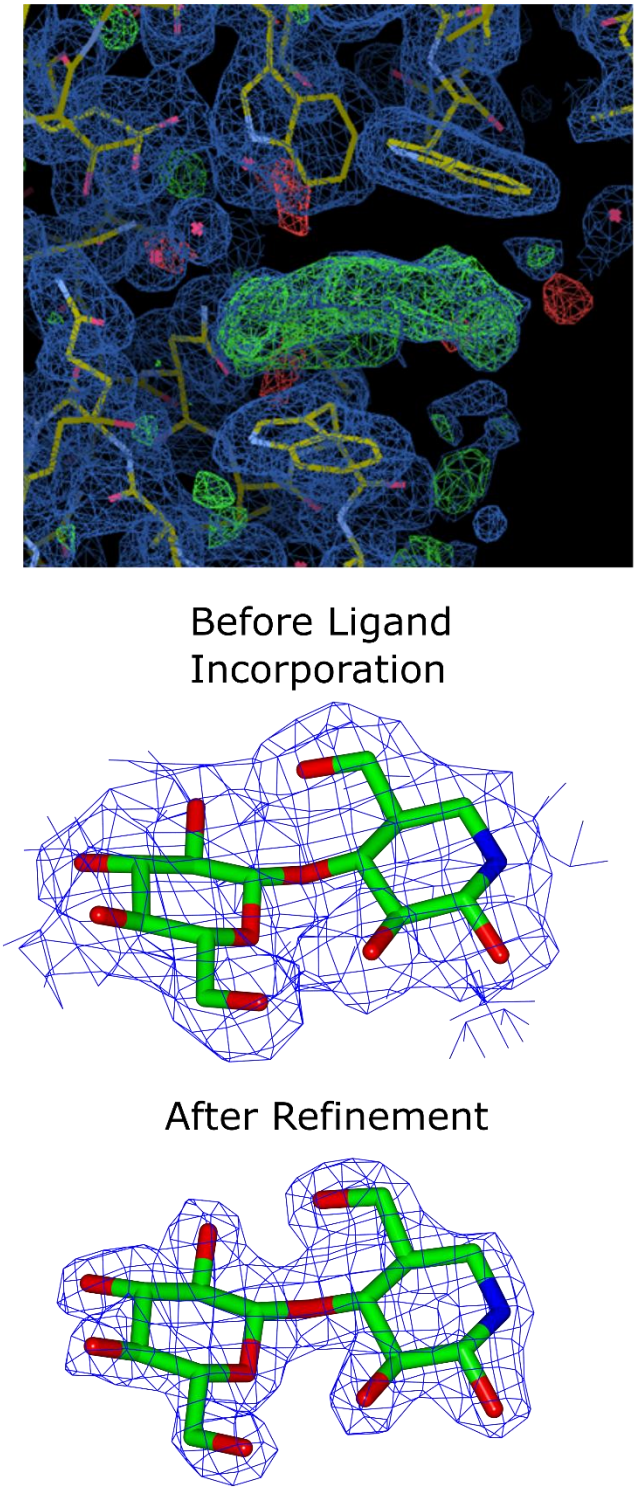


Figure 88 Overview of ligand incorportation into *Tt*GH12 structure Above; Positive electron density within the binding cleft of *Tt*GH12 before the Glc β 1,4 noeuromycin ligand was manually fitted and refined using COOT[159] and REFMAC[160], image taken in COOT. Middle; Electron density around ligand before refinement. Below; electron density of ligand after refinement (REFMAC[160]). Latter two images produced in CCP4mg[55].

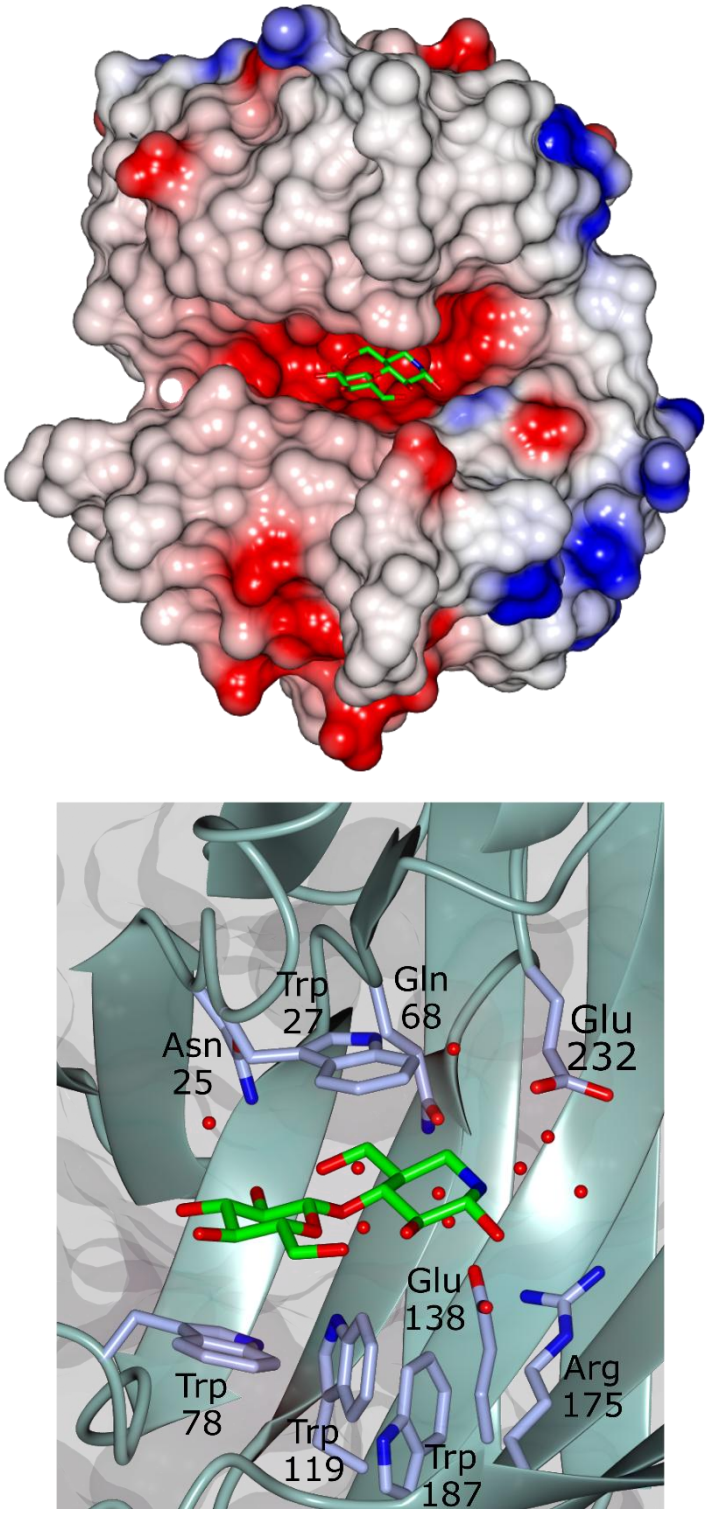


Figure 89 *Tt*GH12 in complex with Glc β 1,4 noeuromycin , shown above bound centrally in the active site, and below surrounded by nearest neighbours (with 4 Å). The series of four conserved Trp residues surround the ligand, with proposed catalytic residues Glu232 and Glu138 positioned on the right. Images produced in CCP4mg[55]

4.5.4 Comparison of *Tt*GH12 Structure with other Family Members

The structure of *Tt*GH12 was compared with several other GH12-complex structures in the PDB in an attempt to discern more clues about the possible activity of the enzyme. *Tt*GH12 was found to be display slight activity on mixed linkage glucan and mixed linkage xylan, but not on cellulose or xyloglucan. The *Tt*GH12-Glc β 1,4 noeuromycin inhibitor complex has provided structural information regarding the -2 to -1 subsites within the active site. Comparison of the active site of *Tt*GH12 with that of the xyloglucanase structure *Bl*GH12 used for molecular replacement and structure solution of *Tt*GH12 shows some significant differences. This may indicate why *Tt*GH12 does not display activity on xyloglucan. Figure 90 shows a comparison of the positions of the tryptophan residues (cylinders) in both *Tt*GH12 and *Bl*GH12 proteins, with ligands bound (thin bonds). There is significant movement around some of the conserved tryptophan residues, specifically Trp78 and Trp187 which are shown to be rotated in *Tt*GH12 compared with the equivalent residues in *Bl*GH12.

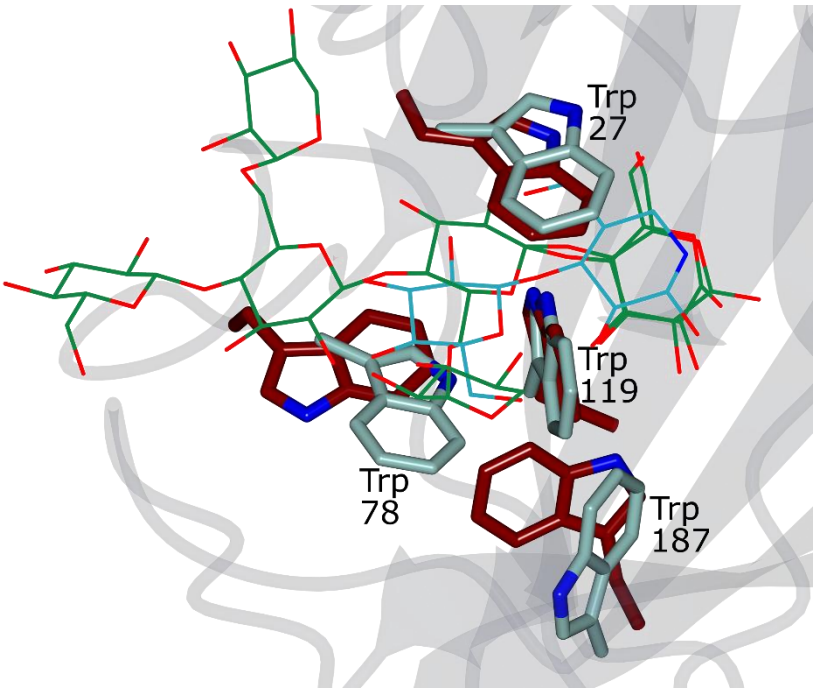


Figure 90 Comparison of the positioning of the four GH12 conserved tryptophan residues in *Tt*GH12 and *Bl*GH12. *Tt*GH12 residues are labelled and Glc β1,4 noeuromycin is shown in blue. The xyloglucan ligand bound in the *Bl*GH12 structure is shown in green.

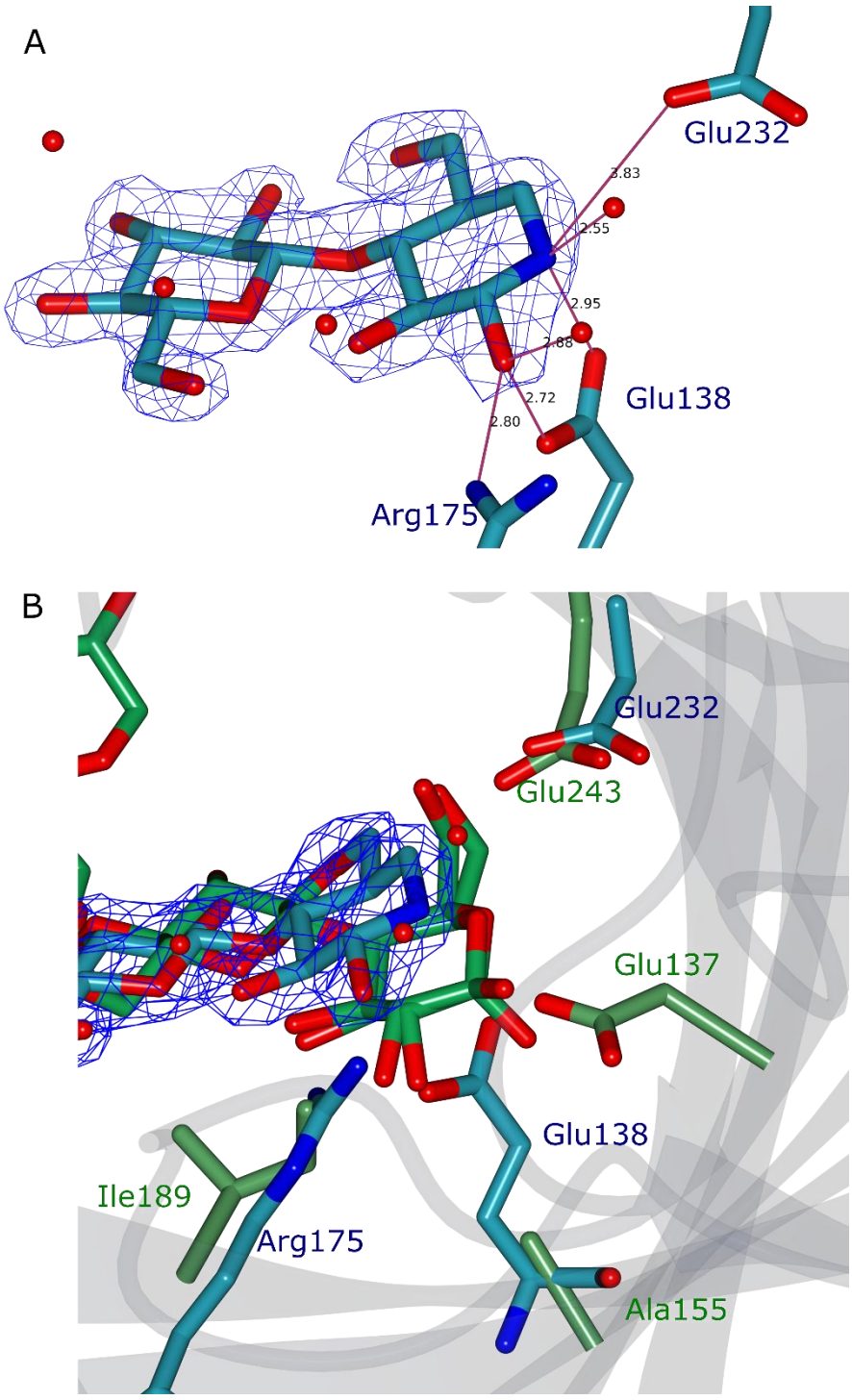


Figure 91 A) Close up view of Glc β1,4 noeuromycin bound in the active site of *Tt*GH12 , showing the distances between the two catalytic glutamate residues, a nearby arginine and two water molecules. B) Structure of *Tt*GH12-Glc β 1,4 noeuromycin overlaid with *Bl*GH12, a catalytically inactive mutant xlyoglucanase with bound substrate (tan, PDB;2JEM). Extra hydrogen bonding to the Glc β 1,4 noeuromycin is provided by a close arginine residue, which is not present in the overlaid xyloglucanase structure, where an isoleucine is found instead.

Figure 91 shows the binding environment of Glc β 1,4 noeuromycin in *Tt*GH12, where distances to the nearby residues are shown. These distances are likely to represent the hydrogen bonding network, where the bound inhibitor is able to interact with the nearby Arg 175 and Glu 138 (likely the catalytic acid/base) as well as several water molecules. Comparison of the catalytic residues of *Tt*GH12 with *Bl*GH12 shows that the catalytic acid/base residues are in equivalent positions. The structure of BlGH12 in this instance is a catalytic mutant, E155A, where the alanine residue can be seen in a similar position to the catalytic nucleophile of *Tt*GH12, **Figure** 91. In *Bl*GH12 the xyloglucan ligand appears to bind in a slightly 'deeper' position than glc β 1,4 noeuromycine in *Tt*GH12, which may be a result of the catalytic mutation to an inactive alanine residue providing more space for binding. A significant difference is in the presence of Arg 175 in TtGH12 instead of Ile 189 in *Bl*GH12. The change in residue type from Ile to Arg provides the bound inhibitor with more hydrogen bonding interactions.

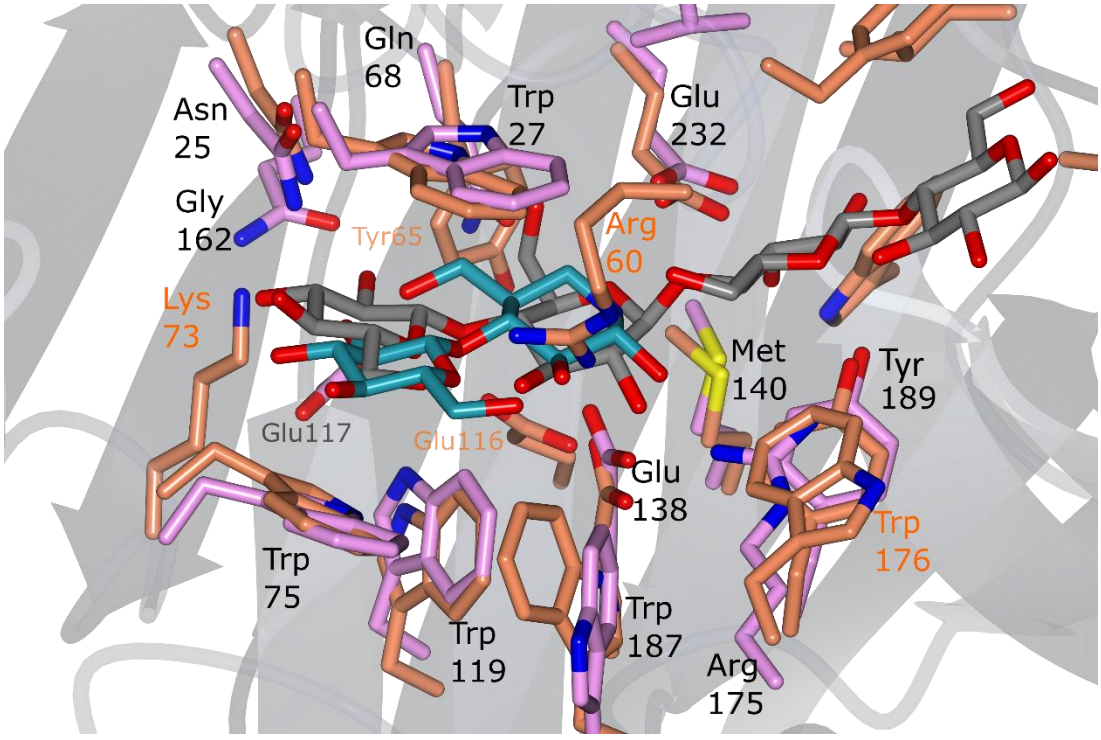


Figure 92 Active site comparison of *Tt*GH12 (pink) and *Tm*GH12 (corel, PDB; 3AMM). Residues in *Tm*GH12 that are different to those in *Tt*GH12 are labelled in corel, whereas all other black labels are residues in *Tt*GH12. Only one tryptophan residue is rotated, but there are some other differences, including Arg60 which appears to bend over the bound glucan ligand, as well as replacement of Arg175 with another Trp residue.

Comparison of *Tt*GH12 with a cellulose active GH12 enzyme, *Tm*GH12 (*Thermotoga maritima,* PDB*;* 3AMM), produced a similar alignment. Figure 92 shows that both catalytic glutamate residues occupy the same position within the active site, with the catalytic acid/base sitting below the -1 subsite. There are some significant differences between the two active sites, with a series of different residues around the peripheries of the binding cleft in *Tm*GH12, such as addition of Lys73 and a lack of Arg175 and subsequent replacement with another residue unable to provide hydrogen bonding interactions to ligands, Trp176. Movements in residue positions are also observed, Glu116/Glu117 of *Tm*GH12 and *Tt*GH12 respectively are in altered locations and as observed previously Typ187 of *Tt*GH12 is rotated compared with the equivalent residue in *Tm*GH12. A major difference in the active site of *Tm*GH12 is the positioning of an unusual loop (Tyr61-Arg60) which is shown to cross over the bound cellooligosacchaide, forming a pocket which may be responsible for helping maintain the linear path of the substrate through the active site. Figure 93 shows the surface view around the active site of *Tm*GH12 and highlights how the glycosyl unit of Glc β 1,4 noeuromycin would clash with the Arg60 residue if it undertook the same binding position in *Tt*GH12 as the cellooligosaccharide does in *Tm*GH12.

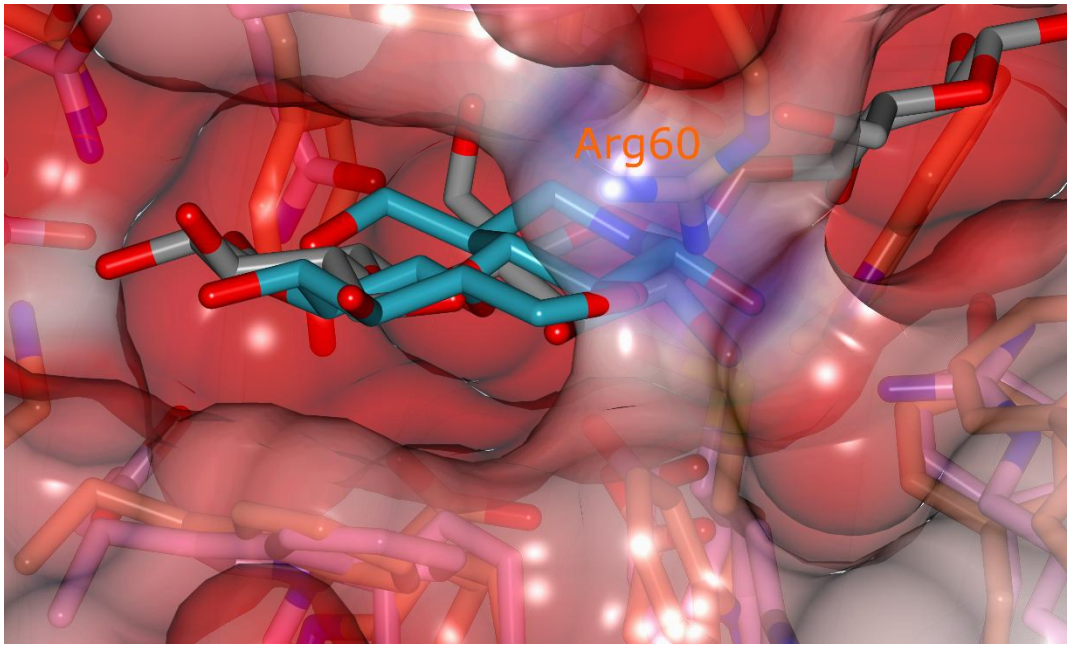


Figure 93 Surface of *Tm*GH12 (70% transparent, coloured by electrostatic potential) showing the Arg 60 creating a pocket behind which the cellooligosaccharide runs through the binding site. The Glc β 1,4 noeuromycin inhibitor as bound in *Tt*GH12 is shown overlaid, and would clash with the arginine of *Tm*GH12 if it followed the same binding configuration.

4.5.4.1 Sequence Analysis

The sequence of *Tt*GH12 was compared to other GH12 sequences (taken from CAZy) by sequence homology alignment using T-Coffee. The structure of the *Tt*GH12-Glc β 1,4 noeuromycin complex was compared with other structures with bound ligands to assist in identifying the catalytic residues. There were some noted residue differences in the comparison of *Tt*GH12 with two other GH12 as described above, so a broader sequence analysis was carried out to assess if there were any other significant changes that were not immediately apparent in the structural analysis. Interestingly, *Tt*GH12 displayed little similarity to eukaryotic sequences, of which there was little overall sequence alignment between any of the sequences. However, comparison with bacterial GH12 sequences produced an alignment that was relevant to further interpretation. As discussed previously, the catalytic and tryptophan residues are conserved in all the bacterial sequences and Figure 94 clearly shows this trend across the selected sequences. Several coloured asterisks mark residues in *Tt*GH12 which deviate from the seemingly invariant residues within the other proteins; for example, the pink marker shows that in *Tt*GH12 there is a phenylalanine adjacent to the catalytic nucleophile instead of the common Leu or Ile residue. The purple marker shows *Tt*GH12 has a serine instead of a highly conserved tyrosine. Finally, the grey marker shows *Tt*GH12 having a serine instead of phenylalanine. Surprisingly, when looking for conservation around Arg 175 (marked with a red asterisk), which sat close to the Glc β 1,4 noeuromycin ligand in the *Tt*GH12 ligand complex, all other sequences exhibit I or L residues.

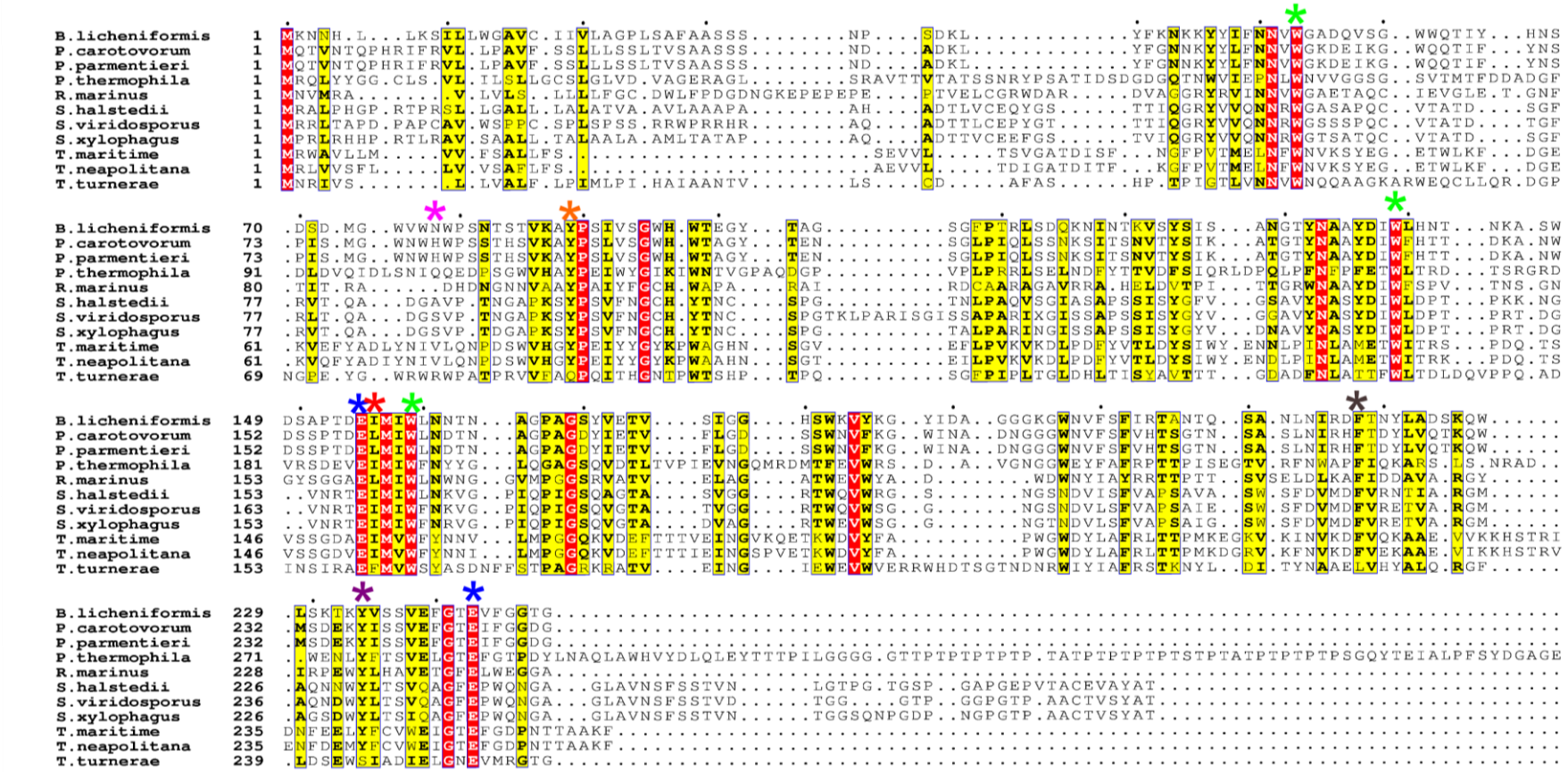


Figure 94 Sequence alignment (T-coffee) of *Tt*GH12, labelled as *T.turnerae* against GH12 proteins from several other bacteria. Note, the numbering of TtGH12 is slightly different in this alignment due to the use of the entire sequence as taken from the CAZy database, whereas previous numbering used in the text and structural analysis is based on the catalytic domain only. Stars; Blue) Catalytic residues, Green) Conserved tryptophans, Pink) Arg175, Red, Purple and Grey) variant residues against those strongly conserved.

During collaboration with the Brumer group at the University of British Columbia, Alexander Holm Viborg carried out an extensive sequence analysis with 762 full GH12 sequences from the CAZy database. This work highlighted that *Tt*GH12 is one of 16 entries that display a different residue at a position near the catalytic residues, an aspartate thought to act as part of a catalytic triad (an example of which is shown in **Figure** 91B). GH12 sequences normally have an aspartate residue at position 134 which is occasionally replaced with a glutamate. *Tt*GH12, along with two other sequences instead display a threonine residue at this position, as shown in **Figure** 95. Phylogenetic analysis carried out by Holm Viborg suggests that *Tt*GH12 and the other sequences with a Thr variant clearly represent their own branch within the GH12 phylogenetic tree.

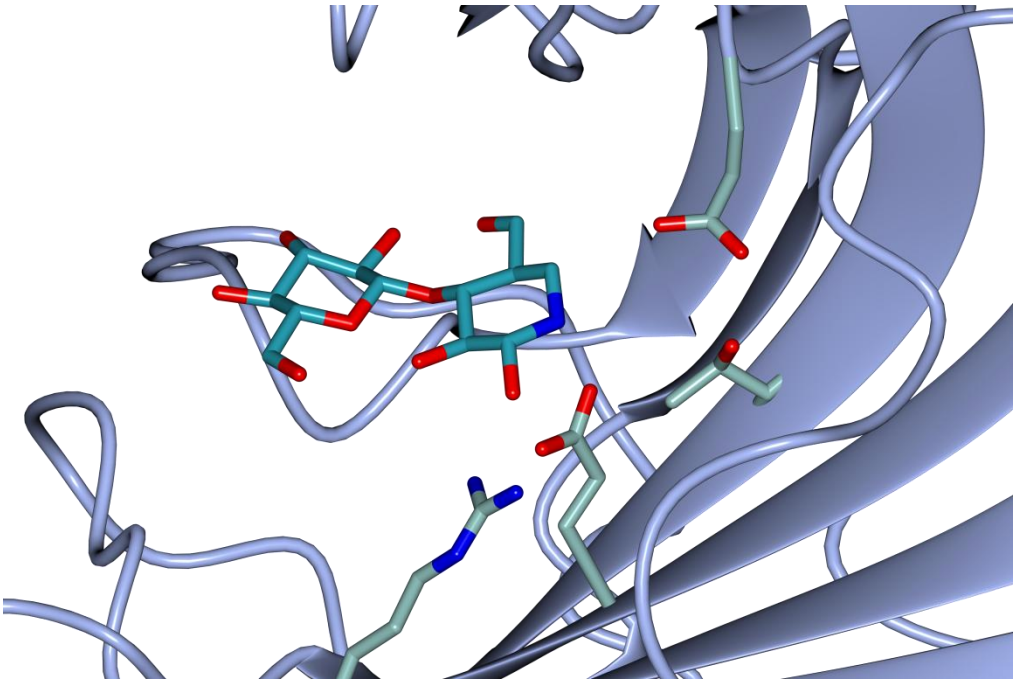


Figure 95 *Tt*GH12-Glc β1,4 noeuromycin complex showing the catalytic residues and Arg175 as previously discussed as well as a threonine residue between the two glutamate residues. This position is normally taken up by a third glutamate, or more rarely an aspartate and is thought to act as a catalytic triad.

4.6 Discussion

4.6.1 Stability and Activity

Within the GH12 family, enzymes are only known to exhibit function as cellulases, working on β(1-4) and mixed β(1-3)/(1-4) glucans, and as xyloglucanases. Looking at the organisation of characterised sequences within the CAZy database it is also clear that the majority of bacterial GH12 proteins are specific endoglucanases; only one sequence from *Bacillus licheniformis* is described as displaying xyloglucanase activity, and is noted alongside the endoglucanase annotation. Sequences grouped coming from eukaryotic sources contain a mixture of the three functions, with many proteins being described as xyloglucan specific. Hence, one might expect that another bacterial sequence this time coming from *T. turnerae* would follow this pattern and exhibit activity on cellulose based substrates. To test this hypothesis, TSA was used to evaluate the effect of certain substrates on the melting temperature of the protein. Firstly, it was noted that the apo protein possesses two melting curves, one at 49 °C and the second at 61 °C. A similar occurrence was described by Damasio *et al*, where a GH12 from *Aspergillus niveus* also exhibited a biphasic melting temperature. The authors suggest that the fungal GH12 undergoes an activation event upon increasing temperature, as the activity of the protein was found to be optimal at temperatures higher than the first T_M.[163] During TSA experiments where the *Tt*GH12 protein was incubated with different substrates it was found that introduction of substrates bMLG, bX and MLX removed the biphasic temperature profile; bX and MLX, both xylan substrates, induced a state in which only the higher protein melting temperature occurred. Incubation with bMLG removed the lower melting temperature and raised the overall T_M by 5 °C, suggesting formation of a stabilised protein-substrate complex. Furthermore, incubation of the protein with tXyG produced one melting temperature, but at a 7 °C lower T_M that that of the other substrate conditions, suggesting the substrate is able to destabilise the protein in some way. No change in thermal profile was observed on introduction of the crystalline cellulose substrates, PASC or Avicel. The TSA experiment indicated that the protein was strongly stabilised by introduction of a mixed linkage glucan, and possibly destabilised by interaction with a branched substrate such as tXyG. The fact that there was no interaction with PASC or Avicel suggests that the protein is incapable of activity against crystalline cellulose; the strong interaction with bMLG may be due to the nature of the substrate, in that it is amorphous, so is easily solubilised and as such accessible to the protein. PASC and Avicel are

crystalline forms of cellulose, where chains are tightly packed together in a regular configuration which severely limits solubility and can prevent GH activity.

The dual melting temperature of *Tt*GH12 was tested in controlled heating experiments and showed that a quantity of protein undertook a melting event above 41°C but the sample still maintained a large amount of intact protein which went on to melt at 61 °C in the TSA assay. In line with the suggestion posed by Damasio *et al*, it is likely that the second melting event of *Tt*GH12 is simply the protein denaturing, whilst the first event could be a temperature induced transition from a less stable form, possibly oligomer to monomer. The interactive power of the non-crystalline substrates with *Tt*GH12 may have disrupted the oligomeric form; substrates bX and MLX produced a melting temperature similar to that of the apo protein second melting event which suggests that the substrates may have induced a transition to the monomer form, but not provided any binding stability.

To further assess the possible function of *Tt*GH12, activity tests were carried out on tXyG, bX, bMLG and MLX at three different pH values and at different temperatures. The protein was shown to degrade bX and bMLG over an 18 hr timescale at 37 °C, but the reaction was incomplete, with insoluble substrate remaining at the bottom of the TLC plate. However, the degradation pattern of bX and MLG showed discrete bands and suggested products lower than C3. Despite the results of the TSA assay which indicated the protein was stable at higher temperatures, more activity was observed at 37 °C than at 45 °C or 50 °C. In the latter experiment it was obvious that the protein had precipitated out of solution during incubation. The stability index of *Tt*GH12 determined by Expasy ProtPara[164] suggested the protein to be unstable, and it may be the case that during the experiment the protein denatured naturally. It was also found that a lower pH increased the amount of degradation products observed on the TLC plate. Typically, when a GH is tested on the substrate it is specific for, one would expect complete degradation of substrate. As such, the slow rate of substrate degradation by *Tt*GH12 indicates that the tested substrates may not be appropriate for the enzyme. Furthermore, *Tt*GH12 was tested by Gregory Arnal at the University of British Columbia, whereby it was found to be inactive over a 2 hour timescale on a range of other substrates; glucomannan, laminarin, HE-cellulose, CM-cellulose, arabinan, curdlan, pustulan, galactomannan, cellooligosaccharides as well as sulfated carrageenan.

4.6.2 Structural Analysis

Crystal structures of *Tt*GH12 and a *Tt*GH12-Glc β 1,4 noeuromycin complex were solved at 1.6 Å and 1.9 Å respectively. *Tt*GH12 displays the typical structural fold seen in clan GH-C, with a curved β-sandwich structure upon which sits a large open active site cleft on the concave face of the protein. The 'cord' structure is well defined in density and the structure lacks the 'thumb' region typically seen in GH11 proteins. Sequence alignment with other GH12 proteins shows that there are a series of strongly conserved residues, specifically the catalytic glutamates and several tryptophans as has been discussed in the literature[148]. *Tt*GH12 was found to bind the inhibitor Glc β 1,4 noeuromycin within its active site, in the -2 to -1 subsites. Glc β 1,4 noeuromycin is similar in structure to isofagamine, where addition of an OH group on the carbon adjacent to the anomeric centre confers nano molar binding affinity to canonical cellulases; the compound is thought to mimic the transition state oxocarbenium ion formed during the inverting mechanism of glycosidic cleavage.[165] The binding of this particular inhibitor does pose an interesting question as GH12 proteins all perform retaining mechanisms during cleavage of the scissile bond. One possibility into the binding of the inhibitor is in the presence of Arg175, which is positioned pointing towards the -1 subsite and is within hydrogen bonding distance (2.8 Å) with the amide group of the Glc □□1,4 noeuromycin ligand. Structural overlays with *Tm*GH12 and *Bl*GH12 show non-hydrogen bonding groups Thr and Trp respectively at a similar position to Arg175. As such, sequence analysis was carried out to determine whether there was any conservation around this position, and it was found that there was no conservation of residues around this region of the sequence when the selected GH12 bacterial sequences were compared.

Structural comparison of the *Tt*GH12-Glc β 1,4 noeuromycin complex with other GH12 proteins identified some adaptations within the active site cleft. Firstly, significant rotational movement in Trp78 and Trp187 was observed compared with the equivalent residues in *Bl*GH12. The set of conserved tryptophan residues which line the active site of GH12 proteins are known to provide ring stacking interactions with the pyranose rings of the substrate chain. GH12 endoglucanases have a deeper cleft (through pockets created by other residues) than GH11 proteins which allows them to preferentially accommodate the CH_2OH groups at position 6 of glucose moieties. It may be possible therefore that the rotation of Trp78 and Trp187 helps define the substrate tolerance of the protein. The rotation of Trp78 and Trp187 may also explain why the glycosyl moiety of Glc β 1,4 noeuromycin had appeared to stick out of the binding site compared to other ligands in

similar GH12 structures, as ring stacking is still likely to occur between the residue and substrate rings.

The both catalytic residues were shown to adopt similar positions as seen in other GH12 structures active on both cellulose and xyloglucan substrates, with the catalytic acid/base occupying a position below the -1 subsite. Positioning of this residue defines the spatial positioning of the sugar moiety within the -1 subsite, pushing the observed Glc β 1,4 noeuromycin ligand out of the binding site slightly. TtGH12 was compared with the structure of *Bl*GH12, an inactive mutant in complex with a xyloglucan ligand. The catalytic acid/base of *Bl*GH12 was mutated to an alanine, which possibly allowed the xyloglucan substrate to move deeper into the active site, as Glc β 1,4 noeuromycin was found to sit further out of the active site. A major difference between the two structures, and to others was noted as the positioning of Arg 175 near the -1 subsite. Residues unlikely to provide hydrogen bonding interactions were observed in the same spatial position in other GH12 structures. Arg 175 was shown to be off a sufficient distance to hydrogen bond to Glc β 1,4 noeuromycin ligand, which could suggest a role in substrate binding or in catalysis.

Despite sharing a similar position of catalytic acid/base, the ligand structures of *Tt*GH12 and *Tm*GH12 show the ligands binding in different orientations. The unusual loop observed crossing the binding cleft of *Tm*GH12 was investigated by Cheng *et al*, whereby mutations of the two loop residues, Try61 and Arg60 were actually found to increase the activity of the protein. The authors suggest this to be due to tipping the balance between substrate binding and product leaving the active site; mutation of Try61, which they note occupies an unusual conformation, to a Gly improved the activity of the protein by 70%. The removal of the Try and its associated OH group removes a potential hydrogen bonding group, making it easier for product to dissociate from the binding cleft, whilst the natural configuration of the glycine is thought to stabilise the loop structure.[166] The loop structure forms a pocket into which the glucan chain binds behind and as such forces the chain to stick closely to the surface of the active site cleft. In *Tt*GH12 there is no such boundary and the Glc β 1,4 noeuromycin ligand is found to bind with the glucose moiety protruding slightly from the cleft.

4.6.3 Sequence Analysis

Sequence and phylogenetic analysis of TtGH12 carried out through collaboration with the Brumer group leads us to the suggestion that *Tt*GH12 forms a new clade within the GH12 family. Extensive sequence alignment highlighted the lack of a conserved aspartate residue, thought to act as part of a catalytic triad during glycosidic bond cleavage. This aspartate residue is occasionally substituted with a glutamate, but in *Tt*GH12 this was replaced by a threonine. A similar occurrence was found in a few other sequences, notably of marine origin, from *Microbulbifer* sp., *Alteromonadaceae*, *Paraglaciecola arctica* and *Aestuariibacter* aggregates, whereby each sequence contained the variant threonine residue. The top NCBI BLAST hits of *Tt*GH12 also contained this variation, although the highest sequence identity homologue comes from a species related to *T. turnerae*, *Alteromonadaceae* bacterium. It was also found that Arg175 was not conserved in the set of sequences, but a charged residue was conserved. Phylogenetic analysis sets these sequences aside within their own clade of the GH12 family. As such it is not unlikely to suggest that *Tt*GH12 is active on a completely different substrate, or functions in a different way to typical GH12 proteins.

4.7 Conclusion

Despite lack of any definitive activity on cellulose substrates, binding of a typical cellulose inhibitor, Glc β 1,4 noeuromycin occurred in the active site of *Tt*GH12 during crystallisation. It would be wise for future studies to analyse the binding affinity of this inhibitor for the protein to determine whether it was opportunistic during crystallisation or specific to the active site. This leads back to the questions posed earlier about the function and specificity of this protein. Structural analysis saw an arginine residue forming interactions with the bound inhibitor, which could indicate a possible role in catalysis. Furthermore, sequence and phylogenetic analysis suggests a distinct alteration in the aspartate residue normally perceived as playing a role in the catalytic triad required for glycosidic bond cleavage. This Thr residue, seen in the apo and inhibitor complex of *Tt*GH12 is suggested to have been replaced by Arg175. The phylogenetic tree shows *Tt*GH12 and three other proteins with the same alteration as their own clade within the family. This, combined with the lack of activity on typical GH12 substrates suggests that this may be a new type of GH12 functionality or allow degradation of a completely different substrate.

5

Characterisation of *Tt*GH8

5.1 Abstract

The marine shipworm symbiont, *Teredinibacter turnerae* has a broad array of carbohydrate active enzymes within its genome which cover a range of GH families. This work shows the characterisation of a glycoside hydrolase from family 8, hereafter *Tt*GH8, found to be an efficient endoxylanase, degrading β 1-4 linked xylan polysaccharides into shorter oligosaccharides, mostly xylotriose. *Tt*GH8 was most active on red algal marine (mixed β-1,3, β-1,4) xylan. Kinetic parameters, obtained using high-performance anion-exchange chromatography with pulsed amperometric detection (HPAEC-PAD) and DNSA reducing sugar assays shows that TtGH8 catalyses the hydrolysis of β-1,4 xylohexaose with a k_{cat}/K_m of 7.5 x 10^7 M^{-1} min^{-1} but displays maximal activity against mixed-linkage polymeric xylans, hinting at a possible primary role in the degradation of marine polysaccharides. The 3-D structure of *Tt*GH8 was solved providing insight into the both the native structure and ligand-complexes (xylobiose and xylotroise) at approximately 1.5 Å resolution. A catalytic mutant enabled the structure of a xylohexaose ligand complex to be captured, whereby binding of six hexose units across the binding site is consistent with the greater k_{cat}/K_m for hexasaccharide substrates. A $^{2,5}B$ conformation observed in the -1 position of bound xylotriose is consistent with the proposed conformational itinerary for this class of enzyme. This work shows *Tt*GH8 to be effective at degradation of xylan based substrates, notably marine xylan, further exemplifying the potential of *T. turnerae* for effective and diverse biomass degradation.

All work in this chapter has been published (Appendix 4) and is written as an adaption of the following;
Fowler, C.A., Hemsworth, G.R., Cuskin, F., Hart, S., Turkenburg, J., Gilbert, H., Walton, P.H. & Davies, G.J. (2018). Structure and function of a glycoside hydrolase family 8 endoxylanase from Teredinibacter turnerae, Acta Cryst. D74, https://doi.org/10.1107/S2059798318009737.

5.2 Introduction

5.2.1 The GH8 Family

Glycoside hydrolase family 8 contains 3642 entries in the CAZy database at the time of writing, 75 these classified sequences have been characterised and 12 of which have had their 3-D structures determined. The GH8 family is one of the more historical CAZyme families, known as cellulase family D through the original HCA analysis,[27-28, 129] before the subsequent classification into the present CAZy database.[34] Currently, the GH8 family is composed of sequences of bacterial origin, and all members utilise an inverting mechanism, as shown originally by Fierobe *et al* for Cel*C*[167] and later by QM/MM metadynamics[168], in their hydrolytic activity against a range of substrates, cellulose, xylan, chitosan and lichenan particularly.[34] Cleavage occurs at β (1-4) linkages, but some enzymes are known to accommodate the mixed linkage, β (1-3)-β (1-4) lichenan substrates within their active sites.[34]GH8 members belong to clan GH-M, in which there is only one other family represented, GH48. GH48 enzymes are typically found in bacterial cellulosomes,[169] and have been shown to display processive endo and reducing end attacking exo action against amorphous cellulose.[170] Both families GH8 and GH48 have the same overall structure, an $(\alpha/\alpha)_6$ barrel fold.[169, 171-173] This common protein fold is represented in 5 other clans, GH-G,L,O,P and Q.[34] The barrel structure is formed by a set of 6 pairs of antiparallel α-helices. The repeating motif of helix-loop-helix creates both an inner and outer ring, forming an overall circular 3D shape, in which the inner core is populated by aliphatic residues[174 172, 174-175], Figure 96.

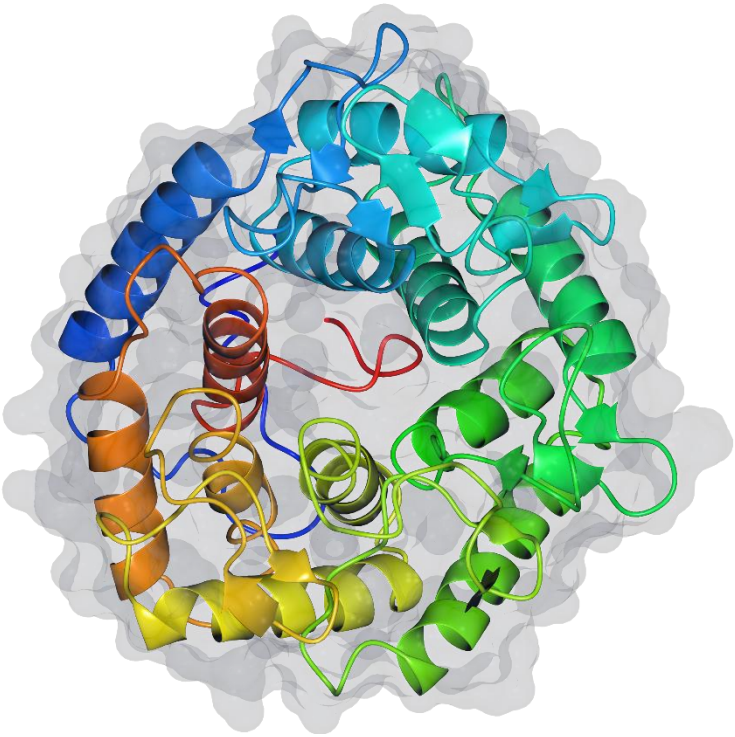


Figure 96 Structure of GH8 endoglucanase CelA (PDB code: 1KWF) shown as ribbon (coloured blue to red). Image made in CCP4mg.[55]

The first structural insight about a GH8 protein was accomplished by Alzari *et al* in 1996, which provided a first look at the overall structure of CelA, an endoglucanase from *Clostridium thermocellum.*[172] Crystallisation of complexes containing glucose, cello-oligosaccharides and an inhibitor provided models which mapped at least 5 subsites across the acidic active site cleft. The cleft is found on the opposite side of the structure to the C and N-terminus, which was thought to be involved in dockerin binding, and as such enable the protein to be involved in multi-protein amalgamations commonly employed in the cellulosome of the source organism[176]. Individual binding of glucose occurred in each site except site *C* which is now conventionally known as the -1 subsite, according to general nomenclature about GH active sites.[177] Lack of binding in the -1 site of a single pyranose ring is likely due to the need for larger substrates, where high affinity binding of a second linked moiety supports the distortion of the pyranose ring held in the -1 subsite, as distortion is normally a prerequisite for hydrolytic cleavage; a discussion of the hydrolytic inverting cleavage reaction can be found in Chapter 1. This high degree of binding affinity is established by a conserved series of tryptophan residues that are able to provide stabilising ring stacking interactions with the pyranose rings of the substrate. Further binding of the substrate is accomplished by hydrogen bonding interactions with nearby residues and those in the water molecule-protein networks located within the active site.[176] Subsequent work

on CelA by Guerin *et al*, under supervision of Alzari, went onto to provide an atomic resolution substrate complex of CelA.[173] Introduction of cellopentaose produced models in which binding of both substrate and product, cellotriose, occurred at different positions in the reducing end of the cleft, highlighting the different affinities of the protein active site for substrate and product, **Figure** 97.[173]

As found in more recent crystallographic studies of GH8 proteins, the binding of long glucans in the active site of CelA causes a large bend, or kink, in structure of the linear polysaccharide.[178] This is thought to be due to the disruption of the normal alternating pattern occurring between sugar moieties in linear chains, at the cleavage position, causing a change in direction, which is strongly encouraged by the shape of the active site cleft [176, 178]. The long deep cleft is common throughout the GH8 family, and other crystallographic studies, including the work presented in the forthcoming **Chapter**, describes the active site as forming a distinct V shape, where the central point comprises of the catalytic residues, causing a bending of the substrate [173, 178], **Figure** 97.

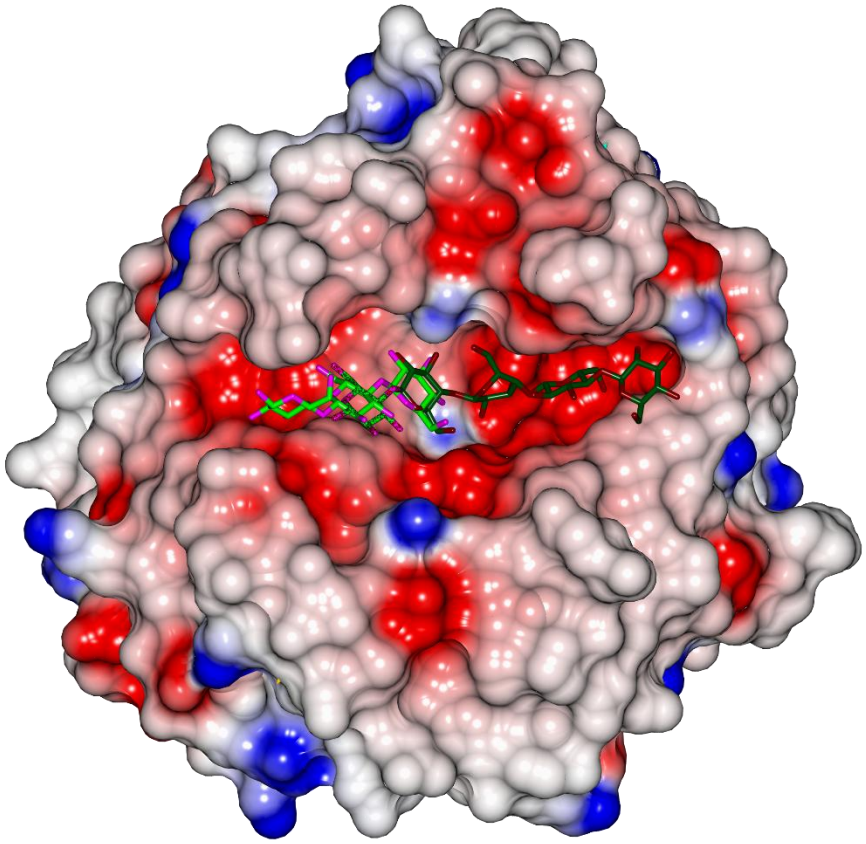


Figure 97 Structure of GH8 endoglucanase CelA , shown as surface (coloured by electrostatic potential). Binding of substrate and product is observed; cellopentaose is shown in dark green and tan, whereas the cellotriose product is shown in green and pink. The wide open active site is visible and forms a distinct V shape with the scissile bond in the centre.

5.2.2 GH8 Subfamilies

Like many other GH families, classification of the members can be broken down further into subfamilies. Adachi *et al* proposed three subfamilies for GH8, a-c, based on the positioning and residue type of the catalytic general base.[175] The work analysed the structure of a GH8 chitosanase and compared the catalytic residues to those known in CelA. It was observed that only the catalytic general acid (Glu) was conserved, whilst the general base (Asp) was interchangeable [179]. Subfamily GH8a is defined by the general base, Asp, being positioned at the n-terminal end of helix α8, as observed in CelA and by QM/MM metadynamics.[168, 172] Subfamily GH8b differs in that the expected general base Asp residue is replaced by Asn, effectively inactivating the catalytic position; however the general base role is taken up by a nearby Glu located on a long loop between helices α7 and α8.[175] Finally, the third subfamily, GH8c is yet to be characterised in terms of general base positioning.

5.2.3 This Work

Teredinibacter turnerae possess a variety of different glycoside hydrolases, but only one representative of the GH8 family is found in its genome, *Tt*GH8. The goal of this work was to study this enzyme and define its role within the CAZyme of *T. turnerae* using various biochemical techniques. Homology searches suggested that *Tt*GH8 was most similar to other GH8 enzymes shown to be active on xylan. Creation of 'inactive mutants' coupled with X-ray crystallography was applied in an effort to study the binding cleft of *Tt*GH8, creating a vivid picture of enzymatic catalysis of marine polysaccharides.

5.3 Methods

5.3.1 Materials

Substrates used in this study were tamarind xyloglucan (tXyG), mixed linkage barley beta glucan (bMLG), Avicel (Sigma), Phosphoric acid swollen cellulose (PASC) (Sigma), birchwood xylan (bX), mixed linkage β 1-3, β 1-4 xylan (MLX) (Elicityl/Oligotech). Xylo-oligosaccharides were bought commercially; xylobiose (X2) from Sigma and TCP, Xylotriose (X3), xylotetraose (X4), xylopentaose (X5) and xylohexaose (X6) from Megazyme. Crystallisation screens were obtained from Molecular Dimensions (PACT) and Hampton Research (PEG/ION, Index and Crystal Screen HT). The mutagenesis kit was purchased from New England Biolabs, and primers purchased from Eurofins-GATC. The HPAEC-PAD experiment was run using a HPLC-Dionex machine belonging to the Prof. Harry Gilbert group at Newcastle University and was used under guidance. 3,5-Dinitrosalicylic acid was purchased from Sigma. Pure recombinant *Tt*GH8 (*ACR14722.1)* was produced as described in Chapter 2

5.3.2 Mutant *Tt*GH8 production

Catalytically inactive mutants of *Tt*GH8 were designed, using structural and activity information from the literature on similar GH8 proteins as well as structures obtained of the apo protein during this work. The mutations were designed using custom primers and implemented using the Q5 site directed mutagenesis kit (New England Biolabs), Table 17. Expression testing for all constructs was carried out prior to large scale production, which was carried out as described for *Tt*GH8. To check purity, the samples were analysed by SDS PAGE throughout purification. The mutants were checked by ESI MS, to ascertain whether there has been an appropriate mass change.

Table 17 Point mutations for TtGH8 catalytic mutants, where Glu73 and Asp281 are altered to catalytically inactive residues. Primers for both the forward and reverse directions are shown and correspond to the *Tt*GH8 construct described in **Chapter** 2

Point Mutation	Primers	
	Forward	Reverse
TtGH8Glu73-Gln	TGTGCGTAGCCAAGGTATGAGCTAC	TCGTTGCTGTTAACGTCATAC
TtGH8Glu73-Ala	TGTGCGTAGCGCCGGTATGAGCTAC	
TtGH8Asp281-Asn	CTTTCGTTACAACGCGTGGCGTAGCG	TCCACGCTCTCCGGACGC
TtGH8Asp281-Ala	CTTTCGTTACGCCGCGTGGCGTAGCG	

5.3.3 TSA of *Tt*GH8 and Mutants

Samples containing SYPRO-orange dye (1000X stock, 15 µL) and either *Tt*GH8, *Tt*GH8D281N or *Tt*GH8E73Q (final concentration of 1 mg/ml) with either buffer or substrate (30 µl total) were analysed using a TSA programme run on a Stratagene rtPCR machine. Samples were heated from 20°C to 91° in increments of 1 °C over 71 cycles. The fluorescence of SYPRO orange was monitored throughout and the data used to calculate the protein melting temperature. Curves were fitted using a free online tool developed by Paul Bond at the University of York and is available at; http://paulsbond.co.uk/jtsa.

5.3.4 TLC and LCMS Analysis of *Tt*GH8 Hydrolysis Products

Overnight hydrolysis reactions with xylo-oligosaccharides X2-X6 (1 mM), wAX, rAX, cX, bX, MLX and tXyG at 1 mg/ml were incubated at 37 °C with 1 µM *Tt*GH8. The samples were heated at 90 °C prior to spotting on a TLC plate (total 4 µl). Standard containing X2-X6 at 1 mM each were run on the same plate. TLC plates were run as described in Chatper 3. Hydrolysis samples for LCMS were prepared using ammonium acetate buffer (pH 6, 50 mM), approximately 1 mg/ml of substrate and 10 mM enzyme. Samples were incubated at 37 °C overnight and shaken at 500 rpm. If required, samples were centrifuged to remove any solid materials and 100 µl loaded onto a Cosmosil Sugar-D HPLC column using the LC-MS Dionex system where the separated products were analysed by ESI or PAD mass spectrometry. Running buffers were a mixture of water and acetronitrile with some test runs also including 1 % formic acid.

5.3.5 Substrate Depletion Kinetics

Substrate depletion kinetics was performed on *Tt*GH8 with xylohexaose, xylopentaose and xylotetraose and measured using high performance anion exchange chromatography-pulsed ampiometric detection (HPAEC-PAD) Dionex system. Hydrolysis reactions were run at 37 °C, for 25 minutes using different substrate and enzyme concentrations to find conditions which yielded linear results. Final enzyme- substrate conditions were as follows; X6 (50 µM substrate, 0.72 nm *Tt*GH8), X5 (100 µM substrate, 3.2 nM *Tt*GH8) and X4 (100 µM substrate, 64 nM *Tt*GH8). Reactions were run in 20 mM sodium phosphate, 100 mM NaCl buffer, pH 7. Aliquots were removed at set time points and immediately placed into a heat block at 90°C

for 10 mins to inactivate *Tt*GH8; all samples were treated in the same manor and as such any delay in stopping the enzyme activity during heating would have been the same throughout the samples. All samples were mixed with fucose (250 µM) which acted as an internal standard. Samples were run using HPLC Dionex system, on an anion exchange column (CARBOPAC) using a sodium acetate gradient. Data was normalised using the peak detection for the internal standard to mitigate any loss in signal between runs.

5.3.6 3,5-Dinitrosalicylic acid reducing sugar assay

Hydrolysis reactions with enzyme and substrate were run at 37 °C and aliquots removed and mixed with DNSA agent (1 % (w/v) 3,5-Dinitrosalicylic acid , 0.2 % (v/v) phenol, 1 % (w/v) NaOH, 0.002 % glucose (w/v), 0.05 % (w/v) $NaSO_3$) at set time points. Phenol was not used during the analysis of *Tt*GH8 *D*281*N*. The colour was developed by heating the reaction aliquots at 90 °C for 20 min before cooling on ice for 10 min. The absorbance of each sample was measured at 575 nm. A standard curve of 0-500 µg/ml xylose plus 1 mg/ml polysaccharide substrate was used to quantify the released reducing sugar chain ends. Absorbance at 575 nm was collected for each sample and data plotted and analysed using the method described in Appendix 3.

5.3.7 Crystallisation Structure Solution of *Tt*GH8 and *Tt*GH8D281N

Initial crystallisation screening was performed robotically using a Mosquito crystal robot and commercial screens including Crystal screen HT, Index and PACT. Crystal hits were obtained for *Tt*GH8 and *Tt*GH8D281N. A 24-well optimisation screen containing sodium acetate (0.1 M, pH 4.6 -5.2), NaCl (0.2 M), polyethylene glycol 14-24% was used to produce the final crystallisation condition for *Tt*GH8: sodium acetate (0.1M, pH 5.0), NaCl (0.2 M), polyethylene glycol 16%. Thin rod shaped crystals were fished and protected with a cryo solution (mother liquor plus 30% ethylene glycol) and soaked in ligands. *Tt*GH8 structures containing X2 and X3 resulted from soaking 'apo' crystals for 30 seconds in a solution containing the mother liquor, 30% ethylene glycol and 150 mM X2 or X3. TtGH8D281N 'apo' crystals grown in a condition containing HEPES (0.1 M, pH 6.8), ammonium sulfate (0.2 M) and 20 % PEG 6000 were soaked in X6 (20 mM) for approximately 10 seconds and then frozen in liquid nitrogen. Crystal data sets were collected at the Diamond synchrotron by remote access using IO2 and IO4 beamlines. Molecular replacement (PHASER[157]) and

refinement (REFMAC[160]) were carried out using the CCP4i2 pipeline. *Tt*GH8 was modelled using PDB entry 1WU4 and data cut to 1.4 Å. Manual manipulation in COOT followed by refinement using REFMAC was cycled several times until the R_{cryst} and R_{free} were consistent (R_{cryst} =0.15, R_{free}=0.18). *Tt*GH8-ligand complexes and TtGH8D281N-X6 were solved by molecular replacement using MOLREP[180] with the 'apo' structure of *Tt*GH8 as the search model *Tt*GH8. Structural analysis was carried out in CCP4mg[181] by comparing apo with protein ligand complexes and several similar PDB entries.

5.4 Results

5.4.1 Protein Production

*Tt*GH8 was produced to a high yield as described in Chapter 2. Only two of the *Tt*GH8 mutants were successfully expressed in a soluble form and were purified using the same methods as carried out for production of the apo protein; *Tt*GH8D281N and *Tt*GH8E73Q. Both mutants where either Glu73 or Asp281 were mutated to an alanine were not expressed in a soluble form. The catalytic mutants were analysed by ESI MS to ascertain that the point mutations had been successful, Figure 98.

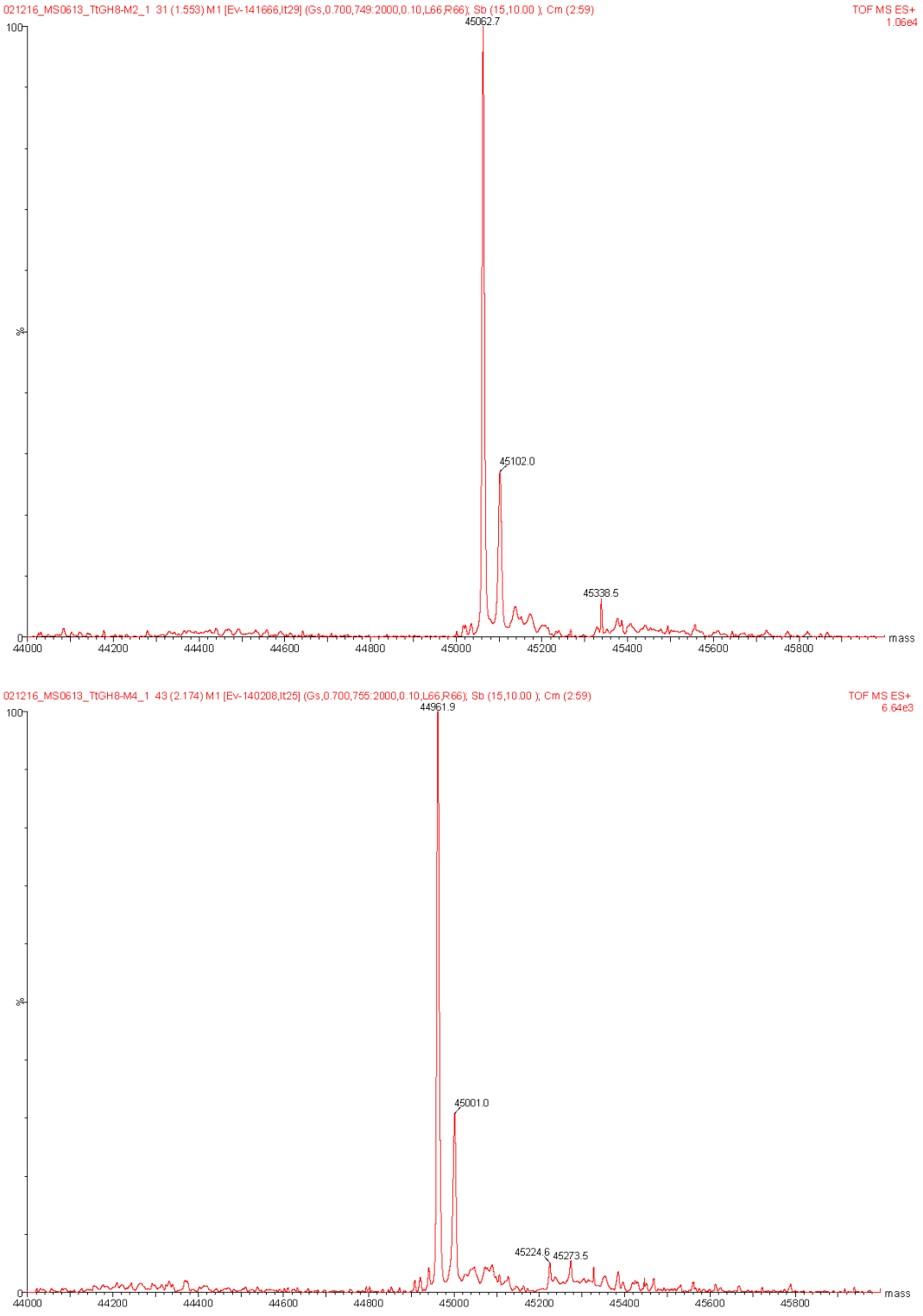


Figure 98: ESI-MS of *Tt*GH8*D*281*N* (top) and *Tt*GH8*E*73*Q* (bottom). Masses are 45068 Da and 44967.2 Da respectively.

5.4.2 Thermal Shift Analysis

Protein substrate activity was initially probed by TSA which showed a positive increase in protein melting temperature upon mixing *Tt*GH8 with BX, 2.1 °C and X6, 2.9 ° C, **Figure** 99. Both catalytic mutants displayed lower protein melting temperatures than native *Tt*GH8 (57.2 °C), indicating lower protein stability due to alterations of the catalytic residues. However both mutants were stabilised by addition of X6, with melting temperatures increase by 7.2 °C for *Tt*GH8D281N and 15.3 °C for *Tt*GH8E73Q, as shown in **Figure** 100 and summarised in **Table** 18.

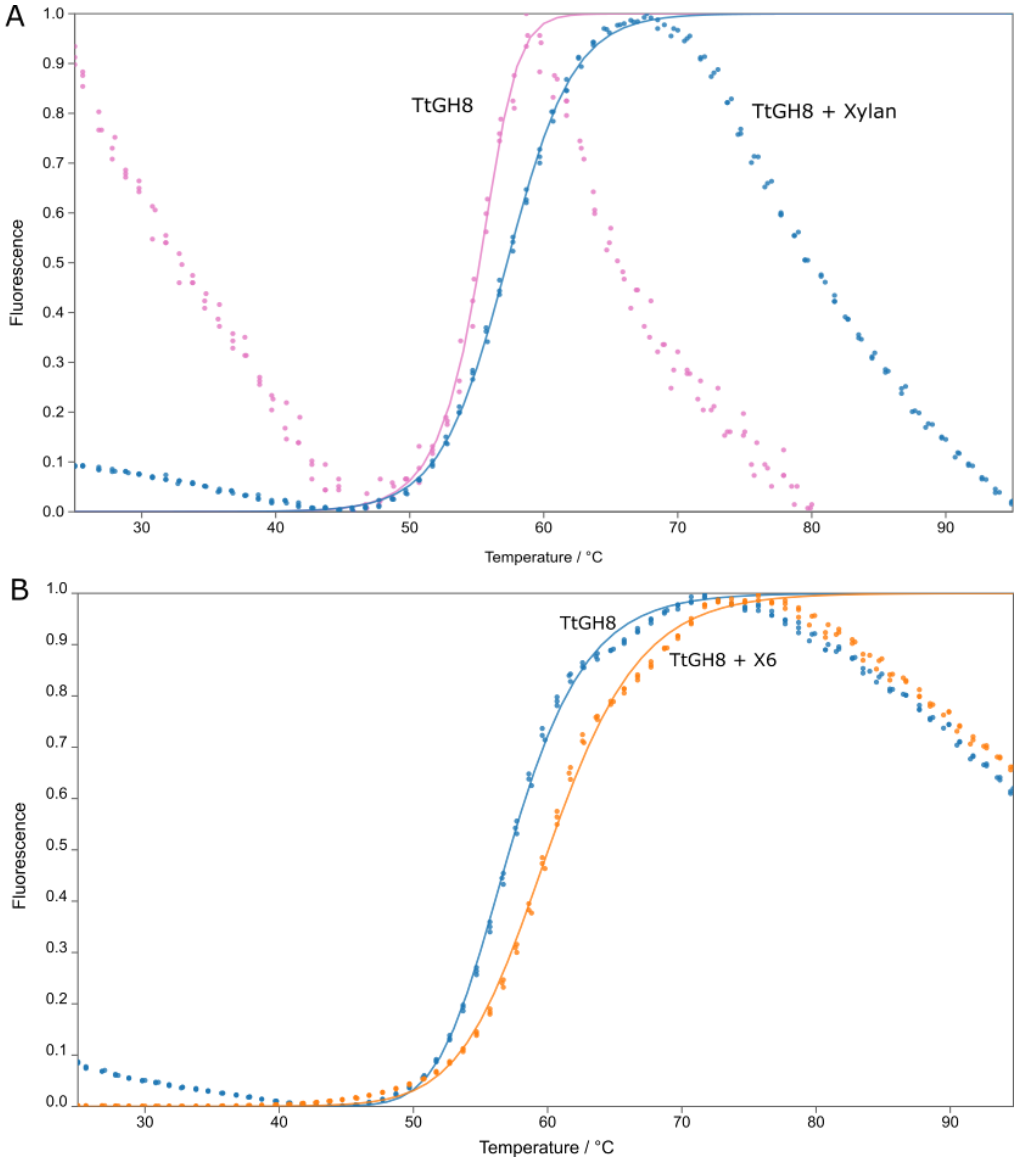


Figure 99 TSA of *Tt*GH8 (22 µM) and incubation with A) BX and B) X6 in the presence of SYPRO orange. Analysis of the protein melting curves was completed using the JTSA online tool developed by Paul Bond, available at URL http://paulsbond.co.uk/jtsa/#/input. Addition of BX and X6 shifted the melting temperature of the TtGH8 from 55.2 -57.3 °C and 57.2 - 60.1 °C respectively. Figure as published in Fowler *et al.*[182]

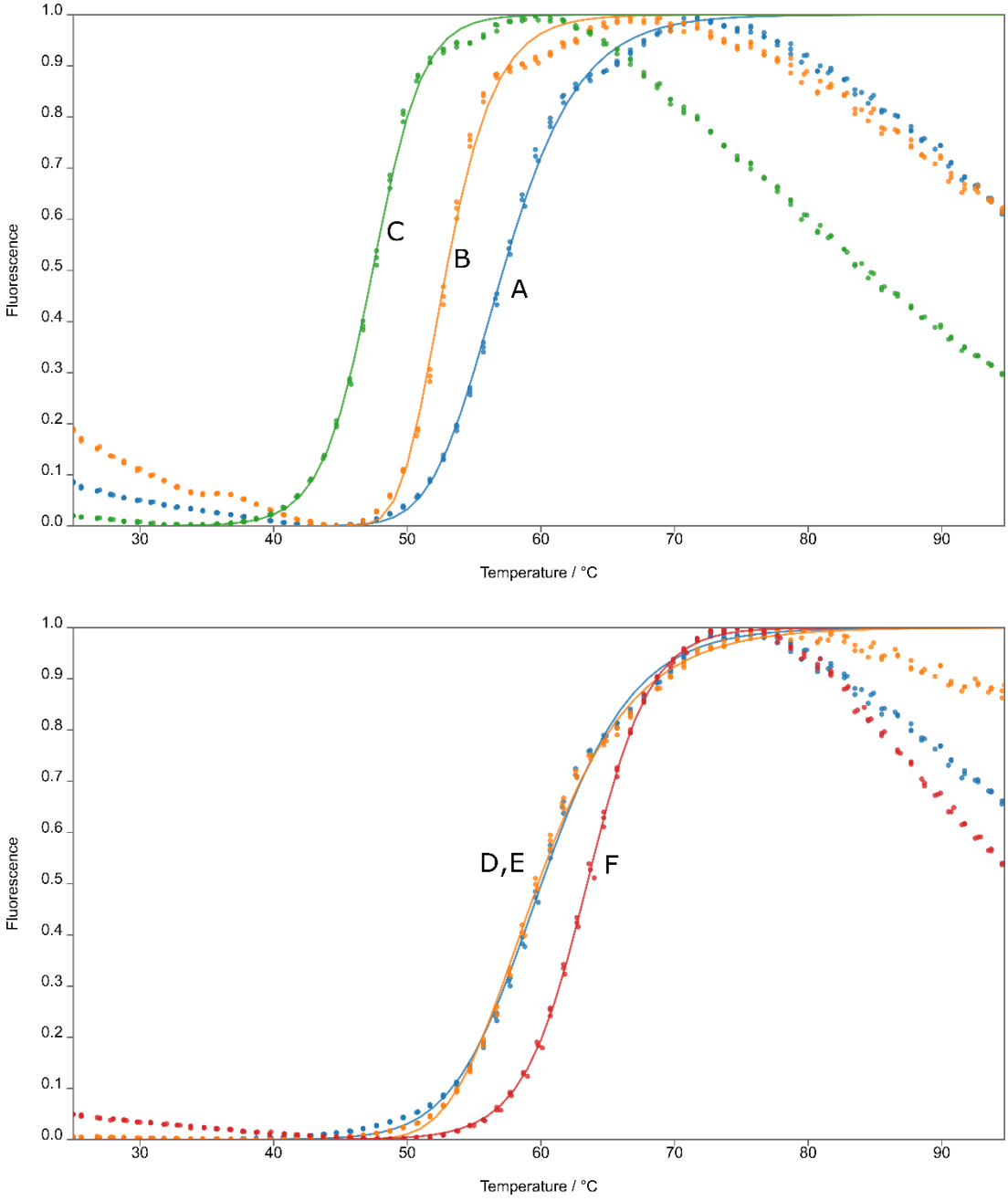


Figure 100 TSA of A) *Tt*GH8, B) *Tt*GH8*D*281*N*, C) *Tt*GH8*E*73*Q*, D) *Tt*GH8 + X6 (blue line), E) *Tt*GH8*D*281*N* + X6 (orange line), and F) *Tt*GH8*E*73*Q* + X6.

Table 18 Average melting temperature calculated from the TSA curves with and without incubation with X6 of TtGH8 and the two catalytic mutants.

Sample	Melting temperature/ °C	Sample	Melting temperature/ °C
*Tt*GH8	57.2	*Tt*GH8 + X6	60.5
*Tt*GH8*D281N*	53	*Tt*GH8*D281N* + X6	60.2
*Tt*GH8*E73Q*	48.2	*Tt*GH8*E73Q* + X6	63.5

5.4.3 TLC and LCMS Activity Analysis

Overnight hydrolysis reactions of *Tt*GH8 with soluble xylo-oligosaccharides and polysaccharides wAX, rAX, cAX, bX, MLX and tXyG were run and soluble samples analysed by TLC, Figure 101. A set of xylo-oligosaccharides X2-X6 were used as standards for comparison of the degradation products. *Tt*GH8 showed activity on wAX, rAX, bX and MLX. Complete conversion of X4-X6 into X3 and X2 products was observed. No reaction was seen for *Tt*GH8 with X3 or X2.

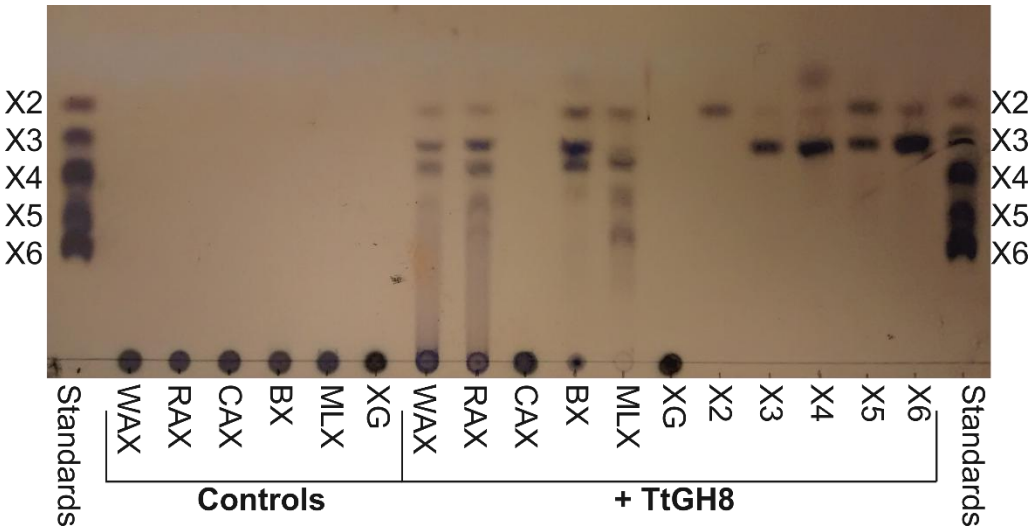


Figure 101 *Tt*GH8 activity was explored using a range of different xylan polysaccharides as well as xyloglucan and analysed by TLC. Standard xylo-oligosaccharide X2-X6 are shown on both sides of the plate to provide a weight comparison ladder. Controls of wAX, rAX, cAX, bX, MLG and tXyG (XG in above) show no soluble products, with all material spotted onto the plate, remaining in the same position. Soluble products were observed for all polysaccharides except CAX and XG. TtGH8 was also tested on xylo-oligosaccharides, with complete breakdown into smaller products observed only for X4, X5 and X6. Figure as published in Fowler *et al*.[182]

The catalytic mutants were also tested for activity to determine whether the mutations prevented the enzyme from breaking down the substrate. Figure 102 shows a TLC comparing the activity of native *Tt*GH8, *Tt*GH8D281N and *Tt*GH8E73Q on X6. For the native protein and the catalytic mutant containing the alteration D281N, complete degradation of X6 in under 5 mins. On the other hand, mutation of the catalytic glutamate residue abolished all enzymatic activity on this substrate.

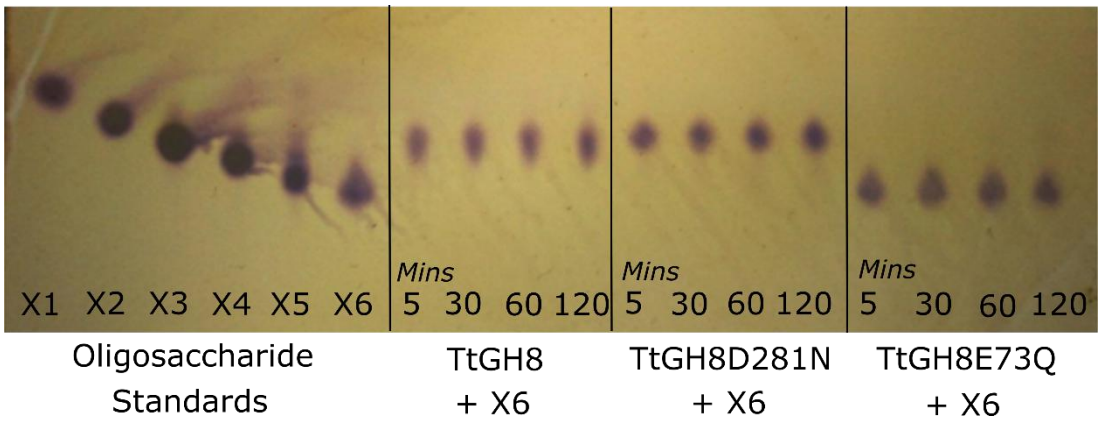


Figure 102 TLC of *Tt*GH8, *Tt*GH8*D*281*N* and *Tt*GH8*E*73*Q* activity on X6 over two hours at 37 °C. Both *Tt*GH8 and *Tt*GH8*D*281*N* were able to degrade the substrate completely in under 5 mins, whereas *Tt*GH8*E*73*Q* showed not activity towards X6 over two hours incubation.

The reactions of *Tt*GH8 on bX and X4-X6 were probed in more detail using LCMS in order to assign the products observed in the TLC accurately. **Figure** 103 shows the LCMS profiles obtained from the activity of *Tt*GH8 on X4-X6. After incubation overnight at 37°C, samples showed complete degradation of X4 to X3, X5 to X3 + X2, and X6 to X3. All xylo-oligosaccharides substrates and products were observed as formic acid adducts during LCMS experimentation due to the buffer used to separate samples on the size exclusion column. Activity on BX was tested by analysing the soluble fraction of the reaction mixture after overnight incubation with *Tt*GH8 at 37°C. The LCMS profile suggested a high abundance of X3 as the major degradation product, along with a smaller amount of X2, **Figure** 104. X3 and X2 are observed as formic acid adducts at 327.2 and 459.2 m/z respectively.

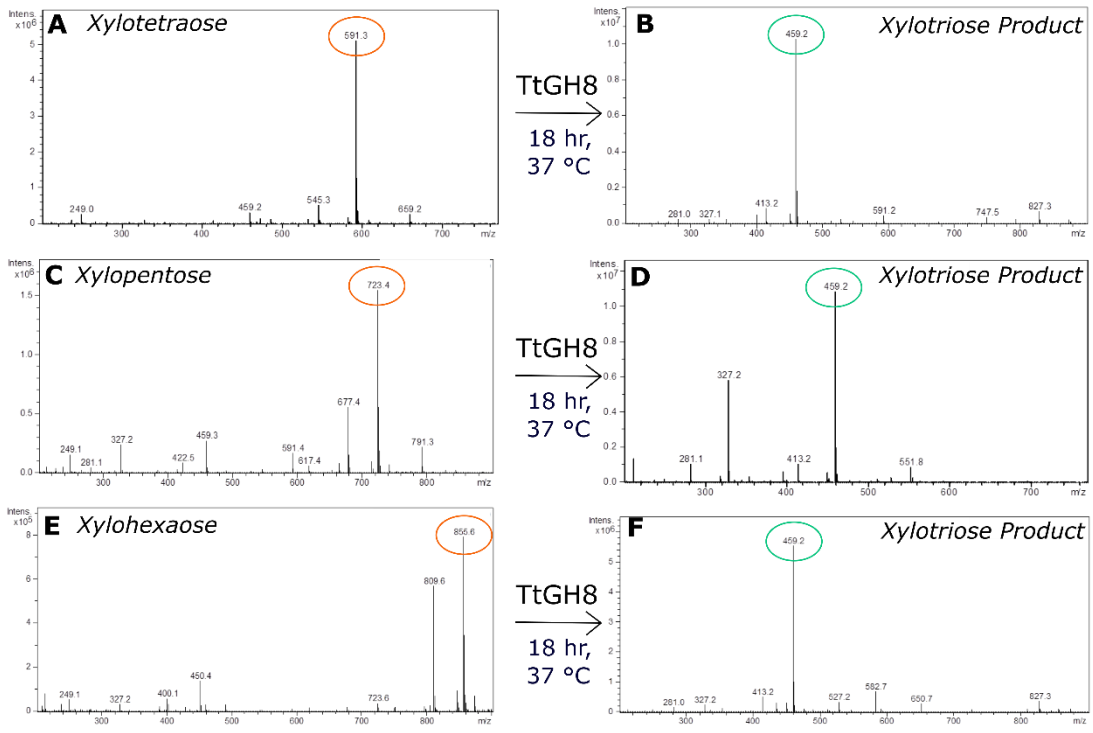


Figure 103 LCMS results of *Tt*GH8 against larger xylo-oligosaccharides. A) X4 (591.3 m/z, formic acid adduct) B) Reaction products of *Tt*GH8 and X4, producing X3 (459.2 m/z formic acid adduct). X1 is not seen due to detections limits of the equipment. C) X5 (723.4 m/z formic acid adduct) D) Reaction products of *Tt*GH8 and X5, producing X3 (459.2 m/z formic acid adduct) and X2 (327.2 m/z formic acid adduct). E) X6 (855.6 m/z formic acid adduct) F) Reaction products of *Tt*GH8 and X6, producing only X3 (459.2 m/z formic acid adduct).Figure as published in Fowler *et al.*[182]

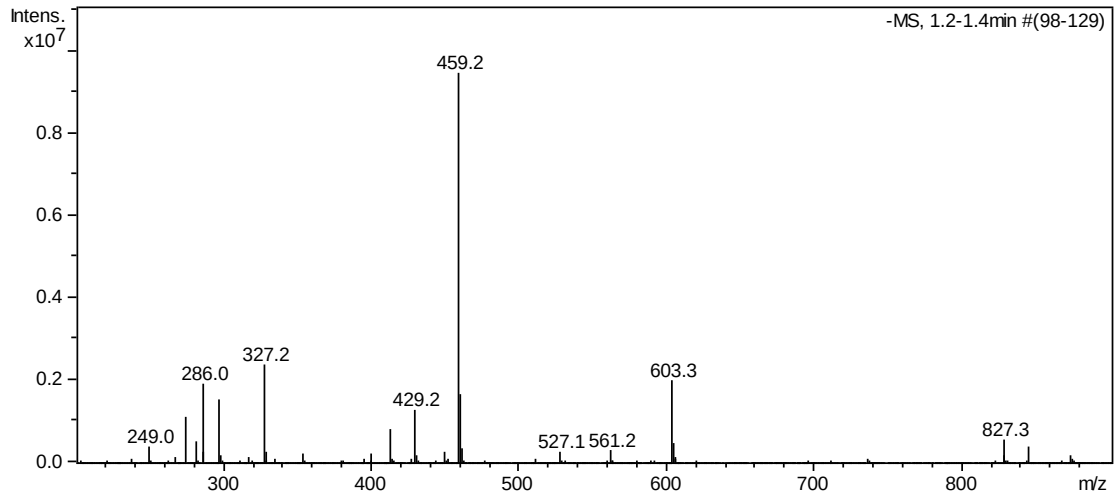


Figure 104 LCMS of soluble reaction products generated after incubation of *Tt*GH8 with BX for 18 hr at 37 °C. X2 and X3 are observed as formic acid adducts at 327.2 and 459.2 m/z respectively. **Figure** as published in Fowler *et al.*[182]

5.4.4 Kinetic Analysis on xylose based substrates

Following on from the TLC experiment, which indicated *Tt*GH8 activity on 4 polysaccharides with xylose backbones, HPAEC-PAD was used to evaluate the soluble oligosaccharides produced when *Tt*GH8 was incubated with bX, wAX, bMLG and rAX. The soluble species naturally present in the substrates were first analysed and HPAEC-PAD spectra are shown in Figure 105. All 4 polysaccharides show new peaks compared with the control samples, after incubation with the enzyme, indicating newly formed soluble oligosaccharides are due to the action of *Tt*GH8, Figure 106.

Kinetic measurements on the activity of *Tt*GH8 on short chain xylo-oligosaccharides were made using HPAEC-PAD and an experimental method known as substrate depletion as described in Appendix 2; whereby reduction in a specific peak is monitored over certain time points of the reaction as shown in Figure 107.[183-185]

Substrate depletion analysis was performed on X6, X5 and X4 using appropriate experimentally determined concentrations of *Tt*GH8 and substrate, Figure 108. Substrate depletion of X6 was measured using only 0.72 nM of enzyme and produced the highest k_{cat}/K_m value out of the three xylo-oligosaccharides tested. Interestingly the observed rate of reaction was lowered on incubation with higher substrate concentrations. X4 was the slowest to be hydrolysed, (k_{cat}/K_M, 6.1 x 10^5 ± 3.1 x 10^4 M^{-1} min^{-1}), requiring approximately 100 fold higher enzyme concentration than that needed for hydrolysis of X6. The kinetics shows that *Tt*GH8 has a higher affinity for X6 (k_{cat}/K_M, 7.5 x 10^7 ± 1.1 x 10^6 M^{-1} min^{-1}) and a

slightly lower affinity for X5 (k_{cat}/K_M, 1.4 x 10^7 ± 1.4 x 10^6 M^{-1} min^{-1}). For comparison, all kinetic parameters measured are also summarized in **Table** 19.

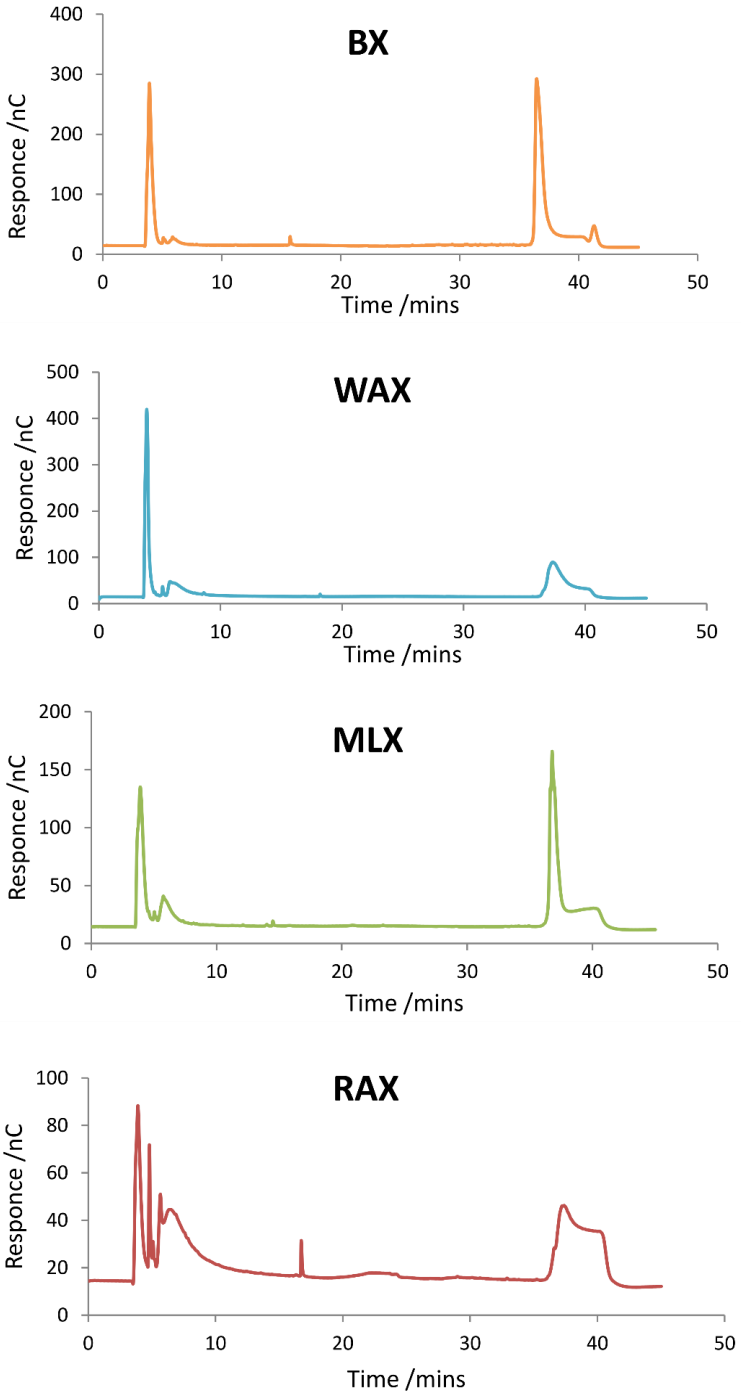


Figure 105 HPAEC-PAD analysis of the soluble fraction taken from un-reacted polysaccharides showing few peaks.

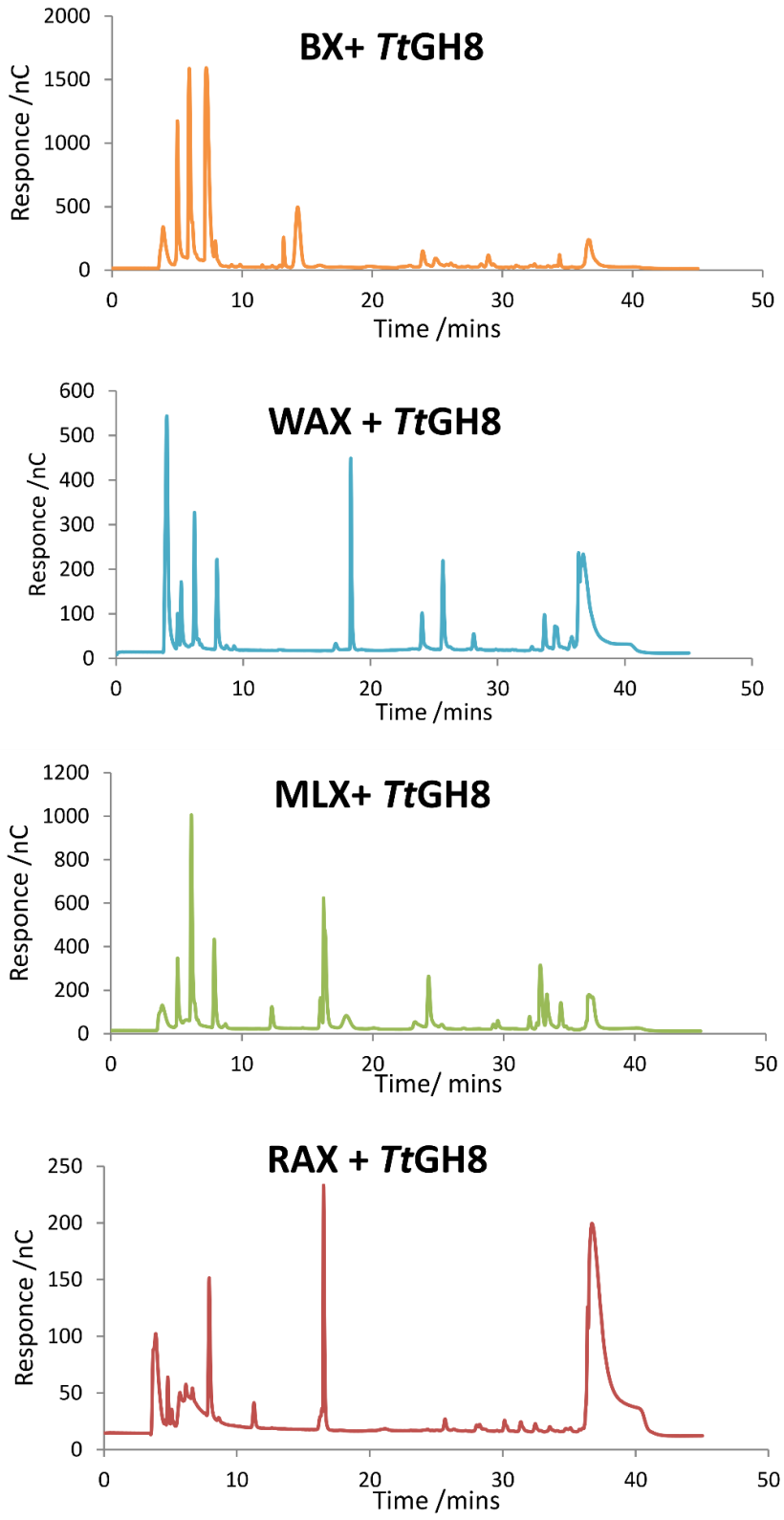


Figure 106 HPAEC-PAD analysis of the soluble fraction taken from overnight reactions of *Tt*GH8 with bX, wAX, bMLG and rAX . Many more peaks are observed compared with the control reactions analysed in the previous figure. More soluble products were observed from the reaction of *Tt*GH8 with MLG and BX than with either wAX or rAX, where the soluble products in the latter two substrates have a lower response.

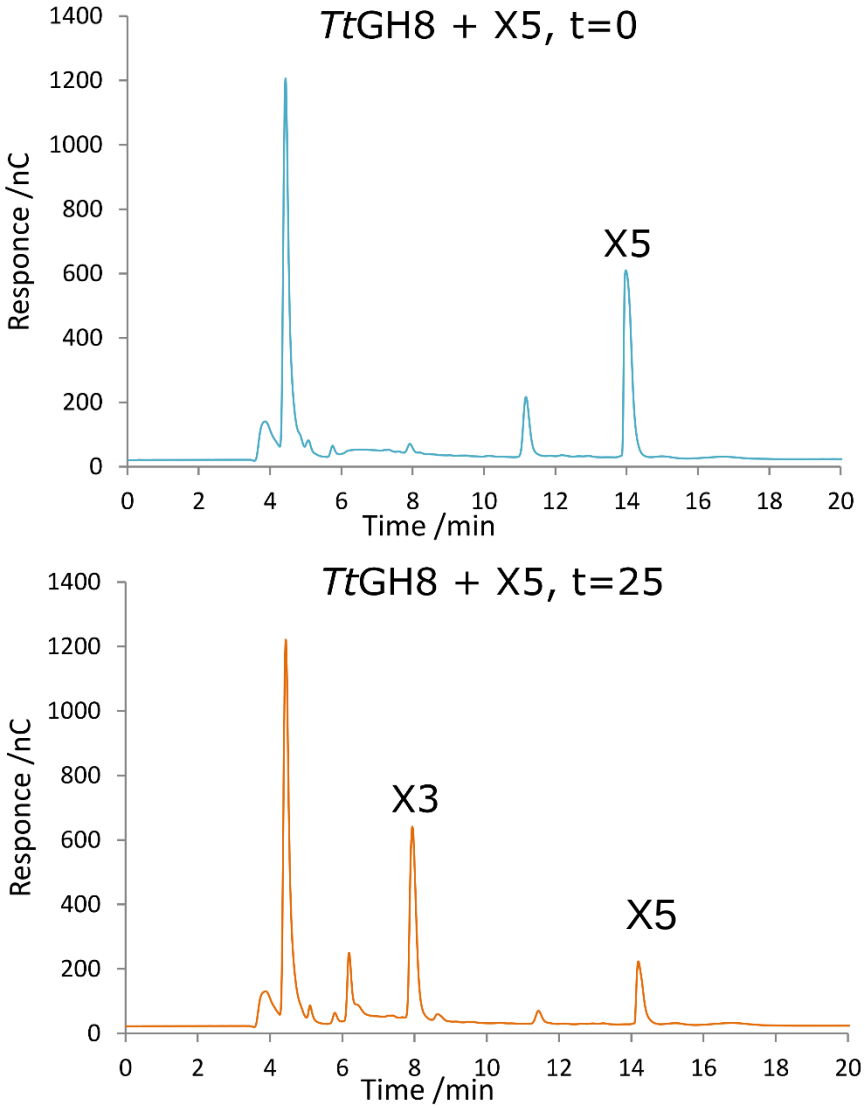


Figure 107 HPAEC-PAD substrate depletion assay of *Tt*GH8 on xylopentaose . Aliquots were removed from the reaction at set time points and later analysed by HPAEC-PAD; the above chromatograms show the samples taken at the beginning and end points of the reaction. A large amount of X5 is degraded to X3. Note, there is a small amount of contaminating xylotetraose within the xylopentaose sample used for this experiment, seen to the left of the X5 peak.

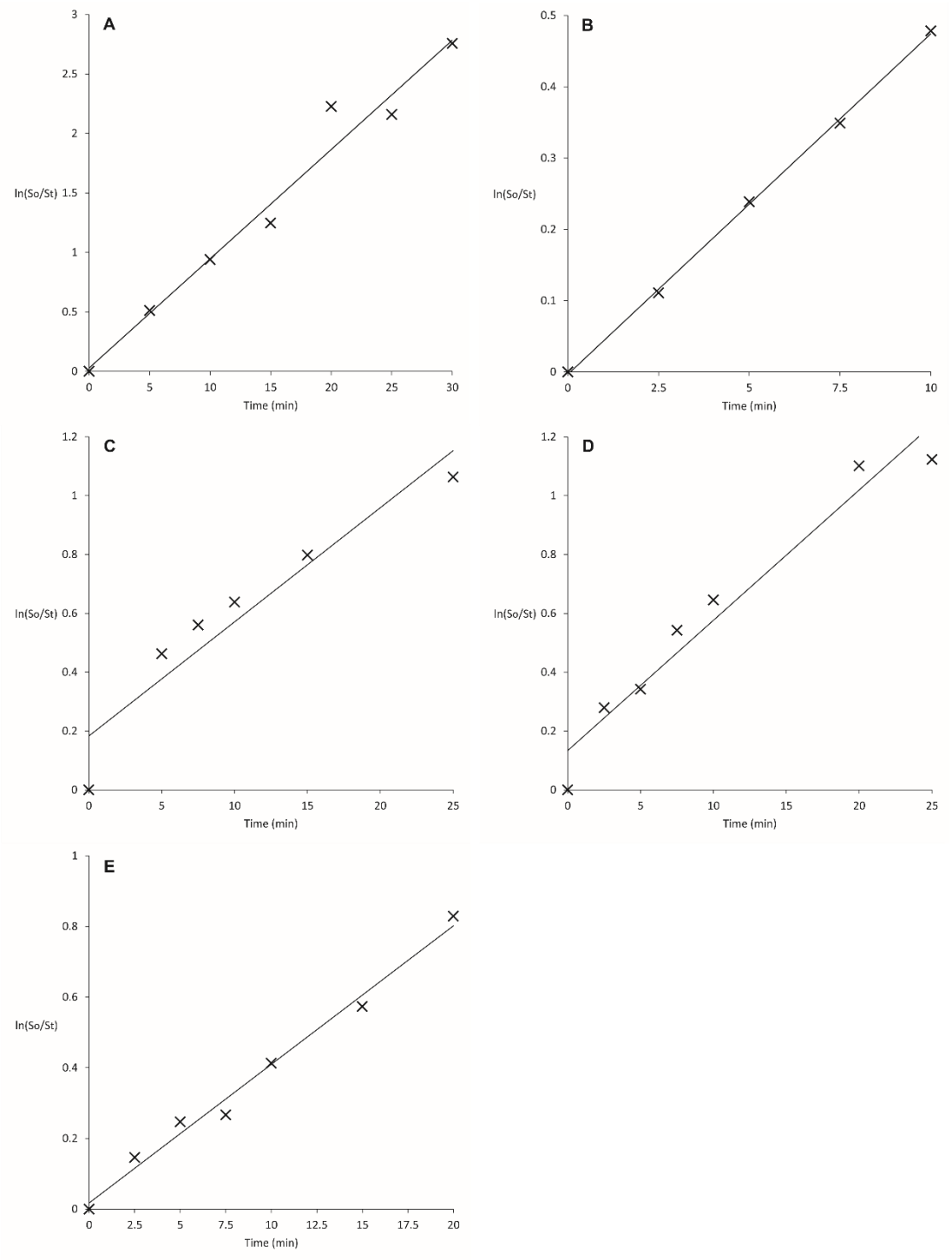


Figure 108 *Tt*GH8 HPAEC-PAD substrate depletion analysis . A) 1.28 nM, 100uM X6, B) 0.64nm, 100 um X6, C)3.2 nm, 100 um X5, D)3.2nm, 100 X5, E)64 nm, 100 um X4. Figure taken from Fowler *et al.*[182]

The activity of *Tt*GH8 of xylose based polysaccharides was also measured kinetically using the DNSA reducing sugar assay on bX, wAX and bMLG for both *Tt*GH8 and *Tt*GH8D281N. This method quantifies the amount of reducing sugar produced during enzymatic hydrolysis of polysaccharides. **Figure** 109 shows the reaction as photographed after absorbance measurements had been taken and is a strong visual representation of the reaction progress over time; where the colour change from yellow to brown indicates an increasing amount of reducing sugar present in the sample. **Figure** 109 also compares two different substrates, where the difference in rate of reaction of *Tt*GH8 on MLX and wAX is evident visually and graphically as described in the methods section) as shown in **Figure** 110. The activity of the catalytic mutant which still maintained the ability to hydrolyse xylose based substrates in activity assays was also tested by DNSA reducing sugar assay and graphically represented in **Figure** 110. The calculated value of k_{cat}/K_M is significantly affected by the removal of the catalytic aspartate residue, showing only 0.1 % retention in activity compared to the native enzyme. All calculated values of k_{cat}/K_M by DNSA reducing sugar assay are summarized in **Table** 19 allowing comparison of the activity of *Tt*GH8 on different substrates, where an apparent preference was observed for bMLG.

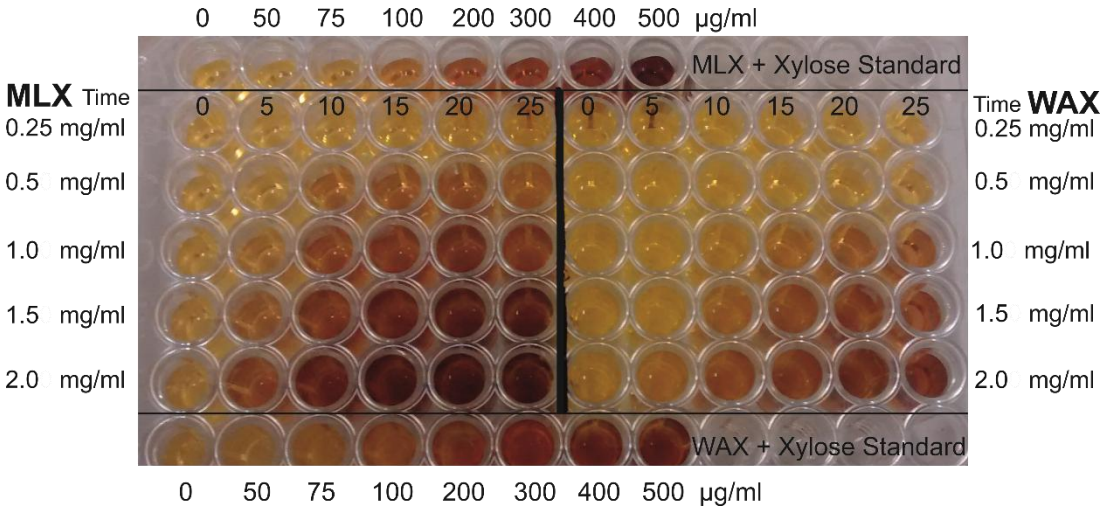


Figure 109 *Tt*GH8 activity on polysaccharides was determined using the DNSA reducing sugar assay. Photograph of plate containing aliquots representing different reaction time points (0-25 mins) and substrate concentrations (mixed linkage xylan and wheat arabinoxylan 0.5-2.0 mg/ml). Reaction of the DNSA agent with reducing sugars causes a visible colour change from light yellow to deep brown, with the amount of colour change measured by absorbance at 575 nm being equivalent to the amount of reducing sugar (chain ends) produced during the enzyme reaction. **Figure** as published in Fowler *et al.*[182]

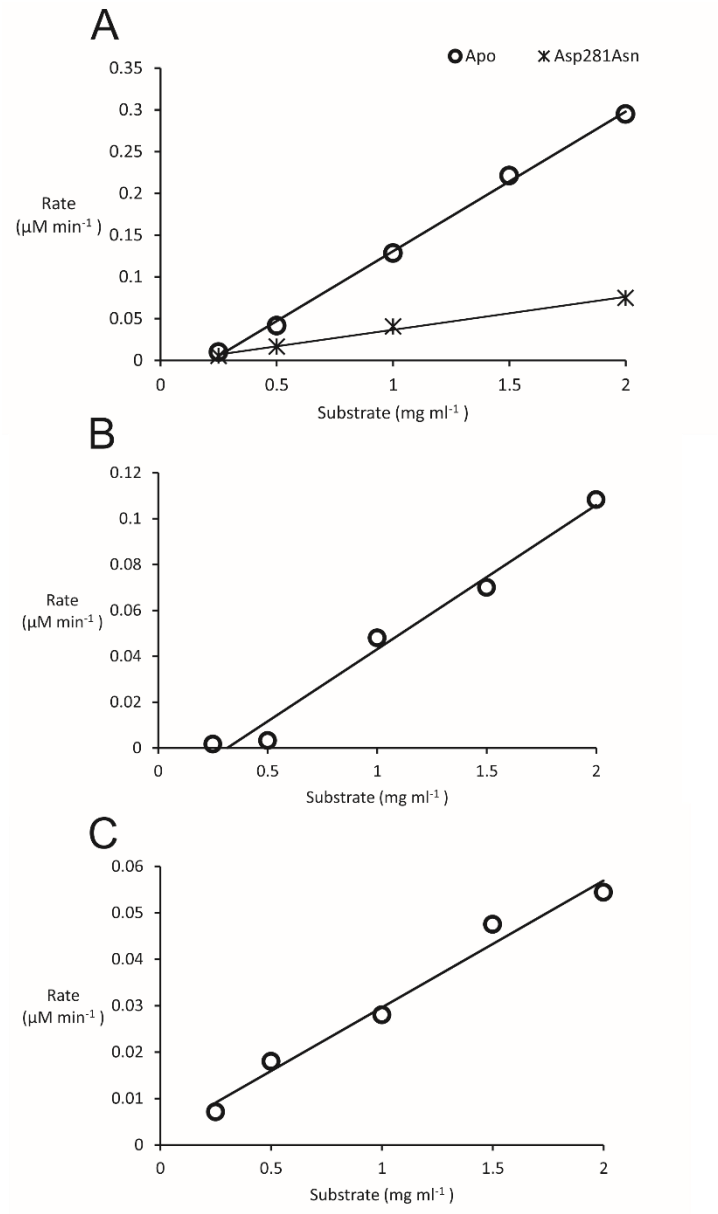


Figure 110 DNSA reducing sugar assay of *Tt*GH8 and *Tt*GH8 catalytic mutant on A) MLX, B) *Tt*GH8 on WAX and C) *Tt*GH8 on BX. The rates for each reaction condition are plotted against the substrate concentration, where the gradient once divided by enzyme concentration is equivalent to k_{cat}/K_M. Figure as published in Fowler *et al*.[182]

Table 19 Catalytic activities of *Tt*GH8 on xylan substrates, as determined by HPAEC-PAD (oligosaccharides) or DNSA reducing-sugar assay (polysaccharides). Activity of *Tt*GH8*D281N* on MLX was also measured on bMLG and found to be significantly slower than the native enzyme. Table as published in Fowler *et al.*[182]

Oligosaccharides	k_{cat}/K_M (M^{-1} min^{-1})	Polysaccharides	k_{cat}/K_M (mg^{-1} ml min^{-1})
Xylohexaose	$7.5 \times 10^7 \pm 1.1 \times 10^6$	**bX**	$1.8 \times 10^7 \pm 4 \times 10^6$
Xylopentaose	$1.4 \times 10^7 \pm 1.4 \times 10^6$	**wAX**	$6.3 \times 10^6 \pm 5 \times 10^5$
Xylotetraose	$6.1 \times 10^5 \pm 3.1 \times 10^4$	**MLX**	$1.6 \times 10^8 \pm 4 \times 10^6$
		MLX (*Tt*GH8*D281N*)	$1.8 \times 10^4 \pm 1 \times 10^3$

5.4.5 Crystallisation and Structure Solution

*Tt*GH8 was crystallised very easily and in high quantities in several well conditions in different commercial screens. Figure 111 shows an image of one drop where the overall crystal morphology is captured; long needles. The crystal hit conditions were evaluated and optimised using the hanging drop method focused around slight deviations in pH and PEG 6000 concentration. Crystals were fished from wells and soaked briefly in a cryo-protectant before being frozen in liquid nitrogen. Protein ligand complexes with xylobiose and xylotriose were achieved by briefly soaking *Tt*GH8 crystals in the solution of the mother-liquor and a very high concentration of ligand which proved successful in producing enzyme-substrate complexes. Crystal soaking experiments using the catalytic mutant *Tt*GH8*D*281*N* were carried out in an attempt to capture the binding of X6 across the protein active site. As the catalytic mutant still retained some activity, as shown previously, it was different to soak X6 into the crystals without it being hydrolysed and observed as a product complex. The soaking times were reduced significantly from 15 minute crystal soaks within mother liqueur and X6 in the first instance to the final successful soak timing of 10 seconds. All crystals were tested for diffraction in house before being sent to the Diamond synchrotron and collection statistics for the different data sets corresponding to *Tt*GH8, *Tt*GH8-X2, *Tt*GH8-X3 and *Tt*GH8*D*281*N*-X6 are shown in Table 20.

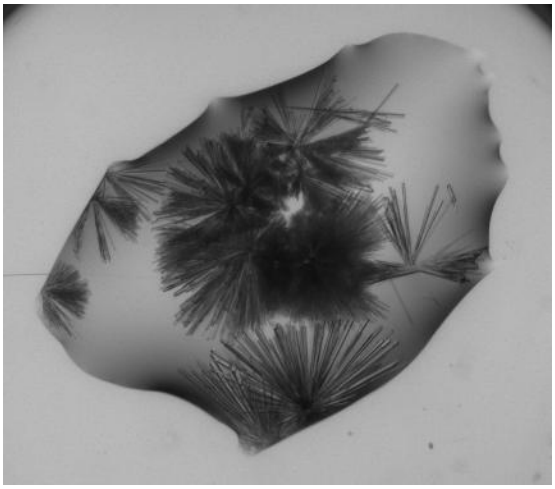


Figure 111 *Tt*GH8 crystallised using the Hampton Screen. Successful well condition contained Sodium acetate (pH 4.6), NaCl and polyethylene glycol 6000.

The 'apo' *Tt*GH8 structure was solved using molecular replacement (MOLREP).[180] A PDB BLAST search identified a GH8 'reducing end xylose releasing exo-oligoxylanase' from *Bacillus halodurans* C-125 (PDB entry 1WU4) as the closest sequence homology match.[113, 186] Refinement cycles involving REFMAC[160] and manual alteration in COOT[159] were applied several times before the R_{cryst} and R_{free} remained constant. All other data sets were solved using the model of *Tt*GH8, where the catalytic mutation in *Tt*GH8D281N was manually changed in COOT. After sufficient modelling of the main chain and water molecules within the structure solutions, data sets where there may have been possible ligand binding in the active site were analysed in COOT to check for any unmodelled electron density. The ligands X2, X3 and X6 were built in by incorporating the correct CIF dictionary into the PDB file. The refinement statistics and PDB codes for *Tt*GH8, *Tt*GH8-X2, *Tt*GH8X3 and *Tt*GH8*D*281*N*-X6 are shown in Table 21.

Table 20 *Tt*GH8 and *Tt*GH8D281N statistics from data collection and processing. Values for the outer shell are given in parentheses. Tables as published in Fowler *et al.*[182]

	***Tt*GH8 Native**	***Tt*GH8 X2**	***Tt*GH8 X3**	***Tt*GH8*D*281*N* X6**
Diffraction source	Diamond Light Source	Diamond Light Source	Diamond Light Source	Diamond Light Source
Wavelength (Å)	0.98	0.98	0.98	0.98
Temperature (K)	100 K	100 K	100 K	100 K
Space group	P 2_1 2_1 2_1	P 2_1 2_1 2_1	P 2_1 2_1 2_1	P 2_1 2_1 2_1
***a*, *b*, *c* (Å)**	61.5 73.0 90.9	61.7 79.0 87.6	59.7 80.5 88.0	62.0 79.7 88.0
α, β, γ (°)	90.0 90.0 90.0	90.0 90.0 90.0	90.0 90.0 90.0	90.0 90.0 90.0
Resolution range (Å)	51.0-1.4 (1.42-14.0)	50.4-1.4 (1.42-1.40)	59.4-1.80 (1.84-1.80)	60.97 -1.80 (1.84-1.80)
Total No. of reflections	342816	530153	316157	313138
No. of unique reflections	80610	84123	40092	40046
Completeness (%)	99.4 (99.9)	99.1 (89.6)	100.0 (100.0)	99.0 (98.0)
Redundancy	4.3 (4.3)	6.3 (4.1)	7.9 (7.7)	7.8 (8.0)
⟨ *I*/σ(*I*)⟩	9.6 (1.7)	10.1 (1.3)	8.5 (1.6)	8.2 (2.0)
CC(1/2)	0.997 (0.670)	0.997 (0.503)	0.994 (0.528)	0.992 (0.639)
$R_{p.i.m.}$	0.045 (0.414)	0.051 (0.549)	0.080 (0.611)	0.109 (0.970)
Overall *B* factor from Wilson plot (Å^2)	9	9	13	12

Table 21 TtGH8 and TtGH8D281N mutant refinement statistics. Values for the outer shell are given in parentheses. Table as published in Fowler *et al.*[182]

	*Tt*GH8 Native	*Tt*GH8 X2	*Tt*GH8 X3	*Tt*GH8*D*281*N* X6
Resolution range (Å)	51.0-1.4 (1.42-1.40)	50.4-1.40 (1.42-1.40)	59.4-1.80 (1.84-1.80)	60.97 -1.80 (1.84-1.80)
Completeness (%)	99.4 (99.9)	99.1 (89.6)	100.0 (100.0)	99.0 (98.0)
No. of reflections, working set	80540	84054	40032	39993
No. of reflections, test set	3924	4274	1868	1942
Final R_{cryst}	0.15	0.16	0.17	0.17
Final R_{free}	0.17	0.18	0.20	0.20
Cruickshank DPI	0.052	0.051	0.111	0.116
No. of non-H atoms				
Protein	3163	3145	3125	3136
Ion	1	-	-	
Ligand	25	23	28	61
Water	310	330	164	194
R.m.s. deviations				
Bonds (Å)	0.017	0.016	0.012	0.012
Angles (°)	1.70	1.68	1.51	1.50
Average *B* factors (Å^2)				
Protein	14	12	17	16
Ion	18	-	-	-
Ligand	28	15	20	26
Water	27	22	24	22
Ramachandran plot				
Most favoured (%)	98.2	97.5	98.0	97.5
Allowed (%)	1.8	2.5	2.0	2.5
PDB Code	6G00	6G09	6G09	6G0N

The structure of 'apo' *Tt*GH8, displayed as ribbon view in **Figure** 112, shows the classical fold associated with the GH8 family, an $(\alpha/\alpha)_6$ barrel. The ribbon structure clearly shows the inner core of the protein surrounded by 6 separate alpha helices, each themselves connected by short β-sheet like linker regions. When looking at the surface of the protein, coloured by electrostatic potential in the same figure, one can see the deep binding groove along one face of structure, which in this case is the catalytic mutant *Tt*GH8D281N. The mutant structure is shown with X6 bound across the active site and resulted from 10 second mutant crystal soaks in a solution of X6. Previous attempts in which *Tt*GH8D281N crystals were soaked for longer amounts of time results in product complexes bound within the active site. One structure (data not presented) which resulted from crystals soaked in X6 for 15 minute, showed clear electron density for two molecules of X3 on either side of the 'catalytic' residues, indicating formation of a product complex. Crystal structures of the native *Tt*GH8 with smaller ligands, X2 and X3 were also accomplished and showed a preference of ligand binding in the negative subsites as shown in **Figure** 113. X2 was found to bind only in positions -3 to -2, X3 was found to bind -3 to -1, whereas X6 in the mutant structure was found to bind -3 to +3. Interestingly, binding of X3 caused movement of the catalytic residues Glu73, as shown in **Figure** 113 where the catalytic residues in ligand complexes are compared with those in the native. Several residues are thought to be involved in substrate binding, as shown in **Figure** 114, where 'nearest neighbours' (selected in CCP4mg[181]) within 4 Å to the bound X6 ligand are shown as cylinders and labelled; several aromatic residues are likely to provide ring stacking interactions, as well as hydrogen bonding interactions coming from a network of water molecules and hydrophilic residues. Binding of a sugar moiety in position -1 induced a conformation change in both the *Tt*GH8-X3 and *Tt*GH8D281N-X6 structures, to $^{2,5}B$ and $^{4}C_{1}$ conformations respectively, shown clearly in **Figure** 115.

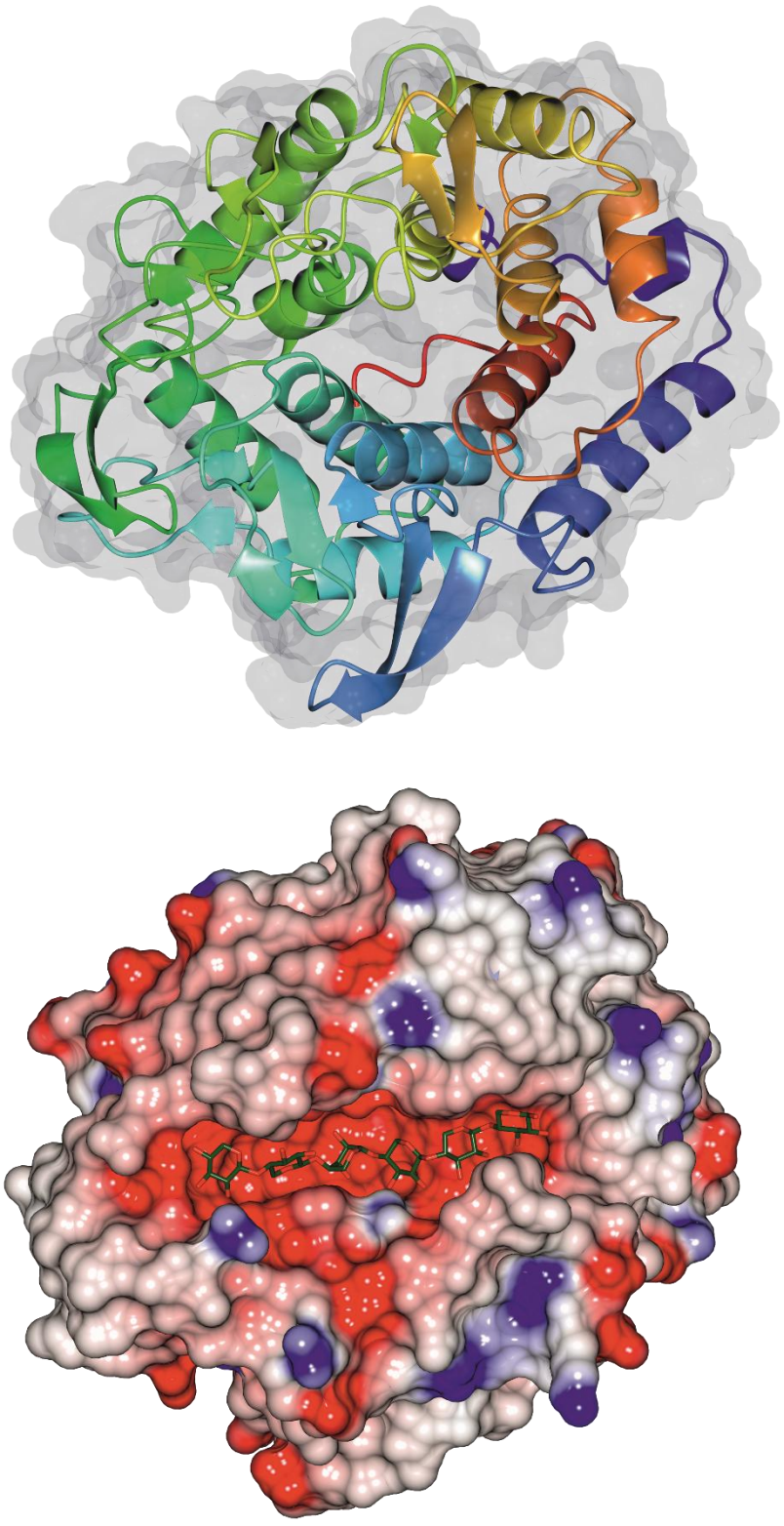


Figure 112 Ribbon structure of 'apo' *Tt*GH8, with surface shown as semi-transparent (above). Surface view, coloured by electrostatic potential of *Tt*GH8D281N in complex with X6 (below).

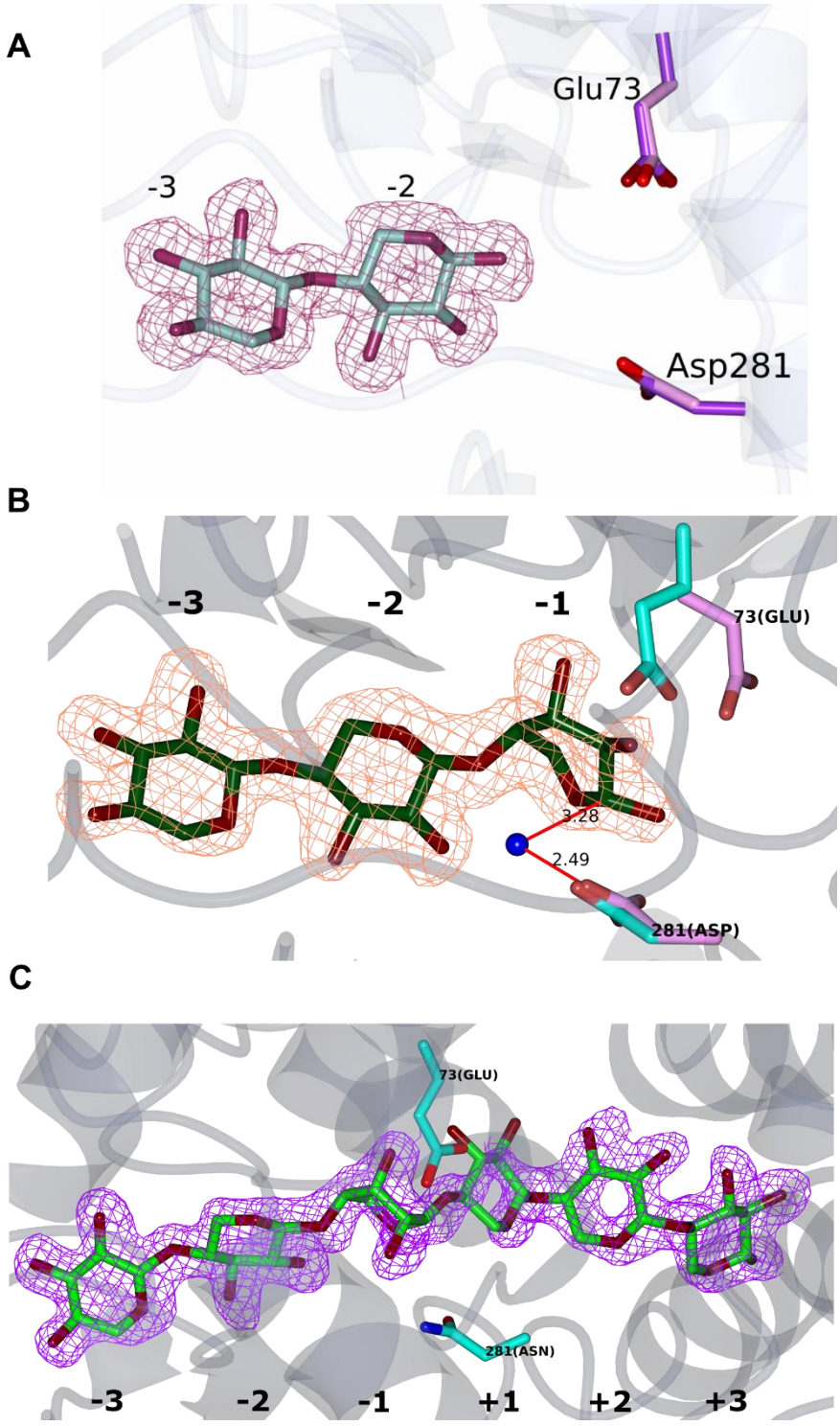


Figure 113 Ligand complexes of apo *Tt*GH8 and catalytic mutant, *Tt*GH8D281N with the corresponding maximum likelihood weighted Fo-Fc "difference" electron density, calculated prior to any incorporation of ligands in refinement, at contour level of 0.35 electrons / Å3 (approx. 2.5σ). A) Xylobiose complex with TtGH8, showing xylobiose bound in sites -3 to -2. B) Xylotriose complex with TtGH8, where the ligand is bound in the -3 to -1 sites. The catalytic residues in both the xylobiose and xylotriose complexs are overlaid with those in the apo structure, shown in purple and light pink respectively. Movement of Glu73 is observed in the xylotriose complex. C) Xylohexaose complex with TtGH8D281N, showing the ligand occupying all 6 subsites within the binding site, -3 to +3.

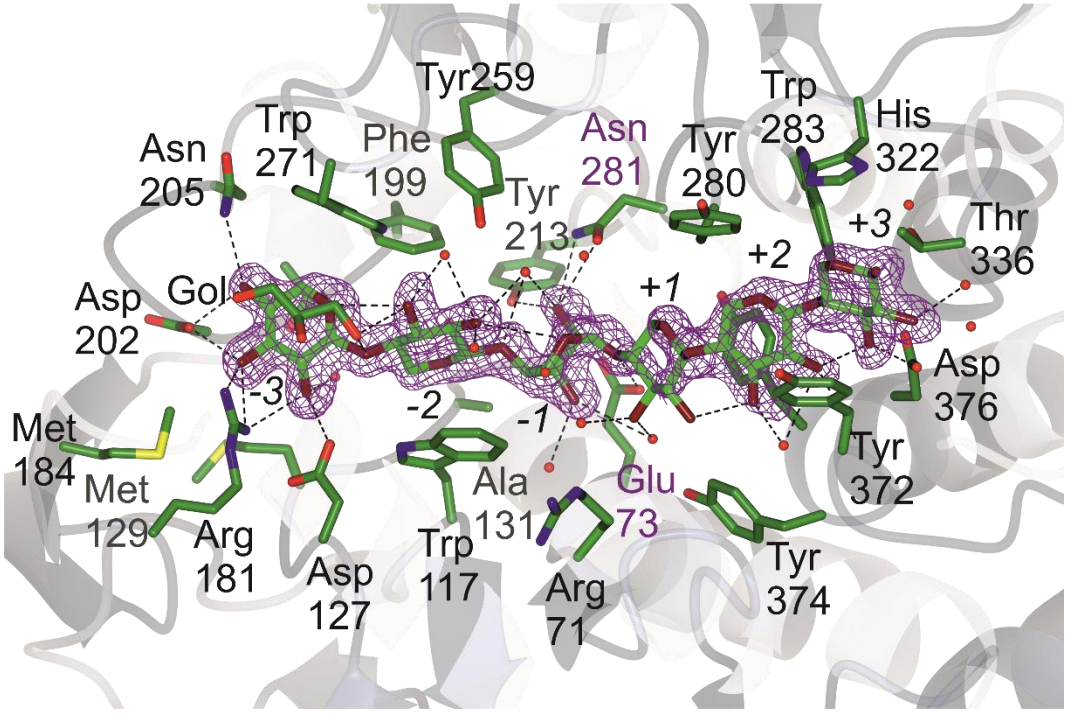


Figure 114 Nearest neighbours, within 4 Å to the bound xylohexaose ligand in complex with *Tt*GH8D281N. Hydrogen bonding to water molecules is also shown with dashed lines. The catalytic residue Glu73 and mutated Asn281 are labelled in purple. Subsites are labelled -3 to +3.

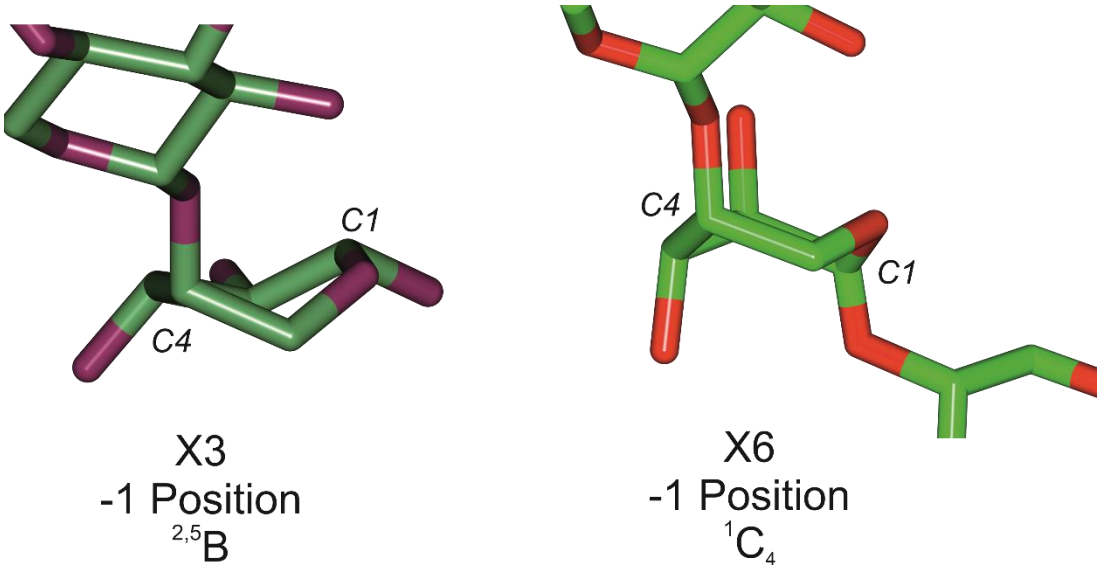


Figure 115 Close up view of the ligands, X3 and X6 and their conformation within the -1 subsite of *Tt*GH8 and *Tt*GH8D281N respectively. X3 forms a skew boat in the -1 position, whereas X6 takes the form of an inverted chair.

5.5 Discussion

5.5.1 Stability and Ligand Interactions

*Tt*GH8 was predicted to be associated with xylan degradation from sequence homology modelling using NCBI BLAST.[113] The enzyme was first tested using TSA to determine the protein stability with and without substrates. TtGH8 on its own displayed a clear protein melting curve, with a T_M of 57.2 °C. The enzyme was incubated with a broad array of polysaccharide substrates but the only significant increase was observed with BX, where the protein T_M increased by 2.1 °C. A similar increase was observed during the TSA assay when the protein was incubated with X6, whereby the melting temperature increased by 2.9 ° C. No other significant changes in melting temperature were observed for *Tt*GH8 mixed with polysaccharides such as cellulose. The GH8 family does contain members which display endo-glucanase activity,[34] but the lack of interaction of *Tt*GH8 on cellulose as shown by no observable changes in melting temperature suggested that the protein does not have this functionality. The catalytic mutants *Tt*GH8D281N and *Tt*GH8E73Q were also analysed by TSA and found to have significantly reduced 'apo' protein melting temperatures. The changes made to the catalytic residues appear to have seriously affected the stability of the protein, more so in *Tt*GH8E73Q than *Tt*GH8D281N which displayed melting temperatures of 48.2 °C and 53.0 °C respectively. Consistent change in 'apo' *Tt*GH8 Tm, incubation of the catalytic mutants with X6 had a large stabilising affect increasing the melting temperatures to approximately the same value as determined for *Tt*GH8 (*Tt*GH8 + X6,60.5 °C; *Tt*GH8D281N + X6,60.2 °C; *Tt*GH8E73Q + X6,63.5 °C). The large shift in melting temperature for all three protein samples suggests that X6 binds within the active site of the protein.

5.5.2 Activity Assays

In order to provide more experimental evidence for the activity of *Tt*GH8 on xylose based polysaccharides, TLC and LCMS assays were carried out. Hydrolysis reaction mixtures containing enzyme and various polysaccharides (bX, wAX, rAX, cAX, tXyG) were separated into soluble and insoluble fractions. When compared to the control samples, TtGH8 was able to solubilise and thus degrade bX, wAX and rAX as shown by products on the TLC plate. The enzyme was also tested against X2-X6 and found to degrade X4-X6. The hydrolysis reaction products of bX and X4-X6 were also evaluated using LCMS to determine accurate masses for the main products observed in the TLC analysis. The xylo-oligsaccharides were found to be

completely degraded into smaller products, whereby X4, X5 and X6 were hydrolysed into X3, X3+X2 and X3 respectively. Analysing the solubilised fraction of BX also showed X3 as the predominant oligosaccharide product; other products were observed on the TLC analysis, but these other soluble products were not as clear in the LCMS trace. These results indicated that *Tt*GH8 is able to hydrolyse xylan based polysaccharides with the dominant product being xylotriose.

As *Tt*GH8 had shown preferential activity towards xylans and on soluble xylo-oligsaccharides, kinetic parameters to assess the degree of enzyme efficiency were collected using HPAEC-PAD and the DNSA reducing sugar assay. The kinetic data analysed by HPAEC-PAD involved a method known as substrate depletion, whereby the amount of soluble substrate present during the enzymatic reaction over time is monitored. *Tt*GH8 was found to follow similar patterns observed with other GH8 enzymes with the same substrate specificity.[187] Whilst no activity was observed for X2 or X3, increasing enzyme activity was observed with progressively larger substrates; with values for k_{cat}/K_m determined for X4 as $6.1 \times 10^5 \pm 3.1 \times 10^4$ M^{-1} min^{-1}, X5 as $1.4 \times 10^7 \pm 1.4 \times 10^6$ M^{-1} min^{-1} and X6 as $7.5 \times 10^7 \pm 1.1 \times 10^6$ M^{-1} min^{-1}. Due to larger substrates being unavailable commercially it is unknown as to what point this trend ceases, but the difference in efficiency between X5 and X6 is less than the difference between X4 and X5. Therefore, one may assume a similar reducing trend between X6 and X7. Studies of other GH8 catalytic mutants had observed binding of 5 sugar units in the active site, whereas the kinetic data indicated that *Tt*GH8 preferentially bound X6, where the extra sugar unit likely provided a higher substrate binding affinity.

The kinetic efficiency of *Tt*GH8 on three polysaccharides was also determined using the DSA reducing sugar assay. The measured k_{cat}/K_m values showed a clear preference of the enzyme for the mixed linkage xylan substrate ($1.6 \times 10^8 \pm 4 \times 106$ mg^{-1} ml min^{-1}) which interestingly was a product isolated from a marine red algae. This was followed by bX ($1.8 \times 10^7 \pm 4 \times 10^6$ mg^{-1} mL min^{-1}) and then wAX ($6.3 \times 10^6 \pm 5 \times 105$ mg^{-1} mL min^{-1}). The catalytic mutant *Tt*GH8*D*281*N*, found to still retain a portion of its activity was also measured on MLX and found to display relatively efficient kinetics ($1.8 \times 10^4 \pm 1 \times 10^3$ mg^{-1} mL min^{-1}). Although the mutant structure displays kinetics at a significantly lower value that the native enzyme, in context this mutant was still able to hydrolyse X6 within 5 minute as monitored by TLC, and described previously.

The preference of the enzyme for the marine xylan (MLX) over the terrestrial xylan (bX) may be due to an adaption of the enzyme to available substrates within the environment of the

shipworm. In a discussion of shipworm larvae, Turner states that whilst shipworm larvae are quick to settle into burrows after extrusion from the adult, most wooden structures are covered in a 'protective forest' of various organisms including algae in which the young shipworm larvae may swim in before settlement.[84] Bacterial symbionts are passed onto shipworm young, so it is possible that before or during larvae settlement, algal particles are digested using enzymes such as *Tt*GH8. However, one must not be quick to jump to convenient conclusions. Little research into marine xylan polysaccharides has been carried out, so there are limited available resources to which the activity of *Tt*GH8 can be compared. The mixed linkage marine xylan used in this study is found as a component of the red alga, *Palmaria palmata,* a polysaccharide involved in mechanical support, development and defence.[188] Analysis of this polysaccharide suggests a ratio of 1:3 for β–1,3: β–1,4 moieties. Whilst pure β–1,4 bonding of xylose residues would result in a twisted ribbon structure, the irregular distribution of β–1,3 between variable lengths of β–1,4 sections may cause disruption.[188] Optical rotation alignment further suggests mixed linkage xylan exhibits a 'random coil' structure, unlike linear β–1,4 xylan which may form interactions with other chains, the presence of β–1,3 introduces flexibility which may assist in the solubility.[188-189] Flexibility may improve the fitting of the polysaccharide into the V-shaped binding site of TtGH8. Improvement in solubility due to the flexible nature of the xylan chain may be a factor in the increased degradation rate exhibited by *Tt*GH8.

5.5.3 Structural Insights

The structure of *Tt*GH8 was determined by molecular replacement at a resolution of 1.4 Å, using the protein coordinates only from the "reducing-end-xylose releasing exo-oligoxylanase" from *Bacillus halodurans* C-125 (PDB code 1WU4)[186] as the search model. *Tt*GH8 exhibits the classical $(\alpha/\alpha)_6$ fold found within the GH8 family, whereby there is a clear deep substrate binding grove across one face.[176] The suspected position of the catalytic residues was confirmed by comparing the structure of native *Tt*GH8 with other GH8 family members; the residues Glu73 and Asp281 were found to be in the expected conserved catalytic positions.[176] *Tt*GH8 crystals were soaked in both X2 and X3, and the ligand complexes refined at 1.4 Å and 1.8 Å respectively. X2 was bound in the –3 and –2 subsites whereas X3 was observed bound across all three negative subsites.[20] Structural analysis carried out on a GH8 from *Pseudoalteromonas haloplanktis*, an enzyme found to be a cold-adapted xylanase [178, 190-191] had revealed binding of X3 across the 3 positive binding sites, +1 to +3 subsites. Combining structural insight from both *P. haloplanktis* and *Tt*GH8 suggests

that there are in fact 6 binding sites available, -3 to + 3. This view is further clarified by structural study of an inactive variant (Asp144Ala) of *P. haloplanktis*, where a complex with X5 was captured by De Vos *et al*.[178] The 'inactive' catalytic mutant of TtGH8 was used to capture a structure containing xylohexaose bound across the whole active site. The *Tt*GH8*D*281*N* crystal soaking experiments with xylohexaose showed that, as observed in the activity assays, the mutant was still able to hydrolyse the substrate. The DNSA reducing sugar assay had indicated that *Tt*GH8*D*281*N* retained around 0.1% catalytic activity (k_{cat}/K_M of 1.8 x10^4 mg^{-1} ml min^{-1}), with MLX as substrate. Retention of activity upon mutation of the catalytic base is consistent with similar mutants on other GH8 systems.[178, 190] As such product complexes (data not shown) were observed in experiments consisting of crystal long soaking timescales with X6; X3 in the negative subsites in longer soaking times, or shorter 15 min soaks, separate X3 molecules bound either side of the remaining catalytic Glu73; -3 to -1 and +1 to +3. The successful soaking process involved submerging the crystal for no more than 10 seconds within a solution of xylohexaose, before immediate freezing. Thus, the TtGH8 Asp281-Asn structure in complex with X6 was obtained at 1.6 Å, with X6 bound across the substrate binding grove from -3 to +3, mapping out one extra binding position as compared with the cold adapted xylanase of *P. haloplanktis*.

5.5.4 Ligand Complex Mechanistic Insight

Hydrolytic cleave by GH8 family members leads to inversion of the anomeric configuration at the scissile glycosidic bond (as described in Chapter 1).[167] Comparison of the native *Tt*GH8 structure with that of the X3 complex, shows that binding of substrate induces rotation of Glu73 into a more catalytic relevant position. Glu73 interacts with the O1 hydroxyl of the sugar moiety in the -1 subsite whilst the catalytic base, Asp218 was found to be interacting with a water molecule "below" C1. The structural arrangement of the bound ligand in the -1 subsite and the positioning of both catalytic residues, and a water molecule mimics the pathway expected for hydrolysis with inversion of anomeric configuration. Furthermore, the –1 subsite sugar is not observed in its low-energy $^4C_1$ chair conformation, but is instead observed distorted to a $^{2,5}B$ conformation. This distortion to a boat conformation is consistent with the expected catalytic itinerary of inverting GH8 enzymes in which the catalytic mechanism moves through an oxocarbenium-like transition state.[23, 168, 192-193] The active site arrangement likely forces the -1 subsite bound sugar into a conformation closer to that of the transition state to favour catalytic movement towards product formation.

Indeed, similar conformations have been observed in ligands complexes of CelA endoglucanase[173], and supported by QM/MM metadynamics.[168].

5.5.5 Retention of Activity

Interestingly, the *Tt*GH8*D*281*N*-X6 complex, where the substrate was found spanning 6 subsites displayed a -1 conformation unlikely to assist in hydrolysis. In a similar occurrence the X5 ligand complex of GH8 from *P. haloplanktis* enzyme (PDB 2B4F) by De Vos [178] displayed an undistorted $^{4}C_{1}$ (although with scant density) sugar conformation in the -1 subsite. Furthermore the position of the catalytic acid was found in a similar position that Glu73 in the native *Tt*GH8 structure – a position not commensurate with the residues role as proton donor. Surprisingly, the conformation of the -1 bound sugar moiety in the *Tt*GH8 catalytic mutant was found to be a ring flipped to a southern hemisphere $^{1}C_{4}$ chair conformation (see Chapter 1 for discussion of conformational analysis of carbohydrates in terms of hemispheres). It is unlikely that this X6 complex with the mutant *Tt*GH8 reflects any mechanistic pathway that follows the 'traditional' GH8 inverting conformational itinerary. The $^{1}C_{4}$ chair conformation blocks the incoming nucleophile water, by positioning O2 in an axial position. Hence, whilst the Asp281Asn variant has allowed definition of the interactions of the –3 to –2 and +1 to +3 subsites well, it highlights the occasional dangers of using inactive variants to study substrate distortion in –1.

*Tt*GH8*D*281*N* was found to retain some activity on both polysaccharide and oligosaccharide substrates, and indeed there is already precedence for GH8 catalytic mutants retaining activity. The GH8 family is split into three subfamilies, one of which, GH8b members lack the expected Asp catalytic base and activity is thought to occur through a nearby Glu residue.[175] Removal of the catalytic base in *Tt*GH8 may have caused a shift in which residue is carrying out hydrolysis. The large reduction in activity compared with the native suggests that the change is very unfavourable, potentially due to the flipped chair conformation observed in the X6 complex, or the positioning of the acting base. A highly hydrogen bonded water, sitting above and slightly left of the scission site can be observed in line with Tyr259 in the structure of *Tt*GH8*D*281*N*-X6. Guerin *et al* noted that a tyrosine residue close to the catalytic residues in CelA was able to interact with the nucleophilic water via hydrogen bonding and this residue was conserved in the majority of GH8 enzymes.[173] This may be a possible route to catalysis whilst the use of the 'normal' catalytic base (Asp281) is not possible, with the tyrosine residue acting as a poor base, as suggested by Guerin those GH8 enzymes which lack a catalytic base display similar activity despite being an 'inactive mutant'.[173]

5.6 Conclusion

The single GH8 enzyme found in the genome of *T. turnerae* has been characterised as a highly efficient endo-xlyanase (k_{cat}/K_m for X6; 7.5 x 10^7 ± 1.1 x 10^6 M^{-1} min^{-1}, for MLX; 1.6 x 10^8 ± 4 x 10^6 mg^{-1} ml min^{-1}), with maximal activity observed on a marine xylan polysaccharide known to consist of both β (1-4) and β (1-3) linkages. Due to unavailable comparisons in the literature, it is unknown whether *Tt*GH8 is highly specific for mixed linkage marine xylans, or displays a higher activity due to the increased soluble nature of the polysaccharide as a result of its intrinsic linear structure. Structural analysis of native enzyme and ligand complexes showed a shift in the catalytic Glu73 residue upon binding on X3. The X3 ligand was in the ^{2,5}B conformation in the -1 binding site, consistent with the catalytic itinerary for an inverting GH8 enzyme. A structure with xylohexaose was accessed through mutation of the catalytic base, and found to display an unusual ring flipped chair conformation in the -1 binding site. Furthermore, the catalytic mutant maintained some residual activity, meaning it was still capable of hydrolysing the substrate through an unknown catalytic pathway. In conclusion, this highly efficient GH8 is adept at depolymerisation a variety of different xylan substrates and due to its ease of production and stability could be a useful asset in industrial biomass degradation.

6

Characterisation of *Tt*AA10

6.1 Abstract

Lytic polysaccharide monooxygenase (LPMO) is a relatively new class of lignocellulolytic enzymes, which are capable of improving insoluble polysaccharide degradation through copper dependent oxidative chemistry. LPMO enzyme discovery and research is a constantly and fast evolving field, where novel enzymes and new insights are providing tools for future efforts towards utilising the full energy benefits of waste lignocellulosic materials. A gene which displays an N-terminal histidine, noted in the genome of *T. turnerae,* an endosymbiotic bacterial found within the gills of marine bivalve molluscs was proposed to be an LPMO, subfamily AA10. The LPMO, *Tt*AA10 (ACR14100.1) is the single representative of this CAZYme family within the genome of *T. turnerae*, and has been characterised through this work as able to degrade crystalline cellulose. Growth of a *T.turnerae* culture on a cellulose substrate was found to cause over expression of native *Tt*AA10 by secretion into the media. Subsequently, the activity of recombinantly produced *Tt*AA10 was analysed by MALDI-TOF-MS and the enzyme found to be active on PASC and Avicel when mixed with a reducing agent, with major product peaks between DP 5-8. Both native oligosaccharides and oxidised products were observed, with mass spectrometry indicating that *Tt*AA10 is able to cleave a glucan chain at both C1 and C4 positions around the scissile bond. TTAA10 therefore belong to AA10 subgroup Type 3. *Tt*AA10 was found to bind to crab (alpha) chitin but not degrade the substrate, and follows the hypothesis that the AA10 subfamily is moving away from chitinolytic activity. The overall structure of *Tt*AA10 is significantly different to other LPMOS, with the highest AA10 sequence homology being 30 %. As such *Tt*AA10 was solved at 1.4 Å by copper phasing; the C-terminal strep tag provided a crystal contact through formation of a secondary copper site, enhancing the copper signal which allowed

for structure phasing and solution. *Tt*AA10 displays a typical Histidine brace active site, as shown by EPR and structural analysis, albeit with a glycine replacing the normal AA10 alanine residue observed in the apical position above the active site copper. This work has analysed the only LPMO encoded in the genome of a symbiotic bacteria from a host animal shown to be able to survive off a diet consisting of only lignocellulose, despite not harbouring any other bacterial microorganisms in its digestive system. Characterisation of *Tt*AA10 will add to the ever growing collection of knowledge and chemical insights towards understanding this highly important class of enzymes.

6.2 Introduction

6.2.1 Lytic polysaccharide monooxygenases (LPMOs)

LPMOs augment the functions of their lignocellulosic degrading counterparts, GHs. LPMOs provide an oxidative mechanism to break up the most resilient, highly ordered, crystalline polysaccharide structures such as cellulose and chitin, as discussed further in Chapter 1.[39, 194] LPMOs are a family of proteins which are only related by a set of highly conserved features, which allow the proteins to use oxidative chemistry on the surface of polysaccharide structures, breaking glycosidic bonds to produce breaks in the otherwise ordered structures. The most important of these features is the active site, commonly known as the Histidine brace.[195] The Histidine brace creates the perfect scaffold for T-shaped coordination of a copper ion. The brace further positions the metal on the "flat" surface of the protein, in a prime location to interact and react with chains on the polysaccharide surface.[65, 195-196] In the same way as GHs, LPMOs are classified by sequence homology in the CAZy database and are described within a set of auxiliary activity (AA) enzymes.[34] There are currently 6 published subfamilies of LPMOs; AA9, AA10, AA11, AA13, AA14 and AA15 which were described in Chapter 1.[51] Each family is extremely different in amino acid sequence and sequence homology. However, all contain the specific and invariant features of an LPMO to enable it to carry out oxidative chemistry on the polysaccharide chain.

6.2.2 Auxiliary Activity 10

This Chapter will focus on an AA10 LPMO. The AA10 subfamily is large, containing at present (September 2018) 3623 entries, of which only 21 have been structurally characterised.[34] There are no AA9 or AA11 enzymes found within bacterial genomes (they are mostly fungal)

and LPMO proteins classified into the AA10 subfamily originate largely from bacterial sources. A smaller subset is found in viral genomes whilst a couple of sequences have been identified as coming from eukaryotes and archae.[34] Phylogenetic analysis by Book *et al* created a clear pictorial representation of the extent to which there is no cross over between AA9 and AA10 proteins and by using an extremely relaxed similarity threshold, it was shown that there is no sequence link between the two LPMO subfamilies, even at the edges of the phylogenetic tree.[197] AA10 proteins can be active on either chitin (AA10 clade I) or cellulose (AA10 clade II), with the latter likely to be a more 'recently' gained functionality considering that chitin as a substrate has been present in oceanic systems before the formation of cellulose structures in land plants.[197-198] Book's analysis supports this interesting hypothesis, suggesting that whilst there was a negative selection pressure for alteration to the key features of AA10 proteins, notably the active site, within the AA10 subfamily there is currently underway a positive selection pressure to move towards cellulolytic function over chitinolytic.[197] Whilst the AA9 and AA10 proteins may share a common, albeit ancient ancestor, it is clear that the continual change in function overtime further separates the two families, whilst maintaining their primary function.

In an overall sense, the structures of the different LPMOs families are similar in that they create a near planar surface involved in the binding of the substrate. Almost at the centre of this flat surface, a molecule (e.g oxygen) bound to the Cu active site is perfectly positioned to react with the C-H bonds of the polysaccharide chains.[43, 199] The tight binding of the copper (a K_d of 43 nM ± 2 nM at pH 5.0 was measured for BaAA10 using displacement isothermal titration calorimetry)[61, 65] is provided by the Histidine brace, where positioning of two histidine residues forms a T-shaped geometry in which the Cu is coordinated by three nitrogen atoms; two ring nitrogen atoms from the side chains and a single nitrogen from the amino terminus, as described in Chapter 1. The coordination sphere of the Cu is completed by a fourth equatorial (exogenous) ligand, typically a water/hydroxide molecule. In the apical position, a further water molecule can be observed in the Cu(II) structures, while a conserved Tyr occupies the other axial position in the AA9, AA11, AA13, AA14 and AA15 classes. In contrast, some AA10 active site lacks this conserved tyrosine residue, which is instead replaced by a phenylalanine. Furthermore, AA10s present a conserved Ala residue positioned above the copper ion. As a consequence of the steric hindrance given from this residue, the apical water molecule is tilted away from the normal to the His brace plane, as observed in the Cu(II) structures and confirmed by EPR studies. The difference in surrounding residues changes the coordination of the copper-ligand sphere compared with

AA9s, which is evident during EPR analysis[65] EPR analysis simulates spin Hamilitonian parameters which can be used to assess coordination around a metal ion such as Cu. Analysis of the parameters measured in the Cu centre of many AA10 proteins have shown separation into two broad classes. The first class exhibits a high value of A_z (similar to AA9 proteins) which is consistent with more axial coordination geometry. The second class presents more rhombic g values, with a reduction in the A_z component suggesting that the active site geometry around the Cu is distorted away from axial. Comparison of AA10s to AA9 spectra, with those AA10 proteins acting on cellulose found to be more similar in their EPR copper signal to AA9s, than AA10s which are active on chitin.[200] Mutation of the phenylalanine in active site of *Sc*LPMO10C (CelS2) to a tyrosine still maintained copper binding, but negatively affected the activity of the protein and thus failing to mimic a partial AA9 active site (as the axial water position would still have been blocked by the alanine). It is likely that other subtle changes beyond the differences in apical residues are also important in producing an 'active' active site.[200] Differences in active site configuration may be indicative of differences in enzymatic mechanism. Crystallography studies of *Ls*AA9, which is active on soluble cellooligosaccharides, combined with EPR show binding of substrate to the surface of the LPMO. The binding of cellotriose causes alterations in the copper coordination sphere, as observed in the structure and by EPR. Substrate binding induced movement of the so called pocket water, away from the normal axial coordination position, where it was then linked within a hydrogen bonding network between substrate and the amino terminus of His1, ultimately linking to the Cu ion, **Figure** 116.[201] In high salt conditions, binding of chloride was observed as an equatorial ligand to the Cu and was found to lower the dissociation constant of cellohexaose (3.7 ± 0.1 μM, measured by ITC) to the surface of the LPMO. The changes in coordination sphere were reflected in the measured EPR parameters, with g_z shifting from 2.28 in the native protein, to 2.23 in the presence of substrate and chloride. A similar shift in EPR values was observed upon mixing the LPMO with insoluble crystalline cellulose, indicating that the protein likely works in a similar way on both substrates, potentially attacking exposed corners of the substrate.[201] Small differences in the active site and binding mode may prove to be important factor in determining the mechanistic pathways utilised by different LPMO families. Indeed, shifts in EPR parameters, or lack therefore, can provide insight into the binding interaction of substrates inaccessible by crystallography.

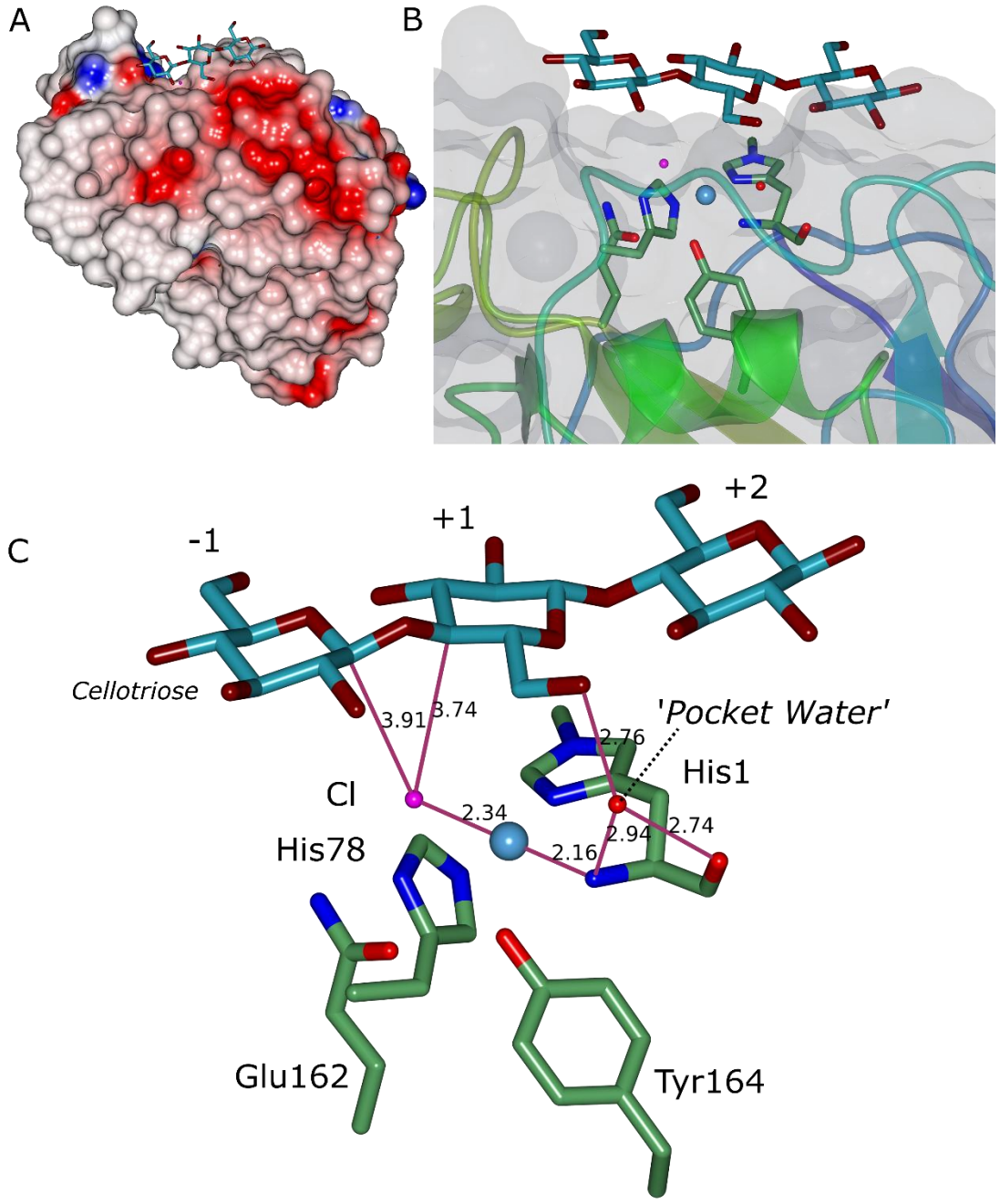


Figure 116 Structure of LsAA9 in complex with cellotriose (PDB code: 5ACF), taken at a low X-ray dose. A) Surface structure showing contours at the binding surface. B) Position of substrate over active site. C) Detailed view of Cu active site, showing presence of Cl positioned 2.3 Å away from copper and 3.9 and 3.7 Å away from C4 and C1 respectively (see **Chapter** 1 for description of C4 vs. C1 oxidation). The so called pocket water, has been shifted out of its normal apical binding position into a hydrogen bonding network linking the substrate to the Cu through the water's interaction with the N-terminus.[201] Image produced in CCP4mg.[55]

The near-planar surface of the LPMO is the region of the protein which interacts with the substrate. Unlike GHs, which often have a groove or cleft able to isolate a single substrate strand, the flat binding surface of most LPMOs allows it to dock onto the ordered, extended regions of polysaccharide surface. The usually flat surface is formed by one end of the β-sandwich and the top of the single helical bundle, with the active site in a central position. In AA10s a conserved set of residues sits either side of the copper active site, likely inducing a substrate specific orientation through hydrogen bonding interactions and aromatic residues.[65] Indeed, several site directed mutagenesis experiments have shown how mutation of certain conserved surface features reduced LPMO function.[53, 61] Recent computational modelling by Bissaro *et al* also suggests that the surface is responsible for the precise positioning of the activated Cu-oxygen intermediate close to the C-H bond to be broken; the modelling of chitin active *Sm*AA10a mimics the stereochemical cleavage observed during activity assays.[202] AA10 family members have been found to degrade cellulose and chitin and as mentioned previously the family can be split into two clades based on this substrate preference.[197-198] Sequences alignments have shown that chitin active AA10s are more similar to each other than they are to cellulose active AA10s, and vice versa. Small differences in the planar region surrounding the active site most likely contribute to the substrate specificity of AA10s.[200]

Both AA9 and AA10 LPMOs are able to act upon substrates in a regioselective manor, and as summarized in **Table** 2 (Chapter 1), the clades can also be split into type. Cleavage of substrate by LPMOs has been shown to occur at positions C1 and C4 (described in **Figure** 13, **Chapter** 1) – some LPMOS are only capable of cleaving at C1 or C4, whereas others seem to lack specificity and have been shown to cleave the same substrate at both positions. Hence, AA10s have been shown to have the following activites, described as four types;[71]

- Type 1A – cleavage of chitin at C1
- Type 1B – cleavage of chitin at C1
- Type 2 – cleavage of celloluse at C1
- Type 3 – cleavage of celloluse at C1 and C4

As of yet, no AA10 enzyme has been shown to only cleave at the C4 position, unlike in Type 2 AA9s. Type 1A and Type 1B are split based on phylogenetic analysis which shows the only Type 1B member, *Cj*LPMO10A, as a distinct branch point from other chitin degrading AA10s.[71]

Catalytic breakage of glycosidic bonds by LPMOs functions through an oxidative mechanism, however, reduction of the copper by an external reducing agent is required before activation of dioxygen (or another substrate such as H_2O_2) and subsequence catalysis can occur. Fungi use AA9s to illicit changes to substrate morphology, before using GHs to break down the substrate further. In some fungal systems, cellobiose dehydrogenase (CDH) is upregulated along with AA9s in response to presence of substrate[203-204], and is able to act as a reducing agent towards AA9s, leading to a possible conclusion that this protein is the natural reducing agent of fungal LPMOs.[58, 205] The natural function of CDH converts disaccharides (or larger oligosaccharides) into their corresponding aldonic acids via an oxidative mechanism. CDH is a two-domain protein consisting of a flavin adenine dinucleotide (FAD) binding dehydrogenase (DH) in complex with a heme binding cytochrome (CYT).[205] CDH can store electrons in the DH domain, due to reduction of FAD and transfer them via a 2 e^- reduction step to a redox partner during a reaction. Alternatively, CDH can perform a 1 e^- reduction via internal transfer of an electron to the iron held within the CYT domain, which can then be subsequently transferred to a redox partner such as an LPMO.[205-206] The single electron transfer ability of CDH is thought to couple with the reduction of the LPMO active site from Cu(II) to Cu(I). The 1 e^{--} reduction of copper primes the system for activation of O_2 into an Cu(II) oxygen intermediate capable of oxidative cleavage of the glycosidic bond. The hydroxylation and subsequent elimination steps result in recycling of the copper site back into the Cu(I) form and breakage of the glycosidic bond.[206] Interestingly, CDH is only found in fungal systems and an equivalent protein in non-fungal organisms have not yet been identified. The redox activity of CDH can be mimicked with small molecule reductants such as ascorbate (Vitamin C) or gallic acid which are commonly used in the study of LPMO activity. However, use of small molecule reductants has issues. Auto-oxidation can occur and lead to reductions in known concentrations, and may result in damage to the LPMO or surrounding enzymes through formation of reactive oxygen species. Furthermore, addition of a small molecule reducing agent to an LPMO may cause damage unless substrate is already present. In nature, small molecule reductants may be present in lower quantities which would limit damage over time and whilst introduction of excess reducing agent may not mimic natural conditions effectively this process is likely to be used in the industrial enzymatic degradation of biomaterials. Experiments which identified the oxidative action of LPMOs did not use small molecule reductants, which suggests that LPMOs have the potential ability to recruit electrons from their substrates.[52-53, 206]

6.2.3 *Tt*AA10 Domain architecture

There is only a single LPMO found within the genome of *T. turnerae* and sequence analysis predicts it to be homologous to other AA10 protein sequences, with the closest match being *Cj*LPMO10A, a chitin active AA10 from *Cellvibrio japonicas*, Figure 117. LPMOs are known to exist as single domains or be fused with other activities, such as CBMs and often domains with unknown functions (termed X-domains). AA10 proteins are often associated with binding domains from CBM2, 3, 5 or 12; with CBM2 and 3 conferring cellulose binding, whilst 5 and 12 are known to interact with chitin.[207] The full length gene sequence of the LPMO in *T. turnerae*, *Tt*AA10, predicts a domain architecture containing a signal peptide, the AA10 LPMO domain and a linker region followed by a CBM10 domain. The CBM10 domain is thought to be cellulose binding or provide cellulose recognition that is unlikely to be associated with a catalytic event, [208] and whilst different from the CBMs commonly found attached to AA10 proteins, [197] its presence in the domain structure of the *Tt*AA10 gene gives a strong indication that this protein will be primarily active on glucose based polysaccharides.

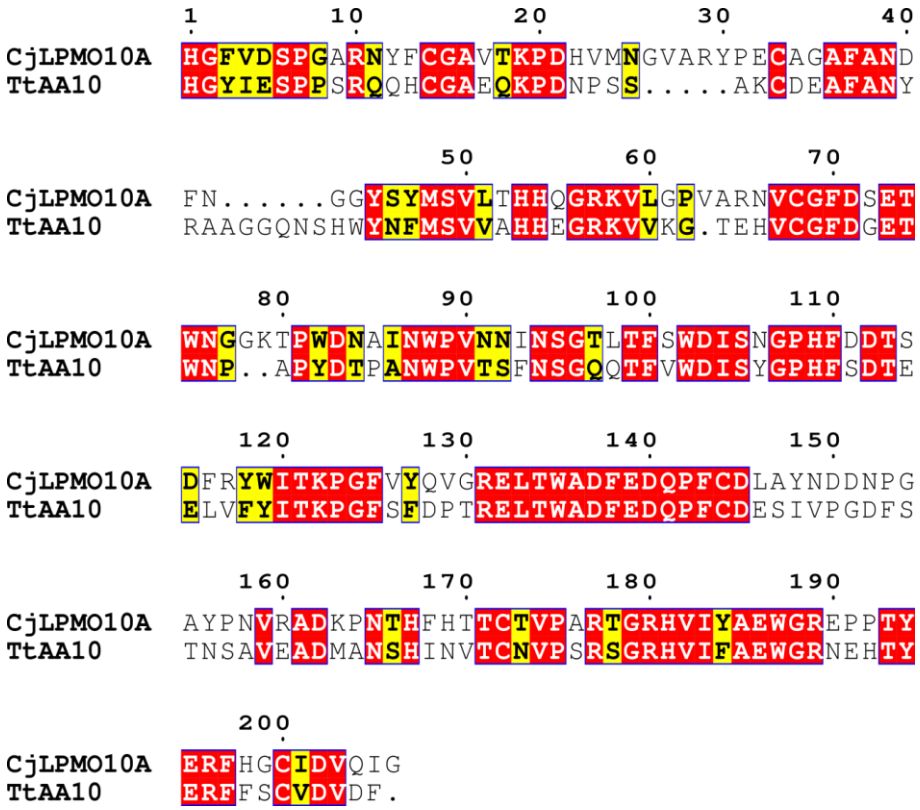


Figure 117 Sequence alignment (T-coffee) of *Tt*AA10 against *Cj*LPMO10A formatted using ESpript. *Cj*LMPO10A is currently the sole member of the AA10 subfamily Type 1B. *Tt*AA10 shares 50% sequence homology with *Cj*LPMO10A, which is the closest sequence match to *Tt*AA10 using NCBI BLAST.

6.3 This Work

The catalytic LPMO domain from the *Tt*AA10 gene was synthesised in *E. coli,* in several batches, as shown in Chapter 2. The work presented herein describes the characterisation of *Tt*AA10 in terms of both structure and function, testing the hypothesis that *Tt*AA10 would display activity on cellulose, based on the positioning of a cellulose binding domain within the gene open reading frame. EPR spectroscopy was used to assess the copper containing Histidine brace active site. The stability of the protein with and without copper as well as the effect of introducing a reducing agent was assessed by TSA. The activity of the protein is analysed by MALDI-TOF-MS, to find out what type of cellulose *Tt*AA10 is specific for and as such carry out product analysis to determine the AA10 subgroup. Finally, the structure of the protein is determined and used to compare with other AA10 structures in the literature.

6.4 Methods

6.4.1 Materials

Pure recombinant protein, *Tt*AA10 (ACR14100.1) was produced as described in Chapter 2, using the construct *Tt*AA10-strep, henceforth written only as *Tt*AA10. Substrates used in this work were Avicel (Sigma), PASC (Sigma), Carboxymethylcellulose (CMC), crab chitin (Sigma), squid pen chitin, and bMLG. *T. turnerae* (strain T7901, ATCC 39867) was purchased from the American Type Culture Collection (ATCC) through LGC Standards Limited.

6.4.2 *T.* turnerae Culture Growth

Growth media (100 mL) was made using LB (10 mL, 10 X stock), seawater (80 mL) (origin, Gulf of Mexico, sterilised, Sigma-Aldrich), Sigmacell Cellulose type 101 (10% *w/v*, 10 mL) (Sigma-Aldrich). Media was autoclaved prior to addition of $MgSO_4$ (1 mM final concentration) and $CaCl_2$ (1 mM final concentration). *T. turnerae* was introduced to the media and cultures grown at 35 °C, 50 rpm for 5 days. Cells were harvested by centrifugation and the media collected. Samples of the media and cell pellet (soluble fraction) were sent for proteomic analysis at the Technology Facility, Department of Biology, University of York. A pull down experiment was carried out by growing 50 mL of culture which was separated into media and pellet after 5 days growth at 35 °C. Cell pellet was resuspended in buffer (1 X PBS, pH 7.4), lysed and the supernatant collected by centrifugation. Both media and lysed cell pellet samples were loaded onto individual strep columns (Strep Trap HP, GE Healthcare) pre-bound with recombinant *Tt*AA10 (1 mg total). Columns were washed with 1 X PBS buffer and *Tt*AA10 eluted with desthiobiotin (20 mL, 2.5 mM in 1 X PBS, pH 7.4) (Sigma-Aldrich). The eluted protein was concentrated down and all samples collected run on SDS PAGE.

6.4.3 EPR

Frozen solution EPR spectra of *Tt*AA10 were collected at 160 K on a Bruker EMX continuous wave X-band spectrometer operating at approximately 9.3 GHz, modulation amplitude of 4 G and 10.02 mW microwave power. Protein samples used during single EPR analysis were all in the concentration range 100-200 µM, in 1X PBS buffer at pH 7.4. Simulations of the collected spectra were carried out in Easy Spin 5.2.6[209] integrated into MatLab 2016a software by Dr Luisa Ciano to determine the *g* and A-tensor parameters. EPR of different

batches of protein shows that some samples contained two copper species, whilst other samples contained only one. EPR Copper titrations ($CuSO_4$ 1 M) were carried out on protein free of copper (10 mM EDTA treated, followed by extensive buffer exchange), with spectra taken before the titration was started confirming the lack of coordinated copper in the protein. Copper was added to the protein solution whilst contained within the EPR tubes, in additions of 0.2 equivalents (to the concentration of the protein in the sample). Spectra were measured after each addition of copper solution. EPR was also used to monitor metal displacement of the second copper species. The sample obtained at the end of the previous copper titration experiment was used as it was known to contain the second copper species. 0.5 to 1.5 equivalents of Fe, Ni, Zn and Mn (from high concentration metal salt solutions) were added one after another and EPR spectra taken in between additions to monitor any changes in the copper signal. EPR spectra were also taken on *Tt*AA10 samples that had been mixed with PASC or chitin.

6.4.4 TSA

*Tt*AA10 (43 µM) was tested by TSA using Sypro-orange as described previously. Incubation of the protein with EDTA (10 mM), ascorbate (10 mM), gallic acid (10 mM) and solid substrates (PASC, Avicel and chitin) during TSA was also carried out.

6.4.5 Activity Assays

*Tt*AA10 (50 µM) was tested on PASC and Avicel at pH 7 using gallic acid (1 mM). Two protein batches were used, one of which contained the second copper species, and reaction products analysed by TLC (run as described previously). Further reactions of *Tt*AA10 (0.1 µM) with substrate and ascorbate (1 µM) was analysed by MALDI-TOF MS (as described previously).

6.4.6 Crystallography and Structure Solution

*Tt*AA10 was screened for crystallisation hits using Hampton HT and PEG/Ion commercial screens. A hit was obtained in the PEG/Ion screen (well B8) with the condition; 0.2 M magnesium formate dehydrate, 20 % *w/v* PEG 3350, pH 7. Conditions were optimised by changing the concentration of both magnesium formate dehydrate and PEG 3350. Crystals were tested in-house for x-ray diffraction before being sent to the Diamond Synchrotron, using beamline I04. Data collected was collected at a maximum resolution of 1.17 Å and

processed in the CCP4i2 pipeline. Data was reduced using Aimless [210] to 1.4 Å in order to improve data completeness. Copper phasing was used to produce a structure solution using SHELX [211], followed by model building in ARP/wARP [212] and CRANK2. Repetitive iterations of manual model building in COOT [159] and refinement by REFMAC [160] were carried out.

6.5 Results

6.5.1 Culturing *T.turnerae* with cellulose as a food source

A small culture of *T.turnerae* was successfully grown using nutrient supplemented sea water and a cellulose source. After 5 days growth, the media and cell pellet were collected and protein content analysed. Proteomics identified 58 sequences corresponding to known *T. turnerae* genes within the cell pellet and 130 sequences secreted into the growth media. *Tt*AA10 was found secreted into the growth media, likely in response to the availability of a cellulose substrate in the media. Protein hits and their associated accession codes and potential functions (as taken from NCBI or Uniprot) are shown in Tables 25 and 26, Appendix 5.

A pull down experiment was designed to elucidate any potential protein binding partners of *Tt*AA10. *Tt*AA10 was bound to a Strep Trap HP column and soluble protein from the lysed *T.turnerae* cell pellet flowed over the column. The bound *Tt*AA10 was eluted using desthiobiotin and the collected fractions analysed by SDS PAGE (not shown). The SDS PAGE analysis showed only elution of *Tt*AA10 and no other protein bands could be observed on the gel. This experiment indicated that no protein within the cell pellet bound to the immobilised *Tt*AA10.

6.5.2 EPR Analysis of the Copper Active Site

The catalytic domain of the *Tt*AA10 gene was recombinantly expressed using a C-terminal strep tag as described in Chapter 2. During production of the recombinant protein, the sample was incubated with an excess of copper to ensure that the protein active site was fully saturated. This step seemingly caused an ambiguity between different batches of protein. The first batch of pure protein (made by Luisa Elias, University of York) which was used for crystallisation and initial activity assays contained two different coordination environments for the copper ion as shown by EPR, Figure 118. The EPR trace clearly shows two sets of hyperfine peaks in the parallel region of the spectrum, indicating either flexibility in the coordination of the His brace or the presence of a second coordination site. Simulation of the EPR spectra was carried out by Dr L Ciano, whereby the spin Hamiltonian tensor values in the parallel direction (z) for the first species were, $g_z = 2.269$ and $A_z = 405$ MHz. The second species displayed different values, with $g_z = 2.314$ and $A_z = 465$ MHz. Simulation indicated that the ratio between species 1 and 2 was approximately 3:2. The g_z

value of the second species is high compared to what one might expect for the typical LPMO copper coordination in the active site.

Subsequent batches of protein did not have this second copper species present at the end of the purification and it still remains unclear what caused the first incidence; **Figure** 119 shows the EPR analysis of a subsequent batch of protein, made by myself, which consists of a only single copper species in the EPR spectrum (no specific changes were made to the purification protocol which yielded a different metal content). The simulated parameters (Dr L Ciano) were as follows; g_x = 2.035, g_y = 2.100, g_z = 2.267; A_x = 110 MHz; A_y = 80 MH_z and A_z = 420 MHz (parameters in the xy region are estimates due to the broad nature of the perpendicular region and should not be taken as accurate). The data was simulated to include 2 nitrogen atoms coupling to the copper ion. Comparison of the g_z value from this analysis with the previous sample containing two copper species suggests that it is the same as species 1. The Cu active site in *Tt*AA10 is most similar to that of the closest homology match, *Cj*LPMO10A, **Table** 22.

Table 22 Comparison of EPR spin Hamiltonian parameters between *Tt*AA10 and its closest homologue. The perpendicular region, XY are italicized in some instances and values are estimates only.

LPMO (Substrate)	Parameters			*G, x10^{-4} cm^{-1}*		
	g_x	g_y	g_z	A_x	A_y	A_z*
*Tt*AA10(species 1)(cellulose)	*2.035*	*2.100*	2.267	*39.3*	*28.6*	150
TtAA10 (species 2)			2.314			166
*Cj*LPMO10A[213]	*2.036*	*3.091*	2.260	*31.8*	*26.1*	154

*A values have been converted to G (Gaus), where the TtAA10 parameters in MHz were divided by 2.8 to give an approximate value in G for comparison with literature values.

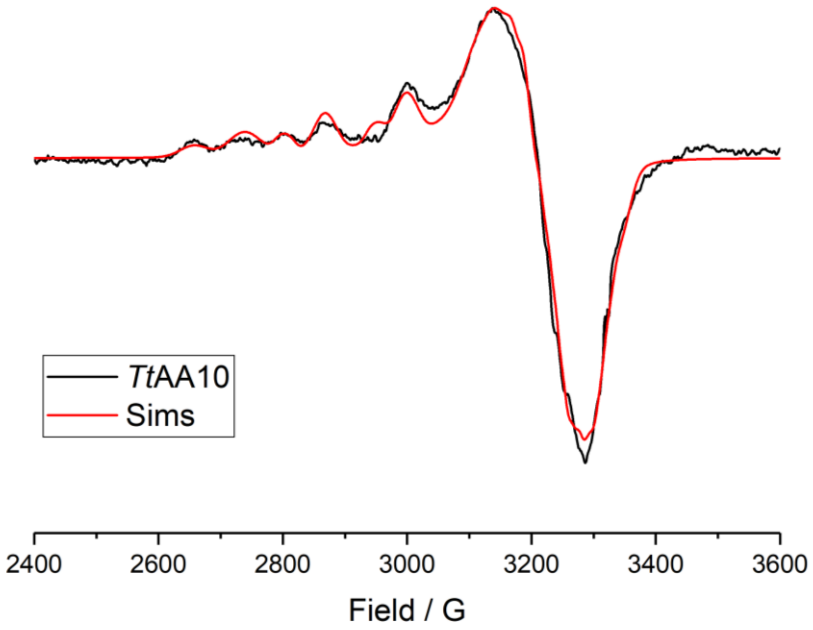


Figure 118 EPR analysis of *Tt*AA10 which shows that the sample contains two different copper species in different environments. The black line is the observed signal spectrum whilst the red is the simulation (carried out by Dr. L Ciano).

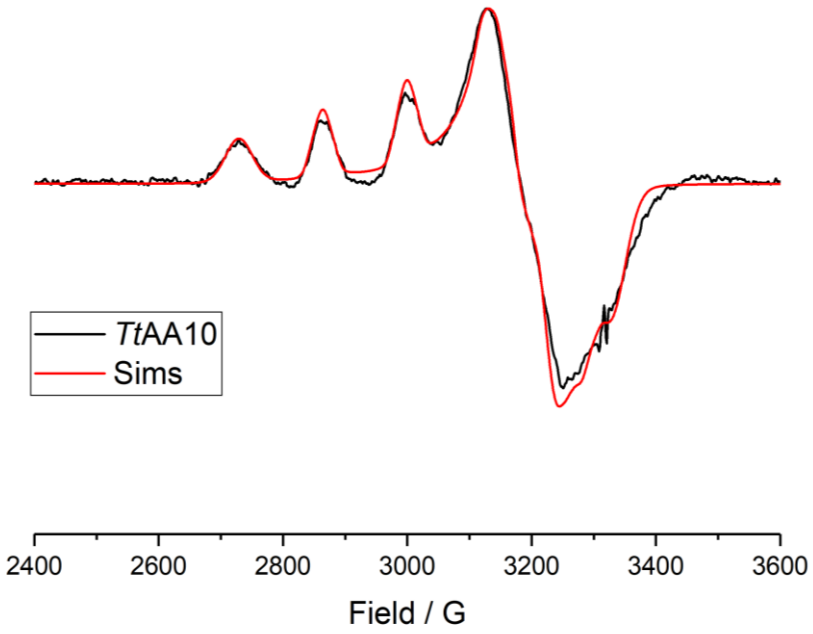


Figure 119 EPR analysis of a different batch of *Tt*AA10, where the black line is the observed signal and red line the simulation (carried out by Dr. L Ciano, University of York). The signal shows the presence of one copper ion, with spin Hamiltonian parameters consistent with what previously observed for a LPMO histidine brace coordination sphere.

A copper titration EPR experiment was designed to test at what point a second copper species becomes present in the EPR signal, Figure 120. The protein was pre-treated with EDTA (10 X the protein concentration) to remove the copper and buffer exchanged to remove any EDTA. This copper-free protein sample was tested and showed no copper signal in the EPR. Addition of 0.2 equivalents of copper (compared to the protein concentration) to the protein sample showed a single signal, likely to be due to coordination to the Histidine brace active site. Further additions of copper increased this Histidine brace copper signal, but a second species also started to grow in, which is most evident after 0.6 equivalents of copper. Further additions of copper produced spectra that weren't as clean which may have been caused by dilution of the protein.

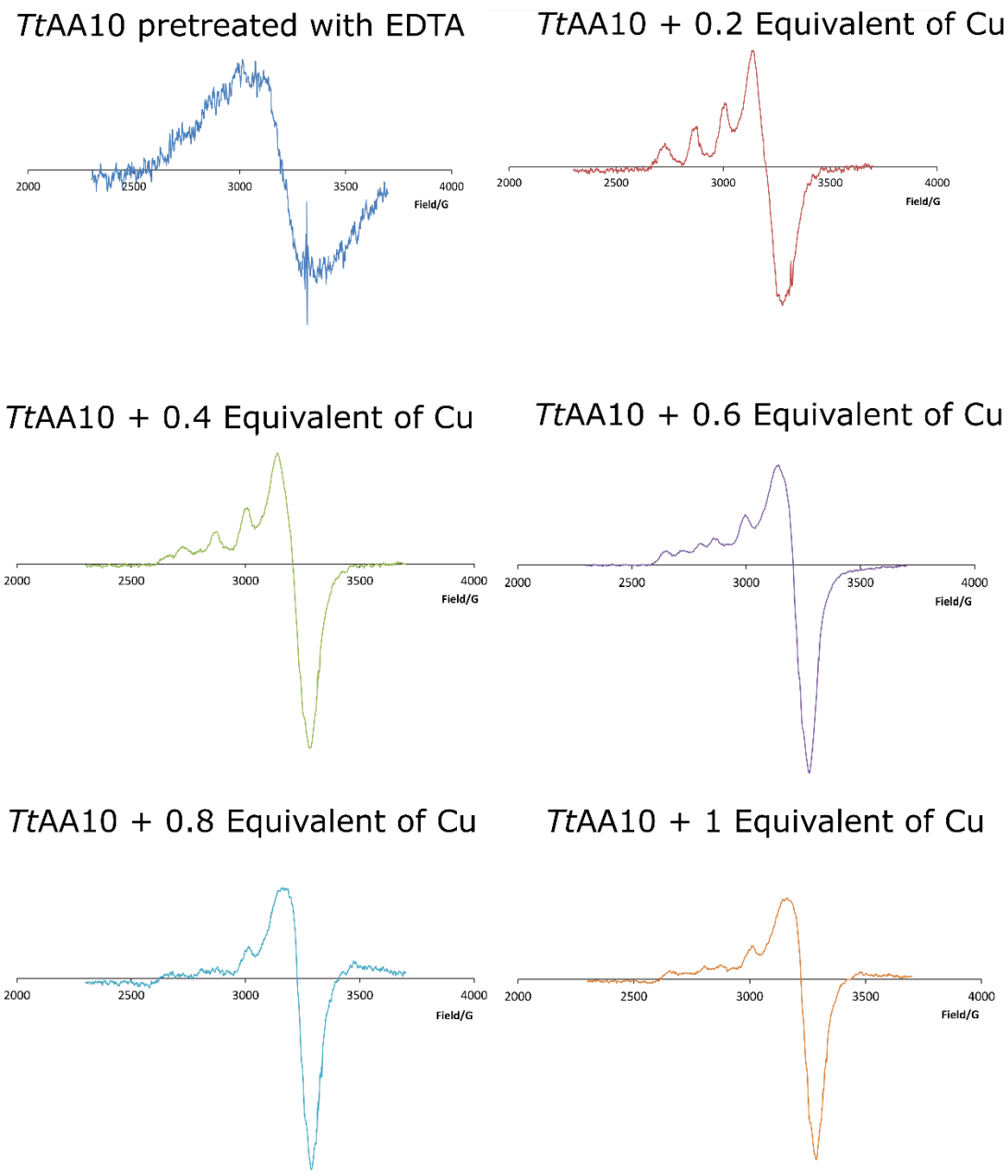


Figure 120 EPR copper titration experiment in which the protein was pre-treated with EDTA to remove the copper (as soon by the lack of EPR signal). Addition of 0.2 equivalents of Cu (to the concentration of protein) produced a single Cu species indicative of the histidine brace. Addition of a further 0.2 equivalents caused a change in the spectra with appearance of a second set of hyperfine peaks in the parallel region. The clearest change is seen after 0.6 equivalents of Cu have been added, where multiple peaks in the spectra indicate two Cu species. Loss of resolution in the last two spectra may be due to signal dilution.

As protein stocks were too low to perform isothermal titration calorimetry, the binding affinity of the second copper site was tested by another simple titration experiment. The copper associated with the histidine brace of an LPMO is known to bind with high affinity. If the second copper species bound with low affinity, it may be displaced by other metals. Similarly, this experiment was used to assess whether any other metals may have preferential binding to this second metal site. A series of metals were added to the protein sample to see if there would be any loss of signal that would equate to copper ion displacement. The protein was mixed with Fe, Ni, Zn and Mn (from 0.5 to 1.5 equivalents to the protein concentration) following the Irving-Williams series, with the only exception of Mn, which was added last in order not to cover the EPR spectrum with the Mn signal. No change in the appearance of the spectra could be observed. **Figure** 121 shows the four spectra taken after sequential additions of metal (all to the same sample) and compared with the original *Tt*AA10 EPR single showing two copper species.

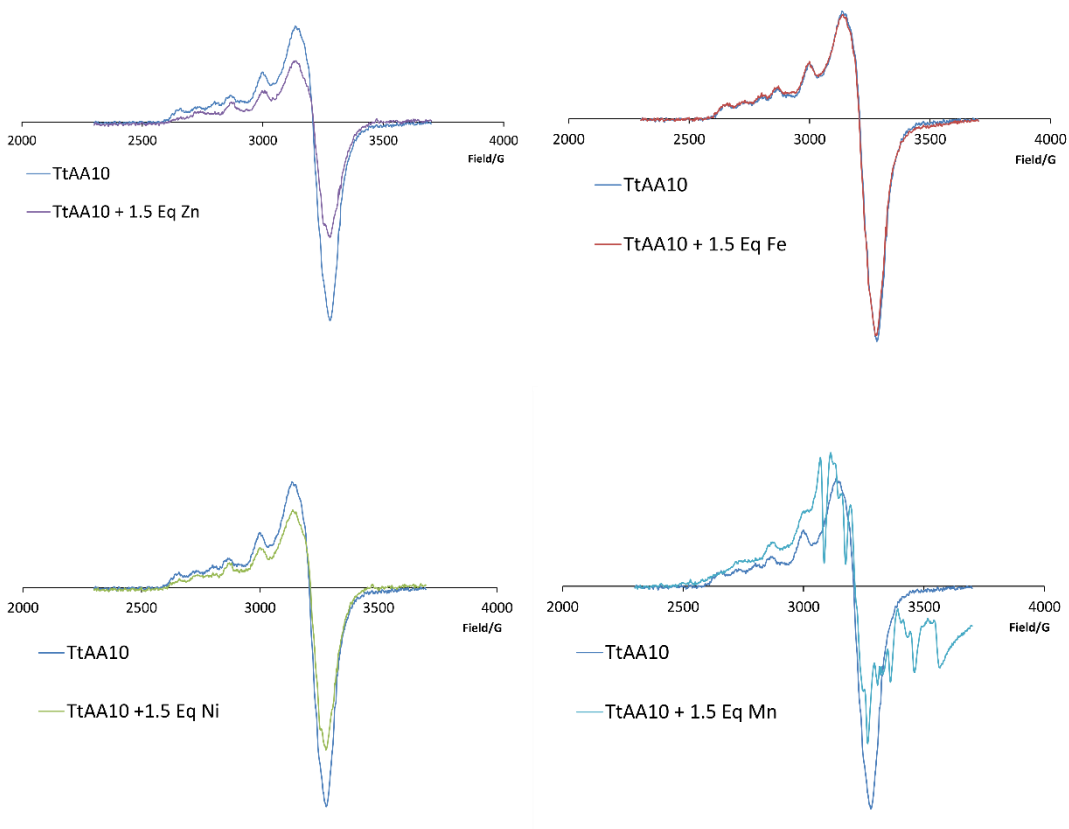


Figure 121 Metal titration EPR. Displacement of the second Cu was probed with 4 metals sequentially, Fe, Ni, Zn and Mn. Addition of each metal did not change the copper signal other than by dilution. Addition of Mn caused the appearance of the Mn EPR signal but with no clear displacement of the Cu ion form the second binding site.

6.5.3 TSA

The stability of the protein was tested under different conditions by TSA. **Figure** 122 shows the comparison of *Tt*AA10 only and protein incubated with EDTA and ascorbate. The protein itself exhibited a melting temperature of approximately 50 °C and was found to be consistent across many different TSA experiments. Introduction of EDTA to the protein sample, which would remove the copper from the active site, reduced the melting temperature by 3.4 °C. This strong negative change in temperature indicates how important the copper ion is within the scaffold of the Histidine brace. Furthermore, addition of ascorbate to the protein reduced the melting temperature by nearly 20 °C. This radical shift in stability of the protein indicates how introduction of a reducing agent can destabilise an LPMO. Interestingly, gallic acid was shown to have less of an effect on the stability of *Tt*AA10, with a negative change in melting temperature only of 2.1 °C (T_M of 47.7 °C curve not shown).

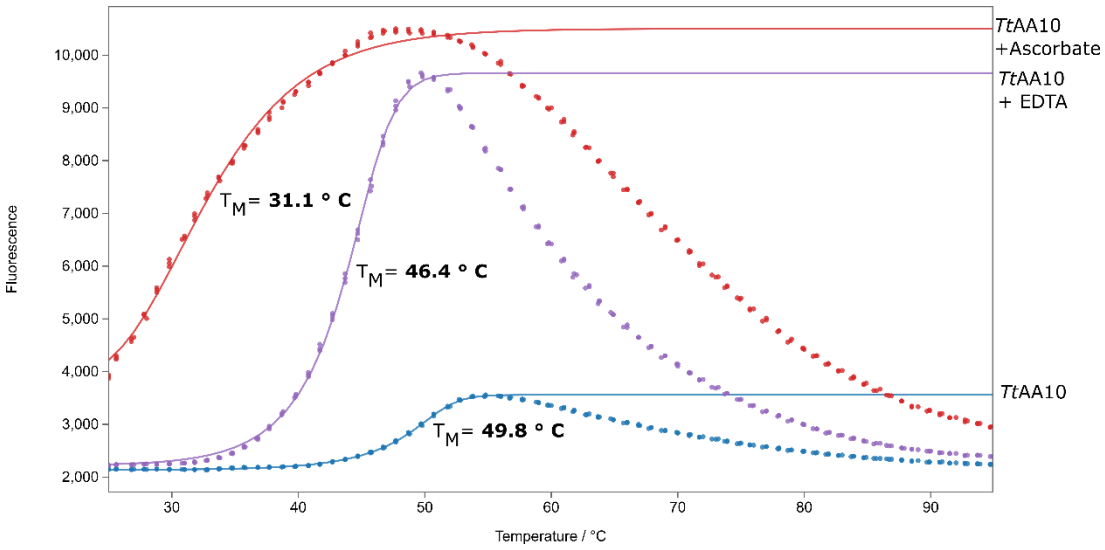


Figure 122 TSA of *Tt*AA10 with melting curves obtained from samples containing protein only, *Tt*AA10 + ascorbate (1 mM) and *Tt*AA10 + EDTA (1mM). The fluorescence level for *Tt*AA10 is lower due to the occurrence of copper quenching of the dye.

*Tt*AA10 was incubated with both cellulose and chitin substrates to assess what affect they had on the protein, and as an indication of whether any binding to substrate occurs. **Figure** 123 shows the TSA experiment comparing the melting temperature of *Tt*AA10 with and without chitin. A positive shift in melting temperature was observed for the protein, with an increase of 6.6 °C. The same test was carried out with both PASC and Avicel and a biphasic melting curve was observed. For each substrate, two melting events could be seen within the data and the curves were modelled to give two distinct melting temperatures. The first melting event for each substrate corresponds to that of protein which was likely not interacting with the substrate, at 45.7 and 45.2 °C for PASC and Avicel respectively. The second events showed increased melting temperatures of 59.0 and 60.4 °C for PASC and Avicel respectively. The large increases in melting temperature, 13.3 and 15.2°C for PASC and Avicel respectively indicate a strong interaction between the protein and the substrate.

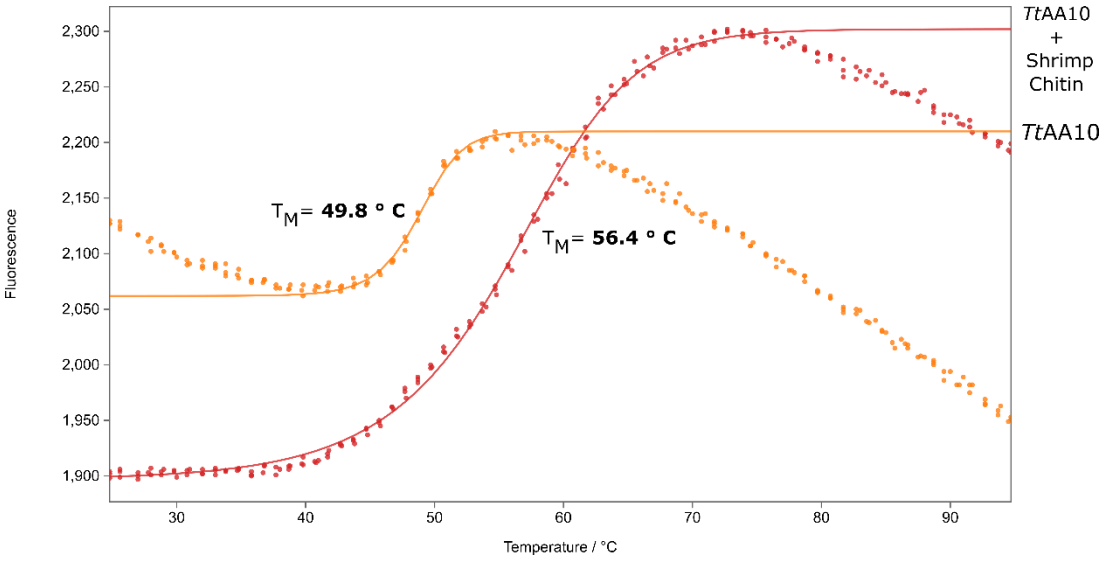


Figure 123 TSA of *Tt*AA10 with crab chitin , showing a 6.6 °C increase in protein melting temperature on incubation with the substrate.

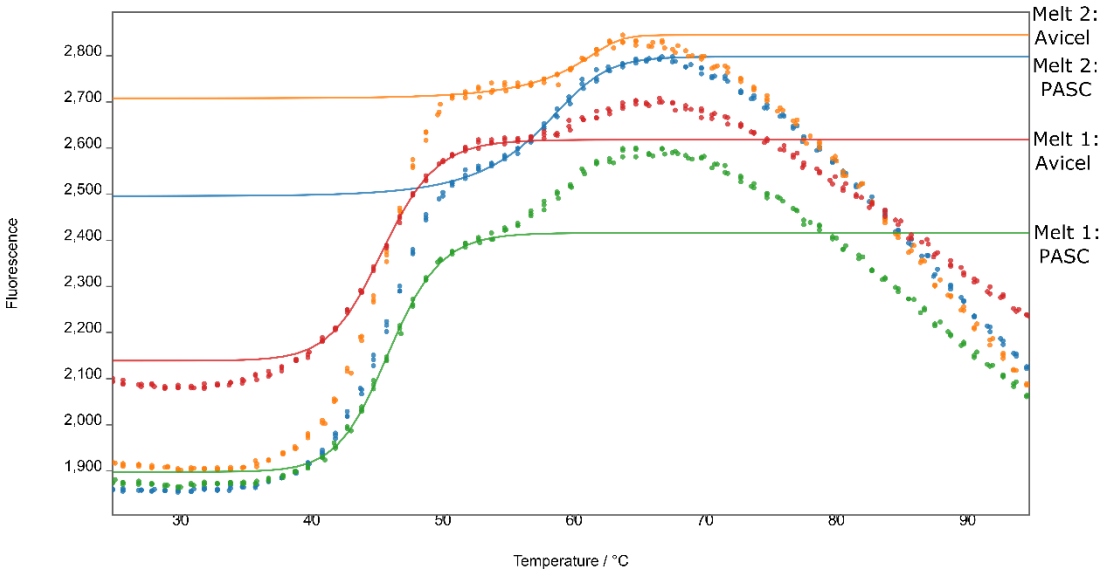


Figure 124 TSA of *Tt*AA10 mixed with PASC and Avicel , producing two melting events for each substrate. Protein T_M with PASC, Melt 1 and 2 were measured at 45.7 and 59.0 °C respectively. Protein T_M with Avicel, Melt 1 and 2 were measured at 45.2 and 60.4 °C respectively. The protein on its own during this experimental run had a melting temperature of 49.1 °C, curve not shown.

6.5.4 Activity analysis

The ability of *Tt*AA10 to degrade cellulose and chitin was tested by MALDI-TOF mass spectrometry using crystalline substrates PASC, Avicel and chitin from crab shells. *Tt*AA10 was found to be active on PASC using both gallic acid and ascorbate as the reducing agent; whereby the deffinition of activity in this case was the ability of the protein to produce oxidised oligosaccharide species. The enzyme was only found to be active on Avicel when using ascorbate as the reducing agent. Controls were run to ensure that reducing agent alone was not able to produce degradation products. The product profile observed for PASC and Avicel is shown in **Figure** 125 and **Figure** 127 and close up views of DP7 for each substrate shown in **Figure** 126 and **Figure** 128 respectively. No oxidative activity was observed when *Tt*AA10 was incubated with chitin.

The peak profile observed for the breakdown of PASC covers the DP range 4-11 with clarity, **Figure** 125, whereas for the breakdown of Avicel the range DP 4-8 is produced, **Figure** 127. Other larger peaks were observed in the mass spectrum, but were close to the baseline noise so have not been included in this discusssion. It is possible that larger substrates are produced and do not show up as well due to the limit of their solubility. It is clear that a broad range of chain lenghts are produced, with the clearest species belonging to DP 7 for the PASC profile, and DP 6 and 7 for Avicel, although these species could not be quantitatively assigned as the major products using this method, the ratio of species suggests these to be the most prevalent in the sample. To identify the peaks within the DP 7 profile, **Figure** 126 and **Figure** 128 show 'zoomed in' views of the region 1160-1230 *m/z*. The same three key peaks are observed for both substrates, which have their own isostope patterns, namely 1173.2, 1191.2 and 1214.2 *m/z*.

Each individual peak in the DP 7 product example was identified and assigned to a potential species as summerized in **Table** 23 for PASC. All peaks are likely to be Na adducts, due to the buffer system used, but examples in **Table** 23 are given which could correspond to K adducts instead. The suggested species are assigned to whether the products correspond to C1 or C4 oxidation of the glycosidic bond. The **Table** 23 shows a mixture of products produced by both C1 and C4 oxidation, with some product m/z values indicating species arising from two separative oxidative cleavage events at both ends of the chain. The same peaks are observed in the degradation of Avicel, albeit with species assigned as the ketoaldose and the native oligosacchardie at 1173 and 1175 *m/z* respectively at a smaller ratio that observed in PASC (see peak heights in **Figure** 128).

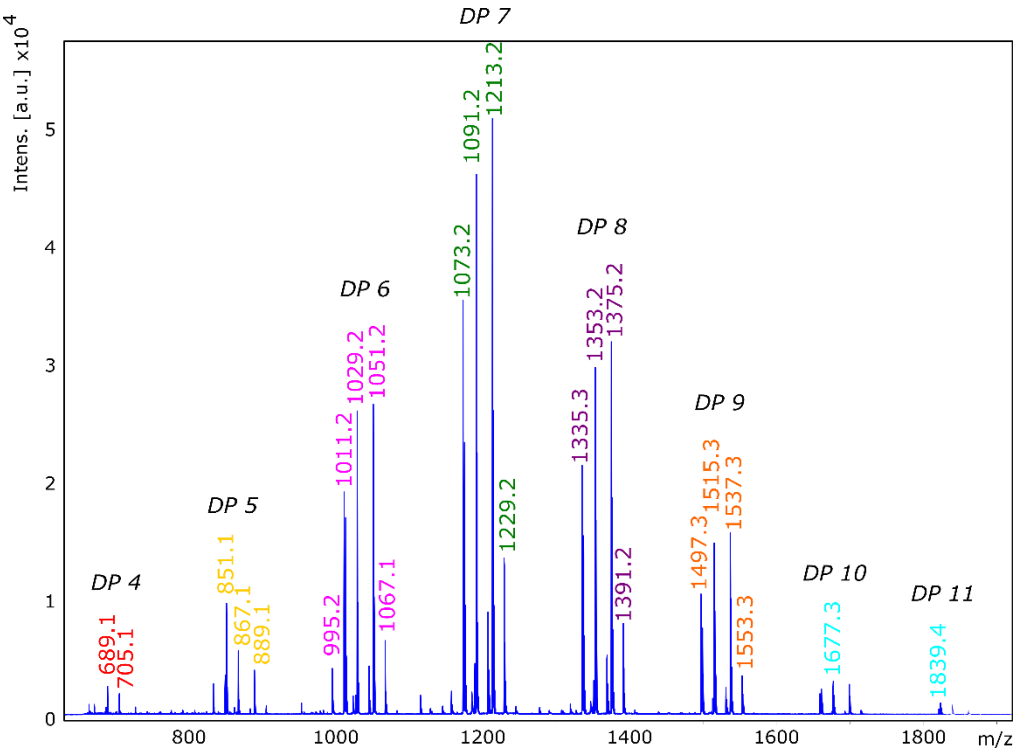


Figure 125 MALDI-TOF MS of *Tt*AA10 (0.1 μM), incubated with PASC and ascorbate (1 μM) , showing soluble reaction products. Reaction products for glucose DP 4-10 can be observed with clarity and main peaks are labelled. DP 4,10 and 11 are observed but in lower amounts and are close to the baseline noise.

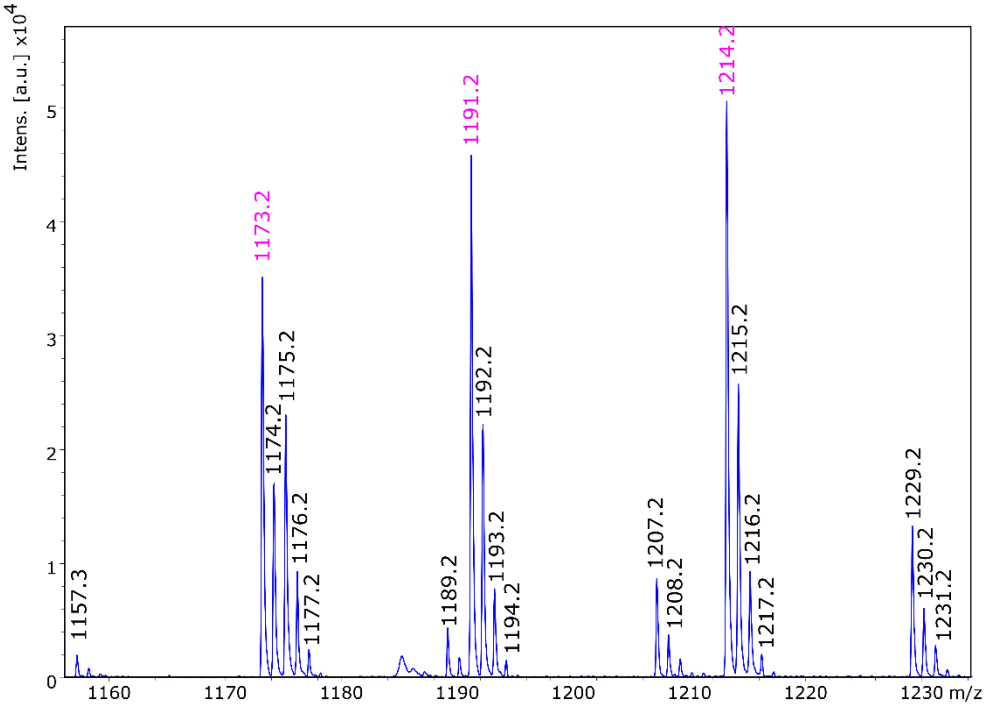


Figure 126 MALDI-TOF MS of *Tt*AA10 activity assay on PASC using ascorbate, focusing on the products peaks of DP 7. The major product peaks are coloured in pink and are assigned as ketoaldose + Na (1173.2 *m/z*), aldonic acid + Na (1191.2 *m/z*) and aldonic acid + 2Na (1214.2 *m/z*). Other peaks are isotope peaks.

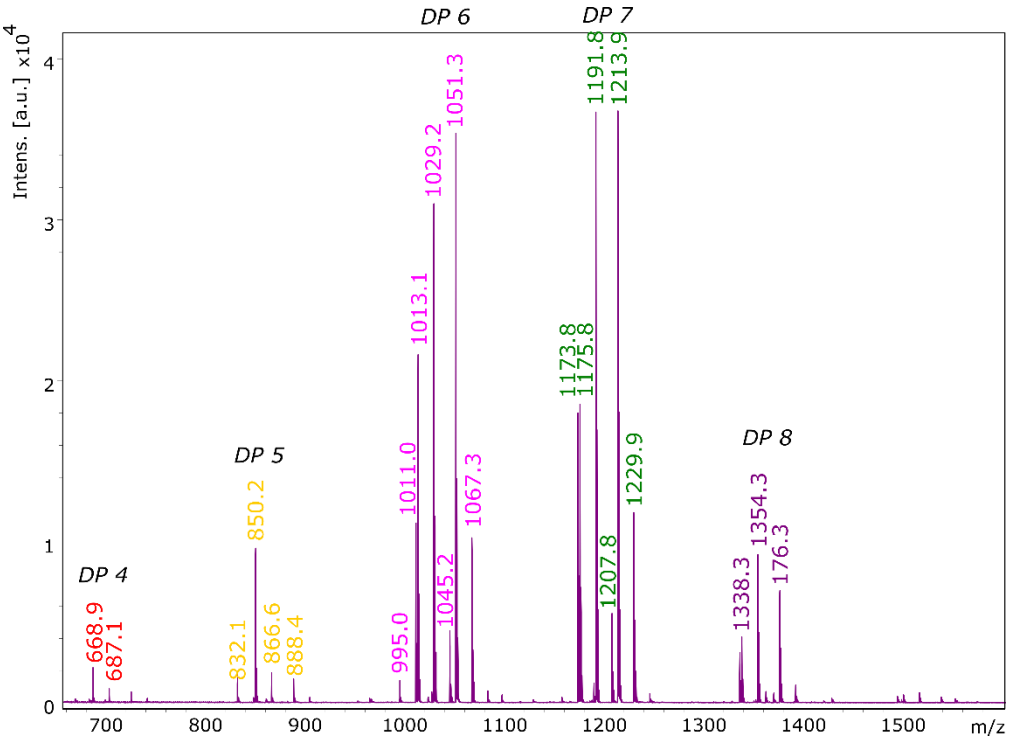


Figure 127 MALDI-TOF MS *Tt*AA10 (0.1 µM), incubated with Avicel and ascorbate (1 µM), showing soluble reaction products. Reaction products for glucose DP 4-8 can be observed with clarity and main peaks are labelled. Higher DP are observed but in very low amounts and are close to the baseline noise.

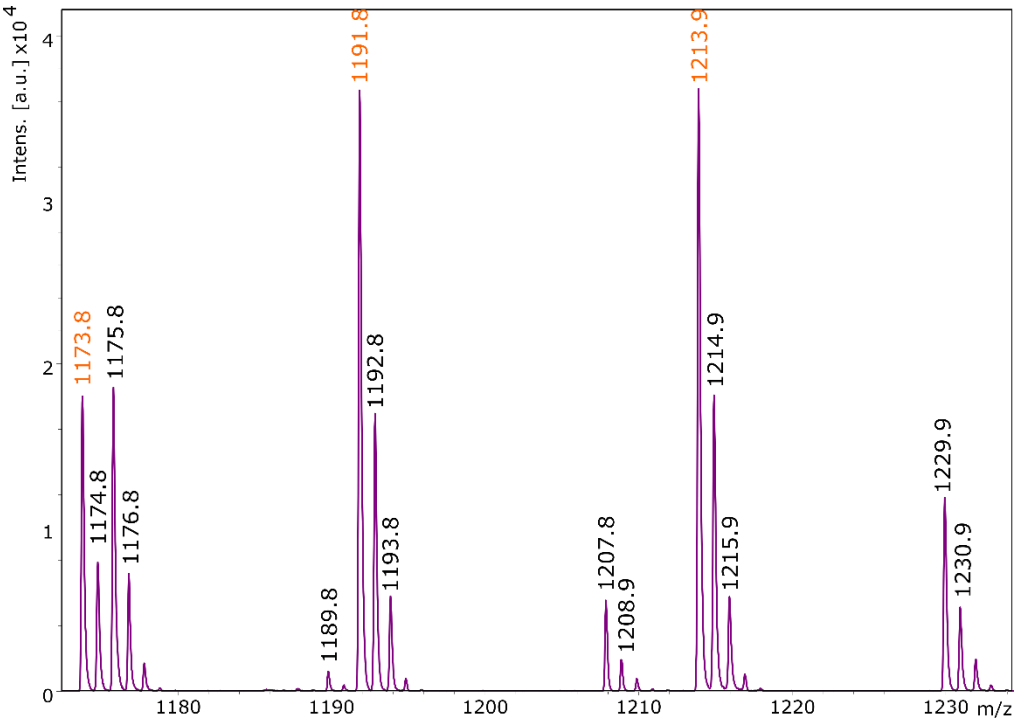


Figure 128 MALDI-TOF MS of *Tt*AA10 activity assay on Avicel using ascorbate, zoomed in view of the products peaks of DP 7. The profile is very similar to that observed for the degradation of PASC. The major product peaks are labelled in orange and are assigned as ketoaldose + Na (1173.2 *m/z*), aldonic acid + Na (1191.2 *m/z*) and aldonic acid + 2Na (1214.2 *m/z*). Other peaks are isotope peaks.

Table 23 Species assignments of PASC product peaks in the region 1173-1231 *m/z*, corresponding to the ion cluster of DP 7. Species coloured in grey indicates a possible alternative assignment for the peak.

Peak (m/z)	PASC Degradation Products: Possible Species	Ion Adduct	Cleavage Pattern
1173.2	**Ketoaldose**	Na	C4
	Lactone	Na	C1
1174.2	Isotope	Na	C4
1175.2	**Native oligosaccharide**	Na	C1 or C1/C4
1176.2	Isotope	Na	C1/C4
1177.2	Isotope	Na	C1/C4
1189.2	**Ketoaldose, aldonic acid**	Na	C4/C1
	Ketoaldose	K	C4
	Lactone	K	C1
1190.2	Isotope	Na	C4
1191.2	**Aldonic acid**	Na	C1
	Native oligosaccharide	*K*	C1/C4
	*Gemdiol**	*Na*	*C4*
1192.2	Isotope	Na	C1
1193.2	Isotope	Na	C1
1194.2	Isotope	Na	C1
1207.2	**Gemdiol, Aldonic acid**	Na	C4/C1
	Aldonic acid	K	C1
1208.2	Isotope	Na	C1
1209.2	Isotope	Na	C1
1214.2	Aldonic acid	2Na	C1
1215.2	Isotope	2Na	C1
1216.2	Isotope	2Na	C1
1229.2	**Gemdiol, Aldonic acid**	2Na	C4/C1
1230.2	Isotope	2Na	C4/C1
1231.2	Isotope	2Na	C4/C1

6.5.5 Crystallisation and Structure Solution

Commercial screens were set up using *Tt*AA10 at a concentration of 7 mg mL^{-1}. A hit was obtained in the PEG/Ion screen (0.2 M magnesium formate dehydrate, 20 % *w/v* PEG 3350, pH 7), where small clusters of very thin 'bow' shaped crystals grew, as shown in **Figure** 129. Optimisation of the crystals by changing the mother liqueur conditions (altering the concentration of magnesium formate dehydrate and PEG 3350) did little to improve the size or definition of the crystals. The fan-like clusters were manually broken apart and crystals fished using loops (see **Figure** 129) from the detached crystal sections. Despite their size and difficult morphology, the crystals were found to diffract to high resolution (**Figure** 130, left) and as such were sent for data collection on the I04 beamline at the Diamond Light Source.

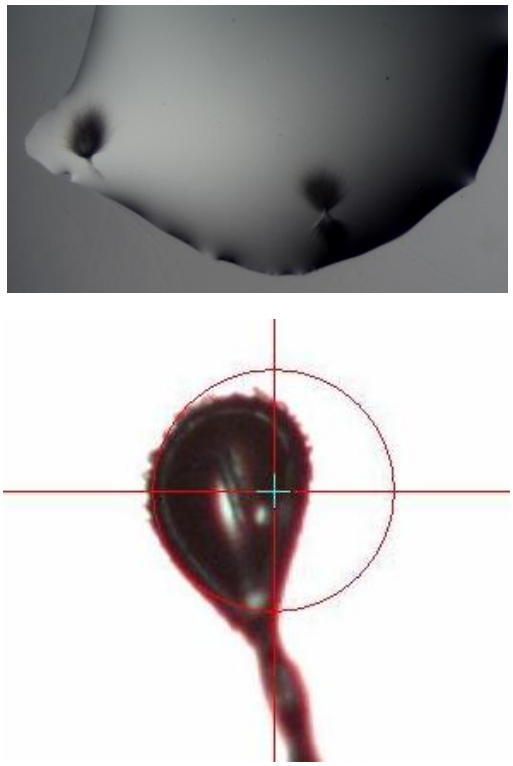


Figure 129 Crystallisation of *Tt*AA10 Top; *Tt*AA10 crystallization screening, with a hit in the screen PEG Ion, well B8 which contained the following conditions; 0.2 M Magnesium formate dehydrate, 20% *w/v* polyethylene glycol 3,350, pH 7.0. The crystals produced were very small bow/fan shaped crystals. Bottom: optimised *Tt*AA10 crystal held in loop during in-house diffraction testing.

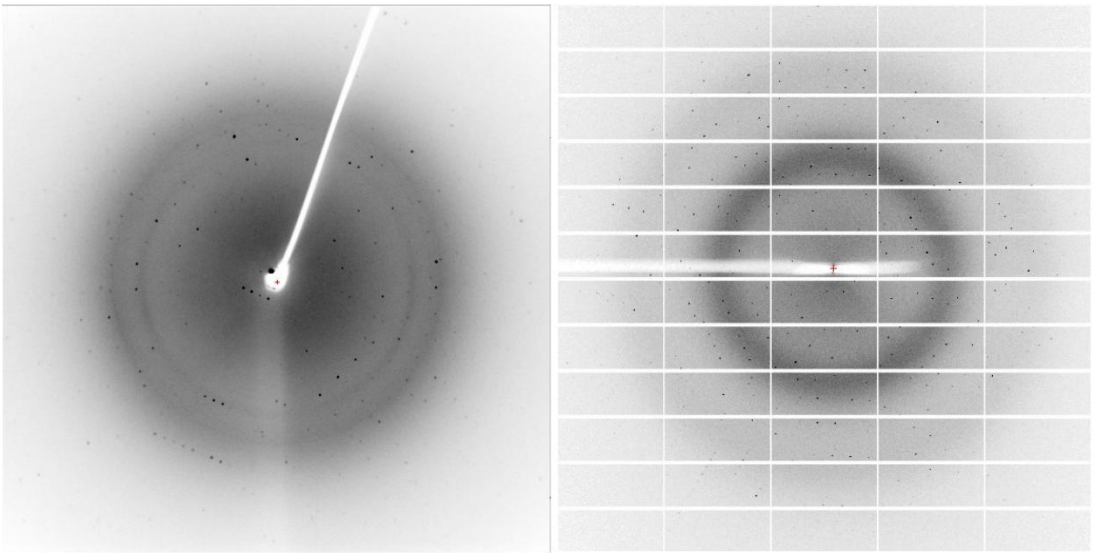


Figure 130 Diffraction pattern of *Tt*AA10. Left: *Tt*AA10 crystal tested for diffraction in house. Right: Image taken from the start of data collection of the *Tt*AA10 crystal on the I04 beamline at the Diamond synchrotron.

Data collected at the Diamond synchrotron (**Figure** 130, right) contained a maximum resolution of 1.17 Å in the outer shell, but had a very poor completeness of approximately 4 %. The data was reduced using Aimless [210] and the resolution cut off chosen based on the graphical representation of $CC_{1/2}$ vs. resolution shown by Pimple in Aimless. Copper phasing was used produce a structure solution using SHELX [211], followed by model building in ARP/wARP [212] and CRANK2. Repetitive iterations of manual model building in COOT [159] and refinement by REFMAC [160] were carried out. The copper ions were built into the structure in COOT. In COOT, displaying a symmetry molecule of the structure showed that a histidine on the strep tag of the symmetry molecule was binding in the secondary copper site of the model. Collection and refinement statistics are shown in **Table** 24.

Table 24 Data collection and refinement statistics for *Tt*AA10. Data was collected at 1.17 Å but reduced to 1.42 Å to improve data quality.

		*Tt*AA10
Data collection	Space group	P 1 2_1 1
	a, *b*, *c* (Å)	36.64 62.52 43.19
	(°)	90.00 90.00 90.00
	Resolution (Å)	36.18 (1.42)
	R_{merge}	0.059 (0.722)
	R_{pim}	0.050 (0.672)
	CC(1/2)	0.998 (0.606)
	I / I	9.6 (0.4)
	Completeness (%)	99.7 (95.6)
	Redundancy	3.9 (2.7)
Refinement	Resolution (Å)	1.4
	No. reflections	147995
	R_{cryst} / R_{free}	0.16/0.18
No. atoms	Protein	1687
	Ligand/ion	2
	Water	157
B-factors (Å^2)	Protein	13.4
	Ligand/ion	15.2
	Water	21.9
R.m.s. deviations	Bond lengths (Å)	0.013
	Bond angles (°)	1.64
Ramachandran plot	Most favoured (%)	94.2
	Allowed (%)	4.9

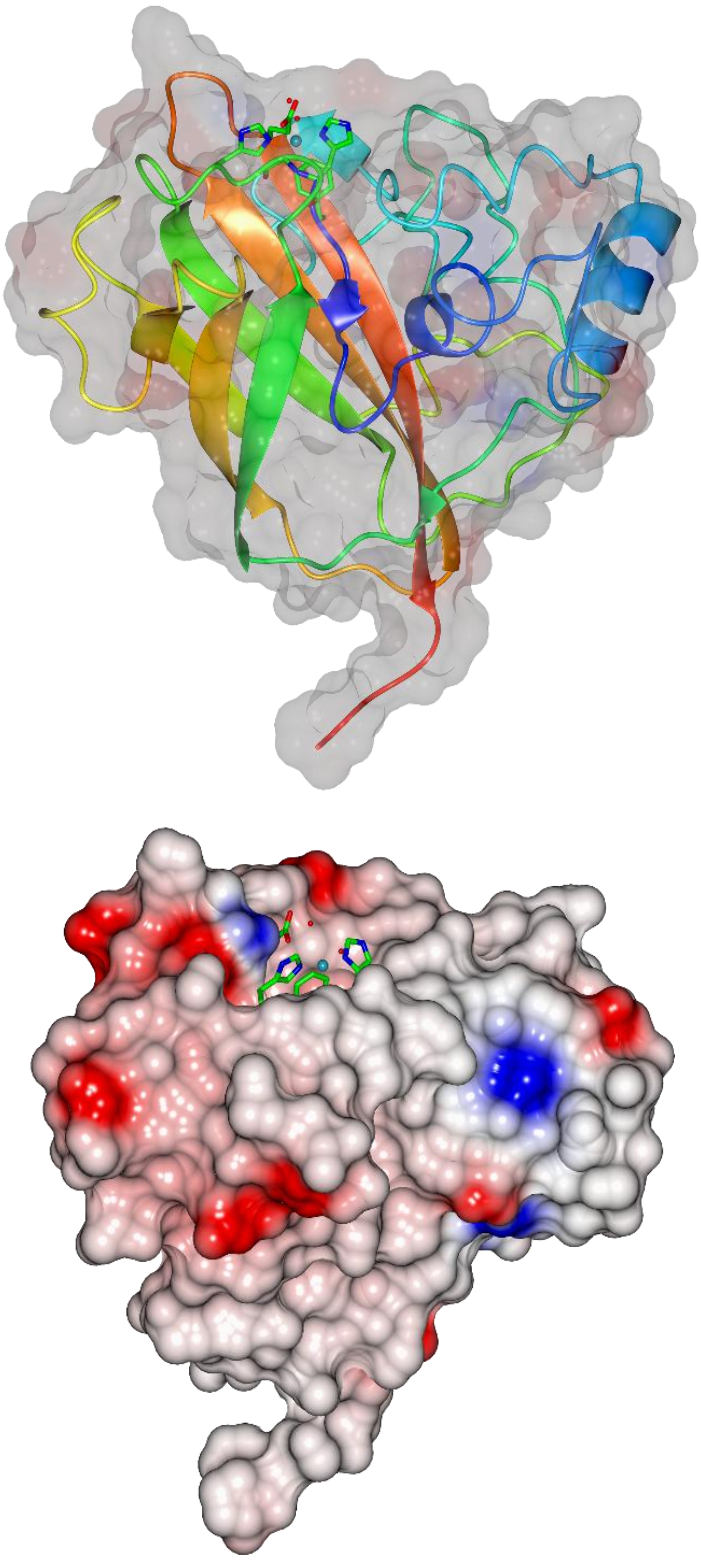


Figure 131 Above: Structure of *Tt*AA10, coloured red-blue for N to C-terminus, with histidine brace shown as cylinders and slight transparent surface. Below: Structure shown as a surface, coloured by electrostatic potential, with the histidine brace shown as cylinders rather than surface. Images made in CCP4mg.[55]

The structure of *Tt*AA10, shown in **Figure** 131, displays the classic features of an AA10 LPMO. The core β-sandwich is surrounded by several variable loops and a single defined α-helix. The active site is positioned on the 'top' of the structure, on the flat surface. The surface view of the protein, based on electrostatic potential shows that there is a short protrusion to the left of the active site. The strep tag is observed at the C-terminal end (red) of the protein (ribbon view). **Figure** 132 highlights the planar nature of the binding face of *Tt*AA10.

Molecular replacement using several AA10 models, as well as ensemble models, failed to find a structure solution for *Tt*AA10 (using MOLREP and Phaser[157, 214]). The structure of *Tt*AA10 was overlaid with the model of CjAA10 (5FJQ), which displays the highest PDB sequence homology, which had been used during structure solution trials. This is a chitin active AA10 from *Cellvibrio japonicas.* It is clear that whilst the core features of both LPMOs remain the same, significant differences arise in the loop structures, **Figure** 133. These differences are likely to have prevented a solution arising from molecular replacement.

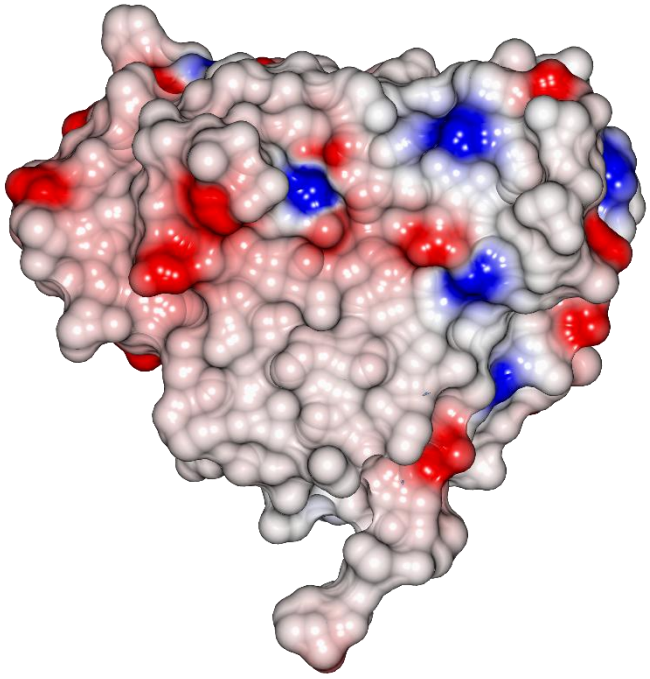


Figure 132 Surface view only of *Tt*AA10, highlighting the planar binding surface at the 'top' of the protein. However, on the left hand side is a protrusion due to the side chain of His190. Image made in CCP4mg.[55]

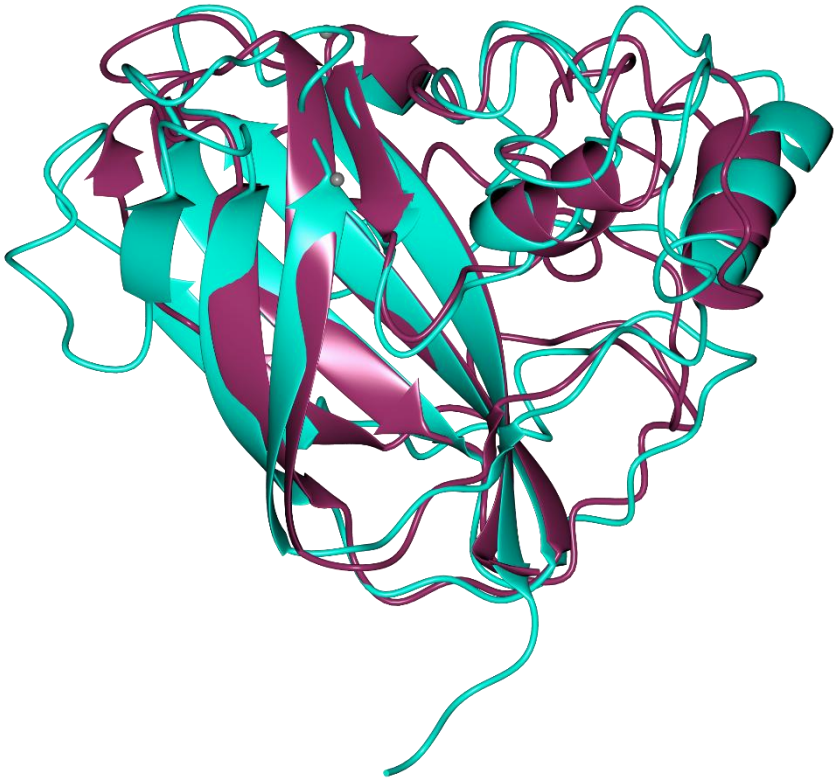


Figure 133 *Tt*AA10 (cyan) overlaid with *CJ*AA10 (purple, PDB code 5FJQ). *Tt*AA10 copper ions shown as grey spheres. Image made in CCP4mg.[55]

The active site of *Tt*AA10 contains the typical LPMO histidine brace motif, formed by the N-terminal residue, His1 and a midchain residue, His107, **Figure** 134. The secondary coordination sphere of the copper ion is formed of Phe195, Gly105 and slightly further away, Glu193. Density for a water ligand was also present in the structure. Copper-N distances to the coordination histidine residues are shown in **Table** 25, alongside distances to other neighbouring residues and ligands.

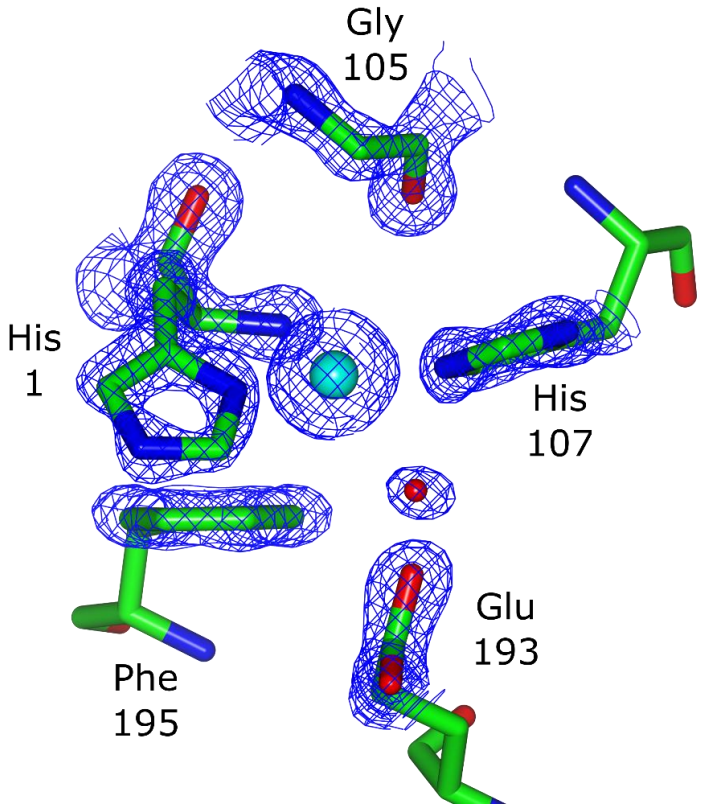


Figure 134 *Tt*AA10 histidine brace with observed electron density map shown at contour level 1.0 σ (equivalent to 0.48 electrons/Å3). Image made in CCP4mg.[55]

Table 25 Distances between the Cu ion and nearby atoms on coordinating residues (bold) and secondary coordination sphere.

Copper-histidine brace	Distance/ Å
Cu-N (ring, His1)	1.9
Cu-NH_2 (amino terminus, His1)	2.2
Cu-N (ring, Hi107)	2.0
Nearby Molecules	
Cu-C (nearest C on ring, Phe195)	3.5
Cu-O (Glu193)	3.7
Cu-O (Water)	3.2
Cu-O (Gly105)	4.0

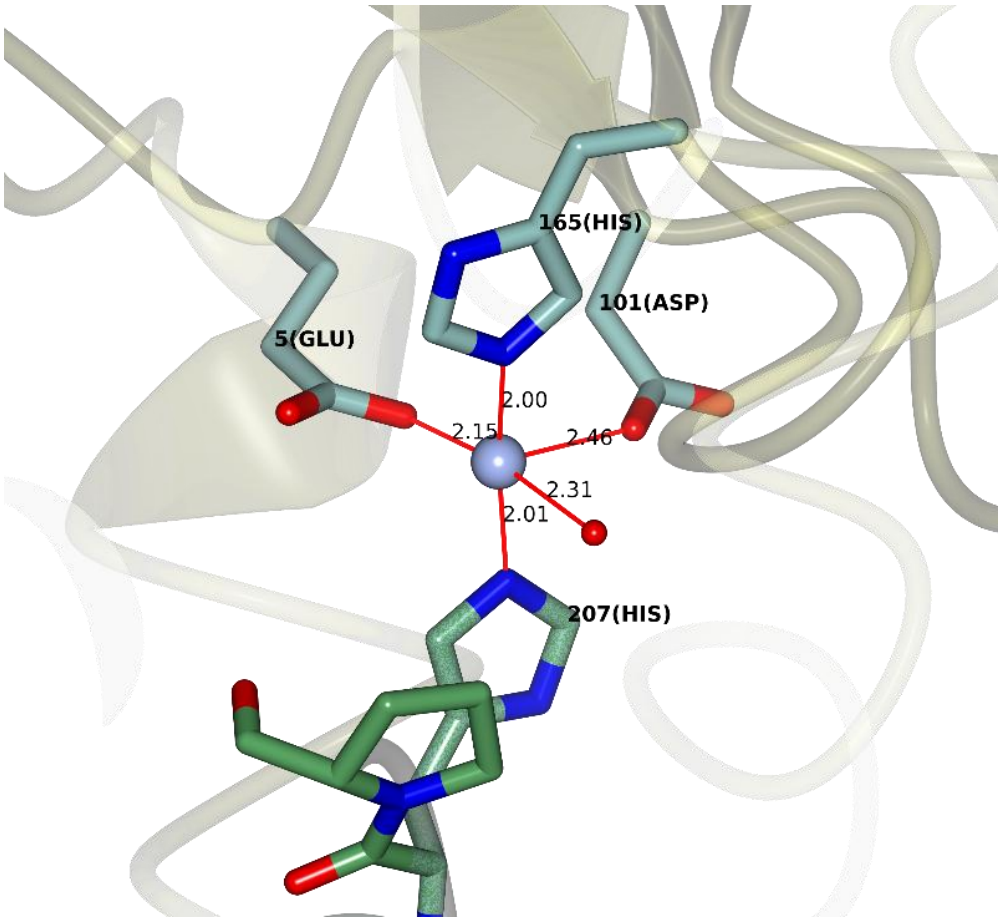


Figure 135 *Tt*AA10 Secondary copper site formed by three residues from the protein chain (Glu5, His165 Asp101), a water molecule and a histidine (207) residue from the strep tag of another molecule in the crystal lattice.

The second copper coordination site is shown in detail in Figure 135, complete with bond distances. Coordination of His207 shown in dark green comes from a symmetry molecule within the crystal lattice. Coordination of Glu5, His165 and Asp101 complete the copper coordination site. It was shown in Chapter 2 that *Tt*AA10 formed a monomer in solution by analysis with SECMALS. It is therefore likely that the coordination of the histidine is replaced in solution by coordination of a water molecule, or ion such as chloride.

6.6 Discussion

6.6.1 Native expression of *TtAA10*

The source organism of the *Tt*AA10 gene, *T.turnerae* was grown under similar conditions to those used upon first collection and identification of the organism in the 1983 by Waterbury *et al.*[215] A sample of the original collected organism, from which the genome had been derived [104], was acquired and grown using sterilised seawater. The media was supplemented with a cellulose substrate to provide a source of food for the bacteria. Growth was very slow but successful and cells were harvested after 5 days of culture growth. Proteomic analysis yielded a large abundance of different proteins, both in the cell pellet and the media. Yang *et al* performed a similar experiment; looking at the secreted cellulosic enzymes by LC-MS when *T. turnerae* was grown on surcrose, CMC and sigmacell.[104] A broad range of lignocellulosic enzymes was observed, as shown in Appendix 5. As expected due to the signal peptide found on the *Tt*AA10 gene, the LPMO was found secreted into the media by proteomic analysis and supports the theory that the enzyme is functional on cellulose. *Tt*AA10 was also observed secreted when the organism was grown on CMC by Yang *et al*. No external reducing agents were provided to the cells to act as an electron source for *Tt*AA10, indicating that the cells must have been providing their own reducing agent to aid the oxidative action of *Tt*AA10 if they were to gain from its expression.

Pure recombinant *Tt*AA10 protein was used in an experiment to try and 'catch' any potential binding partners (electron source) produced by *T.turnerae*. The pure protein had been produced with a C-terminal strep tag, and this tag was used to imobilise the protein on a streptavadin column. A sample of the media, into which many different proteins had been secreted, and soluble fraction of lysed cells was flowed over the streptavadin column containing the bound *Tt*AA10 protein. SDS PAGE analysis was used to determine whether anything had bound to *Tt*AA10, but no extra bands were observed in the sample of *Tt*AA10 removed from the column. This result suggested that nothing had adhered to *Tt*AA10 with a strong enough interaction to keep it bound during the washing step. If a partner protein used for electron transfer was within the media, it may not interact with the LPMO strongly enough to remain bound on a column. Secondly, a protein may not necessarily be used by *T.turnerae* as an electron source, instead the bacteria may utilise a small organic molecule, which would be unlikely to be observed on an SDS PAGE. As such, the experiment was a promising idea to try to identify any possible electron donating partners produced, or used

by *T.turnerae*. Further work using this idea could test whether the pure *Tt*AA10 exposed to the harvested cell culture media would be active without an added reducing agent.

6.6.2 Copper Coordination

During EPR analysis of *Tt*AA10 it was found that the first batch of pure recombinant *Tt*AA10 produced contained a mixture of copper species, whilst subsequent batches contained only a single copper species in the active site of the protein. This single species was similar spectroscopically to other Cu-histidine brace coordination sites, so this species was deemed to be the copper within the active site of the protein. The data was simulated to include 2 nitrogen atoms coupling to the Cu ion and parameters found to be g_x = 2.035, g_y = 2.100, g_z = 2.267; A_x = 110 MHz; A_y = 80 MH_z and A_z = 420 MHz. The histidine brace contains three possible nitrogen atoms through which the Cu ion can coordinate, and simulation of the EPR of *Tt*AA10 with only 2 nitrogen atoms does not mean that the Cu ion is lacking a coordinating ligand. The measured EPR spectrum is broad and thus coupling to a third nitrogen atom was not accurately identifiable by EPR, but still likely to be occurring. A recent review of LPMO coordination chemistry by Vu and Ngo[71] has complied together EPR parameters (among other useful measurements) of characterised LPMOs. Comparison of the g_z value of *Tt*AA10 (the most reliable measured parameter from the data) with the complied data, shows the most similar are *Sm*LPMO10A, *Bl*LPMO10A, *Sli*LPMO10A, *Sc*LPMO10C (CelS2) and *Cj*LMPO10A, whereby the latter is the most homologous protein to *Tt*AA10. All these LPMO AA10s present a g_z value of approximately 2.26. The correlation between g_z and A_z values for LPMOS can be plotted to show different 'zones' which relate to the geometry of the copper centre. This plot, known as the Peisach–Blumberg plot[216], shows that *Tt*AA10 fits into zone 2; the A_z value, converted to G is approximately 150 $x10^{-4}$ cm^{-1}. LPMOs in zone 2 have character in which the unpaired electron of the copper ion inhabits the $d_{x^2-y^2}$ orbital.[71] Again, the A_z value is most similar to that of *Cj*LPMO10A. *Cj*LPMO10A is a chitin active AA10, but interestingly displays features that are more consistent with LPMOs active on cellulose than on chitin.[213]

The identity of the second copper species was analysed further by using protein, which was shown by EPR to contain only one copper species, and then treated with EDTA to remove all trace of copper from the protein sample. EPR was used to probe the ability of the protein to take up a second copper ion through a titration experiment on the copper-free protein. The titration experiment suggests the second copper site is opportunistic; addition of a very concentrated, but small amount of copper solution was required during the titration

experiment in order to limit dilution of the protein (to avoid EPR signal loss) and as such would result in a very high concentration gradient of copper before the sample was mixed. As such it seems likely that the first additions of copper were taken up by the high affinity binding offered by the Histidine brace, whilst later additions of copper were opportunistically taken up by the second copper site. However, the secondary copper site was not affected by introduction of other metal ions. Different batches of protein were used during characterisation of the activity of *Tt*AA10, and so it was found that the presence or absence of the secondary metal site did not affect the activity of the protein.

The occurance of a second copper site was intriguing, especially as the second copper was only 11 Å away from the active copper within the histidine brace. LPMOs are thought to have potential electron transport pathways within the enzyme that can facilitate movement of an electron from a reducing agent to the active copper ion within the histidine brace.[201] The positioning of this second metal binding site, so close to the active site of the protein thus suggested that it could have been a potential binding site for a reducing partner. As discussed earlier, through immobilised binding studies, no electron donating partner was captured from the expressed protein content when *T.turnerae* was grown on cellulose. If such a binding partner exists, it may have functioned through the metal binding site found adjacent to the histidine brace. There is precedence for secondary metal binding sites within LPMOs, with a similar occurance found in *Sli*LPMO10E, where a second metal site facilitated dimer formation between two LPMO molecules during crystallisation. [217]

6.6.3 Enzyme Stability

*Tt*AA10 was expected [61] and found to be stabilised by the addition of copper as shown in the TSA experiment. It was also shown to be susceptible to damage by reducing agents with ascorbate reducing the protein melting temperature more than incubation with gallic acid. The majority of studies on LPMOs use ascorbate as the default reducing agent, and it is likely that whilst ascorbate provides a good source of electrons for the protein to work with, it is may to cause more damage over time and may reduce overall efficiency. The real electron transfer source for bacterial LPMOs is still unknown, with some indications that CDH proteins could be responsible in some cases in providing fungal AA9s with an electron source. Gallic acid was found to reduce the melting temperature of the protein to a lesser degree and was used during activity assays. However, gallic acid has a tendency to polymerise over time and often resulted in a lower activity than if ascorbate had been used initially during activity assays with *Tt*AA10.

*Tt*AA10 was found to interact with chitin (crab shells), but not cause degradation of the substrate. The melting temperature of the protein was found to shift significantly (an increase of 6.6 °C) upon incubation with chitin, but no degradation products were ever observed during activity assays. Similar occurrences have been noted with other AA10s; in a study by Forsberg *et al*[200], an AA10, CelS2, was made with and without the associated CBM2 chitin binding domain, and both versions of the protein were found to bind strongly to chitin but only to display activity on cellulose. Ability to bind to a substrate does not therefore indicate activity.[200] As discussed during the introduction of this work, the AA10 family has been evolving more towards a cellulolytic function [197], which may indicate why AA10s with substrate specificity for cellulose are still able to bind to the ancestral chitin substrate.

Incubation of the protein with cellulose substrates PASC and Avicel created biphasic melting curves with large shifts in melting temperature. The calculated T_M values of each melting event suggested that a portion of the sample retains its 'unbound' melting temperature (49.1 °C) whilst another portion of protein displayed a T_M increased by 13 °C for PASC and 15 °C for Avicel. The increase in protein stability of *Tt*AA10 when incubated with a cellulose substrate suggests that a binding interaction is taking place. However, the fact that not all the sample was associated with this binding event may support the hypothesis that it is a requirement of the protein to only transiently bind to the substrate in order for continual activity to take place. Hypotheses surrounding how binding of the substrate may activate an LPMO carry some weight to this argument; if the protein remains bound to the surface of the polysaccharide, as shown by the strong interaction of *Tt*AA10 with chitin, it is not being reactivated ready to perform another catalytic event. This idea was touched upon early in the discovery of LPMO action, where Moser *et al* suggested that weaker binding affinity occurs to enable re-use of the proteins and favour equilibrium binding of the LPMO domain.[54] However, the concept of transient binding, by our understanding of the word does not reflect the observations made by Eibinger *et al*, whose atomic force microscopy (AFM) studies on two AA9 proteins active on cellulose (*Nc*LPMO9F and *Nc*LPMO9C) showed that adsorption and subsequent desorption at the surface of cellulose had an average timeframe of minutes rather than seconds. Indeed, the proteins studied were found to be relatively immobile, even more so on the side faces of the cellulose chain, thought to be the preferred substrate region of these proteins.[218] This study also visually highlighted the difference in mode of activity between LPMOs and cellobiohydrolases, which were shown moving along the cellulose fibrils in a unidirectional mode of movement. Similar visualisations had been made by Igarashi *et al*,[219] who noted that without sufficient pre-

treatment of the crystalline substrate there were not enough free movement 'lanes' leading to instances of protein 'traffic jams'. Studies by Frandsen[201] and Simmons[199] on the activity of *Ls*AA9 and *Cv*AA9A on soluble glucans and xylans showed an appreciable change in the *Ls*AA9 Cu active site of the protein upon interaction with substrate, by EPR. These studies using LPMOs capable of cleaving soluble substrates show that introduction of the substrate does have an effect on the Cu active site and may support the hypothesis that introduction of the correct substrate can mediate activation of the protein. However, differences in mechanism may be related to the type of substrate; soluble substrates have a larger degree of free movement, an LPMO may therefore requires a stronger binding interaction to hold the substrate long enough for oxidative cleavage to take place, possibly indicating why changes in the active site arrangement occur. With crystalline substrates, where the preferred substrate is collectively in one region, it may be more favourable for an LPMO to adsorb to the preferable surface and 'bore downwards' so to speak, rather than move on after a single catalytic event. EPR studies of *Tt*AA10 showed no change in the observed parameters upon incubation with crystalline cellulose, even after many days of incubation. The binding interactions depicted by the change in T_M upon introduction of substrate however did show that the LPMO was interacting with the substrate to an extent that caused a positive change in protein stability. As no change in EPR parameters was observed under similar incubation conditions, it is unlikely that a binding of substrate to *Tt*AA10 causes any changes in the active site, 'in preparation' for oxidative cleavage. Hence, *Tt*AA10 may mechanistically act more like those LPMO AA9 enzymes observed by Eibinger, largely immobile whilst carrying out activity.

As to why only a portion of the protein sample used during the TSA assay increased in melting temperature, settling of the substrate within the assay tube did occur, which likely limited the potential surface area upon which *Tt*AA10 could interact. Subsequent assays using different ratios of protein to substrate may show a different ratio of 'bound to unbound'. Furthermore, TSA experiments in which *Tt*AA10 was incubated with reducing agent showed destabilisation of the protein, but if protein was mixed with substrate as well, this showed a limited degree of destabilisation of the protein. This suggests that the interaction of the substrate does have a beneficial effect on the protein, potentially preventing degradation caused by the effect of the reducing agent. LPMOs are known to degrade themselves when provided reducing agent with no substrate, as the enzyme has no other outlet for their enzymatic potential than to produce reactive oxygen species capable of degrading the LPMO protein itself.[199] A recent study by *Loose et al* described the effect of

site directed mutatgenesis on surface residues thought to provide chitin binding interactions to CBP21 (also known as *Sm*LPMO10A).[220] Alterations to certain residues thought to be involved in chitin binding (but not copper binding) actually lowered the stability of the protein and thus affected activity. The lessening of activity was proposed to be due to oxidative damage occurring to residues involved in the active site of the protein. The authors suggest that binding of the protein to the substrate effectively prevents 'off pathway' interactions which could lead to oxidative damage.[220] Perhaps binding of an LPMO to the substrate as shown by Eibinger for such long periods of time, limits time spent in solution, whereby activation by the as yet unknown natural AA10 reducing agent could cause self-inflicted damage unless substrate is already bound.

6.6.4 Enzyme Activity

Initial activity assays of the oxidative degradation ability of *Tt*AA10 were carried out on PASC using gallic acid as the reducing agent and were found to produce oligomers by analysis with TLC. Further analysis of the samples by MALDI-TOF MS confirmed the presence of oligosaccharides resulting from oxidative degradation. Tests with Avicel were at first inconclusive, with limited products being observed. When the reducing agent was changed to ascorbate, the protein was found to work much better on this substrate. Controls were carried out with and without the use of reducing agent to determine if the degradation products were indeed due to the action of the enzyme and not through redox activity of the reducing agent. All substrates were found to be clear of oligosaccharides before incubation with the enzyme and did not react to the reducing agent alone. Activity tests with chitin showed no production of species relating to oxidative products and as such *Tt*AA10 is deemed not active on chitin substrates.

A clear pattern of oxidative degradation was seen for cellulose when the protein was incubated with a reducing agent.[38] Degradation of Avicel to a range of products DP 4-8 was observed, whereas PASC displayed a product profile up to DP 12. Analysis of the individual peaks within the DP 7 peak profile firstly lead to the identification of C1 oxidising ability of *Tt*AA10. This was based on the aldonic acid +Na product resulting from C1 oxidation of the glycosidic bond observed at 1191 *m/z* in the example shown in **Figure** 126. Other peaks of interest were also exhibited in high ratio to the C1 aldonic acid product and may present C4 oxidation activity. Firstly, the peak at 1175 *m/z* was identified as the sodium adduct of the native heptamer oligosaccharide peak. This species in theory can be produced in two ways; firstly, through C1 oxidation of a chain whereby the leaving group alcohol possesses a

reducing chain end (*i.e.* a natural chain end), or through cleavage at both ends using two instances of C1 and C4 oxidation where the product is the leaving group alcohol of both oxidative reactions. Only the latter method uses C4 oxidation and thus the occurrence of a native oligosaccharide peak alone does not constitute enough evidence to suggest a protein is able to cleave at the C4 position. In the case of *Tt*AA10, two other peaks point to C4 oxidation occurring; the species at 1173 *m/z* could be indicative of both a C4 oxidised ketoaldose + Na, and a C1 oxidised lactone + Na. The intensity of the peak at 1173 *m/z* indicates a large amount of the species present within the material tested, however the C1 oxidised lactone is an unstable species and readily undergoes hydrolysis into the more commonly observed aldonic acid (1191 *m/z*). This suggests that the peak at 1173 *m/z* is more likely to be the ketoaldose resulting from C4 oxidation. Furthermore, this species can undergo hydrolysis as well, resulting in a gemdiol (+ Na, 1191 *m/z*). However, this species has the same *m/z* value as the aldonic acid species so cannot be identified with any certainty. Two other peaks in the profile of the heptamer however do cooperate with the possibility of C4 oxidation. The peak at 1189 *m/z* indicates oxidation at both ends of the resulting oligosaccharide with a ketoaldose and aldonic acid functional groups present. Whilst the intensity of this peak is small, two further peaks are observed which can be assigned as the ketoaldose hydrolysed form, gemdiol-aldonic acid + Na (1207 *m/z*) and gemdiol-aldonic acid + 2Na (1229 *m/z*). This degradation profile is similar to that observed by Forsberg *et al* for *Sc*LPMO10B, where mass spec data as well as further enzymatic fragmentation of products was used to determine that the AA10 was performing oxidation at both C1 and C4 positions.[221] A similar analysis could be undertaken for the activity of *Tt*AA10 on Avicel, where the only difference observed was in the intensity of the native oligosaccharide and C4 oxidised ketoaldose at 1175 and 1173 *m/z* respectively – both peaks were of a lower intensity compared with the activity of *Tt*AA10 on PASC. It is likely that *Tt*AA10 is acting in a similar fashion, producing a range of different species due to its lack of regioselectivity for C1 or C4. Despite high sequence similarity to *Cj*LPMO10A, a Type 1B chitin degrading AA10, *Tt*AA10 is likely to fall into the Type 3 AA10 subgroup; whereby enzymes oxidise cellulose at either the C1 or C4 position.

6.6.5 Lack of CBM

Characterisation of *Tt*AA10 was done using expression of a gene that had been truncated to express only the catalytic domain. The truncation removed both the native signal peptide and a long linker region ending in a C-terminal CBM10 domain. The CBM10 domain

indicated in the first instance that this particular AA10 would be active on cellulose. This binding module is associated with cellulose binding within aerobic bacteria, grouped into Type A CBM which contains modules capable of interacting with crystalline cellulose, and displays a flat facing binding site not unlike an LPMO.[222] Contradictory to this, in 1995 Millward-Sadler *et al* identified the first CBM10 domain as appended to a xylanase, which showed no catalytic activity on cellulose. Binding modules often do not provide any catalytic function and are simple recognition modules which are able to bind to a particular substrate. The localisation of the CBM10 domain on the C-terminal end of the *Tt*AA10 gene may suggest a recognition function as the lack of this module was not found to prevent the enzyme from adhering to the cellulose substrate and causing degradation. However, as only the catalytic domain was characterised, it is not known if the presence of this CBM10 domain would improve the activity of *Tt*AA10. A comprehensive study carried out by Crouch *et al* investigated the effect of removal and replacement of CBMs from two recombinantly expressed LPMOs from *Cellulomonas fimi* and *Thermobispora bispora.*[222] In this study it was found that removal or replacement with a non-native CBM had a negative effect on the activity of the LPMO on cellulose. One possible mechanism of movement during LPMO binding to a polysaccharide surface has been linked to whether a CBM is present, with a hypothesis being that the interaction of the CBM tethers the LPMO to the substrate before it binds,[223] therefore increasing the likelihood of LPMO oxidative action on the polysaccharides in the correct orientation. Crouch *et al* suggest that the removal of CBMs from LPMOs which naturally possess them may reduce the exposure time the enzymes have to the surface of the polysaccharide and so reduce their activity.[222] The increase in melting temperature observed during the TSA assay, upon incubation of the protein with substrate does provide some weight towards this hypothesis. Only partial binding of substrate was observed with a higher melting temperature, whilst the remainder of the protein remained in an 'unbound' state. Protein samples were analysed by EPR after several days of incubation with substrate, but no change was observed in the copper coordination. Perhaps the transient nature of the binding, over a long timescale is not sufficient to cause an observation change in the EPR spectra. Interestingly, it should be noted that one of the non-native CBMs used in the study by Crouch *et al*, was in fact a CBM10 domain, and it was found that incorporation of this type of module was able to increase the number of non-oxidised products formed through LPMO action. The authors suggest this to be due to the CBM10 directing the LPMO to the reducing ends of the chain, whereby C1 oxidation is able to produce a non-oxidised product.[222] *Tt*AA10 was found to produce non-oxidised products

in a smaller ratio to the oxidised oligosaccharides. Perhaps this is a non-coincidental pairing, but expression and activity analysis of the full gene would be required to assess whether there is a link between these ideas.

6.6.6 Structure Solution

Whilst the crystal morphology was not improved by optimisation of the conditions, *Tt*AA10 crystals diffracted to a high resolution. Indeed, the data set collected at the Diamond synchrotron displayed highest resolution at 1.17 Å in the outer shell. Despite this, the completeness of the model at this resolution was extremely low and data was reduced to 1.4 Å as a compromise. Initial efforts to solve the structure of *Tt*AA10 failed when using molecular replacement. The closest sequence homology match of *Tt*AA10 to a model in the PDB was that of *Cj*AA10 (30%, PDB code, 5FJQ), a chitin active AA10. Molecular replacement using this model as well as truncated versions with certain aspects of the model removed (*i.e* water molecules, flexible loops etc.) as well as combinations of other models (models of lower similarity combined to create an ensemble) failed to produce a solution. It should be recalled that LPMO homology centres around specific features of the structure, whereby the active site copper-histidine brace is invariant and the general fold and shape are maintained, but beyond that there can be large differences. The difficulty with molecular replacement suggested that the structure of *Tt*AA10 would be considerably different to other AA10s in some respect other than the core features.

In a strong coincidence, crystallisation was carried out using only the first batch of protein produced, which was shown to contain two individual copper species by EPR. The structure of *Tt*AA10 was solved by phasing, where the copper anomalous signal was used to produce an electron density map. Even though the data set was not collected at the copper edge, the wavelength normally used to carry out phasing with copper atoms, the high amount of anomalous signal resulting from two copper atoms per molecule was strong enough to enable structure solution at 1.4 Å. In fact, after structure solution it became clear that the non-removable strep tag was actually serendipitous in forming a crystal lattice; a histidine on the strep tag was found to create a crystal contact through coordination to the secondary copper site of another molecule. Considering the poor crystallization hits produced by *Tt*AA10 it is likely that without this crystal lattice contact, crystals would not have formed. The protein had been previously analysed by SECMALS to check the protein molecular weight, and this information also proved useful in ruling out the possibility that this coordination interaction was strong enough to occur in solution. SECMALS showed that the

protein existed as a monomer in solution and as such it is likely that the histidine residue observed binding in the crystal structure is replaced by a water molecule in solution.

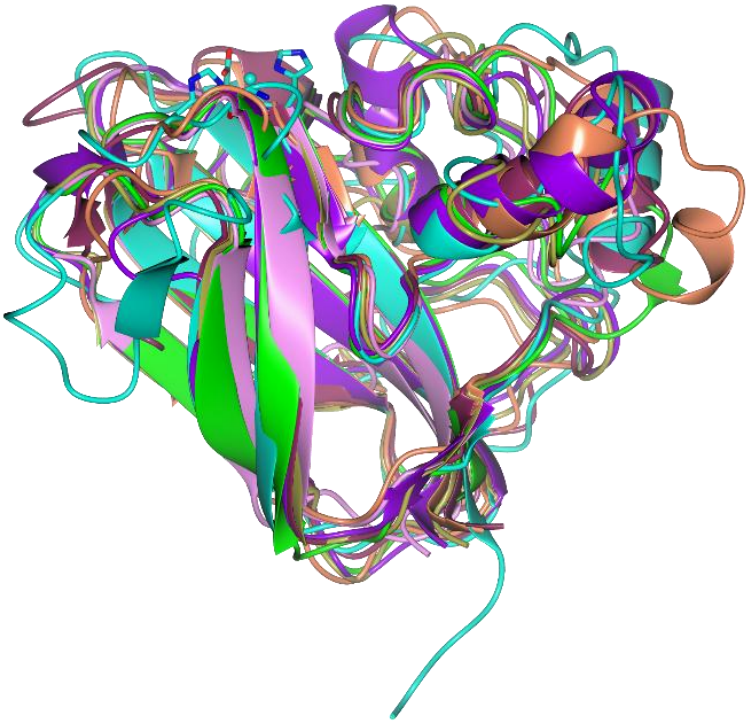


Figure 136 Whole structure overlay of several AA10s showing the general conserved fold ; *Tt*AA10 (cyan), *Ba*AA10 (purple, PDB code; 5IJU), *Cj*AA10(A)(dark purple, PDB code; 5FJQ), *Ef*AA10(A) (gold, PDB code; 4ALC), *Sc*AA10(Bc) (coral, PDB code; 4OY6), *Jd*AA10(Ad) (pink, PDB code; 5AA7), *Sl*AA10(E) (green, PDB code; 5FTZ). The active site of *Tt*AA10 is shown in cylinder view. Image made in CCP4mg.[55]

The structure of *Tt*AA10 was found to adhere to the typical AA10 structural features observed in other LPMO proteins, **Figure** 136. The histidine brace is positioned on the flat surface of the protein, created by the top region of the β-sandwich core. Overlays of the *TtAA10* structure with the highest homology model (*Cj*AA10, 5FJQ) showed significant differences in the loop regions of the protein, whilst the core β-sheets were conserved. The T-shaped coordination geometry around the copper created by the histidine brace was very similar to other structural models, **Figure** 137, and parameters such as bond length as quoted for other LMPOs. The bond lengths between the copper ion and the ring nitrogen of His1, amino terminus nitrogen of His1 and the ring nitrogen of His107 were 1.9, 2.2 and 2.0 Å respectively. The bond lengths were found to match well with values reported for Cu-histidine brace coordination in many other LMPO models as collated by Ciano *et al.*[39] The ligand sphere surrounding the histidine brace was similar to other AA10s in that there were two residues found in the apical positions. Distances measured between the copper ion and the residues Phe193 (ring carbon atom closest to Cu), Gly107 (oxygen atom of main chain)

as 3.5 and 3.7 Å respectively. The phenylalanine was found, as expected, presenting from underneath the copper-histidine brace. It has been found that AA10s either have a phenylalanine or tyrosine residue in this position from sequence homology and structural comparisons. The incorporation of a hydroxyl group in the tyrosine side chain compared with phenylalanine, making it possible for ligand coordination to the copper Several AA9 and one AA10 structures quote the Cu-OH distance as 3.3 Å, whilst two AA10 structures display a shorter distance of 2.5 Å, indicating a likely difference in the copper coordination state. The second apical position was taken up by a glycine residue in *Tt*AA10, which is uncommon based on the current literature data, where AA10s normally present with an alanine residue in this position, **Figure** 136. Interestingly, the only known structure of an AA13 LPMO, active on starch with a grooved binding site, also has a glycine in this coordination position.[224] The position of the glycine, whose side chain simply lacks a carbon linkage compared to alanine, may be the reason as to why the EPR signal of the active site was deemed to be more similar to EPR spectroscopy of AA9 proteins; whereby these descriptions come from analysis of where the copper coordination aligns in the Pleisch-Blumberg plot. LPMOs copper coordination can be split into type based on the coordination form, type 1 or type 2. However, some LPMOs displaying rhombic character in the copper coordination geometry, fall in-between the two classic copper types. In AA9 LPMOs, the apical coordination position is taken up by a water molecule, and it is thought that the positioning of the alanine in AA10s near the copper prevents typical geometric coordination of a water molecule[65]. Alteration to a glycine residue, may allow enough space for the binding of water (or another species such as chloride), although evidence for this was not observed in the crystal structure.

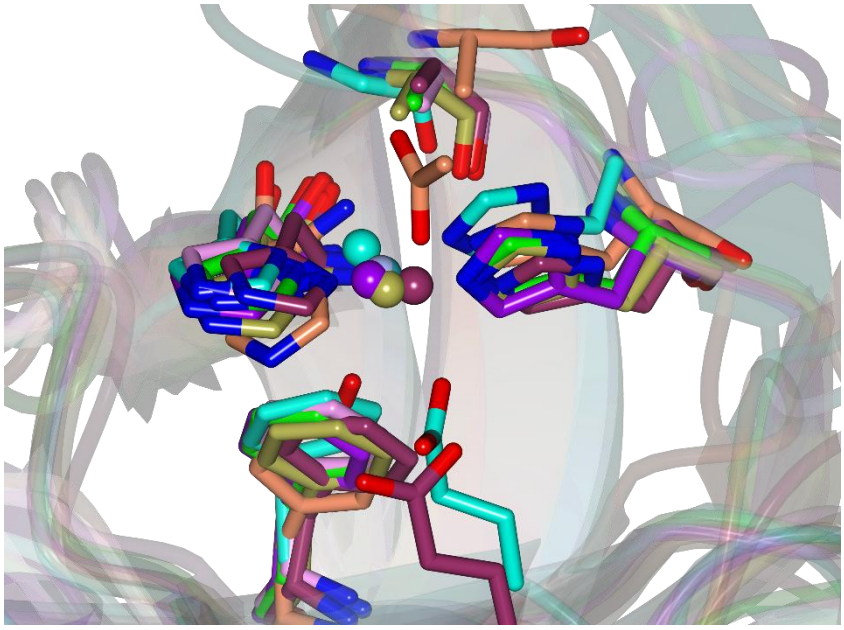


Figure 137 Overlay of several AA10 structures with a focus on the active site residues ; *Tt*AA10 (cyan), *Ba*AA10 (purple, PDB code; 5IJU), *Cj*AA10(A)(dark purple, PDB code; 5FJQ), *Ef*AA10(A) (gold, PDB cpde; 4ALC), *Sc*AA10(Bc) (coral, PDB code; 4OY6), *Jd*AA10(Ad) (pink, PDB code; 5AA7), *Sl*AA10(E) (green, PDB code; 5FTZ). *Sc*AA10(Bc) is modelled with an acetate ion coordinating from above the copper. Apart from *Tt*AA10, only *Cj*AA10 displays a Glu residue in the secondary ligand sphere of the active site. Of the structures available, only *Sc*AA10 has a Tyr instead of Phe in the apical position. Only *Tt*AA10 has a glycine where all other structures display an Ala. Image made in CCP4mg.[55]

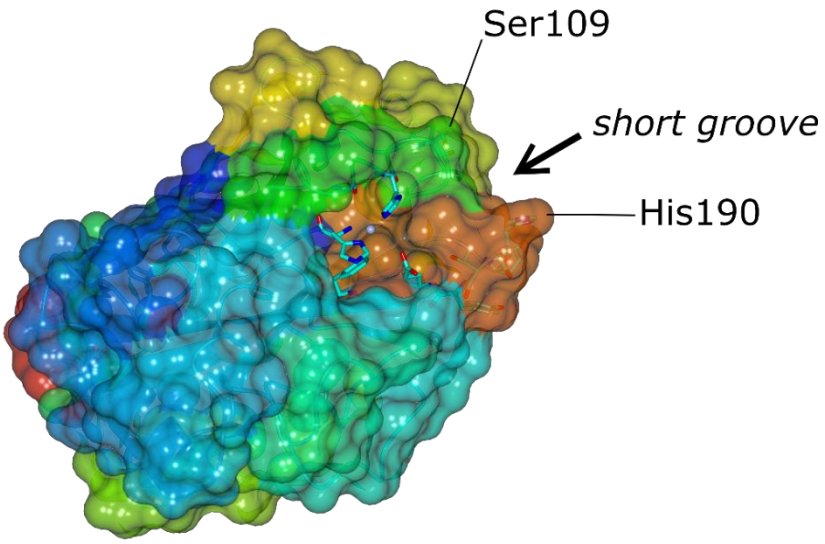


Figure 138 Coloured surface (blended through the model blue-red) of *Tt*AA10 showing the exposed active site (no surface, for clarity) and the potential shallow groove created by His190 and Ser109.

On closer inspection of the planar binding surface of *Tt*AA10, two protrusions exist on one edge of the surface. These protrusions are formed by a single histidine (His190) side chain sticking out into the solvent space around the molecule and a serine residue (Ser109), Figure 138. The two residues form a small channel which points towards the active site. Bound soluble oligosaccharides in crystal structures of *Ls*AA9, has shown that the substrate binds across long edge of the copper active site, flanked by interacting residues on both side.[201] This binding mode is most likely to occur in *Tt*AA10, with interacting residues along the length of the flat surface. Orientation EPR, using a fixed substrate with a defined axis in reference to experimental set up has also shown that the copper active site of *Ls*AA9 binds in a parallel sense to the cellulose chain (Ciano and Walton, unpublished). Hence, it would be expected for the substrate to bind in the same fashion to *Tt*AA10, which would likely involve the polysaccharide chain running through this small His-Ser channel towards the copper active site. The channel may be able to 'select' a specific cellulose chain through hydrogen bonding interactions provided by the oxygen of serine and ring stacking of the histidine, positioning the chain so that it binds over the copper active site.

6.7 Conclusion

This work has successfully proved the hypothesis that the LPMO gene expressed by *T. turnerae* is active on cellulose – through culture growth experiments and testing of recombinant *Tt*AA10 enzyme activity on various plant polysaccharides. *Tt*AA10 has been shown to lack regioselectivity over which carbon position around the glycosidic bond the oxidative chemistry occurs at. Mass spectrometry strongly indicated the presence of species formed through C1, C4 and both C1 and C4 catalytic oxidation events on PASC and Avicel.

The structure solution of *Tt*AA10 was carried out by copper phasing which relied on the strength of two copper species for SAD phasing; a crystal contact between two molecules was formed by a histidine residue of the strep tag of one molecule coordinating with a second copper site on the surface of a second molecule. This second metal binding site was probed by EPR and found to contain copper in the first batch of protein, but not in subsequent batches. The second copper site was found not to affect activity of the protein. EPR experiments in which titrations of different metals could not displace the bound copper species, and copper titration observed binding of the second species after 0.4 equivalents of copper had been added to the sample. It was deemed likely that the second metal site was an opportunistic copper binding site and did not contribute significantly to the LPMO itself. However, further analysis is required to determine if this is truely the case; the positioning of the second metal site in such close proximetry to the active site of *Tt*AA10 may constitute a potential binding site for either an electron donating protein partner, or small organic reducing agent. Efforts to pull out a protein binding partner from cultured *T.turnerae* using immobilised *Tt*AA10 did not work, but the experiment does not rule out whether such a protein binding partner exists – the binding nature may be too transient to capture any potential AA10 electron donating agent.

The structure fits all the classifications to be 'LPMO-like' in that *Tt*AA10 forms a small globular structure with a central β-sheet core surrounded by various loops, which support a flat binding surface in which a centrally placed copper active site it positioned. The loops infer significant structural differences to the protein compared to other AA10s. The binding site of *Tt*AA10 was planar, with the copper active site centrally placed and near to a slight surface protrusion creature by two residues His190 and Ser109, which may form a shallow channel through which a cellulose chain may bound.

7

Conclusions and Perspectives

7.1 In context

For a better tomorrow the world must move towards sustainable fuel sources. Biomass, especially that considered waste or '2nd generation' feedstocks, provides an ample resource from which high value carbohydrates can be obtained. Accessing the full chemical potential of these energy-rich carbohydrates requires the controlled breakdown of a heterogeneous mixture of many different natural recalcitrant polymers. However, being biological in origin means that these complex polymers are also susceptible to breakdown by nature's many tools, namely lignocellulosic enzymes. From this perspective, a wide variety of glycoside hydrolases are produced by many different organisms which specialise in the breakdown of various types of polysaccharides; their mode of action uses hydrolysis to break apart individual glycosidic bonds between sugar units.[15, 17]

Lytic polysaccharide monooxygenases are a comparatively new addition to our knowledge of how organisms break down recalcitrant materials. LPMOs use copper mediated oxidative chemistry to introduce chain breaks in crystalline regions of polysaccharides, regions typically inaccessible by GHs.[40] Our understanding of lignocellulosic breakdown in nature is that a consortium of different enzymatic activities is required to effectively cover the vast number of potential substrates held within biomass. Indeed, one organism may specialise in production of a certain type of enzyme specific for a substrate that other organisms may not be able to efficiently degrade; however, together the combined efficiency of multiple organisms is effective at realising all the potential energy sources from a material. Therefore, we must find combinations of enzymes that can work synergistically, on a given substrate, in an industrial process. In this regard, nature provides a wealth of information on how to carry out the process of biomass degradation to fermetable sugars which can be

processed further into high value molecules such as ethanol. Enzymatic degradation in industrial processes is an area where scientists can focus to improve efficiency and bring down overall costs of running bio-renewables plants. Research often looks to organisms which inhabit very specific niches. For example, the most commonly used organism in industrial enzymatic processes is *Trichoderma reesei*, described as the 'gold standard' in lignocellulose degradation by Bischof *et al*; who describe, in a review, the 70s years' worth of advances since discovery of *T. reesei* (originally named *T. viride* [225] and often now known as *Hypocrea jecorina* in some literature) an organism isolated during World War II, where US army tents in the Solomon Islands made of cellulose fibres were being ravaged by a destructive filamentous fungus.[77] 70 years of research on this organism has led to industrial strains capable of producing 100 g L^1 cellulytic enzymes. Nevertheless, notwithstanding the efficiency of *T. reesei,* and the many improvements carried out over the years through strain development,[226](reviewed by Peterson *et al*[227]) we need to seek ways of improvements, in which a promising method is to artificially combine multiple different enzymatic activities from different organisms that can work synergistically with the variety of polymers found in biomass.[77]

In this regard, *T. Turnerae* is an organism of great interest. It is a symbiotic bacteria found in the gills of the shipworm; an animal which uses its clam like head to burrow into submerged wooden substrates, living off the wooden material as a food source. Knowledge of *T. turnerae* has existed since the 1980s where it was isolated and successfully grown in liquid culture on a cellulose substrate.[79] A burst of interest in this symbiotic-host relationship occurred in the early 2000s when the genome of *T. turnerae* was published. From then it became known that, of the large variety of CAZymes within the genome of this bacterium, over 50% were thought to function as lignocellulosic related or degrading enzymes, making it one of the leading bacteria in the competition for 'best lignocellulosic enzyme producer'.[101, 103] Questions still remain as to why the bacteria are housed so far away from the site of digestion – whether their primary role is in the fixation of nitrogen and the production of lignocellulosic enzymes is simply a secondary role is unknown. How the shipworm signals the bacteria to express the correct enzymes for digestion of its food substrates is an interesting line of enquiry and, if so, how are these enzymes selectively transported to the gut of the animal? Investigating the relationship of the shipworm with its symbiotic microbial communities is an interesting topic on its own merit and theories about this coexistence have been updated recently by work from Sabbadin *et al*, who look at the role of endogenous proteins in shipworm digestion.[70] Investigating the shipworm means

studying two co-existing organisms, which have co-evolved to be highly efficient within environments where survival is based on the ability to obtain energy from typically recalcitrant materials. As such, studying the role of *T. turnerae* in digestion of recalcitrant substrates is likely to provide new and novel enzymatic functions which may benefit improvements in knowledge or industrially processed renewable bio-fuels.

7.2 7.2 This Work

7.2.1 Target Success

Over the course of this work the production of several likely lignocellulosic proteins was attempted using recombinant expression in *E. coli*. The genome of *T. turnerae* encodes many different GHs which most likely cover a wide range of functions. This work selected 14 GH ORFs using information available on the CAZy database, through which the carbohydrate active enzymes of individual genomes are grouped together and classified into their relative GH families.[34] The genome of *T. turnerae* was known to contain 245 CAZymes, over 50% of which were expected to be involved in the degradation of biomass.[103] A wide variety of GH5 proteins were encoded and several of these, where BLAST searches suggested differing activities, were chosen as targets. A couple of ORFs were classified as proteins which were the sole representative of a particular GH family within the *T. turnerae* genome and, as such, two proteins from the GH8 and GH12 families were selected. A single LPMO gene is found in the genome of *T. turnerae*, which made for an obvious research target based on the high level of industrial interest in these powerful oxidative enzymes. Attempts to produce all 15 target proteins were carried out using different expression and purification strategies. As small quantities of *Tt*AA10 had been produced prior to the start of this project (by Dr G. Hemsworth), optimisation of protein production was trialled with a focus on creating fusion constructs containing various solubility and affinity tags. Of the 15 target sequences, six proteins were successfully produced in a soluble form and taken onto further characterisation, as shown in Table 26. The GH constructs were designed to produce only the catalytic domain, and these were chosen based on sequence analysis (and confirmed with CAZy module boundaries kindly provided by Professor B. Henrissat) and involved a simple hexahistidine tag engineered to be removable by 3C protease.

Of the remaining sequences, one protein, *Tt*GH5_53 (GH5-0427, *ACR13327.1*) was successfully produced in a soluble form using the same purification strategy as outlined to produce the 5 other GHs. However, the protein degraded during storage before sufficient

characterisation of substrate specificity and function could be determined. The other sequences, pertaining to proteins from GH families 5, 6, 9 and 45 were not produced in a soluble form, as shown in Table 26. As discussed in Chapter 2, production of inclusion bodies was observed for some of the remaining eight protein constructs and optimisation of expression and purification was not sufficient to capture these proteins in a soluble form for further analysis. Despite this, the remaining eight proteins continue to be of research relevance; further changes in expression system or alterations in construct design are likely to result in improvements in production of these proteins. As shown in the main body of work, the characterisation of six soluble recombinant *T. turnerae* CAZymes was carried out and involved a principal focus on substrate specificity, activity and, where possible, structural analysis.

Table 26 Carbohydrate active enzymes selected from the genome of *T. turnerae* for analysis. Those highlighted in bold and orange were successfully produced and characterised and their functions updated from 'potential' to characterised. Those written in grey were not characterisd, and their functions remain based on sequence homology.

Name	ORF	Genbank Code	Family _subfamily	Function	Characterisation carried out in this work
AA10	**TERTU_0046**	***ACR14100.1***	**AA10**	**Type 3 AA10 (C1/C4 cellulose cleaving)**	**Structure, MALDI product profile, TSA**
GH5-2895	**TERTU_2895**	***ACR12145.1***	**GH5_2**	**Cellulase**	**Kinetic assays, MALDI product profile, TSA**
GH5-3565	TERTU_3565	*ACR11017.1*	GH5_1	B-1,4-glycan cleaving enzyme	-
GH5-0183	TERTU_0183	*ACR12128.1*	GH5_26	Endo β-1,4-glycanase	-
GH5-3751	**TERTU_3751**	***ACR11279.1***	**GH5_un**	**Cellulase**	**Kinetic assays, MALDI product profile, TSA**
GH5-3361	**TERTU_3361**	***ACR12247.1***	**GH5_4**	**Xyloglucanase**	**Kinetic assays, MALDI product profile, TSA**
GH5-0428	TERTU_0428	*ACR12792.1*	GH5_53	Cellodextrinase/β-glycanase	-
GH5-0427	TERTU_0427	*ACR13327.1*	GH5_53	Cellodextrinase/β-glycanase	-
GH6-2898	TERTU_2898	*ACR12723.1*	GH6	Cellobiohydrolase	-
GH6-3996	TERTU_3996	*ACR14000.1*	GH6	Cellobiohydrolase	-
GH6-2895	TERTU_2895	*ACR12145.1*	GH6	Cellobiohydrolase	-
GH8-4506	**TERTU_4506**	***ACR14722.1***	**GH8**	**Endo-β-1,4-xylanase**	**Structure, mutant/native kinetic assays, MALDI product profile, TSA**
GH9-0607	TERTU_0607	*ACR11786.1*	GH9	Cellulase/endo-β-1,4-D-glucanase	-
GH9-0645	TERTU_0645	*ACR14629.1*	GH9	Endoglucanase	-
GH12-0353	**TERTU_0353**	***ACR14297.1***	**GH12**	**Activity not determined**	**Structure, TSA**
GH45-3400	TERTU_3400	*ACR13005.1*	GH45_A	Endoglucanase	-

Note: GH5_2895 and GH6_2895 are now described in the CAZy database as bifunctional β-1,4-endoglucanase / cellobiohydrolase (CelAB), with combined genbank code ABS72374.1

7.2.2 GH5

The three GH5 proteins, *Tt*GH5_2, *Tt*GH5_4 and *Tt*GH5_un, represented three different GH5 subfamilies, two of which contain several members with known substrate activities. Members of the GH5_2 subfamily are mostly glucanases and often exist as part of multimodular domains. In the genome of *T. turnerae* the *Tt*GH5_2 ORF is part of a bi-functional gene construct whereby the GH5_2 domain is positioned between two CBMs from families 5 and 10, upstream of a GH6 module. During the writing up process of this work it came to light, through updates to the CAZy database, that partial experimental analysis of the full gene construct had been carried out previously by Ekborg *et al*.[105] Ekborg and colleagues successfully produced and tested the activity of the gene, named as CelAB, which contained the two suspected GH domains and CBM domains. The multimodular CelAB protein was suggested to be celluloytic through endoglucosidase and cellobiohydrolase activity for GH5 and GH6 respectively. Indeed, their work supported the expected activities suggested based on sequence analysis. However, the work in this thesis focused on the characterisation of catalytic domains only, hence, both the GH5 and GH6 modules (originally written with CAZy codes GH5_2895 and GH6_2895 at the start of this work) were expressed individually. Ekborg *et al*, did state a need for further study to clarify the individual activities of each GH domain in the multimodular protein complex.[105] In this work, the GH6 domain was unsuccessfully produced as a separate catalytic domain but *Tt*GH5_2 was produced in high amounts and its activity studied against several polysaccharides. Analysis of *Tt*GH5_2 found that the catalytic module retained its activity on cellulosic substrates, breaking bMLG down to cellotetraose, and kinetic studies on 4-MU-C3 showed a high level of activity with k_{cat}/K_m measured to be $2.43 \times 10^{-2} \pm 7.2 \times 10^{-3}$ $\mu M^{-1} s^{-1}$. Initial experimentation using mannans had shown that *Tt*GH5_2 was capable of degrading glucomannan but not ivory nut mannan or manno-oligosaccharides. The glucomannan activity was probed through collaboration with Dr T. Tryfona and Prof. P. Dupree at the University of Cambridge, who used PACE to analyse the product profile of glucomannan degradation. This analysis confirmed the suspected activity profile, that the catalytic domain *Tt*GH5_2 was able to tolerate mannose moieties in its active site, but only cleave at glucose residues. Interestingly, this analysis found *Tt*GH5_2 to be active on PASC and Avicel, where a broad array of products were observed from DP 4-11. The previous study by Ekborg *et al* suggested that the multimodular construct was only active on amorphous regions of cellulose.[105]It is possible that the reduction in protein size may have allowed the protein to have obtain activity against

crystalline substrates through better access to the substrate. However, larger oligosaccharide products were being produced in the degradation of the crystalline cellulose substrates, suggesting that incomplete degradation of the substrate was occurring, or that the protein may have only been able to access certain regions of the substrate. Removal of the CBM domains may have lifted some degree of substrate restraint; the two CBM domains in the full protein sequence may restrict the activity of the multimodular construct to substrates upon which both GH domains may act synergistically. Therefore, the separation of *Tt*GH5_2 into a single protein may have extended its endo-acting ability. As such, whilst characterisation of the *Tt*GH5_2 had been carried out previously, unknown to the author at the time of experimentation, this work has provided more insight into the individual function of one of the GH domains; where *Tt*GH5_2 was not only confirmed to be an endo-glucanase, but also shown to be able to tolerate mannose to enable effective degradation of glucomannan, and as an individual protein, present activity on crystalline substrates.

Sequence analysis of *Tt*GH5_4 had predicted that the enzyme was likely to exhibit activity on xyloglucan, with the majority of GH5 xyloglucan activities found in this subfamily. Indeed, this was found to be the case during activity assays, where *Tt*GH5_4 was only able to degrade xyloglucan. The product profile as analysed by MALDI-TOF-MS confirmed the presence of typical degradation products associated with GH5_4 family activity on xyloglucan.

*Tt*GH5_un remains a particularly interesting target, as the sequence at the time of this work is not classified into a subfamily of GH5. As such, this protein likely represents a new GH5 subfamily, with biochemical analysis of *Tt*GH5_un suggesting members will have activity on glucans. The activity of *Tt*GH5_un was in fact similar to *Tt*GH5_2, where the protein was found to tolerate inclusion of mannose units within the substrate but only cleave at glucose residues (as shown by activity assays and PACE). Crystallisation of TtGH5_un was carried out and hits were successfully optimised to produce crystals which diffracted to approximately 1.8 Å. A data set was collected at the Diamond Light Source, but unfortunately no structure solution was determined. The lack of significant homology (< 30 %) with models available in the PDB meant that the structure could not be solved by molecular replacement. Future work will likely need to rely on structure solution via phasing to bypass the lack of available homology structures. Despite this unsuccessful structural outcome, *Tt*GH5_un remains a protein likely to dictate the direction of future GH5 analysis.

7.2.3 GH12

The only GH12 within the genome of *T. turnerae* was successfully crystallised and modelled at 1.6 Å using molecular replacement. The structure was found to display the classic 'jelly roll' tertiary fold associated with the GH family 12. A protein-ligand complex, solved at 1.9 Å, mapped out a portion of the active site, where the disaccharide Glc β–1,4 Glc β–1,4 noeuromycin (generally a cellulase inhibitor) was found to bind in the -2 to -1 subsites, with the glucose moiety ring stacking with nearby tryptophan residues, two of which are rotated compared to other GH12 enzymes. *Tt*GH12 contains an alteration in the normally conserved catalytic triad, and it sits with two other sequences as a single branch within the GH12 phylogentic tree. Substrate specificity analysis did not yield any consistent activity, however protein melting temperatures were increased on addition of xylan and xyloglucan suggesting some interaction was occurring between the substrate and the protein. Activities within the GH12 family consist mostly of cellulases and some xyloglucanases. The lack of confirmed activity of *Tt*GH12 coupled with the interesting alteration in the active site, and the unusual binding of the disaccharide indicate a new subclass for the GH12 family, which will be an interesting branch for future studies.

7.2.4 GH8

The most well characterised target protein within this work was TtGH8, data on which has been recently published. The protein was successfully characterised in terms of function, substrate specificity, efficiency and structure. Interestingly, *Tt*GH8 was the only GH8 within the genome of *T. turnerae* and shown in this work to be a highly efficient degrader of xylose based polymers. *Tt*GH8 catalysed the hydrolysis of β-1,4 xylohexaose with a k_{cat}/K_m value of 7.5×10^7 M^{-1} min^{-1} but displays maximal activity against mixed-linkage polymeric xylans, with a k_{cat}/K_m value of $1.6 \times 10^8 \pm 4 \times 10^6$ mg^{-1} ml min^{-1}, hinting at a possible primary role in the degradation of marine polysaccharides. Structural analysis of *Tt*GH8 showed a classical tertiary $(\alpha/\alpha)_6$ barrel fold, typically associated with the family. Ligand soaking experiments were able to bind X2 and X3 in the negative subsites of the active site, where the binding of X3 showed a conformation in the -1 subsite consistent with the proposed catalytic itinerary. Mapping of the full binding site was achieved through creation of a catalytic mutant, in which the catalytic base Asp281 was altered to Asn. However, this mutant still retained a small amount of activity relative to the native protein, and was still efficient at cleaving bMLG with a k_{cat}/K_m value of $1.8 \times 10^4 \pm 1 \times 10^3$ mg^{-1} ml min^{-1}. Rapid crystal soaking with X6

yielded a structure whereby X6 was bound from the -3 to +3 subsites and displayed a flipped chair conformation in the -1 position. The analysis of *Tt*GH8 showed it to be stable and effective at degrading xylans and thus its potential use in biomass degradation processes could be explored by testing the activity of the protein on real heterogeneous materials likely to be used in biomass processing.

7.2.5 AA10

The only LPMO produced by *T. turnerae*, known to be an AA10, was analysed for substrate specificity and found to exhibit binding capacity for chitin (as predicted from sequence analysis) but showed no catalytic activity on this substrate. Instead activity was observed on crystalline cellulose substrates PASC and Avicel. Growth of the source bacteria *T. turnerae* in liquid culture using a cellulose substrate as a food source found the LPMO to be abundant in the media by proteomic analysis. Activity assays whereby the enzyme and substrate were mixed with reducing agents (ascorbate or gallic acid) showed a product profile consisting of DP4-11. The product profile as analysed by MALDI-TOF MS suggested that *Tt*AA10 was able to oxidatively attack at either the C1 or C4 position, thus further experimentation (i.e. enzymatic digestion, NMR or permethylation [228]) should be used to confirm the 'lack' of regioselectivity, as some ambiguity was present due to MS peaks of different products having the same *m/z* value. Thus, it is tentatively suggested that *Tt*AA10 is a Type 3 AA10. Structural analysis of *Tt*AA10 was carried out and a 3D-model at 1.4 Å was solved using copper phasing which took advantage of the presence of two copper ions in the structure. Classic LPMO features were observed, including the immunoglobulin-like fold and presence of a histidine brace. Furthermore, a second copper site was present on one side of the structure. Significant structural differences, likely from the loops in the protein, meant solution by molecular replacement with other AA10 molecules was not possible.

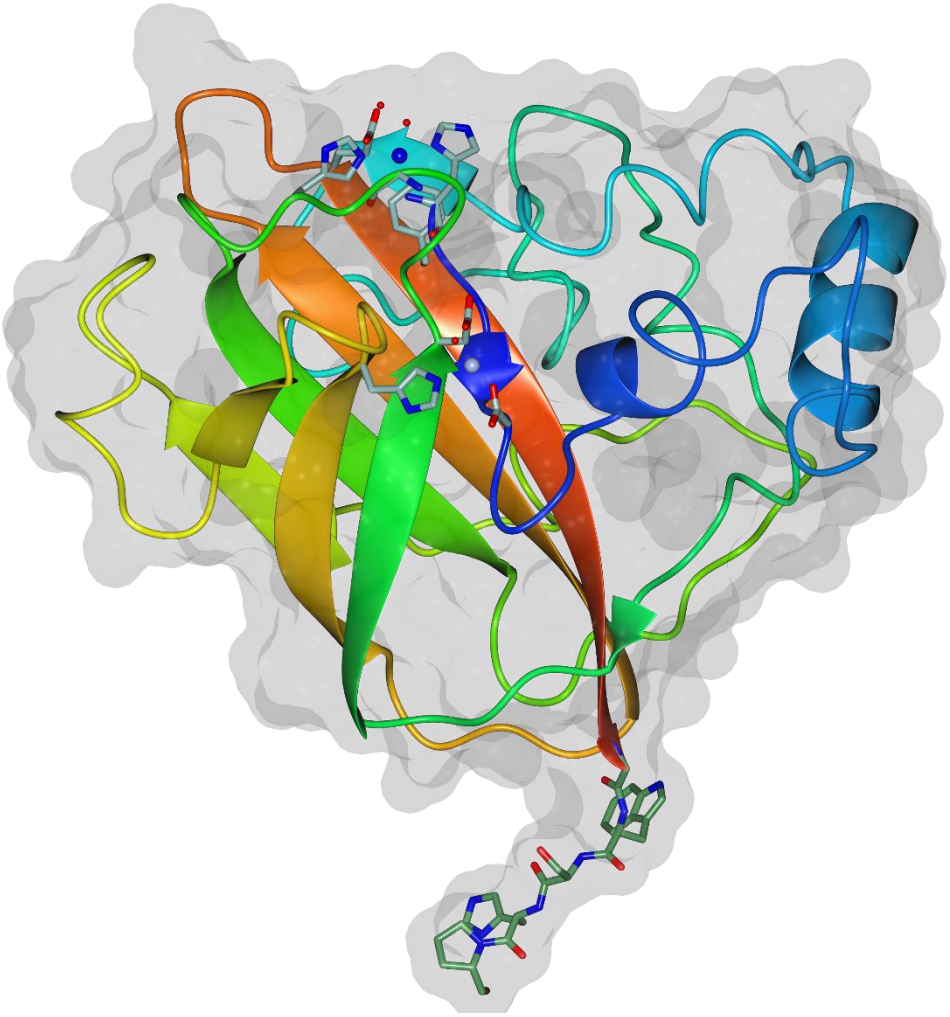


Figure 139 Structure of TtAA10 showing both copper sites. Top is the active site, whilst the front centred is the secondary metal site. The strep tag, shown as cylinders. A histidine from the strep tag of another molecule makes a crystal contact through coordination with the second copper site.

Spectroscopic analysis to assess the copper active site was carried out by frozen solution CW-EPR spectroscopy. Spin Hamiltonian parameters (obtained from simulations by Dr L. Ciano) showed that the active site was similar to other AA10s within the literature; with g_z and A_z values equal to 2.267 and 420 MHz respectively. Occupancy of the second copper site was variable between protein batches, but the copper was found to be taken up during EPR monitored titrations. A change in the secondary coordination sphere of the copper was observed crystallographically with a glycine residue found in the apical position around the copper ion instead of a typically conserved alanine normally present in AA10 LPMOs.

The second copper site may act as a binding site for an electron donating protein partner, similar to how CDH provides electrons to AA9 LPMOs, or small molecule reducing agents. A pull down experiment using immobilised *Tt*AA10 was carried out in an effort to 'catch' any

such AA10 electron donating partner, but this was not successful. Repetitions of this experiment should be carried out however in an effort to elucidate whether there may be any transiently binding protein partners.

It would be prudent to determine the enzymatic efficiency of *Tt*AA10 acting on cellulose. However, the insoluble nature of the substrate complicates this task, preventing experimentation normally used to gain kinetic parameters such as those determined for the GHs in this study. One can instead focus on the rate of sugar release by coupling LPMO activity with traditional GHs and monitor the 'boosting effect' to determine the synergistic[38, 68] capabilities of this AA10 with industrial enzyme mixtures. Gaining kinetic insight on LPMO reactions has historically been difficult due to the insoluble nature of the substrates, but recent techniques have shown methods which bypass these problems.[229, 230] Determination of *Tt*AA10 kinetics would therefore assess the efficiency of the individual enzyme but also provide LPMO activity parameters to a field which has so far lacked sufficient amounts of such data for comparison. Finally, in a similar vein to the discussion of *Tt*GH5_2, only the catalytic domain of the *Tt*AA10 gene was characterised. A CBM binding domain from family 10 is appended to *Tt*AA10 in the native sequence. Sequence homology suggests that this CBM10 domain is likely to bind cellulose and it would be interesting to assess the effect of this binding domain on the activity of the protein. Analysis of *Tt*AA10 has provided more information towards the ever-growing knowledge of LPMOs, as well as structural insight into an LPMO with a slightly variant active site.

7.3 7.3 Future Directions

7.3.1 The enzymes

The symbiosis between the shipworm and *T. turnerae* represents a very small area of research from which new enzymatic functions can be found. The availability of the *T. turnerae* genome initiated this research into 15 carbohydrate active enzymes, of which six were successfully characterised, through what is essentially a process of genome mining; analysing how genomic sequences relate to characterised homologues can suggest potential avenues of inquiry. However, as shown by the analysis of *Tt*GH5_un and *Tt*GH12, novel proteins with potentially new functions can be still be discovered, even if classifications into their likely activities have already been assigned. Characterisation of other enzymes with activities well-known in the literature, such as those from families GH8, GH5_2 and GH5_4 still provide valuable insight into enzymatic function and the knowledge gained through this

work adds to the ever-increasing scientific understanding of the workings of the various types of glycoside hydrolases. Research on the selection of ORFs made during this work, however, was limited by time and resources and, as such, there remains a large pool of likely lignocellulosic enzymes encoded by this interesting shipworm symbiont, which can be probed for their functions and activities. Indeed, the expression of *T. turnerae* in pure culture carried out in the analysis of *Tt*AA10 was an interesting experiment as it provided insight into a variety of proteins that were secreted into the media when the cells were being fed on a source of cellulose. Whilst the experiment was not optimised, *Tt*AA10 was found in the media in 'high' amounts, as well as some other GHs – however, none of the genes selected for characterisation in this work was identified. A protein in high abundance in the media was classified as a GH1, based on BLAST[113] characterisation and thus suggested to be active on xylan by sequence homology analysis, yet was expressed whilst the bacteria was feeding on cellulose! This experiment suggests there may be more enzymes of interest within the genome of *T. turnerae*, which could be identified through similar experiments; by monitoring the secreted proteome in response to different types of substrate, gives clues to up-regulation of enzyme activities and potentially identifies new enzymes of interest.

An original thought prior to this culture growth experiment was that it may help identify a possible electron donor of *Tt*AA10, an area of importance for research into AA10 LPMOS. AA9 LPMOs, from fungal systems have been suggested to work in partnership with celliobiose dehydrogenase (CDH) [204, 231], which are often co-secreted at the same time as LPMOs. They have been shown to be able to provide electrons to AA9 enzymes, likely through electron transfer pathways involving their own haem domain.[205] However, not all fungal systems use CDH and some species encode multiple CDH genes, thus coupling of CDH and LPMOs may not be a simple 'one size fits all' approach.[232] Analysis of secretome data (like that collected for *Tt*AA10) is vital to understanding the coupling process, as shown recently by Adlakha *et al*, who identified a CDH redox partner of a specific AA9 from *Botrytis cinerea* using a holistic approach to secretome analysis.[232] Such enzymatic electron donating partners have not yet been identified for AA10 enzymes, as bacteria lack expression of celliobiose dehydrogenase and an equivalent has remained elusive. Potential external electron donors (useable by both fungal and bacterial systems) may come from phenolic compounds (i.e. gallic acid), compounds released upon degradation of lignin,[233] lignosulfonates [234] oxioreductase type enzymes [235] and small molecule reducing agents such as ascorbate. Small molecule reducing agents are the common choice in experiments, where they 'artificially' provide LPMOs with a source of electrons, but this system may not

necessarily mimic LPMO systems in nature, as the general availability of small molecule reducing agents during biomass degradation is unknown.[235] The secretion of *Tt*AA10 into the media during culturing of *T. turnerae* on cellulose does suggest that the protein was in use during biomass degradation, and accordingly an electron donating partner would likely be found in the media. One suggestion had been that the second metal site observed on *Tt*AA10 could have played a role in protein binding or docking, followed by electron transfer – the two copper sites were only 11 Å away. Immobilisation of *Tt*AA10 on a column, failed to 'trap' any potential binding partners, although if occurring, binding is likely to be transient.

As shown in the introduction to this work, LPMOs are a hot topic in terms of research, with new classes and substrate specificities emerging regularly, and as such there is a wide range of interest in their chemical action and abilities. Electron donating partners are one of those currently elusive topics regarding AA10 enzymes, and identification of what actually provides electrons to these powerful oxidative enzymes may influence how they are used in human-directed situations – specifically, their use in industrial enzyme cocktails during biomass degradation.

Further topics of interest relate to understanding how the LPMO active site works, in a mechanistic sense. The potential mechanisms of LPMOs have not been discussed in detail during this work. Debates about the role of activating molecules, for example oxygen vs. hydrogen peroxide is underway and identification of the *true* activator of the copper site is as yet unproven.[236-237] Different theories use different sequential combinations of LPMO, substrate, reducing agent and activator to present catalytic itineraries which go through various copper states and copper-oxygen intermediates – all of which are chemically plausible, but lack experimental confirmation.[39] Further complications arise when one considers that different classes of LPMO may present catalysis through different pathways and thus the direct comparison of one LPMO to another may not always be appropriate.[39, 44, 74]Combination of enzymatic analysis with small molecular mimics of the active site is likely to aid the discussion over which mechanism may occur, however, creating these compounds is a difficult task. Success has been made whereby complexes containing Cu(III)-OH species have been found capable of attacking strong C-H bonds, with products resulting from hydrogen atom abstraction found to contain a newly formed strong O-H bond.[39, 238] Small molecular mimics should provide a unique perspective on how changes in the coordination sphere of copper (for example in the various LPMO subclasses) can effect formation of reaction oxygen species, and therefore probe thermodynamic favourability towards

breaking strong C-H bonds[239], such as those in polysaccharides, as well as provide information from which real proteins could be spectroscopically compared.

Particulate methane monooxygenases (pMMO) have recently been reclassified as mono-copper enzymes[240], and are able to oxidise methane, meaning they play an important role in the carbon cycle and have potential industrial usage through the production of methanol based fuels.[39, 62-63]Although, their structure was always known to contain a histidine brace type scaffold, the originally suggested dual copper active site, as depicted in the 2011 crystal structure of *Methylococcus capsulatus* (PDB code: 3RGB), led to extensive research on small molecule mimics containing dicopper sites.[39] Due to increasing evidence from LPMO structures the nature of the copper active site in pMMOs was under scrutiny and recently the pMMO structure of *M. Capsulatus* was revisited, and subsequent re-analysis of data now showed that it is instead is a mono-copper species. Future work therefore should focus on comparing the active sites and tertiary structure of pMMOs to those of LPMOs, to help define which alterations in protein design change the properties of the enzyme so boldly, yet both still rely on the same basic copper scaffold and likely, oxidative chemistry, **Figure** 140.

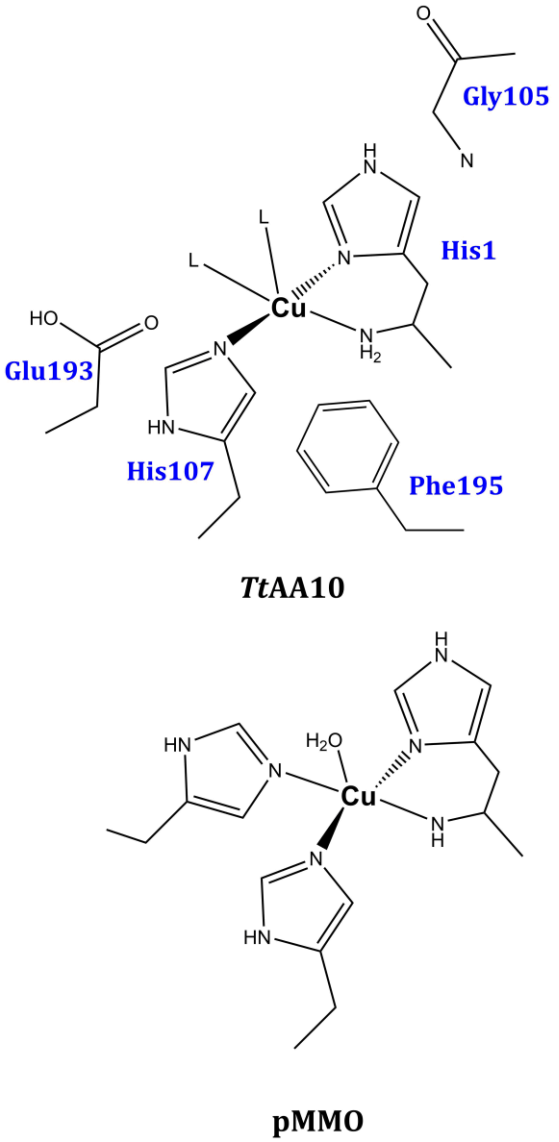


Figure 140 Diagrams comparing the active sites of TtAA10 and pMMO (reclassified, mononuclear site[241]) from *M. capsulatus*, where both consist of a copper ion held in the histidine brace scaffold.

7.3.2 A broader outlook

The life cycle of the shipworm is an attractive avenue for biochemical characterisation, due to its ability to survive on lignocellulosic biomass.[70] However, the shipworm represents just one source from which we can gain scientific insight into lignocellulosic degradation. Across the tree of life there is a wide array of ecological niches in which organisms have evolved exquisite ways to circumnavigate the recalcitrant nature of polysaccharides – and as such represent places in which we can apply genome mining for effective 'bio-prospecting', finding potentially new proteins, in a similar way to work carried out on the shipworm's symbiotic bacteria. Indeed, shipworm species discussed in this work are closely related to

similar marine bivalves which have evolved to live at higher pressures in the oceanic environment; Xylophagainae are found burrowing in sunken wood deposits and are thought to contain symbiotic bacteria, which could be a source of interesting enzymatic functions due to their ability to exist up to 7500 m below sea level.[242-243] Even more recently, a species of shipworm, *Kuphus polythalamia* has been found living in mud flats, whereby its bacterial symbiosis is centred on utilising sulfur rather than lignocellulose.[244]

Likely the most well thought of source for lignocellulosic digestion is that of fungi, where *T. reesei* is the number one example. Fungi are traditionally split into white-rot and brown-rot classes, two phylogenetic clusters that use different approaches to the degradation of lignocellulosic material. White-rot fungi mainly attack wood cell walls close to the surface of the material through enzymatic degradation, using large numbers of complementary cellulases as well as oxidative enzymes capable of acting on lignin and perhaps then somewhat unsurprisingly, a large selection of LPMOs.[80] In contrast, brown-rot fungi, whose evolutionary path from white-rot fungi may have caused a significant loss of cellulolytic machinery, exploits oxygen radical chemistry to effectively deconstruct lignocellulosic material; whereby oxygen radical chemistry is carried out away from the organism to prevent self-inflicted damage through a concentration gradient, with a recent study suggesting that gene regulation is staggered to effectively protect the remaining GHs found in the genome from the damage, when they are ultimately expressed and secreted. This process can be thought of as nature's own version of lignocellulosic pre-treatment prior to enzymatic degradation![245] Other forms of fungal 'rot' exist, such as that of *T. reesei*, a soft rot species; whereby the fungus colonizes the outer layer of substrate, before extending fine hyphae through the wood cell walls to begin the process of enzymatic degradation.[246] Whilst *T. reesei* may not possess an extraordinary amount of cellulolytic genes, its capability of expressing them in large quantities is why it is so highly regarded in the industrial sector.[247] Engineering of *T. reesei* over the past several decades has led to ever more exaggerated levels of production of native proteins, termed hyperproducing strains.

Some relevant areas from which bio-prospecting research is carried out, are those habitats consisting of harsh conditions, which inevitably cause accumulation of organisms capable of tolerating issues such as high temperature, pH, lack of oxygen or light and so on. One such seemingly innocuous environment is compost; whether this is natural composting events such as accumulation of leaf litter, or in man-made habitats. The improvements in so called 'meta-omics' make it possible to map the genomic data of many different species (meta-genomics) found within a single habitat and analyse the over expression of particular genes

(meta-proteomics) which provides a targeted approach to identifying lignocellulosic enzymes of interest, specifically those capable of withstanding harsher conditions which may be of industrial relevance.[248]

As is likely the case with most organisms (the shipworm, amongst other marine organisms such as the gribble, being an interesting outlier[70, 80, 87, 96, 101]), the co-evolution of the host and its gut microbiome have produced a mutualistic symbiosis in which bacterial communities are supported in attractive, food-rich, stable, moist environments and the host with nutrients inaccessible by endogenous digestion.[81, 249-250] The gut microbiome refers to all the genetic material of the various microbes that reside within the digestive system of the host.[250] There are many excellent examples of organisms which have maximised on bacterial habitation for efficient degradation of substrates. Termites for example, ingest 'mechanically pre-treated' materials (grinding of the insects mandibles) which are effectively broken down over the course of 24 hours through enzymatic degradation by an extensive selection of gut microbiota, housed in a series of gut chambers.[251] Ruminants, larger animals like sheep, cattle and antelope, also support a widely diverse range of microbes in a series of gut chambers; and by providing mechanically degraded food substrates and a constant temperature of 39 °C, the host animal is able to survive solely on vegetation – a substrate containing mostly inaccessible carbohydrates. Varieties of bacterial species are found across the different digestive chambers of the animal's anatomy, all of which have slightly differing conditions (i.e. pH)[252] and thus produce enzymes with different condition tolerances. Understanding digestion in ruminants doesn't just provide an attractive avenue of bio-prospecting, it also leads to improve in management of cattle, which are one of the most important human agricultures; improving the health of the cattle will naturally increase the efficiency from which they (and their gut microbiome) can convert the energy held in plants into other useful substances such as milk.[253]

Understanding the gut microbiome and its link to overall health of the host is currently a promising area of research in terms of improving human health and understanding disease. The fluctuating nature of the microbiota involves microbes which either simply pass through the gut during general digestive processing of food intake (digestion of passing microbes in ruminants provides the animals with the vast majority of their required amino acid intake [252]) or be more permanent residents within the gut; the microbiome is subject to change over time in response to many different factors such as ageing, diet, alcohol consumption, use of drugs and occurrence of disease etc.[250] Current thoughts are that the microbiome can be considered as an individual's personalised 'extra organ', which provides a variety of

health benefits and can be used as a marker for overall health. Indeed, studies already suggest that transplant of one healthy person's microbiome into a person suffering from ill-health may assist in re-growth of normal gut microbiota which has been lost for various reasons such as disease. For example, *Clostridium difficile* infections are endemic and becoming resistant to typical methods of treatment via antibiotics. The infection is often caught in hospitals where patients may already be in a poor state of health, but transplantation of healthy gut microbiota in the form of faecal enema (or more recently pill form) to the patient can clear an infection by providing the gut with an influx of healthy bacteria which are able to prevent re-infection of *C. difficile*, which can often lie dormant and reoccur.[254-255]

Being omnivorous, the human diet requires digestive capability of both animal and plant products. Somewhat surprisingly, the human genome only encodes CAZymes capable of breaking down sucrose, lactose and starch,[256] despite our intake of a vast quantities of therefore, in terms of the human genome, 'un-digestible fibre'. Ingestion of these plants fibres, containing complex carbohydrates that are not able to be broken down by our own enzymes, is termed dietary fibre. In fact, the digestion of this dietary fibre is carried out, as one might now expect, by a variety of bacterial communities; the most prevalent of which come from the bacterial phyla, *Firmicutes* and *Bacteroidetes*.[257] Degradation of polysaccharides into fermentable sugars by various microbes produces short chain fatty acids, from which we gain a significant portion of our daily calories.[258] The gut microbiome is likely to have evolved as humans have over time, and differences are found in the prevalence of bacterial species in people from different areas of the world. For example, a recent study suggested that high levels of consumption of raw food in Japanese culture may have caused transfer of bacterial genes, over time, to the gut microbiome, which are normally associated with the bacterial degradation of marine polysaccharides, specifically from algae. [259] Genes encoding for proteins with ability to degrade polysaccharides are often positioned together on the bacterial genome, to provide efficient regulation of multiple proteins in response to a specific substrate, such coalescence of genes is called polysaccharide utilisation loci, or PULs.[259] The *Bacteroidetes* species found in human, and many other mammalian digestive systems has gained a wide variety of PULs, likely through gene transfer events, which enable it to have activity against the complex picture of polysaccharides presented by our diets.[260] Comprehensive understanding of PULs and the roles they play in lignocellulosic digestion relies on correct annotation of genes. Traditional experiments in which individual genes contained within PULs are recombinantly expressed

and characterised is time consuming although high throughput screening of substrates, as described by Vidal-Melgosa *et al* may be the way forward[261]; whereby micro-arrays are used to screen large numbers of substrates at once, using small amounts of enzymes and even samples of liquid cultures.[261] Whilst these techniques are further developed and made accessible to more research groups, one must couple traditional wet chemistry approaches, with computational analysis, where prediction can be used with high levels of accuracy to define PULs and individual gene activities, which naturally shines a light on those involved genes that have unknown functions.[260] Study of PULs being used in human gut digestive processes will both give rise to information about potentially new enzyme activities as well as link to the broadening field of microbiome driven digestive health. The effect of enzymes and our understanding of them encompass a very broad area of research and of life in general, hence it is difficult to provide a picture which captures their overall effects, effectively. Human health and its reliance on bacterial enzymes is a major topic for potential future medicine and control of certain diseases. This slowly emerging understanding is gaining momentum, due to advances in our technical ability to handle such large amounts of genomic data[256] and is likely to continue advancing in directions related to potential therapeutics and possibly personalised medicine.

These advances in understanding how enzymes are important to health mirror how the biofuels industry has been steadily growing over the past couple of years since the introduction of LPMOs. It is amazing how discovery of a single enzyme type can kick-start an industry into something that is becoming more cost-effective and a major competitor of non-renewable fuels. This optimism is however, narrow. Scientists have put the vast amount of effort relating to improving sustainable fuels into researching ways to make enzymatic treatments better, and thus other areas required for efficient production of cellulosic ethanol have been ignored. Indeed, at the time of writing several major cellulosic ethanol producing plants have closed their doors, citing issues with the biomass supply chain. This consideration, where does the biomass actually come from, is something that is usually only merely discussed in terms of whether substrates are 1st or 2nd generation feedstocks – scientists such as myself, who focus narrowly on novel enzyme discovery to improve biomass processing lack the understanding of the whole chain process. Upstream of enzymatic biomass processing, the material must be pre-treated, using methods that clearly require much optimisation to avoid high costs resulting from high temperatures, or use of large amounts of solvents. The greatest problem however, is the source of the biomass. Land use needs to be appropriated for use in generation of crops suitable for biomass –

farmers are not keen to use their fertile, food growing land for the growth of fast growing weeds such as the invasive *Arundo donax*, a bamboo like plant able to grow up to 30 ft a year![262] One alternative is to use marginal land, coupled with the growth of grasses such as switchgrass (*Panicum virgatum*), a common sight in the Great British landscape and the giant reed, Carrizo cane, which are fast growing and able to tolerate land unsuitable for crop production.[263] Genetic engineering of plants may also make biofuel production viable, for example altering the structure of carbon deposition in plants to disfavour production of hard to degrade compounds such as lignin, the saccharification yield (processed polysaccharides) can be increased.[264-265] However, advances such as this are all very well, but still need to bypass those lingering consumer worries based on genetic modification of organisms.

Another potential source, is microalgae, and many companies over the past decade had predicted that use of algae based biofuels would be commercially viable, by now. Unfortunately, as is perhaps the case with cellulosic ethanol, unpredicted difficulties have prevented this idea from being reality.[266] Microalgae can produce a wide variety of high value natural products, which are used extensively in human and animal nutrition for example. The organisms naturally produce large amounts of lipids, and it is this oil which can be used for biodiesel production. Microalgae have a short life cycle, between 1-4 days and can produce relatively large amounts of biomass; the essential gain of using such organisms is that they are very efficient compared to plants in capitalising sunlight. In optimised conditions, microalgae could reach approximately 10 % photosynthetic efficiency,[267] an attractive figure considering the maximum theoretical efficiency rate of plants is below 6 %.[268] Microalgae are also naturally adept at sequestering carbon dioxide to use during metabolism. Logically, it would make sense to build industrial algae ponds near current sources of carbon dioxide output such as factories or other fuel stations.[269] However, climate, land type and water availability are important factors in choosing a plant site. Successful cultivation of algae in classic open ponds, or tubes on an industrial scale has many problems, such as water usage, effect of climate, contamination and control of nutrients.[270] Indeed, to make fuels efficiently from microalgae, its production needs to be maximised and the cost of such scale up has led to many companies abandoning biofuels and turning instead to commercialising microalgae by-products.[270] As has occurred with the continued genetic appraisal and improvement of *T. reesei*,[271] the future of microalgae for biofuels will likely rely on creation of new industrial strains which have maximised lipid production, without adversely affecting growth rate – as shown recently by Ajjawi *et al*.[272] To induce lipid production in microalgae, traditional methods starve the organisms of nitrogen during batch

culture. Ajjwai *et al*, using CRISPR-Cas9[273] reverse genetics, identified and modulated expression of a transcriptional regulator involved in lipid accumulation. The modulation of the regulator, named ZnCys in *Nannochloropsis gaditana* led to doubling of lipid production compared with the wild type without the normal negative effect on growth.[272] Advances in regulation of lipid production inside the organism rather than relying on the batch culture control of nitrogen levels is likely to lead to strains which are more suitable for large scale growth and production of lipids for fuels.

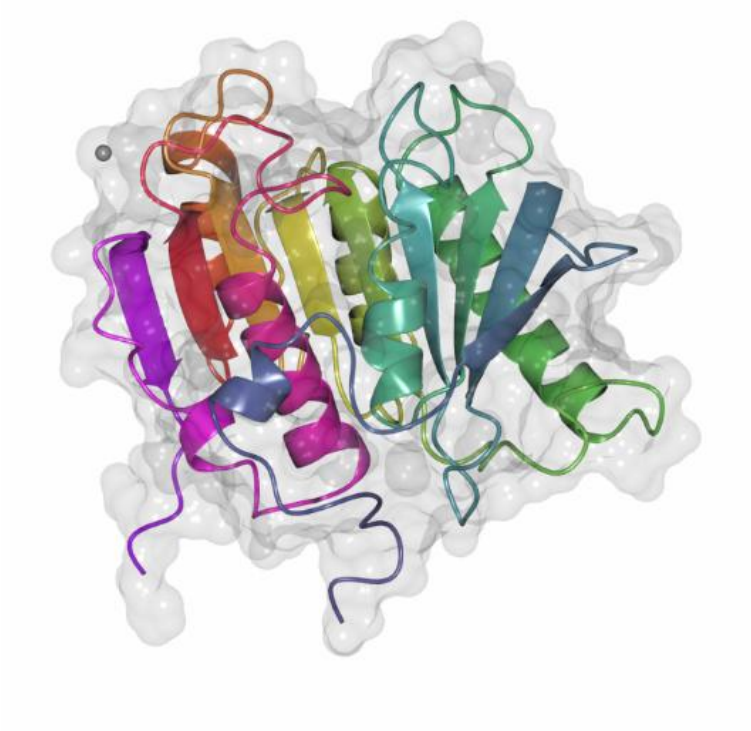


Figure 141 PET-ase from *Ideonella sakaiensis*, displaying an α/β structure containing 6 α-helices and 6 β-sheets, The active site is centred at the top, shown as a narrow groove in the surface structure (grey)

One must be practical however and understand the net benefit of producing biofuels – if more energy is used transporting biomass feedstocks to a bioprocessing plant, is there any net energy gain in doing it? Research into organisation of where plants are built and how materials can be efficiently collected and transported is of vital importance if such ventures are to succeed. Our natural associations lead us to believe that biofuel plants need to be massive, much like coal or oil refineries – perhaps the solution to the transport issue is to have more small scale plants that can be operated on the site of biomass growth – or as in the case of algae, coupled to other industrial sites to make use of outputs such as carbon dioxide.[269] This way of thinking can also be applied to other sustainability issues, such as recycling of plastics and municipal wastes. Canadian company, Enerkem convert municipal waste into ethanol and are set to build 100 new plants in China, a country suggested likely to

reach peak carbon dioxide emissions within the next decade,[274] and as such reduce their reliance on oil through more sustainable measures. The non-biodegradability of plastics is an extreme cause for concern, with generations of improper disposal causing accumulation of plastics with lasting effects on soil infertility as well as infiltration into oceanic food networks via micro plastics[275]. The properties of plastic, so well-renowned for its durable nature mean it is inevitably difficult to remove from the environment and break down. Our over production of plastic may have caused natural selection events, in which various microbiota are showing promise in developing 'natural' enzymatic breakdown pathways of common plastics. Indeed, evidence shown in studies of various creatures such as waxworms and meal worms identified bacteria capable of degrading plastics.[276-277] More recently, directed structural biology has improved the degradative ability of an enzyme, named PET-ase, from the bacterium, *Ideonella sakaiensis*, which was found able to use PET as a carbon and energy source, Figure 141.[278] Further identification of plastic-eating organisms, and subsequent analysis of the enzymatic actions required may aid in improving our recycling economy – potentially, as with LPMOs helping form a new industry centred around improving the sustainability of our plastic world.

This broader outlook hopefully draws together how analysis of enzymes in a single seemingly narrow field of study can have out reaching effects on a broader scientific landscape. Understanding of the microbiome (or lack of it in the shipworm) and the role bacteria play in degradation of biomass, is important to a wide range of diverse research areas from human health, breakdown of plastics to production of bioethanol. All these questions rely on an interdisciplinary approach, whereby enzymatic analysis needs to be considered in context. Thus, such understanding may improve our knowledge of how nature degrades biomass in such efficient ways, allowing us as a species to hijack this information and provide ourselves with a greater understanding of our own co-evolving microbiome 'organ' and more sustainable and, importantly, renewable fuel sources for the future.

Appendix 1: Initial Work on PHM

A1.1 Abstract

A protein found within the Shipworm, S-PHM, is homologues to the oxygen binding C-terminal domain of Peptidylglycine-alpha-hydroxylating monooxygenase (PHM). The closest bacterial and eukaryotic homologues to S-PHM have been expressed under a variety of different expression conditions, purified under different buffer conditions, gene truncated according to sequence analysis and had the addition of a SUMO tag. Further testing involving on-column refolding and western blot analysis lead to the conclusion that the S-PHM homologues were intrinsically misfolded and unavailable for further characterisation at this point.

A1.2 Introduction

A1.2.1 An Interesting Link

LPMOs are known for their incorporation of a single copper ion within the active site and their use of molecular oxygen in its enzymatic activity. A protein found within the hindgut of the Shipworm has been highlighted due to its classification as a 'copper monooxygenase domain' during sequence homology searches. Sequence analysis shows the protein is related to one commonly found in higher eukaryotes, peptidylglycine-alpha-hydroxylating monooxygenase (PHM).

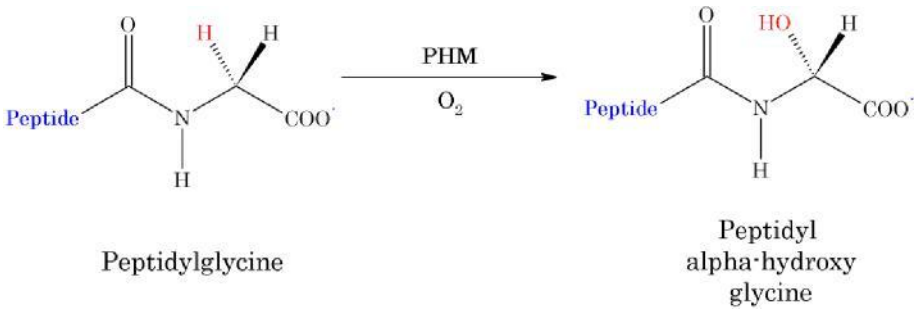


Figure 142 General reaction scheme of PHM actin upon peptidylglycine extended peptides.

PHM, a bi-copper protein, catalyses the stereospecific hydroxylation of the glycine α-carbon of peptidylglycine substrates (prohormones). This is the first step in the post-translational amidation of glycine extended peptides; a process used for activation of peptides such as growth factors, hormones and neurotransmitters, **Figure** 142 General reaction scheme of PHM actin upon peptidylglycine extended peptides.**Figure** 142 Depending on the situation, PHM can act as a single, independent enzyme or as part of the bi-functional PAM protein. [279-281]

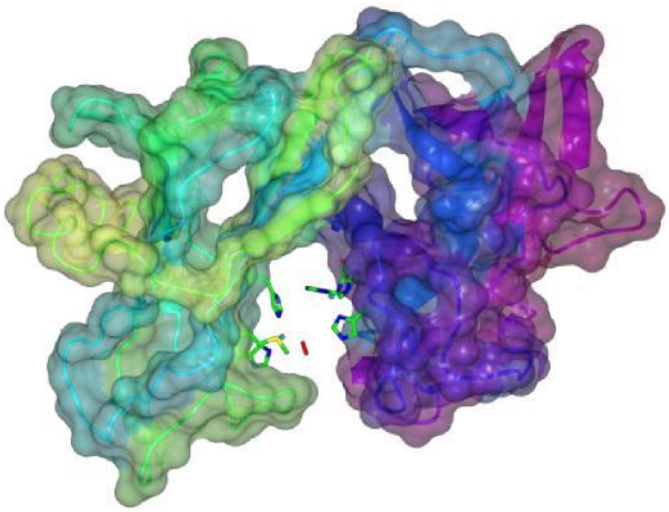


Figure 143 PHM from *Rattus norvegicus* (PDB: 1sdw). The image clearly shows the cleft separating the two copper ions. Green/blue refers to the C-terminal domain whereas the purple/blue half is the N-terminal domain. Image produced in CCP4mg.[55]

Structurally PHM can be described as a scaffold holding two non-coupled copper ions 11 Å apart, **Figure** 143 and **Figure** 144. show the active site of PHM, where the two copper ions, Cu_A and Cu_B are shown separated by a water filled cleft, **Figure** 143. Cu_A coordinates three histidine residues in a 'T-shape' geometry (square pyramidal with two unoccupied positions). Cu_B forms a distorted tetrahedral geometry, binding two histidines, a methionine and a water molecule. The methionine ligand has been linked to increased catalytic activity and overall structural integrity of the protein.[282]

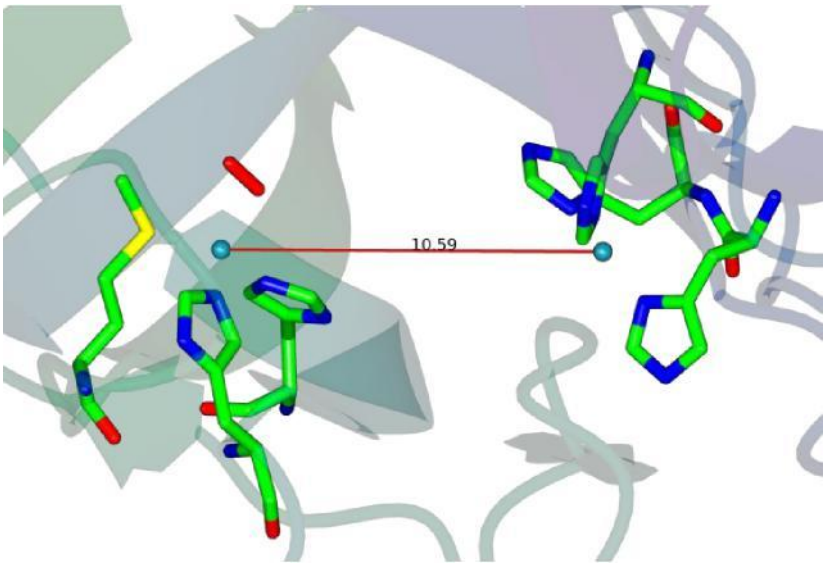


Figure 144 Active site of PHM from *Rattus norvegicus*, showing the two copper ions 10.59 Å apart. Cu_A on the right interacts with three histidine residues whereas Cu_B binds to two histidines and a methionine (a water molecule is also thought to bind tbut this is not shown). An oxygen molecule is also shown in the ligand sphere around the Cu_B ion, thought to bind in an end-on manner. Image produced in CCP4mg[55] from PDB file 1SDW.

Both copper sites are reduced to Cu^{I} prior to catalysis via interactions with two molecules of ascorbate. Physical and computational studies have suggested the formation of a cupric superoxo (Cu^{II}-$O_2^{\cdot -}$) reactive species by end-on coordination of molecular oxygen to Cu_B in a tetrahedral geometry, **Figure** 144 . It has been suggested that the large cleft separating both sites prevents peroxide formation, allowing formation of the more active superoxo complex. Formation of the oxy-PHM adduct combined with binding of a nearby substrate is thought to initiate electron transport from the decoupled Cu_A ion to the awaiting Cu_B ion. This results in a C-H activation reaction via hydrogen abstraction.[279-281, 283]

Several studies have looked at producing stable superoxo Cu_B mimics. Several exhibit limited reactivity or form different Cu-active species, but examples involving stabilised superoxo species exhibits enhanced reactivity.[284] This is thought to arise due to coordination with a thioether, which imitates the methionine ligand, producing what is thought to be the first real biomimetic of the PHM Cu_B site. This is an important step in furthering the understanding of other potential copper intermediates thought to be involved in PHM C-H activation.[284-286]

Interestingly, sequence analysis of the Shipworm protein (S-PHM) against a variety of PHM homologues from various organisms predicts presence of only the C-terminal domain (Cu_B). A previous study has looked at the two copper active sites using PHM mutants containing either deactivated Cu_A or Cu_B to gain independent measurements for the catalytic activity of

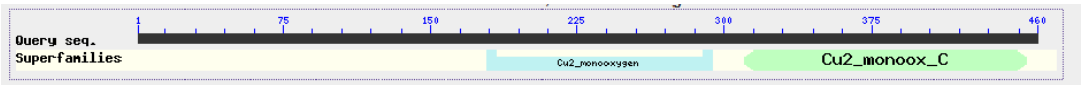
each site; the close proximity of the copper ions often causes overlap of spectral signals during analysis. It was noted that both mutants were devoid of catalytic activity.[287] The presence of this single 'copper monooxygenase domain' in the Shipworm, which is similar to the oxygen binding Cu_B site, could therefore indicate a new protein with a different function. The shipworm could potentially utilise this 'copper monooxygenase domain' in the degradation of cellulosic material within its hindgut.

This work aims express S-PHM homologues using an *E. coli* expression system. The majority of PHM studies use mammalian cell lines. Production of soluble homologues may prove troublesome, as previous studies have observed only insoluble protein expression when using *E. coli*.[288] Several molecular cloning strategies and expression systems shall therefore be employed with the aims of finding a suitable system.

A1.3 Aims

Three PHM homologues from different organisms were chosen based on their size and similarity to Shipworm PHM (S-PHM); *Sorangium cellulosum*, a soil dwelling gram negative bacteria; *Crassostrea gigas*, a Pacific Oyster and *Nasonia vitirpennis,* a type of wasp. These three S-PHM homologues shall be expressed in *E. coli* through a variety of molecular cloning strategies. Once a suitable expression system is found, characterisation and analysis of the S-PHM homologues will ensue.

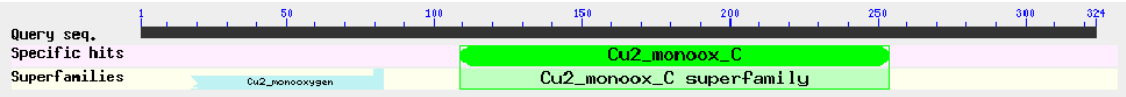
S.cellulosum



Blue – Residues 179-294 **N-Terminal Domain**

Green – Residues 310 to 454 **C-Terminal Domain**

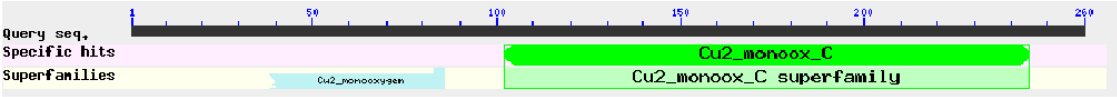
C.gigas



Blue – Residues 18-83 **N-Terminal Domain**

Green – Residues 109 to 254 **C-Terminal Domain**

N.vitripennis



Blue – Residues 38-86 **N-Terminal Domain**

Green – Residues 102 to 245 **C-Terminal Domain**

Figure 145 Three homologues of Shipworm PHM, showing expected copper monooxygenase domains as classified in BLAST searches[113]

A1.4 Methods

A1.4.1 Shipworm PHM Homologues

Plasmids purchased from Genscript of PHM homologues from *Sorangium cellulosum*, *Crassostrea gigas* and *Nasonia vitripennis* were transformed into Top10 cells and grown in LB Agar containing the appropriate antibiotic at 37 °C overnight. Fresh LB (5 ml plus Kanamycin 30 mg mL^{-1} or Ampicillin 100 mg mL^{-1}) was inoculated with individual colonies and grown at 37 °C, 180 rpm, overnight. Cultures were centrifuged and plasmids extracted using the Q1 Aprep Spin Miniprep kit. Plasmids were then transformed into Bl21* competent cells and grown in LB Agar containing either Kanamycin or Ampicillin. Liquid cultures were subcultured and O_D monitored during growth at 37 °C, 180 rpm. At O_{D600}, uninduced samples were taken and cultures cooled to 16 °C before inoculation with IPTG (1 mM). Conditions for the expression tests are shown in table 3. Soluble and insoluble protein was extracted using Bugbuster cell lysis solution of 1 mL culture samples. Samples were mixed with SDS loading dye and analysed by SDS PAGE (12%, 50 mins, 200 mV). All gels were stained with Magic dye.

Table 27 Summary of construct expression conditions

Construct	Cell Strain	Growth Temp. (°C)	IPTG (mM)	Expression Temp. (°C)	Expression Time and other information	Lysis Solution
*Cg*PHM, *Nv*PHM, *Sc*PHM	BL21*	37	1	16	18	Bugbuster
*Cg*PHM, *Nv*PHM, *Sc*PHM	BL21*	37	1	16	18	Tris (pH8), Imidazole (30 mM), NaCl (0-750 mM)
*Cg*PHM, *Nv*PHM, *Sc*PHM	T7 SHuffle	37	1	16	18	Bugbuster
*Nv*PHM	TUNER	37	1	16	18	Bugbuster
*Cg*PHM, *Nv*PHM, *Sc*PHM	BL21*	37	1	16	1.5-18 Cultures centrifuged after 1.5 and 3 hr and pellets resuspended in LB containing chloroamphenicol and grown further at 28 °C	Bugbuster
*Sc*PHM, *Cg*PHM	Bl21*	37	1, 0.5, 0.1	16	18	Bugbuster
*Sc*PHM, *Cg*PHM	Bl21*	37	1	30	18	Bugbuster
*Sc*PHM,*Cg*PHM	Bl21*	37	1	37	18	Bugbuster
*Sc*PHM, *Cg*PHM	Bl21*	37	0.5	16	18, *Cells harvested at OD = 0.4*	Bugbuster
*Cg*PHM, *Nv*PHM	BL21*	37	0.1	10	5	Bugbuster
CgPHM	LEMO	37	1	16	18 Cultures were tested for expression at 30 °C using an increasing concentration of L-rhamnose (0-0.25,0.5,1 mM) and a set concentration of IPTG (400 µM).	Bugbuster

A1.4.2 Production of other PHM Constructs

Eight truncated variants were designed and specific primers procured. *Cg*PHM, *Nv*PHM and *Sc*PHM were made through various PCR reaction and subsequently re-circulised using the Gibson Assembly master mix (in a 1:1 reaction, 50 °C, 15 minutes). Plasmids were transformed into NEW ENGLAND BIOLABS 5-Alpha competent cells and prepped colonies sequenced (GATC, Sanger sequencing). Correctly sequenced plasmids were used to inoculated BL21* competent cells. Expression tests were carried out under standard expression conditions and by using a low temperature expression test (10 °C, 0.1 mM IPTG, 5 hr).

Inserts for *Cg*PHM and *Nv*PHM variants were incorporated into the Champion SUMO pET vector using New England Biolabs Hifi Assembly and colonies prepped and sequenced (GATC, Sanger sequencing). Expression tests on *Cg*PHM SUMO, *Cg*PHM 79 SUMO, *Nv*PHM 65 SUMO and *Nv*PHM 103 SUMO were performed under three different conditions in Bl21*: 37 °C, 1 mM IPTG, 16 °C; 37 °C, 0.5 mM IPTG, 16 °C; 37 °C, 1 mM IPTG, 37 °C. SUMO CgPHMO was also tested in T7 Shuffle. A larger scale purificiation of Cg PHM SUMO was attempted using Ni affinity chromatography. Cell pellets were resuspended in buffer containing 20 mM Tris (pH 8), 500 mM NaCl, 30 mM imidazole, 1 mM DTT and 5% glycerol, but nothing was observed eluting from the column under high imidazole concentrations.

A construct for expression in the periplasm was created using NEW ENGLAND BIOLABS HiFi method by inserting NvPHM into a pET11a vector containing a pelB leader peptide and transformed into BL21* competent cells. Expression was tested as normal (37 °C, 1 mM IPTG, 16 °C).

A1.4.3 On-Column Refolding of Cg PHM SUMO

E. coli containing plasmids of CgPHM-SUMO were grown as usual (37°C, induction with IPTG, incubation overnight at 16°C), harvested and the pellet resuspended in buffer 1 (30 ml, 20 mM Tris (pH 8), 500 mM NaCl, 30 mM imidazole). The suspension was sonicated and lysed cells centrifuged at 15 g. The supernatant was discarded and pellet resuspended in buffer 2 (20 ml, 20 mM Tris (pH 8), 500 mM NaCl, 30 mM imidazole, 1% Triton X-100). The sample was sonicated further and pelleted by centrifugation. The supernatant was again discard and the pellet resupended in buffer 3 (20 ml, 20 mM Tris (pH 8), 500 mM NaCl, 30 mM imidazole, 8 M Urea) and left stiring at room temperature for 20 minutes until fully dissolved. The sample was loaded on a pre-equilibrated Ni HisTrap FF Crude column before a gradient of buffer 3 to buffer 1 was applied. Elution with buffer 4 (20 mM Tris (pH 8), 500 mM NaCl, 500 mM imidazole) saw no protein elution peaks. In an attempt to improve protein recovery, several alterations were made to the procedure;Pellet was left stiring in buffer 2 for 1hr before centrifugation; Buffer 3 was changed to include only 4 M Urea and the pellet left stiring for 2 hrs in this solution. This was also centrifuged and the supernatant only, loaded onto a Ni HisTrap FF Crude column; the concentration of urea was reduced from 4 M to 0 M over an hour and elution with imidazole over 30 minutes. However, none of the changes resulted in protein binding or eluting from the column.

A1.4.4 Buffer Screening of Cg PHM SUMO

Cells were grown as usual, harvested and the pellet frozen. Pellet was resupended in the resuspension buffer before being separated into 30, 1 ml aliquots. These were then centrifuged and each pellet resuspended in the different buffers as quoted by Lindwall.[289] Suspensions were loaded onto a 24 well plate and sonicated (60 sec – pulse 1 s ec, off 1 sec) Lysed cells were centrifuged and the supernatant removed to a clean tube. The insoluble material was resuspended in the same buffers and each fraction mixed with SDS loading dye. Samples were analysed by SDS PAGE. Nine samples were then analysed by Western Blott. Large scale purification of one of the buffer conditions (100 mM Tris, pH 7.6, 10% Glycerol) was done a few times without replicable results. Nothing significant was observed eluting during Ni affinity chromatrography.

A1.5 Results

A1.5.1 Expression of bacterial and eukaryotic S-PHM homologues

Native *S. cellulosum* PHM (*Sc*PHM), *C. gigas* PHM (*Cg*PHM), *N. vitirpennis* PHM (*Nv*PHM) proteins produced no soluble expression in the BL21*, T7 SHuffle or TUNER strains of *E. Coli*. Different conditions such as temperature, IPTG concentrations, culture growth levels and buffer conditions were explored but none were favourable, Figure 146. All three proteins expressed well in the insoluble fraction of purification in the majority of conditions. However the BL21* strain exhibited the highest levels of insoluble expression.

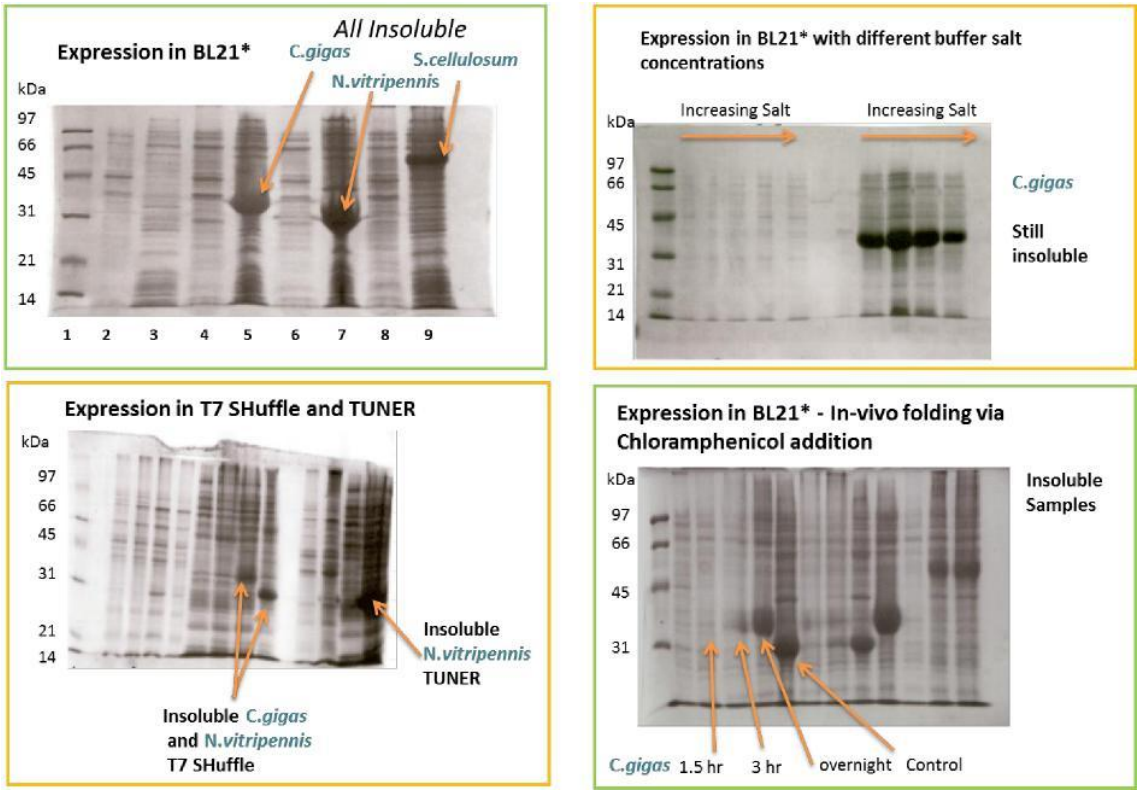


Figure 146 SDS PAGE analysis of expression tests of S-PHM homologues under different conditions. All lead to insoluble expression.

A1.5.2 Bioinformatic anaylsis of Hydrophobicity

To access whether the insolubiliy of CgPHM was due to a high level of hydrophobic surface residues, bioinfomatical analysis was performed, **Figure** 147. A web tool was used to produce a Kyle-Doolittle Scale or hydrophobicity plot. The plots shows no high degree of hydrophobic surface residues as observed by the lack of positive peaks.

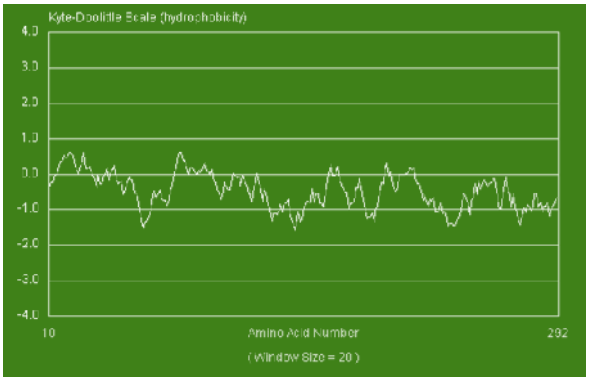


Figure 147 Kyle-Doolittle plot of hydrophobicity . Scores above zero relate to hydrophobic regions. Produced using an online tool at
http://www.vivo.colostate.edu/molkit/hydropathy/index.html.

A1.5.3 Expression of truncated S-PHM homologues in *E. coli*

Protein sequence alignments with S-PHM and its suggested domain boundaries were used to determine potential truncation positions for the PHM homologues, **Figure** 148. Sequence alignments with the S-PHM and each homologue yielded several possible truncations based around the N and C-terminal domains. Three truncations were designed See Table 1. First, the amino acid sequence prior to the predicted N-terminus was removed. Second, the whole N-terminal domain was removed, leaving only the predicted C-terminal domain. Third, alignments with S-PHM saw an optimum alignment starting position, any sequence prior to this was removed. Table 28 shows the expected protein weights upon completion of the described mutations.

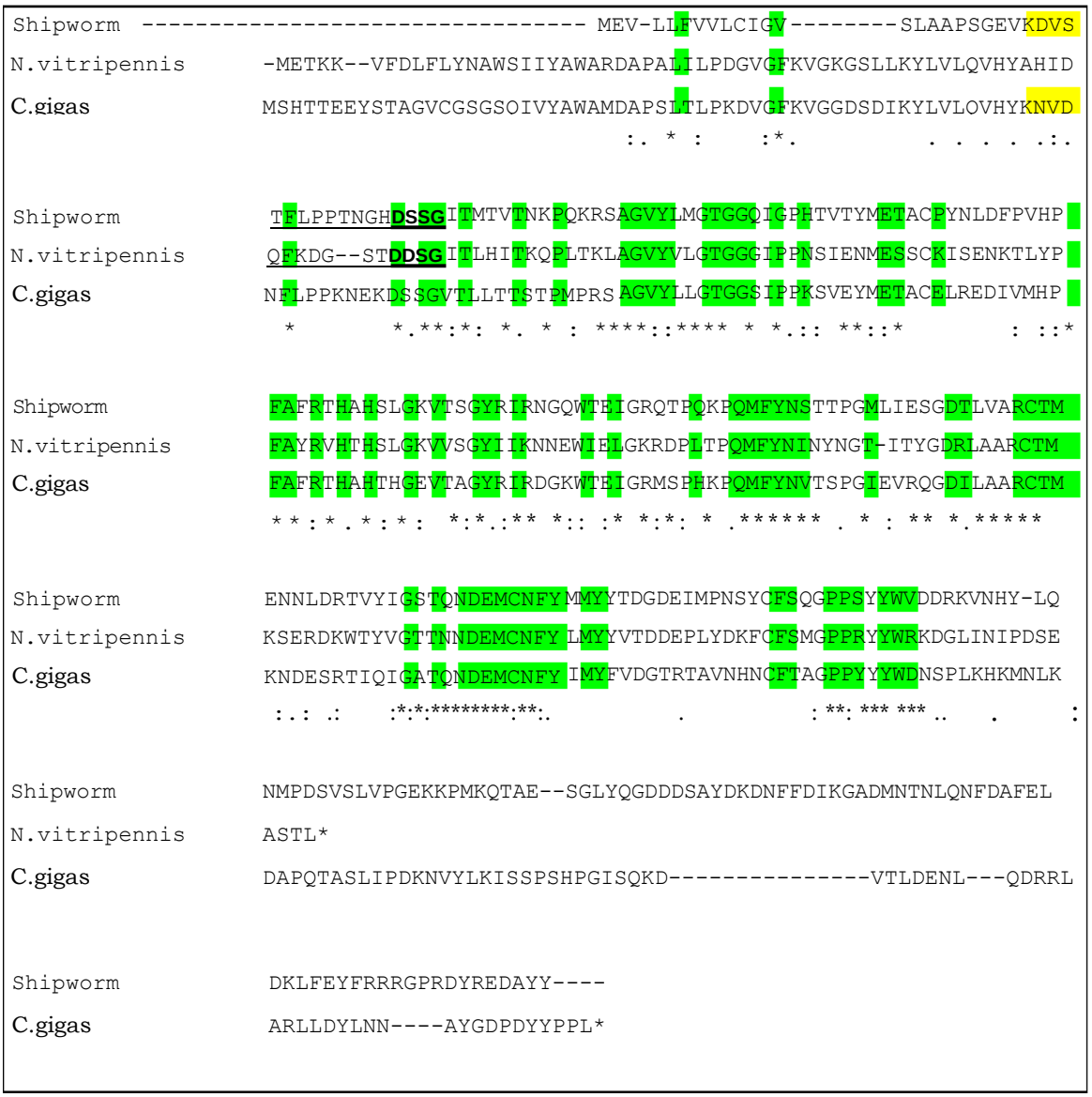


Figure 148 Sequence alignment of Shipworm PHM, N.vitripennis PHM and C.gigas PHM using Clustal Omega. Residues highlighted in green show high sequence similarity. Yellow highlight and bold underline is the position of optimum sequence alignment as identified by BLAST for each alignment pair.

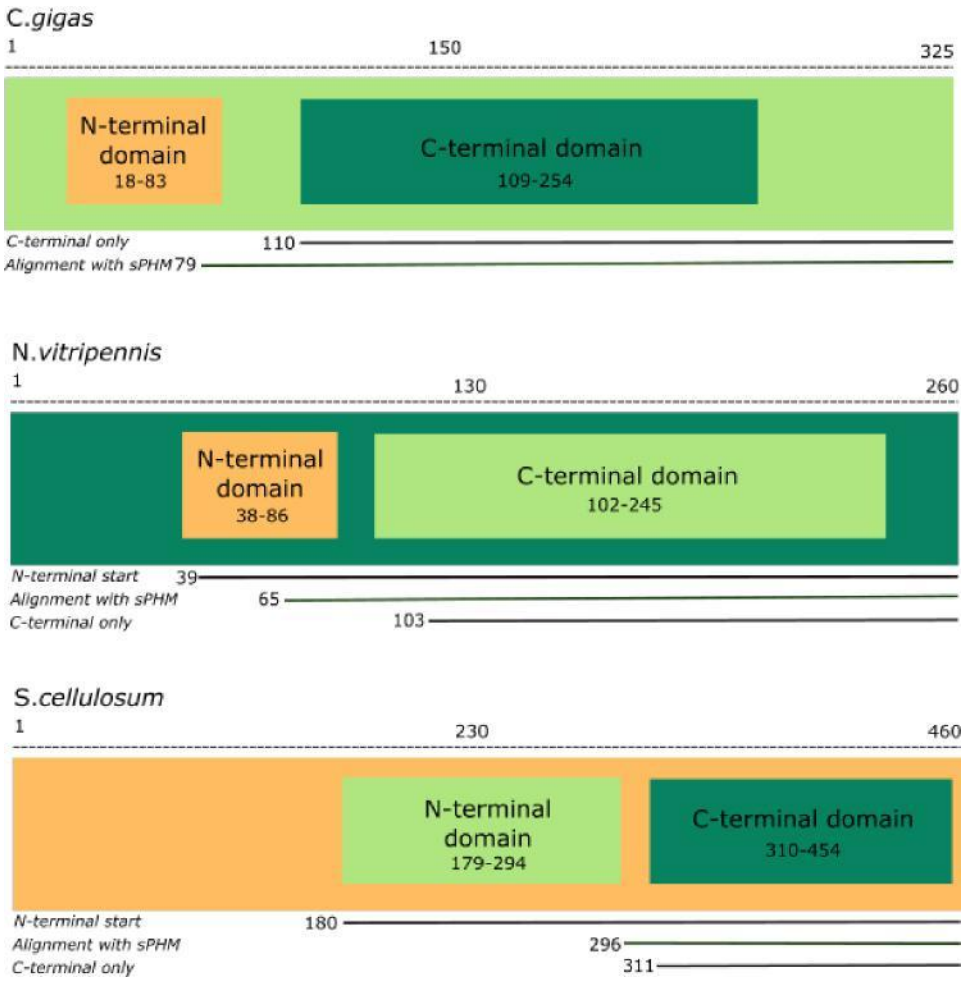


Figure 149 Diagram showing truncation positions relative to the N-terminal and C-terminal domains and to the best alignment against S-PHM. Numbers indicate amino acids.

Table 28 Expected protein weights in KDa after truncations.

Protein	Weight K/da	Protein	Weight K/da	Protein	Weight K/da
C.gigas	36.06	*N.vitripennis*	29.45	*S.cellulosum*	48.95
C.gigas 110	26.81	*N.vitripennis* 65	22.22	*S.cellulosum* 180	35.27
C.gigas 79	30.06	*N.vitripennis* 39	27.57	*S.cellulosum* 311	21.36
		N.vitripennis 103	20.6	*S.cellulosum* 296	22.85

Expression test results on the truncated proteins were varied. *Sc*PHM constructs showed neither soluble nor insoluble expression in BL21*. The truncated versions of *Cg*PHM and *Nv*PHM however retained their insoluble expression, Figure 150.

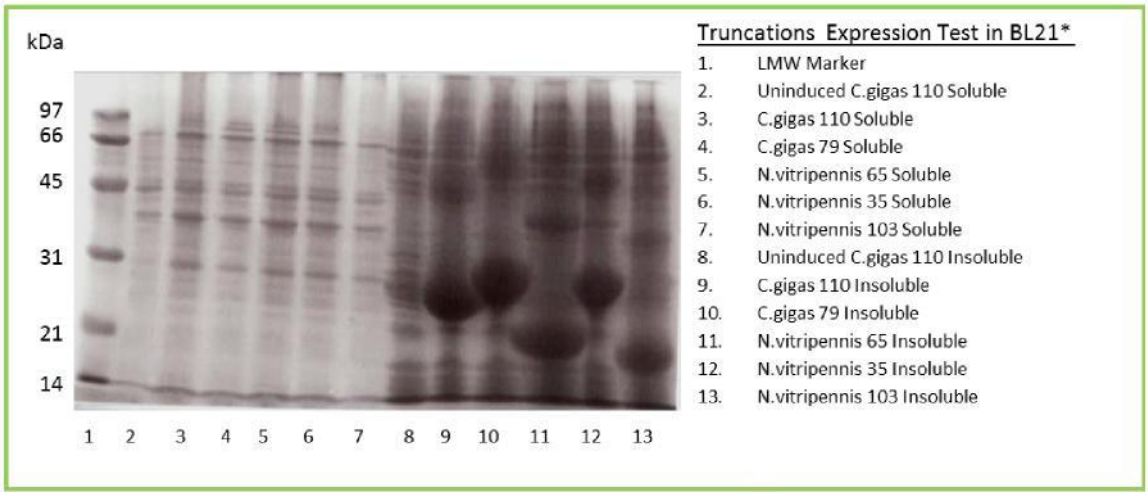


Figure 150 SDS PAGE showing soluble and insoluble fractions of the truncated constructs for *Cg*PHM and *Nv*PHM. A high degree of insoluble expression in BL21* under standard expression conditions is observed (growth at 37°C, induction with 1 mM IPTG and expression at 16°C).

Handa *et al*, successfully expressed the catalytic core of human PHM in E. coli by coupling a thioredoxin fusion system with low temperature (10 °C) and low IPTG concentration (0.1 mM conditions).[41] Expression tests on CgPHM and NvPHM constructs in different cell strains, with and without truncations were subjected to the same low temperature expression conditions. Cells were harvested after 5 hours but the only difference resulted in the amount of insoluble expression observed. These truncated S-PHM homologues lack a fusion system which can often improve protein folding and solubility. This may have helped achieve the notoriously difficult soluble protein expression of PHM by Handa.

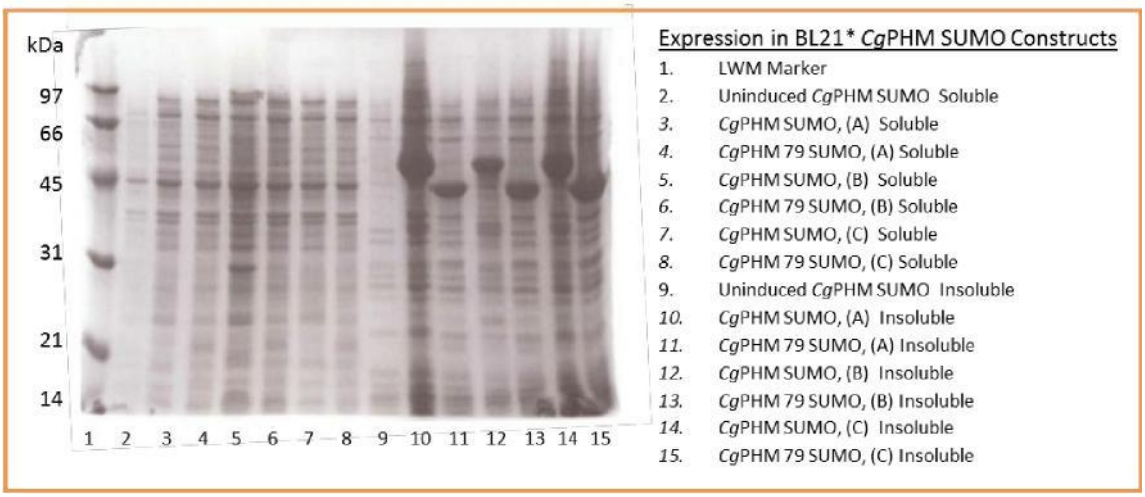


Figure 151 SDS PAGE of soluble and insoluble fractions of *Cg*PHM SUMO and *Cg*PHM 79 SUMO under three different expression conditions. A) Growth at 37°C, induction with 1 mM IPTG and expression at 16°C. B) Growth at 37°C, induction with 0.1 mM IPTG and expression at 37°C. C)) Growth at 37°C, induction with 0.5 mM IPTG and expression at 16°C.

A fusion system is required that can be easily removed, leaving behind an active form of PHM. Thioredoxin can be removed but will leave left over sequence prior to the start of PHM. Addition of a SUMO tag may improve solubility, protein folding and allow efficient removal of non-native protein sequence. Figure 151 shows the SDS PAGE results of expression tests for *Cg*PHM SUMO and *Cg*PHM 79 SUMO. The difference in weight observed between the two constructs can be clearly seen on the gel as well as the occurrence of only strong insoluble expression. The same result was obtained for *Nv*PHM 65 SUMO and *Nv*PHM 103 SUMO. A larger scale purificiation of *Cg*PHM SUMO was attempted using Ni affinity chromatography but nothing was observed eluting from the column.

A1.5.4 Lemo21(DE3) and T7 SHuffle cells

*Cg*PHM was transformed into Lemo21(DE3) competent E. coli tuneable expression cells which have been shown to increase soluble protein production. Addition of small amounts of L-rhamnose controls expression of the target gene by encouraging the cells to produce lyzsoyme, an inhibitor of T7 RNA polymerase. Control using L-rhamnose did not improve soluble expression. A consistant expression level was observed for the insoluble protein despite addition of L-rhamnose suggesting the insolubility of the protein may naturally limit its expression. *Cg*PHM SUMO was transformed into T7 SHuffle competent *E. coli* cells and tested for expression using an increasing concentration of IPTG (0.05, 0.1 ,1 mM) . The

S-PHM constructs do contain disulfide bonds which should be favoured within the reducing environment of T7 SHuffle cells however, only insoluble expression was observed.

A1.5.5 Periplasmic Expression

A periplasmic construct for NvPHM in plasmid pET11a using a pelB signal peptide was designed. Expression was tested using standard conditions but neither insoluble or soluble protein was observed at the expected weight.

A1.5.6 On- Column Refolding of *Cg*PHM SUMO

Purification of insoluble proteins can sometimes be achieved through a process of refolding. The insoluble protein is unfolded under denaturing conditions such as a high concentration of urea. These conditions are gradually reversed in the hope that the protein will correctly fold. On-column refolding utilises a protein's His tag and reduces the need for excessive amounts of buffer normally required during dialysis, **Figure** 152. The denatured protein can still bind to the column due to the sequential repetition of histidine residues. The high concentration of urea is gradually removed. In principle the bound protein can refold during the process and be eluted from the column using an increasing concentration of imidazole. [42]

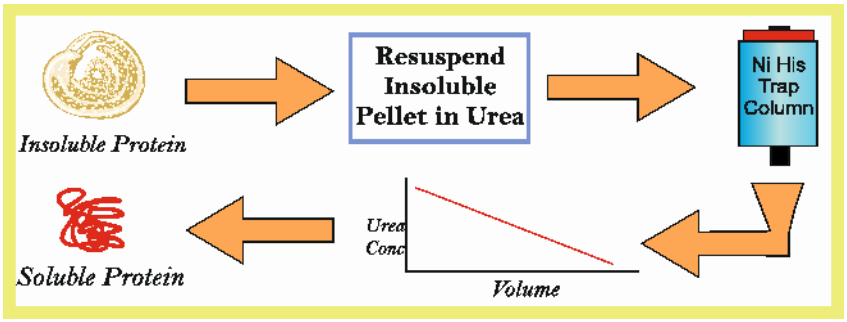


Figure 152 Simple schematic explaining on-column refolding

Some success has been observed through the use of this method. However its use of proteins with no known function makes the method harder due to the lack of any available assay to test for successful refolding. One such recent example is a hypothetical protein found in the genome of S.*degradans* purified using this method. The authors prepositioned that this protein, with an N-terminal CBM similar to a CBM 2 from *T.turnerae* and a C-terminal domain with no observed similarity to any known carbohydrate active proteins may be of novel function. The protein was successfully purified using the on-column refolding

mechanism and a structured determined. However, assessment of function with regard to polysaccharide degradation produced no conclusive results and the function of the protein remains unknown. Hence, in principle refolding can produce soluble protein. However if the function is unknown there will be no known assay to assess whether the protein has refolded correctly, producing active protein.

On-column refolding of CgPHM SUMO was attempted. It was thought that if soluble protein could be ilsolated and the SUMO tag cleaved off successfully through specific recognition of the C-terminal tertiary structure by the corresponding protease, then there may be a good chance of the protein refolding correctly. Unfortunately, the assay was not able to be investigated as the refolding stage was not successful. Some soluble protein was observed during suspension in a strong detergent containing buffer but unfortunately the protein, once denatured with urea failed to bind to the column. The His tag was present, as observed in the sequencing data during the cloning stage. Despite this the protein was only observed in the column flow through as shown in Figure 153.

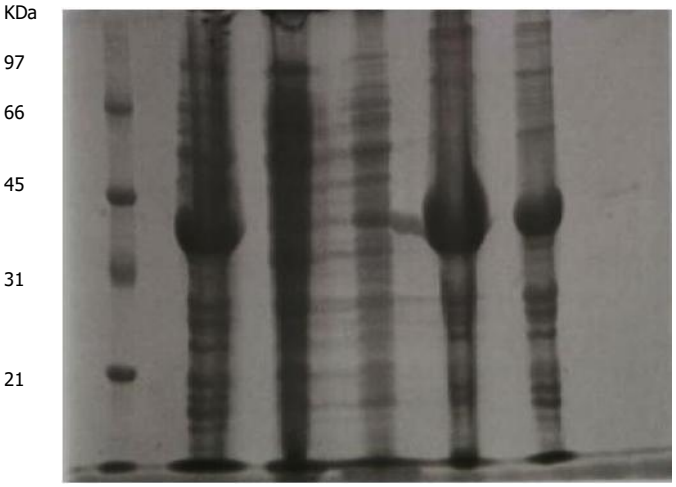


Figure 153 SDS PAGE of samples taken from the attempted on-column refolding of CgPHM SUMO. Lane1: Marker. Lane 2: reference sample taken from an insoluble fraction during a previous expression test. Lane 3: After cell lysis. Lane 4: Supernatant collected from re-suspension in 1% Triton X100 buffer: Lane 5: Ni Column load after being denatured in 8M Urea. Lane 6: Column flow through.

The use of detergent in the refolding protocol did form a small amount of soluble protein. It is possible that this relates to the faint band observed the last lane of the SDS PAGE. However, the use of strong detergents such as Triton X100 is not ideal for techniques further down the line. As such it was decided that a relatively large buffer screen containing 30 different buffer conditions would be tested.[289]

A1.5.7 Buffer Screening for CgPHM SUMO

A buffer screening protocol involving 30 buffers, as described by Lindwall et al, was tested. 43 Small amounts of potentially soluble protein was observed and analysed further by western blot as shown in Figure 18. An antibody with specific binding towards his tags was used to evaluate the presence of soluble protein in the supernatant fractions of each buffer screen. Large scale purification of one of these conditions (100 mM Tris, pH 7.6, 10% Glycerol) failed to produce any eluted, purified protein in replicable amounts. This lead to the conclusion that the protein is intrinsically misfolded and the His tag is potentially buried within the structure, making it inaccessible.

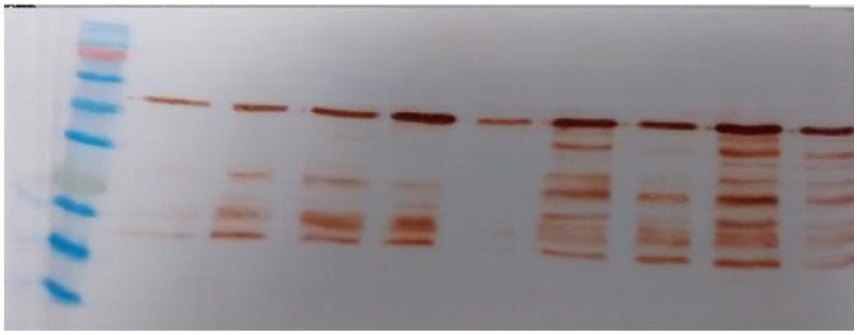


Figure 154 Western Blot of 9 buffer conditions producing small amounts of soluble protein as shown by the top bands at approximately 47 kDa.

A1.6 Conclusion

Expression of S-PHM homologues has failed to produce soluble protein, despite testing under various different conditions. Truncations, addition of a SUMO tag and periplasmic expression has thus far not aided solubility. Insoluble protein expression was however quite high in the majority of cases. Refolding under denaturing conditions also failed due to insufficient protein binding to the Ni column. Soluble protein was observed during buffer screening but the small amounts were not successfully scaled up during larger scale purification.

Appendix 2: Methodology in Theory

A2.1 Abstract

In this work, several different experimental methods have been used to probe enzyme function and to analyse the structure of the 6 different target proteins. As such this Appendix presents information the reader may find helpful prior to reading the main results Chapters. Topics include cloning methods, affinity chromatography, thermal shift analysis (TSA), thin layer chromatography (TLC), matrix assisted laser desorption-time of flight mass spectrometry (MALDI-TOF-MS), high performance anion exchange chromatography with pulsed amperometric dection (HPAEC-PAD), 3, 5-dinitrosalicyclic acid (DNSA) reducing sugar assay, protein crystallisation and x-ray structure determination and EPR.

A2.2 Cloning

The polymerase chain reaction (PCR) is a technique used to produce many copies (amplification) of a target DNA sequence. Amplified plasmids for expression of a target protein generally consist of two types, the individual gene sequence for the protein of interest, and a vector DNA backbone. Separate PCRs are carried out to produce large amounts of vector DNA and target gene DNA (also called a DNA insert), which when annealed together through overlapping base pair ends (sticky ends), form plasmids which can be taken up by bacterial cells during the process of transformation. Conditions during a PCR experiment rely on changes in termpature (thermal cycling) between several specific temperatures which cause the DNA to behave in a certain manor. At high temperature, 94 °C, double stranded DNA is known to melt, or in other words the DNA double helix becomes separated. The temperature is then lowered to a temperature suitable for a certain length oligonucleotide (a primer) to bind to the single stranded DNA; temperature is dependent on the length of the primer and on its G/C base content and is between 50-70 °C. A primer is designed to bind to a specific portion of the DNA, this is typically used when designing expression constructs, to target certain regions of a gene or vector. Primers consist of a set of bases that are complimentary to the target gene base sequence and are always used in pairs. The first primer binds in the 'forward' direction where DNA polymerase works from 5' to 3'. The forward primer is simple to design as it is merely the complimentary sequence to the N-terminal end of the target gene's coding strand (sense DNA strand). The second primer, known as the reserve also requires DNA polymerase to work from the 5' to 3' direction, meaning that this second primer must be designed in 'reverse'. The reserve primer is complementary to the C-terminus, but the oligosaccharide must be designed to be antiparallel. This allows DNA polymerase to work along from the reverse primer as if it were the same sense as a coding DNA strand. Once the primers have annealed to their individual complimentary DNA sequence, on both separated DNA strands, the temperature is increased to 72 °C. A higher temperature allows a DNA polymerase to work efficiently and recognise the newly formed double stranded DNA segments produced by primer annealing. The activity of DNA polymerase is initiated by it's own binding to the double stranded segments, causing it to process along the two template single stranded DNA strands, adding complimentary free nucleotides in a process known as extenstion. Eventually, two completely new double stranded DNA fragments are produced. The PCR machine cycler then begins the thermal cycle again, melting the newly formed DNA fragments apart. There

are now four single stranded fragments to which the primers can anneal to. After annealing and extention, the number of fragments will double. Continuation of thermal cycling means the number of DNA fragemtns will increase exponentially, until millions of double stranded DNA fragments are formed.

To make a full circular plasmid, the primers can be designed so that they have overhangs – sections at the start or end of the primer that do not anneal to the DNA sequence present in the PCR, but are instead complimentary to a second DNA fragment not used in the PCR. In order to expose the overhanging primers as single stranded DNA regions once again, an exonuclease can be used which removes up to 15 bases from either the 5' or 3' end of the sequence dependent on type. Once overhangs (also called sticky ends) have been formed, the annealing can take place between a target protein sequence and its intended vector backbone, by overlap of the two complimentary single stranded DNA overhangs. Any gaps are filled in by DNA polymerase before a ligase enzyme seals the nicks in the DNA backbone. A circular plasmid is formed if both ends of the insert and vector sequence are designs to join together.

A2.3 Affinity Chromatography

Protein production relies heavily on the use of affinity chromatography. If a gene construct expresses well in the chosen host system, cells can be grown to over express the target protein. Once harvested, the target protein is subjected to a variable series of chromatography steps to ensure that the final product is pure. Analysis using impure protein may lead to erroneous results due to contamination or degradation which can hinder progress on understanding the function of the target. One typical method used by researchers, is to use a gene construct containing a tag – a region of the protein that is able to interact with the solid phase of a specific affinity chromatography column. In this work, GHs were designed to contained be purified using Ni affinity chromatography, whereas the final LPMO construct was purified using streptavidin affinity chromatography. Ni affinity chromatography relies on columns packed with a matrix of agarose beads, attached to which are coordinated Ni ions. This technique is often called immobilised metal affinity chromatography, as the metal (which can be changed to other divalent metals such as cobalt) is held within the resin and the sample is flowed through the column. Metal chelatation to histidine residues was a known to occur in aqueous solutions and had been tipped as a method for protein fractionation the 1970's.[290] in 1987 Hochuli *et al* designed a

matrix system that is now extensively used in the preparation of biological molecules, were insoluble agarose beads are linked with nitrilotriacetic acid (NTA), which coordinates to a Ni ion through three oxygen atoms whilst the two remaining coordination sites are reversibly taken up by water.[291] A year later, Hochuli *et al* presented the first purification of a protein engineered to contain 6 consecutive histidine residues, a tag which interacted with this new NTA-Ni bound resin.[292] During the design stage of a gene construct it is now commonplace to incorporate such a tag at the N or C-terminal end of the protein sequence. The histidine side chain has a high binding affinity for Ni and is able to displace the water ligands around the NTA-Ni coordination site. The consecutive histidine repeats give a high likelihood of the protein becoming trapped within the column as the histidine residues interact with the immobilised Ni ion. Other non-tagged proteins and small molecule impurities will flow through the column with little interaction. During the chromatography procedure, one can use absorbance at 280 nm to monitor the amount of protein eluting from the column and once this reaches a baseline level during loading of the sample, it can be assumed that only the target protein (and potential other contaminants if they contain regions of surface histidine residues) remains within the column matrix. To collect the purified tagged protein, the column is washed with imidazole at either a specific concentration (if the preparation has been carried out previously) or using an increasing gradient. At a certain imidazole concentration that is very specific for the target protein, the histidine tag will be displaced from its coordination with the Ni ions in favour of the imidazole molecules, which have a greater binding affinity.[293]

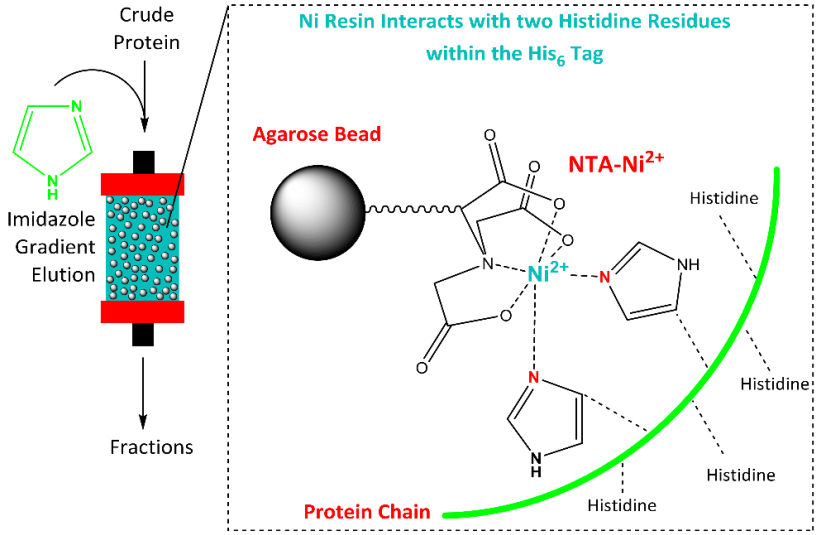


Figure 155 Ni affinity chromatography using agarose beads linked with nitrilotriacetic acid (NTA). A single Ni coordination site can interact with two histidine residues.[294]

Streptavidin is a robust protein from *Streptomyces avidinii* known to exhibit extremely high affinity for a small molecule known as biotin (also known as vitamin B7). Purification strategies that use columns containing specifically engineered streptavidin (Strep-Tactin) use the same principle, in that the target protein contains a tag able to interact with the column matrix. In this case, the tag is known as a strep tag and consists of the specific amino acid sequence: Trp-Ser-His-Pro-Gln-Phe-Glu-Lys. The strep tag protein remains bound to the Strep-Tactin column whilst other contaminants are removed. The tagged protein is displaced with desthiobiotin, a derivative of biotin which has a strong binding affinity for the Strep-Tactin. The column is washed with a low concentration of desthiobiotin which elutes the protein typically with very high peak resolution.[295-296]

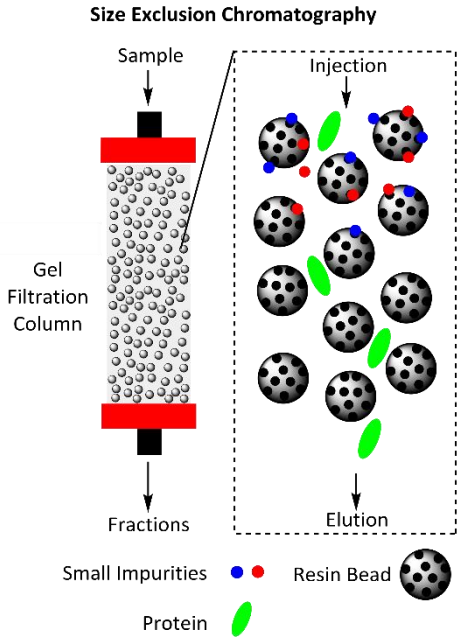


Figure 156 Size exclusion chromatography showing a diagram of the agarose bead resin containing small pores in which small impurities move through.[294]

Affinity chromatography often preludes size exclusion chromatography, where the latter often used to 'polish' the purity of the final sample. Size exclusion chromatography uses a matrix containing a preparation of agarose beads known as Sepharose. The beads contain pores in which molecules can move into which efficiently causes size separation as molecules of different sizes interact with the pores to different extents. Smaller molecules are more prone to interaction with the bead pores and thus take longer to diffuse through the column matrix. Larger molecules are quick to proceed through the column as they bypass the pores and move between the packed beads. Size exclusion chromatography will often produce high separation resolution and is an excellent method to remove

contaminants or degradation products that have been retained during previous chromatography steps.[297]

A2.4 Thermal Shift Analysis

The primary amino acid sequence of a protein defines its secondary structure, producing α-helices, β-sheets and loop regions, which in turn define its tertiary structure (the overall shape). Proteins require a well defined structure to be able to carry out very specific functions. Proteins also evolve to work in a specific environment and any changes away from the preferred environmental conditions may affect the stability and homogeneity of the protein; the point at which the defined structure begins to turn towards disorder. Environmental changes in temperature, pH, salt concentration, and buffer conditions can cause alterations in the stability of the protein. In a discussion on the 'induced fit theory' of enzymatic action, Koshland notes that the *"flexible nature of portions, if not all, of the protein chain is adduced from many sources, and these changes can be caused by small molecules, charged or uncharged"*.[298] The 'ligand-induced conformational stabilisation' of proteins by various chemical agents, such as substrates, co-factors, inhibitors, metal ions and other proteins, is a well known and an experimentally observed occurrence.[299] Koshland goes on to postulate the importance of the precise orientation of the amino acids involved in the catalysis of a reaction and how, upon binding the correct substrate, the residues will move into their optimum catalytic position. Small movements in the position of a catalytic acid or base will affect the coordination of other residues or disrupt the hydrogen bonding network of water normally found within a binding pocket; this is commonly observed when one tries to design protein-mutants to assess the importance of particular residues, where changes in one residue can cause a protein miss fold or disrupt function Protein stability can move in three 'directions'; improvement, tolerance or disruption. A positive shift in protein stability improves the temperature range in which a protein normally remains stable and more often than not, active form. A negative change in the stability of a protein is usually in response to unfavourable conditions such as high temperatures, changes in pH or non-specific binding events. Certain protein are more able to tolerate changes in environment that other, for example, proteins evolved to work in specifically difficult living environments such as hydrothermal vents.

The thermal shift assay relies on the concept of protein disorder and as the name suggests, uses changes in temperature to denature a protein. This is a simple and cost effective

method of protein analysis, with the experiment being carried out in a standard qPCR machine. SYPRO orange is a fluorophore that is quenched in water, but is able to fluoresce when binding occurs to a hydrophobic surface such as an aromatic side chain.[300] SYPRO orange, is mixed with a protein and the fluorescence of the dye monitored, as this parallels the stability of the protein as the temperature is increased. The hydrophobic nature of aromatic amino acid side chains mean they are typically held buried within a proteins structure, away from the solvent assessable surface. As such, introducing SYPRO orange to a protein in a preferable set of conditions will yield a low level of background fluorescence when tested. Changing the conditions away from the optimal will result in a negative effect on protein stability, leading towards the exposure of aromatic side chains normally buried within the protein structure as it denatures.

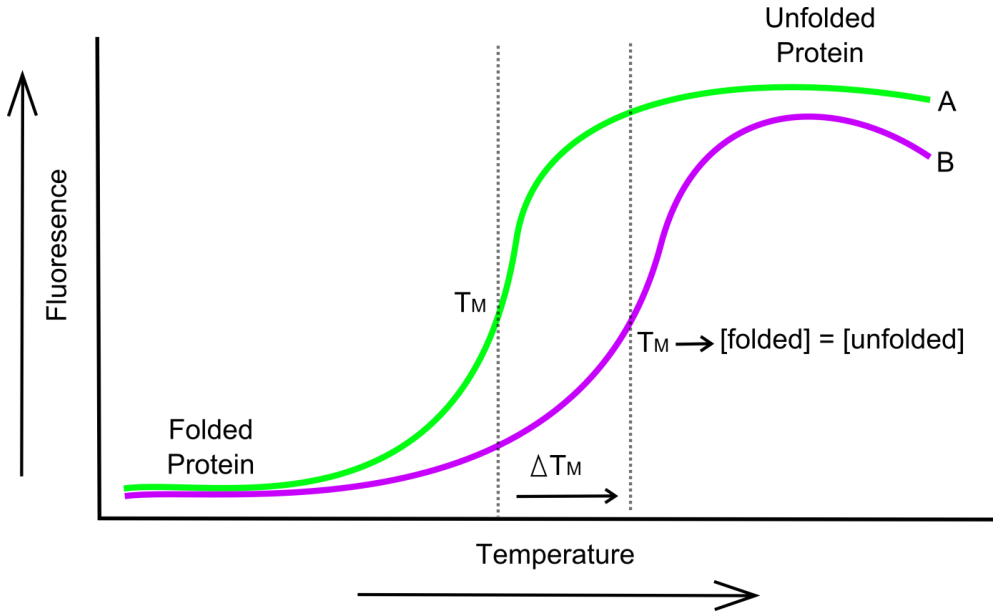


Figure 157 Representation of a thermal shift experiment where curve A is 'apo' protein and curve B is protein with a ligand. The curve shows the change in protein state as the temperature increases the protein denatures, allowing the interaction of the dye molecule with aromatic residues Causing fluorescence. The protein melting temperature is the midpoint of the sigmoidal curve and can be mathematically calculated.

Normally, a protein is first tested for its 'apo' melting temperature (TM), this gives an excellent indication of its inherent stability and environmental working preference. For instance, a protein produced by bacteria living within hydrothermal vents will display a drastically different thermal profile to an enzyme produced by an organism living in the arctic tundra. To cover the wide range of potential thermal activity, the protein is mixed with the dye and the temperature is increased in incremental amounts from 20°C to 91° in increments of 1 °C over 71 cycles. During this time the fluorescence of the dye is monitored (usually in a qPRC machine) at the wavelengths for excitation and emission, 490 nm and 580

nm, respectively. All proteins will denature with increasing temperature, but the temperature at which they start to denature is specific for individual proteins. As the temperature starts to increase, the amount of fluorescence will move away from the base line. As unfolding begins more aromatic residues become exposed, increasing the fluorescence. Once the protein is completely denatured and all tertiary structure lost, the level of fluorescence will plateau (and/or drop due to protein aggregation). In an ideal experiment, in which a protein at low temperature is completely folded with no hydrophobic residues exposed, a classic sigmoidal curve can be plotted as fluorescence against temperature. The midpoint of the sigmoidal curve is defined as the apparent protein melting temperature.[301] A non-ideal result is observed when there is a lack of the classic sigmoidal shape; for example, there may instead be a straight line with a decreasing gradient, indicating the protein is already in an unfolded state and aggregation is starting (and possibly an indication of insufficient storage conditions) or there may be a high fluorescence baseline, due to surface exposed hydrophobic residues or an interaction of the dye with the buffer or any additives being tested, **Figure** 158, for generic examples.[301] Careful consideration is needed in non-ideal experiments and analysis can continue even if curves become complex; for example, proteins have been known to display multiple melting events. If one can pinpoint the start of a melting curve despite a high background signal, the curve can still be analysed by applying a fitting procedure such as the Boltzmann model, **Figure** 158, curve B. The key is carrying out all the necessary controls to determine why there is a move away from the idea shape; principally controls focusing on whether the buffer or ligand affects the fluorescent signal.

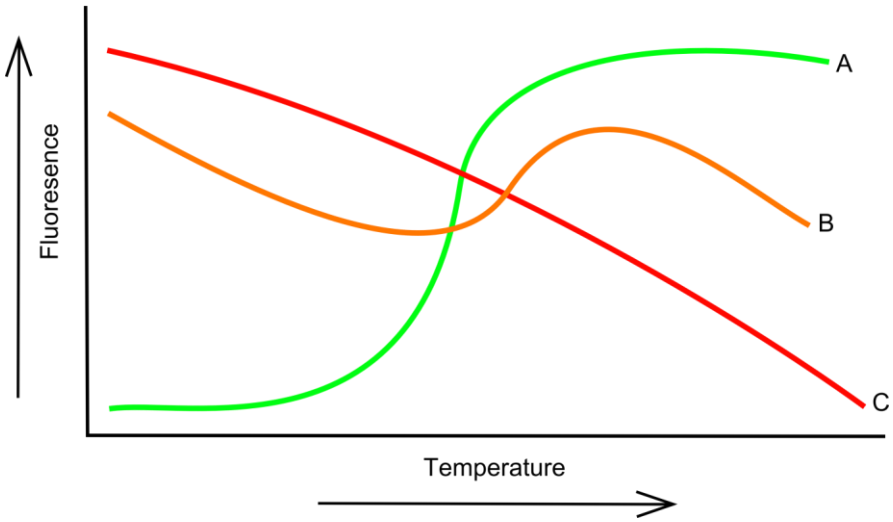


Figure 158 Representation of three typically observed curve types during TSA experiments. Curve A represents the ideal sigmoidal shape, in which there is little background fluorescence at the start at the experiment, peaking as the protein unfolds, followed by a plateau. Curve B represents an experiment where there is a significant background fluorescence before the protein begins unfolding. Curve C shows no increase in fluorescence with temperature, indicating that the protein of interest is already in an unfolded state.

Once the melting temperature of the 'apo' protein has been defined, alterations can be made to see how the protein tolerates changes in conditions. Experiments such as this are important in finding out the optimal working conditions for a protein of interest. Industrial enzymes are normally famed for their ability to work at certain temperatures. One excellent example is the use of biological washing powders – the enzymes are more efficient at 30 °C. Creating industrial enzyme cocktails for biomass degradation requires efficiency of all the component enzymes. There is little point in producing an enzyme for this type of work, if when combined with the typical enzymatic mixture, it is unable to work efficiently in the pre-set conditions. Information gained during these experiments can ascertain whether a new protein could work in the same conditions as others and as such whether it is worth investing both time and cost producing it.

The use of thermal shift assay experiments in this work has mostly centred on the identification of protein-substrate interactions. Introduction of a substrate to a protein can lead to a change in protein melting temperature if a binding event occurs. Binding of a substrate within an active site is generally favourable; otherwise this would not lead towards a protein successfully carrying out its intended function. A change in protein melting temperature is considered significant if there is a shift of 1 °C in either a positive or negative direction away from the melting temperature of the 'apo' protein and this is consistent with experiment repetition. However, the assay is not able to discriminate between the changes in melting temperature which can occur from specific and non-specific binding events.

Therefore, the technique can often produce false positive results. Despite this, in the majority of cases, thermal shift assays can be extremely helpful in identifying potential substrates for an enzyme with an as of yet uncharacterised function.

A2.5 Thin Layer Chromatography

Thin layer chromatography is an analytical technique capable of separating out the components of a mixture and in this work has been used to identify the reaction products of the enzymatic hydrolysis of polysaccharides.

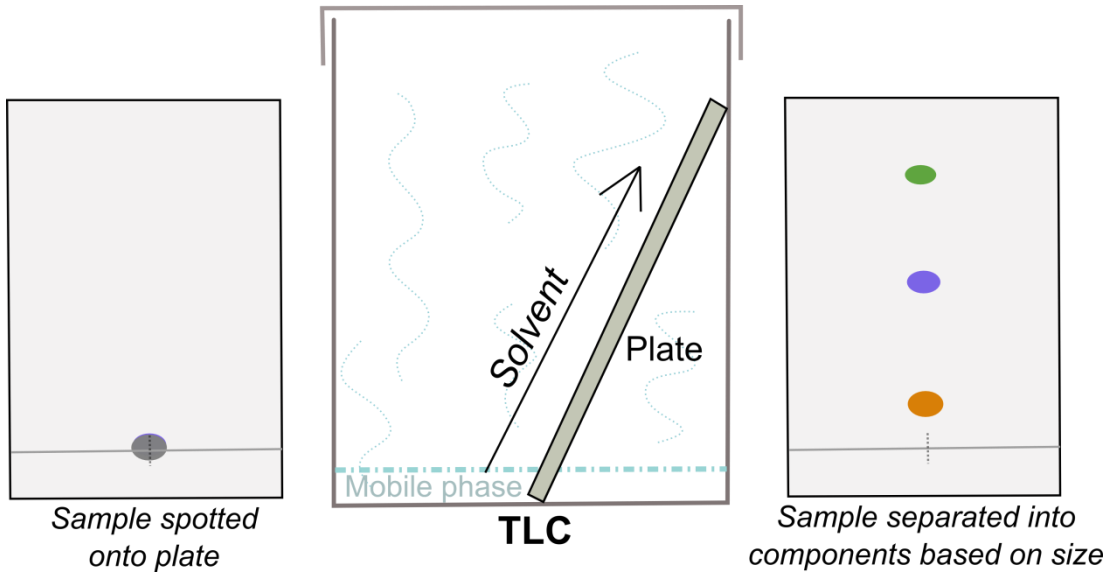


Figure 159 Diagram showing general set up of a TLC experiment. The stationary phase plate is placed in a tank that has been pre-equilibrated with the mobile phase, meaning the air within the tank is saturated with solvent vapour. Samples spotted onto the plate are above the level of the mobile phase, and solvent moves up the plate over time. On the right is an example of a typical plate after visualisation, where three samples have moved at different rates through the stationary phase as carried by the mobile phase. This is related to the properties of the components and in the case of this work is more often than not due to size. The green spot represents a small carbohydrate molecule, whereas the orange spot is a larger carbohydrate that has not travelled far up the plate.

The principles of TLC are not dissimilar to many other forms of chromatography used within protein production or other analytic techniques. A typical experimental set up is shown in Figure 159, where a plate, coated in what is known as the stationary phase (often alumina) is placed in a tank containing a solvent mixture, the mobile phase. The diffusion of the solvent up the plate carries the components of the sample, however, the rates at which they are carried depends on the properties of the specific component. For the analysis of carbohydrates, the most important factor is size. The plate is either coated with silica gel or alumina, where the exposed surface has a layer of hydroxyl groups, creating a polar environment. Larger oligosaccharides have a higher adsorption affinity for the plate surface

as they have more functional groups capable of forming hydrogen bonds with the OH groups on the stationary phase surface. As such a larger oligosaccharide will not move very far up the plate, whereas a smaller single sugar unit will move a lot further, providing good separation between the two components. Separation of components also depends on their solubility in the mobile phase. Careful consideration is required in the first instance to ensure the correct mobile phase is used. Carbohydrates are polar molecules and as such can be run on the TLC in their native form without alteration, but they do require a polar solvent for efficient dissolution. All TLC experiments in this work were carried out using a solvent system containing butanol, acetic acid and water in a ratio of 2:1:1. Plate set up requires precision; in the analysis of oligosaccharide products it is often the case that the product concentration is not very high, so multiple applications of a sample to the same spot is often required. The samples must also be spotted on a straight line, and the plate placed into the solvent completely parallel to the solvent surface to ensure there is no skewed movement of solvent up the plate. The samples must be spotted so they sit above the level of the solvent within the tank, otherwise dissolution into the mobile phase of the sample from the plate will most likely occur. Despite these requirements, TLC is a quick and simple method of checking the activity of a protein on a substrate.[302] Staining of the plate is required in order to visualise the product bands that have moved up the plate. For carbohydrate analysis, a mixture of sulphuric acid, ethanol and orcinol is sprayed onto the plate, before it is dried of excess solution and heated to around 100 °C. Once developed, bands often take on a purple hue and can be photographed before the colour fades.

A2.6 MALDI-TOF Mass spectrometry

Matrix assisted laser desorption/ionisation time of flight mass spectrometry is the ideal method for carbohydrate analysis. MALDI-TOF-MS is a 'soft ionisation' technique and is able to analyse complex mixtures of carbohydrates at both low and high degrees of polymerisation with little effect on the analyte structure. Combination of a sample with a matrix is thought to improve desorption of non-volatile compounds, such as carbohydrates, when the spot is irradiated by the laser and enhance the observable signal of the components with little compromise on the signal to noise ratio.[303]

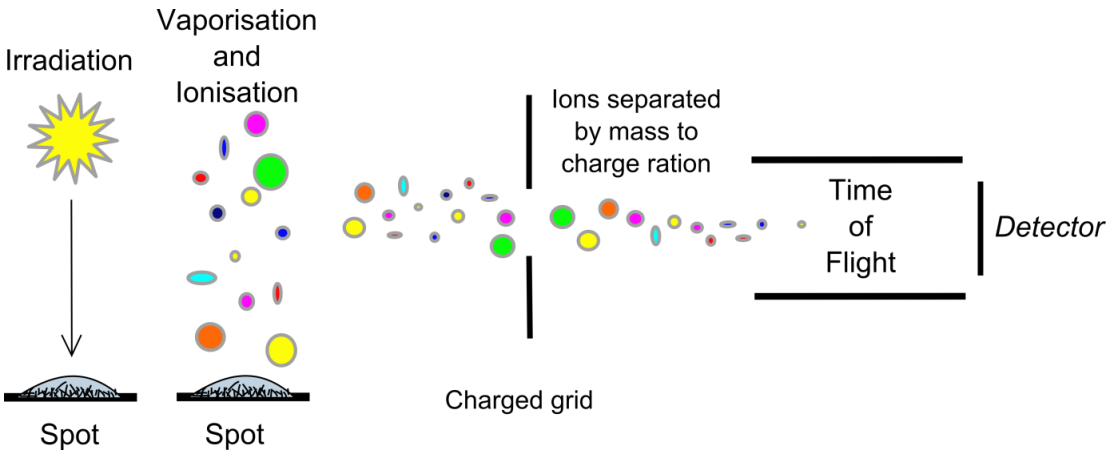


Figure 160 Diagram of MALDI-TOF mass spectrometry. Sample is held within a dried matrix spot and upon irradiation with a laser, vaporisation of matrix and sample occurs, producing ions. The ions are charged (positive) and then separated by their mass to charge ratio (m/z) during their progress through the time of flight chamber, before being detected.

Carbohydrate samples are mixed with the matrix, often 2,5-Dihydroxybenzoic acid (DHB), spotted onto a specialised plate and left to dry. Solvent evaporation causes the matrix to crystallise, trapping the carbohydrate molecules in place. DHB normally crystallises from the edges of the plate spot, where long needle shaped crystals can be seen pointing towards to the centre of the spot. The efficiency of matrix crystallisation can often leave regions of the spot with little signal from the desired sample. Issues can occur in which crystallisation does not occur and the spot is classed as 'glassy', in which the laser simply bounces off. Often is the case that the amount of products within a sample is unknown and changing the ratio of matrix to sample can alleviate issues with poor crystallisation; it is worth testing both low and high ratios of matrix to sample to obtain the best spectra. The user can also move the laser around the spot to target matrix crystals by eye, and collect spectra in localised regions of the spot. Considering these limitations and biases caused by human interpretation of the spot, the signal strength observed for a component can vary significantly from spectra to

spectra. However, if several spectra are taken, averaging produces a spectrum that provides a good indication on the ratio of molecules in the sample. The major ion peak is consistently $[M + Na]^+$, unless there is an abundance of other metals such as potassium. Little fragmentation of carbohydrate structures is observed and the spectra can analyse a wide range of molecular weights. Generally speaking, there is a loss of sensitivity below 1 kDa and at high molecular weights, above 3000 kDa. MALDI-TOF-MS is generally very tolerant of contaminating agents, and does not suffer from the same issues as other MS techniques in the use of buffers. Hence, a sample can be taken straight from an enzymatic reaction with little need to prepare the sample in any way. In a review by Harvey, he notes an excellent example of this; a study was carried out in which a shop bought gummy bear was simply dissolved in water and filtered, before being mixed with matrix, led to the spectra showing maltose oligomers up to 35 units in length! Whilst the power of MALDI-TOF-MS is self evident in the analysis of carbohydrates, it does have its limitations. For example, carbohydrates of a certain mass can be due to several different isomeric configurations and peaks can often indicate a molecular weight, but not the configuration. It can be hard to identify these isomer patterns from MALDI-TOF-MS alone.[304] However, this is a small caveat and MALDI-TOF has been used throughout this work in the analysis of soluble reaction products resulting from enzymatic degradation of many different polysaccharides, providing key evidence for substrate specificity of a target enzyme.

A2.7 High-performance anion-exchange chromatography with pulsed amperometric detection

High-performance anion-exchange chromatography with pulsed amperometric detection (HPAEC-PAD) has been used in this work to analyse the kinetic efficiency and range of hydrolysis products formed during enzymatic substrate degradation. This versatile technique can be used to separate out the different sugar products obtained from an enzymatic reaction, and assist in the identification of the observed products through comparison with external standards. The HPAEC-PAD system is simple in overview; samples are injected onto a specific anion-exchange column (CARBOPAC), flowed over the column using a buffer gradient and finally detected electrochemically upon exiting the column. In much the same way as TLC, the rate of movement of particular sugars through the column matrix is dependent on several factors; such as size, structure and conformation. The running buffer (or mobile phase) contains sodium hydroxide at a high pH which causes the sugars in the sample to become oxyanions. The oxyanion-sugars are then able to bind to the column by interacting with the polystyrene-divinylbenzene beads (or *stationary phase*). Once a sample has been injected onto the column, a sodium acetate gradient is applied as the mobile phase. Movement of the sugars through the stationary phase matrix occurs over time and is controlled by the increasing gradient of acetate ions. Weakly interacting sugars are easily outcompeted by a low concentration of acetate ions, whereas larger sugars will remain bound for a longer period of time before elution. Subtle differences in carbohydrate structure and differences in pKa can change the relative affinity of the sugars for the column stationary phase. As such, even extremely similar sugars, or those moieties with the same molecular weight will have different retention times, as shown in Figure 161.

After separation on the column, the sugars are detected using pulsed amperometric electrochemistry. The carbohydrates released from the column are immediately oxidised on a gold electrode using a series of potentials known as a wave-form, Figure 162. A carbohydrate will be oxidised by the electrode at potential E1, after which there is a slight delay before the current is measure; the current is integrated over the time delay which yields charge, hence the signal response is measured in coulombs. Very sharp narrow peaks are often observed as the sugar is detected and is due to the quick transition between potentials. The series of wave-form potentials used during detection also acts as a self

cleaning mechanism of the electrode surface. This extremely quick method of self maintenance ensures smooth detection during separation.[305-306]

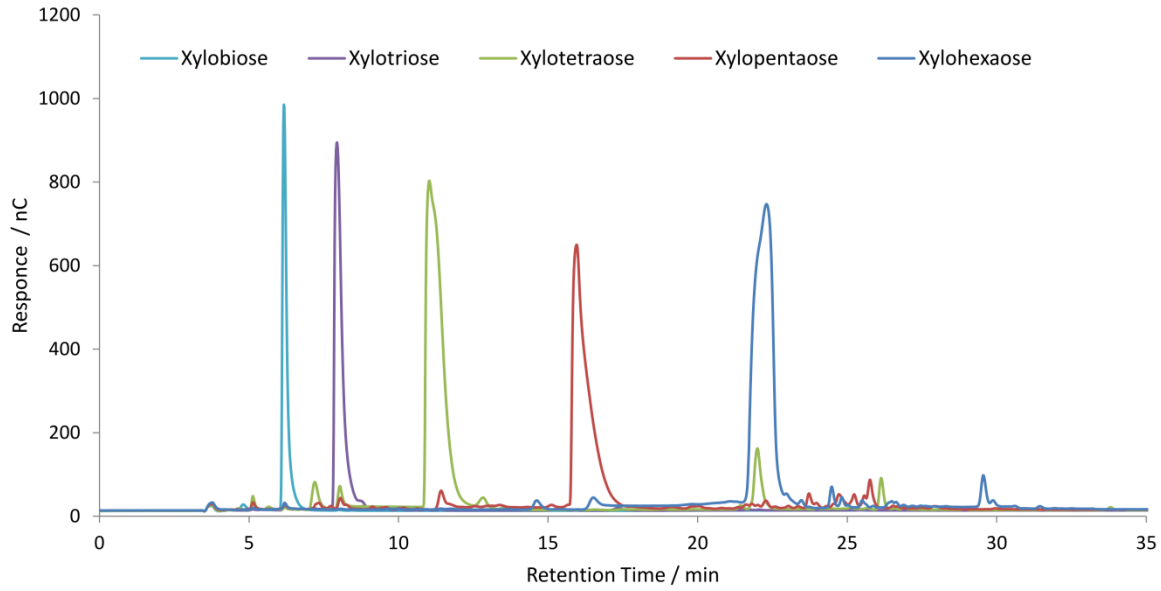


Figure 161 HPAEC-PAD spectra showing the retention times of five xylooligosaccharides. The component sugars were injected into the anion-exchange column as a combined mixture, and as size increases so does retention time of the sugar on the column.

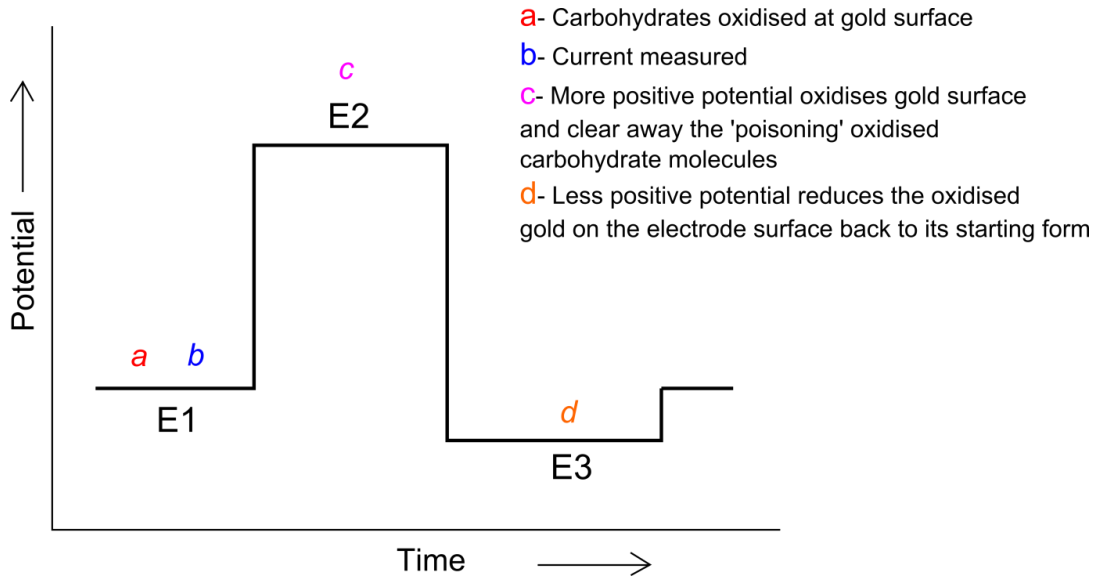


Figure 162 The waveform used during electrochemical detection of carbohydrates eluting from the anion-exchange column. Three potentials are used; binding and measurement of the carbohydrate occurs at E1, removal of the carbohydrate is performed at a more positive potential E2, and finally the surface is reset to a less positive potential, E3.

HPAEC-PAD can be used to evaluate certain kinetic parameters of an enzyme reaction with a substrate. A general equation for an enzymatic reaction is shown below, where E is enzyme,

S is substrate, ES is the enzyme-substrate complex and P is the resulting product of the reaction.

$$E + S \underset{k_{-1}}{\overset{k_1}{\leftrightharpoons}} \mathrm{ES} \overset{k_{cat}}{\rightleftharpoons} \mathrm{E} + \mathrm{P}$$

In any reaction we only known the concentration values of E and S, but not the amount of ES present. However, we do know that the rate of the reaction, V_o is related to the concentration of ES through a rate constant known as k_{cat}.

$$V_o = \mathrm{k_{cat}[ES]}$$

(Equation 1)

Where, k_{cat} is the ***turnover number***, or how many substrate molecules are converted into products by a single enzyme active site per second. Under steady state conditions, the rate of formation of the enzyme-substrate complex, k_1 is equivalent to the rate of dissociation of the enzyme-substrate complex, k_{-1}. As such we can describe the overall rate of reaction as follows;

$$k_1[E][S] = k_{-1}[ES] + k_{cat}[ES]$$

(Equation 2)

which rearranges to;

$$\frac{k_{-1} + k_{cat}}{k_1} = \frac{[E][S]}{[ES]}$$

(Equation 3)

The left hand term in equation 3 can be simplified as K_M and is a ratio of how much free enzyme and substrate exists against those in a complex. It is therefore a measure of the substrate binding affinity for the enzyme active site. A high value of K_M indicates that the substrate exhibits poor binding affinity for the enzyme active site (more free enzyme/substrate), whereas a lower value suggests good binding affinity (more enzyme-substrate complex and a higher value of k_1). As such,

$$K_m = \frac{[E][S]}{[ES]}$$

(Equation 4)

which rearranges to give a term for [ES];

$$[ES] = \frac{[E][S]}{K_m}$$

(Equation 5)

The term for [ES] can be substituted into our original rate of reaction equation (equation 1) as follows;

$$V_o = k_{cat} \frac{[E][S]}{K_m}$$

(Equation 6)

The catalytic efficiency of an enzyme on a particular substrate is described by k_{cat}/K_m. Using HPAEC-PAD and a method known as substrate depletion kinetics, the value of k_{cat}/K_m for a particular enzyme/substrate combination can be calculated. The value of this rate term is always limited by the rate of formation of the enzyme-substrate complex and as such the highest value this can ever be is k_1. Enzymatic hydrolysis is monitored overtime by evaluating the decrease in the peak area of the substrate as determined experimentally by HPAEC-PAD. Several initial runs may be required to find suitable concentrations of enzyme and substrate that yield usable linear results. The substrate concentration used in the evaluation of k_{cat}/K_m must be lower than the value of K_m for the particular enzyme/substrate combination being tested; high amounts of substrate lead to saturation of enzyme binding pockets, even if the natural affinity of the substrate for the enzyme binding site is low, skewing the kinetic analysis of the results.

To provide enough data (rate of change, compared with the detection response) reactions must be slow enough that only 80% of the substrate is hydrolysed within the time frame of the reaction. Enzymatic hydrolysis reactions are set up containing a known concentration of substrate at the start of the reaction and aliquots removed at regular time intervals and enzyme inactivated. All samples taken from a reaction are mixed with a known concentration of an internal standard, commonly fucose. Fucose is a small sugar moiety and under specific experimental conditions, displays a quick retention time, meaning it is eluted first and does not interfere by overlapping with any other sugar peaks. The depletion of substrate during the reaction over time can be related to k_{cat}/K_m through the following equation;

$$kt = ln\,(S_0/S_t\,)$$

where $k = \left(\frac{k_{cat}}{K_m}\right).[E]$, t = time and S_0 and S_t are the substrate peak area at time 0 and t, respectively. Before each set of HPAEC-PAD reaction aliquots are analysed, a fucose external standard is tested to check the performance of the machine. Due to the mechanistic limitations of the instrument, signal loss over the course of several runs can be a commonly encountered issue. To combat this, normalising the signal of the peak area of interest to both an internal and external standard (run at the start of the HPAEC-PAD analysis) removes the ambiguity created by any loss of signal during the subsequent runs. The peak areas observed in the HPAEC-PAD traces can be normalised against the fucose internal standard and external standard as follows;

$$Normalised\ substrate\ peak\ area = \frac{Substrate\ peak\ area}{fucose\ internal\ standard\ peak\ area} \times fucose\ external\ standard\ peak\ area$$

The normalised peak areas can then be manipulated into the form of $\ln(S_0/S_t)$ and plotted against time. A positive gradient is produced as the change in substrate peak area will always increase from 0; at time point 0, S_0/S_t is equal to 1, and ln(1) is equal to 0. From the above equation, plotting $\ln(S_O/S_t)$ against time gives a gradient of *k*. The value of k_{cat}/K_m (M^{-1} min^{-1}) is then simply determined by dividing the gradient by the enzyme concentration.[307-309]

A2.8 DNSA reducing sugar assay

The activity of an enzyme on polysaccharides can be investigated using the 3, 5-dinitrosalicyclic acid (DNSA) reducing sugar assay. The DNSA assays enables quantification of enzyme activity on insoluble substrates. For every chain of a polysaccharide, there are two chain ends, but due to the directional nature of glycosidic linkages there can only be one reducing end per chain. The reducing sugar is classed as the moiety at the end of a chain in which the anomeric carbon in the C1 position is free to ring open, exposing an aldehyde functional group, **Figure** 163. It is this aldehyde at the anomeric carbon that reacts with the DNSA reagent; upon heating, the 3-nitro group on DNSA is reduced to an amino group, whilst the aldehyde on the sugar is oxidised to carboxylic acid, **Figure** 164

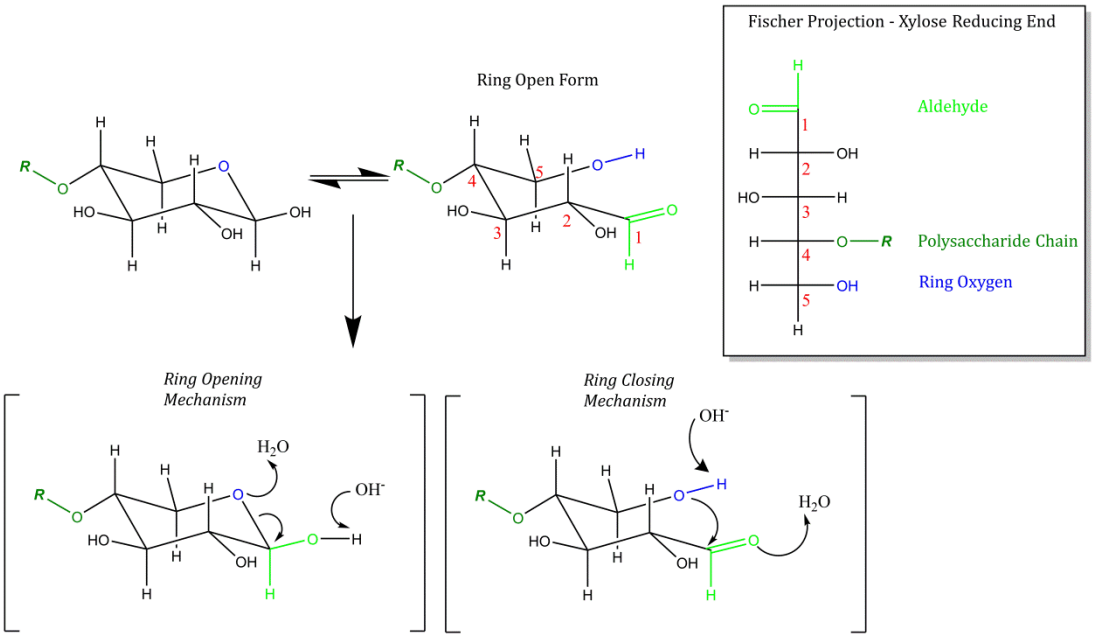


Figure 163 Schematic showing the spontaneous ring opening mechanism of chain end xylose, where R represents the xylan chain. Ring opening and closing in solution is shown by action of an acid and a base, H2O and OH-, respectively. The xylose chain end is also depicted in the insert box as a Fischer projection.

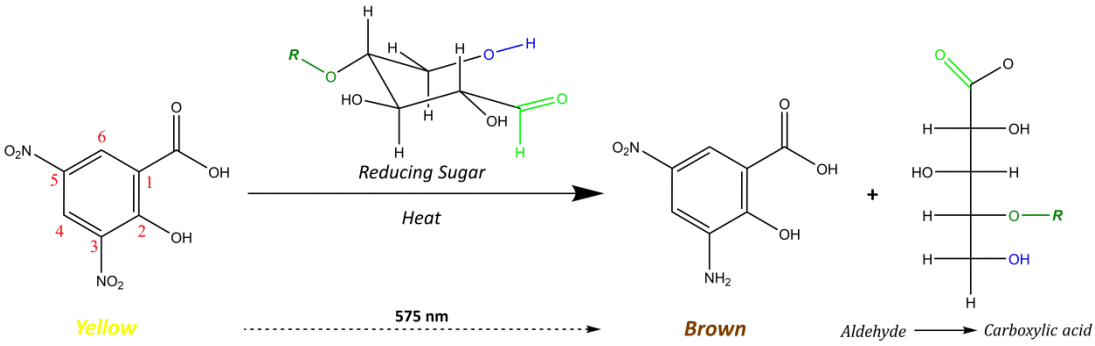


Figure 164 Schematic of DNSA reaction with reducing end xylose, where R is representative of the rest of the polysaccharide (xylan). Upon heating, the 3-nitro group is reduced to an amino group, whilst the xylose aldehyde is oxidised to a carboxylic acid. This is accompanied by a colour change of the DNSA agent from yellow to brown.

The DNSA agent reacts with the reducing ends of polysaccharide chains in a one to one ratio, leading to a colour change of the DNSA from yellow to brown, measurable at 575 nm. As such, monitoring a hydrolysis reaction with the DNSA assay allows the rate of enzymatic activity to be tracked by spectrophotometry. Aliquots removed from the enzymatic hydrolysis reaction are mixed with DNSA agent (1 % (w/v) DNSA, 0.2 % (v/v) phenol, 1 % (w/v) NaOH, 0.002 % glucose, 0.05 % (w/v) $NaSO_3$) at set time points. Phenol can be used to improve the intensity of the colour whilst $NaSO_3$ is used to stabilise the colour. Mixing the removed reaction aliquots with the DNSA agent is efficient in stopping the enzyme reaction

due to the drastic change in pH (alkaline), which usually caused the enzyme to degrade. Development of colour is caused by heating the reaction aliquots at 90 °C. A standard curve of a single sugar substrate, such as glucose, is used to quantify the released reducing sugar chain ends. The polysaccharide is added as well create a baseline of the amount of reducing sugar present in the substrate – as each chain end will account for a potential reaction with the DNSA agent. The standard curves were used to quantify the amount of reducing sugar present to the measured absorbance at 575 nm of each reaction aliquot.

$$k = \frac{([reducing\ sugar]/t)}{[substrate]}$$

(Equation 7)

where

$$k = \left(\frac{k_{cat}}{K_m}\right).[Enzyme]$$

Equation 7 is simply a rearrangement of equation 6 presented earlier and thus can be written as follows,

$$k = \frac{([reducing\ sugar]/t)}{[substrate]} \equiv \frac{V_o}{[S]} = k_{cat}\frac{[E]}{K_m}$$

where the amount of reducing sugar over time is a measure of the rate of reaction, $V_{0.}$

Reaction rates are obtained from linear regression of the calculated amount of reducing sugar against time. Plotting the observed gradients against the substrate concentration at which the reaction was carried out, gives k_{cat}. The gradient divided by the enzyme concentration gives k_{cat}/K_m (mg^{-1} ml min^{-1}) for the reaction of the enzyme on a particular polysaccharide substrate (which is often measured in mg ml^{-1} due to the insoluble nature of many substrates).[310]

A2.9 3D structure determination

The 3D structure of a protein is a vital tool in understanding the complex nature of proteins, and can provide many biochemical insights into the function and activity of a protein. The 3D structure of a protein can be determined using X-ray crystallography if a protein can be successfully crystallised. Crystallisation of a protein means formation of repeating structure, in which there is a common unit cell. The unit cell varies for each protein but always contains at least one copy of the protein within it. A crystal is made up of a repeating network of unit cells, which pack together in a mostly uniform way – although natural inefficiencies in the packing of protein molecules do occur. It is the repetition of the unit cell that allows the structure of proteins to be elucidated using X-ray crystallography.

Crystallisation of proteins in this work was carried out using vapour diffusion. The plates containing either 96 wells for screening, or 24 wells for optimisation can be set up using various conditions. Commercial crystallisation screens are normally used in the first attempts of protein crystallisation, and there are wide range of screens to choose from. Individual screens contain 96 different, or varying conditions containing mixtures of buffers, precipitants, additives at different concentrations and pH. Other factors such as protein concentration (starting concentrations of stock protein work best above 10 mg^{-1} mL) and temperature of plate incubation can be altered to assist crystal growth. Commercial screens generally rely on using a sparse matrix approach, whereby the conditions contained within the screen are those conditions which commonly led to successful crystallisation of other proteins. Several screens can be used at once, in a sort of 'shot gun' method which subjects the protein to a large number of possible conditions at the same time – however, this can be an expensive method and as such particular screens may be favoured for first attempts. If hits are found, the crystals grown in the screening plates are often too small for practical use due to the low volumes used during plate preparation (drop volumes are typically 0.3 nL) but sometimes crystals are large enough that can be removed and used directly for crystallography. Similarly small crystals can be taken and used for seed stocks, in which the crystals are crushed and diluted to provide new drops, containing larger amounts of protein concentration, nucleation points from which new crystals can grow. Optimisation of a hit condition is normally required for production of crystals for diffraction experiments. This involves setting up 24 well plates using the hit condition, with graduated changes across a couple of the conditions variables; for example, the pH of the buffer could be changes so that there are a selection of wells with pH above and below that of the original condition;

the concentration of other components such as precipitant can also be varied in small increments around that of the original value. These small changes can often yield more favourable conditions for crystal growth. Optimisation plates also provide larger crystals due to their use of larger drop volumes. In the hanging drop method, as used throughout this thesis, the well is filled with 1 mL of 'mother liquor' and suspended above is a drop held onto a glass slide by surface tension. The drop contains a mixture of the mother liquor and the protein sample, typically in a 1:1 ratio. This drop ratio can be changed to increase or decrease the protein concentration relative to the mother liquor. Once the well is sealed, vapour diffusion occurs overtime whereby water leaves the suspended drop as vapour as equilibrium forms; the concentration of components in the mother liquor is lower than that of the well solution due to the addition of protein sample. Vapour diffusion causes the protein sample in the drop to become supersaturated, a condition which is often required for crystallisation. [311]

Once a protein crystal has reached an appropriate size and it can be removed from the drop and tested for diffraction. Crystals may need to be broken apart before they can be removed, especially for certain crystal morphologies such as needles, which tend to cluster together. Crystals should always be collected as individuals, as groups of overlapping crystals will provide difficult data sets. Crystal removal is carried out using very small nylon loops. The loops can be chosen based on the crystal size and are small enough that they can be used to scoop up an individual crystal. The crystal is maintained in the loop by surface tension, and suspension in the loop with the mother liquor keeps a crystal from drying out or dissolving. Loops containing the crystals must be frozen quickly in liquid nitrogen; if the well condition doesn't already contain what is known as a cryo-protectant, the loop must first be dipped into a solution of mother liquor containing an appropriate additive such as glycerol or ethylene glycol. Whilst flash freezing in liquid nitrogen vitrifies the water contained in the loop/crystal, ice crystals can still be formed when the process is carried out by inefficient hands. Adding a cryo-protectant can limit formation of ice crystals around the protein crystal, and thus provide better diffraction; ice crystals often interfere with collected diffraction patterns of proteins, as observed by dark ring structures on the diffraction pattern. These ice rings can hinder structure solution by masking reflections from the protein atoms.

Crystals are then tested for diffraction using either a home x-ray source, or through beamline time at facilities such as the Diamond Light Source. The loops containing the crystals are mounted in the path of an x-ray beam and often kept under a flow of liquid

nitrogen to keep them from thawing. X-rays pass through the crystal and diffract based on their interaction with the positions of the atoms within the protein. A diffraction pattern, as shown in Figure 165 consists of a series of spots, or reflections. The position of the spots is related to the unit cell within the crystal, whilst the intensity of the reflections related to the distribution of atoms within the unit cell. As such, each reflection contains information about all the atoms within a protein. A single diffraction image taken from one angel of the protein is one of many thousands taken during an experiment. Crystals are rotated 180 °C with respect to the X-ray beam as this allows all possible reflections to be collected – missing reflections means information about atoms within the protein is lost and can lead to incomplete data sets. The need to take so many diffractions images mean crystals can be exposed to X-rays for a relatively long timeframe, in which they can undergo damage and loose resolution (data quality). X-rays can damage proteins by dislodging electrons, which can start chain reactions which degrades the protein crystal.

A data set therefore contains a large amount of information relating to the intensities of individual reflections. Inidividual reflections are unique, but can be measured multiple times during data collection which leads to higher redundancy within the data set. The greater the redundancy, the more information there is from which the intensities from all the images be averaged, giving a single intensity value for each reflection.

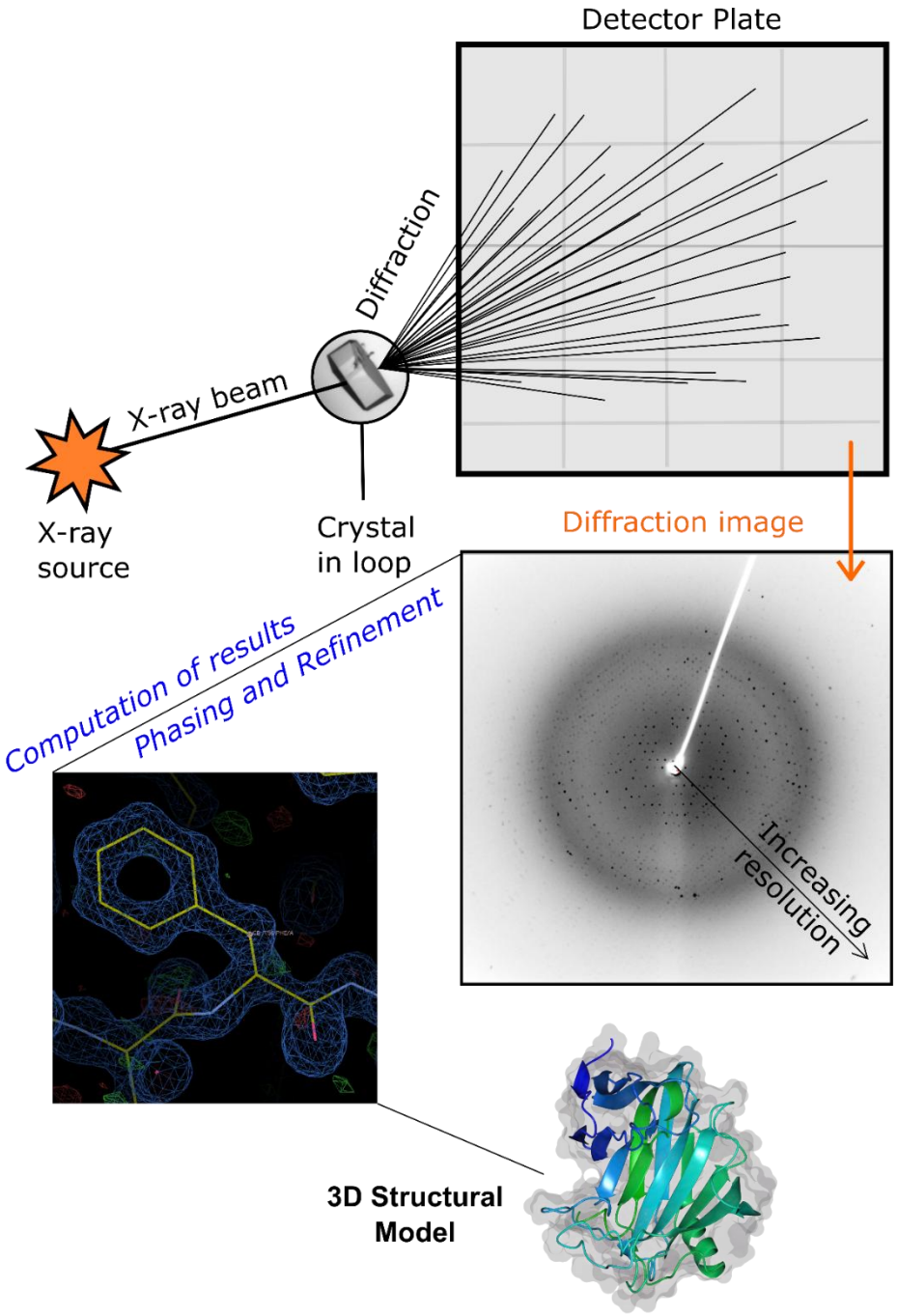


Figure 165 Diagram depicting X-ray crystallography of protein crystals.

A2.10 EPR

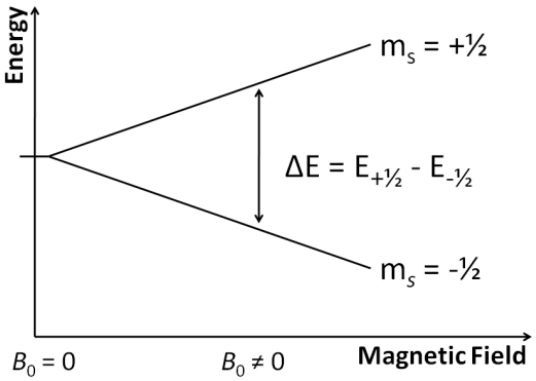


Figure 166 Energy level splitting due to the Zeeman effect of electron spin moments in an applied magnetic field with strength B0. Image taken from www.wikicommons.org

Electron paramagnetic resonance is a spectroscopy technique used for studying materials containing unpaired electrons, such a copper complexes. The experiment is able to analyse the magnetic environment of an unpaired electron, which is affected by the environment of the electron; i.e. coordinating ligands to a metal occur through overlap in molecular orbitals, which affects the magnetic moment of an unpaired electron. EPR excites the spin of electrons, causing a splitting effect as shown in Figure 164; whereby the magnetic moment of an electron pairs parallel or antiparallel to the applied external magnetic field. An unpaired electron can move between these two levels, by adsorption or emission of a photon. The experiment subjects a sample to a fixed microwave frequency, as the magnetic field (B0) is scanned. The microwaves energy can induce transitions in the unpaired electron between the two spin states shown in Figure 166 and as the magnetic field is scanned, different transitions occur. The equation, $\Delta E = g\beta B0 = h\nu$ describes the change in energy between the two spin states. B is a constant and the magnitude of the magnetic field is known, as well as the frequency of the microwave energy. Thus the g-factor can be calculated and is it this that is often compared. The g-factor gives information about the amount of spin-orbit coupling to the unpaired electron, through overlap of molecular orbitals. Changes in the environment (such as different ligands) can affect the spin-orbit coupling and alter the value of g. The g values described in this thesis were denoted with directionality in terms of x,y and z. These frames are axes from which the effect of different electronic orbitals (D for copper) constitute to the value of g. A second parameter, known as the A tensor represents coupling of the unpaired electron with atomic nuclei. The g value and A tensor are coupled in an antiparallel fashion, whereby as g goes up, the A value goes down. Plots of g values against the A tensor yields what is known as a Peisach-Blumberg

plot, from which it can be seen that splitting into groups appears for copper containing complexes; known as type 1 and type 2 copper for example. [216, 312]

Appendix 3:

Protein Sequences, Vector Sequences and Plasmid Maps

A3.1 Protein Sequences

All GH sequences start with CC to move insert into correct reading frame in pET28a. TAA is added as stop codon at the end, which translates as a *.

A3.1.1 *Tt*GH5_2

Codon optimised sequence:

CCATGGGCAGCAGCCATCATCATCATCACAGCAGCGGCCTGGAAGTTCTGTTCCAGGGACCAGCAGATGT
GCCGCCGCTGACCGTTAGCGGTAACCAGGTGCTGAGCGGTGGCGAGGCGAAGAGCTTTGCGGGCAACAGCTT
CTTTTGGAGCAACACCGGTTGGGGCCAAGAGCGTTTTTATAACGCGGAAACCGTTCGTTGGCTGAAAGACGAT
TGGAACGCGACCATTGTTCGTGCGGCGATGGGTGTGGACTTCGATGGCAGCTACATCCCGGAGCACGAAGAC
GCGGATCCGGAGGGTAACGTTGCGCGTGTTCGTGCGCTGGTTGACGCGGCGATTGCGGAGGATATGTATGTG
ATCATTGACTTTCACACCCACCACGCGGAAGACTACCAGGCGGAGAGCATCGAATTCTTTGAGGAAATGGCGA
CCCTGTACGGTGGCTATGATAACGTGATCTACGAGATTTATAACGAACCGCTGCAAATTAGCTGGGACAACGTT
ATCAAGCCGTATGCGGAGAGCGTGATCGGTGCGATTCGTGCGATCGACCCGGATAACCTGATCATTGTTGGCA
CCCCGACCTGGAGCCAAGATGTGGATGCGGCGGCGCGTAACCCGATTACCAGCTACAGCAACATCGCGTATAC
CCTGCACTTTTATGCGGGTACCCACGGCAGCTGGCTGCGTGATAAAGCGCGTAACGCGATGAACAGCGGTATT
GCGCTGTTTGTTACCGAGTGGGGTACCGTGAACGCGGATGGTGATGGTGCGCCGGCGGTTAACGAAACCCAG
CAATGGATGGATTTCCTGAAGCAGAACAACATCAGCCACCTGAACTGGAGCGTTAGCGACAAACTGGAAGGTG
CGAGCATTGTGCAACCGGGTACCCCGATCAGCGGTTGGAACGCGAGCGACCTGACCGCGAGCGGTACCCTGG
TGAAGAACATTGTTAGCAACTGGGGTACCACCATCGGCTAA

Protein sequence:

MGSSHHHHHHSSGLEVLFQGPADVPPLTVSGNQVLSGGEAKSFAGNSFFWSNTGWGQERFYNAETVRWLK
DDWNATIVRAAMGVDFDGSYIPEHEDADPEGNVARVRALVDAAIAEDMYVIIDFHTHHAEDYQAESIEFFEE
MATLYGGYDNVIYEIYNEPLQISWDNVIKPYAESVIGAIRAIDPDNLIIVGTPTWSQDVDAAARNPITSYSNIAYTL
HFYAGTHGSWLRDKARNAMNSGIALFVTEWGTVNADGDGAPAVNETQQWMDFLKQNNISHLNWSVSDKL
EGASIVQPGTPISGWNASDLTASGTLVKNIVSNWGTTIG*

A3.1.2 *Tt*GH5_4

Codon optimised sequence:

CCATGGGCAGCAGCCATCATCATCATCATCACAGCAGCGGCCTGGAAGTTCTGTTCCAGGGACCAGCAGACAT
GACCGGTATGGAGGCGGATGCGGTGACCCTGGCGAGCCGTATCAGCCTGGGTTGGAACATTGGCAACACCCT
GGAGGCGATCGGTGGCGAAACCGCGTGGGGTAACCCGATGATCACCAACCAGCTGATTGAGCTGGTGAAGCA
AAGCGGCTTCGACGCGATCCGTATTCCGGCGAGCTGGGATCAGTACGCGAACCAAACCACCGCGGAAATTGAC
GATGCGTGGCTGGCGCGTGTTAAAGACGTGGTTCAGCTGTGCATCGACAACGATATGCCGGTGGTTCTGAACA
TTCACTGGGATGGTGGCTGGCTGGAGAACAACATCACCCCGGAAATGCAGGCGGCGAACAACGCGAAGCAAC
GTGCGTTCTGGCAGCAAATTGCGACCGAGATGCGTGACTTCGATGGTCGTCTGATGTTTGCGAGCGCGAACGA
GCCGAACGTGGACAGCGCGGAACAGATGGCGGTTCTGATGAGCTACCACCAAACCTTTGTGGATGCGGTTCGT
GAAACCGGTGGCCGTAACGCGTACCGTGTGCTGGTTATCCAGGGCCCGAACACCGACATTGAGCGTACCTATG
AACTGATGAGCAGCATGCCGAGCGATACCGTGGCGGGTCGTCTGATGGCGGAAGTGCACTTCTACACCCCGTA
TCAGTTTACCCTGATGGGCGAAGACGCGGGTTGGGGCAACCAATTCTTTTATTGGGGTGCGGGTAACCACAGC
GCGACCGATACCGCGCACAACCCGACCTGGGGTGAGGAAAGCTTCGTGGACGACCTGTTCGCGATGATGCAG
AGCCAATTCGTTGACAACGGTATCCCGGTGGTTCTGGGCGAGTTTAGCGCGATGCGTCGTACCGACCAGCTGA
GCGGCGATAACCTGCAGCTGCACCTGCAAAGCCGTGCGTACTGGCACAAATATGTTACCCAAAGCGCGATCGA
ACACGGTCTGATGCCGTTTTATTGGGACGCGGGTGGCCTGAACAACCACAGCAGCGGCATTTTCGACCGTAAC
AGCAACACCGTGTTTGATACCCAGACCATGAACGCGCTGAACGACGGTGTTGATGCGGCGCAATAA

Protein sequence:

MGSSHHHHHHSSGLEVLFQGPADMTGMEADAVTLASRISLGWNIGNTLEAIGGETAWGNPMITNQLIELVKQS
GFDAIRIPASWDQYANQTTAEIDDAWLARVKDVVQLCIDNDMPVVLNIHWDGGWLENNITPEMQAANNAKQR
AFWQQIATEMRDFDGRLMFASANEPNVDSAEQMAVLMSYHQTFVDAVRETGGRNAYRVLVIQGPNTDIERTYE
LMSSMPSDTVAGRLMAEVHFYTPYQFTLMGEDAGWGNQFFYWGAGNHSATDTAHNPTWGEESFVDDLFAM
MQSQFVDNGIPVVLGEFSAMRRTDQLSGDNLQLHLQSRAYWHKYVTQSAIEHGLMPFYWDAGGLNNHSSGIFD
RNSNTVFDTQTMNALNDGVDAAQ*

A3.1.3 *Tt*GH5_un

Codon optimised sequence:

CCATGGGCAGCAGCCATCATCATCATCATCACAGCAGCGGCCTGGAAGTTCTGTTCCAGGGACCAGCAGGTCC
GCTGTTCCAGTTTCATCATGCGCAAGGTACCGCGATTGCGGACGAGAAGAACCAGCCGGTGTTCCTGCGTGGT
GTTAGCTTTGGCAACCGTGTGTGGGTTAACGACCGTATTCCGGTGACCCACCACAGCGCGGAAGATTACAGCC
GTGTGCGTGCGATGGGTATGAACCTGGTTCGTTTCTACCTGAACTATCAGACCCTGGAGCTGGATGCGGCGCC
GTTTGAATACCAGGCGGACGGTTGGCAATGGCTGGATACCAACATTGCGTGGGCGCGTGCGGCGGGCGTGTA
TCTGATTCTGAACGTGCACGTTCCGCAAGGTGGCTTCCAGAGCCAGGGTAACGGCCGTAAGCTGTGGCAGGAC
GTTGATCTGCAAAAACGTTTCATCGCGATGTGGCGTGCGATTGCGGAGCGTTACCGTGACGAACCGGTGGTTT
TTGGTTATGATCTGCTGAACGAGCCGGGCGTTACCCGTGCGAAACAGCAATGGCAGCGTCTGGCGCAACGTAC
CGTGGACGCGATCCGTCTGGTTGATAAGAAACACCCGATCATTGTGGAGCGTGTTAACAGCATTAACCGTCGTT
GGGAAAACGACGCGGATATGAACTTCGTGACCGTTGACGGTAACAACATCATTTACACCTTCCACAGCTATGCG
CCGTACTTTTATAGCCACCAGGGCATCCCGTGGGATGCGAGCATGAAGAACCGTGATGGTGGCGTGTGGCCG
GATCCGAGCCGTCAGCACACCCGTGCGTTTCTGGCGCGTACCATTGACCAATACCTGGCGTGGGGCAAGGCGC
ACAACGTTCCGCTGTACTTCGGCGAGTGGGGCACCTATAAAGCGAACTTTGAAGCGGACCGTGGTGGCCTGAA
CTATCTGCGTGATATGCTGAGCGTGCTGGAGGAACGTAAACTGACCAACACCTTCCACGTTTACCACGAGGAA
AGCTTTGGCATCTATCGTGGTGACGGTCCGCTGGATCCGGCGAACGTGAACCAACCGGTTATTGATATGTTTAC
CACCTACCTGAAGCCGAAACACAACGAAAGCTAA

Protein sequence:

MGSSHHHHHHSSGLEVLFQGPAGPLFQFHHAQGTAIADEKNQPVFLRGVSFGNRVWVNDRIPVTHHSAEDYSRV
RAMGMNLVRFYLNYQTLELDAAPFEYQADGWQWLDTNIAWARAAGVYLILNVHVPQGGFQSQGNGRKLWQDV
DLQKRFIAMWRAIAERYRDEPVVFGYDLLNEPGVTRAKQQWQRLAQRTVDAIRLVDKKHPIIVERVNSINRRWEND
ADMNFVTVDGNNIIYTFHSYAPYFYSHQGIPWDASMKNRDGGVWPDPSRQHTRAFLARTIDQYLAWGKAHNVPL
YFGEWGTYKANFEADRGGLNYLRDMLSVLEERKLTNTFHVYHEESFGIYRGDGPLDPANVNQPVIDMFTTYLKPKH
NES*

A3.1.4 *Tt*GH12

Codon optimised sequence:
CCATGGGCAGCAGCCATCATCATCATCATCACAGCAGCGGCCTGGAAGTTCTGTTCCAGGGACCAGCAGCGAA
CACCGTTCTGAGCTGCGATGCGTTTGCGAGCCACCCGACCCCGATTGGTACCCTGGTGAACAACGTTTGGAAC
CAACAAGCGGCGGGCAAGGCGCGTTGGGAGCAGTGCCTGCTGCAACGTGATGGTCCGAACGGCCCGGAATA
CGGTTGGCGTTGGCGTTGGCCGGCGACCCCGCGTGTGGTTTTTGCGCAGCCGCAAATTACCCACGGCAACACC
CCGTGGACCAGCCACCCGACCCCGCAGAGCGGTTTTCCGATTCCGCTGACCGGTCTGGACCACCTGACCATTA
GCTATGCGGTGACCACCACCGGTGATGCGGATTTCAACCTGGCGACCACCTTTTGGCTGACCGACCTGGATCA
GGTGCCGCCGCAAGCGGACATCAACAGCATTCGTGCGGAGTTCATGGTTTGGAGCTATGCGAGCGATAACTTC
TTTAGCACCCCGGCGGGTCGTAAGCGTGCGACCGTTGAGATCAACGGCATTGAGTGGGAAGTGTGGGTTGAA
CGTCGTTGGCACGACACCAGCGGCACCAACGATAACCGTTGGATCTACATTGCGTTCCGTAGCACCAAAAACTA
CCTGGACATCACCTATAACGCGGCGGAGCTGGTGCACTATGCGCTGCAACGTGGTTTTCTGGACAGCGAATGG
AGCATCGCGGATATTGAGCTGGGCAACGAAGTTATGCGTGGTACCGGCGAAACCTGGGTGCACAAATTTCGTG
TTAGCACCCCGGCGCTGAACACCGATAACATTTAA

Protein sequence:
MGSSHHHHHHSSGLEVLFQGPAANTVLSCDAFASHPTPIGTLVNNVWNQQAAGKARWEQCLLQRDGPNGPEYG
WRWRWPATPRVVFAQPQITHGNTPWTSHPTPQSGFPIPLTGLDHLTISYAVTTTGDADFNLATTFWLTDLDQVPP
QADINSIRAEFMVWSYASDNFFSTPAGRKRATVEINGIEWEVWVERRWHDTSGTNDNRWIYIAFRSTKNYLDITYN
AAELVHYALQRGFLDSEWSIADIELGNEVMRGTGETWVHKFRVSTPALNTDNI*

A3.1.5 *Tt*GH8

Codon optimised sequence:
CCATGGGCAGCAGCCATCATCATCATCACAGCAGCGGCCTGGAAGTTCTGTTCCAGGGACCAGCAGCGGG
TGCGGTGGCGACCGGTGAATACCGTAACCTGTTCGCGGAGATTGGCAAGAGCGAAATCGACATTCAACGTAAA
ATCGATGAAGCGTTCCAGCACCTGTTTTATGGCGACGCGAAGGATGCGGCGGTTTACTATCAAGCGGGTGGCA
ACGAGAACGGTCCGCTGGCGTACGTGTATGACGTTAACAGCAACGATGTGCGTAGCGAGGGTATGAGCTACG
GCATGATGATTACCGTTCAAATGGACAAGAAAGCGGAATTTGATGCGATCTGGAACTGGGCGAAAACCTACAT
GTATCAAGACAGCCCGACCCACCCGGCGTTCGGTTATTTTGCGTGGAGCATGCGTCGTGATGGTGTGGCGAAC
GATGATATGCCGGCGCCGGATGGCGAGGAATACTTCGTTACCGCGCTGTATTTTGCGGCGGCGCGTTGGGGT
AACGGCGAGGGTATCTTCAACTACCAGCAAGAAGCGGATACCATTCTGAGCCGTATGCGTCACCGTCAAGTGA
TCACCGGTCCGACCAACCGTGGCGTTATGACCGCGACCAACCTGTTCCACCCGGAGGAAGCGCAAGTGCGTTT
TACCCCGGACATTAACAACGCGGACCACACCGATGCGAGCTACCACCTGCCGAGCTTTTATGAGATCTGGGCG
CGTGTTGCGCCGCAGGAAGATCGTGCGTTTTGGGCGAAGGCGGCGGATGTGAGCCGTGATTACTTTGCGAAG
GCGGCGCACCCGGTTACCGCGCTGACCCCGGACTATGGTAACTTCGATGGTACCCCGTGGGCGGCGAGCTGG
CGTCCGGAGAGCGTGGACTTTCGTTACGATGCGTGGCGTAGCGTTATGAACTGGAGCATGGACTATGCGTGGT
GGGGTAAAGATAGCGGTGCGCCGGCGCGTAGCGACAAACTGCTGGCGTTCTTTGAGACCCAAGAAGGTAAAA
TGAACCACCTGTACAGCCTGGACGGTAAGCCGCTGGGTGGCGGTCCGACCCTGGGTCTGATTAGCATGAACGC
GACCGCGGCGATGGCGGCGACCGACCCGCGTTGGCACAACTTCGTGGAAAAGCTGTGGCAGCAACAGCCGCC
GACCGGCCAGTACCGTTACTATGACGGCGTTCTGTATCTGATGGCGCTGCTGCACTGCGCGGGCGAGTACAAA
GCGTGGATCCCGGATGGCGAATAA

Protein sequence:
MGSSHHHHHHSSGLEVLFQGPAAGAVATGEYRNLFAEIGKSEIDIQRKIDEAFQHLFYGDAKDAAVYYQAGGNENG
PLAYVYDVNSNDVRSEGMSYGMMITVQMDKKAEFDAIWNWAKTYMYQDSPTHPAFGYFAWSMRRDGVANDD
MPAPDGEEYFVTALYFAAARWGNGEGIFNYQQEADTILSRMRHRQVITGPTNRGVMTATNLFHPEEAQVRFTPDI
NNADHTDASYHLPSFYEIWARVAPQEDRAFWAKAADVSRDYFAKAAHPVTALTPDYGNFDGTPWAASWRPESVD
FRYDAWRSVMNWSMDYAWWGKDSGAPARSDKLLAFFETQEGKMNHLYSLDGKPLGGGPTLGLISMNATAAMA
ATDPRWHNFVEKLWQQQPPTGQYRYYDGVLYLMALLHCAGEYKAWIPDGE*

A3.1.6 *Tt*AA10

Catalytic domain:
CATGGCTATATTGAATCGCCGCCGTCCCGTCAGCAGCATTGTGGTGCCGAACAGAAACCGGATAACCCGAGC
AGCGCAAAATGTGATGAAGCATTTGCTAATTATCGTGCGGCCGGCGGTCAGAATAGTCATTGGTACAACTTCA
TGTCCGTGGTTGCGCATCACGAAGGCCGCAAAGTCGTGAAAGGTACCGAACACGTTTGCGGCTTTGATGGTG
AAACGTGGAACCCGGCTCCGTATGACACCCCGGCGAATTGGCCGGTCACGAGTTTTAACTCCGGCCAGCAAA
CCTTCGTGTGGGATATTAGCTATGGTCCGCATTTTTCTGACACCGAAGAACTGGTCTTCTACATCACGAAACCG
GGCTTTAGCTTCGATCCGACCCGTGAACTGACGTGGGCCGATTTTGAAGACCAACCGTTCTGCGATGAAAGCA
TTGTGCCGGGTGACTTTTCAACCAATTCGGCGGTTGAAGCCGATATGGCAAACTCTCATATTAATGTTACGTGT
AACGTCCCGAGCCGTTCTGGCCGCCACGTGATCTTCGCAGAATGGGGTCGCAATGAACATACCTACGAACGCT
TCTTCTCCTGTGTGGACGTGGACTTTGGT

Protein sequence:
HGYIESPPSRQQHCGAEQKPDNPSSAKCDEAFANYRAAGGQNSHWYNFMSVVAHHEGRKVVKGTEHVCG
FDGETWNPAPYDTPANWPVTSFNSGQQTFVWDISYGPHFSDTEELVFYITKPGFSFDPTRELTWADFEDQPF
CDESIVPGDFSTNSAVEADMANSHINVTCNVPSRSGRHVIFAEWGRNEHTYERFFSCVDVDFG

A3.2 Vector Sequences

A3.2.1 pET11a

TGGCGAATGGGACGCGCCCTGTAGCGGCGCATTAAGCGCGGCGGGTGTGGTGGTTACGCGCAGCGTGACCGCTACACTTGCCAGCGCCCT
AGCGCCCGCTCCTTTCGCTTTCTTCCCTTCCTTTCTCGCCACGTTCGCCGGCTTTCCCCGTCAAGCTCTAAATCGGGGGCTCCCTTTAGGGTTC
CGATTTAGTGCTTTACGGCACCTCGACCCCAAAAAACTTGATTAGGGTGATGGTTCACGTAGTGGGCCATCGCCCTGATAGACGGTTTTTCGC
CCTTTGACGTTGGAGTCCACGTTCTTTAATAGTGGACTCTTGTTCCAAACTGGAACAACACTCAACCCTATCTCGGTCTATTCTTTTGATTTATA
AGGGATTTTGCCGATTTCGGCCTATTGGTTAAAAAATGAGCTGATTTAACAAAAATTTAACGCGAATTTTAACAAAATATTAACGTTTACAATT
TCAGGTGGCACTTTTCGGGGAAATGTGCGCGGAACCCCTATTTGTTTATTTTTCTAAATACATTCAAATATGTATCCGCTCATGAATTAATTCTT
AGAAAAACTCATCGAGCATCAAATGAAACTGCAATTTATTCATATCAGGATTATCAATACCATATTTTTGAAAAAGCCGTTTCTGTAATGAAGG
AGAAAACTCACCGAGGCAGTTCCATAGGATGGCAAGATCCTGGTATCGGTCTGCGATTCCGACTCGTCCAACATCAATACAACCTATTAATTT
CCCCTCGTCAAAAATAAGGTTATCAAGTGAGAAATCACCATGAGTGACGACTGAATCCGGTGAGAATGGCAAAAGTTTATGCATTTCTTTCCA
GACTTGTTCAACAGGCCAGCCATTACGCTCGTCATCAAAATCACTCGCATCAACCAAACCGTTATTCATTCGTGATTGCGCCTGAGCGAGACG
AAATACGCGATCGCTGTTAAAAGGACAATTACAAACAGGAATCGAATGCAACCGGCGCAGGAACACTGCCAGCGCATCAACAATATTTTCAC
CTGAATCAGGATATTCTTCTAATACCTGGAATGCTGTTTTCCCGGGGATCGCAGTGGTGAGTAACCATGCATCATCAGGAGTACGGATAAAA
TGCTTGATGGTCGGAAGAGGCATAAATTCCGTCAGCCAGTTTAGTCTGACCATCTCATCTGTAACATCATTGGCAACGCTACCTTTGCCATGT
TTCAGAAACAACTCTGGCGCATCGGGCTTCCCATACAATCGATAGATTGTCGCACCTGATTGCCCGACATTATCGCGAGCCCATTTATACCCA
TATAAATCAGCATCCATGTTGGAATTTAATCGCGGCCTAGAGCAAGACGTTTCCCGTTGAATATGGCTCATAACACCCCTTGTATTACTGTTTA
TGTAAGCAGACAGTTTTATTGTTCATGACCAAAATCCCTTAACGTGAGTTTTCGTTCCACTGAGCGTCAGACCCCGTAGAAAAGATCAAAGGA
TCTTCTTGAGATCCTTTTTTTCTGCGCGTAATCTGCTGCTTGCAAACAAAAAAACCACCGCTACCAGCGGTGGTTTGTTTGCCGGATCAAGAG
CTACCAACTCTTTTTCCGAAGGTAACTGGCTTCAGCAGAGCGCAGATACCAAATACTGTCCTTCTAGTGTAGCCGTAGTTAGGCCACCACTTC
AAGAACTCTGTAGCACCGCCTACATACCTCGCTCTGCTAATCCTGTTACCAGTGGCTGCTGCCAGTGGCGATAAGTCGTGTCTTACCGGGTTG
GACTCAAGACGATAGTTACCGGATAAGGCGCAGCGGTCGGGCTGAACGGGGGGTTCGTGCACACAGCCCAGCTTGGAGCGAACGACCTAC
ACCGAACTGAGATACCTACAGCGTGAGCTATGAGAAAGCGCCACGCTTCCCGAAGGGAGAAAGGCGGACAGGTATCCGGTAAGCGGCAGG
GTCGGAACAGGAGAGCGCACGAGGGAGCTTCCAGGGGGAAACGCCTGGTATCTTTATAGTCCTGTCGGGTTTCGCCACCTCTGACTTGAGC
GTCGATTTTTGTGATGCTCGTCAGGGGGGCGGAGCCTATGGAAAAACGCCAGCAACGCGGCCTTTTTACGGTTCCTGGCCTTTTGCTGGCCT
TTTGCTCACATGTTCTTTCCTGCGTTATCCCCTGATTCTGTGGATAACCGTATTACCGCCTTTGAGTGAGCTGATACCGCTCGCCGCAGCCGAA
CGACCGAGCGCAGCGAGTCAGTGAGCGAGGAAGCGGAAGAGCGCCTGATGCGGTATTTTCTCCTTACGCATCTGTGCGGTATTTCACACCG
CATATATGGTGCACTCTCAGTACAATCTGCTCTGATGCCGCATAGTTAAGCCAGTATACACTCCGCTATCGCTACGTGACTGGGTCATGGCTG
CGCCCCGACACCCGCCAACACCCGCTGACGCGCCCTGACGGGCTTGTCTGCTCCCGGCATCCGCTTACAGACAAGCTGTGACCGTCTCCGGG
AGCTGCATGTGTCAGAGGTTTTCACCGTCATCACCGAAACGCGCGAGGCAGCTGCGGTAAAGCTCATCAGCGTGGTCGTGAAGCGATTCAC
AGATGTCTGCCTGTTCATCCGCGTCCAGCTCGTTGAGTTTCTCCAGAAGCGTTAATGTCTGGCTTCTGATAAAGCGGGCCATGTTAAGGGCG
GTTTTTTCCTGTTTGGTCACTGATGCCTCCGTGTAAGGGGGATTTCTGTTCATGGGGGTAATGATACCGATGAAACGAGAGAGGATGCTCAC
GATACGGGTTACTGATGATGAACATGCCCGGTTACTGGAACGTTGTGAGGGTAAACAACTGGCGGTATGGATGCGGCGGGACCAGAGAAA
AATCACTCAGGGTCAATGCCAGCGCTTCGTTAATACAGATGTAGGTGTTCCACAGGGTAGCCAGCAGCATCCTGCGATGCAGATCCGGAACA
TAATGGTGCAGGGCGCTGACTTCCGCGTTTCCAGACTTTACGAAACACGGAAACCGAAGACCATTCATGTTGTTGCTCAGGTCGCAGACGTT
TTGCAGCAGCAGTCGCTTCACGTTCGCTCGCGTATCGGTGATTCATTCTGCTAACCAGTAAGGCAACCCCGCCAGCCTAGCCGGGTCCTCAAC
GACAGGAGCACGATCATGCGCACCCGTGGGGCCGCCATGCCGGCGATAATGGCCTGCTTCTCGCCGAAACGTTTGGTGGCGGGACCAGTG
ACGAAGGCTTGAGCGAGGGCGTGCAAGATTCCGAATACCGCAAGCGACAGGCCGATCATCGTCGCGCTCCAGCGAAAGCGGTCCTCGCCG
AAAATGACCCAGAGCGCTGCCGGCACCTGTCCTACGAGTTGCATGATAAAGAAGACAGTCATAAGTGCGGCGACGATAGTCATGCCCCGCG
CCCACCGGAAGGAGCTGACTGGGTTGAAGGCTCTCAAGGGCATCGGTCGAGATCCCGGTGCCTAATGAGTGAGCTAACTTACATTAATTGC
GTTGCGCTCACTGCCCGCTTTCCAGTCGGGAAACCTGTCGTGCCAGCTGCATTAATGAATCGGCCAACGCGCGGGGAGAGGCGGTTTGCGT
ATTGGGCGCCAGGGTGGTTTTTCTTTTCACCAGTGAGACGGGCAACAGCTGATTGCCCTTCACCGCCTGGCCCTGAGAGAGTTGCAGCAAGC
GGTCCACGCTGGTTTGCCCCAGCAGGCGAAAATCCTGTTTGATGGTGGTTAACGGCGGGATATAACATGAGCTGTCTTCGGTATCGTCGTAT
CCCACTACCGAGATATCCGCACCAACGCGCAGCCCGGACTCGGTAATGGCGCGCATTGCGCCCAGCGCCATCTGATCGTTGGCAACCAGCAT
CGCAGTGGGAACGATGCCCTCATTCAGCATTTGCATGGTTTGTTGAAAACCGGACATGGCACTCCAGTCGCCTTCCCGTTCCGCTATCGGCTG
AATTTGATTGCGAGTGAGATATTTATGCCAGCCAGCCAGACGCAGACGCGCCGAGACAGAACTTAATGGGCCCGCTAACAGCGCGATTTGC
TGGTGACCCAATGCGACCAGATGCTCCACGCCCAGTCGCGTACCGTCTTCATGGGAGAAAATAATACTGTTGATGGGTGTCTGGTCAGAGAC
ATCAAGAAATAACGCCGGAACATTAGTGCAGGCAGCTTCCACAGCAATGGCATCCTGGTCATCCAGCGGATAGTTAATGATCAGCCCACTGA
CGCGTTGCGCGAGAAGATTGTGCACCGCCGCTTTACAGGCTTCGACGCCGCTTCGTTCTACCATCGACACCACCACGCTGGCACCCAGTTGA
TCGGCGCGAGATTTAATCGCCGCGACAATTTGCGACGGCGCGTGCAGGGCCAGACTGGAGGTGGCAACGCCAATCAGCAACGACTGTTTGC
CCGCCAGTTGTTGTGCCACGCGGTTGGGAATGTAATTCAGCTCCGCCATCGCCGCTTCCACTTTTTCCCGCGTTTTCGCAGAAACGTGGCTGG
CCTGGTTCACCACGCGGGAAACGGTCTGATAAGAGACACCGGCATACTCTGCGACATCGTATAACGTTACTGGTTTCACATTCACCACCCTGA
ATTGACTCTCTTCCGGGCGCTATCATGCCATACCGCGAAAGGTTTTGCGCCATTCGATGGTGTCCGGGATCTCGACGCTCTCCCTTATGCGAC
TCCTGCATTAGGAAGCAGCCCAGTAGTAGGTTGAGGCCGTTGAGCACCGCCGCCGCAAGGAATGGTGCATGCAAGGAGATGGCGCCCAAC
AGTCCCCCGGCCACGGGGCCTGCCACCATACCCACGCCGAAACAAGCGCTCATGAGCCCGAAGTGGCGAGCCCGATCTTCCCCATCGGTGA
TGTCGGCGATATAGGCGCCAGCAACCGCACCTGTGGCGCCGGTGATGCCGGCCACGATGCGTCCGGCGTAGAGGATCGAGATCTCGATCCC
GCGAAATTAATACGACTCACTATAGGGGAATTGTGAGCGGATAACAATTCCCCTCTAGAAATAATTTTGTTTAACTTTAAGAAGGAGATATAC
CATGGGCAGCAGCCATCATCATCATCATCACAGCAGCGGCCTGGTGCCGCGCGGCAGCCATATGGCTAGCATGACTGGTGGACAGCAAATG
GGTCGCGGATCCGAATTCGAGCTCCGTCGACAAGCTTGCGGCCGCACTCGAGCACCACCACCACCACCACTGAGATCCGGCTGCTAACAAA
GCCCGAAAGGAAGCTGAGTTGGCTGCTGCCACCGCTGAGCAATAACTAGCATAACCCCTTGGGGCCTCTAAACGGGTCTTGAGGGGTTTTTT
GCTGAAAGGAGGAACTATATCCGGAT

A3.2.2 pNT-TrxT

CTAGAAATAATTTTGTTTAACTTTAAGAAGGAGATATACATATGCACCATCATCATCATCATTCTTCTGGTATGAGCGATAAAATTATTCACCT
GACTGACGACAGTTTTGACACGGATGTACTCAAAGCGGACGGGGCGATCCTCGTCGATTTCTGGGCAGAGTGGTGCGGTCCGTGCAAAATG
ATCGCCCCGATTCTGGATGAAATCGCTGACGAATATCAGGGCAAACTGACCGTTGCAAAACTGAACATCGATCAAAACCCTGGCACTGCGCC
GAAATATGGCATCCGTGGTATCCCGACTCTGCTGCTGTTCAAAAACGGTGAAGTGGCGGCAACCAAAGTGGGCGCACTGTCTAAAGGTCAG
TTGAAAGAGTTCCTCGACGCTAACCTGGCCGGTACCGAGAACTTGTACTTCCAATCCATGGAGACCGACGTCCACATATACCTGCCGTTCACT
ATTATTTAGTGAAATGAGATATTATGATATTTTCTGAATTGTGATTAAAAAGGCAACTTTATGCCCATGCAACAGAAACTATAAAAAATACAGA
GAATGAAAAGAAACAGATAGATTTTTTAGTTCTTTAGGCCCGTAGTCTGCAAATCCTTTTATGATTTTCTATCAAACAAAAGAGGAAAATAGA
CCAGTTGCAATCCAAACGAGAGTCTAATAGAATGAGGTCGAAAAGTAAATCGCGCGGGTTTGTTACTGATAAAGCAGGCAAGACCTAAAAT
GTGTAAAGGGCAAAGTGTATACTTTGGCGTCACCCCTTACATATTTTAGGTCTTTTTTTATTGTGCGTAACTAACTTGCCATCTTCAAACAGGA
GGGCTGGAAGAAGCAGACCGCTAACACAGTACATAAAAAAGGAGACATGAACGATGAACATCAAAAAGTTTGCAAAACAAGCAACAGTATT
AACCTTTACTACCGCACTGCTGGCAGGAGGCGCAACTCAAGCGTTTGCGAAAGAAACGAACCAAAAGCCATATAAGGAAACATACGGCATTT
CCCATATTACACGCCATGATATGCTGCAAATCCCTGAACAGCAAAAAAATGAAAAATATAAAGTTCCTGAGTTCGATTCGTCCACAATTAAAA
ATATCTCTTCTGCAAAAGGCCTGGACGTTTGGGACAGCTGGCCATTACAAAACACTGACGGCACTGTCGCAAACTATCACGGCTACCACATC
GTCTTTGCATTAGCCGGAGATCCTAAAAATGCGGATGACACATCGATTTACATGTTCTATCAAAAAGTCGGCGAAACTTCTATTGACAGCTGG
AAAAACGCTGGCCGCGTCTTTAAAGACAGCGACAAATTCGATGCAAATGATTCTATCCTAAAAGACCAAACACAAGAATGGTCAGGTTCAGC
CACATTTACATCTGACGGAAAAATCCGTTTATTCTACACTGATTTCTCCGGTAAACATTACGGCAAACAAACACTGACAACTGCACAAGTTAAC
GTATCAGCATCAGACAGCTCTTTGAACATCAACGGTGTAGAGGATTATAAATCAATCTTTGACGGTGACGGAAAAACGTATCAAAATGTACA
GCAGTTCATCGATGAAGGCAACTACAGCTCAGGCGACAACCATACGCTGAGAGATCCTCACTACGTAGAAGATAAAGGCCACAAATACTTAG
TATTTGAAGCAAACACTGGAACTGAAGATGGCTACCAAGGCGAAGAATCTTTATTTAACAAAGCATACTATGGCAAAAGCACATCATTCTTCC
GTCAAGAAAGTCAAAAACTTCTGCAAAGCGATAAAAAACGCACGGCTGAGTTAGCAAACGGCGCTCTCGGTATGATTGAGCTAAACGATGA
TTACACACTGAAAAAAGTGATGAAACCGCTGATTGCATCTAACACAGTAACAGATGAAATTGAACGCGCGAACGTCTTTAAAATGAACGGCA
AATGGTACCTGTTCACTGACTCCCGCGGATCAAAAATGACGATTGACGGCATTACGTCTAACGATATTTACATGCTTGGTTATGTTTCTAATTC
TTTAACTGGCCCATACAAGCCGCTGAACAAAACTGGCCTTGTGTTAAAAATGGATCTTGATCCTAACGATGTAACCTTTACTTACTCACACTTC
GCTGTACCTCAAGCGAAAGGAAACAATGTCGTGATTACAAGCTATATGACAAACAGAGGATTCTACGCAGACAAACAATCAACGTTTGCGCC
TAGCTTCCTGCTGAACATCAAAGGCAAGAAAACATCTGTTGTCAAAGACAGCATCCTTGAACAAGGACAATTAACAGTTAACAAATAAAAAC
GCAAAAGAAAATGCCGATATCCTATTGGCATTGACGGTCTCCAGTAAAGGTGGATACGGATCCGAATTCGAGCTCCGTCGACAAGCTTGCG
GCCGCACTCGAGCACCACCACCACCACCACTGAGATCCGGCTGCTAACAAAGCCCGAAAGGAAGCTGAGTTGGCTGCTGCCACCGCTGAGC
AATAACTAGCATAACCCCTTGGGGCCTCTAAACGGGTCTTGAGGGGTTTTTTGCTGAAAGGAGGAACTATATCCGGATTGGCGAATGGGAC
GCGCCCTGTAGCGGCGCATTAAGCGCGGCGGGTGTGGTGGTTACGCGCAGCGTGACCGCTACACTTGCCAGCGCCCTAGCGCCCGCTCCTT
TCGCTTTCTTCCCTTCCTTTCTCGCCACGTTCGCCGGCTTTCCCCGTCAAGCTCTAAATCGGGGGCTCCCTTTAGGGTTCCGATTTAGTGCTTTA
CGGCACCTCGACCCCAAAAAACTTGATTAGGGTGATGGTTCACGTAGTGGGCCATCGCCCTGATAGACGGTTTTTCGCCCTTTGACGTTGGA
GTCCACGTTCTTTAATAGTGGACTCTTGTTCCAAACTGGAACAACACTCAACCCTATCTCGGTCTATTCTTTTGATTTATAAGGGATTTTGCCGA
TTTCGGCCTATTGGTTAAAAAATGAGCTGATTTAACAAAAATTTAACGCGAATTTTAACAAAATATTAACGTTTACAATTTCAGGTGGCACTTT
TCGGGGAAATGTGCGCGGAACCCCTATTTGTTTATTTTTCTAAATACATTCAAATATGTATCCGCTCATGAATTAATTCTTAGAAAAACTCATC
GAGCATCAAATGAAACTGCAATTTATTCATATCAGGATTATCAATACCATATTTTTGAAAAAGCCGTTTCTGTAATGAAGGAGAAAACTCACC
GAGGCAGTTCCATAGGATGGCAAGATCCTGGTATCGGTCTGCGATTCCGACTCGTCCAACATCAATACAACCTATTAATTTCCCCTCGTCAAA
AATAAGGTTATCAAGTGAGAAATCACCATGAGTGACGACTGAATCCGGTGAGAATGGCAAAAGTTTATGCATTTCTTTCCAGACTTGTTCAAC
AGGCCAGCCATTACGCTCGTCATCAAAATCACTCGCATCAACCAAACCGTTATTCATTCGTGATTGCGCCTGAGCGAGACGAAATACGCGATC
GCTGTTAAAAGGACAATTACAAACAGGAATCGAATGCAACCGGCGCAGGAACACTGCCAGCGCATCAACAATATTTTCACCTGAATCAGGAT
ATTCTTCTAATACCTGGAATGCTGTTTTCCCGGGGATCGCAGTGGTGAGTAACCATGCATCATCAGGAGTACGGATAAAATGCTTGATGGTC
GGAAGAGGCATAAATTCCGTCAGCCAGTTTAGTCTGACCATCTCATCTGTAACATCATTGGCAACGCTACCTTTGCCATGTTTCAGAAACAAC
TCTGGCGCATCGGGCTTCCCATACAATCGATAGATTGTCGCACCTGATTGCCCGACATTATCGCGAGCCCATTTATACCCATATAAATCAGCA
TCCATGTTGGAATTTAATCGCGGCCTAGAGCAAGACGTTTCCCGTTGAATATGGCTCATAACACCCCTTGTATTACTGTTTATGTAAGCAGAC
AGTTTTATTGTTCATGACCAAAATCCCTTAACGTGAGTTTTCGTTCCACTGAGCGTCAGACCCCGTAGAAAAGATCAAAGGATCTTCTTGAGAT
CCTTTTTTTCTGCGCGTAATCTGCTGCTTGCAAACAAAAAAACCACCGCTACCAGCGGTGGTTTGTTTGCCGGATCAAGAGCTACCAACTCTTT
TTCCGAAGGTAACTGGCTTCAGCAGAGCGCAGATACCAAATACTGTCCTTCTAGTGTAGCCGTAGTTAGGCCACCACTTCAAGAACTCTGTAG
CACCGCCTACATACCTCGCTCTGCTAATCCTGTTACCAGTGGCTGCTGCCAGTGGCGATAAGTCGTGTCTTACCGGGTTGGACTCAAGACGAT
AGTTACCGGATAAGGCGCAGCGGTCGGGCTGAACGGGGGGTTCGTGCACACAGCCCAGCTTGGAGCGAACGACCTACACCGAACTGAGAT
ACCTACAGCGTGAGCTATGAGAAAGCGCCACGCTTCCCGAAGGGAGAAAGGCGGACAGGTATCCGGTAAGCGGCAGGGTCGGAACAGGA
GAGCGCACGAGGGAGCTTCCAGGGGGAAACGCCTGGTATCTTTATAGTCCTGTCGGGTTTCGCCACCTCTGACTTGAGCGTCGATTTTTGTG
ATGCTCGTCAGGGGGGCGGAGCCTATGGAAAAACGCCAGCAACGCGGCCTTTTTACGGTTCCTGGCCTTTTGCTGGCCTTTTGCTCACATGT
TCTTTCCTGCGTTATCCCCTGATTCTGTGGATAACCGTATTACCGCCTTTGAGTGAGCTGATACCGCTCGCCGCAGCCGAACGACCGAGCGCA
GCGAGTCAGTGAGCGAGGAAGCGGAAGAGCGCCTGATGCGGTATTTTCTCCTTACGCATCTGTGCGGTATTTCACACCGCATATATGGTGCA
CTCTCAGTACAATCTGCTCTGATGCCGCATAGTTAAGCCAGTATACACTCCGCTATCGCTACGTGACTGGGTCATGGCTGCGCCCCGACACCC
GCCAACACCCGCTGACGCGCCCTGACGGGCTTGTCTGCTCCCGGCATCCGCTTACAGACAAGCTGTGACCGTCTCCGGGAGCTGCATGTGTC
AGAGGTTTTCACCGTCATCACCGAAACGCGCGAGGCAGCTGCGGTAAAGCTCATCAGCGTGGTCGTGAAGCGATTCACAGATGTCTGCCTG
TTCATCCGCGTCCAGCTCGTTGAGTTTCTCCAGAAGCGTTAATGTCTGGCTTCTGATAAAGCGGGCCATGTTAAGGGCGGTTTTTTCCTGTTT
GGTCACTGATGCCTCCGTGTAAGGGGGATTTCTGTTCATGGGGGTAATGATACCGATGAAACGAGAGAGGATGCTCACGATACGGGTTACT
GATGATGAACATGCCCGGTTACTGGAACGTTGTGAGGGTAAACAACTGGCGGTATGGATGCGGCGGGACCAGAGAAAAATCACTCAGGGT
CAATGCCAGCGCTTCGTTAATACAGATGTAGGTGTTCCACAGGGTAGCCAGCAGCATCCTGCGATGCAGATCCGGAACATAATGGTGCAGG
GCGCTGACTTCCGCGTTTCCAGACTTTACGAAACACGGAAACCGAAGACCATTCATGTTGTTGCTCAGGTCGCAGACGTTTTGCAGCAGCAG
TCGCTTCACGTTCGCTCGCGTATCGGTGATTCATTCTGCTAACCAGTAAGGCAACCCCGCCAGCCTAGCCGGGTCCTCAACGACAGGAGCAC
GATCATGCGCACCCGTGGGGCCGCCATGCCGGCGATAATGGCCTGCTTCTCGCCGAAACGTTTGGTGGCGGGACCAGTGACGAAGGCTTGA
GCGAGGGCGTGCAAGATTCCGAATACCGCAAGCGACAGGCCGATCATCGTCGCGCTCCAGCGAAAGCGGTCCTCGCCGAAAATGACCCAG
AGCGCTGCCGGCACCTGTCCTACGAGTTGCATGATAAAGAAGACAGTCATAAGTGCGGCGACGATAGTCATGCCCCGCGCCCACCGGAAGG
AGCTGACTGGGTTGAAGGCTCTCAAGGGCATCGGTCGAGATCCCGGTGCCTAATGAGTGAGCTAACTTACATTAATTGCGTTGCGCTCACTG
CCCGCTTTCCAGTCGGGAAACCTGTCGTGCCAGCTGCATTAATGAATCGGCCAACGCGCGGGGAGAGGCGGTTTGCGTATTGGGCGCCAGG

GTGGTTTTTCTTTTCACCAGTGAGACGGGCAACAGCTGATTGCCCTTCACCGCCTGGCCCTGAGAGAGTTGCAGCAAGCGGTCCACGCTGGT
TTGCCCCAGCAGGCGAAAATCCTGTTTGATGGTGGTTAACGGCGGGATATAACATGAGCTGTCTTCGGTATCGTCGTATCCCACTACCGAGA
TATCCGCACCAACGCGCAGCCCGGACTCGGTAATGGCGCGCATTGCGCCCAGCGCCATCTGATCGTTGGCAACCAGCATCGCAGTGGGAAC
GATGCCCTCATTCAGCATTTGCATGGTTTGTTGAAAACCGGACATGGCACTCCAGTCGCCTTCCCGTTCCGCTATCGGCTGAATTTGATTGCG
AGTGAGATATTTATGCCAGCCAGCCAGACGCAGACGCGCCGAGACAGAACTTAATGGGCCCGCTAACAGCGCGATTTGCTGGTGACCCAAT
GCGACCAGATGCTCCACGCCCAGTCGCGTACCGTCTTCATGGGAGAAAATAATACTGTTGATGGGTGTCTGGTCAGAGACATCAAGAAATAA
CGCCGGAACATTAGTGCAGGCAGCTTCCACAGCAATGGCATCCTGGTCATCCAGCGGATAGTTAATGATCAGCCCACTGACGCGTTGCGCGA
GAAGATTGTGCACCGCCGCTTTACAGGCTTCGACGCCGCTTCGTTCTACCATCGACACCACCACGCTGGCACCCAGTTGATCGGCGCGAGAT
TTAATCGCCGCGACAATTTGCGACGGCGCGTGCAGGGCCAGACTGGAGGTGGCAACGCCAATCAGCAACGACTGTTTGCCCGCCAGTTGTT
GTGCCACGCGGTTGGGAATGTAATTCAGCTCCGCCATCGCCGCTTCCACTTTTTCCCGCGTTTTCGCAGAAACGTGGCTGGCCTGGTTCACCA
CGCGGGAAACGGTCTGATAAGAGACACCGGCATACTCTGCGACATCGTATAACGTTACTGGTTTCACATTCACCACCCTGAATTGACTCTCTT
CCGGGCGCTATCATGCCATACCGCGAAAGGTTTTGCGCCATTCGATGGTGTCCGGGATCTCGACGCTCTCCCTTATGCGACTCCTGCATTAG
GAAGCAGCCCAGTAGTAGGTTGAGGCCGTTGAGCACCGCCGCCGCAAGGAATGGTGCATGCAAGGAGATGGCGCCCAACAGTCCCCCGGC
CACGGGGCCTGCCACCATACCCACGCCGAAACAAGCGCTCATGAGCCCGAAGTGGCGAGCCCGATCTTCCCCATCGGTGATGTCGGCGATA
TAGGCGCCAGCAACCGCACCTGTGGCGCCGGTGATGCCGGCCACGATGCGTCCGGCGTAGAGGATCGAGATCTCGATCCCGCGAAATTAAT
ACGACTCACTATAGGGGAATTGTGAGCGGATAACAATTCCCCT

A3.2.3 pETFPP2-His-MBP

CTGCCACCATACCCACGCCGAAACAAGCGCTCATGAGCCCGAAGTGGCGAGCCCGATCTTCCCCATCGGTGATGTCGGCGATATAGGCGCC
AGCAACCGCACCTGTGGCGCCGGTGATGCCGGCCACGATGCGTCCGGCGTAGAGGATCGAGATCTCGATCCCGCGAAATTAATACGACTCA
CTATAGGGGAATTGTGAGCGGATAACAATTCCCCTCTAGAAATAATTTTGTTTAACTTTAAGAAGGAGATATACCATGGGCAGCAGCCATCAT
CATCATCATCACAGCAGCATGAAAATCGAAGAAGGTAAACTGGTAATCTGGATTAACGGCGATAAAGGCTATAACGGTCTCGCTGAAGTCG
GTAAGAAATTCGAGAAAGATACCGGAATTAAAGTCACCGTTGAGCATCCGGATAAACTGGAAGAGAAATTCCCACAGGTTGCGGCAACTGG
CGATGGCCCTGACATTATCTTCTGGGCACACGACCGCTTTGGTGGCTACGCTCAATCTGGCCTGTTGGCTGAAATCACCCCGGACAAAGCGT
TCCAGGACAAGCTGTATCCGTTTACCTGGGATGCCGTACGTTACAACGGCAAGCTGATTGCTTACCCGATCGCTGTTGAAGCGTTATCGCTG
ATTTATAACAAAGATCTGCTGCCGAACCCGCCAAAAACCTGGGAAGAGATCCCGGCGCTGGATAAAGAACTGAAAGCGAAAGGTAAGAGCG
CGCTGATGTTCAACCTGCAAGAACCGTACTTCACCTGGCCGCTGATTGCTGCTGACGGGGGTTATGCGTTCAAGTATGAAAACGGCAAGTAC
GACATTAAAGACGTGGGCGTGGATAACGCTGGCGCGAAAGCGGGTCTGACCTTCCTGGTTGACCTGATTAAAAACAAACACATGAATGCAG
ACACCGATTACTCCATCGCAGAAGCTGCCTTTAATAAAGGCGAAACAGCGATGACCATCAACGGCCCGTGGGCATGGTCCAACATCGACACC
AGCAAAGTGAATTATGGTGTAACGGTACTGCCGACCTTCAAGGGTCAACCATCCAAACCGTTCGTTGGCGTGCTGAGCGCAGGTATTAACGC
CGCCAGTCCGAACAAAGAGCTGGCGAAAGAGTTCCTCGAAAACTATCTGCTGACTGATGAAGGTCTGGAAGCGGTTAATAAAGACAAACCG
CTGGGTGCCGTAGCGCTGAAGTCTTACGAGGAAGAGTTGGCGAAAGATCCACGTATTGCCGCCACCATGGAAAACGCCCAGAAAGGTGAA
ATCATGCCGAACATCCCGCAGATGTCCGCTTTCTGGTATGCCGTGCGTACTGCGGTGATCAACGCCGCCAGCGGTCGTCAGACTGTCGATGA
AGCCCTGAAAGACGCGCAGACTCGTATCACCAAGGGCCTGGAAGTTCTGTTCCAGGGACCAGCAAGGCGCGCCTTCTCCTCACATATGGCTA
GCATGACTGGTGGACAGCAAATGGGTCGCGGATCCGAATTCGAGCTCCGTCGACAAGCTTGCGGCCGCACTCGAGCACCACCACCACCACC
ACTGAGATCCGGCTGCTAACAAAGCCCGAAAGGAAGCTGAGTTGGCTGCTGCCACCGCTGAGCAATAACTAGCATAACCCCTTGGGGCCTC
TAAACGGGTCTTGAGGGGTTTTTTGCTGAAAGGAGGAACTATATCCGGATTGGCGAATGGGACGCGCCCTGTAGCGGCGCATTAAGCGCGG
CGGGTGTGGTGGTTACGCGCAGCGTGACCGCTACACTTGCCAGCGCCCTAGCGCCCGCTCCTTTCGCTTTCTTCCCTTCCTTTCTCGCCACGT
TCGCCGGCTTTCCCCGTCAAGCTCTAAATCGGGGGCTCCCTTTAGGGTTCCGATTTAGTGCTTTACGGCACCTCGACCCCAAAAAACTTGATT
AGGGTGATGGTTCACGTAGTGGGCCATCGCCCTGATAGACGGTTTTTCGCCCTTTGACGTTGGAGTCCACGTTCTTTAATAGTGGACTCTTGT
TCCAAACTGGAACAACACTCAACCCTATCTCGGTCTATTCTTTTGATTTATAAGGGATTTTGCCGATTTCGGCCTATTGGTTAAAAAATGAGCT
GATTTAACAAAAATTTAACGCGAATTTTAACAAAATATTAACGTTTACAATTTCAGGTGGCACTTTTCGGGGAAATGTGCGCGGAACCCCTAT
TTGTTTATTTTTCTAAATACATTCAAATATGTATCCGCTCATGAATTAATTCTTAGAAAAACTCATCGAGCATCAAATGAAACTGCAATTTATTC
ATATCAGGATTATCAATACCATATTTTTGAAAAAGCCGTTTCTGTAATGAAGGAGAAAACTCACCGAGGCAGTTCCATAGGATGGCAAGATCC
TGGTATCGGTCTGCGATTCCGACTCGTCCAACATCAATACAACCTATTAATTTCCCCTCGTCAAAAATAAGGTTATCAAGTGAGAAATCACCAT
GAGTGACGACTGAATCCGGTGAGAATGGCAAAAGTTTATGCATTTCTTTCCAGACTTGTTCAACAGGCCAGCCATTACGCTCGTCATCAAAAT
CACTCGCATCAACCAAACCGTTATTCATTCGTGATTGCGCCTGAGCGAGACGAAATACGCGATCGCTGTTAAAAGGACAATTACAAACAGGA
ATCGAATGCAACCGGCGCAGGAACACTGCCAGCGCATCAACAATATTTTCACCTGAATCAGGATATTCTTCTAATACCTGGAATGCTGTTTTC
CCGGGGATCGCAGTGGTGAGTAACCATGCATCATCAGGAGTACGGATAAAATGCTTGATGGTCGGAAGAGGCATAAATTCCGTCAGCCAGT
TTAGTCTGACCATCTCATCTGTAACATCATTGGCAACGCTACCTTTGCCATGTTTCAGAAACAACTCTGGCGCATCGGGCTTCCCATACAATCG
ATAGATTGTCGCACCTGATTGCCCGACATTATCGCGAGCCCATTTATACCCATATAAATCAGCATCCATGTTGGAATTTAATCGCGGCCTAGA
GCAAGACGTTTCCCGTTGAATATGGCTCATAACACCCCTTGTATTACTGTTTATGTAAGCAGACAGTTTTATTGTTCATGACCAAAATCCCTTA
ACGTGAGTTTTCGTTCCACTGAGCGTCAGACCCCGTAGAAAAGATCAAAGGATCTTCTTGAGATCCTTTTTTTCTGCGCGTAATCTGCTGCTTG
CAAACAAAAAAACCACCGCTACCAGCGGTGGTTTGTTTGCCGGATCAAGAGCTACCAACTCTTTTTCCGAAGGTAACTGGCTTCAGCAGAGC
GCAGATACCAAATACTGTCCTTCTAGTGTAGCCGTAGTTAGGCCACCACTTCAAGAACTCTGTAGCACCGCCTACATACCTCGCTCTGCTAATC
CTGTTACCAGTGGCTGCTGCCAGTGGCGATAAGTCGTGTCTTACCGGGTTGGACTCAAGACGATAGTTACCGGATAAGGCGCAGCGGTCGG
GCTGAACGGGGGGTTCGTGCACACAGCCCAGCTTGGAGCGAACGACCTACACCGAACTGAGATACCTACAGCGTGAGCTATGAGAAAGCG
CCACGCTTCCCGAAGGGAGAAAGGCGGACAGGTATCCGGTAAGCGGCAGGGTCGGAACAGGAGAGCGCACGAGGGAGCTTCCAGGGGG
AAACGCCTGGTATCTTTATAGTCCTGTCGGGTTTCGCCACCTCTGACTTGAGCGTCGATTTTTGTGATGCTCGTCAGGGGGGCGGAGCCTAT
GGAAAAACGCCAGCAACGCGGCCTTTTTACGGTTCCTGGCCTTTTGCTGGCCTTTTGCTCACATGTTCTTTCCTGCGTTATCCCCTGATTCTGT
GGATAACCGTATTACCGCCTTTGAGTGAGCTGATACCGCTCGCCGCAGCCGAACGACCGAGCGCAGCGAGTCAGTGAGCGAGGAAGCGGA
AGAGCGCCTGATGCGGTATTTTCTCCTTACGCATCTGTGCGGTATTTCACACCGCATATATGGTGCACTCTCAGTACAATCTGCTCTGATGCC
GCATAGTTAAGCCAGTATACACTCCGCTATCGCTACGTGACTGGGTCATGGCTGCGCCCCGACACCCGCCAACACCCGCTGACGCGCCCTGA
CGGGCTTGTCTGCTCCCGGCATCCGCTTACAGACAAGCTGTGACCGTCTCCGGGAGCTGCATGTGTCAGAGGTTTTCACCGTCATCACCGAA
ACGCGCGAGGCAGCTGCGGTAAAGCTCATCAGCGTGGTCGTGAAGCGATTCACAGATGTCTGCCTGTTCATCCGCGTCCAGCTCGTTGAGTT
TCTCCAGAAGCGTTAATGTCTGGCTTCTGATAAAGCGGGCCATGTTAAGGGCGGTTTTTTCCTGTTTGGTCACTGATGCCTCCGTGTAAGGG
GGATTTCTGTTCATGGGGGTAATGATACCGATGAAACGAGAGAGGATGCTCACGATACGGGTTACTGATGATGAACATGCCCGGTTACTGG
AACGTTGTGAGGGTAAACAACTGGCGGTATGGATGCGGCGGGACCAGAGAAAAATCACTCAGGGTCAATGCCAGCGCTTCGTTAATACAGA

TGTAGGTGTTCCACAGGGTAGCCAGCAGCATCCTGCGATGCAGATCCGGAACATAATGGTGCAGGGCGCTGACTTCCGCGTTTCCAGACTTT
ACGAAACACGGAAACCGAAGACCATTCATGTTGTTGCTCAGGTCGCAGACGTTTTGCAGCAGCAGTCGCTTCACGTTCGCTCGCGTATCGGT
GATTCATTCTGCTAACCAGTAAGGCAACCCCGCCAGCCTAGCCGGGTCCTCAACGACAGGAGCACGATCATGCGCACCCGTGGGGCCGCCA
TGCCGGCGATAATGGCCTGCTTCTCGCCGAAACGTTTGGTGGCGGGACCAGTGACGAAGGCTTGAGCGAGGGCGTGCAAGATTCCGAATA
CCGCAAGCGACAGGCCGATCATCGTCGCGCTCCAGCGAAAGCGGTCCTCGCCGAAAATGACCCAGAGCGCTGCCGGCACCTGTCCTACGAG
TTGCATGATAAAGAAGACAGTCATAAGTGCGGCGACGATAGTCATGCCCCGCGCCCACCGGAAGGAGCTGACTGGGTTGAAGGCTCTCAAG
GGCATCGGTCGAGATCCCGGTGCCTAATGAGTGAGCTAACTTACATTAATTGCGTTGCGCTCACTGCCCGCTTTCCAGTCGGGAAACCTGTC
GTGCCAGCTGCATTAATGAATCGGCCAACGCGCGGGGAGAGGCGGTTTGCGTATTGGGCGCCAGGGTGGTTTTTCTTTTCACCAGTGAGAC
GGGCAACAGCTGATTGCCCTTCACCGCCTGGCCCTGAGAGAGTTGCAGCAAGCGGTCCACGCTGGTTTGCCCCAGCAGGCGAAAATCCTGT
TTGATGGTGGTTAACGGCGGGATATAACATGAGCTGTCTTCGGTATCGTCGTATCCCACTACCGAGATATCCGCACCAACGCGCAGCCCGGA
CTCGGTAATGGCGCGCATTGCGCCCAGCGCCATCTGATCGTTGGCAACCAGCATCGCAGTGGGAACGATGCCCTCATTCAGCATTTGCATGG
TTTGTTGAAAACCGGACATGGCACTCCAGTCGCCTTCCCGTTCCGCTATCGGCTGAATTTGATTGCGAGTGAGATATTTATGCCAGCCAGCCA
GACGCAGACGCGCCGAGACAGAACTTAATGGGCCCGCTAACAGCGCGATTTGCTGGTGACCCAATGCGACCAGATGCTCCACGCCCAGTCG
CGTACCGTCTTCATGGGAGAAAATAATACTGTTGATGGGTGTCTGGTCAGAGACATCAAGAAATAACGCCGGAACATTAGTGCAGGCAGCTT
CCACAGCAATGGCATCCTGGTCATCCAGCGGATAGTTAATGATCAGCCCACTGACGCGTTGCGCGAGAAGATTGTGCACCGCCGCTTTACAG
GCTTCGACGCCGCTTCGTTCTACCATCGACACCACCAGCGCTGGCACCCAGTTGATCGGCGCGAGATTTAATCGCCGCGACAATTTGCGACGG
CGCGTGCAGGGCCAGACTGGAGGTGGCAACGCCAATCAGCAACGACTGTTTGCCCGCCAGTTGTTGTGCCACGCGGTTGGGAATGTAATTC
AGCTCCGCCATCGCCGCTTCCACTTTTTCCCGCGTTTTCGCAGAAACGTGGCTGGCCTGGTTCACCACGCGGGAAACGGTCTGATAAGAGAC
ACCGGCATACTCTGCGACATCGTATAACGTTACTGGTTTCACATTCACCACCCTGAATTGACTCTCTTCCGGGCGCTATCATGCCATACCGCGA
AAGGTTTTGCGCCATTCGATGGTGTCCGGGATCTCGACGCTCTCCCTTATGCGACTCCTGCATTAGGAAGCAGCCCAGTAGTAGGTTGAGGC
CGTTGAGCACCGCCGCCGCAAGGAATGGTGCATGCAAGGAGATGGCGCCCAACAGTCCCCCGGCCACGGGGC

A3.2.4 pET28 YSBL–His-3C-LIC[313]

TGGCGAATGGGACGCGCCCTGTAGCGGCGCATTAAGCGCGGCGGGTGTGGTGGTTACGCGCAGCGTGACCGCTACACTTGCCAGCGCCCT
AGCGCCCGCTCCTTTCGCTTTCTTCCCTTCCTTTCTCGCCACGTTCGCCGGCTTTCCCCGTCAAGCTCTAAATCGGGGGCTCCCTTTAGGGTTC
CGATTTAGTGCTTTACGGCACCTCGACCCCAAAAAACTTGATTAGGGTGATGGTTCACGTAGTGGGCCATCGCCCTGATAGACGGTTTTTCGC
CCTTTGACGTTGGAGTCCACGTTCTTTAATAGTGGACTCTTGTTCCAAACTGGAACAACACTCAACCCTATCTCGGTCTATTCTTTTGATTTATA
AGGGATTTTGCCGATTTCGGCCTATTGGTTAAAAAATGAGCTGATTTAACAAAAATTTAACGCGAATTTTAACAAAATATTAACGTTTACAATT
TCAGGTGGCACTTTTCGGGGAAATGTGCGCGGAACCCCTATTTGTTTATTTTTCTAAATACATTCAAATATGTATCCGCTCATGAATTAATTCTT
AGAAAAACTCATCGAGCATCAAATGAAACTGCAATTTATTCATATCAGGATTATCAATACCATATTTTTGAAAAAGCCGTTTCTGTAATGAAGG
AGAAAACTCACCGAGGCAGTTCCATAGGATGGCAAGATCCTGGTATCGGTCTGCGATTCCGACTCGTCCAACATCAATACAACCTATTAATTT
CCCCTCGTCAAAAATAAGGTTATCAAGTGAGAAATCACCATGAGTGACGACTGAATCCGGTGAGAATGGCAAAAGTTTATGCATTTCTTTCCA
GACTTGTTCAACAGGCCAGCCATTACGCTCGTCATCAAAATCACTCGCATCAACCAAACCGTTATTCATTCGTGATTGCGCCTGAGCGAGACG
AAATACGCGATCGCTGTTAAAAGGACAATTACAAACAGGAATCGAATGCAACCGGCGCAGGAACACTGCCAGCGCATCAACAATATTTTCAC
CTGAATCAGGATATTCTTCTAATACCTGGAATGCTGTTTTCCCGGGGATCGCAGTGGTGAGTAACCATGCATCATCAGGAGTACGGATAAAA
TGCTTGATGGTCGGAAGAGGCATAAATTCCGTCAGCCAGTTTAGTCTGACCATCTCATCTGTAACATCATTGGCAACGCTACCTTTGCCATGT
TTCAGAAACAACTCTGGCGCATCGGGCTTCCCATACAATCGATAGATTGTCGCACCTGATTGCCCGACATTATCGCGAGCCCATTTATACCCA
TATAAATCAGCATCCATGTTGGAATTTAATCGCGGCCTAGAGCAAGACGTTTCCCGTTGAATATGGCTCATAACACCCCTTGTATTACTGTTTA
TGTAAGCAGACAGTTTTATTGTTCATGACCAAAATCCCTTAACGTGAGTTTTCGTTCCACTGAGCGTCAGACCCCGTAGAAAAGATCAAAGGA
TCTTCTTGAGATCCTTTTTTTCTGCGCGTAATCTGCTGCTTGCAAACAAAAAAACCACCGCTACCAGCGGTGGTTTGTTTGCCGGATCAAGAG
CTACCAACTCTTTTTCCGAAGGTAACTGGCTTCAGCAGAGCGCAGATACCAAATACTGTCCTTCTAGTGTAGCCGTAGTTAGGCCACCACTTC
AAGAACTCTGTAGCACCGCCTACATACCTCGCTCTGCTAATCCTGTTACCAGTGGCTGCTGCCAGTGGCGATAAGTCGTGTCTTACCGGGTTG
GACTCAAGACGATAGTTACCGGATAAGGCGCAGCGGTCGGGCTGAACGGGGGGTTCGTGCACACAGCCCAGCTTGGAGCGAACGACCTAC
ACCGAACTGAGATACCTACAGCGTGAGCTATGAGAAAGCGCCACGCTTCCCGAAGGGAGAAAGGCGGACAGGTATCCGGTAAGCGGCAGG
GTCGGAACAGGAGAGCGCACGAGGGAGCTTCCAGGGGGAAACGCCTGGTATCTTTATAGTCCTGTCGGGTTTCGCCACCTCTGACTTGAGC
GTCGATTTTTGTGATGCTCGTCAGGGGGGCGGAGCCTATGGAAAAACGCCAGCAACGCGGCCTTTTTACGGTTCCTGGCCTTTTGCTGGCCT
TTTGCTCACATGTTCTTTCCTGCGTTATCCCCTGATTCTGTGGATAACCGTATTACCGCCTTTGAGTGAGCTGATACCGCTCGCCGCAGCCGAA
CGACCGAGCGCAGCGAGTCAGTGAGCGAGGAAGCGGAAGAGCGCCTGATGCGGTATTTTCTCCTTACGCATCTGTGCGGTATTTCACACCG
CATATATGGTGCACTCTCAGTACAATCTGCTCTGATGCCGCATAGTTAAGCCAGTATACACTCCGCTATCGCTACGTGACTGGGTCATGGCTG
CGCCCCGACACCCGCCAACACCCGCTGACGCGCCCTGACGGGCTTGTCTGCTCCCGGCATCCGCTTACAGACAAGCTGTGACCGTCTCCGGG
AGCTGCATGTGTCAGAGGTTTTCACCGTCATCACCGAAACGCGCGAGGCAGCTGCGGTAAAGCTCATCAGCGTGGTCGTGAAGCGATTCAC
AGATGTCTGCCTGTTCATCCGCGTCCAGCTCGTTGAGTTTCTCCAGAAGCGTTAATGTCTGGCTTCTGATAAAGCGGGCCATGTTAAGGGCG
GTTTTTTCCTGTTTGGTCACTGATGCCTCCGTGTAAGGGGGATTTCTGTTCATGGGGGTAATGATACCGATGAAACGAGAGAGGATGCTCAC
GATACGGGTTACTGATGATGAACATGCCCGGTTACTGGAACGTTGTGAGGGTAAACAACTGGCGGTATGGATGCGGCGGGACCAGAGAAA
AATCACTCAGGGTCAATGCCAGCGCTTCGTTAATACAGATGTAGGTGTTCCACAGGGTAGCCAGCAGCATCCTGCGATGCAGATCCGGAACA
TAATGGTGCAGGGCGCTGACTTCCGCGTTTCCAGACTTTACGAAACACGGAAACCGAAGACCATTCATGTTGTTGCTCAGGTCGCAGACGTT
TTGCAGCAGCAGTCGCTTCACGTTCGCTCGCGTATCGGTGATTCATTCTGCTAACCAGTAAGGCAACCCCGCCAGCCTAGCCGGGTCCTCAAC
GACAGGAGCACGATCATGCGCACCCGTGGGGCCGCCATGCCGGCGATAATGGCCTGCTTCTCGCCGAAACGTTTGGTGGCGGGACCAGTG
ACGAAGGCTTGAGCGAGGGCGTGCAAGATTCCGAATACCGCAAGCGACAGGCCGATCATCGTCGCGCTCCAGCGAAAGCGGTCCTCGCCG
AAAATGACCCAGAGCGCTGCCGGCACCTGTCCTACGAGTTGCATGATAAAGAAGACAGTCATAAGTGCGGCGACGATAGTCATGCCCCGCG
CCCACCGGAAGGAGCTGACTGGGTTGAAGGCTCTCAAGGGCATCGGTCGAGATCCCGGTGCCTAATGAGTGAGCTAACTTACATTAATTGC
GTTGCGCTCACTGCCCGCTTTCCAGTCGGGAAACCTGTCGTGCCAGCTGCATTAATGAATCGGCCAACGCGCGGGGAGAGGCGGTTTGCGT
ATTGGGCGCCAGGGTGGTTTTTCTTTTCACCAGTGAGACGGGCAACAGCTGATTGCCCTTCACCGCCTGGCCCTGAGAGAGTTGCAGCAAGC
GGTCCACGCTGGTTTGCCCCAGCAGGCGAAAATCCTGTTTGATGGTGGTTAACGGCGGGATATAACATGAGCTGTCTTCGGTATCGTCGTAT
CCCACTACCGAGATATCCGCACCAACGCGCAGCCCGGACTCGGTAATGGCGCGCATTGCGCCCAGCGCCATCTGATCGTTGGCAACCAGCAT
CGCAGTGGGAACGATGCCCTCATTCAGCATTTGCATGGTTTGTTGAAAACCGGACATGGCACTCCAGTCGCCTTCCCGTTCCGCTATCGGCTG
AATTTGATTGCGAGTGAGATATTTATGCCAGCCAGCCAGACGCAGACGCGCCGAGACAGAACTTAATGGGCCCGCTAACAGCGCGATTTGC
TGGTGACCCAATGCGACCAGATGCTCCACGCCCAGTCGCGTACCGTCTTCATGGGAGAAAATAATACTGTTGATGGGTGTCTGGTCAGAGAC

ATCAAGAAATAACGCCGGAACATTAGTGCAGGCAGCTTCCACAGCAATGGCATCCTGGTCATCCAGCGGATAGTTAATGATCAGCCCACTGA
CGCGTTGCGCGAGAAGATTGTGCACCGCCGCTTTACAGGCTTCGACGCCGCTTCGTTCTACCATCGACACCACCACGCTGGCACCCAGTTGA
TCGGCGCGAGATTTAATCGCCGCGACAATTTGCGACGGCGCGTGCAGGGCCAGACTGGAGGTGGCAACGCCAATCAGCAACGACTGTTTGC
CCGCCAGTTGTTGTGCCACGCGGTTGGGAATGTAATTCAGCTCCGCCATCGCCGCTTCCACTTTTTCCCGCGTTTTCGCAGAAACGTGGCTGG
CCTGGTTCACCACGCGGGAAACGGTCTGATAAGAGACACCGGCATACTCTGCGACATCGTATAACGTTACTGGTTTCACATTCACCACCCTGA
ATTGACTCTCTTCCGGGCGCTATCATGCCATACCGCGAAAGGTTTTGCGCCATTCGATGGTGTCCGGGATCTCGACGCTCTCCCTTATGCGAC
TCCTGCATTAGGAAGCAGCCCAGTAGTAGGTTGAGGCCGTTGAGCACCGCCGCCGCAAGGAATGGTGCATGCAAGGAGATGGCGCCCAAC
AGTCCCCCGGCCACGGGGCCTGCCACCATACCCACGCCGAAACAAGCGCTCATGAGCCCGAAGTGGCGAGCCCGATCTTCCCCATCGGTGA
TGTCGGCGATATAGGCGCCAGCAACCGCACCTGTGGCGCCGGTGATGCCGGCCACGATGCGTCCGGCGTAGAGGATCGAGATCTCGATCCC
GCGAAATTAATACGACTCACTATAGGGGAATTGTGAGCGGATAACAATTCCCCTCTAGAAATAATTTTGTTTAACTTTAAGAAGGAGATATAC
CATGGGCAGCAGCCATCATCATCATCATCACAGCAGCGGCCTGGAAGTTCTGTTCCAGGGACCAGCAAGGCGCGCCTTCTCCTCACATATGG
CTAGCATGACTGGTGGACAGCAAATGGGTCGCGGATCCGAATTCGAGCTCCGTCGACAAGCTTGCGGCCGCACTCGAGCACCACCACCACC
ACCACTGAGATCCGGCTGCTAACAAAGCCCGAAAGGAAGCTGAGTTGGCTGCTGCCACCGCTGAGCAATAACTAGCATAACCCCTTGGGGC
CTCTAAACGGGTCTTGAGGGGTTTTTTGCTGAAAGGAGGAACTATATCCGGAT

A3.2.5 pSF 1477 BdSUMO-MBP^{122}

CTCGAGAAATCATAAAAAATTTATTTGCTTTGTGAGCGGATAACAATTATAATAGATTCAATTGTGAGCGGATAACAATTTCACACAGAATTCA
TTAAAGAGGAGAAATTAACCATGAGCAAGCATCACCATCATTCAGGCCATCACCATACCGGACACCACCATCATTCAGGCAGTCATCACCATT
CCGGATCTGCTGCGGGTGGCGAAGAAGATAAGAAACCGGCAGGTGGCGAAGGTGGCGGTGCCCATATCAACCTGAAAGTGAAAGGTCAA
GACGGCAACGAAGTCTTTTTCCGCATCAAACGTTCTACCCAGCTGAAAAAGCTGATGAACGCATACTGTGACCGTCAGTCTGTAGACATGAC
CGCAATTGCTTTCCTGTTTGATGGTCGTCGCCTGCGTGCGGAACAGACCCCGGATGAACTGGAGATGGAAGATGGCGACGAAATCGACGCA
ATGCTGCACCAGACTGGTGGCGCCGGTACCAAAACTGAAGAAGGTAAACTGGTAATCTGGATTAACGGCGATAAAGGCTATAACGGTCTCG
CTGAAGTCGGTAAGAAATTCGAGAAAGATACCGGAATTAAAGTCACCGTTGAGCATCCGGATAAACTGGAAGAGAAATTCCCACAGGTTGC
GGCAACTGGCGATGGCCCTGACATTATCTTCTGGGCACACGACCGCTTTGGTGGCTACGCTCAATCTGGCCTGTTGGCTGAAATCACCCCGG
ACAAAGCGTTCCAGGACAAGCTGTATCCGTTTACCTGGGATGCCGTACGTTACAACGGCAAGCTGATTGCTTACCCGATCGCTGTTGAAGCG
TTATCGCTGATTTATAACAAAGATCTGCTGCCGAACCCGCCAAAAACCTGGGAAGAGATCCCGGCGCTGGATAAAGAACTGAAAGCGAAAG
GTAAGAGCGCGCTGATGTTCAACCTGCAAGAACCGTACTTCACCTGGCCGCTGATTGCTGCTGACGGGGGTTATGCGTTCAAGTATGAAAAC
GGCAAGTACGACATTAAAGACGTGGGCGTGGATAACGCTGGCGCGAAAGCGGGTCTGACCTTCCTGGTTGACCTGATTAAAAACAAACACA
TGAATGCAGACACCGATTACTCCATCGCAGAAGCTGCCTTTAATAAAGGCGAAACAGCGATGACCATCAACGGCCCGTGGGCATGGTCCAAC
ATCGACACCAGCAAAGTGAATTATGGTGTAACGGTACTGCCGACCTTCAAGGGTCAACCATCCAAACCGTTCGTTGGCGTGCTGAGCGCAG
GTATTAACGCCGCCAGTCCGAACAAAGAGCTGGCAAAAGAGTTCCTCGAAAACTATCTGCTGACTGATGAAGGTCTGGAAGCGGTTAATAA
AGACAAACCGCTGGGTGCCGTAGCGCTGAAGTCTTACGAGGAAGAGTTGGCGAAAGATCCACGTATTGCCGCCACCATGGAAAACGCCCAG
AAAGGTGAAATCATGCCGAACATCCCGCAGATGTCCGCTTTCTGGTATGCCGTGCGTACTGCGGTGATCAACGCCGCCAGCGGTCGTCAGAC
TGTCGATGAAGCCCTGAAAGACGCGCAGACTAATGGCACCGGTTGTTAAGGATCTCATCACCATCACCATCACTAAGCTTAATTAGCTGAGC
TTGGACTCCTGTTGATAGATCCAGTAATGACCTCAGAACTCCATCTGGATTTGTTCAGAACGCTCGGTTGCCGCCGGGCGTTTTTTATTGGTG
AGAATCCAAGCTAGCCATGAAAATAAACTGTCTGCTTACATAAACAGTAATACAAGGGGTGTTATGAGCCATATTCAACGGGAAACGTCTTG
CTCTAGGCCGCGATTAAATTCCAACATGGATGCTGATTTATATGGGTATAAATGGGCTCGCGATAATGTCGGGCAATCAGGTGCGACAATCT
ATCGATTGTATGGGAAGCCCGATGCGCCAGAGTTGTTTCTGAAACATGGCAAAGGTAGCGTTGCCAATGATGTTACAGATGAGATGGTCAG
ACTAAACTGGCTGACGGAATTTATGCCTCTTCCGACCATCAAGCATTTTATCCGTACTCCTGATGATGCTTGGTTACTCACGACTGCGATCCCC
GGCAAAACAGCATTCCAGGTATTAGAAGAATATCCTGATTCAGGTGAAAATATTGTTGATGCGCTGGCAGTGTTCCTGCGCCGGTTGCATTC
GATTCCTGTTTGTAATTGTCCTTTTAACAGCGATCGCGTATTTCGTCTCGCTCAGGCGCAATCACGAATGAATAACGGTTTGGTTGATGCGAG
TGATTTTGATGACGAGCGTAATGGCTGGCCTGTTGAACAAGTCTGGAAAGAAATGCACAAACTTTTGCCATTCTCACCGGATTCAGTCGTCAC
TCATGGTGATTTCTCACTTGATAACCTTATTTTTGACGAGGGGAAATTAATAGGTTGTATTGATGTTGGACGAGTCGGAATCGCAGACCGATA
CCAGGATCTTGCCATCCTATGGAACTGCCTCGGTGAGTTTTCTCCTTCATTACAGAAACGGCTTTTTCAAAAATATGGTATTGATAATCCTGAT
ATGAATAAATTGCAGTTTCATTTGATGCTCGATGAGTTTTTCTAAGAATTAATTCATGGGCAAATATTATACGCAAGGCGACAAGGTGCTGAT
GCCGCTGGCGATTCAGGTTCATCATGCCGTTTGTGATGGCTTCCATGTCGGCAGAATGCTTAATGAATTACAACAGTACTGCGATGAGTGGC
AGGGCGGGGCGTAATTTTTTTAAGGCAGTTATTGGTGCCCTTAAACGCCTGGGGTAATGACTCTCTAGCTTGAGGCATCAAATAAAACGAAA
GGCTCAGTCGAAAGACTGGGCCTTTCGTTTTATCTGTTGTTTGTCGGTGAACGCTCTCCTGAGTAGGACAAATCCGCCGCTCTAGATTACGTG
CAGTCGATGATAAGCTGTCAAACATGAGAATTGTGCCTAATGAGTGAGCTAACTTACATTAATTGCGTTGCGCTCACTGCCCGCTTTCCAGTC
GGGAAACCTGTCGTGCCAGCTGCATTAATGAATCGGCCAACGCGCGGGGAGAGGCGGTTTGCGTATTGGGGGCCAGGGTGGTTTTTCTTTT
CACCAGTGAGACGGGCAACAGCTGATTGCCCTTCACCGCCTGGCCCTGAGAGAGTTGCAGCAAGCGGTCCACGCTGGTTTGCCCCAGCAGG
CGAAAATCCTGTTTGATGGTGGTTAACGGCGGGATATAACATGAGCTGTCTTCGGTATCGTCGTATCCCACTACCGAGATATCCGCACCAAC
GCGCAGCCCGGACTCGGTAATGGCGCGCATTGCGCCCAGCGCCATCTGATCGTTGGCAACCAGCATCGCAGTGGGAACGATGCCCTCATTC
AGCATTTGCATGGTTTGTTGAAAACCGGACATGGCACTCCAGTCGCCTTCCCGTTCCGCTATCGGCTGAATTTGATTGCGAGTGAGATATTTA
TGCCAGCCAGCCAGACGCAGACGCGCCGAGACAGAACTTAATGGGCCCGCTAACAGCGCGATTTGCTGGTGACCCAATGCGACCAGATGCT
CCACGCCCAGTCGCGTACCGTCTTCATGGGAGAAAATAATACTGTTGATGGGTGTCTGGTCAGAGACATCAAGAAATAACGCCGGAACATTA
GTGCAGGCAGCTTCCACAGCAATGGCATCCTGGTCATCCAGCGGATAGTTAATGATCAGCCCACTGACGCGTTGCGCGAGAAGATTGTGCA
CCGCCGCTTTACAGGCTTCGACGCCGCTTCGTTCTACCATCGACACCACCACGCTGGCACCCAGTTGATCGGCGCGAGATTTAATCGCCGCG
ACAATTTGCGACGGCGCGTGCAGGGCCAGACTGGAGGTGGCAACGCCAATCAGCAACGACTGTTTGCCCGCCAGTTGTTGTGCCACGCGGT
TGGGAATGTAATTCAGCTCCGCCATCGCCGCTTCCACTTTTTCCCGCGTTTTCGCAGAAACGTGGCTGGCCTGGTTCACCACGCGGGAAACG
GTCTGATAAGAGACACCGGCATACTCTGCGACATCGTATAACGTTACTGGTTTCACATTCACCACCCTGAATTGACTCTCTTCCGGGCGCTAT
CATGCCATACCGCGAAAGGTTTTGCACCATTCGATGGTGTCGGAATTTCGGGCAGCGTTGGGTCCTGGCCACGGGTGCGCATGATCTAGAG
CTGCCTCGCGCGTTTCGGTGATGACGGTGAAAACCTCTGACACATGCAGCTCCCGGAGACGGTCACAGCTTGTCTGTAAGCGGATGCCGGG
AGCAGACAAGCCCGTCAGGGCGCGTCAGCGGGTGTTGGCGGGTGTCGGGGCGCAGCCATGACCCAGTCACGTAGCGATAGCGGAGTGTA
TACTGGCTTAACTATGCGGCATCAGAGCAGATTGTACTGAGAGTGCACCATATGCGGTGTGAAATACCGCACAGATGCGTAAGGAGAAAAT
ACCGCATCAGGCGCTCTTCCGCTTCCTCGCTCACTGACTCGCTGCGCTCGGTCGTTCGGCTGCGGCGAGCGGTATCAGCTCACTCAAAGGCG
GTAATACGGTTATCCACAGAATCAGGGGATAACGCAGGAAAGAACATGTGAGCAAAAGGCCAGCAAAAGGCCAGGAACCGTAAAAAGGCC
GCGTTGCTGGCGTTTTTCCATAGGCTCCGCCCCCCTGACGAGCATCACAAAAATCGACGCTCAAGTCAGAGGTGGCGAAACCCGACAGGACT

ATAAAGATACCAGGCGTTTCCCCCTGGAAGCTCCCTCGTGCGCTCTCCTGTTCCGACCCTGCCGCTTACCGGATACCTGTCCGCCTTTCTCCCT
TCGGGAAGCGTGGCGCTTTCTCATAGCTCACGCTGTAGGTATCTCAGTTCGGTGTAGGTCGTTCGCTCCAAGCTGGGCTGTGTGCACGAACC
CCCCGTTCAGCCCGACCGCTGCGCCTTATCCGGTAACTATCGTCTTGAGTCCAACCCGGTAAGACACGACTTATCGCCACTGGCAGCAGCCAC
TGGTAACAGGATTAGCAGAGCGAGGTATGTAGGCGGTGCTACAGAGTTCTTGAAGTGGTGGCCTAACTACGGCTACACTAGAAGGACAGTA
TTTGGTATCTGCGCTCTGCTGAAGCCAGTTACCTTCGGAAAAAGAGTTGGTAGCTCTTGATCCGGCAAACAAACCACCGCTGGTAGCGGTGG
TTTTTTTGTTTGCAAGCAGCAGATTACGCGCAGAAAAAAAGGATCTCAAGAAGATCCTTTGATCTTTTCTACGGGGTCTGACGCTCAGTGGAA
CGAAAACTCACGTTAAGGGATTTTGGTCATGACATTAACCTATAAAAATAGGCGTATCACGAGGCCCTTTCGTCTTCAC

A3.2.6 pSF 1478 BdNEDD8-AGT

CTCGAGAAATCATAAAAAATTTATTTGCTTTGTGAGCGGATAACAATTATAATAGATTCAATTGTGAGCGGATAACAATTTCACACAGAATTCA
TTAAAGAGGAGAAATTAACCATGAGCAAGCATCACCATCATTCAGGCCATCACCATACCGGACACCACCATCATTCAGGCAGTCATCACCATT
CCGGAACCATGATTAAAGTAAAGACTCTGACCGGCAAGGAAATTGAGATCGACATCGAACCGACCGACACCATCGATCGTATCAAAGAACG
TGTGGAGGAAAAAGAAGGCATTCCTCCGGTTCAGCAGCGTCTCATTTACGCGGGTAAACAGCTTGCAGATGACAAAACCGCCAAAGATTAC
AATATCGAGGGCGGTAGCGTACTGCACCTTGTTCTGGCCCTGCGTGGTGGCGCCGGTACCAAAACTGAAGAAGGTAAACTGGTAATCTGGA
TTAACGGCGATAAAGGCTATAACGGTCTCGCTGAAGTCGGTAAGAAATTCGAGAAAGATACCGGAATTAAAGTCACCGTTGAGCATCCGGA
TAAACTGGAAGAGAAATTCCCACAGGTTGCGGCAACTGGCGATGGCCCTGACATTATCTTCTGGGCACACGACCGCTTTGGTGGCTACGCTC
AATCTGGCCTGTTGGCTGAAATCACCCCGGACAAAGCGTTCCAGGACAAGCTGTATCCGTTTACCTGGGATGCCGTACGTTACAACGGCAAG
CTGATTGCTTACCCGATCGCTGTTGAAGCGTTATCGCTGATTTATAACAAAGATCTGCTGCCGAACCCGCCAAAAACCTGGGAAGAGATCCC
GGCGCTGGATAAAGAACTGAAAGCGAAAGGTAAGAGCGCGCTGATGTTCAACCTGCAAGAACCGTACTTCACCTGGCCGCTGATTGCTGCT
GACGGGGGTTATGCGTTCAAGTATGAAAACGGCAAGTACGACATTAAAGACGTGGGCGTGGATAACGCTGGCGCGAAAGCGGGTCTGACC
TTCCTGGTTGACCTGATTAAAAACAAACACATGAATGCAGACACCGATTACTCCATCGCAGAAGCTGCCTTTAATAAAGGCGAAACAGCGAT
GACCATCAACGGCCCGTGGGCATGGTCCAACATCGACACCAGCAAAGTGAATTATGGTGTAACGGTACTGCCGACCTTCAAGGGTCAACCAT
CCAAACCGTTCGTTGGCGTGCTGAGCGCAGGTATTAACGCCGCCAGTCCGAACAAAGAGCTGGCAAAAGAGTTCCTCGAAAACTATCTGCTG
ACTGATGAAGGTCTGGAAGCGGTTAATAAAGACAAACCGCTGGGTGCCGTAGCGCTGAAGTCTTACGAGGAAGAGTTGGCGAAAGATCCA
CGTATTGCCGCCACCATGGAAAACGCCCAGAAAGGTGAAATCATGCCGAACATCCCGCAGATGTCCGCTTTCTGGTATGCCGTGCGTACTGC
GGTGATCAACGCCGCCAGCGGTCGTCAGACTGTCGATGAAGCCCTGAAAGACGCGCAGACTAATGGCACCGGTTGTTAAGGATCTCATCAC
CATCACCATCACTAAGCTTAATTAGCTGAGCTTGGACTCCTGTTGATAGATCCAGTAATGACCTCAGAACTCCATCTGGATTTGTTCAGAACGC
TCGGTTGCCGCCGGGCGTTTTTTATTGGTGAGAATCCAAGCTAGCCATGAAAATAAACTGTCTGCTTACATAAACAGTAATACAAGGGGTGTT
ATGAGCCATATTCAACGGGAAACGTCTTGCTCTAGGCCGCGATTAAATTCCAACATGGATGCTGATTTATATGGGTATAAATGGGCTCGCGA
TAATGTCGGGCAATCAGGTGCGACAATCTATCGATTGTATGGGAAGCCCGATGCGCCAGAGTTGTTTCTGAAACATGGCAAAGGTAGCGTT
GCCAATGATGTTACAGATGAGATGGTCAGACTAAACTGGCTGACGGAATTTATGCCTCTTCCGACCATCAAGCATTTTATCCGTACTCCTGAT
GATGCTTGGTTACTCACGACTGCGATCCCCGGCAAAACAGCATTCCAGGTATTAGAAGAATATCCTGATTCAGGTGAAAATATTGTTGATGC
GCTGGCAGTGTTCCTGCGCCGGTTGCATTCGATTCCTGTTTGTAATTGTCCTTTTAACAGCGATCGCGTATTTCGTCTCGCTCAGGCGCAATCA
CGAATGAATAACGGTTTGGTTGATGCGAGTGATTTTGATGACGAGCGTAATGGCTGGCCTGTTGAACAAGTCTGGAAAGAAATGCACAAAC
TTTTGCCATTCTCACCGGATTCAGTCGTCACTCATGGTGATTTCTCACTTGATAACCTTATTTTTGACGAGGGGAAATTAATAGGTTGTATTGA
TGTTGGACGAGTCGGAATCGCAGACCGATACCAGGATCTTGCCATCCTATGGAACTGCCTCGGTGAGTTTTCTCCTTCATTACAGAAACGGCT
TTTTCAAAAATATGGTATTGATAATCCTGATATGAATAAATTGCAGTTTCATTTGATGCTCGATGAGTTTTTCTAAGAATTAATTCATGGGCAA
ATATTATACGCAAGGCGACAAGGTGCTGATGCCGCTGGCGATTCAGGTTCATCATGCCGTTTGTGATGGCTTCCATGTCGGCAGAATGCTTA
ATGAATTACAACAGTACTGCGATGAGTGGCAGGGCGGGGCGTAATTTTTTTAAGGCAGTTATTGGTGCCCTTAAACGCCTGGGGTAATGACT
CTCTAGCTTGAGGCATCAAATAAAACGAAAGGCTCAGTCGAAAGACTGGGCCTTTCGTTTTATCTGTTGTTTGTCGGTGAACGCTCTCCTGAG
TAGGACAAATCCGCCGCTCTAGATTACGTGCAGTCGATGATAAGCTGTCAAACATGAGAATTGTGCCTAATGAGTGAGCTAACTTACATTAAT
TGCGTTGCGCTCACTGCCCGCTTTCCAGTCGGGAAACCTGTCGTGCCAGCTGCATTAATGAATCGGCCAACGCGCGGGGAGAGGCGGTTTG
CGTATTGGGGGCCAGGGTGGTTTTTCTTTTCACCAGTGAGACGGGCAACAGCTGATTGCCCTTCACCGCCTGGCCCTGAGAGAGTTGCAGCA
AGCGGTCCACGCTGGTTTGCCCCAGCAGGCGAAAATCCTGTTTGATGGTGGTTAACGGCGGGATATAACATGAGCTGTCTTCGGTATCGTCG
TATCCCACTACCGAGATATCCGCACCAACGCGCAGCCCGGACTCGGTAATGGCGCGCATTGCGCCCAGCGCCATCTGATCGTTGGCAACCAG
CATCGCAGTGGGAACGATGCCCTCATTCAGCATTTGCATGGTTTGTTGAAAACCGGACATGGCACTCCAGTCGCCTTCCCGTTCCGCTATCGG
CTGAATTTGATTGCGAGTGAGATATTTATGCCAGCCAGCCAGACGCAGACGCGCCGAGACAGAACTTAATGGGCCCGCTAACAGCGCGATT
TGCTGGTGACCCAATGCGACCAGATGCTCCACGCCCAGTCGCGTACCGTCTTCATGGGAGAAAATAATACTGTTGATGGGTGTCTGGTCAGA
GACATCAAGAAATAACGCCGGAACATTAGTGCAGGCAGCTTCCACAGCAATGGCATCCTGGTCATCCAGCGGATAGTTAATGATCAGCCCAC
TGACGCGTTGCGCGAGAAGATTGTGCACCGCCGCTTTACAGGCTTCGACGCCGCTTCGTTCTACCATCGACACCACCACGCTGGCACCCAGT
TGATCGGCGCGAGATTTAATCGCCGCGACAATTTGCGACGGCGCGTGCAGGGCCAGACTGGAGGTGGCAACGCCAATCAGCAACGACTGTT
TGCCCGCCAGTTGTTGTGCCACGCGGTTGGGAATGTAATTCAGCTCCGCCATCGCCGCTTCCACTTTTTCCCGCGTTTTCGCAGAAACGTGGC
TGGCCTGGTTCACCACGCGGGAAACGGTCTGATAAGAGACACCGGCATACTCTGCGACATCGTATAACGTTACTGGTTTCACATTCACCACC
CTGAATTGACTCTCTTCCGGGCGCTATCATGCCATACCGCGAAAGGTTTTGCACCATTCGATGGTGTCGGAATTTCGGGCAGCGTTGGGTCCT
GGCCACGGGTGCGCATGATCTAGAGCTGCCTCGCGCGTTTCGGTGATGACGGTGAAAACCTCTGACACATGCAGCTCCCGGAGACGGTCAC
AGCTTGTCTGTAAGCGGATGCCGGGAGCAGACAAGCCCGTCAGGGCGCGTCAGCGGGTGTTGGCGGGTGTCGGGGCGCAGCCATGACCC
AGTCACGTAGCGATAGCGGAGTGTATACTGGCTTAACTATGCGGCATCAGAGCAGATTGTACTGAGAGTGCACCATATGCGGTGTGAAATA
CCGCACAGATGCGTAAGGAGAAAATACCGCATCAGGCGCTCTTCCGCTTCCTCGCTCACTGACTCGCTGCGCTCGGTCGTTCGGCTGCGGCG
AGCGGTATCAGCTCACTCAAAGGCGGTAATACGGTTATCCACAGAATCAGGGGATAACGCAGGAAAGAACATGTGAGCAAAAGGCCAGCA
AAAGGCCAGGAACCGTAAAAAGGCCGCGTTGCTGGCGTTTTTCCATAGGCTCCGCCCCCCTGACGAGCATCACAAAAATCGACGCTCAAGTC
AGAGGTGGCGAAACCCGACAGGACTATAAAGATACCAGGCGTTTCCCCCTGGAAGCTCCCTCGTGCGCTCTCCTGTTCCGACCCTGCCGCTT
ACCGGATACCTGTCCGCCTTTCTCCCTTCGGGAAGCGTGGCGCTTTCTCATAGCTCACGCTGTAGGTATCTCAGTTCGGTGTAGGTCGTTCGC
TCCAAGCTGGGCTGTGTGCACGAACCCCCCGTTCAGCCCGACCGCTGCGCCTTATCCGGTAACTATCGTCTTGAGTCCAACCCGGTAAGACA
CGACTTATCGCCACTGGCAGCAGCCACTGGTAACAGGATTAGCAGAGCGAGGTATGTAGGCGGTGCTACAGAGTTCTTGAAGTGGTGGCCT
AACTACGGCTACACTAGAAGGACAGTATTTGGTATCTGCGCTCTGCTGAAGCCAGTTACCTTCGGAAAAAGAGTTGGTAGCTCTTGATCCGG
CAAACAAACCACCGCTGGTAGCGGTGGTTTTTTTGTTTGCAAGCAGCAGATTACGCGCAGAAAAAAAGGATCTCAAGAAGATCCTTTGATCT
TTTCTACGGGGTCTGACGCTCAGTGGAACGAAAACTCACGTTAAGGGATTTTGGTCATGACATTAACCTATAAAAATAGGCGTATCACGAGG
CCCTTTCGTCTTCAC

A3.2.7 pSF-1479 SsNEDD8-AGT

CTCGAGAAATCATAAAAAATTTATTTGCTTTGTGAGCGGATAACAATTATAATAGATTCAATTGTGAGCGGATAACAATTTCACACAGAATTCA
TTAAAGAGGAGAAATTAACCATGAGCAAGCATCACCATCATTCAGGCCATCACCATACCGGACACCACCATCATTCAGGCAGTCATCACCATT
CCGGAATGCTGATTAAAGTTAAAACCCTGACTGGTAAGGAAATTGAGATTGACATCGAACCTACCGACAAGGTTGAACGTATTAAAGAACGT
GTGGAAGAAAAGGAAGGTATCCCGCCGCAGCAACAGCGTCTGATCTACTCTGGTAAACAGATGAACGACGAAAAGACTGCGGCTGACTATA
AAATCCAGGGTGGCTCCGTGCTGCACCTGGTACTGGCCCTGCGTGGTGGCGCCGGTACCAAAACTGAAGAAGGTAAACTGGTAATCTGGAT
TAACGGCGATAAAGGCTATAACGGTCTCGCTGAAGTCGGTAAGAAATTCGAGAAAGATACCGGAATTAAAGTCACCGTTGAGCATCCGGAT
AAACTGGAAGAGAAATTCCCACAGGTTGCGGCAACTGGCGATGGCCCTGACATTATCTTCTGGGCACACGACCGCTTTGGTGGCTACGCTCA
ATCTGGCCTGTTGGCTGAAATCACCCCGGACAAAGCGTTCCAGGACAAGCTGTATCCGTTTACCTGGGATGCCGTACGTTACAACGGCAAGC
TGATTGCTTACCCGATCGCTGTTGAAGCGTTATCGCTGATTTATAACAAAGATCTGCTGCCGAACCCGCCAAAAACCTGGGAAGAGATCCCG
GCGCTGGATAAAGAACTGAAAGCGAAAGGTAAGAGCGCGCTGATGTTCAACCTGCAAGAACCGTACTTCACCTGGCCGCTGATTGCTGCTG
ACGGGGGTTATGCGTTCAAGTATGAAAACGGCAAGTACGACATTAAAGACGTGGGCGTGGATAACGCTGGCGCGAAAGCGGGTCTGACCT
TCCTGGTTGACCTGATTAAAAACAAACACATGAATGCAGACACCGATTACTCCATCGCAGAAGCTGCCTTTAATAAAGGCGAAACAGCGATG
ACCATCAACGGCCCGTGGGCATGGTCCAACATCGACACCAGCAAAGTGAATTATGGTGTAACGGTACTGCCGACCTTCAAGGGTCAACCATC
CAAACCGTTCGTTGGCGTGCTGAGCGCAGGTATTAACGCCGCCAGTCCGAACAAAGAGCTGGCAAAAGAGTTCCTCGAAAACTATCTGCTG
ACTGATGAAGGTCTGGAAGCGGTTAATAAAGACAAACCGCTGGGTGCCGTAGCGCTGAAGTCTTACGAGGAAGAGTTGGCGAAAGATCCA
CGTATTGCCGCCACCATGGAAAACGCCCAGAAAGGTGAAATCATGCCGAACATCCCGCAGATGTCCGCTTTCTGGTATGCCGTGCGTACTGC
GGTGATCAACGCCGCCAGCGGTCGTCAGACTGTCGATGAAGCCCTGAAAGACGCGCAGACTAATGGCACCGGTTGTTAAGGATCTCATCAC
CATCACCATCACTAAGCTTAATTAGCTGAGCTTGGACTCCTGTTGATAGATCCAGTAATGACCTCAGAACTCCATCTGGATTTGTTCAGAACGC
TCGGTTGCCGCCGGGCGTTTTTTATTGGTGAGAATCCAAGCTAGCCATGAAAATAAACTGTCTGCTTACATAAACAGTAATACAAGGGGTGTT
ATGAGCCATATTCAACGGGAAACGTCTTGCTCTAGGCCGCGATTAAATTCCAACATGGATGCTGATTTATATGGGTATAAATGGGCTCGCGA
TAATGTCGGGCAATCAGGTGCGACAATCTATCGATTGTATGGGAAGCCCGATGCGCCAGAGTTGTTTCTGAAACATGGCAAAGGTAGCGTT
GCCAATGATGTTACAGATGAGATGGTCAGACTAAACTGGCTGACGGAATTTATGCCTCTTCCGACCATCAAGCATTTTATCCGTACTCCTGAT
GATGCTTGGTTACTCACGACTGCGATCCCCGGCAAAACAGCATTCCAGGTATTAGAAGAATATCCTGATTCAGGTGAAAATATTGTTGATGC
GCTGGCAGTGTTCCTGCGCCGGTTGCATTCGATTCCTGTTTGTAATTGTCCTTTTAACAGCGATCGCGTATTTCGTCTCGCTCAGGCGCAATCA
CGAATGAATAACGGTTTGGTTGATGCGAGTGATTTTGATGACGAGCGTAATGGCTGGCCTGTTGAACAAGTCTGGAAAGAAATGCACAAAC
TTTTGCCATTCTCACCGGATTCAGTCGTCACTCATGGTGATTTCTCACTTGATAACCTTATTTTTGACGAGGGGAAATTAATAGGTTGTATTGA
TGTTGGACGAGTCGGAATCGCAGACCGATACCAGGATCTTGCCATCCTATGGAACTGCCTCGGTGAGTTTTCTCCTTCATTACAGAAACGGCT
TTTTCAAAAATATGGTATTGATAATCCTGATATGAATAAATTGCAGTTTCATTTGATGCTCGATGAGTTTTTCTAAGAATTAATTCATGGGCAA
ATATTATACGCAAGGCGACAAGGTGCTGATGCCGCTGGCGATTCAGGTTCATCATGCCGTTTGTGATGGCTTCCATGTCGGCAGAATGCTTA
ATGAATTACAACAGTACTGCGATGAGTGGCAGGGCGGGGCGTAATTTTTTTAAGGCAGTTATTGGTGCCCTTAAACGCCTGGGGTAATGACT
CTCTAGCTTGAGGCATCAAATAAAACGAAAGGCTCAGTCGAAAGACTGGGCCTTTCGTTTTATCTGTTGTTTGTCGGTGAACGCTCTCCTGAG
TAGGACAAATCCGCCGCTCTAGATTACGTGCAGTCGATGATAAGCTGTCAAACATGAGAATTGTGCCTAATGAGTGAGCTAACTTACATTAAT
TGCGTTGCGCTCACTGCCCGCTTTCCAGTCGGGAAACCTGTCGTGCCAGCTGCATTAATGAATCGGCCAACGCGCGGGGAGAGGCGGTTTG
CGTATTGGGGGCCAGGGTGGTTTTTCTTTTCACCAGTGAGACGGGCAACAGCTGATTGCCCTTCACCGCCTGGCCCTGAGAGAGTTGCAGCA
AGCGGTCCACGCTGGTTTGCCCCAGCAGGCGAAAATCCTGTTTGATGGTGGTTAACGGCGGGATATAACATGAGCTGTCTTCGGTATCGTCG
TATCCCACTACCGAGATATCCGCACCAACGCGCAGCCCGGACTCGGTAATGGCGCGCATTGCGCCCAGCGCCATCTGATCGTTGGCAACCAG
CATCGCAGTGGGAACGATGCCCTCATTCAGCATTTGCATGGTTTGTTGAAAACCGGACATGGCACTCCAGTCGCCTTCCCGTTCCGCTATCGG
CTGAATTTGATTGCGAGTGAGATATTTATGCCAGCCAGCCAGACGCAGACGCGCCGAGACAGAACTTAATGGGCCCGCTAACAGCGCGATT
TGCTGGTGACCCAATGCGACCAGATGCTCCACGCCCAGTCGCGTACCGTCTTCATGGGAGAAAATAATACTGTTGATGGGTGTCTGGTCAGA
GACATCAAGAAATAACGCCGGAACATTAGTGCAGGCAGCTTCCACAGCAATGGCATCCTGGTCATCCAGCGGATAGTTAATGATCAGCCCAC
TGACGCGTTGCGCGAGAAGATTGTGCACCGCCGCTTTACAGGCTTCGACGCCGCTTCGTTCTACCATCGACACCACCACGCTGGCACCCAGT
TGATCGGCGCGAGATTTAATCGCCGCGACAATTTGCGACGGCGCGTGCAGGGCCAGACTGGAGGTGGCAACGCCAATCAGCAACGACTGTT
TGCCCGCCAGTTGTTGTGCCACGCGGTTGGGAATGTAATTCAGCTCCGCCATCGCCGCTTCCACTTTTTCCCGCGTTTTCGCAGAAACGTGGC
TGGCCTGGTTCACCACGCGGGAAACGGTCTGATAAGAGACACCGGCATACTCTGCGACATCGTATAACGTTACTGGTTTCACATTCACCACC
CTGAATTGACTCTCTTCCGGGCGCTATCATGCCATACCGCGAAAGGTTTTGCACCATTCGATGGTGTCGGAATTTCGGGCAGCGTTGGGTCCT
GGCCACGGGTGCGCATGATCTAGAGCTGCCTCGCGCGTTTCGGTGATGACGGTGAAAACCTCTGACACATGCAGCTCCCGGAGACGGTCAC
AGCTTGTCTGTAAGCGGATGCCGGGAGCAGACAAGCCCGTCAGGGCGCGTCAGCGGGTGTTGGCGGGTGTCGGGGCGCAGCCATGACCC
AGTCACGTAGCGATAGCGGAGTGTATACTGGCTTAACTATGCGGCATCAGAGCAGATTGTACTGAGAGTGCACCATATGCGGTGTGAAATA
CCGCACAGATGCGTAAGGAGAAAATACCGCATCAGGCGCTCTTCCGCTTCCTCGCTCACTGACTCGCTGCGCTCGGTCGTTCGGCTGCGGCG
AGCGGTATCAGCTCACTCAAAGGCGGTAATACGGTTATCCACAGAATCAGGGGATAACGCAGGAAAGAACATGTGAGCAAAAGGCCAGCA
AAAGGCCAGGAACCGTAAAAAGGCCGCGTTGCTGGCGTTTTTCCATAGGCTCCGCCCCCCTGACGAGCATCACAAAAATCGACGCTCAAGTC
AGAGGTGGCGAAACCCGACAGGACTATAAAGATACCAGGCGTTTCCCCCTGGAAGCTCCCTCGTGCGCTCTCCTGTTCCGACCCTGCCGCTT
ACCGGATACCTGTCCGCCTTTCTCCCTTCGGGAAGCGTGGCGCTTTCTCATAGCTCACGCTGTAGGTATCTCAGTTCGGTGTAGGTCGTTCGC
TCCAAGCTGGGCTGTGTGCACGAACCCCCCGTTCAGCCCGACCGCTGCGCCTTATCCGGTAACTATCGTCTTGAGTCCAACCCGGTAAGACA
CGACTTATCGCCACTGGCAGCAGCCACTGGTAACAGGATTAGCAGAGCGAGGTATGTAGGCGGTGCTACAGAGTTCTTGAAGTGGTGGCCT
AACTACGGCTACACTAGAAGGACAGTATTTGGTATCTGCGCTCTGCTGAAGCCAGTTACCTTCGGAAAAAGAGTTGGTAGCTCTTGATCCGG
CAAACAAACCACCGCTGGTAGCGGTGGTTTTTTTGTTTGCAAGCAGCAGATTACGCGCAGAAAAAAAGGATCTCAAGAAGATCCTTTGATCT
TTTCTACGGGGTCTGACGCTCAGTGGAACGAAAACTCACGTTAAGGGATTTTGGTCATGACATTAACCTATAAAAATAGGCGTATCACGAGG
CCCTTTCGTCTTCAC

A3.2.8 pelB pSF 1477 BdSUMO-MBP

CTCGAGAAATCATAAAAAATTTATTTGCTTTGTGAGCGGATAACAATTATAATAGATTCAATTGTGAGCGGATAACAATTTCACACAGAATTCA
TTAAAGAGGAGAAATTAACCATGAAATACCTGCTGCCGACGGCTGCTGCTGGTCTGCTGCTGCTGGCTGCCCAACCGGCGATGGCGAGCAA
GCATCACCATCATTCAGGCCATCACCATACCGGACACCACCATCATTCAGGCAGTCATCACCATTCCGGATCTGCTGCGGGTGGCGAAGAAG
ATAAGAAACCGGCAGGTGGCGAAGGTGGCGGTGCCCATATCAACCTGAAAGTGAAAAGGTCAAGACGGCAACGAAGTCTTTTTCCGCATCAA
ACGTTCTACCCAGCTGAAAAAGCTGATGAACGCATACTGTGACCGTCAGTCTGTAGACATGACCGCAATTGCTTTCCTGTTTGATGGTCGTCG
CCTGCGTGCGGAACAGACCCCGGATGAACTGGAGATGGAAGATGGCGACGAAATCGACGCAATGCTGCACCAGACTGGTGGCGCCGGTAC
CAAAACTGAAGAAGGTAAACTGGTAATCTGGATTAACGGCGATAAAGGCTATAACGGTCTCGCTGAAGTCGGTAAGAAATTCGAGAAAGAT
ACCGGAATTAAAGTCACCGTTGAGCATCCGGATAAACTGGAAGAGAAATTCCCACAGGTTGCGGCAACTGGCGATGGCCCTGACATTATCTT
CTGGGCACACGACCGCTTTGGTGGCTACGCTCAATCTGGCCTGTTGGCTGAAATCACCCCGGACAAAGCGTTCCAGGACAAGCTGTATCCGT
TTACCTGGGATGCCGTACGTTACAACGGCAAGCTGATTGCTTACCCGATCGCTGTTGAAGCGTTATCGCTGATTTATAACAAAGATCTGCTGC
CGAACCCGCCAAAAACCTGGGAAGAGATCCCGGCGCTGGATAAAGAACTGAAAGCGAAAGGTAAGAGCGCGCTGATGTTCAACCTGCAAG
AACCGTACTTCACCTGGCCGCTGATTGCTGCTGACGGGGGTTATGCGTTCAAGTATGAAAACGGCAAGTACGACATTAAAGACGTGGGCGT
GGATAACGCTGGCGCGAAAGCGGGTCTGACCTTCCTGGTTGACCTGATTAAAAACAAACACATGAATGCAGACACCGATTACTCCATCGCAG
AAGCTGCCTTTAATAAAGGCGAAACAGCGATGACCATCAACGGCCCGTGGGCATGGTCCAACATCGACACCAGCAAAGTGAATTATGGTGT
AACGGTACTGCCGACCTTCAAGGGTCAACCATCCAAACCGTTCGTTGGCGTGCTGAGCGCAGGTATTAACGCCGCCAGTCCGAACAAAGAG
CTGGCAAAAGAGTTCCTCGAAAACTATCTGCTGACTGATGAAGGTCTGGAAGCGGTTAATAAAGACAAACCGCTGGGTGCCGTAGCGCTGA
AGTCTTACGAGGAAGAGTTGGCGAAAGATCCACGTATTGCCGCCACCATGGAAAACGCCCAGAAAGGTGAAATCATGCCGAACATCCCGCA
GATGTCCGCTTTCTGGTATGCCGTGCGTACTGCGGTGATCAACGCCGCCAGCGGTCGTCAGACTGTCGATGAAGCCCTGAAAGACGCGCAG
ACTAATGGCACCGGTTGTTAAGGATCTCATCACCATCACCATCACTAAGCTTAATTAGCTGAGCTTGGACTCCTGTTGATAGATCCAGTAATG
ACCTCAGAACTCCATCTGGATTTGTTCAGAACGCTCGGTTGCCGCCGGGCGTTTTTTATTGGTGAGAATCCAAGCTAGCCATGAAAATAAACT
GTCTGCTTACATAAACAGTAATACAAGGGGTGTTATGAGCCATATTCAACGGGAAACGTCTTGCTCTAGGCCGCGATTAAATTCCAACATGG
ATGCTGATTTATATGGGTATAAATGGGCTCGCGATAATGTCGGGCAATCAGGTGCGACAATCTATCGATTGTATGGGAAGCCCGATGCGCCA
GAGTTGTTTCTGAAACATGGCAAAGGTAGCGTTGCCAATGATGTTACAGATGAGATGGTCAGACTAAACTGGCTGACGGAATTTATGCCTCT
TCCGACCATCAAGCATTTTATCCGTACTCCTGATGATGCTTGGTTACTCACGACTGCGATCCCCGGCAAAACAGCATTCCAGGTATTAGAAGA
ATATCCTGATTCAGGTGAAAATATTGTTGATGCGCTGGCAGTGTTCCTGCGCCGGTTGCATTCGATTCCTGTTTGTAATTGTCCTTTTAACAGC
GATCGCGTATTTCGTCTCGCTCAGGCGCAATCACGAATGAATAACGGTTTGGTTGATGCGAGTGATTTTGATGACGAGCGTAATGGCTGGCC
TGTTGAACAAGTCTGGAAAGAAATGCACAAACTTTTGCCATTCTCACCGGATTCAGTCGTCACTCATGGTGATTTCTCACTTGATAACCTTATT
TTTGACGAGGGGAAATTAATAGGTTGTATTGATGTTGGACGAGTCGGAATCGCAGACCGATACCAGGATCTTGCCATCCTATGGAACTGCCT
CGGTGAGTTTTCTCCTTCATTACAGAAACGGCTTTTTCAAAAATATGGTATTGATAATCCTGATATGAATAAATTGCAGTTTCATTTGATGCTC
GATGAGTTTTTCTAAGAATTAATTCATGGGCAAATATTATACGCAAGGCGACAAGGTGCTGATGCCGCTGGCGATTCAGGTTCATCATGCCG
TTTGTGATGGCTTCCATGTCGGCAGAATGCTTAATGAATTACAACAGTACTGCGATGAGTGGCAGGGCGGGGCGTAATTTTTTTAAGGCAGT
TATTGGTGCCCTTAAACGCCTGGGGTAATGACTCTCTAGCTTGAGGCATCAAATAAAACGAAAGGCTCAGTCGAAAGACTGGGCCTTTCGTT
TTATCTGTTGTTTGTCGGTGAACGCTCTCCTGAGTAGGACAAATCCGCCGCTCTAGATTACGTGCAGTCGATGATAAGCTGTCAAACATGAGA
ATTGTGCCTAATGAGTGAGCTAACTTACATTAATTGCGTTGCGCTCACTGCCCGCTTTCCAGTCGGGAAACCTGTCGTGCCAGCTGCATTAAT
GAATCGGCCAACGCGCGGGGAGAGGCGGTTTGCGTATTGGGGGCCAGGGTGGTTTTTCTTTTCACCAGTGAGACGGGCAACAGCTGATTG
CCCTTCACCGCCTGGCCCTGAGAGAGTTGCAGCAAGCGGTCCACGCTGGTTTGCCCCAGCAGGCGAAAATCCTGTTTGATGGTGGTTAACG
GCGGGATATAACATGAGCTGTCTTCGGTATCGTCGTATCCCACTACCGAGATATCCGCACCAACGCGCAGCCCGGACTCGGTAATGGCGCGC
ATTGCGCCCAGCGCCATCTGATCGTTGGCAACCAGCATCGCAGTGGGAACGATGCCCTCATTCAGCATTTGCATGGTTTGTTGAAAACCGGA
CATGGCACTCCAGTCGCCTTCCCGTTCCGCTATCGGCTGAATTTGATTGCGAGTGAGATATTTATGCCAGCCAGCCAGACGCAGACGCGCCG
AGACAGAACTTAATGGGCCCGCTAACAGCGCGATTTGCTGGTGACCCAATGCGACCAGATGCTCCACGCCCAGTCGCGTACCGTCTTCATGG
GAGAAAATAATACTGTTGATGGGTGTCTGGTCAGAGACATCAAGAAATAACGCCGGAACATTAGTGCAGGCAGCTTCCACAGCAATGGCAT
CCTGGTCATCCAGCGGATAGTTAATGATCAGCCCACTGACGCGTTGCGCGAGAAGATTGTGCACCGCCGCTTTACAGGCTTCGACGCCGCTT
CGTTCTACCATCGACACCACCACGCTGGCACCCAGTTGATCGGCGCGAGATTTAATCGCCGCGACAATTTGCGACGGCGCGTGCAGGGCCA
GACTGGAGGTGGCAACGCCAATCAGCAACGACTGTTTGCCCGCCAGTTGTTGTGCCACGCGGTTGGGAATGTAATTCAGCTCCGCCATCGCC
GCTTCCACTTTTTCCCGCGTTTTCGCAGAAACGTGGCTGGCCTGGTTCACCACGCGGGAAACGGTCTGATAAGAGACACCGGCATACTCTGC
GACATCGTATAACGTTACTGGTTTCACATTCACCACCCTGAATTGACTCTCTTCCGGGCGCTATCATGCCATACCGCGAAAGGTTTTGCACCAT
TCGATGGTGTCGGAATTTCGGGCAGCGTTGGGTCCTGGCCACGGGTGCGCATGATCTAGAGCTGCCTCGCGCGTTTCGGTGATGACGGTGA
AAACCTCTGACACATGCAGCTCCCGGAGACGGTCACAGCTTGTCTGTAAGCGGATGCCGGGAGCAGACAAGCCCGTCAGGGCGCGTCAGC
GGGTGTTGGCGGGTGTCGGGGCGCAGCCATGACCCAGTCACGTAGCGATAGCGGAGTGTATACTGGCTTAACTATGCGGCATCAGAGCAG
ATTGTACTGAGAGTGCACCATATGCGGTGTGAAATACCGCACAGATGCGTAAGGAGAAAATACCGCATCAGGCGCTCTTCCGCTTCCTCGCT
CACTGACTCGCTGCGCTCGGTCGTTCGGCTGCGGCGAGCGGTATCAGCTCACTCAAAGGCGGTAATACGGTTATCCACAGAATCAGGGGAT
AACGCAGGAAAGAACATGTGAGCAAAAGGCCAGCAAAAGGCCAGGAACCGTAAAAAGGCCGCGTTGCTGGCGTTTTTCCATAGGCTCCGC
CCCCCTGACGAGCATCACAAAAATCGACGCTCAAGTCAGAGGTGGCGAAACCCGACAGGACTATAAAGATACCAGGCGTTTCCCCCTGGAA
GCTCCCTCGTGCGCTCTCCTGTTCCGACCCTGCCGCTTACCGGATACCTGTCCGCCTTTCTCCCTTCGGGAAGCGTGGCGCTTTCTCATAGCTC
ACGCTGTAGGTATCTCAGTTCGGTGTAGGTCGTTCGCTCCAAGCTGGGCTGTGTGCACGAACCCCCCGTTCAGCCCGACCGCTGCGCCTTAT
CCGGTAACTATCGTCTTGAGTCCAACCCGGTAAGACACGACTTATCGCCACTGGCAGCAGCCACTGGTAACAGGATTAGCAGAGCGAGGTAT
GTAGGCGGTGCTACAGAGTTCTTGAAGTGGTGGCCTAACTACGGCTACACTAGAAGGACAGTATTTGGTATCTGCGCTCTGCTGAAGCCAGT
TACCTTCGGAAAAAGAGTTGGTAGCTCTTGATCCGGCAAACAAACCACCGCTGGTAGCGGTGGTTTTTTTGTTTGCAAGCAGCAGATTACGC
GCAGAAAAAAAGGATCTCAAGAAGATCCTTTGATCTTTTCTACGGGGTCTGACGCTCAGTGGAACGAAAACTCACGTTAAGGGATTTTGGTC
ATGACATTAACCTATAAAAATAGGCGTATCACGAGGCCCTTTCGTCTTCAC

A3.2.9 pelB pSF 1478 BdNEDD8-AGT

CTCGAGAAATCATAAAAAATTTATTTGCTTTGTGAGCGGATAACAATTATAATAGATTCAATTGTGAGCGGATAACAATTTCACACAGAATTCA
TTAAAGAGGAGAAATTAACCATGAAATACCTGCTGCCGACGGCTGCTGCTGGTCTGCTGCTGCTGGCTGCCCAACCGGCGATGGCGAGCAA
GCATCACCATCATTCAGGCCATCACCATACCGGACACCACCATCATTCAGGCAGTCATCACCATTCCGGAACCATGATTAAAGTAAAGACTCT
GACCGGCAAGGAAATTGAGATCGACATCGAACCGACCGACACCATCGATCGTATCAAAGAACGTGTGGAGGAAAAAGAAGGCATTCCTCCG
GTTCAGCAGCGTCTCATTTACGCGGGTAAACAGCTTGCAGATGACAAAACCGCCAAAGATTACAATATCGAGGGCGGTAGCGTACTGCACCT
TGTTCTGGCCCTGCGTGGTGGCGCCGGTACCAAAACTGAAGAAGGTAAACTGGTAATCTGGATTAACGGCGATAAAGGCTATAACGGTCTC
GCTGAAGTCGGTAAGAAATTCGAGAAAGATACCGGAATTAAAGTCACCGTTGAGCATCCGGATAAACTGGAAGAGAAATTCCCACAGGTTG
CGGCAACTGGCGATGGCCCTGACATTATCTTCTGGGCACACGACCGCTTTGGTGGCTACGCTCAATCTGGCCTGTTGGCTGAAATCACCCCG
GACAAAGCGTTCCAGGACAAGCTGTATCCGTTTACCTGGGATGCCGTACGTTACAACGGCAAGCTGATTGCTTACCCGATCGCTGTTGAAGC
GTTATCGCTGATTTATAACAAAGATCTGCTGCCGAACCCGCCAAAAACCTGGGAAGAGATCCCGGCGCTGGATAAAGAACTGAAAGCGAAA
GGTAAGAGCGCGCTGATGTTCAACCTGCAAGAACCGTACTTCACCTGGCCGCTGATTGCTGCTGACGGGGGTTATGCGTTCAAGTATGAAAA
CGGCAAGTACGACATTAAAGACGTGGGCGTGGATAACGCTGGCGCGAAAGCGGGTCTGACCTTCCTGGTTGACCTGATTAAAAACAAACAC
ATGAATGCAGACACCGATTACTCCATCGCAGAAGCTGCCTTTAATAAAGGCGAAACAGCGATGACCATCAACGGCCCGTGGGCATGGTCCA
ACATCGACACCAGCAAAGTGAATTATGGTGTAACGGTACTGCCGACCTTCAAGGGTCAACCATCCAAACCGTTCGTTGGCGTGCTGAGCGCA
GGTATTAACGCCGCCAGTCCGAACAAAGAGCTGGCAAAAGAGTTCCTCGAAAACTATCTGCTGACTGATGAAGGTCTGGAAGCGGTTAATA
AAGACAAACCGCTGGGTGCCGTAGCGCTGAAGTCTTACGAGGAAGAGTTGGCGAAAGATCCACGTATTGCCGCCACCATGGAAAACGCCCA
GAAAGGTGAAATCATGCCGAACATCCCGCAGATGTCCGCTTTCTGGTATGCCGTGCGTACTGCGGTGATCAACGCCGCCAGCGGTCGTCAG
ACTGTCGATGAAGCCCTGAAAGACGCGCAGACTAATGGCACCGGTTGTTAAGGATCTCATCACCATCACCATCACTAAGCTTAATTAGCTGA
GCTTGGACTCCTGTTGATAGATCCAGTAATGACCTCAGAACTCCATCTGGATTTGTTCAGAACGCTCGGTTGCCGCCGGGCGTTTTTTATTGG
TGAGAATCCAAGCTAGCCATGAAAATAAACTGTCTGCTTACATAAACAGTAATACAAGGGGTGTTATGAGCCATATTCAACGGGAAACGTCT
TGCTCTAGGCCGCGATTAAATTCCAACATGGATGCTGATTTATATGGGTATAAATGGGCTCGCGATAATGTCGGGCAATCAGGTGCGACAAT
CTATCGATTGTATGGGAAGCCCGATGCGCCAGAGTTGTTTCTGAAACATGGCAAAGGTAGCGTTGCCAATGATGTTACAGATGAGATGGTCA
GACTAAACTGGCTGACGGAATTTATGCCTCTTCCGACCATCAAGCATTTTATCCGTACTCCTGATGATGCTTGGTTACTCACGACTGCGATCCC
CGGCAAAACAGCATTCCAGGTATTAGAAGAATATCCTGATTCAGGTGAAAATATTGTTGATGCGCTGGCAGTGTTCCTGCGCCGGTTGCATT
CGATTCCTGTTTGTAATTGTCCTTTTAACAGCGATCGCGTATTTCGTCTCGCTCAGGCGCAATCACGAATGAATAACGGTTTGGTTGATGCGA
GTGATTTTGATGACGAGCGTAATGGCTGGCCTGTTGAACAAGTCTGGAAAGAAATGCACAAACTTTTGCCATTCTCACCGGATTCAGTCGTC
ACTCATGGTGATTTCTCACTTGATAACCTTATTTTTGACGAGGGGAAATTAATAGGTTGTATTGATGTTGGACGAGTCGGAATCGCAGACCGA
TACCAGGATCTTGCCATCCTATGGAACTGCCTCGGTGAGTTTTCTCCTTCATTACAGAAACGGCTTTTTCAAAAATATGGTATTGATAATCCTG
ATATGAATAAATTGCAGTTTCATTTGATGCTCGATGAGTTTTTCTAAGAATTAATTCATGGGCAAATATTATACGCAAGGCGACAAGGTGCTG
ATGCCGCTGGCGATTCAGGTTCATCATGCCGTTTGTGATGGCTTCCATGTCGGCAGAATGCTTAATGAATTACAACAGTACTGCGATGAGTG
GCAGGGCGGGGCGTAATTTTTTTAAGGCAGTTATTGGTGCCCTTAAACGCCTGGGGTAATGACTCTCTAGCTTGAGGCATCAAATAAAACGA
AAGGCTCAGTCGAAAGACTGGGCCTTTCGTTTTATCTGTTGTTTGTCGGTGAACGCTCTCCTGAGTAGGACAAATCCGCCGCTCTAGATTACG
TGCAGTCGATGATAAGCTGTCAAACATGAGAATTGTGCCTAATGAGTGAGCTAACTTACATTAATTGCGTTGCGCTCACTGCCCGCTTTCCAG
TCGGGAAACCTGTCGTGCCAGCTGCATTAATGAATCGGCCAACGCGCGGGGAGAGGCGGTTTGCGTATTGGGGGCCAGGGTGGTTTTTCTT
TTCACCAGTGAGACGGGCAACAGCTGATTGCCCTTCACCGCCTGGCCCTGAGAGAGTTGCAGCAAGCGGTCCACGCTGGTTTGCCCCAGCA
GGCGAAAATCCTGTTTGATGGTGGTTAACGGCGGGATATAACATGAGCTGTCTTCGGTATCGTCGTATCCCACTACCGAGATATCCGCACCA
ACGCGCAGCCCGGACTCGGTAATGGCGCGCATTGCGCCCAGCGCCATCTGATCGTTGGCAACCAGCATCGCAGTGGGAACGATGCCCTCAT
TCAGCATTTGCATGGTTTGTTGAAAACCGGACATGGCACTCCAGTCGCCTTCCCGTTCCGCTATCGGCTGAATTTGATTGCGAGTGAGATATT
TATGCCAGCCAGCCAGACGCAGACGCGCCGAGACAGAACTTAATGGGCCCGCTAACAGCGCGATTTGCTGGTGACCCAATGCGACCAGATG
CTCCACGCCCAGTCGCGTACCGTCTTCATGGGAGAAAATAATACTGTTGATGGGTGTCTGGTCAGAGACATCAAGAAATAACGCCGGAACAT
TAGTGCAGGCAGCTTCCACAGCAATGGCATCCTGGTCATCCAGCGGATAGTTAATGATCAGCCCACTGACGCGTTGCGCGAGAAGATTGTGC
ACCGCCGCTTTACAGGCTTCGACGCCGCTTCGTTCTACCATCGACACCACCACGCTGGCACCCAGTTGATCGGCGCGAGATTTAATCGCCGC
GACAATTTGCGACGGCGCGTGCAGGGCCAGACTGGAGGTGGCAACGCCAATCAGCAACGACTGTTTGCCCGCCAGTTGTTGTGCCACGCG
GTTGGGAATGTAATTCAGCTCCGCCATCGCCGCTTCCACTTTTTCCCGCGTTTTCGCAGAAACGTGGCTGGCCTGGTTCACCACGCGGGAAAC
GGTCTGATAAGAGACACCGGCATACTCTGCGACATCGTATAACGTTACTGGTTTCACATTCACCACCCTGAATTGACTCTCTTCCGGGCGCTA
TCATGCCATACCGCGAAAGGTTTTGCACCATTCGATGGTGTCGGAATTTCGGGCAGCGTTGGGTCCTGGCCACGGGTGCGCATGATCTAGA
GCTGCCTCGCGCGTTTCGGTGATGACGGTGAAAACCTCTGACACATGCAGCTCCCGGAGACGGTCACAGCTTGTCTGTAAGCGGATGCCGG
GAGCAGACAAGCCCGTCAGGGCGCGTCAGCGGGTGTTGGCGGGTGTCGGGGCGCAGCCATGACCCAGTCACGTAGCGATAGCGGAGTGT
ATACTGGCTTAACTATGCGGCATCAGAGCAGATTGTACTGAGAGTGCACCATATGCGGTGTGAAATACCGCACAGATGCGTAAGGAGAAAA
TACCGCATCAGGCGCTCTTCCGCTTCCTCGCTCACTGACTCGCTGCGCTCGGTCGTTCGGCTGCGGCGAGCGGTATCAGCTCACTCAAAGGC
GGTAATACGGTTATCCACAGAATCAGGGGATAACGCAGGAAAGAACATGTGAGCAAAAGGCCAGCAAAAGGCCAGGAACCGTAAAAAGGC
CGCGTTGCTGGCGTTTTTCCATAGGCTCCGCCCCCCTGACGAGCATCACAAAAATCGACGCTCAAGTCAGAGGTGGCGAAACCCGACAGGA
CTATAAAGATACCAGGCGTTTCCCCCTGGAAGCTCCCTCGTGCGCTCTCCTGTTCCGACCCTGCCGCTTACCGGATACCTGTCCGCCTTTCTCC
CTTCGGGAAGCGTGGCGCTTTCTCATAGCTCACGCTGTAGGTATCTCAGTTCGGTGTAGGTCGTTCGCTCCAAGCTGGGCTGTGTGCACGAA
CCCCCCGTTCAGCCCGACCGCTGCGCCTTATCCGGTAACTATCGTCTTGAGTCCAACCCGGTAAGACACGACTTATCGCCACTGGCAGCAGCC
ACTGGTAACAGGATTAGCAGAGCGAGGTATGTAGGCGGTGCTACAGAGTTCTTGAAGTGGTGGCCTAACTACGGCTACACTAGAAGGACAG
TATTTGGTATCTGCGCTCTGCTGAAGCCAGTTACCTTCGGAAAAAGAGTTGGTAGCTCTTGATCCGGCAAACAAACCACCGCTGGTAGCGGT
GGTTTTTTTGTTTGCAAGCAGCAGATTACGCGCAGAAAAAAAGGATCTCAAGAAGATCCTTTGATCTTTTCTACGGGGTCTGACGCTCAGTGG
AACGAAAACTCACGTTAAGGGATTTTGGTCATGACATTAACCTATAAAAATAGGCGTATCACGAGGCCCTTTCGTCTTCAC

A3.2.10 pelB pSF-1479 SsNEDD8-AGT

CTCGAGAAATCATAAAAAATTTATTTGCTTTGTGAGCGGATAACAATTATAATAGATTCAATTGTGAGCGGATAACAATTTCACACAGAATTCA
TTAAAGAGGAGAAATTAACCATGAAATACCTGCTGCCGACGGCTGCTGCTGGTCTGCTGCTGCTGGCTGCCCAACCGGCGATGGCGAGCAA
GCATCACCATCATTCAGGCCATCACCATACCGGACACCACCATCATTCAGGCAGTCATCACCATTCCGGAATGCTGATTAAAGTTAAAACCCT
GACTGGTAAGGAAATTGAGATTGACATCGAACCTACCGACAAGGTTGAACGTATTAAAGAACGTGTGGAAGAAAAGGAAGGTATCCCGCCG
CAGCAACAGCGTCTGATCTACTCTGGTAAACAGATGAACGACGAAAAGACTGCGGCTGACTATAAAATCCAGGGTGGCTCCGTGCTGCACCT
GGTACTGGCCCTGCGTGGTGGCGCCGGTACCAAAACTGAAGAAGGTAAACTGGTAATCTGGATTAACGGCGATAAAGGCTATAACGGTCTC
GCTGAAGTCGGTAAGAAATTCGAGAAAGATACCGGAATTAAAGTCACCGTTGAGCATCCGGATAAACTGGAAGAGAAATTCCCACAGGTTG
CGGCAACTGGCGATGGCCCTGACATTATCTTCTGGGCACACGACCGCTTTGGTGGCTACGCTCAATCTGGCCTGTTGGCTGAAATCACCCCG
GACAAAGCGTTCCAGGACAAGCTGTATCCGTTTACCTGGGATGCCGTACGTTACAACGGCAAGCTGATTGCTTACCCGATCGCTGTTGAAGC
GTTATCGCTGATTTATAACAAAGATCTGCTGCCGAACCCGCCAAAAACCTGGGAAGAGATCCCGGCGCTGGATAAAGAACTGAAAGCGAAA
GGTAAGAGCGCGCTGATGTTCAACCTGCAAGAACCGTACTTCACCTGGCCGCTGATTGCTGCTGACGGGGGTTATGCGTTCAAGTATGAAAA
CGGCAAGTACGACATTAAAGACGTGGGCGTGGATAACGCTGGCGCGAAAGCGGGTCTGACCTTCCTGGTTGACCTGATTAAAAACAAACAC
ATGAATGCAGACACCGATTACTCCATCGCAGAAGCTGCCTTTAATAAAGGCGAAACAGCGATGACCATCAACGGCCCGTGGGCATGGTCCA
ACATCGACACCAGCAAAGTGAATTATGGTGTAACGGTACTGCCGACCTTCAAGGGTCAACCATCCAAACCGTTCGTTGGCGTGCTGAGCGCA
GGTATTAACGCCGCCAGTCCGAACAAAGAGCTGGCAAAAGAGTTCCTCGAAAACTATCTGCTGACTGATGAAGGTCTGGAAGCGGTTAATA
AAGACAAACCGCTGGGTGCCGTAGCGCTGAAGTCTTACGAGGAAGAGTTGGCGAAAGATCCACGTATTGCCGCCACCATGGAAAACGCCCA
GAAAGGTGAAATCATGCCGAACATCCCGCAGATGTCCGCTTTCTGGTATGCCGTGCGTACTGCGGTGATCAACGCCGCCAGCGGTCGTCAG
ACTGTCGATGAAGCCCTGAAAGACGCGCAGACTAATGGCACCGGTTGTTAAGGATCTCATCACCATCACCATCACTAAGCTTAATTAGCTGA
GCTTGGACTCCTGTTGATAGATCCAGTAATGACCTCAGAACTCCATCTGGATTTGTTCAGAACGCTCGGTTGCCGCCGGGCGTTTTTTATTGG
TGAGAATCCAAGCTAGCCATGAAAATAAACTGTCTGCTTACATAAACAGTAATACAAGGGGTGTTATGAGCCATATTCAACGGGAAACGTCT
TGCTCTAGGCCGCGATTAAATTCCAACATGGATGCTGATTTATATGGGTATAAATGGGCTCGCGATAATGTCGGGCAATCAGGTGCGACAAT
CTATCGATTGTATGGGAAGCCCGATGCGCCAGAGTTGTTTCTGAAACATGGCAAAGGTAGCGTTGCCAATGATGTTACAGATGAGATGGTCA
GACTAAACTGGCTGACGGAATTTATGCCTCTTCCGACCATCAAGCATTTTATCCGTACTCCTGATGATGCTTGGTTACTCACGACTGCGATCCC
CGGCAAAACAGCATTCCAGGTATTAGAAGAATATCCTGATTCAGGTGAAAATATTGTTGATGCGCTGGCAGTGTTCCTGCGCCGGTTGCATT
CGATTCCTGTTTGTAATTGTCCTTTTAACAGCGATCGCGTATTTCGTCTCGCTCAGGCGCAATCACGAATGAATAACGGTTTGGTTGATGCGA
GTGATTTTGATGACGAGCGTAATGGCTGGCCTGTTGAACAAGTCTGGAAAGAAATGCACAAACTTTTGCCATTCTCACCGGATTCAGTCGTC
ACTCATGGTGATTTCTCACTTGATAACCTTATTTTTGACGAGGGGAAATTAATAGGTTGTATTGATGTTGGACGAGTCGGAATCGCAGACCGA
TACCAGGATCTTGCCATCCTATGGAACTGCCTCGGTGAGTTTTCTCCTTCATTACAGAAACGGCTTTTTCAAAAATATGGTATTGATAATCCTG
ATATGAATAAATTGCAGTTTCATTTGATGCTCGATGAGTTTTTCTAAGAATTAATTCATGGGCAAATATTATACGCAAGGCGACAAGGTGCTG
ATGCCGCTGGCGATTCAGGTTCATCATGCCGTTTGTGATGGCTTCCATGTCGGCAGAATGCTTAATGAATTACAACAGTACTGCGATGAGTG
GCAGGGCGGGGCGTAATTTTTTTAAGGCAGTTATTGGTGCCCTTAAACGCCTGGGGTAATGACTCTCTAGCTTGAGGCATCAAATAAAACGA
AAGGCTCAGTCGAAAGACTGGGCCTTTCGTTTTATCTGTTGTTTGTCGGTGAACGCTCTCCTGAGTAGGACAAATCCGCCGCTCTAGATTACG
TGCAGTCGATGATAAGCTGTCAAACATGAGAATTGTGCCTAATGAGTGAGCTAACTTACATTAATTGCGTTGCGCTCACTGCCCGCTTTCCAG
TCGGGAAACCTGTCGTGCCAGCTGCATTAATGAATCGGCCAACGCGCGGGGAGAGGCGGTTTGCGTATTGGGGGCCAGGGTGGTTTTTCTT
TTCACCAGTGAGACGGGCAACAGCTGATTGCCCTTCACCGCCTGGCCCTGAGAGAGTTGCAGCAAGCGGTCCACGCTGGTTTGCCCCAGCA
GGCGAAAATCCTGTTTGATGGTGGTTAACGGCGGGATATAACATGAGCTGTCTTCGGTATCGTCGTATCCCACTACCGAGATATCCGCACCA
ACGCGCAGCCCGGACTCGGTAATGGCGCGCATTGCGCCCAGCGCCATCTGATCGTTGGCAACCAGCATCGCAGTGGGAACGATGCCCTCAT
TCAGCATTTGCATGGTTTGTTGAAAACCGGACATGGCACTCCAGTCGCCTTCCCGTTCCGCTATCGGCTGAATTTGATTGCGAGTGAGATATT
TATGCCAGCCAGCCAGACGCAGACGCGCCGAGACAGAACTTAATGGGCCCGCTAACAGCGCGATTTGCTGGTGACCCAATGCGACCAGATG
CTCCACGCCCAGTCGCGTACCGTCTTCATGGGAGAAAATAATACTGTTGATGGGTGTCTGGTCAGAGACATCAAGAAATAACGCCGGAACAT
TAGTGCAGGCAGCTTCCACAGCAATGGCATCCTGGTCATCCAGCGGATAGTTAATGATCAGCCCACTGACGCGTTGCGCGAGAAGATTGTGC
ACCGCCGCTTTACAGGCTTCGACGCCGCTTCGTTCTACCATCGACACCACCACGCTGGCACCCAGTTGATCGGCGCGAGATTTAATCGCCGC
GACAATTTGCGACGGCGCGTGCAGGGCCAGACTGGAGGTGGCAACGCCAATCAGCAACGACTGTTTGCCCGCCAGTTGTTGTGCCACGCG
GTTGGGAATGTAATTCAGCTCCGCCATCGCCGCTTCCACTTTTTCCCGCGTTTTCGCAGAAACGTGGCTGGCCTGGTTCACCACGCGGGAAAC
GGTCTGATAAGAGACACCGGCATACTCTGCGACATCGTATAACGTTACTGGTTTCACATTCACCACCCTGAATTGACTCTCTTCCGGGCGCTA
TCATGCCATACCGCGAAAGGTTTTGCACCATTCGATGGTGTCGGAATTTCGGGCAGCGTTGGGTCCTGGCCACGGGTGCGCATGATCTAGA
GCTGCCTCGCGCGTTTCGGTGATGACGGTGAAAACCTCTGACACATGCAGCTCCCGGAGACGGTCACAGCTTGTCTGTAAGCGGATGCCGG
GAGCAGACAAGCCCGTCAGGGCGCGTCAGCGGGTGTTGGCGGGTGTCGGGGCGCAGCCATGACCCAGTCACGTAGCGATAGCGGAGTGT
ATACTGGCTTAACTATGCGGCATCAGAGCAGATTGTACTGAGAGTGCACCATATGCGGTGTGAAATACCGCACAGATGCGTAAGGAGAAAA
TACCGCATCAGGCGCTCTTCCGCTTCCTCGCTCACTGACTCGCTGCGCTCGGTCGTTCGGCTGCGGCGAGCGGTATCAGCTCACTCAAAGGC
GGTAATACGGTTATCCACAGAATCAGGGGATAACGCAGGAAAGAACATGTGAGCAAAAGGCCAGCAAAAGGCCAGGAACCGTAAAAAGGC
CGCGTTGCTGGCGTTTTTCCATAGGCTCCGCCCCCCTGACGAGCATCACAAAAATCGACGCTCAAGTCAGAGGTGGCGAAACCCGACAGGA
CTATAAAGATACCAGGCGTTTCCCCCTGGAAGCTCCCTCGTGCGCTCTCCTGTTCCGACCCTGCCGCTTACCGGATACCTGTCCGCCTTTCTCC
CTTCGGGAAGCGTGGCGCTTTCTCATAGCTCACGCTGTAGGTATCTCAGTTCGGTGTAGGTCGTTCGCTCCAAGCTGGGCTGTGTGCACGAA
CCCCCCGTTCAGCCCGACCGCTGCGCCTTATCCGGTAACTATCGTCTTGAGTCCAACCCGGTAAGACACGACTTATCGCCACTGGCAGCAGCC
ACTGGTAACAGGATTAGCAGAGCGAGGTATGTAGGCGGTGCTACAGAGTTCTTGAAGTGGTGGCCTAACTACGGCTACACTAGAAGGACAG
TATTTGGTATCTGCGCTCTGCTGAAGCCAGTTACCTTCGGAAAAAGAGTTGGTAGCTCTTGATCCGGCAAACAAACCACCGCTGGTAGCGGT
GGTTTTTTTGTTTGCAAGCAGCAGATTACGCGCAGAAAAAAAGGATCTCAAGAAGATCCTTTGATCTTTTCTACGGGGTCTGACGCTCAGTGG
AACGAAAACTCACGTTAAGGGATTTTGGTCATGACATTAACCTATAAAAATAGGCGTATCACGAGGCCCTTTCGTCTTCAC

A3.3 Example GH Plasmid Map - *Tt*GH8 in pET28a

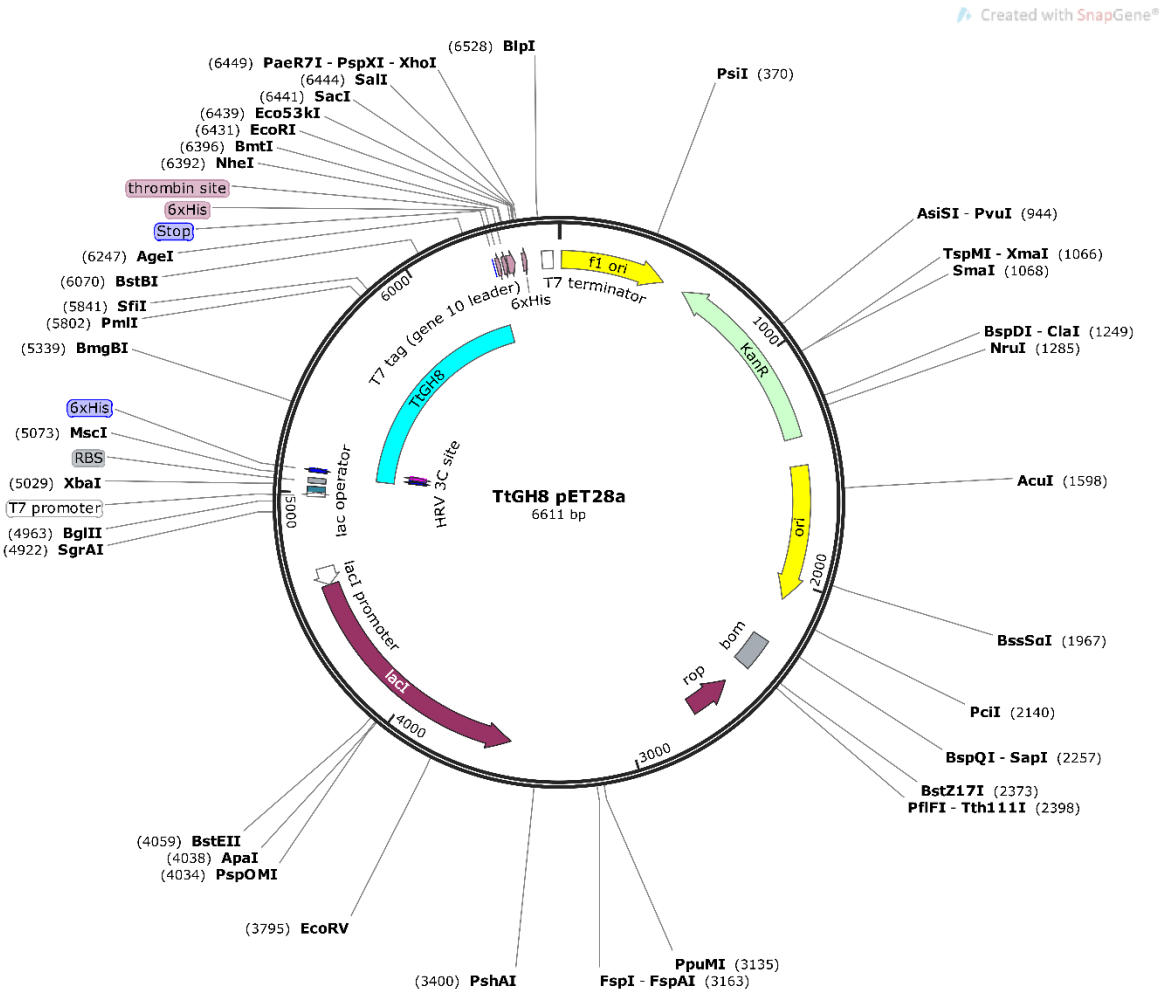


Figure 167 pET28a plasmid map showing position of TtGH8 as an example. Map created in Snapgene viewer.

A3.4 Original DNA source of *Tt*AA10: Champion SUMO vector

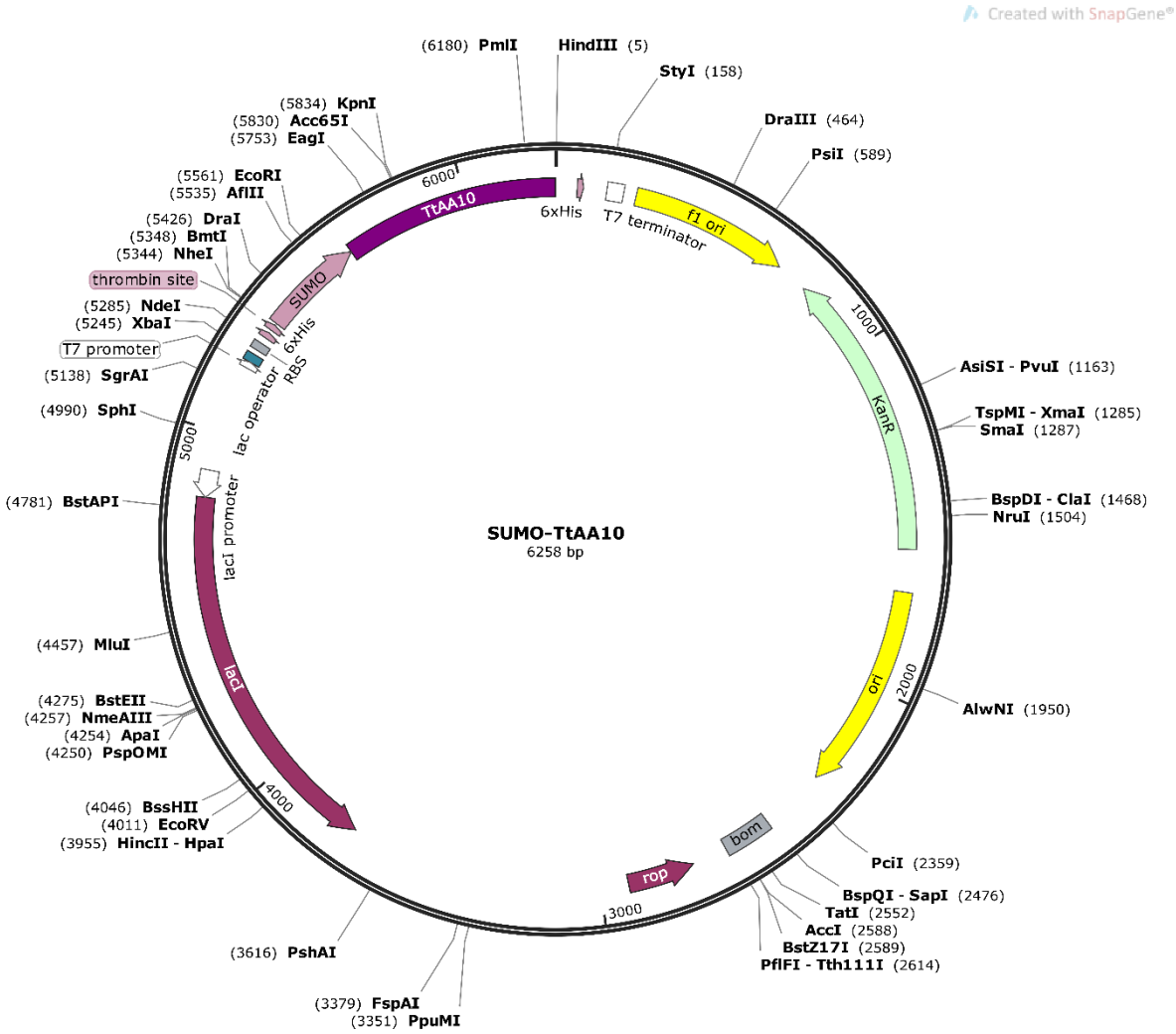


Figure 168 Plasmid map of SUMO-TtAA10 in Champion SUMO vector, made by Dr. G Hemsworth. Map created in Snapgene viewer

Appendix 4:

Published Work

Fowler, C. A., Hemsworth, G. R., Cuskin, F., Hart, S., Turkenburg, J., Gilbert, H. J., Walton, P. H. & Davies, G. J. (2018). Structure and function of a glycoside hydrolase family 8 endoxylanase from Teredinibacter turnerae, *Acta Cryst*. D74, 946-955. https://doi.org/10.1107/S2059798318009737

STRUCTURAL BIOLOGY

ISSN 2059-7983

Received 20 March 2018
Accepted 9 July 2018

Edited by C. S. Bond, University of Western Australia, Crawley, Australia

Keywords: glycoside hydrolase; biomass; marine polysaccharide; enzyme; shipworm; *Teredinibacter turnerae*.

PDB references: GH8 xylanase from *Teredinibacter turnerae*, 6g00; complex with xylobiose, 6g09; complex with xylotriose, 6g0b; catalytic mutant, complex with xylohexaose, 6g0n

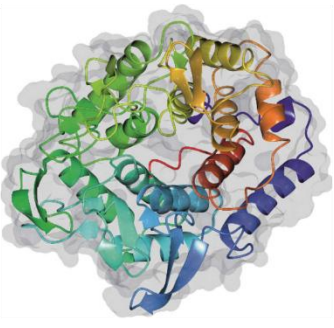
Supporting information: this article has supporting information at journals.iucr.org/d

Structure and function of a glycoside hydrolase family 8 endoxylanase from *Teredinibacter turnerae*

Claire A. Fowler,[a] Glyn R. Hemsworth,[b] Fiona Cuskin,[c] Sam Hart,[a] Johan Turkenburg,[a] Harry Gilbert,[d] Paul H. Walton[e] and Gideon J. Davies[a]*

[a]York Structural Biology Laboratory, Department of Chemistry, The University of York, York YO10 5DD, England, [b]School of Molecular and Cellular Biology, The Faculty of Biological Sciences, University of Leeds, Leeds LS2 9JT, England, [c]School of Natural and Environmental Science, Newcastle University, Newcastle upon Tyne NE1 7RU, England, [d]Institute for Cell and Molecular Biosciences, Newcastle University, Newcastle upon Tyne NE2 4HH, England, and [e]Department of Chemistry, The University of York, York YO10 5DD, England. *Correspondence e-mail: gideon.davies@york.ac.uk

The biological conversion of lignocellulosic matter into high-value chemicals or biofuels is of ncreasing industrial importance as the sector slowly transitions away from nonrenewable sources. Many industrial processes involve the use of cellulolytic enzyme cocktails – a selection of glycoside hydrolases and, increasingly, polysaccharide oxygenases – to break down recalcitrant plant polysaccharides. ORFs from the genome of *Teredinibacter turnerae*, a symbiont hosted within the gills of marine shipworms, were identified in order to search for enzymes with desirable traits. Here, a putative *T. turnerae* glycoside hydrolase from family 8, hereafter referred to as TtGH8, is analysed. The enzyme is shown to be active against β-1,4-xylan and mixed-linkage (β-1,3,β-1,4) marine xylan. Kinetic parameters, obtained using high-performance anion-exchange chromatography with pulsed amperometric detection and 3,5-dinitrosalicyclic acid reducing-sugar assays, show that TtGH8 catalyses the hydrolysis of β-1,4-xylohexaose with a k_{cat}/K_m of $7.5 \times 10^7\ M^{-1}\ min^{-1}$ but displays maximal activity against mixed-linkage polymeric xylans, hinting at a primary role in the degradation of marine polysaccharides. The three-dimensional structure of TtGH8 was solved in uncomplexed and xylobiose-, xylotriose- and xylohexaose-bound forms at approximately 1.5 Å resolution; the latter was consistent with the greater k_{cat}/K_m for hexasaccharide substrates. A $^{2,5}B$ boat conformation observed in the −1 position of bound xylotriose is consistent with the proposed conformational itinerary for this class of enzyme. This work shows TtGH8 to be effective at the degradation of xylan-based substrates, notably marine xylan, further exemplifying the potential of *T. turnerae* for effective and diverse biomass degradation.



OPEN ACCESS

1. Introduction

The demand for biofuels is increasing amid positive shifts in political and public opinion regarding the growing need for more sustainable fuel sources (discussed, for example, in Somerville, 2007; Pauly & Keegstra, 2008). A barrier towards the sustainable and efficient usage of plant biomass for fuel conversion lies in the complexity and recalcitrance of plant cell walls (Himmel *et al.*, 2007; Bomble *et al.*, 2017). The heterogeneous matrix of various carbohydrate compounds hinders the enzymatic breakdown of plant cell walls into energy-rich carbohydrate monomers. Crystalline cellulose regions are interspersed with a web of more soluble polysaccharides, known as hemicelluloses. The plant cell-wall matrix is strengthened by a hydrophobic and insoluble barrier: a mixture of phenolic compounds known as lignin (Li *et al.*,

2015). In nature, the breakdown of lignocellulosic material is achieved through the synergistic action of a wide variety of different enzymes, including glycoside hydrolases and polysaccharide oxygenases (Hemsworth *et al.*, 2015; Walton & Davies, 2016). A single organism can produce a consortium of enzymes capable of lignocellulosic degradation or can utilize the symbiotic behaviour of other smaller organisms such as bacteria and fungi (Cragg *et al.*, 2015). One organism of increasing interest is the marine symbiont *Teredinibacter turnerae*, which has been found in the gills (Fig. 1) of at least 24 species of bivalve molluscs (Ekborg *et al.*, 2007; Horak & Montoya, 2014).

Marine bivalve molluscs of the family Teredinidae, more commonly referred to as shipworms (Fig. 1), are considered pests by the maritime community as they inflict damage onto wooden structures such as piers and ship hulls by burrowing into the material (Cragg *et al.*, 2015). In 1927, Boynton and Miller observed the disappearance of 80% of cellulose and 15–56% of hemicellulose from Douglas fir piling during its transit through the digestive tract of *Teredo navalis* (Boynton & Miller, 1927), highlighting the ability of these organisms to degrade recalcitrant plant biomass. Much later, in 1983, Waterbury and coworkers observed, isolated and cultured the symbiotic bacterium *T. turnerae* (Waterbury *et al.*, 1983) from specialized cells called bacteriocytes found inside an internal region of the shipworm's gills. In its genome, *T. turnerae* encodes >100 glycoside hydrolases, enzymes that are capable of breaking glycosidic bonds, 54% of which have been classed as 'wood-specific' (Yang *et al.*, 2009). Furthermore, *T. turnerae* enzymes have been identified in the shipworm gut (O'Connor *et al.*, 2014), thus establishing the resident bacteria and the enzymes that they produce as key players in shipworm biology. It is therefore now clear that lignocellulose degradation may occur through the synergistic action of host enzymes and enzymes from this community of endosymbiotic bacteria housed within the gills of the shipworm.

In this light, we sought to mine the uncharacterized *T. turnerae* glycoside hydrolases to find enzymes that would be active under more biotechnologically relevant conditions and on useful substrates. *T. turnerae* exists primarily in high-salt conditions, and so its enzymes may be additionally stabilized towards harsher environments such as might be experienced in a biorefinery. As a proof of principle, we chose to focus on the consortia of enzymes harnessed by this organism to degrade xylans, given the interest in this substrate in the context of biofuels (Somerville, 2007; Pauly & Keegstra, 2008; Biely *et al.*, 2016), and also in the pulp and paper, animal feed and bread-making sectors. Additionally, following the recent insights gained from studying xylan-degrading enzyme consortia from new ecological niches (Rogowski *et al.*, 2015), work that revealed both new enzymes and specificities, the study of xylan-degrading enzymes from this shipworm symbiont represents fertile new ground for further discoveries in this area.

Applying the CAZY classification for carbohydrate-active enzymes (http://www.cazy.org; see Lombard *et al.*, 2014) to the *T. turnerae* genome reveals many potential xylanases, including ten GH10 enzymes, five GH11 enzymes and two enzymes from the potential xylanase-containing family GH30 (all of these glycoside hydrolase families are reviewed in CAZYpedia; for a review, see The Cazypedia Consortium, 2018). We were particularly drawn to family GH8, a cellulase/xylanase-containing family, from which *T. turnerae* has just a single representative, TtGH8. We sought to use this enzyme as an exemplar for whether *T. turnerae* can provide enzymes with biological and chemical utility, cast in terms of three-dimensional structure and notably, in this case, in the context of a diverse array of other potential xylanolytic enzymes.

Here, we report the cloning, expression and characterization of this GH8 carbohydrate-active enzyme (Lombard *et al.*, 2014). We show that TtGH8 is a xylan-active endoxylanase with six catalytically relevant subsites and notably a maximal activity towards mixed-linkage (β-1,3,β-1,4) marine xylan. The three-dimensional structure of TtGH8, determined to 1.5 Å resolution, reveals intimate details of the substrate-binding sites and the distortions of xylose within the catalytic centre. The work thus highlights the potential of the *T. turnerae* genome for enzyme discovery and adds to the growing toolbox of enzymes that may be used to tackle the recalcitrant hemicellulose xylan (Biely *et al.*, 2016).

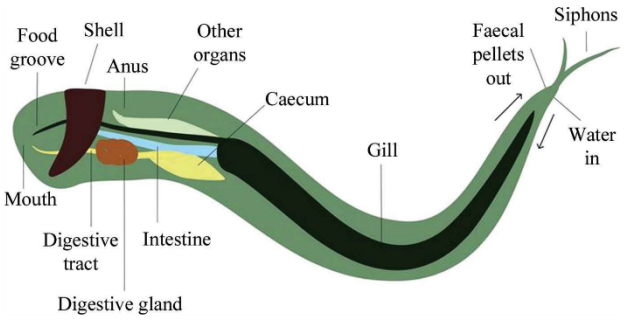


Figure 1
Basic diagram of a shipworm (image not to scale) showing the general layout of the main organs within the animal. More information on shipworm physiology can be found in Betcher (2011), Gallager *et al.* (1981) and Waterbury *et al.* (1983), and in-depth illustrations and descriptions can be found in Turner (1966).

2. Materials and methods

2.1. Purification of TtGH8 and catalytic mutants

The sequence of TtGH8 (TERTU_4506; UniProt C5BJ89) was analysed and the catalytic domain was identified (Supplementary Table S1) and truncated to remove the original signal peptide and linker region and to include an N-terminal hexahistadine tag and 3C protease cleavage site. The sequence was codon-optimized and produced for *Escherichia coli* by GenScript (New Jersey, USA) and the plasmid was transformed into *E. coli* BL21 competent cells (Supplementary Table S2). Catalytic residues were identified using the literature and structures of similar GH8 proteins deposited in the PDB. The mutations were designed using custom primers and implemented using the Q5 site-directed

mutagenesis kit (NEB) to alter Asp281 to Asn (TtGH8 D281N; Supplementary Table S3). Expression testing for both constructs was carried out prior to large-scale production. Cultures (6 × 500 ml LB, 30 mg ml^{-1} kanamycin) were inoculated with 500 µl overnight culture and grown at 37°C and 200 rev min^{-1} until an OD of 0.6 was obtained. IPTG (1 m*M* final concentration) was then added and the cultures were cooled to 16°C and left shaking overnight. The cultures were harvested by centrifugation and the pellet was resuspended in 50 m*M* HEPES, 250 m*M* NaCl, 30 m*M* imidazole pH 7. The cells were lysed by sonication and centrifuged at 15*g* for 30 min. The supernatant was collected and loaded onto a pre-equilibrated Nickel HiTrap Crude 5 ml affinity column. Nickel-affinity chromatography was run on an ÄKTA start, with an elution gradient of 30–300 m*M* imidazole over 25 column volumes. The collected protein sample was treated with a 1:100 ratio of 3C protease to TtGH8 and 5 m*M* DTT and left shaking at room temperature overnight. The sample was loaded onto a pre-equilibrated Nickel HiTrap Crude 5 ml affinity column and the flowthrough and wash were collected. The collected sample was buffer-exchanged into 20 m*M* HEPES, 200 m*M* NaCl before being concentrated to approximately 300 µl and loaded onto a Superdex 200 gel-filtration column. Pure fractions were combined, concentrated and buffer-exchanged into 10 m*M* HEPES. To check their purity, samples were analysed by SDS–PAGE throughout purification and the final sample was analysed by electrospray ionization mass spectrometry.

2.2. Thermal shift analysis (TSA)

Samples (30 µl total) containing SYPRO Orange dye (15 µl) and enzyme (final concentration of 1 mg ml^{-1}) with either buffer or substrate were prepared and thermal shift analyses were run on a Stratagene Mx3005P qPCR device. The substrates tested included the solid substrates (small amount, unmeasured) tamarind xyloglucan, rye arabinoxylan, birchwood xylan, shrimp shell chitin and Avicel and the soluble oligosaccharides cellohexaose, cellobiose, xylobiose and xylohexaose (10 m*M*). Samples were heated from 20 to 91°C in increments of 1°C over 71 cycles. The fluorescence of SYPRO Orange was monitored throughout and the data were used to calculate the melting temperature of the protein (Supplementary Fig. S1). The data were analysed using the *JTSA* fitting program developed by Paul Bond, which is available at http://paulsbond.co.uk/jtsa/#/input.

2.3. Thin-layer chromatography (TLC) and liquid chromatography–mass spectrometry (LCMS)

Xylo-oligosaccharides were purchased commercially: xylobiose (X2) from Sigma and TCP, and xylotriose (X3), xylotetraose (X4), xylopentaose (X5), xylohexaose (X6) and polysaccharides from Megazyme (unless stated otherwise). Overnight hydrolysis reactions with the xylo-oligosaccharides X2–X6 (1 m*M*), wheat arabinoxylan (WAX), rye arabinoxylan (RAX), corn xylan (CAX), birchwood xylan (BX), mixed-linkage β1–3,β1–4 xylan (MLX, purchased from Elicityl/Oligotech) and xyloglucan (XG) at 1 mg ml^{-1} were incubated at 37°C with 1 µ*M* TtGH8. The samples were heated at 90°C prior to spotting onto a TLC plate (total 4 µl). Standards containing X2–X6 at 1 m*M* each were run on the same plate. TLC plates were placed in chromatography tanks containing the running buffer [50%(*v*/*v*) *n*-butanol, 25%(*v*/*v*) acetic acid, 25%(*v*/*v*) water]. Plates were run once, dried and then re-run to improve the separation of sugars. The plates were visualized using a staining solution [3%(*v*/*v*) sulfuric acid, 75%(*v*/*v*) ethanol, 0.1%(*w*/*v*) orcinol monohydrate], dried and then heated to approximately 100°C (Fig. 2*a*). Hydrolysis samples for LCMS were prepared using 50 m*M* ammonium acetate buffer pH 6, approximately 1 mg ml^{-1} substrate and 1 µ*M* enzyme. Samples were incubated at 37°C overnight and shaken at 500 rev min^{-1}. If required, samples were centrifuged to remove any solid materials and 100 µl was loaded onto a Cosmosil Sugar-D HPLC column using the LC-MS Dionex system, where the separated products were analysed by ESI or PAD mass spectrometry. Running buffers were a mixture of water and acetronitrile, with some test runs also including 1% formic acid (Supplementary Figs. S2 and S3).

2.4. Kinetics measurements using high-performance anion-exchange chromatography with pulsed amperometric detection (HPAEC-PAD)

Substrate-depletion kinetics measurements were performed on TtGH8 with xylohexaose, xylopentaose and xylotetraose and were measured using an HPAEC-PAD Dionex system. Hydrolysis reactions were run at 37°C using different substrate and enzyme concentrations, with aliquots removed at set time points and boiled to inactivate TtGH8. All samples were mixed with a fucose internal standard and run on an anion-exchange column (CarboPac) using a sodium acetate gradient. The depletion of substrate during the reaction can be related to k_{cat}/K_m through

$$kt = \ln\left(\frac{S_0}{S_t}\right), \tag{1}$$

where $k = (k_{cat}/K_m)$[enzyme], t is time and S_0 and S_t are the substrate-peak areas at time 0 and t, respectively. The substrate-peak areas observed in the HPAEC PAD traces were normalized against both an external and internal fucose standard and the resulting values for $\ln(S_0/S_t)$ were plotted against time, producing positive gradients (the change in the substrate-peak area increases from 0, as the substrate-peak area at $t = 0$ is at its maximum). Linear regression analysis was used to measure the gradient that represents the rate of reaction, and k_{cat}/K_m (M^{-1} min^{-1}) was determined by dividing this gradient by the enzyme concentration (Figs. 2*b*, 2*c* and 2*d*). The use of substrate depletion to obtain k_{cat}/K_m has been widely used in the glycoside hydrolase field (see, for example, Pell *et al.*, 2004; Charnock *et al.*, 1998; Matsui *et al.*, 1991).

2.5. Reducing-sugar assay

The activity of TtGH8 towards several polysaccharides was tested using the 3,5-dinitrosalicyclic acid (DNSA) reducing-

sugar assay. Hydrolysis reactions with enzyme and substrate were run at 37°C and 150 µl aliquots were removed and immediately mixed in a 1:1 volume ratio with the DNSA agent [comprising 1%(*w*/*v*) DNSA, 0.2%(*v*/*v*) phenol, 1%(*w*/*v*) NaOH, 0.002% glucose and 0.05%(*w*/*v*) Na_2SO_3] at set time points to stop the reaction. The aliquots were heated at 90°C for 20 min and cooled on ice for 10 min. The absorbance of each sample was measured at 575 nm. A standard curve of 0–500 µg ml^{-1} xylose (plus 1 mg ml^{-1} polysaccharide substrate) was used to quantify the released reducing sugar. Reaction rates determined for each different substrate-concentration condition were plotted against the substrate concentration, where the gradient divided by the enzyme concentration is k_{cat}/K_m (Fig. 3).

2.6. Crystallization and X-ray crystallography of TtGH8 and mutants

Initial crystallization screening was performed robotically using a Mosquito crystal robot and commercial screens including Crystal Screen HT and Index from Hampton Research and PACT from Molecular Dimensions. Several crystal hits were obtained for TtGH8 and TtGH8 D281N. A 24-well optimization screen containing 0.1 *M* sodium acetate pH 4.6–5.2, 0.2 *M* NaCl, 14–24% polyethylene glycol (PEG) 6000 was used to produce the final crystallization condition for TtGH8: 0.1 *M* sodium acetate pH 5.0, 0.2 *M* NaCl, 16% polyethylene glycol. Thin rod-shaped crystals were harvested in nylon loops and cryoprotected by soaking them in mother liquor plus 30%(*v*/*v*) ethylene glycol. The TtGH8 crystals containing xylobiose and xylotriose were both soaked for 30 s in a solution consisting of the well solution, 30% ethylene glycol and 150 m*M* xylobiose/xylotriose. TtGH8 mutant crystals were soaked in 20 m*M* xylohexose (0.1 *M* HEPES pH 6.8, 0.2 *M* ammonium sulfate, 20% PEG 6000) for 15 min to produce a product complex and for 10 s to produce a substrate complex and were dipped in a cryosolution consisting of 30% ethylene glycol before cooling. Crystal data sets were collected on beamlines I02 and I04 at Diamond Light Source.

Data sets were processed using *xia2*/*DIALS* (Winter, 2010; Winter *et al.*, 2018), with the outer resolution limits defined by $CC_{1/2} > 0.5$. Molecular replacement (*Phaser*; McCoy *et al.*, 2007) and refinement (*REFMAC*; Murshudov *et al.*, 2011) were carried out using the *CCP4i2* pipeline (Potterton *et al.*, 2018). The TtGH8 structure was solved by molecular replacement using the *CCP4* implementation of *MOLREP* (Vagin

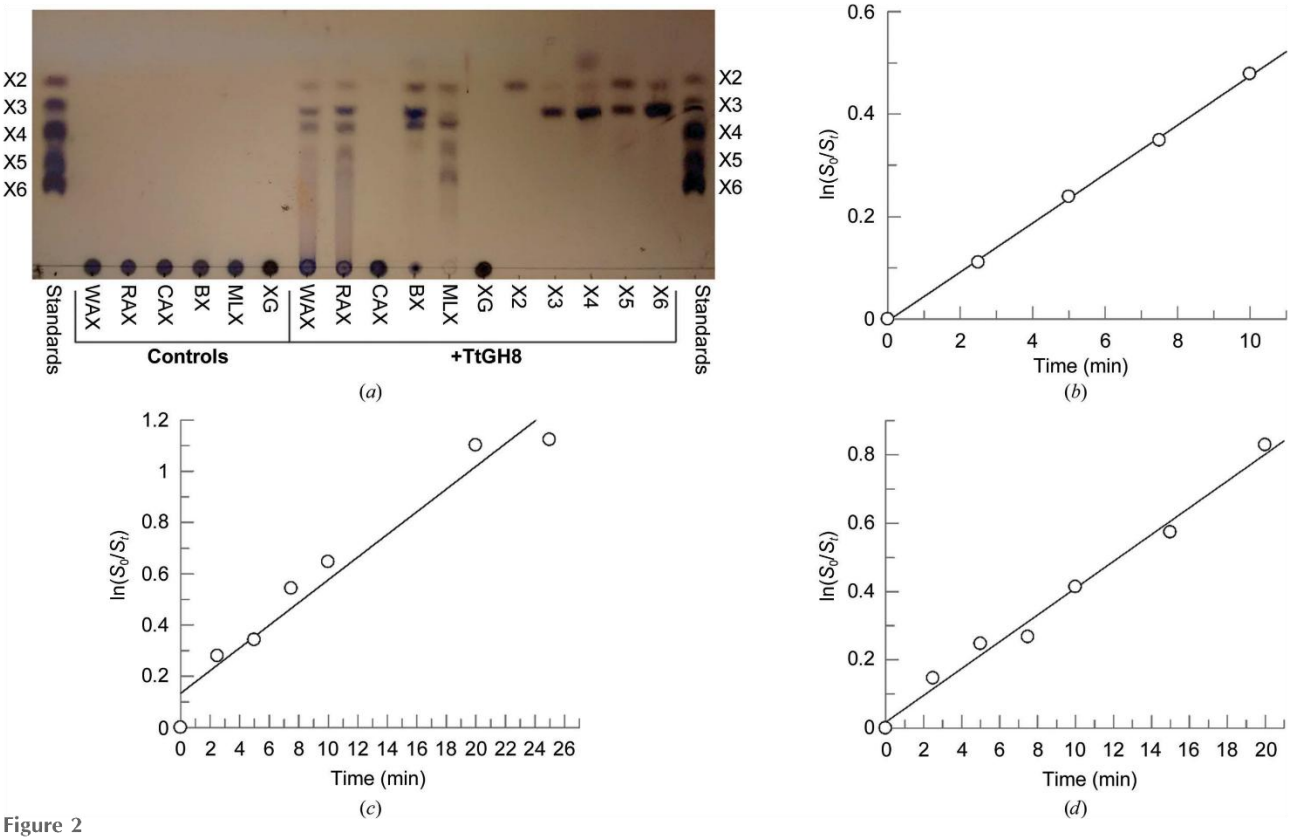


Figure 2
(*a*) TtGH8 activity was explored on a range of substrates using thin-layer chromatography. Hydrolysis activity can be observed on WAX, RAX, BX and MLX, as well as on the oligosaccharides xylotetraose (X4), xylopentaose (X5) and xylohexaose (X6). Activity analysis of TtGH8 on xylohexaose (*b*), xylopentaose (*c*) and xylotetraose (*d*) was carried out with HPAEC-PAD by substrate depletion. The peak areas observed in the HPAEC-PAD traces were normalized against a fucose internal standard and values for $\ln(S_0/S_t)$, where S_0 is the substrate-peak area at time 0 and S_t is the substrate-peak area at time t, were plotted against time. Linear regression analysis was used to calculate k_{cat}/K_m (in M^{-1} min^{-1}).

& Teplyakov, 2010) and default parameters with the protein atoms only of PDB entry 1wu4, the GH8 reducing-end-xylose releasing exo-oligoxylanase from *Bacillus halodurans* C-125 (Fushinobu *et al.*, 2005). Manual manipulation in *Coot* (Emsley *et al.*, 2010) followed by refinement using *REFMAC* was cycled several times to a give a final R factor and R_{free} of 0.15 and 0.17, respectively. Protein–ligand complex structures were solved by molecular replacement using the *CCP*4 implementation of *MOLREP* (Vagin & Teplyakov, 2010) with the unliganded TtGH8 enzyme structure as the search model. Ligands were modelled using *JLigand* (Lebedev *et al.*, 2012). Structural analysis and figure preparation was carried out in *CCP*4*mg* (McNicholas *et al.*, 2011).

3. Results

3.1. Catalytic activity of TtGH8

T. turnerae possesses a wide variety of glycoside hydrolases spanning many different families and potential substrate specificities. Given its possible relevance to xylan degradation and perhaps even marine colonization of wood, DNA encoding the predicted catalytic domain (residues 41–436; Supplementary Table S1) of the GH8 enzyme of the bacterium (TtGH8; locus tag TERTU_4506) was cloned into an *E. coli* expression vector. The gene sequence was codon-optimized by GenScript (DNA sequence given in Supplementary Table S2) and designed with a 3C protease-removable N-terminal hexahistidine tag. *E. coli* containing the recombinant plasmid produced high levels of soluble TtGH8, which was purified using nickel-affinity chromatography and gel-filtration chromatography. The identity of pure TtGH8, as judged by SDS–PAGE, was confirmed by electrospray ionization mass spectrometry (ESI-MS, data not shown).

Initial screening for likely binding ligands/substrates was performed using thermal shift analysis and a panel of oligo/polysaccharides (xyloglucan, rye arabinoxylan, birchwood xylan, shrimp shell chitin, Avicel, cellohexaose, cellobiose, xylobiose and xylohexaose). Significant increases in melting temperature were only observed in conditions containing xylan or xylohexaose, which resulted in a 2.1°C (apo 55.2°C, with xylan 57.3°C) and 2.9°C (apo 57.2°C, with xylohexaose

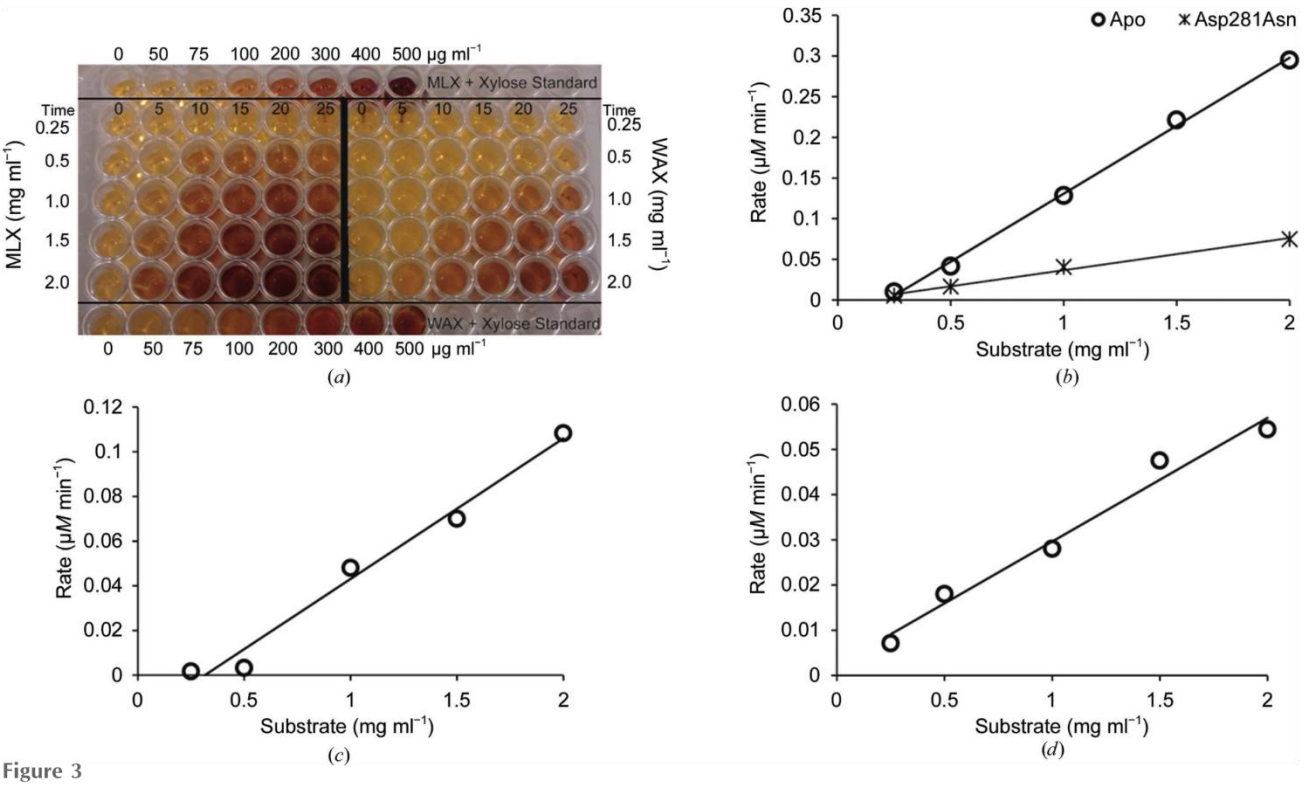


Figure 3
TtGH8 activity on polysaccharides was determined using the 3,5-dinitrosalicyclic acid (DNSA) reducing-sugar assay. (*a*) Photograph of the plate containing aliquots representing different reaction time points (0–25 min) and substrate concentrations [mixed-linkage xylan (MLX) and wheat arabinoxylan (WAX) at 0.5–2.0 mg ml^{-1}]. Reaction of the DNSA agent with reducing sugars causes a visible colour change from light yellow to deep brown, with the amount of colour change measured by the absorbance at 510 nm being equivalent to the amount of reducing sugar (chain ends) produced during the enzyme reaction. Standard curves were calculated using the free xylose controls in the top and bottom rows of the plate. These were MLX (1 mg ml^{-1}) + xylose (0–500 µg ml^{-1}) and WAX (1 mg ml^{-1}) + xylose (0–500 µg ml^{-1}). TtGH8 is more active on MLX than on WAX as there is a greater colour change. Absorption measurements taken for each reaction aliquot can be converted into the amount of reducing sugar present using the standard curve [shown in (*a*)] and plotted against time to give a rate (µ*M* min^{-1}, not shown). The rates for each reaction condition are plotted against the substrate concentration, where the gradient divided by the enzyme concentration is equivalent to k_{cat}/K_m. (*b*) Activity of TtGH8 and TtGH8 catalytic mutant on MLX, (*c*) activity of TtGH8 on WAX and (*d*) activity of TtGH8 on BX.

60.1°C) increase in melting temperature, respectively (Supplementary Fig. S1).

Given the known specificities of family GH8 members (available at http://www.cazy.org/GH8.html), chitosanase (EC 3.2.1.132), cellulase (EC 3.2.1.4), licheninase (EC 3.2.1.73), endo-1,4-β-xylanase (EC 3.2.1.8) and reducing-end-xylose releasing exo-oligoxylanase (EC 3.2.1.156), the increase in melting temperature for TtGH8 suggested activity as a xylanase. TtGH8 was therefore incubated (18 h) with a variety of potential substrates and the reaction products were monitored by thin-layer chromatography (Fig. 2*a*). TtGH8 was tested against wheat arabinoxylan (WAX), rye arabinoxylan (RAX), corn arabinoxylan (in the aleurone layer coating the seeds known as the bran fraction; CAX), birchwood xylan (BX), mixed-linkage xylan (MLX), xyloglucan (XG) and also β-1,4-linked xylo-oligosaccharides with degrees of polymerization (DP) from 2 to 6. TLC clearly demonstrated that TtGH8 acts as a xylanase, with xylotriose (X3) the predominant product and with activity on WAX, RAX, BX and MLX clearly evident. Xylotetraose (X4) was the minimal length xylooligosaccharide that could act as a substrate for TtGH8 observable by TLC. The TLC results were confirmed using liquid chromatography linked to mass spectrometry (Supplementary Fig. S2). TtGH8 was observed to cleave X6 into X3, X5 into X3 and X2, and X4 into X3 (X1 could not be observed). Analysis of the soluble products removed by centrifugation after incubation of the protein with birchwood xylan (1 mg ml^{-1}, 37°C, 18 h) indeed confirmed X3 as the dominant product (Supplementary Fig. S3).

With the knowledge that TtGH8 acts as a xylanase, we next sought to determine the kinetic parameters for the action of TtGH8 both by substrate-depletion analysis using high-performance anion-exchange chromatography with pulsed amperometric detection (HPAEC-PAD) and by reducing-sugar assays with 3,5-dinitrosalicyclic acid (DNSA). HPAEC-PAD analysis (Figs. 2*b* and 2*c*) yielded a maximal activity, k_{cat}/K_m, of 7.5×10^7 M^{-1} min^{-1} for X6, with X5 and X4 being considerably worse substrates, with k_{cat}/K_m values some five and 123 times lower, respectively (Table 1). Such values are typical for endo-xylanases, falling as they do between the k_{cat}/K_m values reported for GH10 enzymes such as CjXyn10A and CyXyn10C (Pell *et al.*, 2004). TtGH8 activity against xylose-based polysaccharides was determined using the DNSA reducing-sugar assay (Fig. 3*a*). Enzyme activity was monitored over time at five different substrate concentrations, allowing the determination of k_{cat}/K_m (Figs. 3*b*, 3*c* and 3*d*). An apparent preference for MLX was observed (1.6×10^8 mg^{-1} ml min^{-1}; Table 1).

Table 1
Catalytic activities of TtGH8 on xylan substrates as determined by HPAEC-PAD (oligosaccharides) or DNSA reducing-sugar assay (polysaccharides).

Substrate	k_{cat}/K_m†
Oligosaccharides	
Xylohexaose	$7.5 \times 10^7 \pm 1.1 \times 10^6$
Xylopentaose	$1.4 \times 10^7 \pm 1.4 \times 10^6$
Xylotetraose	$6.1 \times 10^5 \pm 3.1 \times 10^4$
Polysaccharides	
BX	$1.8 \times 10^7 \pm 4 \times 10^6$
WAX	$6.3 \times 10^6 \pm 5 \times 10^5$
MLX (wild type)	$1.6 \times 10^8 \pm 4 \times 10^6$
MLX (Asp281Asn)	$1.8 \times 10^4 \pm 1 \times 10^3$

† Values for oligosaccharides are given in M^{-1} min^{-1} and those for polysaccharides are given in mg^{-1} ml min^{-1}.

3.2. Three-dimensional structure of TtGH8

In order to provide molecular insight into its catalytic properties, as described above, the three-dimensional structure of TtGH8 was determined. TtGH8 was initially crystallized in an apo (unliganded) form and the structure was determined by molecular replacement at a resolution of 1.4 Å (Tables 2 and 3), using the protein coordinates only from the structure of the GH8 reducing-end-xylose releasing exo-oligoxylanase from *B. halodurans* C-125 (PDB entry 1wu4; Fushinobu *et al.*, 2005) as the search model.

The first three-dimensional structure of a CAZY family GH8 member was that of *Clostridium thermocellum* CelA (Alzari *et al.*, 1996). Consistent with the defining family member, TtGH8 exhibits a classical $(\alpha/\alpha)_6$ fold with a clear deep substrate-binding groove (Fig. 4*a*). *PDBeFold* (Krissinel & Henrick, 2004) analysis reveals that TtGH8 shares close structural similarity to the reducing-end-xylose releasing exo-oligoxylanase from *B. halodurans* C-125 (Fushinobu *et al.*, 2005). TtGH8 shares 43% sequence identity, overlaying 344 C^α atoms with an r.m.s.d. of 1.1 Å, a *Q*-score of 0.7 and a *Z*-score of 15.9. The next closest match is the cold-adapted GH8 xylanase from *Pseudoalteromonas haloplanktis* (Collins *et al.*, 2005) with 35% sequence identity, an r.m.s.d. of 1.4 Å over 341 C^α atoms, a *Q*-score of 0.6 and a *Z*-score of 14.7.

3.3. Mechanism of GH8 endoxylanases

Glycoside hydrolases may act through one of two main mechanisms, leading to retention or inversion of anomeric configuration (reviewed, for example, in Davies & Henrissat, 1995; Henrissat & Davies, 1997; Rye & Withers, 2000). CAZY GH8 enzymes act with inversion of anomeric configuration (Fierobe *et al.*, 1993) and thus feature two essential catalytic residues: a base to activate the water for nucleophilic attack and an acid to protonate the leaving group for departure. Previous studies on GH8 enzymes have identified a conserved glutamic acid found on helix α2 which functions as the catalytic acid in the inverting mechanism (Alzari *et al.*, 1996). The equivalent in TtGH8 is Glu73. As in many inverting enzyme families, the location of the base is less clear, and it has been proposed that GH8 may, in fact, be subdivided into groupings based upon the location of the catalytic base (Adachi *et al.*, 2004). The classical xylanases and endoglucanases (as first reported by Alzari *et al.*, 1996) are believed to have the base at the end of helix α8, with the equivalent residue in TtGH8 being Asp281. In order to clarify the catalytic residues in TtGH8, crystals were soaked in both xylobiose (X2) and xylotriose (X3) and the resulting structures were refined at 1.4 and 1.8 Å resolution, respectively (Tables 2 and 3).

Table 2
Data-collection and processing statistics for TtGH8 and the TtGH8 mutant with and without ligands.

Values in parentheses are for the outer shell.

	Native TtGH8	TtGH8–X2	TtGH8–X3	TtGH8 mutant–X6
Diffraction source	Diamond Light Source	Diamond Light Source	Diamond Light Source	Diamond Light Source
Wavelength (Å)	0.98	0.98	0.98	0.98
Temperature (K)	100	100	100	100
Space group	$P2_12_12_1$	$P2_12_12_1$	$P2_12_12_1$	$P2_12_12_1$
a, b, c (Å)	61.5, 73.0, 90.9	61.7, 79.0, 87.6	59.7, 80.5, 88.0	62.0, 79.7, 88.0
α, β, γ (°)	90.0, 90.0, 90.0	90.0, 90.0, 90.0	90.0, 90.0, 90.0	90.0, 90.0, 90.0
Resolution range (Å)	51.0–1.40 (1.42–14.0)	50.4–1.40 (1.42–1.40)	59.4–1.80 (1.84–1.80)	60.97–1.80 (1.84–1.80)
Total No. of reflections	342816	530153	316157	313138
No. of unique reflections	80610	84123	40092	40046
Completeness (%)	99.4 (99.9)	99.1 (89.6)	100.0 (100.0)	99.0 (98.0)
Multiplicity	4.3 (4.3)	6.3 (4.1)	7.9 (7.7)	7.8 (8.0)
$\langle I/\sigma(I)\rangle$	9.6 (1.7)	10.1 (1.3)	8.5 (1.6)	8.2 (2.0)
$CC_{1/2}$	0.997 (0.670)	0.997 (0.503)	0.994 (0.528)	0.992 (0.639)
$R_{p.i.m.}$	0.045 (0.414)	0.051 (0.549)	0.080 (0.611)	0.109 (0.970)
Overall B factor from Wilson plot (Å^2)	9	9	13	12

Table 3
Structure solution and refinement of TtGH8 and the TtGH8 mutant with and without ligands.

Values in parentheses for the outer shell.

	Native TtGH8	TtGH8–X2	TtGH8–X3	TtGH8 mutant–X6
Resolution range (Å)	51.0–1.40 (1.42–1.40)	50.4–1.40 (1.42–1.40)	59.4–1.80 (1.84–1.80)	60.97–1.80 (1.84–1.80)
Completeness (%)	99.4 (99.9)	99.1 (89.6)	100.0 (100.0)	99.0 (98.0)
No. of reflections, working set	80540	84054	40032	39993
No. of reflections, test set	3924	4274	1868	1942
Final R_{cryst}	0.15	0.16	0.17	0.17
Final R_{free}	0.17	0.18	0.20	0.20
Cruickshank DPI	0.052	0.051	0.111	0.116
No. of non-H atoms				
Protein	3163	3145	3125	3136
Ion	1	—	—	—
Ligand	25	23	28	61
Water	310	330	164	194
R.m.s. deviations				
Bonds (Å)	0.017	0.016	0.012	0.012
Angles (°)	1.70	1.68	1.51	1.50
Average B factors (Å^2)				
Protein	14	12	17	16
Ion	18	—	—	—
Ligand	28	15	20	26
Water	27	22	24	22
Ramachandran plot				
Most favoured (%)	98.2	97.5	98.0	97.5
Allowed (%)	1.8	2.5	2.0	2.5
PDB code	6g00	6g09	6g0b	6g0n

X2 was bound in the −2 and −3 subsites (Supplementary Fig. S4; subsite nomenclature is as discussed by Davies *et al.*, 1997), hinting at the strength of the −3 subsite, and consistent with the product profiles (Fig. 2*a*). X3 was observed bound in the −3 to −1 subsites (Fig. 4*b*) and confirms the catalytic residue proposals, with Glu73 (which has rotated from its position in the apo structure into this more catalytically relevant position) interacting with the O1 hydroxyl and with Asp281 interacting with a water molecule 'below' C1 in a position mimicking that which would be expected for hydrolysis with inversion of anomeric configuration. Notably, the −1 subsite sugar is not observed in its low-energy $^4C_1$ chair conformation, but is instead observed distorted into a $^{2,5}B$ conformation (Fig. 4*b*), consistent with proposals for the catalytic itinerary of GH8 enzymes discussed below.

Enzymatic glycoside hydrolysis involves the distortion of the reactive, −1 subsite, sugar into a variety of skew-boat and boat conformations, reflecting the requirements of inline attack and the stereoelectronic requirements of an oxocarbenium-ion-like transition state (for reviews, see Davies & Williams, 2016; Speciale *et al.*, 2014; Davies *et al.*, 2003, 2012). The GH8 family has been proposed to go through a $^{2,5}B$-like transition state, notably because of the observation of sugars with ^2S_O and $^{2,5}B$ conformations in complexes of the CelA endoglucanase (Guérin *et al.*, 2002) that were subsequently analysed by QM/MM metadynamics (Petersen *et al.*, 2009).

The previous X3 complex, that of the *P. haloplanktis* cold-adapted xylanase (Collins *et al.*, 2005; De Vos *et al.*, 2006), revealed binding in the +1 to +3 subsites, which together with the −3 to −1 observations here approximately defined

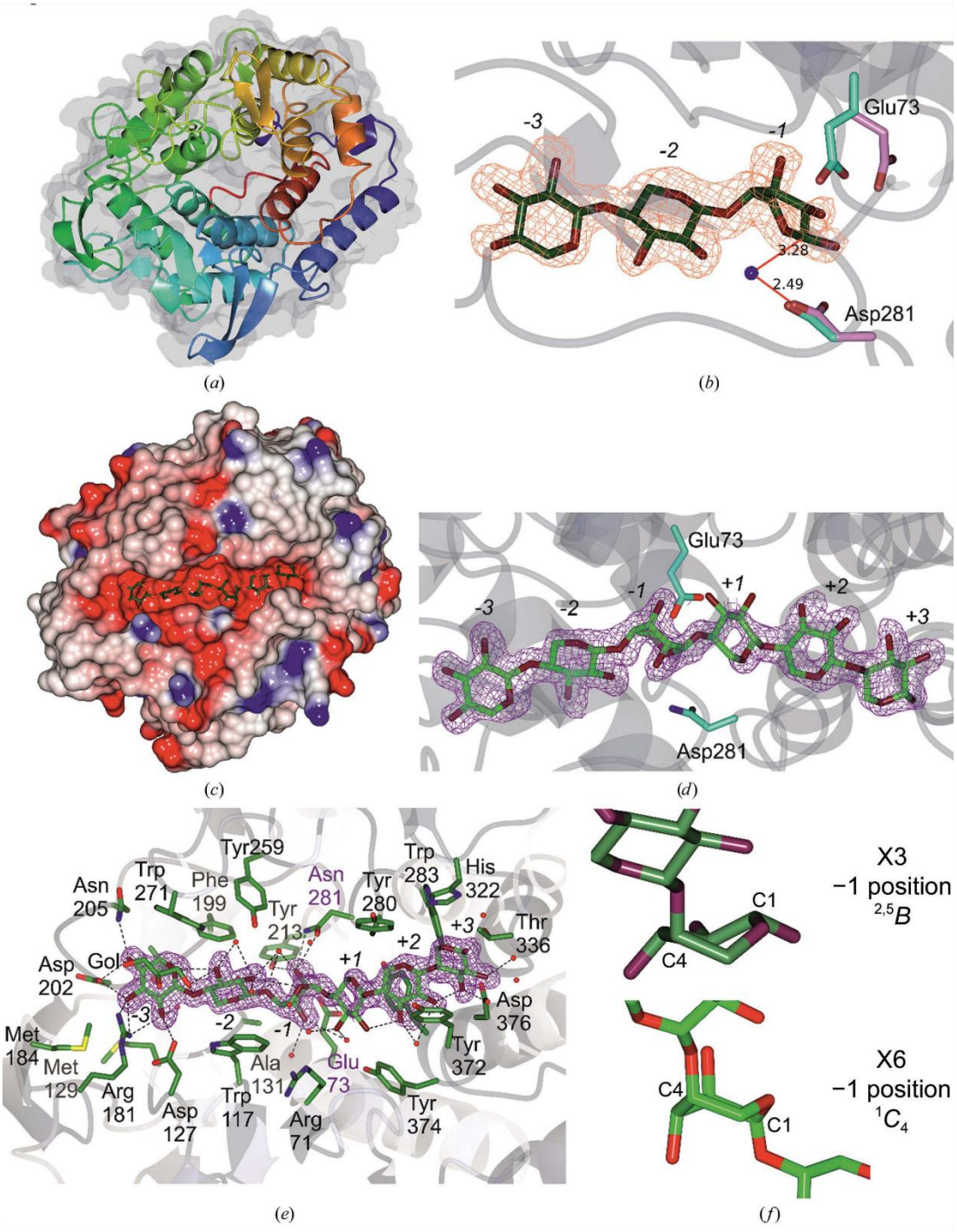


Figure 4

(*a*) Apo TtGH8 ribbon structure (coloured from blue to red) with the overall surface transparent, clearly showing the classical $(\alpha/\alpha)_6$ fold. (*b*) TtGH8–xylotriose complex, showing xylotriose bound in sites −3 to −1, with the corresponding maximum-likelihood-weighted $F_o - F_c$ 'difference' electron density, calculated prior to any incorporation of ligands in refinement, at a contour level of 0.35 e Å^{-3} (approximately 2.5σ). The catalytic residues from the apo form (light blue) and X3 complex (pink) are shown to highlight the position of the scission site and the shift in the catalytic acid (Glu73). The nucleophilic water is shown as a blue sphere, with bonding distances to Asp281 and the catalytic cleavage site (were the substrate longer) depicted. (*c*) TtGH8 D281N–X6 complex shown as a surface (coloured by electrostatic potential), in which xylohexaose can be observed spanning the deep binding cleft. (*d*) TtGH8 D281N–X6 complex shown with xylohexaose spanning the six binding positions −3 to +3 with the associated electron density (contour level 0.35 e$^-$ Å^{-3}; approximately 2.5σ). The catalytic acid and mutated base are shown in light blue. Here, it can be clearly seen that the position normally taken by the nucleophilic water (sitting below and in line with C1) is unavailable owing to the occupation of O2 caused by a ring flip. (*e*) TtGH8 D281N–X6 complex, showing the nearest neighbours of the X6 ligand (coloured by blending through model); waters are shown in blue, the cryoprotectant ethylene glycol (blue) is shown on the left and hydrogen bonds are depicted by dashed lines. (*f*) Close-up views of the sugar conformations found in the −1 position of the two TtGH8 complexes, showing the conformations, with C1 and C4 labelled, as $^{2,5}B$ and $^1C_4$ for xylotriose and xylohexaose, respectively. All images were produced using *CCP4mg* (McNicholas *et al.*, 2011).

binding through the six subsites of the enzyme, which is consistent with the inactive-variant (Asp144Ala) complex with X5 observed for the *P. haloplanktis* enzyme by De Vos *et al.* (2006). In order to observe the hexasaccharide complex of TtGH8 we first made a catalytic variant at the proposed base, TtGH8 D281N; however, this appeared to retain around 0.1% of the catalytic activity (k_{cat}/K_m of 1.8×10^4 mg^{-1} ml min^{-1}; Table 1) in the reducing-sugar assay with MLX as substrate, which is consistent with similar mutants in other GH8 systems (Collins *et al.*, 2005; De Vos *et al.*, 2006). In addition, structures with X6, with long soak times, all revealed xylotriose product complexes (not shown). A complex with unhydrolysed X6 was therefore obtained by rapid soaking/cooling with a total time of approximately 10 s. Thus, a TtGH8 D281N structure in complex with X6 was obtained at 1.6 Å resolution, revealing X6 bound across the entire substrate-binding groove from subsites −3 to +3 (Figs. 4*c*, 4*d* and 4*e*; Tables 2 and 3).

The complex of TtGH8 with xylotriose showed distortion of the −1 xylose into a boat configuration, consistent with past work by others on the conformational itinerary in this family. To our surprise, the TtGH8 D281N–X6 complex revealed something different; namely, the −1 subsite sugar in a completely ring-flipped, southern hemisphere $^1C_4$ chair conformation (Fig. 4*f*). Although this allowed access to a hexasaccharide complex structure, the ring-flipped −1 sugar is unlikely to be representative of a catalytically relevant conformation since its position neither allows protonation of the leaving group by Glu73 nor is there a potential reactive water. Indeed, in the $^1C_4$ chair conformation the now axial (and 'down') O2 occupies the position that should instead be occupied by the nucleophilic water. Indeed, the only other structure in this family in which a substrate spanning the −1 subsite has been observed in anything other than the boat conformation was in the X5 complex of the *P. haloplanktis* GH8 (PDB entry 2b4f; De Vos *et al.*, 2006). Here, the −1 xylose was undistorted $^4C_1$ (although with scant density) and featured a position of the catalytic acid (akin to the TtGH8 apo structure) that was not commensurate with its role as a proton donor. This work, together with our current study, therefore highlights the difficulty in analysing substrate-complex structures with xylose-containing oligosaccharides which may be influenced by the integral conformational flexibility of xylose. Importantly, whilst the Asp281Asn variant has allowed definition of the interactions of the −3 to −2 and +1 to +3 subsites well, it highlights the occasional dangers of using inactive variants to study substrate distortion in the −1 site.

4. Discussion

As with recent studies on xylan utilization by the human microbiota (Rogowski *et al.*, 2015), the digestive system of bivalve molluscs such as marine wood-boring shipworms has the potential to provide a wealth of carbohydrate-active enzymes with potentially beneficial applications. Here, we have studied a potential GH8 xylanase from the shipworm symbiont *T. turnerae*, the genome sequence of which (Yang *et al.*, 2009) unveiled a treasure chest of carbohydrate-active enzymes. We have shown that TtGH8 is a single-domain endo-xylanase with six catalytically relevant subsites that hydrolyses X6 to yield predominantly xylotriose. The enzyme is active on diverse classical β-1,4-xylans, likely reflecting its role in the host digestion of woody biomass after the possible translocation of bacterial proteins from the gills to the connected gastrointestinal tract of its Teredinidae shipworm host (O'Connor *et al.*, 2014). The enzyme thus bears similarities to the well studied *P. haloplanktis* 'cold-adapted' xylanase, with which TtGH8 shares 33% identity. Intriguingly, TtGH8 shows the highest activity on mixed-linkage β-1,3,β-1,4 xylans, which may reflect a genuine biological adaption to, or at least an accommodation of, these marine substrates. The mixed-linkage marine xylan used in this study is found as a component of the red alga *Palmaria palmata*, and is a polysaccharide that is involved in mechanical support, development and defence (Viana *et al.*, 2011). Analysis of this polysaccharide suggests a 1:3 ratio of β-1,3:β-1,4 moieties. Whilst pure β-1,4 bonding of xylose residues would result in a threefold screw-axis helical structure, the irregular distribution of β-1,3 sections between variable lengths of β-1,4 sections may cause disruption of this regular conformation (Viana *et al.*, 2011). Optical rotation alignment further suggests that mixed-linkage xylan exhibits a 'random-coil' structure; unlike linear β-1,4-xylan, which may form interactions with other chains, the presence of β-1,3 linkages introduces flexibility which may assist in the solubility (Viana *et al.*, 2011; Cerezo *et al.*, 1971). Flexibility may improve the fitting of the polysaccharide into the V-shaped binding site of TtGH8. Improvement in solubility owing to the flexible nature of the xylan chain may be a factor in the increased degradation rate exhibited by TtGH8. We would argue, therefore, that the increased activity on MLX may not necessarily reflect the requirement for a β-1,3-linked xylose at one or more of the subsites, but confers increased solubility and thus enzyme access. It is also possible, given its marine environment, that the specificity of the enzyme has adapted to potential terrestrial xylans (β-1,4-xylans) and marine xylans (MLXs). It should be noted, however, that only k_{cat}/K_m values were determined and not the individual kinetic constants. This was because the maximum soluble substrate concentration was much lower than K_m, as indicated by the linear relationship between the rate and the substrate concentration. It is possible, therefore, that the difference in activity reflects a variation in K_m values that may not reflect the binding affinities but the actual concentrations of available substrate in two very different xylans. MLX has not been extensively used as a substrate in the analysis of xylanase activity. Exploring whether MLXs feature as the optimum substrate for all xylanases, or only those enzymes exposed to a marine system, will provide insight into the environmental selection pressures that influence glycoside hydrolase activity. In a discussion of shipworm larvae, Turner states that whilst shipworm larvae are quick to settle into burrows after extrusion from the adult, most wooden structures are covered in a 'protective forest' of various organisms, including algae, in which the young shipworm larvae may swim before settlement

(Turner, 1966). Bacterial symbionts are passed onto shipworm young, so it is possible that algal particles are digested using enzymes such as TtGH8 before or during larvae settlement. Such analyses highlight how shipworms and their symbionts offer a plethora of possibilities for novel enzyme discovery and application for biotechnology and biofuels.

Acknowledgements

The authors would also like to thank Diamond Light Source for beamtime (proposals mx-9948 and mx-13587) and the staff of beamlines I04 and I02 for assistance with crystal testing and data collection.

Funding information

This project was funded by the BBSRC (grant BB/L001926/1). GJD is a Royal Society Ken Murray Research Professor.

References

Adachi, W., Sakihama, Y., Shimizu, S., Sunami, T., Fukazawa, T., Suzuki, M., Yatsunami, R., Nakamura, S. & Takénaka, A. (2004). *J. Mol. Biol.* **343**, 785–795.
Alzari, P. M., Souchon, H. & Dominguez, R. (1996). *Structure*, **4**, 265–275.
Betcher, M. A. (2011). Thesis. Oregon Health and Science University. https://digitalcommons.ohsu.edu/etd/775/.
Biely, P., Singh, S. & Puchart, V. (2016). *Biotechnol. Adv.* **34**, 1260–1274.
Bomble, Y. J., Lin, C.-Y., Amore, A., Wei, H., Holwerda, E. K., Ciesielski, P. N., Donohoe, B. S., Decker, S. R., Lynd, L. R. & Himmel, M. E. (2017). *Curr. Opin. Chem. Biol.* **41**, 61–70.
Boynton, L. C. & Miller, R. C. (1927). *J. Biol. Chem.* **75**, 613–618.
Cerezo, A. S., Lezerovich, A., Labriola, R. & Rees, D. A. (1971). *Carbohydr. Res.* **19**, 289–296.
Charnock, S. J., Spurway, T. D., Xie, H., Beylot, M. H., Virden, R., Warren, R. A., Hazlewood, G. P. & Gilbert, H. J. (1998). *J. Biol. Chem.* **273**, 32187–32199.
Collins, T., De Vos, D., Hoyoux, A., Savvides, S. N., Gerday, C., Van Beeumen, J. & Feller, G. (2005). *J. Mol. Biol.* **354**, 425–435.
Cragg, S. M., Beckham, G. T., Bruce, N. C., Bugg, T. D. H., Distel, D. L., Dupree, P., Etxabe, A. G., Goodell, B. S., Jellison, J., McGeehan, J. E., McQueen-Mason, S. J., Schnorr, K., Walton, P. H., Watts, J. E. M. & Zimmer, M. (2015). *Curr. Opin. Chem. Biol.* **29**, 108–119.
Davies, G. J., Ducros, V. M.-A., Varrot, A. & Zechel, D. L. (2003). *Biochem. Soc. Trans.* **31**, 523–527.
Davies, G. & Henrissat, B. (1995). *Structure*, **3**, 853–859.
Davies, G. J., Planas, A. & Rovira, C. (2012). *Acc. Chem. Res.* **45**, 227–234.
Davies, G. J. & Williams, S. J. (2016). *Biochem. Soc. Trans.* **44**, 79–87.
Davies, G. J., Wilson, K. S. & Henrissat, B. (1997). *Biochem. J.* **321**, 557–559.
De Vos, D., Collins, T., Nerinckx, W., Savvides, S. N., Claeyssens, M., Gerday, C., Feller, G. & Van Beeumen, J. (2006). *Biochemistry*, **45**, 4797–4807.
Ekborg, N. A., Morrill, W., Burgoyne, A. M., Li, L. & Distel, D. L. (2007). *Appl. Environ. Microbiol.* **73**, 7785–7788.
Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. (2010). *Acta Cryst.* D**66**, 486–501.
Fierobe, H. P., Bagnara-Tardif, C., Gaudin, C., Guerlesquin, F., Sauve, P., Belaich, A. & Belaich, J.-P. (1993). *Eur. J. Biochem.* **217**, 557–565.
Fushinobu, S., Hidaka, M., Honda, Y., Wakagi, T., Shoun, H. & Kitaoka, M. (2005). *J. Biol. Chem.* **280**, 17180–17186.
Gallager, S. M., Turner, R. D. & Berg, C. J. (1981). *J. Exp. Mar. Biol. Ecol.* **52**, 63–77.
Guérin, D. M., Lascombe, M. B., Costabel, M., Souchon, H., Lamzin, V., Béguin, P. & Alzari, P. M. (2002). *J. Mol. Biol.* **316**, 1061–1069.
Hemsworth, G. R., Johnston, E. M., Davies, G. J. & Walton, P. H. (2015). *Trends Biotechnol.* **33**, 747–761.
Henrissat, B. & Davies, G. (1997). *Curr. Opin. Struct. Biol.* **7**, 637–644.
Himmel, M. E., Ding, S.-Y., Johnson, D. K., Adney, W. S., Nimlos, M. R., Brady, J. W. & Foust, T. D. (2007). *Science*, **315**, 804–807.
Horak, R. E. A. & Montoya, J. P. (2014). *J. Mar. Biol. Assoc. UK*, **94**, 177–185.
Krissinel, E. & Henrick, K. (2004). *Acta Cryst.* D**60**, 2256–2268.
Lebedev, A. A., Young, P., Isupov, M. N., Moroz, O. V., Vagin, A. A. & Murshudov, G. N. (2012). *Acta Cryst.* D**68**, 431–440.
Li, C., Zhao, X., Wang, A., Huber, G. W. & Zhang, T. (2015). *Chem. Rev.* **115**, 11559–11624.
Lombard, V., Golaconda Ramulu, H., Drula, E., Coutinho, P. M. & Henrissat, B. (2014). *Nucleic Acids Res.* **42**, D490–D495.
Matsui, I., Ishikawa, K., Matsui, E., Miyairi, S., Fukui, S. & Honda, K. (1991). *J. Biochem.* **109**, 566–569.
McCoy, A. J., Grosse-Kunstleve, R. W., Adams, P. D., Winn, M. D., Storoni, L. C. & Read, R. J. (2007). *J. Appl. Cryst.* **40**, 658–674.
McNicholas, S., Potterton, E., Wilson, K. S. & Noble, M. E. M. (2011). *Acta Cryst.* D**67**, 386–394.
Murshudov, G. N., Skubák, P., Lebedev, A. A., Pannu, N. S., Steiner, R. A., Nicholls, R. A., Winn, M. D., Long, F. & Vagin, A. A. (2011). *Acta Cryst.* D**67**, 355–367.
O'Connor, R. M. *et al.* (2014). *Proc. Natl Acad. Sci. USA*, **111**, E5096–E5104.
Pauly, M. & Keegstra, K. (2008). *Plant J.* **54**, 559–568.
Pell, G., Szabo, L., Charnock, S. J., Xie, H., Gloster, T. M., Davies, G. J. & Gilbert, H. J. (2004). *J. Biol. Chem.* **279**, 11777–11788.
Petersen, L., Ardèvol, A., Rovira, C. & Reilly, P. J. (2009). *J. Phys. Chem. B*, **113**, 7331–7339.
Potterton, L. *et al.* (2018). *Acta Cryst.* D**74**, 68–84.
Rogowski, A. *et al.* (2015). *Nature Commun.* **6**, 7481.
Rye, C. S. & Withers, S. G. (2000). *Curr. Opin. Chem. Biol.* **4**, 573–580.
Somerville, C. (2007). *Curr. Biol.* **17**, R115–R119.
Speciale, G., Thompson, A., Davies, G. J. & Williams, S. J. (2014). *Curr. Opin. Chem. Biol.* **28**, 1–13.
The CAZypedia Consortium (2018). *Glycobiology*, **28**, 3–8.
Turner, R. D. (1966). *A Survey and Illustrated Catalogue of the Teredinidae (Mollusca: Bivalvia)*. Cambridge: The Museum of Comparative Biology.
Vagin, A. & Teplyakov, A. (2010). *Acta Cryst.* D**66**, 22–25.
Viana, A. G., Noseda, M. D., Gonçalves, A. G., Duarte, M. E. R., Yokoya, N., Matulewicz, M. C. & Cerezo, A. S. (2011). *Carbohydr. Res.* **346**, 1023–1028.
Walton, P. H. & Davies, G. J. (2016). *Curr. Opin. Chem. Biol.* **31**, 195–207.
Waterbury, J. B., Calloway, C. B. & Turner, R. D. (1983). *Science*, **221**, 1401–1403.
Winter, G. (2010). *J. Appl. Cryst.* **43**, 186–190.
Winter, G., Waterman, D. G., Parkhurst, J. M., Brewster, A. S., Gildea, R. J., Gerstel, M., Fuentes-Montero, L., Vollmar, M., Michels-Clark, T., Young, I. D., Sauter, N. K. & Evans, G. (2018). *Acta Cryst.* D**74**, 85–97.
Yang, J. C. *et al.* (2009). *PLoS One*, **4**, e6085.

STRUCTURAL
BIOLOGY

Volume 74 (2018)

Supporting information for article:

Structure and function of a GH8 endoxylanase from *Teredinibacter turnerae*

Claire A. Fowler, Glyn R. Hemsworth, Fiona Cuskin, Sam Hart, Johan Turkenburg, Harry Gilbert, Paul H. Walton and Gideon J. Davies

Table S1 TtGH8 (CAZy reference TERTU_4506) DNA catalytic domain sequence involving residues 41-436.

GCTGGTGCCGTTGCTACCGGCGAGTACCGCAATCTGTTTGCCGAAATCGGAAAAAGCGAAATAGACATCC
AGCGCAAAATTGACGAGGCGTTTCAGCACTTGTTTTATGGCGACGCGAAAGATGCAGCTGTCTACTATCA
AGCGGGTGGAAACGAGAATGGTCCACTCGCATATGTTTACGATGTGAACAGCAATGACGTGCGCTCAGA
AGGCATGAGCTACGGCATGATGATTACTGTTCAAATGGACAAAAAAGCCGAGTTCGATGCAATCTGGAA
CTGGGCGAAAACCTATATGTATCAAGACTCCCCCACGCATCCAGCGTTTGGTTACTTTGCCTGGTCCATGC
GCCGCGATGGTGTCGCCAATGACGATATGCCAGCGCCAGATGGCGAGGAATATTTCGTGACCGCTCTCTA
TTTCGCCGCCGCCCGCTGGGGTAATGGCGAAGGTATTTTCAACTACCAACAGGAAGCGGACACCATTTTG
AGCCGCATGCGCCACCGCCAGGTGATCACCGGCCCAACCAATCGCGGAGTAATGACTGCGACCAATCTG
TTCCACCCGGAAGAGGCGCAAGTGCGCTTCACGCCCGACATCAATAATGCTGATCATACAGACGCGTCTT
ACCATCTGCCCTCGTTCTATGAAATTTGGGCACGTGTCGCGCCGCAAGAAGATCGCGCGTTTTGGGCCAA
AGCGGCCGATGTGAGCCGCGACTATTTTGCCAAAGCCGCCCACCCTGTCACTGCGTTAACACCGGACTAC
GGTAATTTTGATGGCACCCCGTGGGCGGCATCCTGGCGGCCGGAGTCGGTAGATTTTCGATACGATGCCT
GGCGTTCCGTCATGAACTGGTCCATGGACTATGCCTGGTGGGGCAAAGATTCAGGCGCACCCGCGCGCA
GTGATAAATTACTCGCGTTCTTCGAAACCCAGGAAGGCAAAATGAACCACCTCTATAGCCTGGATGGCAA
ACCGCTGGGTGGTGGACCGACCCTCGGCCTAATTTCCATGAATGCAACGGCAGCTATGGCAGCTACTGAT
CCCCGCTGGCACAATTTTGTGGAAAAGCTCTGGCAACAACAACCCCCCACAGGGCAATACCGGTACTAC
GACGGTGTTCTATACCTGATGGCGCTGCTACATTGCGCTGGGGAGTACAAAGCGTGGATCCCCGACGGGG
AATAA

Table S2 TtGH8 macromolecule production information. The TtGH8 catalytic domain was identified and a construct produced without the associated host periplasmic leader sequence and linker section. A hexahistidine tag and 3C protease cleavage site was built into the construct (flanked by NcoI and XhoI restriction sites) produced by GenScript, which was codon optimised for use in E.coli.

Source organism	*Teredinibacter turnerae*
DNA source	Optimised sequence obtained from GenScript
Forward primer	NcoI
Reverse primer	XhoI
Cloning vector	One Shot Top10
Expression vector	pET-28a
Expression host	*E.coli* BL21 (DE3)
Codon Optimised DNA Sequence	ATGGGCAGCAGCCATCATCATCATCACAGCAGCGG CCTGGAAGTTCTGTTCCAGGGACCAGCAGCGGGTGCGG TGGCGACCGGTGAATACCGTAACCTGTTCGCGGAGATTG GCAAGAGCGAAATCGACATTCAACGTAAAATCGATGAA GCGTTCCAGCACCTGTTTTATGGCGACGCGAAGGATGCG GCGGTTTACTATCAAGCGGGTGGCAACGAGAACGGTCC GCTGGCGTACGTGTATGACGTTAACAGCAACGATGTGC GTAGCGAGGGTATGAGCTACGGCATGATGATTACCGTTC AAATGGACAAGAAAGCGGAATTTGATGCGATCTGGAAC TGGGCGAAAACCTACATGTATCAAGACAGCCCGACCCA CCCGGCGTTCGGTTATTTTGCGTGGAGCATGCGTCGTGA

	TGGTGTGGCGAACGATGATATGCCGGCGCCGGATGGCG AGGAATACTTCGTTACCGCGCTGTATTTTGCGGCGGCGC GTTGGGGTAACGGCGAGGGTATCTTCAACTACCAGCAA GAAGCGGATACCATTCTGAGCCGTATGCGTCACCGTCAA GTGATCACCGGTCCGACCAACCGTGGCGTTATGACCGCG ACCAACCTGTTCCACCCGGAGGAAGCGCAAGTGCGTTTT ACCCCGGACATTAACAACGCGGACCACACCGATGCGAG CTACCACCTGCCGAGCTTTTATGAGATCTGGGCGCGTGT TGCGCCGCAGGAAGATCGTGCGTTTTGGGCGAAGGCGG CGGATGTGAGCCGTGATTACTTTGCGAAGGCGGCGCAC CCGTTACCGCGCTGACCCCGGACTATGGTAACTTCGATG GTACCCCGTGGGCGGCGAGCTGGCGTCCGGAGAGCGTG GACTTTCGTTACGATGCGTGGCGTAGCGTTATGAACTGG AGCATGGACTATGCGTGGTGGGGTAAAGATAGCGGTGC GCCGGCGCGTAGCGACAAACTGCTGGCGTTCTTTGAGAC CCAAGAAGGTAAAATGAACCACCTGTACAGCCTGGACG GTAAGCCGCTGGGTGGCGGTCCGACCCTGGGTCTGATTA GCATGAACGCGACCGCGGCGATGGCGGCGACCGACCCG CGTTGGCACAACTTCGTGGAAAAGCTGTGGCAGCAACA GCCGCCGACCGGCCAGTACCGTTACTATGACGGCGTTCT GTATCTGATGGCGCTGCTGCACTGCGCGGGCGAGTACA AAGCGTGGATCCCGGATGGCGAATAACTCGAG
Complete amino acid sequence of the construct produced	M G S S H H H H H H S S G L E V L F Q G P A A G A V A T G E Y R N L F A E I G K S E I D I Q R K I D E A F Q H L F Y G D A K D A A V Y Y Q A G G N E N G P L A Y V Y D V N S N D V R S E G M S Y G M M I T V Q M D K K A E F D A I W N W A K T Y M Y Q D S P T H P A F G Y F A W S M R R D G V A N D D M P A P D G E E Y F V T A L Y F A A A R W G N G E G I F N Y Q Q E A D T I L S R M R H R Q V I T G P T N R G V M T A T N L F H P E E A Q V R F T P D I N N A D H T D A S Y H L P S F Y E I W A R V A P Q E D R A F W A K A A D V S R D Y F A K A A H P V T A L T P D Y G N F D G T P W A A S W R P E S V D F R Y D A W R S V M N W S M D Y A W W G K D S G A P A R S D K L L A F F E T Q E G K M N H L Y S L D G K P L G G G P T L G L I S M N A T A A M A A T D P R W H N F V E K L W Q Q Q P P T G Q Y R Y Y D G V L Y L M A L L H C A G E Y K A W I P D G E Stop L E

Table S3 TtGH8 mutant macromolecule production information. Primers for point mutation were designed for the TtGH8 construct and implemented using NEB Q5 site mutagenesis kit.

Source organism	*Teredinibacter turnerae*
DNA source	TtGH8 Optimised sequence obtained from GenScript
Forward primer	CTTTCGTTACAACGCGTG GCGTAGCG (Asp 281- Asn)
Reverse primer	TCCACGCTCTCCGGACG C (Asp 281- Asn)
Cloning vector	One Shot Top10
Expression vector	pET-28a
Expression host	*E.coli* BL21 (DE3)
Complete amino acid sequence of the construct produced	M G S S H H H H H H S S G L E V L F Q G P A A G A V A T G E Y R N L F A E I G K S E I D I Q R K I D E A F Q H L F Y G D A K D A A V Y Y Q A G G N E N G P L A Y V Y D V N S N D V R S E G M S Y G M M I T V Q M D K K A E F D A I W N W A K T Y M Y Q D S P T H P A F G Y F A W S M R R D G V A N D D M P A P D G E E Y F V T A L Y F A A A R W G N G E G I F N Y Q Q E A D T I L S R M R H R Q V I T G P T N R G V M T A T N L F H P E E A Q V R F T P D I N N A D H T D A S Y H L P S F Y E I W A R V A P Q E D R A F W A K A A D V S R D Y F A K A A H P V T A L T P D Y G N F D G T P W A A S W R P E S V D F R Y N A W R S V M N W S M D Y A W W G K D S G A P A R S D K L L A F F E T Q E G K M N H L Y S L D G K P L G G G P T L G L I S M N A T A A M A A T D P R W H N F V E K L W Q Q Q P P T G Q Y R Y Y D G V L Y L M A L L H C A G E Y K A W I P D G E Stop L E

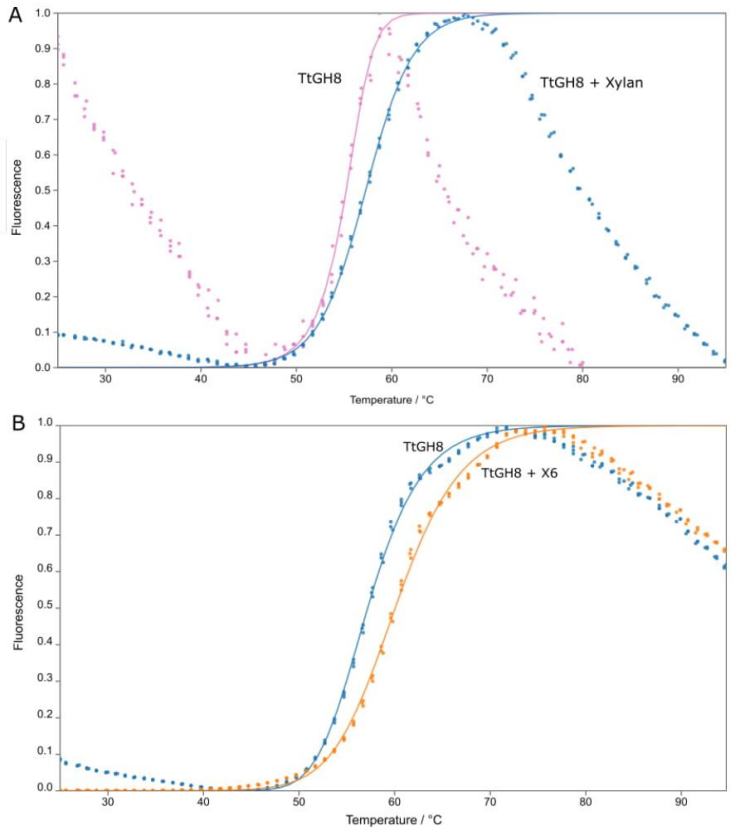


Figure S1 Thermal shift analysis of apo TtGH8 (22 µM) and incubation with A) birchwood xylan (1 mg/ml) and B) xylohexose (X6, 10 mM) in the presence of SYPRO orange. Fluorescence (shown as normalised values) indicates interaction of the SYPRO orange dye with aromatic residues commonly found within a protein structure. Increasing fluorescence as temperature increases models the denaturation of the protein. Analysis of the protein melting curves was completed using the JTSA online tool developed by Paul Bond, available at URL http://paulsbond.co.uk/jtsa/#/input. Addition of xylan and xylohexaose (X6) shifted the melting temperature of the apo protein from 55.2 -57.3 °C and 57.2 - 60.1 °C respectively. In general, a significant shift in temperature is deemed to be at least 1 °C and would indicate an interaction of the protein with the substrate. A positive shift in melting temperature of 2.1 °C due to xylan incubation and 2.9 °C with xylohexaose is indicative of TtGH8 activity/binding.

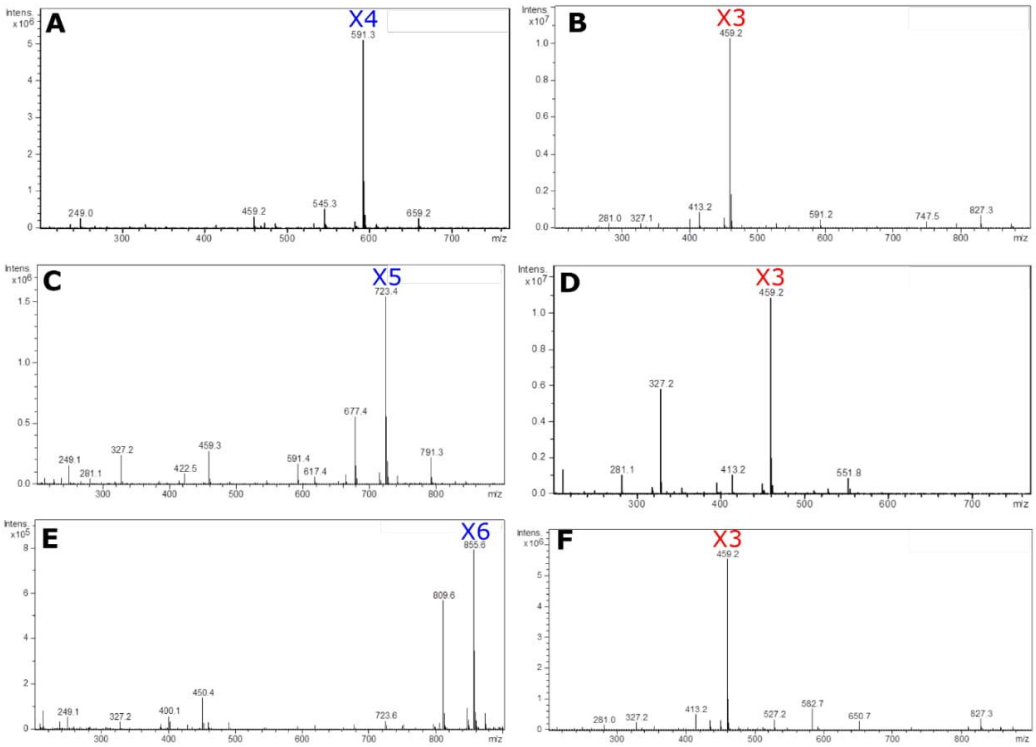


Figure S2 LCMS results of TtGH8 against larger xylooligosaccharides. A) Xylotetraose (591.3 m/z, formic acid adduct) B) Reaction products of TtGH8 and xylotetraose, producing xylotriose (459.2 m/z formic acid adduct). Xylose is not seen due to detections limits of the equipment. C) Xylopentaose (723.4 m/z formic acid adduct) D) Reaction products of TtGH8 and xylopentaose, producing xylotriose (459.2 m/z formic acid adduct) and xylobiose (327.2 m/z formic acid adduct). E) Xylohexaose (855.6 m/z formic acid adduct) F) Reaction products of TtGH8 and xylohexaose, producing only xylotriose (459.2 m/z formic acid adduct).

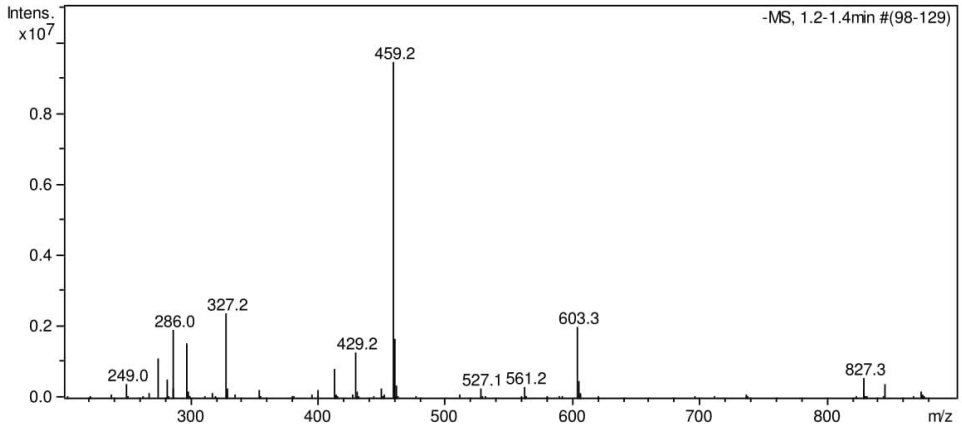


Figure S3 Liquid chromatography mass spectrum (LCMS-Dionex system) of soluble reaction products (run on a Cosmosil Sugar-D HPLC column) generated after incubation of TtGH8 (1 μM) with birchwood xylan (1 mg/ml) for 18 hr at 37 °C. Xylotriose and xylobiose are observed as formic acid adducts at 327.2 and 459.2 m/z respectively.

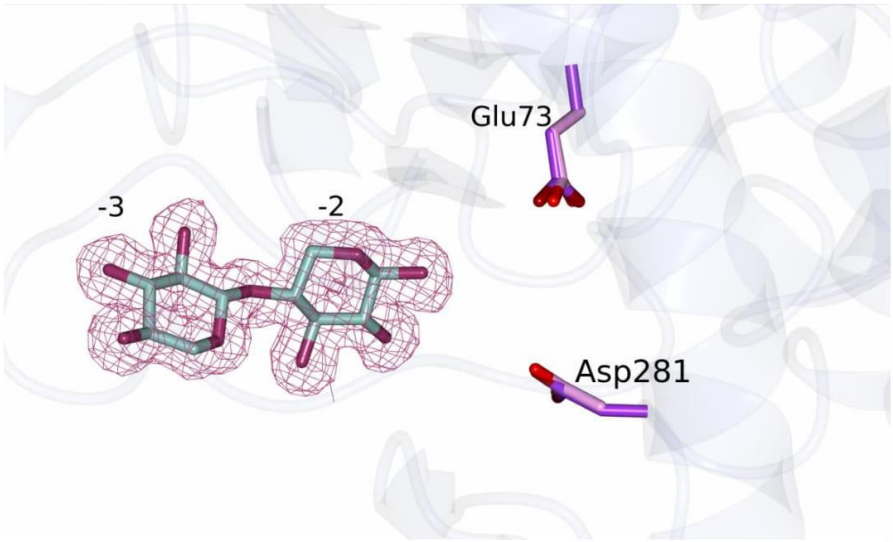


Figure S4 TtGH8 xylobiose complex, showing xylobiose bound in sites -3 to -2, with the corresponding maximum likelihood weighted Fo-Fc "difference" electron density, calculated *prior* to any incorporation of ligands in refinement, at contour level of 0.35 electrons / Å^3 (approx. 2.5σ). The catalytic residues in the xylobiose complex and the unliganded enzyme are shown in purple and light-pink, respectively.

Appendix 5: *T.turnerae* Growth Experiment Proteomic Analysis

Table 29 Proteomic analysis of protein content of the cell pellet collected during growth of *T.turnerae* culture on sigmacell food source. Data collected by MS/MS Ion Search, after Trypsin digestion. Other contaminating proteins were observed in the proteomic analysis but have not been shown in the table below. A potential description of each hit is taken from NCBI. Exponentially Modified Protein Abundance Index is abbreviated as EmPAI.

T.turnerae Cell Pellet Proteomic Analysis									
Family	**Database**	**Accession**	**Score**	**Mass**	**Num. of significant matches**	**Num. of significant sequences**	**emPAI**	**Mol %**	**Description**
9	NCBI_T_turnerae	WP_018277997.1	693	125025	18	18	0.63	0.18	TonB-dependent receptor [Teredinibacter turnerae]
12	NCBI_T_turnerae	WP_018415904.1	567	59360	19	19	1.94	0.55	phosphoenolpyruvate carboxykinase (ATP) [Teredinibacter turnerae]
19	NCBI_T_turnerae	WP_018416100.1	438	49776	16	15	1.76	0.50	xylose isomerase [Teredinibacter turnerae]
30	NCBI_T_turnerae	WP_018275717.1	349	112337	11	11	0.39	0.11	TonB-dependent receptor [Teredinibacter turnerae]
39	NCBI_T_turnerae	ACR11039.1	277	106623	8	8	0.29	0.08	TonB-dependent receptor [Teredinibacter turnerae T7901]
67	NCBI_T_turnerae	WP_018416081.1	196	25464	6	6	1.19	0.34	membrane protein [Teredinibacter turnerae]
77	NCBI_T_turnerae	WP_018277703.1	182	75504	6	6	0.31	0.09	peptidase M13 [Teredinibacter turnerae]
78	NCBI_T_turnerae	WP_018277281.1	179	106885	6	6	0.21	0.06	hypothetical protein [Teredinibacter turnerae]
80	NCBI_T_turnerae	WP_018275583.1	178	49468	6	6	0.5	0.14	biopolymer transporter TonB [Teredinibacter turnerae]
86	NCBI_T_turnerae	ACR14237.1	173	29286	5	5	0.77	0.22	conserved hypothetical protein [Teredinibacter turnerae T7901]
94	NCBI_T_turnerae	WP_018275433.1	157	67503	3	3	0.16	0.05	TonB-dependent receptor [Teredinibacter turnerae]
116	NCBI_T_turnerae	ACR12978.1	131	52058	6	6	0.47	0.13	glutamine synthetase, type I [Teredinibacter turnerae T7901]
118	NCBI_T_turnerae	ACR11393.1	130	118086	4	3	0.12	0.03	ribonuclease E [Teredinibacter turnerae T7901]
132	NCBI_T_turnerae	WP_018276110.1	122	30019	2	2	0.25	0.07	DUF481 domain-containing protein [Teredinibacter turnerae]
138	NCBI_T_turnerae	WP_018416164.1	118	78948	4	4	0.19	0.05	hypothetical protein [Teredinibacter turnerae]

Family	Database	Accession	Score	Mass	Num. of significant matches	Num. of significant sequences	emPAI	Mol %	Description
142	NCBI_T_turnerae	ACR13988.1	116	27911	2	2	0.27	0.08	NADH:ubiquinone oxidoreductase, C subunit [Teredinibacter turnerae T7901]
168	NCBI_T_turnerae	ACR14788.1	102	24195	2	2	0.32	0.09	toluene-tolerance protein [Teredinibacter turnerae T7901]
178	NCBI_T_turnerae	ACR10904.1	98	49853	4	4	0.31	0.09	type I secretion outer membrane protein TolC [Teredinibacter turnerae T7901]
191	Uniprot	C5BU70	91	35398	3	3	0.33	0.09	Malate dehydrogenase OS=Teredinibacter turnerae (strain ATCC 39867 / T7901) GN=mdh PE=3 SV=1
197	NCBI_T_turnerae	ACR13466.1	87	60107	3	3	0.18	0.05	glucose-6-phosphate isomerase [Teredinibacter turnerae T7901]
205	NCBI_T_turnerae	WP_028885230.1	84	20925	2	2	0.37	0.10	DUF4136 domain-containing protein [Teredinibacter turnerae]
212	NCBI_T_turnerae	ACR14024.1	81	44770	1	1	0.08	0.02	D-3-phosphoglycerate dehydrogenase [Teredinibacter turnerae T7901]
230	NCBI_T_turnerae	ACR14800.1	75	50987	2	2	0.14	0.04	ribonucleoside-diphosphate reductase [Teredinibacter turnerae T7901]
255	NCBI_T_turnerae	ACR12999.1	68	49501	2	2	0.15	0.04	acetyl-CoA carboxylase, biotin carboxylase subunit [Teredinibacter turnerae T7901]
273	Uniprot	C5BTJ0	65	23725	2	2	0.32	0.09	ATP-dependent Clp protease proteolytic subunit OS=Teredinibacter turnerae (strain ATCC 39867 / T7901) GN=clpP PE=3 SV=1
277	NCBI_T_turnerae	WP_018277691.1	64	46788	2	2	0.15	0.04	M18 family aminopeptidase [Teredinibacter turnerae]
279	Uniprot	C5BLR8	64	34115	3	3	0.34	0.10	Transaldolase OS=Teredinibacter turnerae (strain ATCC 39867 / T7901) GN=tal PE=3 SV=1
283	NCBI_T_turnerae	ACR12600.1	63	22776	1	1	0.16	0.05	cytochrome c oxidase, Cbb3-type, subunit II [Teredinibacter turnerae T7901]
304	NCBI_T_turnerae	ACR11710.1	58	29433	1	1	0.12	0.03	fumarate reductase / succinate dehydrogenase family protein, iron-sulfur subunit [Teredinibacter turnerae T7901]

Family	Database	Accession	Score	Mass	Num. of significant matches	Num. of significant sequences	emPAI	Mol %	Description
306	NCBI_T_turnerae	ACR12856.1	58	19200	1	1	0.19	0.05	OmpA domain protein [Teredinibacter turnerae T7901]
315	NCBI_T_turnerae	ACR11317.1	55	90819	1	1	0.04	0.01	TonB-dependent receptor [Teredinibacter turnerae T7901]
327	NCBI_T_turnerae	ACR13350.1	54	107294	2	2	0.07	0.02	ribonucleoside-diphosphate reductase alpha subunit, class I [Teredinibacter turnerae T7901]
331	NCBI_T_turnerae	ACR12274.1	53	17541	1	1	0.21	0.06	Dps family protein [Teredinibacter turnerae T7901]
343	NCBI_T_turnerae	ACR13652.1	51	42713	1	1	0.08	0.02	membrane protease subunit HflK [Teredinibacter turnerae T7901]
352	NCBI_T_turnerae	ACR11583.1	50	31555	1	1	0.11	0.03	methylenetetrahydrofolate reductase [Teredinibacter turnerae T7901]
362	NCBI_T_turnerae	ACR14494.1	49	28145	1	1	0.13	0.04	putative secreted protein [Teredinibacter turnerae T7901]
364	NCBI_T_turnerae	ACR14059.1	49	21827	1	1	0.16	0.05	antioxidant, AhpC/Tsa family [Teredinibacter turnerae T7901]
391	NCBI_T_turnerae	ACR12797.1	44	84693	1	1	0.04	0.01	putative transcriptional accessory protein [Teredinibacter turnerae T7901]
403	NCBI_T_turnerae	ACR11324.1	43	78965	1	1	0.04	0.01	putative protein PilJ [Teredinibacter turnerae T7901]
421	NCBI_T_turnerae	WP_018277330.1	41	68167	2	2	0.1	0.03	protein translocase subunit SecD [Teredinibacter turnerae]
424	NCBI_T_turnerae	WP_018417194.1	40	33893	1	1	0.1	0.03	porin [Teredinibacter turnerae]
427	NCBI_T_turnerae	ACS93546.1	40	36170	1	1	0.1	0.03	HTH-type transcriptional regulator CysB [Teredinibacter turnerae T7901]
456	NCBI_T_turnerae	ACR10895.1	37	71039	1	1	0.05	0.01	ABC transporter domain containing protein [Teredinibacter turnerae T7901]
470	NCBI_T_turnerae	ACR11635.1	36	27159	1	1	0.13	0.04	putative membrane protein [Teredinibacter turnerae T7901]
475	NCBI_T_turnerae	WP_028886481.1	35	26949	1	1	0.13	0.04	YciK family oxidoreductase [Teredinibacter turnerae]

Family	Database	Accession	Score	Mass	Num. of significant matches	Num. of significant sequences	emPAI	Mol %	Description
535	NCBI_T_turnerae	ACR12398.1	31	21544	1	1	0.17	0.05	superoxide dismutase [Teredinibacter turnerae T7901]
542	NCBI_T_turnerae	ACR13413.1	30	27757	1	1	0.13	0.04	peptidyl-prolyl cis-trans isomerase, FKBP-type [Teredinibacter turnerae T7901]
567	NCBI_T_turnerae	ACR13432.1	28	115259	1	1	0.03	0.01	TonB-dependent receptor [Teredinibacter turnerae T7901]
605	NCBI_T_turnerae	ACR11760.1	24	44302	1	1	0.08	0.02	putative lipoprotein [Teredinibacter turnerae T7901]
614	NCBI_T_turnerae	ACR14393.1	23	26370	1	1	0.13	0.04	RNA polymerase sigma factor for flagellar operon [Teredinibacter turnerae T7901]
615	NCBI_T_turnerae	ACR12598.1	23	51732	1	1	0.07	0.02	dihydrolipoyl dehydrogenase [Teredinibacter turnerae T7901]
643	NCBI_T_turnerae	ACR10686.1	19	115649	1	1	0.03	0.01	glycoside hydrolase family 3 domain protein [Teredinibacter turnerae T7901]
659	NCBI_T_turnerae	ACR13392.1	18	65718	2	1	0.05	0.01	acyl-CoA dehydrogenase domain protein [Teredinibacter turnerae T7901]
661	NCBI_T_turnerae	ACR11634.1	17	32447	1	1	0.11	0.03	cytochrome c oxidase, Cbb3-type, subunit III [Teredinibacter turnerae T7901]
663	NCBI_T_turnerae	WP_018276910.1	17	65980	1	1	0.05	0.01	right-handed parallel beta-helix repeat-containing protein [Teredinibacter turnerae]
673	NCBI_T_turnerae	ACR12620.1	16	49415	1	1	0.07	0.02	Carboxy-terminal-processing protease precursor [Teredinibacter turnerae T7901]
683	NCBI_T_turnerae	ACR12399.1	16	21391	1	1	0.17	0.05	yggt family protein [Teredinibacter turnerae T7901]
703	NCBI_T_turnerae	ACR13739.1	14	29354	1	1	0.12	0.03	ferredoxin--NADP reductase [Teredinibacter turnerae T7901]
704	NCBI_T_turnerae	ACR14254.1	13	8969	1	1	0.43	0.12	hypothetical protein TERTU_3630 [Teredinibacter turnerae T7901]

Table 30 Proteomic analysis of protein content of the media collected during growth of *T.turnerae* culture on sigmacell food source.Data collected by MS/MS Ion Search, after Trypsin digestion.Other contaminating proteins were observed in the proteomic analysis but have not been shown in the table below. A potential description of each hit is taken from NCBI. Exponentially Modified Protein Abundance Index is abbreviated as EmPAI.

***T.turnerae* secreted proteins**									
Family	**Database**	**Accession**	**Score**	**Mass**	**Num. of significant matches**	**Num. of significant sequences**	**emPAI**	**Mol%**	**Description**
1	NCBI_T_turnerae	WP_018415904.1	718	59360	30	21	2.91	1.69	phosphoenolpyruvate carboxykinase (ATP) [Teredinibacter turnerae]
5	NCBI_T_turnerae	WP_018416100.1	597	49776	23	17	2.62	1.52	xylose isomerase [Teredinibacter turnerae]
6	NCBI_T_turnerae	WP_018277997.1	475	125025	24	23	0.87	0.51	TonB-dependent receptor [Teredinibacter turnerae]
7	NCBI_T_turnerae	WP_018276610.1	387	104109	15	14	0.63	0.37	TonB-dependent receptor [Teredinibacter turnerae]
9	NCBI_T_turnerae	WP_018416437.1	297	79536	13	12	0.74	0.43	dipeptidyl carboxypeptidase II [Teredinibacter turnerae]
11	NCBI_T_turnerae	WP_018016028.1	287	35368	10	7	1.58	0.92	malate dehydrogenase [Teredinibacter turnerae]
12	NCBI_T_turnerae	WP_018275935.1	274	313181	7	7	0.08	0.05	hypothetical protein [Teredinibacter turnerae]
20	NCBI_T_turnerae	WP_018277703.1	191	75504	7	7	0.37	0.22	peptidase M13 [Teredinibacter turnerae]
21	NCBI_T_turnerae	WP_019604011.1	188	51699	8	8	0.68	0.40	adenosylhomocysteinase [Teredinibacter turnerae]
22	NCBI_T_turnerae	WP_018275630.1	184	28401	8	7	1.56	0.91	methyltransferase [Teredinibacter turnerae]
27	NCBI_T_turnerae	ACR12978.1	165	52058	11	9	0.91	0.53	glutamine synthetase, type I [Teredinibacter turnerae T7901]
28	NCBI_T_turnerae	WP_018275363.1	161	70668	6	6	0.33	0.19	oligoendopeptidase F [Teredinibacter turnerae]
29	NCBI_T_turnerae	WP_018276415.1	157	37818	6	6	0.7	0.41	N-acetyl-gamma-glutamyl-phosphate reductase [Teredinibacter turnerae]
32	NCBI_T_turnerae	WP_018274079.1	146	53629	7	7	0.55	0.32	glucuronate isomerase [Teredinibacter turnerae]
37	NCBI_T_turnerae	WP_018275583.1	132	49468	4	4	0.31	0.18	biopolymer transporter TonB [Teredinibacter turnerae]
38	NCBI_T_turnerae	WP_018274710.1	131	100446	6	5	0.22	0.13	aminopeptidase N [Teredinibacter turnerae]
43	NCBI_T_turnerae	WP_018277281.1	124	106885	5	5	0.17	0.10	hypothetical protein [Teredinibacter turnerae]

Family	Database	Accession	Score	Mass	Num. of significant matches	Num. of significant sequences	emPAI	Mol%	Description
49	NCBI_T_turnerae	ACR12598.1	118	51732	5	5	0.38	0.22	dihydrolipoyl dehydrogenase [Teredinibacter turnerae T7901]
51	NCBI_T_turnerae	WP_018416081.1	117	25464	5	5	0.92	0.54	membrane protein [Teredinibacter turnerae]
52	NCBI_T_turnerae	ACR10904.1	113	49853	5	5	0.4	0.23	type I secretion outer membrane protein TolC [Teredinibacter turnerae T7901]
53	NCBI_T_turnerae	ACR14059.1	112	21827	3	3	0.58	0.34	antioxidant, AhpC/Tsa family [Teredinibacter turnerae T7901]
54	NCBI_T_turnerae	WP_018274887.1	112	45483	3	3	0.25	0.15	serine hydroxymethyltransferase [Teredinibacter turnerae]
59	NCBI_T_turnerae	WP_018276110.1	107	30019	4	4	0.56	0.33	DUF481 domain-containing protein [Teredinibacter turnerae]
67	NCBI_T_turnerae	ACR11775.1	98	51404	5	4	0.39	0.23	peptidase, imelysin family [Teredinibacter turnerae T7901]
69	Uniprot	C5BLU3	97	58674	2	2	0.12	0.07	GMP synthase [glutamine-hydrolyzing] OS=Teredinibacter turnerae (strain ATCC 39867 / T7901) GN=guaA PE=3 SV=1
74	NCBI_T_turnerae	ACR12181.1	94	76142	2	2	0.09	0.05	methionine--tRNA ligase [Teredinibacter turnerae T7901]
77	NCBI_T_turnerae	ACR11014.1	92	51594	2	2	0.14	0.08	argininosuccinate lyase [Teredinibacter turnerae T7901]
78	NCBI_T_turnerae	ACR12454.1	91	29072	3	3	0.41	0.24	putative deoxyribonuclease Sll1786 [Teredinibacter turnerae T7901]
79	NCBI_T_turnerae	ACR12829.1	90	40132	3	3	0.28	0.16	oxidoreductase, NAD-binding Rossmann fold family [Teredinibacter turnerae T7901]
80	NCBI_T_turnerae	WP_018274037.1	90	60033	5	5	0.32	0.19	glucose-6-phosphate isomerase [Teredinibacter turnerae]
82	NCBI_T_turnerae	ACR10668.1	89	20611	1	1	0.17	0.10	putative acetyltransferase [Teredinibacter turnerae T7901]
83	Uniprot	C5BQZ6	88	36886	5	4	0.58	0.34	Ketol-acid reductoisomerase (NADP(+)) OS=Teredinibacter turnerae (strain ATCC 39867 / T7901) GN=ilvC PE=3 SV=1
85	NCBI_T_turnerae	ACR12856.1	88	19200	3	3	0.67	0.39	OmpA domain protein [Teredinibacter turnerae T7901]

Family	Database	Accession	Score	Mass	Num. of significant matches	Num. of significant sequences	emPAI	Mol%	Description
88	NCBI_T_turnerae	WP_018274985.1	86	27772	3	3	0.43	0.25	membrane protein [Teredinibacter turnerae]
89	NCBI_T_turnerae	WP_018275433.1	86	67503	3	3	0.16	0.09	TonB-dependent receptor [Teredinibacter turnerae]
91	NCBI_T_turnerae	ACR10811.1	85	20253	3	3	0.63	0.37	inorganic diphosphatase [Teredinibacter turnerae T7901]
94	NCBI_T_turnerae	ACR10686.1	84	115649	4	4	0.12	0.07	glycoside hydrolase family 3 domain protein [Teredinibacter turnerae T7901]
96	NCBI_T_turnerae	ACR14788.1	82	24195	2	2	0.32	0.19	toluene-tolerance protein [Teredinibacter turnerae T7901]
113	NCBI_T_turnerae	ACR12980.1	77	31806	1	1	0.11	0.06	thioredoxin [Teredinibacter turnerae T7901]
114	NCBI_T_turnerae	ACR12150.1	77	19492	1	1	0.18	0.10	thiol-disulfide isomerase and thioredoxins/Thiol-disulfide oxidoreductase resA [Teredinibacter turnerae T7901]
120	NCBI_T_turnerae	ACR12750.1	74	40035	5	5	0.52	0.30	aspartate-semialdehyde dehydrogenase [Teredinibacter turnerae T7901]
128	NCBI_T_turnerae	WP_018275795.1	69	63541	2	2	0.11	0.06	cellulase [Teredinibacter turnerae]
129	Uniprot	C5BQ72	69	11538	3	3	1.32	0.77	50S ribosomal protein L24 OS=Teredinibacter turnerae (strain ATCC 39867 / T7901) GN=rplX PE=3 SV=1
131	NCBI_T_turnerae	WP_018275181.1	69	118387	4	4	0.12	0.07	hypothetical protein [Teredinibacter turnerae]
132	NCBI_T_turnerae	WP_018275819.1	68	65896	2	2	0.11	0.06	hypothetical protein [Teredinibacter turnerae]
136	NCBI_T_turnerae	WP_018274172.1	68	91344	2	2	0.08	0.05	glycoside hydrolase [Teredinibacter turnerae]
139	NCBI_T_turnerae	ACR13760.1	66	26199	3	3	0.46	0.27	thiol:disulfide interchange protein DsbA precursor [Teredinibacter turnerae T7901]
145	NCBI_T_turnerae	ACR12074.1	65	11865	2	2	0.73	0.42	thioredoxin [Teredinibacter turnerae T7901]
146	NCBI_T_turnerae	WP_018274795.1	65	33798	1	1	0.1	0.06	DUF4340 domain-containing protein [Teredinibacter turnerae]
152	NCBI_T_turnerae	ACR11760.1	64	44302	1	1	0.08	0.05	putative lipoprotein [Teredinibacter turnerae T7901]
158	NCBI_T_turnerae	WP_018276910.1	64	65980	2	2	0.11	0.06	right-handed parallel beta-helix repeat-containing protein [Teredinibacter turnerae]

Family	Database	Accession	Score	Mass	Num. of significant matches	Num. of significant sequences	emPAI	Mol%	Description
160	NCBI_T_turnerae	WP_019605879.1	63	38462	6	4	0.69	0.40	alpha-N-arabinofuranosidase [Teredinibacter turnerae]
162	NCBI_T_turnerae	ACR11505.1	63	46380	4	4	0.34	0.20	histidinol dehydrogenase [Teredinibacter turnerae T7901]
164	NCBI_T_turnerae	ACR14074.1	63	56780	5	5	0.35	0.20	glucose-6-phosphate dehydrogenase [Teredinibacter turnerae T7901]
166	NCBI_T_turnerae	ACR12398.1	61	21544	3	2	0.36	0.21	superoxide dismutase [Teredinibacter turnerae T7901]
169	NCBI_T_turnerae	WP_018277807.1	59	82476	3	3	0.13	0.08	alpha-glucuronidase [Teredinibacter turnerae]
171	NCBI_T_turnerae	ACR11620.1	58	47988	2	2	0.15	0.09	Tol-Pal system beta propeller repeat protein TolB [Teredinibacter turnerae T7901]
172	NCBI_T_turnerae	WP_018275650.1	58	41737	2	2	0.17	0.10	hypothetical protein [Teredinibacter turnerae]
173	NCBI_T_turnerae	ACR12004.2	57	72537	2	2	0.1	0.06	arginine 2-monooxygenase [Teredinibacter turnerae T7901]
174	NCBI_T_turnerae	ACR11317.1	57	90819	3	3	0.12	0.07	TonB-dependent receptor [Teredinibacter turnerae T7901]
175	Uniprot	C5BQ61	52	22798	2	2	0.34	0.20	50S ribosomal protein L3 OS=Teredinibacter turnerae (strain ATCC 39867 / T7901) GN=rplC PE=3 SV=1
192	NCBI_T_turnerae	WP_018276526.1	53	32025	2	2	0.23	0.13	putative selenate ABC transporter substrate-binding protein [Teredinibacter turnerae]
196	Uniprot	C5BTJ0	53	23725	1	1	0.15	0.09	ATP-dependent Clp protease proteolytic subunit OS=Teredinibacter turnerae (strain ATCC 39867 / T7901) GN=clpP PE=3 SV=1
199	NCBI_T_turnerae	ACR14100.1	53	35802	2	2	0.21	0.12	carbohydrate binding module family 33 and 10 domain protein [Teredinibacter turnerae T7901]
200	NCBI_T_turnerae	ACR13207.1	53	30460	2	2	0.25	0.15	nicotinate-nucleotide diphosphorylase [Teredinibacter turnerae T7901]
202	NCBI_T_turnerae	WP_018277031.1	52	77014	3	3	0.14	0.08	oligopeptidase A [Teredinibacter turnerae]
207	NCBI_T_turnerae	ACR12274.1	50	17541	2	2	0.46	0.27	Dps family protein [Teredinibacter turnerae T7901]
210	NCBI_T_turnerae	WP_018276818.1	50	20846	1	1	0.17	0.10	YceI family protein [Teredinibacter turnerae]
212	NCBI_T_turnerae	ACR11053.1	49	83036	2	2	0.08	0.05	conserved hypothetical protein Teredinibacter turnerae

Family	Database	Accession	Score	Mass	Num. of significant matches	Num. of significant sequences	emPAI	Mol%	Description
221	NCBI_T_turnerae	ACR14237.1	48	29286	2	2	0.26	0.15	conserved hypothetical protein [Teredinibacter turnerae T7901]
223	NCBI_T_turnerae	ACR13496.1	47	60055	1	1	0.06	0.03	A-type flagellin [Teredinibacter turnerae T7901]
230	Uniprot	C5BQ71	45	13450	2	2	0.62	0.36	50S ribosomal protein L14 OS=Teredinibacter turnerae (strain ATCC 39867 / T7901) GN=rplN PE=3 SV=1
231	NCBI_T_turnerae	WP_018274899.1	44	95999	3	3	0.11	0.06	hypothetical protein [Teredinibacter turnerae]
236	NCBI_T_turnerae	WP_018416164.1	44	78948	3	3	0.14	0.08	hypothetical protein [Teredinibacter turnerae]
238	NCBI_T_turnerae	WP_018417194.1	44	33893	1	1	0.1	0.06	porin [Teredinibacter turnerae]
242	NCBI_T_turnerae	ACR10643.1	43	20926	2	2	0.37	0.22	peroxiredoxin [Teredinibacter turnerae T7901]
249	NCBI_T_turnerae	ACR13552.1	42	44338	2	2	0.16	0.09	xylanase [Teredinibacter turnerae T7901]
258	Uniprot	C5BQA1	41	19316	2	2	0.41	0.24	3-hydroxydecanoyl-[acyl-carrier-protein] dehydratase OS=Teredinibacter turnerae (strain ATCC 39867 / T7901) GN=fabA PE=3 SV=1
265	Uniprot	C5BQ62	38	22564	1	1	0.16	0.09	50S ribosomal protein L4 OS=Teredinibacter turnerae (strain ATCC 39867 / T7901) GN=rplD PE=3 SV=1
266	NCBI_T_turnerae	ACR14510.1	38	53067	1	1	0.07	0.04	glycoside hydrolase family 5 domain protein [Teredinibacter turnerae T7901]
269	NCBI_T_turnerae	WP_018277908.1	38	53258	2	2	0.13	0.08	leucyl aminopeptidase [Teredinibacter turnerae]
277	NCBI_T_turnerae	ACR10734.1	37	20026	1	1	0.18	0.10	intracellular protease, PfpI family [Teredinibacter turnerae T7901]
282	NCBI_T_turnerae	WP_018277691.1	36	46788	1	1	0.07	0.04	M18 family aminopeptidase [Teredinibacter turnerae]
285	NCBI_T_turnerae	ACR12076.1	36	18122	2	2	0.44	0.26	peptidoglycan-associated lipoprotein [Teredinibacter turnerae T7901]
288	NCBI_T_turnerae	ACR12798.1	36	31128	2	2	0.24	0.14	polysaccharide deacetylase family protein [Teredinibacter turnerae T7901]
290	NCBI_T_turnerae	ACR13430.1	35	45608	2	2	0.16	0.09	Starvation-sensing protein [Teredinibacter turnerae T7901]
293	NCBI_T_turnerae	ACR11839.1	35	35705	1	1	0.1	0.06	TonB domain protein [Teredinibacter turnerae T7901]

Family	Database	Accession	Score	Mass	Num. of significant matches	Num. of significant sequences	emPAI	Mol%	Description
294	NCBI_T_turnerae	WP_018274537.1	35	18473	3	3	0.71	0.41	molybdenum cofactor biosynthesis protein B [Teredinibacter turnerae]
295	NCBI_T_turnerae	WP_018276523.1	35	60727	2	2	0.12	0.07	hypothetical protein [Teredinibacter turnerae]
297	NCBI_T_turnerae	ACR12147.1	35	81294	1	1	0.04	0.02	conserved hypothetical protein [Teredinibacter turnerae T7901]
301	NCBI_T_turnerae	WP_018278098.1	35	38961	1	1	0.09	0.05	hypothetical protein [Teredinibacter turnerae]
303	NCBI_T_turnerae	ACR12431.1	34	66836	3	3	0.16	0.09	dihydrolipoyllysine-residue acetyltransferase E2 component of pyruvate dehydrogenase complex [Teredinibacter turnerae T7901]
308	Uniprot	C5BR76	34	13443	2	2	0.62	0.36	50S ribosomal protein L19 OS=Teredinibacter turnerae (strain ATCC 39867 / T7901) GN=rplS PE=3 SV=1
309	NCBI_T_turnerae	ACR10990.1	34	43835	1	1	0.08	0.05	phosphoserine phosphatase [Teredinibacter turnerae T7901]
310	NCBI_T_turnerae	ACR14494.1	34	28145	1	1	0.13	0.08	putative secreted protein [Teredinibacter turnerae T7901]
313	NCBI_T_turnerae	ACR13899.1	33	37768	2	2	0.19	0.11	phosphate binding protein [Teredinibacter turnerae T7901]
314	NCBI_T_turnerae	ACR11046.1	33	39375	1	1	0.09	0.05	xylanase [Teredinibacter turnerae T7901]
322	Uniprot	C5BS85	32	16316	1	1	0.22	0.13	6,7-dimethyl-8-ribityllumazine synthase OS=Teredinibacter turnerae (strain ATCC 39867 / T7901) GN=ribH PE=3 SV=1
325	NCBI_T_turnerae	ACR11303.1	32	15645	1	1	0.23	0.13	ferric uptake regulation protein [Teredinibacter turnerae T7901]
328	NCBI_T_turnerae	ACR13510.1	32	31342	1	1	0.11	0.06	putative lipoprotein [Teredinibacter turnerae T7901]
343	NCBI_T_turnerae	ACR13775.1	30	20375	1	1	0.18	0.10	conserved hypothetical protein [Teredinibacter turnerae T7901]
359	NCBI_T_turnerae	WP_018277586.1	29	29319	1	1	0.12	0.07	VacJ lipoprotein [Teredinibacter turnerae]
365	NCBI_T_turnerae	ACR14029.1	28	20919	1	1	0.17	0.10	putative lipoprotein [Teredinibacter turnerae T7901]
377	NCBI_T_turnerae	ACR12260.1	27	42285	1	1	0.08	0.05	penicillin-binding protein 6 [Teredinibacter turnerae T7901]

Family	Database	Accession	Score	Mass	Num. of significant matches	Num. of significant sequences	emPAI	Mol%	Description
380	NCBI_T_turnerae	ACR13177.1	27	17479	1	1	0.21	0.12	phosphoribosylaminoimidazole carboxylase, catalytic subunit [Teredinibacter turnerae T7901]
398	NCBI_T_turnerae	ACR11959.1	26	28676	1	1	0.12	0.07	methionine aminopeptidase, type I [Teredinibacter turnerae T7901]
402	NCBI_T_turnerae	ACR11162.1	25	43894	1	1	0.08	0.05	xylanase [Teredinibacter turnerae T7901]
405	NCBI_T_turnerae	WP_018278008.1	25	43015	1	1	0.08	0.05	hypothetical protein [Teredinibacter turnerae]
406	NCBI_T_turnerae	ACR10809.1	25	14005	1	1	0.26	0.15	**lactoylglutathione lyase [Teredinibacter turnerae T7901]**
411	NCBI_T_turnerae	ACR12933.1	24	136609	1	1	0.03	0.02	type IV pilus biogenesis protein [Teredinibacter turnerae T7901]
412	Uniprot	C5BS66	24	15862	1	1	0.23	0.13	50S ribosomal protein L13 OS=Teredinibacter turnerae (strain ATCC 39867 / T7901) GN=rplM PE=3 SV=1
415	NCBI_T_turnerae	ACR14614.1	24	114111	1	1	0.03	0.02	TonB-dependent receptor [Teredinibacter turnerae T7901]
420	NCBI_T_turnerae	WP_018275493.1	24	31602	1	1	0.11	0.06	MBL fold metallo-hydrolase [Teredinibacter turnerae]
430	NCBI_T_turnerae	ACR13449.1	23	55851	1	1	0.06	0.03	2-isopropylmalate synthase [Teredinibacter turnerae T7901]
438	NCBI_T_turnerae	ACR13840.1	22	50038	1	1	0.07	0.04	putative peptidase [Teredinibacter turnerae T7901]
443	NCBI_T_turnerae	WP_018275187.1	22	100828	3	3	0.11	0.06	TonB-dependent receptor [Teredinibacter turnerae]
447	NCBI_T_turnerae	ACR14211.1	21	57658	1	1	0.06	0.03	glycoside hydrolase family 51 domain protein [Teredinibacter turnerae T7901]
450	NCBI_T_turnerae	ACR11193.1	21	60712	1	1	0.06	0.03	glycoside hydrolase family 43 domain protein [Teredinibacter turnerae T7901]
452	NCBI_T_turnerae	WP_018274151.1	21	99427	1	1	0.03	0.02	TonB-dependent receptor [Teredinibacter turnerae]
456	NCBI_T_turnerae	ACR12041.1	21	86959	2	2	0.08	0.05	TonB-dependent receptor [Teredinibacter turnerae T7901]
470	Uniprot	C5BSJ0	19	48593	1	1	0.07	0.04	5-methylthioadenosine/S-adenosylhomocysteine deaminase OS=Teredinibacter turnerae (strain ATCC 39867 / T7901) GN=mtaD PE=3 SV=1

Family	Database	Accession	Score	Mass	Num. of significant matches	Num. of significant sequences	emPAI	Mol%	Description
473	NCBI_T_turnerae	ACR11840.1	19	17815	1	1	0.2	0.12	CheW domain protein [Teredinibacter turnerae T7901]
475	Uniprot	C5BU55	19	6847	1	1	0.57	0.33	50S ribosomal protein L32 OS=Teredinibacter turnerae (strain ATCC 39867 / T7901) GN=rpmF PE=3 SV=1
487	NCBI_T_turnerae	WP_018274421.1	18	67017	1	1	0.05	0.03	peptidyl-dipeptidase A [Teredinibacter turnerae]
489	NCBI_T_turnerae	WP_018277477.1	18	73696	1	1	0.05	0.03	TonB-dependent receptor [Teredinibacter turnerae]
493	Uniprot	C5BQ76	17	19068	1	1	0.19	0.11	50S ribosomal protein L6 OS=Teredinibacter turnerae (strain ATCC 39867 / T7901) GN=rplF PE=3 SV=1
504	NCBI_T_turnerae	WP_018277521.1	16	75175	1	1	0.05	0.03	DUF5011 domain-containing protein [Teredinibacter turnerae]
507	NCBI_T_turnerae	ACR14581.1	15	109213	1	1	0.03	0.02	conserved hypothetical protein [Teredinibacter turnerae T7901]
512	NCBI_T_turnerae	ACR11936.1	15	73305	1	1	0.05	0.03	ExsB [Teredinibacter turnerae T7901]
520	NCBI_T_turnerae	WP_018275946.1	14	98401	1	1	0.04	0.02	hypothetical protein [Teredinibacter turnerae]
523	Uniprot	C5BJ25	14	28299	1	1	0.13	0.08	2,3-bisphosphoglycerate-dependent phosphoglycerate mutase OS=Teredinibacter turnerae (strain ATCC 39867 / T7901) GN=gpmA PE=3 SV=1

Abbreviation List

Abbreviation	*Full name*
AA (9-15)	Auxilary activity
abs	Absorbance
BLAST	Basic local alignment search tool
bMLG	Mixed linkage barley glucan
bX	Birchwood xylan
C1	Carbon position 1, glucose ring
C4	Carbon position 4, glucose ring
cAX	Corn arabinoxylan
CAZy	Carbohydrate active enzyme database
CAZyme	Carbohydrate active enzyme
CBM	Carbohydrate binding module
Cu	Copper
DNSA	3,5-Dinitrosalicylic acid
DP (followed by number)	Degree of polymerisation
EDTA	Ethylenediaminetetraacetic acid
CW-EPR	Continuous wave electron paramagnetic resonance
ESI-MS	Electrospray ionisation mass spectrometry
GH	Glycoside hydrolase
HEPES	4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid
HPAEC-PAD	High performance anion exchange with pulsed amperiometric detection
IPTG	Isopropyl β-D-1-thiogalactopyranoside
ITC	Isothermal titration calorimetry

LB	Lysogeny broth
LCMS	Liquid chromatography mass spectrometry
LPMO	Lytic polysaccharide monooxygenase
MALDI-TOF-MS	Matrix assisted laser desorbtion-time of flight-mass spectrometry
MLX	Mixed linkage xylan
OD	Optical density
PASC	Phosphoric acid swollen cellulose
PBS	Phosphate-buffered saline
PEG	Polyethylene glycol
rAX	Rye arabinoxylan
SDS PAGE	Sodium dodecyl sulfate–polyacrylamide gel electrophoresis
SEC	Size exclusion chromatography
SECMALS	Size exclusion chromatography multi angle light scattering
Strep	Strep tag
SUMO	Small ubiquitin like modifier
TLC	Thin layer chromatography
Tm	Melting temperature
TSA	Thermal shift assay
Tt	Teredinibacter turnerae
tXyG	Taramarind xyloglucan
wAX	Wheat arabinoxylan
X1	Xylose
X2	Xylobiose
X3	Xylotriose
X4	Xylotetraose
X5	Xylopentaose
X6	Xylohexaose

References

1. Arrhenius, G.; Caldwell, K. D.; Wold, S.; Ingenjörsvetenskapsakademien, *A Tribute to the Memory of Svante Arrhenius (1859-1927): Presented at the 2008 Annual Meeting of the Royal Swedish Academy of Engineering Sciences*. Royal Swedish Academy of Engineering Sciences (IVA): 2008.
2. Arrhenius, S., *Uber den Einfluss des atmosphärischen Kohlensäure-gehalts auf die Temperatur der Erdoberfläche*. 1896.
3. Sheldon, R. A., Green and sustainable manufacture of chemicals from biomass: state of the art. *Green Chemistry* **2014,** *16* (3), 950-963.
4. Claassen, P. A. M.; van Lier, J. B.; Lopez Contreras, A. M.; van Niel, E. W. J.; Sijtsma, L.; Stams, A. J. M.; de Vries, S. S.; Weusthuis, R. A., Utilisation of biomass for the supply of energy carriers. *Applied Microbiology and Biotechnology* **1999,** *52* (6), 741-755.
5. Dimarogona, M.; Topakas, E.; Christakopoulos, P., Recalcitrant polysaccharide degradation by novel oxidative biocatalysts. *Applied Microbiology and Biotechnology* **2013,** *97* (19), 8455-8465.
6. Cherubini, F., The biorefinery concept: Using biomass instead of oil for producing energy and chemicals. *Energy Conversion and Management* **2010,** *51* (7), 1412-1421.
7. Octave, S.; Thomas, D., Biorefinery: Toward an industrial metabolism. *Biochimie* **2009,** *91* (6), 659-664.
8. Himmel, M. E.; Ding, S. Y.; Johnson, D. K.; Adney, W. S.; Nimlos, M. R.; Brady, J. W.; Foust, T. D., Biomass recalcitrance: Engineering plants and enzymes for biofuels production. *Science* **2007,** *315* (5813), 804-807.
9. Laine, R. A., A calculation of all possible oligosaccharide isomers both branched and linear yields 1.05x10(12) structures for a reducing hexasaccharide: the isomer barrier to development of single-method saccharide sequencing or synthesis systems. *Glycobiology* **1994,** *4*.
10. Varki, A.; Cummings, R.; Aebi, M.; H Packer, N.; H Seeberger, P.; D Esko, J.; Stanley, P.; Hart, G.; Darvill, A.; Kinoshita, T.; J Prestegard, J.; Schnaar, R.; Freeze, H.; D Marth, J.; R Bertozzi, C.; Etzler, M.; Frank, M.; Vliegenthart, J.; Lütteke, T.; Kornfeld, S., *Symbol Nomenclature for Graphical Representations of Glycans*. 2015; Vol. 25, p 1323-1324.
11. Cosgrove, D. J., Growth of the plant cell wall. *Nature Reviews Molecular Cell Biology* **2005,** *6* (11), 850-861.
12. Frommhagen, M.; Westphal, A. H.; van Berkel, W. J. H.; Kabel, M. A., Distinct Substrate Specificities and Electron-Donating Systems of Fungal Lytic Polysaccharide Monooxygenases. *Frontiers in Microbiology* **2018,** *9*.
13. Boerjan, W.; Ralph, J.; Baucher, M., Lignin biosynthesis. *Annual Review of Plant Biology* **2003,** *54*, 519-546.
14. Jarvis, M., Cellulose stacks up. *Nature* **2003,** *426*, 611.
15. Davies, G.; Henrissat, B., Structures and mechanisms of glycosyl hydrolases. *Structure* **1995,** *3* (9), 853-859.
16. Capon, B., Mechanism in carbohydrate chemistry. *Chemical Reviews* **1969,** *69* (4), 407-498.
17. Koshland, D. E., Stereochemistry and the mechanism of enzymatic reactions. *Biological Reviews of the Cambridge Philosophical Society* **1953,** *28* (4), 416-436.
18. McCarter, J. D.; Withers, S. G., Mechanisms of enzymatic glycoside hydrolysis. *Curr Opin Struct Biol* **1994,** *4* (6), 885-92.

19. McIntosh, L. P.; Hand, G.; Johnson, P. E.; Joshi, M. D.; Korner, M.; Plesniak, L. A.; Ziser, L.; Wakarchuk, W. W.; Withers, S. G., The pKa of the general acid/base carboxyl group of a glycosidase cycles during catalysis: a 13C-NMR study of bacillus circulans xylanase. *Biochemistry* **1996,** *35* (31), 9958-66.
20. Davies, G. J.; Wilson, K. S.; Henrissat, B., Nomenclature for sugar-binding subsites in glycosyl hydrolases. *Biochemical Journal* **1997,** *321* (Pt 2), 557-559.
21. Cremer, D.; Pople, J. A., General definition of ring puckering coordinates. *Journal of the American Chemical Society* **1975,** *97* (6), 1354-1358.
22. Jeffrey, G. A.; Yates, J. H., Stereographic representation of the cremer-pople ring-puckering parameters for pyranoid rings. *Carbohydrate research* **1979,** *74* (1), 319-322.
23. Davies, G. J.; Planas, A.; Rovira, C., Conformational Analyses of the Reaction Coordinate of Glycosidases. *Accounts of Chemical Research* **2012,** *45* (2), 308-316.
24. Sulzenbacher, G.; Mackenzie, L. F.; Wilson, K. S.; Withers, S. G.; Dupont, C.; Davies, G. J., The crystal structure of a 2-fluorocellotriosyl complex of the Streptomyces lividans endoglucanase CelB2 at 1.2 A resolution. *Biochemistry* **1999,** *38* (15), 4826-33.
25. Henrissat, B.; Davies, G., Structural and sequence-based classification of glycoside hydrolases. *Current Opinion in Structural Biology* **1997,** *7* (5), 637-644.
26. Lim, V. I., Structural principles of the globular organization of protein chains. A stereochemical theory of globular protein secondary structure. *J Mol Biol* **1974,** *88* (4), 857-72.
27. Henrissat, B.; Claeyssens, M.; Tomme, P.; Lemesle, L.; Mornon, J. P., Cellulase families revealed by hydrophobic cluster analysis. *Gene* **1989,** *81* (1), 83-95.
28. Gaboriaud, C.; Bissery, V.; Benchetrit, T.; Mornon, J. P., Hydrophobic cluster analysis: An efficient new way to compare and analyse amino acid sequences. *FEBS Letters* **1987,** *224* (1), 149-155.
29. Henrissat, B., A classification of glycosyl hydrolases based on amino-acid-sequence similarities. *Biochem J* **1991,** *280*.
30. Coutinho, P. M.; Deleury, E.; Davies, G. J.; Henrissat, B., An evolving hierarchical family classification for glycosyltransferases. *Journal of Molecular Biology* **2003,** *328* (2), 307-317.
31. Davies, G. J.; Sinnott, M., Sorting the diverse: The sequence-based classifications of carbohydrate-active enzymes. *Biochem. J* **2008,** *30*, 26-32.
32. Levasseur, A.; Drula, E.; Lombard, V.; Coutinho, P. M.; Henrissat, B., Expansion of the enzymatic repertoire of the CAZy database to integrate auxiliary redox enzymes. *Biotechnology for Biofuels* **2013,** *6*.
33. Henrissat, B.; Bairoch, A., New families in the classification of glycosyl hydrolases based on amino-acid-sequence similarities. *Biochemical Journal* **1993,** *293*, 781-788.
34. Lombard, V.; Golaconda Ramulu, H.; Drula, E.; Coutinho, P. M.; Henrissat, B., The carbohydrate-active enzymes database (CAZy) in 2013. *Nucleic Acids Research* **2014,** *42* (D1), D490-D495.
35. Cantarel, B. L.; Coutinho, P. M.; Rancurel, C.; Bernard, T.; Lombard, V.; Henrissat, B., The Carbohydrate-Active EnZymes database (CAZy): an expert resource for Glycogenomics. *Nucleic Acids Res* **2009,** *37*.
36. Henrissat, B.; Bairoch, A., Updating the sequence-based classification of glycosyl hydrolases. *Biochem J* **1996,** *316 (Pt 2)*, 695-6.
37. Beeson, W. T.; Phillips, C. M.; Cate, J. H.; Marletta, M. A., Oxidative cleavage of cellulose by fungal copper-dependent polysaccharide monooxygenases. *J Am Chem Soc* **2012,** *134*.
38. Vaaje-Kolstad, G.; Westereng, B.; Horn, S. J.; Liu, Z.; Zhai, H.; Sorlie, M.; Eijsink, V. G. H., An Oxidative Enzyme Boosting the Enzymatic Conversion of Recalcitrant Polysaccharides. *Science* **2010,** *330* (6001), 219-222.

39. Ciano, L.; Davies, G. J.; Tolman, W. B.; Walton, P. H., Bracing copper for the catalytic oxidation of C–H bonds. *Nature Catalysis* **2018,** *1* (8), 571-577.
40. Harris, P. V.; Welner, D.; McFarland, K. C.; Re, E.; Navarro Poulsen, J. C.; Brown, K., Stimulation of lignocellulosic biomass hydrolysis by proteins of glycoside hydrolase family 61: structure and function of a large, enigmatic family. *Biochemistry* **2010,** *49*.
41. Valdivia, M.; Galan, J. L.; Laffarga, J.; Ramos, J. L., Biofuels 2020: Biorefineries based on lignocellulosic materials. *Microbial Biotechnology* **2016,** *9* (5), 585-594.
42. Horn, S. J.; Vaaje-Kolstad, G.; Westereng, B.; Eijsink, V. G. H., Novel enzymes for the degradation of cellulose. *Biotechnology for Biofuels* **2012,** *5*.
43. Frandsen, K. E. H.; Lo Leggio, L., Lytic polysaccharide monooxygenases: a crystallographer's view on a new class of biomass-degrading enzymes. *Iucrj* **2016,** *3*, 448-467.
44. Meier, K. K.; Jones, S. M.; Kaper, T.; Hansson, H.; Koetsier, M. J.; Karkehabadi, S.; Solomon, E. I.; Sandgren, M.; Kelemen, B., Oxygen Activation by Cu LPMOs in Recalcitrant Carbohydrate Polysaccharide Conversion to Monomer Sugars. *Chemical Reviews* **2018,** *118* (5), 2593-2635.
45. Reese, E. T.; Siu, R. G. H.; Levinson, H. S., The biological degradation of soluble cellulose derivatives and its relationship to the mechanism of cellulose hydrolysis. *J Bacteriol* **1950,** *59*.
46. Raguz, S.; Yague, E.; Wood, D. A.; Thurston, C. F., Isolation and characterization of a cellulose-growth-specific gene from Agaricus bisporus. *Gene* **1992,** *119* (2), 183-90.
47. Schrempf, H.; Walter, S., The cellulolytic system of Streptomyces reticuli. *International Journal of Biological Macromolecules* **1995,** *17* (6), 353-355.
48. Saloheimo, M.; Nakari-Setala, T.; Tenkanen, M.; Penttila, M., cDNA cloning of a Trichoderma reesei cellulase and demonstration of endoglucanase activity by expression in yeast. *Eur J Biochem* **1997,** *249* (2), 584-91.
49. Eriksson, K. E.; Pettersson, B.; Westermark, U., Oxidation: an important enzyme reaction in fungal degradation of cellulose. *FEBS Lett* **1974,** *49* (2), 282-5.
50. Schnellmann, J.; Zeltins, A.; Blaak, H.; Schrempf, H., The novel lectin-like protein CHB1 is encoded by a chitin-inducible Streptomyces olivaceoviridis gene and binds specifically to crystalline alpha-chitin of fungi and other organisms. *Molecular microbiology* **1994,** *13* (5), 807-19.
51. Henrissat, B.; Levasseur, A., Expansion of the enzymatic repertoire of theCAZy database to integrate auxiliary redoxenzymes *Biotechnol Biofuels* **2013,** *6* (41).
52. Vaaje-Kolstad, G.; Horn, S. J.; van Aalten, D. M. F.; Synstad, B.; Eijsink, V. G. H., The non-catalytic chitin-binding protein CBP21 from Serratia marcescens is essential for chitin degradation. *Journal of Biological Chemistry* **2005,** *280* (31), 28492-28497.
53. Vaaje-Kolstad, G.; Houston, D. R.; Riemen, A. H.; Eijsink, V. G.; van Aalten, D. M., Crystal structure and binding properties of the Serratia marcescens chitin-binding protein CBP21. *The Journal of biological chemistry* **2005,** *280* (12), 11313-9.
54. Moser, F.; Irwin, D.; Chen, S. L.; Wilson, D. B., Regulation and characterization of Thermobifida fusca carbohydrate-binding module proteins E7 and E8. *Biotechnology and Bioengineering* **2008,** *100* (6), 1066-1077.
55. McNicholas, S.; Potterton, E.; Wilson, K. S.; Noble, M. E., Presenting your structures: the CCP4mg molecular-graphics software. *Acta crystallographica. Section D, Biological crystallography* **2011,** *67*, 386-394.
56. Karkehabadi, S.; Hansson, H.; Kim, S.; Piens, K.; Mitchinson, C.; Sandgren, M., The First Structure of a Glycoside Hydrolase Family 61 Member, Cel61B from Hypocrea jecorina, at 1.6 angstrom Resolution. *Journal of Molecular Biology* **2008,** *383* (1), 144-154.

57. Quinlan, R. J.; Sweeney, M. D.; Lo Leggio, L.; Otten, H.; Poulsen, J. C. N.; Johansen, K. S., Insights into the oxidative degradation of cellulose by a copper metalloenzyme that exploits biomass components. *Proc Natl Acad Sci USA* **2011,** *108*.
58. Phillips, C. M.; Beeson, W. T.; Cate, J. H.; Marletta, M. A., Cellobiose Dehydrogenase and a Copper-Dependent Polysaccharide Monooxygenase Potentiate Cellulose Degradation by Neurospora crassa. *Acs Chemical Biology* **2011,** *6* (12), 1399-1406.
59. Westereng, B.; Cannella, D.; Agger, J. W.; Jorgensen, H.; Andersen, M. L.; Eijsink, V. G. H.; Felby, C., Enzymatic cellulose oxidation is linked to lignin by long-range electron transfer. *Scientific Reports* **2015,** *5*.
60. Couturier, M.; Ladeveze, S.; Sulzenbacher, G.; Ciano, L.; Fanuel, M.; Moreau, C.; Villares, A.; Cathala, B.; Chaspoul, F.; Frandsen, K. E.; Labourel, A.; Herpoel-Gimbert, I.; Grisel, S.; Haon, M.; Lenfant, N.; Rogniaux, H.; Ropartz, D.; Davies, G. J.; Rosso, M. N.; Walton, P. H.; Henrissat, B.; Berrin, J. G., Lytic xylan oxidases from wood-decay fungi unlock biomass degradation. *Nat Chem Biol* **2018,** *14* (3), 306-310.
61. Aachmann, F. L.; Sorlie, M.; Skjak-Braek, G.; Eijsink, V. G. H.; Vaaje-Kolstad, G., NMR structure of a lytic polysaccharide monooxygenase provides insight into copper binding, protein dynamics, and substrate interactions. *Proceedings of the National Academy of Sciences of the United States of America* **2012,** *109* (46), 18779-18784.
62. Lieberman, R. L.; Rosenzweig, A. C., Crystal structure of a membrane-bound metalloenzyme that catalyses the biological oxidation of methane. *Nature* **2005,** *434* (7030), 177-182.
63. Ross, M. O.; Rosenzweig, A. C., A tale of two methane monooxygenases. *Journal of Biological Inorganic Chemistry* **2017,** *22* (2-3), 307-319.
64. MacPherson, I. S.; Murphy, M. E., Type-2 copper-containing enzymes. *Cellular and molecular life sciences : CMLS* **2007,** *64* (22), 2887-99.
65. Hemsworth, G. R.; Taylor, E. J.; Kim, R. Q.; Gregory, R. C.; Lewis, S. J.; Turkenburg, J. P.; Parkin, A.; Davies, G. J.; Walton, P. H., The Copper Active Site of CBM33 Polysaccharide Oxygenases. *Journal of the American Chemical Society* **2013,** *135* (16), 6069-6077.
66. Hemsworth, G. R.; Henrissat, B.; Davies, G. J.; Walton, P. H., Discovery and characterization of a new family of lytic polysaccharide monooxygenases. *Nat Chem Biol* **2014,** *10*.
67. Vu, V. V.; Beeson, W. T.; Phillips, C. M.; Cate, J. H.; Marletta, M. A., Determinants of regioselective hydroxylation in the fungal polysaccharide monooxygenases. *J Am Chem Soc* **2014,** *136*.
68. Lo Leggio, L.; Simmons, T. J.; Poulsen, J.-C. N.; Frandsen, K. E. H.; Hemsworth, G. R.; Stringer, M. A.; von Freiesleben, P.; Tovborg, M.; Johansen, K. S.; De Maria, L.; Harris, P. V.; Soong, C.-L.; Dupree, P.; Tryfona, T.; Lenfant, N.; Henrissat, B.; Davies, G. J.; Walton, P. H., Structure and boosting activity of a starch-degrading lytic polysaccharide monooxygenase. *Nature Communications* **2015,** *6*.
69. Nekiunaite, L.; Isaksen, T.; Vaaje-Kolstad, G.; Abou Hachem, M., Fungal lytic polysaccharide monooxygenases bind starch and β-cyclodextrin similarly to amylolytic hydrolases. *FEBS Letters* **2016,** *590* (16), 2737-2747.
70. Sabbadin, F.; Pesante, G.; Elias, L.; Besser, K.; Li, Y.; Steele-King, C. G.; Stark, M.; Rathbone, D. A.; Dowle, A.; Bates, R.; Shipway, J. R.; Cragg, S. M.; Bruce, N. C.; McQueen Mason, S. J., Uncovering the molecular mechanisms of lignocellulose digestion in shipworms. *Biotechnology for biofuels* **2018,** *11* (59).
71. Vu, V. V.; Ngo, S. T., Copper active site in polysaccharide monooxygenases. *Coordination Chemistry Reviews* **2018,** *368*, 134-157.
72. Smith, S. M.; Rawat, S.; Telser, J.; Hoffman, B. M.; Stemmler, T. L.; Rosenzweig, A. C., Crystal Structure and Characterization of Particulate Methane Monooxygenase from Methylocystis species Strain M. *Biochemistry* **2011,** *50* (47), 10231-10240.

73. Himes, R. A.; Karlin, K. D., Copper-dioxygen complex mediated C-H bond oxygenation: relevance for particulate methane monooxygenase (pMMO). *Current Opinion in Chemical Biology* **2009,** *13* (1), 119-131.
74. Walton, P. H.; Davies, G. J., On the catalytic mechanisms of lytic polysaccharide monooxygenases. *Current Opinion in Chemical Biology* **2016,** *31*, 195-207.
75. Kim, S.; Stahlberg, J.; Sandgren, M.; Paton, R. S.; Beckham, G. T., Quantum mechanical calculations suggest that lytic polysaccharide monooxygenases use a copper-oxyl, oxygen-rebound mechanism. *Proceedings of the National Academy of Sciences of the United States of America* **2014,** *111* (1), 149-154.
76. Kjaergaard, C. H.; Qayyum, M. F.; Wong, S. D.; Xu, F.; Hemsworth, G. R.; Davies, G. J.; Walton, P. H.; Johansen, K. S.; Hodgson, K. O.; Hedman, B.; Solomon, E. I., Spectroscopic and computational insight into the activation of O-2 by the mononuclear Cu center in Polysaccharide monooxygenases. *Abstracts of Papers of the American Chemical Society* **2014,** *248*.
77. Bischof, R. H.; Ramoni, J.; Seiboth, B., Cellulases and beyond: the first 70 years of the enzyme producer Trichoderma reesei. *Microbial Cell Factories* **2016,** *15* (1), 106.
78. Mohanram, S.; Amat, D.; Choudhary, J.; Arora, A.; Nain, L., Novel perspectives for evolving enzyme cocktails for lignocellulose hydrolysis in biorefineries. *Sustainable Chemical Processes* **2013,** *1* (1), 15.
79. Waterbury, J. B.; Calloway, C. B.; Turner, R. D., A cellulolytic nitrogen-fixing bacterium cultured from the gland of deshayes in shipworms (bivalvia, teredinidae). *Science* **1983,** *221* (4618), 1401-1403.
80. Cragg, S. M.; Beckham, G. T.; Bruce, N. C.; Bugg, T. D. H.; Distel, D. L.; Dupree, P.; Etxabe, A. G.; Goodell, B. S.; Jellison, J.; McGeehan, J. E.; McQueen-Mason, S. J.; Schnorr, K.; Walton, P. H.; Watts, J. E. M.; Zimmer, M., Lignocellulose degradation mechanisms across the Tree of Life. *Current Opinion in Chemical Biology* **2015,** *29*, 108-119.
81. Harris, J. M., The presence, nature, and role of gut microflora in aquatic invertebrates - a synthesis. *Microbial Ecology* **1993,** *25* (3), 195-231.
82. Dash, M., The Epic Struggle to Tunnel Under the Thames. *Smithsonian Magazine* 2012.
83. Treneman, N. C.; Carlton, J. T.; Borges, L. M. S.; Shipway, J. R.; Raupach, M. J.; Altermark, B., Species diversity and abundance of shipworms (Mollusca: Bivalvia: Teredinidae) in woody marine debris generated by the Great East Japan Earthquake and Tsunami of 2011. *Aquatic Invasions* **2018,** *13* (1), 87-100.
84. Turner, R. D., A survey and illustrated catalogue of the Teredinidae (Mollusca:Bivalvia). University, H., Ed. The Museum of Comparative Biology: Cambridge, MA, 1966.
85. Shipway, J. R. Aspects of the life history strategies of the Teredinidae. Doctoral Thesis, University of Portmouth, 2013.
86. Pechenik, J. A.; Perron, F. E.; Turner, R. D., Role of phytoplankton in the diets of adult and larval shipworms, *lyrodus-pedicellatus* (Bivalvia, Teredinidae). *Estuaries* **1979,** *2* (1), 58-60.
87. Betcher, M. A.; Fung, J. M.; Han, A. W.; O'Connor, R.; Seronay, R.; Concepcion, G. P.; Distel, D. L.; Haygood, M. G., Microbial distribution and abundance in the digestive system of five shipworm species (Bivalvia: Teredinidae). **2012**.
88. Sundberg, A., Molluscan Explosion: The Dutch Shipworm Epidemic of the 1730s. *Environment & Society Portal* 2015.
89. Carpenter, E. J.; Culliney, J. L., Nitrogen-fixation in marine shipworms. *Science* **1975,** *187* (4176), 551-552.
90. Fiore, C. L.; Jarett, J. K.; Olson, N. D.; Lesser, M. P., Nitrogen fixation and nitrogen transformations in marine symbioses. *Trends in Microbiology* **2010,** *18* (10), 455-463.

91. Gallager, S. M.; Turner, R. D.; Berg, C. J., Physiological aspects of wood consumption, growth, and reproduction in the shipworm Lyrodus pedicellatus Quatrefages (Bivalvia: Teredinidae). *Journal of Experimental Marine Biology and Ecology* **1981,** *52* (1), 63-77.
92. Boynton, L. C.; Miller, R. C., The occurrence of a cellulase in the ship-worm. *Journal of Biological Chemistry* **1927,** *75* (2), 613-618.
93. Dore, W. H. M., R.C, *The digestion of wood by Teredo navalis.* Berkeley, 1923.
94. D. Popham, J.; R. Dickson, M., *Bacterial associations in the teredo Bankia australis (Lamellibranchia: Mollusca)*. 1973; Vol. 19, p 338-340.
95. Griffin, H. L.; Freer, S. N.; Greene, R. V., Extracellular endoglucanase activity by a novel bacterium isolated from marine shipworm. *Biochemical and biophysical research communications* **1987,** *144* (1), 143-51.
96. Distel, D. L.; Beaudoin, D. J.; Morrill, W., Coexistence of multiple proteobacterial endosymbionts in the gills of the wood-boring bivalve Lyrodus pedicellatus (Bivalvia : Teredinidae). *Applied and Environmental Microbiology* **2002,** *68* (12), 6292-6299.
97. Gilkes, N. R.; Warren, R. A.; Miller, R. C., Jr.; Kilburn, D. G., Precise excision of the cellulose binding domains from two Cellulomonas fimi cellulases by a homologous protease and the effect on catalysis. *The Journal of biological chemistry* **1988,** *263* (21), 10401-7.
98. Gill, J.; Rixon, J. E.; Bolam, D. N.; McQueen-Mason, S.; Simpson, P. J.; Williamson, M. P.; Hazlewood, G. P.; Gilbert, H. J., The type II and X cellulose-binding domains of Pseudomonas xylanase A potentiate catalytic activity against complex substrates by a common mechanism. *Biochem J* **1999,** *342 (Pt 2)*, 473-80.
99. Yang, J. C.; Madupu, R.; Durkin, A. S.; Ekborg, N. A.; Pedamallu, C. S.; Hostetler, J. B.; Radune, D.; Toms, B. S.; Henrissat, B.; Coutinho, P. M.; Schwarz, S.; Field, L.; Trindade-Silva, A. E.; Soares, C. A. G.; Elshahawi, S.; Hanora, A.; Schmidt, E. W.; Haygood, M. G.; Posfai, J.; Benner, J.; Madinger, C.; Nove, J.; Anton, B.; Chaudhary, K.; Foster, J.; Holman, A.; Kumar, S.; Lessard, P. A.; Luyten, Y. A.; Slatko, B.; Wood, N.; Wu, B.; Teplitski, M.; Mougous, J. D.; Ward, N.; Eisen, J. A.; Badger, J. H.; Distel, D. L., The Complete Genome of Teredinibacter turnerae T7901: An Intracellular Endosymbiont of Marine Wood-Boring Bivalves (Shipworms). *Plos One* **2009,** *4* (7).
100. Lechene, C. P.; Luyten, Y.; McMahon, G.; Distel, D. L., Quantitative imaging of nitrogen fixation by individual bacteria within animal cells. *Science* **2007,** *317* (5844), 1563-1566.
101. O'Connor, R. M.; Fung, J. M.; Sharp, K. H.; Benner, J. S.; McClung, C.; Cushing, S.; Lamkin, E. R.; Fomenkov, A. I.; Henrissat, B.; Londer, Y. Y.; Scholz, M. B.; Posfai, J.; Malfatti, S.; Tringe, S. G.; Woyke, T.; Malmstrom, R. R.; Coleman-Derr, D.; Altamia, M. A.; Dedrick, S.; Kaluziak, S. T.; Haygood, M. G.; Distel, D. L., Gill bacteria enable a novel digestive strategy in a wood-feeding mollusk. *Proceedings of the National Academy of Sciences* **2014**.
102. Distel, D. L.; Altamia, M. A.; Lin, Z. J.; Shipway, J. R.; Han, A.; Forteza, I.; Antemano, R.; Limbaco, M.; Tebo, A. G.; Dechavez, R.; Albano, J.; Rosenberg, G.; Concepcion, G. P.; Schmidt, E. W.; Haygood, M. G., Discovery of chemoautotrophic symbiosis in the giant shipworm Kuphus polythalamia (Bivalvia: Teredinidae) extends wooden-steps theory. *Proceedings of the National Academy of Sciences of the United States of America* **2017,** *114* (18), E3652-E3658.
103. Distel, D. L.; Morrill, W.; MacLaren-Toussaint, N.; Franks, D.; Waterbury, J., Teredinibacter turnerae gen. nov., sp nov., a dinitrogen-fixing, cellulolytic, endosymbiotic gamma-proteobacterium isolated from the gills of wood-boring molluscs (Bivalvia : Teredinidae). *International Journal of Systematic and Evolutionary Microbiology* **2002,** *52*, 2261-2269.
104. Yang, J. C.; Madupu, R.; Durkin, A. S.; Ekborg, N. A.; Pedamallu, C. S.; Hostetler, J. B.; Radune, D.; Toms, B. S.; Henrissat, B.; Coutinho, P. M.; Schwarz, S.; Field, L.; Trindade-Silva,

A. E.; Soares, C. A. G.; Elshahawi, S.; Hanora, A.; Schmidt, E. W.; Haygood, M. G.; Posfai, J.; Benner, J.; Madinger, C.; Nove, J.; Anton, B.; Chaudhary, K.; Foster, J.; Holman, A.; Kumar, S.; Lessard, P. A.; Luyten, Y. A.; Slatko, B.; Wood, N.; Wu, B.; Teplitski, M.; Mougous, J. D.; Ward, N.; Eisen, J. A.; Badger, J. H.; Distel, D. L., The Complete Genome of Teredinibacter turnerae T7901: An Intracellular Endosymbiont of Marine Wood-Boring Bivalves (Shipworms). *PLOS ONE* **2009,** *4* (7), e6085.

105. Ekborg, N. A.; Morrill, W.; Burgoyne, A. M.; Li, L.; Distell, D. L., CelAB, a multifunctional cellulase encoded by Teredinibacter turnerae T7902(T), a culturable symbiont isolated from the wood-boring marine bivalve Lyrodus pedicellatus. *Applied and Environmental Microbiology* **2007,** *73* (23), 7785-7788.

106. Tano, T.; Okubo, Y.; Kunishita, A.; Kubo, M.; Sugimoto, H.; Fujieda, N.; Ogura, T.; Itoh, S., Redox properties of a mononuclear copper(II)-superoxide complex. *Inorganic chemistry* **2013,** *52* (18), 10431-7.

107. Lima Brito, T.; Barreto Campos, A.; Bastiaan von Meijenfeldt, F. A.; Paulino Daniel, J.; Ribeiro, G. B.; Gueiros Zacarias Silva, G.; Morais, D.; Veras Wilke, D.; Dutilh, B. E.; Meirelles, P. M.; Trindade Silva, A. E., The gill-associated symbiont microbiome is a main source of woody-plant polysaccharide hydrolase genes and secondary metabolite gene clusters in Neoteredo reynei, a unique shipworm from south Atlantic mangroves. *bioRxiv* **2018**.

108. O'Connor, R. M.; Fung, J. M.; Sharp, K. H.; Benner, J. S.; McClung, C.; Cushing, S.; Lamkin, E. R.; Fomenkov, A. I.; Henrissat, B.; Londer, Y. Y.; Scholz, M. B.; Posfai, J.; Malfatti, S.; Tringe, S. G.; Woyke, T.; Malmstrom, R. R.; Coleman-Derr, D.; Altamia, M. A.; Dedrick, S.; Kaluziak, S. T.; Haygood, M. G.; Distel, D. L., Gill bacteria enable a novel digestive strategy in a wood-feeding mollusk. *Proceedings of the National Academy of Sciences* **2014,** *111* (47), E5096-E5104.

109. Senra, M. V. X.; Sung, W.; Ackerman, M.; Miller, S. F.; Lynch, M.; Soares, C. A. G., An Unbiased Genome-Wide View of the Mutation Rate and Spectrum of the Endosymbiotic Bacterium Teredinibacter turnerae. *Genome Biology and Evolution* **2018,** *10* (3), 723-730.

110. A., A. M.; Nicole, W.; M., F. J.; Sandra, D.; W., L. E.; P., C. G.; G., H. M.; L., D. D., Genetic differentiation among isolates of Teredinibacter turnerae, a widely occurring intracellular endosymbiont of shipworms. *Molecular Ecology* **2014,** *23* (6), 1418-1432.

111. Sipe, A. R.; Wilbur, A. E.; Cary, S. C., Bacterial symbiont transmission in the wood-boring shipworm Bankia setacea (Bivalvia : Teredinidae). *Applied and Environmental Microbiology* **2000,** *66* (4), 1685-1691.

112. Altschul, S. F.; Gish, W.; Miller, W.; Myers, E. W.; Lipman, D. J., Basic local alignment search tool. *J Mol Biol* **1990,** *215* (3), 403-10.

113. Boratyn, G. M.; Camacho, C.; Cooper, P. S.; Coulouris, G.; Fong, A.; Ma, N.; Madden, T. L.; Matten, W. T.; McGinnis, S. D.; Merezhuk, Y.; Raytselis, Y.; Sayers, E. W.; Tao, T.; Ye, J.; Zaretskaya, I., BLAST: a more efficient report with usability improvements. *Nucleic Acids Res* **2013,** *41* (Web Server issue), W29-33.

114. Butt, T. R.; Edavettal, S. C.; Hall, J. P.; Mattern, M. R., SUMO fusion technology for difficult-to-express proteins. *Protein Expression and Purification* **2005,** *43* (1), 1-9.

115. Malakhov, M. P.; Mattern, M. R.; Malakhova, O. A.; Drinker, M.; Weeks, S. D.; Butt, T. R., SUMO fusions and SUMO-specific protease for efficient expression and purification of proteins. *Journal of Structural and Functional Genomics* **2004,** *5* (1-2), 75-86.

116. Marblestone, J. G.; Edavettal, S. C.; Lim, Y.; Lim, P.; Zuo, X.; Butt, T. R., Comparison of SUMO fusion technology with traditional gene fusion systems: Enhanced expression and solubility with SUMO. *Protein Science* **2006,** *15* (1), 182-189.

117. Panavas, T.; Sanders, C.; Butt, T. R., SUMO fusion technology for enhanced protein production in prokaryotic and eukaryotic expression systems. *Methods in molecular biology (Clifton, N.J.)* **2009,** *497*, 303-17.

118. Peroutka Iii, R.; Orcutt, S.; Strickler, J.; Butt, T., SUMO Fusion Technology for Enhanced Protein Expression and Purification in Prokaryotes and Eukaryotes. In *Heterologous Gene Expression in E.coli*, Evans, J. T. C.; Xu, M.-Q., Eds. Humana Press: 2011; Vol. 705, pp 15-30.
119. Gregory, R. C.; Hemsworth, G. R.; Turkenburg, J. P.; Hart, S. J.; Walton, P. H.; Davies, G. J., Activity, stability and 3-D structure of the Cu(II) form of a chitin-active lytic polysaccharide monooxygenase from Bacillus amyloliquefaciens. *Dalton Transactions* **2016,** *45* (42), 16904-16912.
120. Nishihara, K.; Kanemori, M.; Kitagawa, M.; Yanagi, H.; Yura, T., Chaperone coexpression plasmids: differential and synergistic roles of DnaK-DnaJ-GrpE and GroEL-GroES in assisting folding of an allergen of Japanese cedar pollen, Cryj2, in Escherichia coli. *Appl Environ Microbiol* **1998,** *64* (5), 1694-9.
121. Nishihara, K.; Kanemori, M.; Yanagi, H.; Yura, T., Overexpression of Trigger Factor Prevents Aggregation of Recombinant Proteins in Escherichia coli. *Applied and Environmental Microbiology* **2000,** *66* (3), 884-889.
122. Frey, S.; Goerlich, D., A new set of highly efficient, tag-cleaving proteases for purifying recombinant proteins. *Journal of Chromatography A* **2014,** *1337*, 95-105.
123. de Marco, A.; Vigh, L.; Diamant, S.; Goloubinoff, P., Native folding of aggregation-prone recombinant proteins in Escherichia coli by osmolytes, plasmid- or benzyl alcohol–overexpressed molecular chaperones. *Cell Stress & Chaperones* **2005,** *10* (4), 329-339.
124. Oganesyan, N.; Ankoudinova, I.; Kim, S.-H.; Kim, R., Effect of Osmotic Stress and Heat Shock in Recombinant Protein Overexpression and Crystallization. *Protein expression and purification* **2007,** *52* (2), 280-285.
125. Aspeborg, H.; Coutinho, P. M.; Wang, Y.; Brumer, H., III; Henrissat, B., Evolution, substrate specificity and subfamily classification of glycoside hydrolase family 5 (GH5). *Bmc Evolutionary Biology* **2012,** *12*.
126. Naas, A. E.; MacKenzie, A. K.; Dalhus, B.; Eijsink, V. G.; Pope, P. B., Structural Features of a Bacteroidetes-Affiliated Cellulase Linked with a Polysaccharide Utilization Locus. *Sci Rep* **2015,** *5*, 11666.
127. Davies, G. J.; Dauter, M.; Brzozowski, A. M.; Bjornvad, M. E.; Andersen, K. V.; Schulein, M., Structure of the Bacillus agaradherans family 5 endoglucanase at 1.6 A and its cellobiose complex at 2.0 A resolution. *Biochemistry* **1998,** *37* (7), 1926-32.
128. Dominguez, R.; Souchon, H.; Spinelli, S.; Dauter, Z.; Wilson, K. S.; Chauvaux, S.; Beguin, P.; Alzari, P. M., A common protein fold and similar active site in two distinct families of beta-glycanases. *Nature structural biology* **1995,** *2* (7), 569-76.
129. Henrissat, B.; Callebaut, I.; Fabrega, S.; Lehn, P.; Mornon, J. P.; Davies, G., Conserved catalytic machinery and the prediction of a common fold for several families of glycosyl hydrolases. *Proc Natl Acad Sci U S A* **1995,** *92* (15), 7090-4.
130. Wierenga, R. K., The TIM-barrel fold: a versatile framework for efficient enzymes. *FEBS Letters* **2001,** *492* (3), 193-198.
131. Silverman, J. A.; Balakrishnan, R.; Harbury, P. B., Reverse engineering the (β/α)₈ barrel fold. *Proceedings of the National Academy of Sciences* **2001,** *98* (6), 3092-3097.
132. Almeida, V. M.; Frutuoso, M. A.; Marana, S. R., Search for independent (β/α)4 subdomains in a (β/α)8 barrel β-glucosidase. *PLOS ONE* **2018,** *13* (1), e0191282.
133. Lang, D.; Thoma, R.; Henn-Sax, M.; Sterner, R.; Wilmanns, M., Structural evidence for evolution of the beta/alpha barrel scaffold by gene duplication and fusion. *Science* **2000,** *289* (5484), 1546-50.
134. Rempel, B. P.; Withers, S. G., Covalent inhibitors of glycosidases and their applications in biochemistry and biology. *Glycobiology* **2008,** *18* (8), 570-586.

135. Davies, G. J.; Mackenzie, L.; Varrot, A.; Dauter, M.; Brzozowski, A. M.; Schulein, M.; Withers, S. G., Snapshots along an enzymatic reaction coordinate: analysis of a retaining beta-glycoside hydrolase. *Biochemistry* **1998,** *37* (34), 11707-13.
136. Jenkins, J.; Lo Leggio, L.; Harris, G.; Pickersgill, R., Beta-glucosidase, beta-galactosidase, family A cellulases, family F xylanases and two barley glycanases form a superfamily of enzymes with 8-fold beta/alpha architecture and with two conserved glutamates near the carboxy-terminal ends of beta-strands four and seven. *FEBS Lett* **1995,** *362* (3), 281-5.
137. Aspeborg, H.; Coutinho, P. M.; Wang, Y.; Brumer, H., 3rd; Henrissat, B., Evolution, substrate specificity and subfamily classification of glycoside hydrolase family 5 (GH5). *BMC evolutionary biology* **2012,** *12*, 186.
138. Fry, S.; S. York, W.; Albersheim, P.; Darvill, A.; Hayashi, T.; Joseleau, J. P.; Kato, Y.; Perez Lorences, E.; A. Maclachlan, G.; McNeil, M.; J. Mort, A.; S. Grant Reid, J.; Ulrich Seitz, H.; R. Selvendran, R.; Voragen, A.; White, A., *An Unambiguous nomenclature for xyloglucan-derived oligosaccharides*. 1993; Vol. 89.
139. Varki, A.; Cummings, R. D.; Esko, J. D.; Freeze, H. H.; Stanley, P.; Marth, J. D.; Bertozzi, C. R.; Hart, G. W.; Etzler, M. E., Symbol nomenclature for glycan representation. *Proteomics* **2009,** *9* (24), 5398-9.
140. Goubet, F.; Jackson, P.; Deery, M. J.; Dupree, P., Polysaccharide analysis using carbohydrate gel electrophoresis: a method to study plant cell wall polysaccharides and polysaccharide hydrolases. *Anal Biochem* **2002,** *300* (1), 53-68.
141. McGregor, N.; Morar, M.; Fenger, T. H.; Stogios, P.; Lenfant, N.; Yin, V.; Xu, X.; Evdokimova, E.; Cui, H.; Henrissat, B.; Savchenko, A.; Brumer, H., Structure-Function Analysis of a Mixed-linkage β-Glucanase/Xyloglucanase from the Key Ruminal Bacteroidetes Prevotella bryantii B14. *Journal of Biological Chemistry* **2016,** *291* (3), 1175-1197.
142. Gloster, T. M.; Ibatullin, F. M.; Macauley, K.; Eklöf, J. M.; Roberts, S.; Turkenburg, J. P.; Bjørnvad, M. E.; Jørgensen, P. L.; Danielsen, S.; Johansen, K. S.; Borchert, T. V.; Wilson, K. S.; Brumer, H.; Davies, G. J., Characterization and Three-dimensional Structures of Two Distinct Bacterial Xyloglucanases from Families GH5 and GH12. *Journal of Biological Chemistry* **2007,** *282* (26), 19177-19189.
143. Hsieh, Y.; Harris, P., *Xyloglucans of Monocotyledons Have Diverse Structures*. 2009; Vol. 2, p 943-65.
144. Kato, K.; Matsuda, K., Studies on the Chemical Structure of Konjac Mannan. *Agricultural and Biological Chemistry* **1969,** *33* (10), 1446-1453.
145. Gao, J.; Wakarchuk, W., Characterization of Five β-Glycoside Hydrolases from Cellulomonas fimi ATCC484. *Journal of bacteriology* **2014**, JB. 02194-14.
146. Tuomivaara, S. T.; Yaoi, K.; O'Neill, M. A.; York, W. S., Generation and structural validation of a library of diverse xyloglucan-derived oligosaccharides, including an update on xyloglucan nomenclature. *Carbohydrate research* **2015,** *402*, 56-66.
147. Larsbrink, J.; Rogers, T. E.; Hemsworth, G. R.; McKee, L. S.; Tauzin, A. S.; Spadiut, O.; Klinter, S.; Pudlo, N. A.; Urs, K.; Koropatkin, N. M., A discrete genetic locus confers xyloglucan metabolism in select human gut Bacteroidetes. *Nature* **2014,** *506* (7489), 498.
148. Gloster, T. M.; Ibatullin, F. M.; Macauley, K.; Eklof, J. M.; Roberts, S.; Turkenburg, J. P.; Bjornvad, M. E.; Jorgensen, P. L.; Danielsen, S.; Johansen, K. S.; Borchert, T. V.; Wilson, K. S.; Brumer, H.; Davies, G. J., Characterization and three-dimensional structures of two distinct bacterial xyloglucanases from families GH5 and GH12. *The Journal of biological chemistry* **2007,** *282* (26), 19177-89.
149. Burton, R. A.; Fincher, G. B., (1,3;1,4)-beta-D-glucans in cell walls of the poaceae, lower plants, and fungi: a tale of two linkages. *Molecular plant* **2009,** *2* (5), 873-82.

150. McGregor, N.; Morar, M.; Fenger, T. H.; Stogios, P.; Lenfant, N.; Yin, V.; Xu, X.; Evdokimova, E.; Cui, H.; Henrissat, B.; Savchenko, A.; Brumer, H., Structure-Function Analysis of a Mixed-linkage β-Glucanase/Xyloglucanase from the Key Ruminal Bacteroidetes Prevotella bryantii B(1)4. *The Journal of biological chemistry* **2016,** *291* (3), 1175-1197.
151. Sulzenbacher, G.; Shareck, F.; Morosoli, R.; Dupont, C.; Davies, G. J., The Streptomyces lividans family 12 endoglucanase: construction of the catalytic cre, expression, and X-ray structure at 1.75 A resolution. *Biochemistry* **1997,** *36* (51), 16032-9.
152. Saarilahti, H. T.; Henrissat, B.; Palva, E. T., CELS - A novel endoglucanase identified from erwinia-carotovora subsp carotovora. *Gene* **1990,** *90* (1), 9-14.
153. Torronen, A.; Harkki, A.; Rouvinen, J., 3-dimensional structure of endo-1,4-beta-xylanase-ii from trichoderma-reesei Conformational states in the active-site. *Embo Journal* **1994,** *13* (11), 2493-2501.
154. Miao, S. C.; Ziser, L.; Aebersold, R.; Withers, S. G., Identification of glutamic-acid-78 as the active-site nucleophile in bacillus-subtilis xylanase using electrospray tandem mass-spectrometry. *Biochemistry* **1994,** *33* (23), 7027-7032.
155. Okada, H.; Mori, K.; Tada, K.; Nogawa, M.; Morikawa, Y., Identification of active site carboxylic residues in Trichoderma reesei endoglucanase Cel12A by site-directed mutagenesis. *Journal of Molecular Catalysis B-Enzymatic* **2000,** *10* (1-3), 249-255.
156. Winn, M. D.; Ballard, C. C.; Cowtan, K. D.; Dodson, E. J.; Emsley, P.; Evans, P. R.; Keegan, R. M.; Krissinel, E. B.; Leslie, A. G.; McCoy, A.; McNicholas, S. J.; Murshudov, G. N.; Pannu, N. S.; Potterton, E. A.; Powell, H. R.; Read, R. J.; Vagin, A.; Wilson, K. S., Overview of the CCP4 suite and current developments. *Acta crystallographica. Section D, Biological crystallography* **2011,** *67*, 235-242.
157. McCoy, A. J.; Grosse-Kunstleve, R. W.; Adams, P. D.; Winn, M. D.; Storoni, L. C.; Read, R. J., Phaser crystallographic software. *J Appl Crystallogr* **2007,** *40*, 658-674.
158. Cowtan, K., The Buccaneer software for automated model building. 1. Tracing protein chains. *Acta crystallographica. Section D, Biological crystallography* **2006,** *62*, 1002.
159. Emsley, P.; Lohkamp, B.; Scott, W. G.; Cowtan, K., Features and development of Coot. *Acta crystallographica. Section D, Biological crystallography* **2010,** *66*, 486-501.
160. Murshudov, G. N.; Skubák, P.; Lebedev, A. A.; Pannu, N. S.; Steiner, R. A.; Nicholls, R. A.; Winn, M. D.; Long, F.; Vagin, A. A., REFMAC5 for the refinement of macromolecular crystal structures. *Acta crystallographica. Section D, Biological crystallography* **2011,** *67*, 355-367.
161. Lebedev, A. A.; Young, P.; Isupov, M. N.; Moroz, O. V.; Vagin, A. A.; Murshudov, G. N., JLigand: a graphical tool for the CCP4 template-restraint library. *Acta crystallographica. Section D, Biological crystallography* **2012,** *68*, 431-440.
162. Murshudov, G. N.; Vagin, A. A.; Dodson, E. J., Refinement of macromolecular structures by the maximum-likelihood method. *Acta crystallographica. Section D, Biological crystallography* **1997,** *53*, 240-255.
163. Damasio, A. R. L.; Ribeiro, L. F. C.; Ribeiro, L. F.; Furtado, G. P.; Segato, F.; Almeida, F. B. R.; Crivellari, A. C.; Buckeridge, M. S.; Souza, T.; Murakami, M. T.; Ward, R. J.; Prade, R. A.; Polizeli, M., Functional characterization and oligomerization of a recombinant xyloglucan-specific endo-beta-1,4-glucanase (GH12) from Aspergillus niveus. *Biochimica Et Biophysica Acta-Proteins and Proteomics* **2012,** *1824* (3), 461-467.
164. Gasteiger, E.; Hoogland, C.; Gattiker, A.; Duvaud, S. e.; Wilkins, M. R.; Appel, R. D.; Bairoch, A., Protein Identification and Analysis Tools on the ExPASy Server. In *The Proteomics Protocols Handbook*, Walker, J. M., Ed. Humana Press: Totowa, NJ, 2005; pp 571-607.

165. Liu, H. Z.; Liang, X. F.; Sohoel, H.; Bulow, A.; Bols, M., Noeuromycin, a glycosyl cation mimic that strongly inhibits glycosidases. *Journal of the American Chemical Society* **2001,** *123* (21), 5116-5117.
166. Cheng, Y.-S.; Ko, T.-P.; Huang, J.-W.; Wu, T.-H.; Lin, C.-Y.; Luo, W.; Li, Q.; Ma, Y.; Huang, C.-H.; Wang, A. H.-J.; Liu, J.-R.; Guo, R.-T., Enhanced activity of Thermotoga maritima cellulase 12A by mutating a unique surface loop. *Applied Microbiology and Biotechnology* **2012,** *95* (3), 661-669.
167. Fierobe, H. P.; Bagnara-Tardif, C.; Gaudin, C.; Guerlesquin, F.; Sauve, P.; Belaich, A.; Belaich, J. P., Purification and characterization of endoglucanase C from Clostridium cellulolyticum. Catalytic comparison with endoglucanase A. *Eur J Biochem* **1993,** *217* (2), 557-65.
168. Petersen, L.; Ardevol, A.; Rovira, C.; Reilly, P. J., Mechanism of cellulose hydrolysis by inverting GH8 endoglucanases: a QM/MM metadynamics study. *The journal of physical chemistry. B* **2009,** *113* (20), 7331-9.
169. Parsiegla, G.; Reverbel-Leroy, C.; Tardif, C.; Belaich, J. P.; Driguez, H.; Haser, R., Crystal structures of the cellulase Cel48F in complex with inhibitors and substrates give insights into its processive action. *Biochemistry* **2000,** *39* (37), 11238-46.
170. Barr, B. K.; Hsieh, Y. L.; Ganem, B.; Wilson, D. B., Identification of two functionally different classes of exocellulases. *Biochemistry* **1996,** *35* (2), 586-92.
171. Guimaraes, B. G.; Souchon, H.; Lytle, B. L.; David Wu, J. H.; Alzari, P. M., The crystal structure and catalytic mechanism of cellobiohydrolase CelS, the major enzymatic component of the Clostridium thermocellum Cellulosome. *J Mol Biol* **2002,** *320* (3), 587-96.
172. Alzari, P. M.; ne Souchon, H.; Dominguez, R., The crystal structure of endoglucanase CelA, a family 8 glycosyl hydrolase from Clostridium thermocellum. *Structure* **1996,** *4* (3), 265-275.
173. Guerin, D. M.; Lascombe, M. B.; Costabel, M.; Souchon, H.; Lamzin, V.; Beguin, P.; Alzari, P. M., Atomic (0.94 A) resolution structure of an inverting glycosidase in complex with substrate. *J Mol Biol* **2002,** *316* (5), 1061-9.
174. Scapin, S. M. N.; Souza, F. H. M.; Zanphorlin, L. M.; de Almeida, T. S.; Sade, Y. B.; Cardoso, A. M.; Pinheiro, G. L.; Murakami, M. T., Structure and function of a novel GH8 endoglucanase from the bacterial cellulose synthase complex of Raoultella ornithinolytica. *PLoS One* **2017,** *12* (4), e0176550.
175. Adachi, W.; Sakihama, Y.; Shimizu, S.; Sunami, T.; Fukazawa, T.; Suzuki, M.; Yatsunami, R.; Nakamura, S.; Takenaka, A., Crystal structure of family GH-8 chitosanase with subclass II specificity from Bacillus sp. K17. *J Mol Biol* **2004,** *343* (3), 785-95.
176. Alzari, P. M.; Souchon, H.; Dominguez, R., The crystal structure of endoglucanase CelA, a family 8 glycosyl hydrolase from Clostridium thermocellum. *Structure* **1996,** *4* (3), 265-75.
177. Davies, G. J.; Wilson, K. S.; Henrissat, B., Nomenclature for sugar-binding subsites in glycosyl hydrolases. *Biochemical Journal* **1997,** *321* (2), 557-559.
178. De Vos, D.; Collins, T.; Nerinckx, W.; Savvides, S. N.; Claeyssens, M.; Gerday, C.; Feller, G.; Van Beeumen, J., Oligosaccharide Binding in Family 8 Glycosidases: Crystal Structures of Active-Site Mutants of the β-1,4-Xylanase pXyl from Pseudoaltermonas haloplanktis TAH3a in Complex with Substrate and Product. *Biochemistry* **2006,** *45* (15), 4797-4807.
179. Adachi, W.; Sakihama, Y.; Shimizu, S.; Sunami, T.; Fukazawa, T.; Suzuki, M.; Yatsunami, R.; Nakamura, S.; Takénaka, A., Crystal Structure of Family GH-8 Chitosanase with Subclass II Specificity from Bacillus sp. K17. *Journal of Molecular Biology* **2004,** *343* (3), 785-795.
180. Vagin, A.; Teplyakov, A., Molecular replacement with MOLREP. *Acta crystallographica. Section D, Biological crystallography* **2010,** *66* (Pt 1), 22-5.

181. McNicholas, S.; Potterton, E.; Wilson, K. S.; Noble, M. E. M., Presenting your structures: the CCP4mg molecular-graphics software. *Acta Crystallographica Section D-Biological Crystallography* **2011,** *67*, 386-394.
182. Fowler, C. A., Hemsworth, G.R., Cuskin, F., Hart, S., Turkenburg, J., Gilbert, H., Walton, P.H. & Davies, G.J., Structure and function of a glycoside hydrolase family 8 endoxylanase from Teredinibacter turnerae. *Acta Cryst. D* **2018,** (74).
183. Pell, G.; Szabo, L.; Charnock, S. J.; Xie, H.; Gloster, T. M.; Davies, G. J.; Gilbert, H. J., Structural and Biochemical Analysis of Cellvibrio japonicus Xylanase 10C: How variation in substrate-binding cleft influences the catalytic profile of family gh-10 xylanases. *Journal of Biological Chemistry* **2004,** *279* (12), 11777-11788.
184. Charnock, S. J.; Spurway, T. D.; Xie, H.; Beylot, M. H.; Virden, R.; Warren, R. A.; Hazlewood, G. P.; Gilbert, H. J., The topology of the substrate binding clefts of glycosyl hydrolase family 10 xylanases are not conserved. *The Journal of biological chemistry* **1998,** *273* (48), 32187-99.
185. Matsui, I.; Ishikawa, K.; Matsui, E.; Miyairi, S.; Fukui, S.; Honda, K., Subsite structure of Saccharomycopsis alpha-amylase secreted from Saccharomyces cerevisiae. *Journal of biochemistry* **1991,** *109* (4), 566-9.
186. Fushinobu, S.; Hidaka, M.; Honda, Y.; Wakagi, T.; Shoun, H.; Kitaoka, M., Structural basis for the specificity of the reducing end xylose-releasing exo-oligoxylanase from Bacillus halodurans C-125. *The Journal of biological chemistry* **2005,** *280* (17), 17180-6.
187. Pollet, A.; Schoepe, J.; Dornez, E.; Strelkov, S. V.; Delcour, J. A.; Courtin, C. M., Functional analysis of glycoside hydrolase family 8 xylanases shows narrow but distinct substrate specificities and biotechnological potential. *Applied Microbiology and Biotechnology* **2010,** *87* (6), 2125-2135.
188. Viana, A. G.; Noseda, M. D.; Gonçalves, A. G.; Duarte, M. E. R.; Yokoya, N.; Matulewicz, M. C.; Cerezo, A. S., β-d-(1→4), β-d-(1→3) 'mixed linkage' xylans from red seaweeds of the order Nemaliales and Palmariales. *Carbohydrate research* **2011,** *346* (8), 1023-1028.
189. Cerezo, A. S.; Lezerovich, A.; Labriola, R.; Rees, D. A., A xylan from the red seaweed chaetangium fastigiatum. *Carbohydrate research* **1971,** *19* (3), 289-296.
190. Collins, T.; Gerday, C.; Feller, G., Xylanases, xylanase families and extremophilic xylanases. *FEMS microbiology reviews* **2005,** *29* (1), 3-23.
191. Van Petegem, F.; Collins, T.; Meuwis, M. A.; Gerday, C.; Feller, G.; Van Beeumen, J., The structure of a cold-adapted family 8 xylanase at 1.3 A resolution. Structural adaptations to cold and investgation of the active site. *The Journal of biological chemistry* **2003,** *278* (9), 7531-9.
192. Speciale, G.; Thompson, A. J.; Davies, G. J.; Williams, S. J., Dissecting conformational contributions to glycosidase catalysis and inhibition. *Current Opinion in Structural Biology* **2014,** *28*, 1-13.
193. Sabini, E.; Sulzenbacher, G.; Dauter, M.; Dauter, Z.; Jørgensen, P. L.; Schülein, M.; Dupont, C.; Davies, G. J.; Wilson, K. S., Catalysis and specificity in enzymatic glycoside hydrolysis: a 2,5B conformation for the glycosyl-enzyme intermediate revealed by the structure of the Bacillus agaradhaerens family 11 xylanase. *Chemistry & Biology* **1999,** *6* (7), 483-492.
194. Solomon, E. I.; Heppner, D. E.; Johnston, E. M.; Ginsbach, J. W.; Cirera, J.; Qayyum, M.; Kieber-Emmons, M. T.; Kjaergaard, C. H.; Hadt, R. G.; Tian, L., Copper Active Sites in Biology. *Chemical Reviews* **2014,** *114* (7), 3659-3853.
195. Quinlan, R. J.; Sweeney, M. D.; Lo Leggio, L.; Otten, H.; Poulsen, J.-C. N.; Johansen, K. S.; Krogh, K. B. R. M.; Jorgensen, C. I.; Tovborg, M.; Anthonsen, A.; Tryfona, T.; Walter, C. P.; Dupree, P.; Xu, F.; Davies, G. J.; Walton, P. H., Insights into the oxidative degradation of

cellulose by a copper metalloenzyme that exploits biomass components. *Proceedings of the National Academy of Sciences of the United States of America* **2011,** *108* (37), 15079-15084.
196. Hemsworth, G. R.; Davies, G. J.; Walton, P. H., Recent insights into copper-containing lytic polysaccharide mono-oxygenases. *Current opinion in structural biology* **2013,** *23* (5), 660-8.
197. Book, A. J.; Yennamalli, R. M.; Takasuka, T. E.; Currie, C. R.; Phillips, G. N.; Fox, B. G., Evolution of substrate specificity in bacterial AA10 lytic polysaccharide monooxygenases. *Biotechnology for Biofuels* **2014,** *7* (1), 109.
198. Agostoni, M.; Hangasky, J. A.; Marletta, M. A., Physiological and Molecular Understanding of Bacterial Polysaccharide Monooxygenases. *Microbiology and Molecular Biology Reviews* **2017,** *81* (3).
199. Simmons, T. J.; Frandsen, K. E. H.; Ciano, L.; Tryfona, T.; Lenfant, N.; Poulsen, J. C.; Wilson, L. F. L.; Tandrup, T.; Tovborg, M.; Schnorr, K.; Johansen, K. S.; Henrissat, B.; Walton, P. H.; Lo Leggio, L.; Dupree, P., Structural and electronic determinants of lytic polysaccharide monooxygenase reactivity on polysaccharide substrates. *Nature Communications* **2017,** *8* (1), 1064.
200. Forsberg, Z.; Røhr, Å. K.; Mekasha, S.; Andersson, K. K.; Eijsink, V. G. H.; Vaaje-Kolstad, G.; Sørlie, M., Comparative Study of Two Chitin-Active and Two Cellulose-Active AA10-Type Lytic Polysaccharide Monooxygenases. *Biochemistry* **2014**.
201. Frandsen, K. E.; Simmons, T. J.; Dupree, P.; Poulsen, J. C.; Hemsworth, G. R.; Ciano, L.; Johnston, E. M.; Tovborg, M.; Johansen, K. S.; von Freiesleben, P.; Marmuse, L.; Fort, S.; Cottaz, S.; Driguez, H.; Henrissat, B.; Lenfant, N.; Tuna, F.; Baldansuren, A.; Davies, G. J.; Lo Leggio, L.; Walton, P. H., The molecular basis of polysaccharide cleavage by lytic polysaccharide monooxygenases. *Nat Chem Biol* **2016,** *12* (4), 298-303.
202. Bissaro, B.; Isaksen, I.; Vaaje-Kolstad, G.; Eijsink, V. G. H.; Røhr, Å. K., How a Lytic Polysaccharide Monooxygenase Binds Crystalline Chitin. *Biochemistry* **2018,** *57* (12), 1893-1906.
203. Yakovlev, I.; Vaaje-Kolstad, G.; Hietala, A. M.; Stefanczyk, E.; Solheim, H.; Fossdal, C. G., Substrate-specific transcription of the enigmatic GH61 family of the pathogenic white-rot fungus Heterobasidion irregulare during growth on lignocellulose. *Applied Microbiology and Biotechnology* **2012,** *95* (4), 979-990.
204. Langston, J. A.; Shaghasi, T.; Abbate, E.; Xu, F.; Vlasenko, E.; Sweeney, M. D., Oxidoreductive Cellulose Depolymerization by the Enzymes Cellobiose Dehydrogenase and Glycoside Hydrolase 61. *Applied and Environmental Microbiology* **2011,** *77* (19), 7007-7015.
205. Tan, T. C.; Kracher, D.; Gandini, R.; Sygmund, C.; Kittl, R.; Haltrich, D.; Hallberg, B. M.; Ludwig, R.; Divne, C., Structural basis for cellobiose dehydrogenase action during oxidative cellulose degradation. *Nat Commun* **2015,** *6*, 7542.
206. Loose, J. S. M.; Forsberg, Z.; Kracher, D.; Scheiblbrandner, S.; Ludwig, R.; Eijsink, V. G. H.; Vaaje-Kolstad, G., Activation of bacterial lytic polysaccharide monooxygenases with cellobiose dehydrogenase. *Protein Science : A Publication of the Protein Society* **2016,** *25* (12), 2175-2186.
207. Crouch, L. I.; Labourel, A.; Walton, P. H.; Davies, G. J.; Gilbert, H. J., The Contribution of Non-catalytic Carbohydrate Binding Modules to the Activity of Lytic Polysaccharide Monooxygenases. *The Journal of biological chemistry* **2016,** *291* (14), 7439-49.
208. Millward-Sadler, S. J.; Davidson, K.; Hazlewood, G. P.; Black, G. W.; Gilbert, H. J.; Clarke, J. H., Novel cellulose-binding domains, NodB homologues and conserved modular architecture in xylanases from the aerobic soil bacteria Pseudomonas fluorescens subsp. cellulosa and Cellvibrio mixtus. *Biochem J* **1995,** *312 (Pt 1)*, 39-48.

209. Stoll, S.; Schweiger, A., EasySpin, a comprehensive software package for spectral simulation and analysis in EPR. *Journal of magnetic resonance (San Diego, Calif. : 1997)* **2006,** *178* (1), 42-55.
210. Evans, P. R.; Murshudov, G. N., How good are my data and what is the resolution? *Acta crystallographica. Section D, Biological crystallography* **2013,** *69*, 1204-1214.
211. Sheldrick, G. M., Experimental phasing with SHELXC/D/E: combining chain tracing with density modification. *Acta crystallographica. Section D, Biological crystallography* **2010,** *66*, 479-485.
212. Langer, G.; Cohen, S. X.; Lamzin, V. S.; Perrakis, A., Automated macromolecular model building for X-ray crystallography using ARP/wARP version 7. *Nat Protoc* **2008,** *3*, 1171-1179.
213. Forsberg, Z.; Nelson, C. E.; Dalhus, B.; Mekasha, S.; Loose, J. S.; Crouch, L. I.; Rohr, A. K.; Gardner, J. G.; Eijsink, V. G.; Vaaje-Kolstad, G., Structural and Functional Analysis of a Lytic Polysaccharide Monooxygenase Important for Efficient Utilization of Chitin in Cellvibrio japonicus. *The Journal of biological chemistry* **2016,** *291* (14), 7300-12.
214. Vagin, A.; Teplyakov, A., Molecular replacement with MOLREP. *Acta Crystallogr D Biol Crystallogr* **2010,** *66*, 22-25.
215. Waterbury, J. B.; Calloway, C. B.; Turner, R. D., A cellulolytic nitrogen-fixing bacterium cultured from the gland of Deshayes in shipworms (Bivalvia: Teredinidae). *Science* **1983,** *221* (4618), 1401-1403.
216. Peisach, J.; Blumberg, W. E., Structural implications derived from the analysis of electron paramagnetic resonance spectra of natural and artificial copper proteins. *Archives of Biochemistry and Biophysics* **1974,** *165* (2), 691-708.
217. Chaplin, A. K.; Wilson, M. T.; Hough, M. A.; Svistunenko, D. A.; Hemsworth, G. R.; Walton, P. H.; Vijgenboom, E.; Worrall, J. A., Heterogeneity in the Histidine-brace Copper Coordination Sphere in Auxiliary Activity Family 10 (AA10) Lytic Polysaccharide Monooxygenases. *The Journal of biological chemistry* **2016,** *291* (24), 12838-50.
218. Eibinger, M.; Sattelkow, J.; Ganner, T.; Plank, H.; Nidetzky, B., Single-molecule study of oxidative enzymatic deconstruction of cellulose. *Nature Communications* **2017,** *8* (1), 894.
219. Igarashi, K.; Uchihashi, T.; Koivula, A.; Wada, M.; Kimura, S.; Okamoto, T.; Penttila, M.; Ando, T.; Samejima, M., Traffic jams reduce hydrolytic efficiency of cellulase on cellulose surface. *Science* **2011,** *333* (6047), 1279-82.
220. Loose, J. S. M.; Arntzen, M. Ø.; Bissaro, B.; Ludwig, R.; Eijsink, V. G. H.; Vaaje-Kolstad, G., Multipoint Precision Binding of Substrate Protects Lytic Polysaccharide Monooxygenases from Self-Destructive Off-Pathway Processes. *Biochemistry* **2018,** *57* (28), 4114-4124.
221. Forsberg, Z.; Mackenzie, A. K.; Sørlie, M.; Røhr, Å. K.; Helland, R.; Arvai, A. S.; Vaaje-Kolstad, G.; Eijsink, V. G. H., Structural and functional characterization of a conserved pair of bacterial cellulose-oxidizing lytic polysaccharide monooxygenases. *Proceedings of the National Academy of Sciences of the United States of America* **2014,** *111* (23), 8446-8451.
222. Crouch, L. I.; Labourel, A.; Walton, P. H.; Davies, G. J.; Gilbert, H. J., The contribution of non-catalytic carbohydrate binding modules to the activity lytic polysaccharide monooxygenases. *Journal of Biological Chemistry* **2016**.
223. Courtade, G.; Le, S. B.; Sætrom, G. I.; Brautaset, T.; Aachmann, F. L., A novel expression system for lytic polysaccharide monooxygenases. *Carbohydrate research*.
224. Lo Leggio, L.; Simmons, T. J.; Poulsen, J.-C. N.; Frandsen, K. E. H.; Hemsworth, G. R.; Stringer, M. A.; von Freiesleben, P.; Tovborg, M.; Johansen, K. S.; De Maria, L.; Harris, P. V.; Soong, C.-L.; Dupree, P.; Tryfona, T.; Lenfant, N.; Henrissat, B.; Davies, G. J.; Walton, P. H., Structure and boosting activity of a starch-degrading lytic polysaccharide monooxygenase. *Nature Communications* **2015,** *6*, 5961.

225. Mandels, M.; Reese, E. T., Induction of cellulase in Trichoderma viride as influenced by carbon sources and metals. *J Bacteriol* **1957,** *73* (2), 269-78.
226. Mandels, M.; Weber, J.; Parizek, R., Enhanced Cellulase Production by a Mutant of Trichoderma viride. *Applied Microbiology* **1971,** *21* (1), 152-154.
227. Peterson, R.; Nevalainen, H., Trichoderma reesei RUT-C30--thirty years of strain improvement. *Microbiology (Reading, England)* **2012,** *158* (Pt 1), 58-68.
228. Westereng, B.; Loose, J. S. M.; Vaaje-Kolstad, G.; Aachmann, F. L.; Sørlie, M.; Eijsink, V. G. H., Analytical Tools for Characterizing Cellulose-Active Lytic Polysaccharide Monooxygenases (LPMOs). In *Cellulases: Methods and Protocols*, Lübeck, M., Ed. Springer New York: New York, NY, 2018; pp 219-246.
229. Breslmayr, E.; Hanžek, M.; Hanrahan, A.; Leitner, C.; Kittl, R.; Šantek, B.; Oostenbrink, C.; Ludwig, R., A fast and sensitive activity assay for lytic polysaccharide monooxygenase. *Biotechnology for Biofuels* **2018,** *11*, 79.
230. Wang, D.; Li, J.; Wong, A. C. Y.; Aachmann, F. L.; Hsieh, Y. S. Y., A colorimetric assay to rapidly determine the activities of lytic polysaccharide monooxygenases. *Biotechnology for Biofuels* **2018,** *11* (1), 215.
231. Phillips, C. M.; Iavarone, A. T.; Marletta, M. A., Quantitative proteomic approach for cellulose degradation by Neurospora crassa. *Journal of proteome research* **2011,** *10* (9), 4177-85.
232. Adlakha, N.; Rajacharya, G. H.; Bhatnagar, R., Identification of an endogenous redox partner for lytic polysaccharide monooxygenase-based oxidative cleavage of polysaccharides. *Journal of Biological Chemistry* **2017**.
233. Harris, P. V.; Welner, D.; McFarland, K. C.; Re, E.; Navarro Poulsen, J. C.; Brown, K.; Salbo, R.; Ding, H.; Vlasenko, E.; Merino, S.; Xu, F.; Cherry, J.; Larsen, S.; Lo Leggio, L., Stimulation of lignocellulosic biomass hydrolysis by proteins of glycoside hydrolase family 61: structure and function of a large, enigmatic family. *Biochemistry* **2010,** *49* (15), 3305-16.
234. Chylenski, P.; Petrović, D. M.; Müller, G.; Dahlström, M.; Bengtsson, O.; Lersch, M.; Siika-aho, M.; Horn, S. J.; Eijsink, V. G. H., Enzymatic degradation of sulfite-pulped softwoods and the role of LPMOs. *Biotechnology for Biofuels* **2017,** *10* (1), 177.
235. Kracher, D.; Scheiblbrandner, S.; Felice, A. K.; Breslmayr, E.; Preims, M.; Ludwicka, K.; Haltrich, D.; Eijsink, V. G.; Ludwig, R., Extracellular electron transfer systems fuel cellulose oxidative degradation. *Science* **2016,** *352* (6289), 1098-101.
236. Bissaro, B.; Røhr, Å. K.; Müller, G.; Chylenski, P.; Skaugen, M.; Forsberg, Z.; Horn, S. J.; Vaaje-Kolstad, G.; Eijsink, V. G. H., Oxidative cleavage of polysaccharides by monocopper enzymes depends on H2O2. *Nature Chemical Biology* **2017,** *13*, 1123.
237. Hangasky, J. A.; Iavarone, A. T.; Marletta, M. A., Reactivity of O₂ versus H₂O₂ with polysaccharide monooxygenases. *Proceedings of the National Academy of Sciences* **2018,** *115* (19), 4915-4920.
238. Dhar, D.; Tolman, W. B., Hydrogen Atom Abstraction from Hydrocarbons by a Copper(III)-Hydroxide Complex. *Journal of the American Chemical Society* **2015,** *137* (3), 1322-1329.
239. Dhar, D.; Yee, G. M.; Spaeth, A. D.; Boyce, D. W.; Zhang, H.; Dereli, B.; Cramer, C. J.; Tolman, W. B., Perturbing the Copper(III)–Hydroxide Unit through Ligand Structural Variation. *Journal of the American Chemical Society* **2016,** *138* (1), 356-368.
240. Cao, L.; Caldararu, O.; Rosenzweig, A. C.; Ryde, U., Quantum Refinement Does Not Support Dinuclear Copper Sites in Crystal Structures of Particulate Methane Monooxygenase. *Angewandte Chemie International Edition* **2018,** *57* (1), 162-166.
241. Cao, L.; Caldararu, O.; Rosenzweig, A. C.; Ryde, U., Quantum Refinement Does Not Support Dinuclear Copper Sites in Crystal Structures of Particulate Methane Monooxygenase. *Angewandte Chemie (International ed. in English)* **2018,** *57* (1), 162-166.

242. Distel, D. L.; Amin, M.; Burgoyne, A.; Linton, E.; Mamangkey, G.; Morrill, W.; Nove, J.; Wood, N.; Yang, J., Molecular phylogeny of Pholadoidea Lamarck, 1809 supports a single origin for xylotrophy (wood feeding) and xylotrophic bacterial endosymbiosis in Bivalvia. *Molecular Phylogenetics and Evolution* **2011,** *61* (2), 245-254.
243. Turner, R. D., Wood-boring bivalves, opportunistic species in the deep sea. *Science* **1973,** *180* (4093), 1377-9.
244. Distel, D. L.; Altamia, M. A.; Lin, Z.; Shipway, J. R.; Han, A.; Forteza, I.; Antemano, R.; Limbaco, M. G. J. P.; Tebo, A. G.; Dechavez, R.; Albano, J.; Rosenberg, G.; Concepcion, G. P.; Schmidt, E. W.; Haygood, M. G., Discovery of chemoautotrophic symbiosis in the giant shipworm Kuphus polythalamia (Bivalvia: Teredinidae) extends wooden-steps theory. *Proceedings of the National Academy of Sciences* **2017,** *114* (18), E3652-E3658.
245. Zhang, J.; Presley, G. N.; Hammel, K. E.; Ryu, J. S.; Menke, J. R.; Figueroa, M.; Hu, D.; Orr, G.; Schilling, J. S., Localizing gene regulation reveals a staggered wood decay mechanism for the brown rot fungus Postia placenta. *Proc Natl Acad Sci U S A* **2016,** *113* (39), 10968-73.
246. Hale, M. D.; Eaton, R. A., The Ultrastructure of Soft Rot Fungi. I. Fine Hyphae in Wood Cell Walls. *Mycologia* **1985,** *77* (3), 447-463.
247. Martinez, D.; Berka, R. M.; Henrissat, B.; Saloheimo, M.; Arvas, M.; Baker, S. E.; Chapman, J.; Chertkov, O.; Coutinho, P. M.; Cullen, D.; Danchin, E. G. J.; Grigoriev, I. V.; Harris, P.; Jackson, M.; Kubicek, C. P.; Han, C. S.; Ho, I.; Larrondo, L. F.; de Leon, A. L.; Magnuson, J. K.; Merino, S.; Misra, M.; Nelson, B.; Putnam, N.; Robbertse, B.; Salamov, A. A.; Schmoll, M.; Terry, A.; Thayer, N.; Westerholm-Parvinen, A.; Schoch, C. L.; Yao, J.; Barbote, R.; Nelson, M. A.; Detter, C.; Bruce, D.; Kuske, C. R.; Xie, G.; Richardson, P.; Rokhsar, D. S.; Lucas, S. M.; Rubin, E. M.; Dunn-Coleman, N.; Ward, M.; Brettin, T. S., Genome sequencing and analysis of the biomass-degrading fungus Trichoderma reesei (syn. Hypocrea jecorina). *Nature Biotechnology* **2008,** *26* (5), 553-560.
248. Simmons, C. W.; Reddy, A. P.; D'haeseleer, P.; Khudyakov, J.; Billis, K.; Pati, A.; Simmons, B. A.; Singer, S. W.; Thelen, M. P.; VanderGheynst, J. S., Metatranscriptomic analysis of lignocellulolytic microbial communities involved in high-solids decomposition of rice straw. *Biotechnology for Biofuels* **2014,** *7* (1), 495.
249. Backhed, F.; Ley, R. E.; Sonnenburg, J. L.; Peterson, D. A.; Gordon, J. I., Host-bacterial mutualism in the human intestine. *Science* **2005,** *307* (5717), 1915-20.
250. Liang, D.; Leung, R. K.-K.; Guan, W.; Au, W. W., Involvement of gut microbiome in human health and disease: brief overview, knowledge gaps and research opportunities. *Gut Pathogens* **2018,** *10*, 3.
251. Konig, H.; Li, L.; Frohlich, J., The cellulolytic system of the termite gut. *Appl Microbiol Biotechnol* **2013,** *97* (18), 7943-62.
252. Bergmann, G. T., Microbial community composition along the digestive tract in forage- and grain-fed bison. *BMC Veterinary Research* **2017,** *13*, 253.
253. Mao, S.; Zhang, M.; Liu, J.; Zhu, W., Characterising the bacterial microbiota across the gastrointestinal tracts of dairy cattle: membership and potential function. *Scientific Reports* **2015,** *5*, 16116.
254. Bakken, J. S.; Borody, T.; Brandt, L. J.; Brill, J. V.; Demarco, D. C.; Franzos, M. A.; Kelly, C.; Khoruts, A.; Louie, T.; Martinelli, L. P.; Moore, T. A.; Russell, G.; Surawicz, C., Treating Clostridium difficile infection with fecal microbiota transplantation. *Clinical gastroenterology and hepatology : the official clinical practice journal of the American Gastroenterological Association* **2011,** *9* (12), 1044-9.
255. Vargason, A. M.; Anselmo, A. C., Clinical translation of microbe-based therapies: Current clinical landscape and preclinical outlook. *Bioengineering & Translational Medicine* **2018,** *3* (2), 124-137.

256. El Kaoutari, A.; Armougom, F.; Gordon, J. I.; Raoult, D.; Henrissat, B., The abundance and variety of carbohydrate-active enzymes in the human gut microbiota. *Nature reviews. Microbiology* **2013,** *11* (7), 497-504.
257. Ley, R. E.; Turnbaugh, P. J.; Klein, S.; Gordon, J. I., Microbial ecology: human gut microbes associated with obesity. *Nature* **2006,** *444* (7122), 1022-3.
258. McNeil, N. I., The contribution of the large intestine to energy supplies in man. *The American journal of clinical nutrition* **1984,** *39* (2), 338-42.
259. Mathieu, S.; Touvrey-Loiodice, M.; Poulet, L.; Drouillard, S.; Vincentelli, R.; Henrissat, B.; Skjåk-Bræk, G.; Helbert, W., Ancient acquisition of "alginate utilization loci" by human gut microbiota. *Scientific Reports* **2018,** *8* (1), 8075.
260. Terrapon, N.; Lombard, V.; Gilbert, H. J.; Henrissat, B., Automatic prediction of polysaccharide utilization loci in Bacteroidetes species. *Bioinformatics* **2015,** *31* (5), 647-655.
261. Vidal-Melgosa, S.; Pedersen, H. L.; Schuckel, J.; Arnal, G.; Dumon, C.; Amby, D. B.; Monrad, R. N.; Westereng, B.; Willats, W. G., A new versatile microarray-based method for high throughput screening of carbohydrate-active enzymes. *The Journal of biological chemistry* **2015,** *290* (14), 9020-36.
262. Dale, B., Time to Rethink Cellulosic Biofuels? *Biofuels, Bioproducts and Biorefining* **2018,** *12* (1), 5-7.
263. Schmer, M. R.; Vogel, K. P.; Mitchell, R. B.; Perrin, R. K., Net energy of cellulosic ethanol from switchgrass. *Proceedings of the National Academy of Sciences* **2008,** *105* (2), 464-469.
264. Shen, H.; Poovaiah, C. R.; Ziebell, A.; Tschaplinski, T. J.; Pattathil, S.; Gjersing, E.; Engle, N. L.; Katahira, R.; Pu, Y.; Sykes, R.; Chen, F.; Ragauskas, A. J.; Mielenz, J. R.; Hahn, M. G.; Davis, M.; Stewart, C. N.; Dixon, R. A., Enhanced characteristics of genetically modified switchgrass (Panicum virgatum L.) for high biofuel production. *Biotechnology for Biofuels* **2013,** *6* (1), 71.
265. Van Acker, R.; Vanholme, R.; Storme, V.; Mortimer, J. C.; Dupree, P.; Boerjan, W., Lignin biosynthesis perturbations affect secondary cell wall composition and saccharification yield in Arabidopsis thaliana. *Biotechnology for Biofuels* **2013,** *6* (1), 46.
266. Barsanti, L.; Gualtieri, P., Is exploitation of microalgae economically and energetically sustainable? *Algal Research* **2018,** *31*, 107-115.
267. Tredici, M. R., Photobiology of microalgae mass cultures: understanding the tools for the next green revolution. *Biofuels* **2010,** *1* (1), 143-162.
268. Zhu, X. G.; Long, S. P.; Ort, D. R., What is the maximum efficiency with which photosynthesis can convert solar energy into biomass? *Current opinion in biotechnology* **2008,** *19* (2), 153-159.
269. Pokoo-Aikins, G.; Nadim, A.; El-Halwagi, M. M.; Mahalec, V., Design and analysis of biodiesel production from algae grown through carbon sequestration. *Clean Technologies and Environmental Policy* **2010,** *12* (3), 239-254.
270. Maeda, Y.; Yoshino, T.; Matsunaga, T.; Matsumoto, M.; Tanaka, T., Marine microalgae for production of biofuels and chemicals. *Current opinion in biotechnology* **2018,** *50*, 111-120.
271. Seiboth, B.; Ivanova, C.; Seidl-Seiboth, V., Trichoderma reesei: a fungal enzyme producer for cellulosic biofuels. In *Biofuel production-recent developments and prospects*, InTech: 2011.
272. Ajjawi, I.; Verruto, J.; Aqui, M.; Soriaga, L. B.; Coppersmith, J.; Kwok, K.; Peach, L.; Orchard, E.; Kalb, R.; Xu, W.; Carlson, T. J.; Francis, K.; Konigsfeld, K.; Bartalis, J.; Schultz, A.; Lambert, W.; Schwartz, A. S.; Brown, R.; Moellering, E. R., Lipid production in Nannochloropsis gaditana is doubled by decreasing expression of a single transcriptional regulator. *Nature Biotechnology* **2017,** *35*, 647.

273. Adli, M., The CRISPR tool kit for genome editing and beyond. *Nature Communications* **2018,** *9* (1), 1911.
274. Green, F.; Stern, N., China's changing economy: implications for its carbon dioxide emissions. *Climate Policy* **2017,** *17* (4), 423-442.
275. Ahmed, T.; Shahid, M.; Azeem, F.; Rasul, I.; Shah, A. A.; Noman, M.; Hameed, A.; Manzoor, N.; Manzoor, I.; Muhammad, S., Biodegradation of plastics: current scenario and future prospects for environmental safety. *Environmental Science and Pollution Research* **2018,** *25* (8), 7287-7298.
276. Yang, Y.; Yang, J.; Wu, W. M.; Zhao, J.; Song, Y.; Gao, L.; Yang, R.; Jiang, L., Biodegradation and Mineralization of Polystyrene by Plastic-Eating Mealworms: Part 2. Role of Gut Microorganisms. *Environmental science & technology* **2015,** *49* (20), 12087-93.
277. Yang, J.; Yang, Y.; Wu, W. M.; Zhao, J.; Jiang, L., Evidence of polyethylene biodegradation by bacterial strains from the guts of plastic-eating waxworms. *Environmental science & technology* **2014,** *48* (23), 13776-84.
278. Austin, H. P.; Allen, M. D.; Donohoe, B. S.; Rorrer, N. A.; Kearns, F. L.; Silveira, R. L.; Pollard, B. C.; Dominick, G.; Duman, R.; El Omari, K.; Mykhaylyk, V.; Wagner, A.; Michener, W. E.; Amore, A.; Skaf, M. S.; Crowley, M. F.; Thorne, A. W.; Johnson, C. W.; Woodcock, H. L.; McGeehan, J. E.; Beckham, G. T., Characterization and engineering of a plastic-degrading aromatic polyesterase. *Proceedings of the National Academy of Sciences* **2018**.
279. Prigge, S. T.; Kolhekar, A. S.; Eipper, B. A.; Mains, R. E.; Amzel, L. M., Amidation of bioactive peptides: The structure of peptidylglycine alpha-hydroxylating monooxygenase. *Science* **1997,** *278* (5341), 1300-1305.
280. Prigge, S. T.; Eipper, B. A.; Mains, R. E.; Amzel, L. M., Dioxygen binds end-on to mononuclear copper in a precatalytic enzyme complex. *Science* **2004,** *304* (5672), 864-867.
281. Sean T. Prigge, A. S. K., Betty A.Eipper, Richard E.Mains, L.MArio Amzel, Substrate-mediated electron transfer in peptidyglycine alpha hydroxylating monooxygenase. *Nature Structural Biology* **1999,** *6* (10).
282. Bauman, A. T.; Broers, B. A.; Kline, C. D.; Blackburn, N. J., A Copper-Methionine Interaction Controls the pH-Dependent Activation of Peptidylglycine Monooxygenase. *Biochemistry* **2011,** *50* (50), 10819-10828.
283. Siebert, X.; Eipper, B. A.; Mains, R. E.; Prigge, S. T.; Blackburn, N. J.; Amzel, L. M., The catalytic copper of peptidylglycine alpha-hydroxylating monooxygenase also plays a critical structural role. *Biophysical Journal* **2005,** *89* (5), 3312-3319.
284. Kim, S.; Lee, J. Y.; Cowley, R. E.; Ginsbach, J. W.; Siegler, M. A.; Solomon, E. I.; Karlin, K. D., A N3S(thioether)-Ligated Cu(II)-Superoxo with Enhanced Reactivity. *Journal of the American Chemical Society* **2015,** *137* (8), 2796-9.
285. Lee, C.-D.; Sun, H.-C.; Hu, S.-M.; Chiu, C.-F.; Homhuan, A.; Liang, S.-M.; Leng, C.-H.; Wang, T.-F., An improved SUMO fusion protein system for effective production of native proteins. *Protein Science* **2008,** *17* (7), 1241-1248.
286. Lee, J. Y.; Karlin, K. D., Elaboration of copper–oxygen mediated CH activation chemistry in consideration of future fuel and feedstock generation. *Current Opinion in Chemical Biology* **2015,** *25* (0), 184-193.
287. Chauhan, S.; Kline, C. D.; Mayfield, M.; Blackburn, N. J., Binding of Copper and Silver to Single-Site Variants of Peptidylglycine Monooxygenase Reveals the Structure and Chemistry of the Individual Metal Centers. *Biochemistry* **2014,** *53* (6), 1069-1080.
288. Kim, K.-H.; Seong, B. L., Peptide amidation: Production of peptide hormonesin vivo andin vitro. *Biotechnology and Bioprocess Engineering* **2001,** *6* (4), 244-251.
289. Lindwall, G.; Chau, M. F.; Gardner, S. R.; Kohlstaedt, L. A., A sparse matrix approach to the solubilization of overexpressed proteins. *Protein Engineering* **2000,** *13* (1), 67-71.

290. Porath, J.; Carlsson, J.; Olsson, I.; Belfrage, G., Metal Chelate Affinity Chromatography, A New Approach To Protein Fractionation *Nature* **1975,** *258* (5536), 598-599.
291. Hochuli, E.; Döbeli, H.; Schacher, A., New metal chelate adsorbent selective for proteins and peptides containing neighbouring histidine residues. *Journal of Chromatography A* **1987,** *411*, 177-184.
292. Hochuli, E.; Bannwarth, W.; Döbeli, H.; Gentz, R.; Stüber, D., Genetic Approach to Facilitate Purification of Recombinant Proteins with a Novel Metal Chelate Adsorbent. *Bio/Technology* **1988,** *6*, 1321.
293. Porath, J., Immobilized Metal-Ion Affinity-Chromatography *Protein Expression and Purification* **1992,** *3* (4), 263-281.
294. Fowler, C. A., Purification and analysis of novel lytic polysaccharide monooxygenases containing an n-terminal arginine residue. In *Advanced Research Project, MChem, University of York*, 2014.
295. Skerra, A.; Schmidt, T. G. M., Use of the Strep- tag and streptavidin for detection and purification of recombinant proteins. In *Methods in Enzymology*, Academic Press: 2000; Vol. 326, pp 271-304.
296. Schmidt, T. G. M.; Skerra, A., The Strep-tag system for one-step purification and high-affinity detection or capturing of proteins. *Nature Protocols* **2007,** *2*, 1528.
297. Porath, J.; Flodin, P., Gel Filtration - Method For Desalting And Group Separation. *Nature* **1959,** *183* (4676), 1657-1659.
298. E., K. D., Enzyme flexibility and enzyme action. *Journal of Cellular and Comparative Physiology* **1959,** *54* (S1), 245-258.
299. Pantoliano, M. W.; Petrella, E. C.; Kwasnoski, J. D.; Lobanov, V. S.; Myslik, J.; Graf, E.; Carver, T.; Asel, E.; Springer, B. A.; Lane, P.; Salemme, F. R., High-density miniaturized thermal shift assays as a general strategy for drug discovery. *Journal of Biomolecular Screening* **2001,** *6* (6), 429-440.
300. Lo, M. C.; Aulabaugh, A.; Jin, G.; Cowling, R.; Bard, J.; Malamas, M.; Ellestad, G., Evaluation of fluorescence-based thermal shift assays for hit identification in drug discovery. *Anal Biochem* **2004,** *332* (1), 153-9.
301. Reinhard, L.; Mayerhofer, H.; Geerlof, A.; Mueller-Dieckmann, J.; Weiss, M. S., Optimization of protein buffer cocktails using Thermofluor. *Acta crystallographica. Section F, Structural biology and crystallization communications* **2013,** *69* (Pt 2), 209-14.
302. Zhang, Z. Q.; Xiao, Z. P.; Linhardt, R. J., Thin Layer Chromatography for the Separation and Analysis of Acidic Carbohydrates. *Journal of Liquid Chromatography & Related Technologies* **2009,** *32* (11-12), 1711-1732.
303. Karas, M.; Bachmann, D.; Bahr, U.; Hillenkamp, F., Matrix-assisted ultraviolet laser desorption of non-volatile compounds. *International Journal of Mass Spectrometry and Ion Processes* **1987,** *78*, 53-68.
304. Harvey, D. J., Matrix-assisted laser desorption/ionization mass spectrometry of carbohydrates and glycoconjugates. *International Journal of Mass Spectrometry* **2003,** *226* (1), 1-35.
305. Rohrer, J. S.; Basumallick, L.; Hurum, D., High-performance anion-exchange chromatography with pulsed amperometric detection for carbohydrate analysis of glycoproteins. *Biochemistry. Biokhimiia* **2013,** *78* (7), 697-709.
306. Cataldi, T. R.; Campa, C.; De Benedetto, G. E., Carbohydrate analysis by high-performance anion-exchange chromatography with pulsed amperometric detection: the potential is still growing. *Fresenius' journal of analytical chemistry* **2000,** *368* (8), 739-58.
307. Pell, G.; Szabo, L.; Charnock, S. J.; Xie, H. F.; Gloster, T. M.; Davies, G. J.; Gilbert, H. J., Structural and biochemical analysis of Cellvibrio japonicus xylanase 10C - How variation

in substrate-binding cleft influences the catalytic profile of family GH-10 xylanases. *Journal of Biological Chemistry* **2004,** *279* (12), 11777-11788.
308. Matsui, I.; Ishikawa, K.; Miyairi, S.; Fukui, S.; Honda, K., An increase in the transglycosylation activity of saccharomycopsis alpha-amylase altered by site-directed mutagenesis. *Biochimica Et Biophysica Acta* **1991,** *1077* (3), 416-419.
309. Hrmova, M.; Garrett, T. P. J.; Fincher, G. B., Subsite affinities and disposition of catalytic amino-acids in the substrate-binding region of barley 1,3-beta-glucanases - implications in plant-pathogen interactions. *Journal of Biological Chemistry* **1995,** *270* (24), 14556-14563.
310. Miller, G. L., Use of dinitrosalicylic acid reagent for determination of reducing sugar. *Analytical Chemistry* **1959,** *31* (3), 426-428.
311. Roberts, S. M.; Davies, G. J., The crystallization and structural analysis of cellulases (and other glycoside hydrolases): strategies and tactics. *Methods Enzymol* **2012,** *510*, 141-68.
312. Peisach, J.; Blumberg, W. E.; Lode, E. T.; Coon, M. J., An Analysis of the Electron Paramagnetic Resonance Spectrum of Pseudomonas oleovorans Rubredoxin : A method for determination of the ligands of ferric iron in completely rhombic sites. *Journal of Biological Chemistry* **1971,** *246* (19), 5877-5881.
313. Fogg, M. J.; Wilkinson, A. J., Higher-throughput approaches to crystallization and crystal structure determination. *Biochemical Society Transactions* **2008,** *36*, 771-775.