Analysis of Protein Crystallisation Parameters

Jobie Samuel Kirkwood

Doctor of Philosophy

University of York

Chemistry

June 2015

Abstract

Proteins are biochemical molecules that are essential for life processes. Their function is linked to their structure and so it follows an understanding of their structure will assist in an understanding of their function. The predominant method of solving protein structures is X-ray crystallography and for this a protein crystal is required. The process of obtaining a crystal is amongst the phases of the structure determination process with the highest rates of attrition. Analyses are performed throughout this thesis, which are intended to help improve output for this bottleneck. It has been possible to develop a method to determine pH using a spectrophotometer and acid-base indicator in an accurate, rapid and efficient manner. A method for predicting the pH of buffered solutions has also been developed and these predicted pH values are linked to the isoelectric point of a protein sequence. The isoelectric point is in turn used in classification, along with many other features, to determine a protein's propensity to crystallise. Finally, the most prevalent and successful chemical species in crystallisation are explored, compared and linked. These chemicals are used to design a new crystallisation screen.

List of Contents

List of Figures

List of Tables

Acknowledgements

I would like to thank: the UK Biotechnology and Biological Sciences Research Council (BBSRC grant BB/ I015868/1) and AstraZeneca for funding this research; my supervisors, David Hargreaves, Julie Wilson, Richard Pauptit and Simon O'Keefe for their tolerance and wisdom; those who I could not have produced the work in this thesis without, Alex Cooper, the Chemistry Graduate Office, David 'Derv' Mitchell, Dominic Friend, Frank von Delft, Heather Deighton, James Cussens, Janet Newman, Jia Tsing Ng, Marek Brzozowski, Martin Rusilowicz, Richard Pauptit, Sarah Christmas, Shirley Roberts, Simon Grist and Tina Howard; my family and friends Alan Millard, Dad and Julie, Elmira Esmaeili, Ian Halliday, Jonpaul Musgrove, Julie Bloor, Kane Hudson, Katie Tasker, Liam Hollinshead, Lucy Milner, Michael Zahn, Mike Robinson, Mum and Tom, Murat Dogan, Nana and Gramps, Peter Nemeth, Poppy Marvin, Richard Gammons, Shulan Zhao, Thomas Kirkwood and Tsui Lan Chen for keeping me (just about) sane and for providing me with laughter, love and food.

Author's Declaration

All of the work contained in this thesis is original and has not been submitted for any other degree at this or any other institution, with the exception of the following chapters:

Chapter 4 | Kirkwood, JS, Hargreaves, D., O'Keefe, S. & Wilson, J. (2015). Bioinformatics, btv011.

Chapter 5 | Kirkwood, JS., Wilson, J., O'Keefe, S. & Hargreaves, D. (2014). Acta Crystallographica Section D: Biological Crystallography 70, 2367-2375.

Chapters 6 and 7 | Kirkwood, JS., Hargreaves, D., O'Keefe, S. & Wilson, J. (2015). Acta Crystallographica Section F: Structural Biology Communications. *In press.*

Kirkwood, JS. was the main contributor to all papers.

1. Introduction

Proteins are biochemical molecules that underpin the processes essential for life (Jurisica *et al.*, 2001). They are found in all life forms from viruses to animals, from bacteria to plants. They have numerous roles such as hormones, enzymes, transporters, receptors and regulators. The function of proteins is related to their structure (Wright & Dyson, 1999). For example, the globular protein haemoglobin surrounds oxygen atoms while transporting them through blood (Eaton *et al.*, 1999) and the fibrous protein collagen provides the backbone for the connective tissues in the heart, tendons and cornea (Bolboaca & Jantschi, 2007). For the majority of proteins their structure is determined by their sequence (Anfinsen *et al.*, 1961). Their sequence is in turn determined by RNA, which is transcribed from genes in DNA.

It is also possible that, although the gene and protein sequence is transcribed correctly, the protein folds in a manner such that a different structure can be obtained from the same sequence. Such proteins are called prions. They can fold into a numerous conformations, some of which can be harmful inducing conformational change amongst other proteins and causing the diseases Creutzfeldt-Jakob's Disease (CJD), scrapie and Bovine Spongiform Encephalopathy (BSE) (Pietzsch, 2002, NHS, 2013). Proteins can also undergo post-translational modification (PTM), meaning their structure is modified in one of many ways, after the protein has been formed. An example of a PTM is the addition of ubiquitin to proteins. This addition is recognised by proteasomes, which in turn begin to degrade and recycle the protein. PTMs of the protein tau have been linked to Alzheimer's disease (Gong *et al.*, 2005).

Structural Genomics seeks to determine the three-dimensional structure of proteins in order to understand their function and assist in developing drugs (Navia & Murcko, 1992). It has been possible to determine the structures of many proteins related to drug development (Tickle *et al.*, 1984) including that of haemoglobin (Perutz *et al.*, 1960). The Nobel Prize was awarded to the scientists who solved the structures of haemoglobin and in total, 24 Nobel Prizes have been awarded for efforts focused on determining the structure of proteins (Jaskolski *et al.*, 2014).

The Protein Data Bank (PDB) is an open access online repository where information pertaining to the three-dimensional structure of macromolecules is stored (Berman *et al.*, 2000). Along with the coordinates describing the atomic structure, each entry in the PDB includes variables such as the organism from which the protein was obtained, the protein sequence and the experimental method used to determine the structure. In April 2015 there were 100,032 protein structures in the PDB (PDB, 2015) and the predominant method used to detemine their structure was X-ray crystallography (X-ray), the recorded method for 89,977 (89.7%) entries, as shown in Figure 1. Nuclear Magnetic Resonance (NMR) was to determine 9,559 (9.5%) structures and other methods were used for just 785 (0.8%) structures.

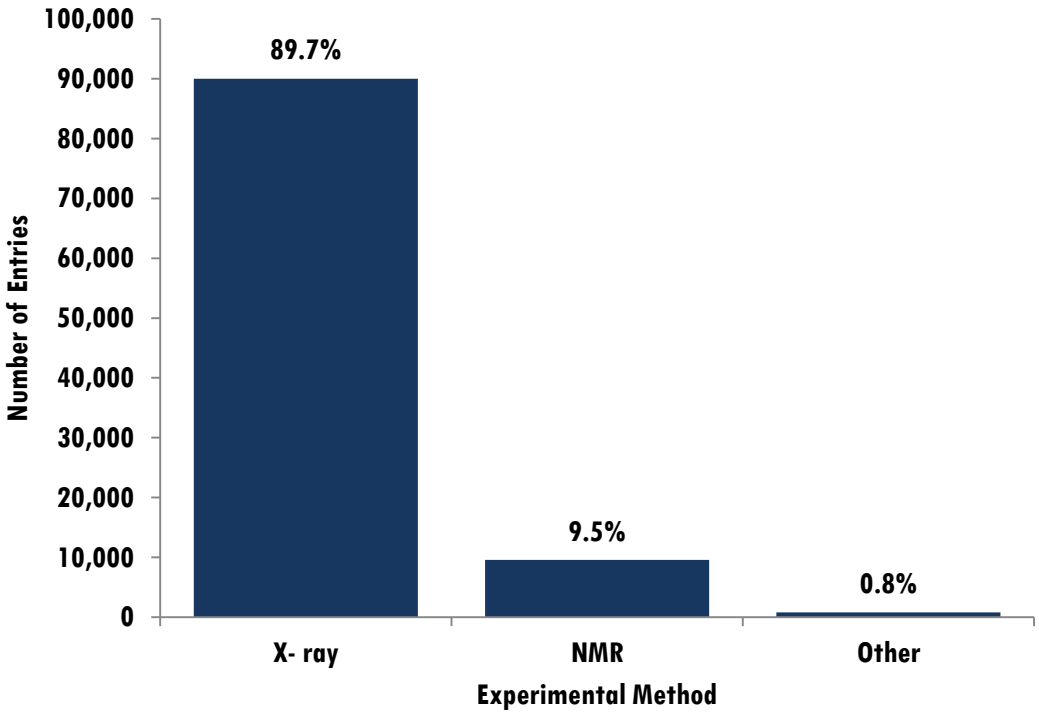


Figure 1: Number of entries in the PDB by experimental method.

In April 2015 the predominant method of protein structure determination was by X-ray, accounting for 89.7% of proteins in the PDB, with NMR accounting for 9.5% and other methods for 0.8%.

Other methods for structure determination include electron microscopy (EM), the most prevalent of the other methods (Elands & Hax, 2004, Morikawa *et al.*, 2015), electron crystallography (Yonekura *et al.*, 2015) and electron paramagnetic resonance (Fleissner *et al.*, 2009); fibre diffraction (Tewary *et al.*, 2011) and neutron

diffraction (Coates *et al.*, 2014); solution scattering (Wu *et al.*, 2009); powder diffraction (Von Dreele *et al.*, 2000) and hybrid methods (Howard *et al.*, 2011). The number of structures solved using these methods has accounted for 0.5 to 2% of the structures deposited each year from 2004.

The number of NMR solved structures has fallen in recent years. Of all the structures solved and deposited in the PDB in 2004, NMR was used for 14%. In 2014 it accounted for just 4%. Conversely, the number of X-ray solved structures is still growing. In 2004, it was used to determine 86% of structures and in 2014 this had increased to 94%. X-ray, however, is a much older method than the others, with the first X-ray diffraction of a crystal being achieved in 1913 (Bragg & Bragg, 1913), allowing for a century of refinement. The first NMR and EM entries were in 1986 and 1997 respectively, compared to the first X-ray entry in 1971. This means NMR and EM are not currently advanced enough to be able to determine the structure of all proteins. NMR has limitations on the size of protein that can be used (Smialowski *et al.*, 2006) and EM currently lacks powerful resolution (Milne *et al.*, 2013). The limitation on size and resolution, coupled with the demand for protein complexes (Aloy & Russell, 2006) currently ensures the continued use of X-ray crystallography. As many of the recently deposited structures have a sequence similar to those previously determined by X-ray crystallography, there has been no reason to change the method. However, a major advantage of non-X-ray techniques is that they do not require a protein crystal and can be used in structure determination for proteins that cannot be crystallised (Elands & Hax, 2004). NMR can also provide dynamic information and is less destructive to a sample (McDermott, 2004). These techniques can be complementary to X-ray crystallography, with NMR used to determine the structures of small protein-binding structures and X-ray crystallography used to determine the larger protein structures (Jahnke & Widmer, 2004).

In order to collect X-ray diffraction data it is essential to obtain suitable crystals via crystallisation. Before crystallisation, a protein has to be obtained. When a protein target is identified for which the structure is to be determined, the gene that encodes for the protein is cloned and then inserted into a vector within a host cell. The protein is then over-expressed to provide many copies, which are then extracted from the cell by several steps of purification that might include sonication, centrifugation (Lesley,

2001) and fast protein liquid chromatography (FPLC). The significance of the purity of a sample was raised by Kam *et al.* (1978) who suggested that impurity eventually makes further growth energetically unfavourable and therefore could influence the terminal size of a crystal. It has also been reported that crystal quality is positively correlated to protein purity (Ducruix & Giegé, 1992). In addition to the problem of whether impurities are detrimental to crystallisation, protein purity is important for the reproducibility of experiments (Lorber *et al.*, 1993). If further buffers or additives are required to stabilise the protein these are added to create a protein solution ready for crystallisation.

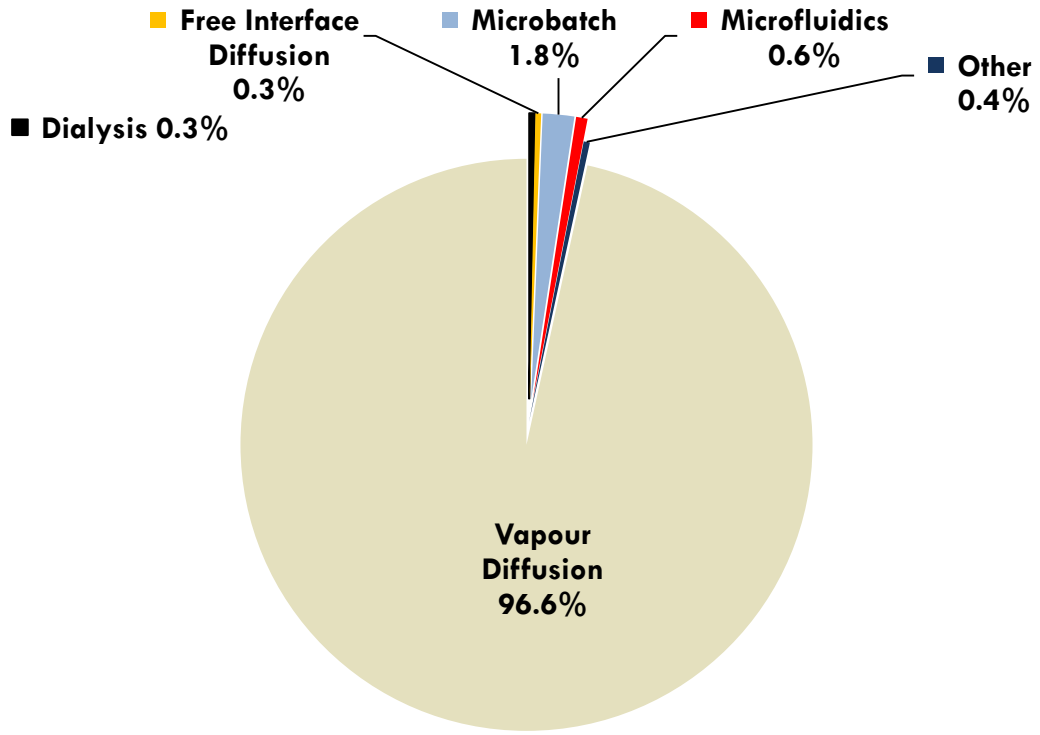


Figure 2: Methods of crystallisation recorded in the PDB.

The predominant method of crystallisation in the PDB is vapour diffusion accounting for 96.6% of entries, from a sample of 68,202 obtained in March 2015.

Protein crystallisation was first referred to by Hünefeld in 1840, who crystallised earthworm haemoglobin. Hünefeld suggested that it was possible to obtain protein crystals through a method of controlled evaporation. In 1851, Fünke devised a reproducible method that involved the use of alcohol (the first use of organic solvent in crystallisation). Throughout the latter part of the 19th century and the early 20th century crystallisation experiments were undertaken that incorporated many of the

chemical species used in crystallisation today. It was in the early 1930s when X-ray crystallographers began to look at protein crystals as a method of obtaining structural information about proteins (McPherson, 1991).

There are several methods of crystallisation recorded in the PDB. Grouping these methods together into the broad categories described by Chayen and Saridakis (2008) it can be seen that the predominant method of crystallisation is vapour diffusion (Figure 2), accounting for 96.6% (65,870/68,202) (Bolanos-Garcia & Chayen, 2009).

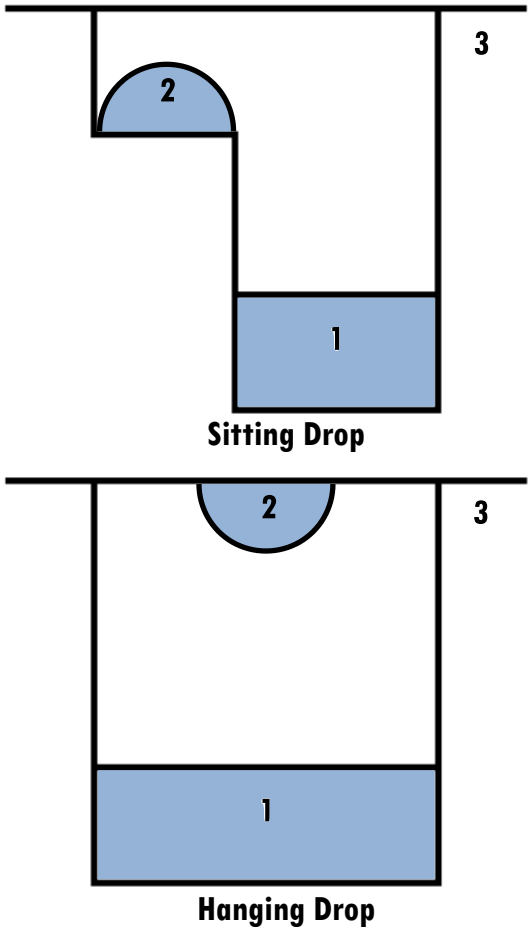
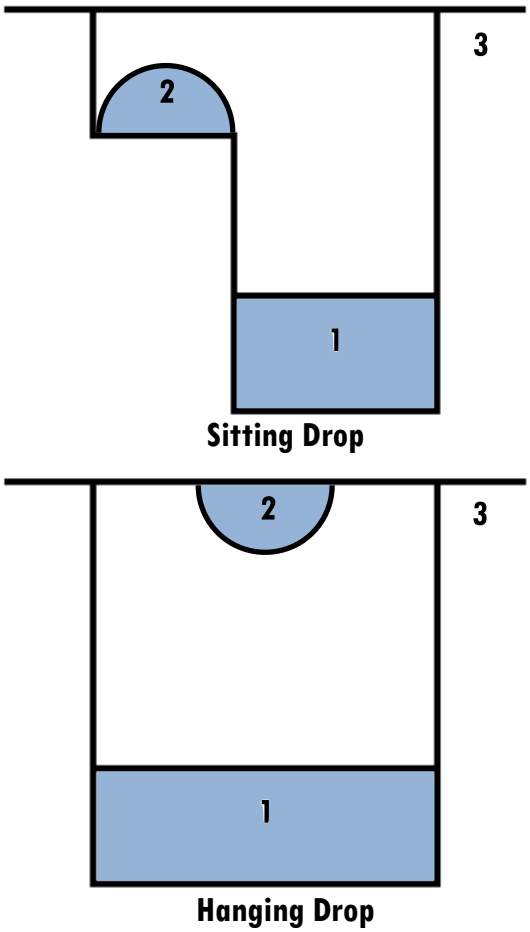


Figure 3: A schematic of vapour diffusion.

Two methods of vapour diffusion, sitting and hanging drop. The different components of the setup are indicated by the following numbers:

1. The mother liquor, a mixture of crystallisation chemicals.
2. A mixture of the mother liquor and the protein for crystallisation.
3. The system is sealed to allow vapour diffusion.

There are two predominant methods of vapour diffusion, either sitting drop or



hanging drop, both shown in Figure 3. In a vapour diffusion experiment, chemicals used to promote crystallisation are mixed and this chemical cocktail, referred to as the crystallisation solution, is placed in a crystallisation well (a plastic container). Throughout this thesis the terms well, crystallisation solution and condition are used interchangeably. It should be noted that there are no standard names for various components of crystallisation experiments and as such they vary throughout the literature. A sample of this crystallisation solution is taken and combined, typically in a 1:1 ratio with the protein solution (Chirgadze, 2001). This mixture of protein and chemical cocktail is then offset from the reservoir containing the majority of the crystallisation solution either by hanging from a cover slip above (hanging drop) or by sitting in a smaller well (sitting drop). The system is then sealed using a water impervious barrier such as transparent pressure-sensitive tape. The solutions then equilibrate and the concentration of the components in the droplets changes over time. For an indication of scale, custom crystallisation experiments described throughout this thesis are

sitting drop where crystallisation solution is 80µl and from this 100nl is taken and mixed with 100nl of protein solution.

Some of the lesser used methods to induce crystallisation include: microbatch, in which both protein and crystallisation solution are dispensed together under a layer of low density paraffin, silicone or a mixture of both; dialysis, in which a semipermeable membrane separates the protein and crystallisation solutions; free interface diffusion, in which protein and crystallisation solutions sit side by side; and microfluidics, which use much fewer nanolitres of protein and chemical solutions than standard experiments (Chayen & Saridakis, 2008).

The physics of crystallisation dictate that for a crystal to form the protein solution must reach supersaturation via the diffusion process. In vapour diffusion this occurs by the evaporation of water from the sitting/hanging drop to the crystallisation solution in the reservoir, increasing the concentration within the drop. Too much supersaturation will result in precipitation, too little and nucleation will not occur. Once nucleation has occurred the protein solution needs to move into a metastable state where the growth of crystals can occur. This transition from nucleation to crystal growth is where the combined crystallisation parameters such as method, pH, precipitants and temperature have their effect (Asherie, 2004, DeLucas *et al.*, 2003, Weber, 1997).

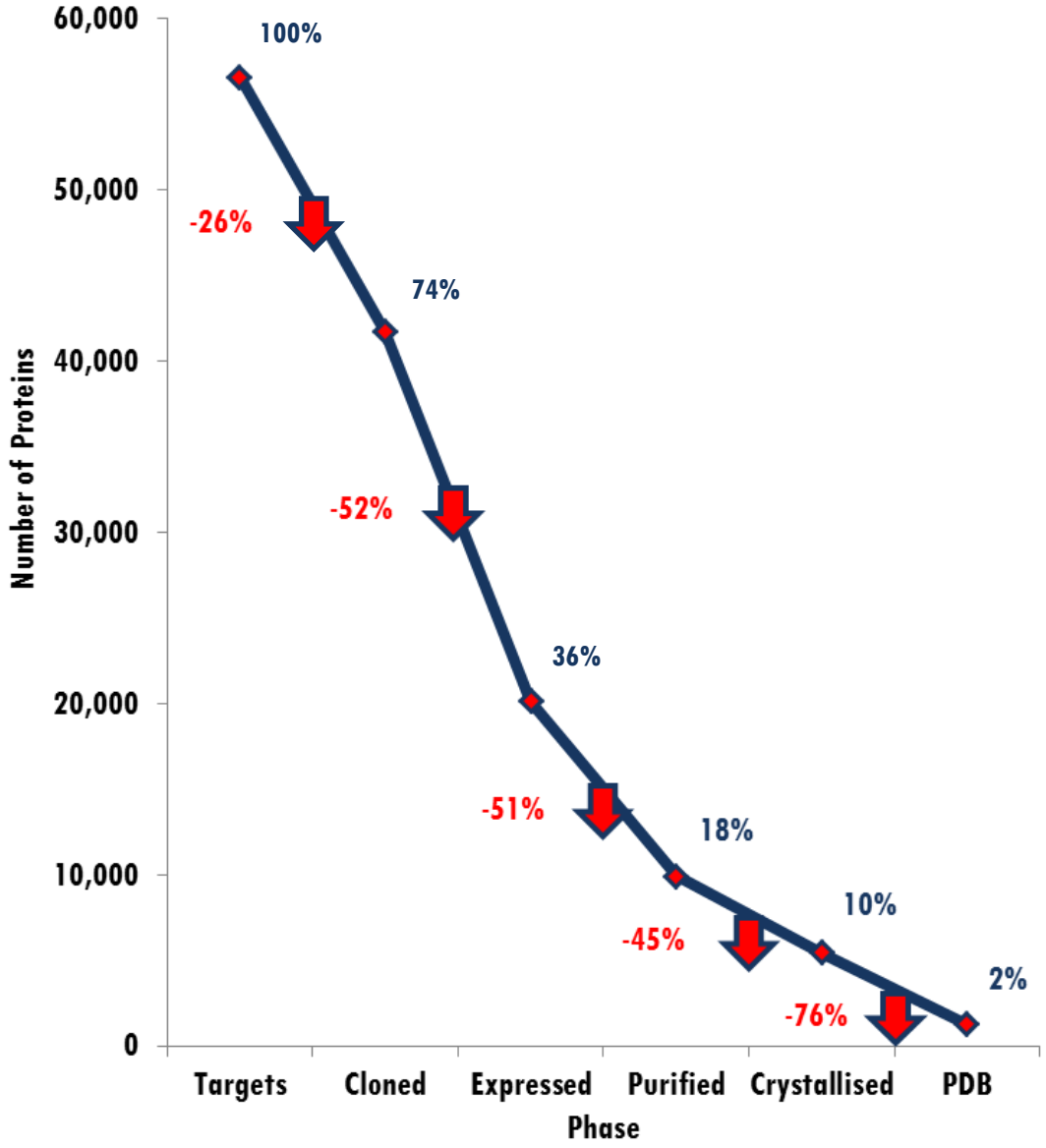


Figure 4: Attrition rates of the structure determination pipeline.

The percentages on the right of the phase marker show the proportion of proteins that have passed each phased of the process. Those on the left of the arrow show the percentage drop in proteins from the previous phase. The data was obtained from TargetDB at http://sbkb.org/metrics/ in March 2015.

Despite some understanding of the processes that control protein crystal growth, obtaining crystals from which a structure can be determined has the highest rate of attrition for any phase of the structure determination pipeline (Chayen, 2003). Just 24% of those proteins that could crystallise become structures deposited in the PDB, as shown in Figure 4. The stage before this, from purification to crystallisation, is also challenging with roughly 1 in 2 (55%) proteins forming crystals. This step,

crystallisation, has been the stage with the highest rate of attrition over the past decade (Chayen, 2004, TargetDB, 2010) and is, therefore, described as a bottleneck in the protein structure determination process (Chayen and Saridakis (2008), D'Arcy (1994), Stevens (2000)).

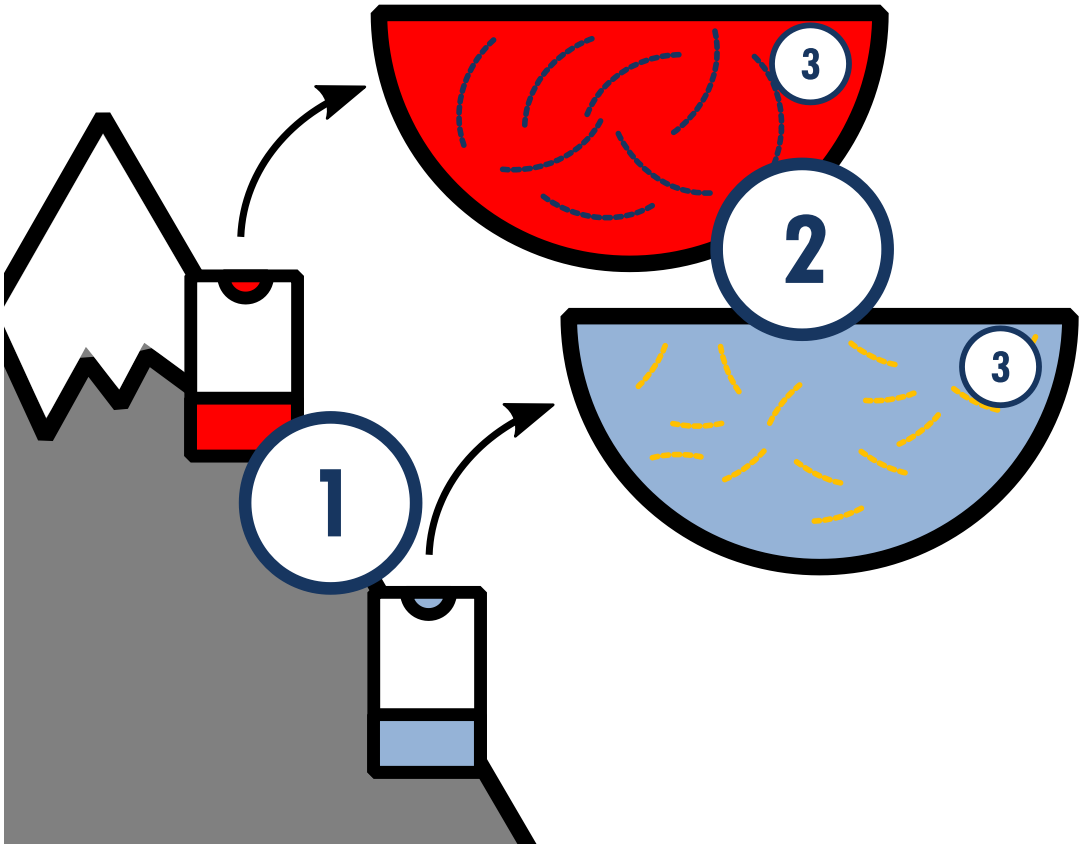


Figure 5: Parameters affecting crystallisation.

Two different crystallisation experiments described in terms of three major parameters: (1) physical parameters - experiments performed at different altitude, both gravity, temperature and pressure effected; (2) chemical parameters - one solution (the upper) is red indicative of an acidic pH, the other solution (the lower) is blue indicative of a basic pH; (3) biochemical parameters - two different proteins, one longer than the other and at lower concentration.

The parameters affecting a protein crystallisation experiment (shown in Figure 5) can be categorised into one of the following groups: (1) physical parameters, such as the temperature and method of experiment; (2) chemical parameters, such as the pH and the precipitant used; and (3) biochemical parameters, such as the isoelectric point and chemical modifications to the protein sample (McPherson, 1999).

Over time, each of the groups has been explored for ways of reducing the number of protein crystallisation experiments that do not result in a crystal. Physical parameters have been studied in collaboration with the North American Space Agency (NASA) and experiments have been performed in microgravity (Gilliland *et al.*, 1996, CASIS, 2013). Some researchers have explored the use of desiccation (Yin *et al.*, 2010, Xie *et al.*, 2012) and others have used computer simulations (Yoshizaki *et al.*, 2004). Samudzi *et al.* (1992) found that most crystallisation experiments were either attempted at 3ºc or 21ºc with very little exploration of the temperatures in-between. They suggested that temperature needed to be studied further, as 86% of proteins display temperature dependence (Christopher *et al.*, 1998). Nucleation is also dependent on temperature. A study of four temperatures with the protein SmFru-1,6-P2ase showed that nucleation only occurs in a narrow range of concentrations at 15 ºC, but this range increases at 30 ºC (Zhu *et al.*, 2006). Physical parameters are arguably the most difficult and expensive to change and may require long term planning into infrastructure. The majority of experiments are presumably undertaken with constant gravity, pressure and similar amount of vibration. In data pertaining to crystallisation trials, unless physical parameters are the focus of the study, they are not recorded, therefore, physical parameters are not explored further here.

Usually, there are multiples of 96 combinations of chemicals trialled at any given time, each combination in its own well. The collective term for these 96 wells is a screen. Although any number of conditions tested at the same time can be called a screen, 96 conditions is the number in commercially available screens. An early logical approach to screening employed the use of incomplete factorial design (Fisher, 1942). A set of conditions was selected that sampled the chemical parameter space in a 'statistically effective manner' (Carter & Carter, 1979). Following on from this other attempts have been made to effectively sample chemical parameter space by: searching a small region of chemical parameter space in detail (McPherson, 1989b); sampling regions known to be favourable for crystal growth (Jancarik & Kim, 1991); systematically searching distinct regions of parameter space (Stura *et al.*, 1992); focusing the selection of chemicals for particular proteins (Brzozowski & Walton, 2001); using minimal spanning set theory to obtain the theoretically most efficient screen (Kimber *et al.*, 2003, Page *et al.*, 2003); and including the use of

ligands and additives into a screen (Gorrec, 2009). All screens vary the chemical species, concentration and pH with the most successful crystallisation species including polyethylene glycol (PEG) of various molecular weights, salts and buffers (Fazio et al., 2014). Typically each screen is trialled with one protein, which is encompassed by the biochemical parameters. Chemical and biochemical parameters are explored in this thesis.

There are those who argue that the protein is the most important variable (Longenecker *et al.*, 2001) and that it is often overlooked (Dale *et al.*, 2003). Researchers have sought to determine the structure of a protein solely from its sequence using only computational methods (Chou & Fasman, 1977, Baker & Sali, 2001, Garnier *et al.*, 1996); used protein properties, such as its hydropathy value (Kyte & Doolittle, 1982), to determine its propensity to crystallise (Smialowski *et al.*, 2006, Overton & Barton, 2006, Jahandideh & Mahdavi, 2012) and those who have used such properties to determine under which conditions a protein will crystallise (Samudzi *et al.*, 1992, Hennessy *et al.*, 2000, Kantardjieff & Rupp, 2004). This thesis also considers the relationship between a protein's physical properties and the likelihood of its crystallisation.

1.1. Thesis Summary

The difficulties surrounding crystallisation are due to the complex nature of the interactions between proteins and parameters such as pH, precipitants and temperature. Varying the conditions with numerous chemicals in combination makes the crystallisation parameter space exceptionally large and impossible to sample fully. Similarly, the number of properties which can be calculated for a protein sequence is also large, as there are many thousands of combinations of di- and tri-peptide pairs (Charoenkwan *et al.*, 2013). Fortunately, high-throughput structure determination generates lots of data that can be mined. Throughout this thesis we make use of data from several repositories to analyse protein crystallisation parameters. In those instances where data was not available, we created our own through experimentation. The datasets are described in detail in Chapter 2 and the methods used are outlined in Chapter 3. Using this data it has been possible to implement a new method of determining pH rapidly and accurately as described in

Chapter 4. In Chapter 5 a method to predict pH for buffered solutions is used to investigate a fiercely contested link between a protein's isoelectric point and the pH at which it crystallises. The use of predictors for determining a protein’s propensity to crystallise is challenged in Chapter 6. The most widely used chemicals and their combinations, which crystallise many proteins, are explored in Chapter 7 and used to design a new screen as described in Chapter 8. Finally, similarities in crystallisation parameter space are explored in Chapter 9 with the aim of reducing the search space.

2. Definitions and Data

The following chapter defines acronyms and chemicals terms used throughout this thesis along with a detailed description of the four datasets used for analyses. The datasets were obtained from AstraZeneca, the Structural Genomics Consortium, the Protein Data Bank and a dataset produced following customised experiments.

2.1. Abbreviations and Definitions

AHA	Alpha hydroxy acid.
Anion	A negatively charged ion.
AZ	AstraZeneca (Alderley Park, Cheshire).
Bis Tris	Bis-(2-hydroxyethyl)imino-tris(hydroxymethyl)methane.
BMCD	Biological Macromolecule Crystallisation Database.
CAPS	N-Cyclohexyl-3-aminopro- panesulfonic acid.
Cation	A positively charged ion.
Centrifugation	The process of separating particles by weight or settling solids from a solution by using a centrifuge.
Chromatography	See HPLC.
Construct ID (SGC)	Identifies the specific sequence of amino acids that form a (section of a) protein. The construct ID is the same whether or not the sequence contains a purification tag, usually comprising of six histidines genetically engineered to the end of a sequence to assist in purification.
Counterion	The ion that maintains electric neutrality.
Divalent	A molecule with a valence of two.
DTT	Dithiothreitol.
Dynamic Light Scattering	A technique that measures fluctuations in light intensity from a sample of proteins, which are assumed to be spherical.
GRAVY	Grand Average of Hydropathy.
HCL	Hydrogen chloride.

HEPES	2-(4-(2-Hydroxyethyl)-1- piperazinyl)ethanesulfonic Acid.
HPLC	High-Performance Liquid Chromatography is a technique in which samples are forced at high pressure through a matrix that binds to specific particles in the solution.
LDA	Linear Discriminant Analysis.
MDS	Multidimensional Scaling.
MES	2-(N-morpholino)ethanesulfonic acid.
MPD	2-methyl-2,4-pentanediol.
PCA	Principal Components Analysis.
PCTP	A broad range buffer system comprising propionic acid, cacodylate, bis-trispropane system.
PDB	Protein Data Bank.
PEG	Polyethylene glycol.
pI	Isoelectric point.
Plate barcode (SGC)	The identifier for a specific crystallisation plate in which a purified protein is screened. The plate barcode can be used to trace: the screen type (random/filter/grid/custom); the screen name (the particular sparse/filter screen used); the concentration of the protein; the temperature of the plate; any added compounds; whether the protein sample was fresh or frozen; the name of the crystallographer and the date of the experiment.
Project (AZ)	The target protein for which the structure is to be determined. All of the protein sequences within a project have a fixed percentage sequence similarity. A sequence may undergo more than one purification protocol. However, in most instances it is assumed that each new project relates to a new protein sequence.
Purification ID (SGC)	Each construct may be purified by more than one method and the purification ID identifies the particular method. A construct ID may be associated with multiple purification IDs.

SGC	Structural Genomics Consortium (Oxford).
TCEP	Tris(2-carboxyethyl)phosphine.
Tris	2-Amino-2-(hydroxymethyl)propane- 1,3-diol.

2.2. The AstraZeneca Dataset

In March 2012, AstraZeneca (AZ) provided a dataset associated with the crystallisation of macromolecules at their site at Alderley Park, Cheshire. The dataset contained information regarding 655,806 experiments (with each experiment relating to a well in a screen) from 26 screens and 163 projects. For each experiment, the dataset has several recorded fields, described in Table 1.

Field	Description
Project ID	An identifier of the protein
Trial and Session IDs	Identifier of relative time of experiment in relation to other experiments
Matrix Name	Identification of screen type
Well ID	Location of well within screen
Annotation (Crystal Size and Type)	Manually annotated outcome of experiment
Chemical Name and Concentration	Description of chemicals in the well
Buffer pH	The pH of the buffer component of the crystallisation solution (where applicable)

Table 1: Summary of fields contained within the AZ dataset.

The data did not have any information on:

- temperature of the experiment;
- inclusion of ligands;
- purification details;
- protein sequence;
- the final outcome of experiment- diffraction quality/structure solved.

The screens for which the largest number of projects were trialled are five evolutions of a filter screen (a hybrid of grid and footprint screens, which are described in Chapter 8) named Filter 2, Filter 3, Filter 4, Filter 5 and Filter 6 and four generations of sparse matrix screens (random screening) named Sparse 0, Sparse 1, Sparse 2 and Sparse 3. Other screens were either custom screens or follow-up screens containing many experiments with no successful outcome (crystals) and only specific to a single project. The nine main screens together provide 87% (568,957 entries) of the total data covering 152 projects. Throughout this thesis, reference to the AZ dataset means the data from these nine screens.

AstraZeneca Annotation	York Scale
Null, Clear	0
Skin	1
Precipitate	2
Phase	3
Urchin	4
Plate, Needle, Leaf	5
Pyramid, Hexagon, Block	6

Table 2: New annotation of crystallisation results.

The AstraZeneca annotation is given by a crystallographer. On receiving the data the annotations were grouped and assigned a number (York Scale).

The outcome of each crystallisation experiment was scored as shown in Table 2. This allowed the number of classes to be reduced so that each class had more examples. In this classification system an experiment with a score ≥ 4 is considered to be a

successful experiment (crystalline/crystal) and conversely those ◁ 4 are considered to be a fail.

2.3. The Structural Genomics Consortium Dataset

Data was also obtained from the Structural Genomics Consortium (SGC), Oxford in November 2012. In its entirety it relates to 62,605 construct IDs, for which there are 17,591 purification IDs and 54,383 plate barcodes. In total, 608 structures have been solved and PDB IDs obtained. The relationships between the terms construct ID, purification ID and plate barcode are shown in Figure 6. Descriptive metadata includes information pertaining to protein families of proteins, protein sequences, purification methods, crystallisation conditions and whether a solved structure has been deposited in the PDB.

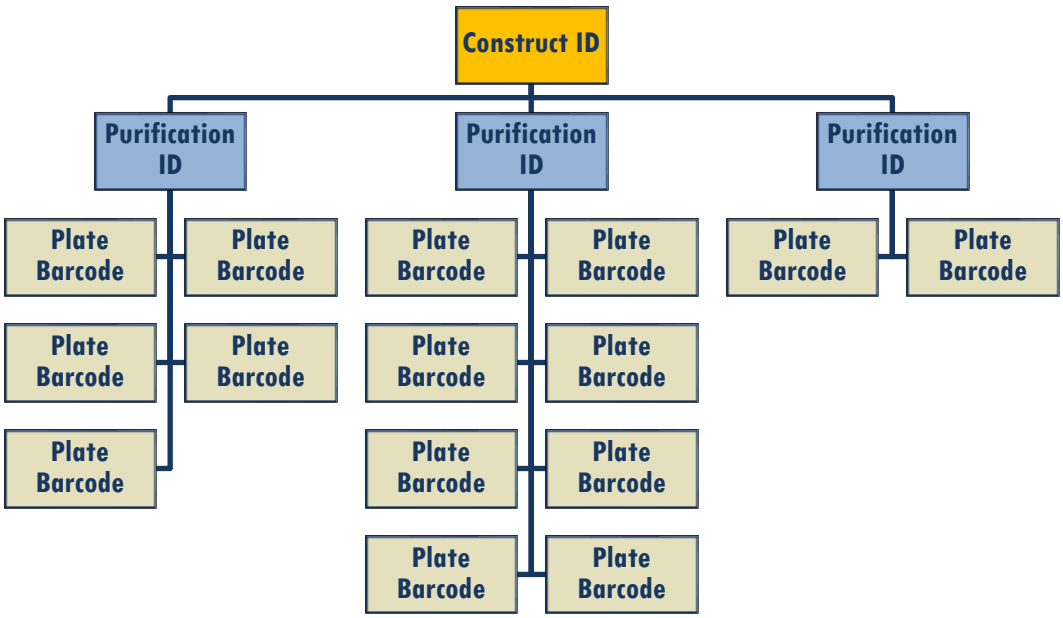


Figure 6: Structure of SGC data.

Each construct ID can be mapped to several purification IDs which in turn can be mapped to multiple plate barcodes. This is because a construct can undergo several different purification processes and the purified protein then trialled tested in various screens, differing in factors such as temperature or whether the protein sample is fresh or has been frozen.

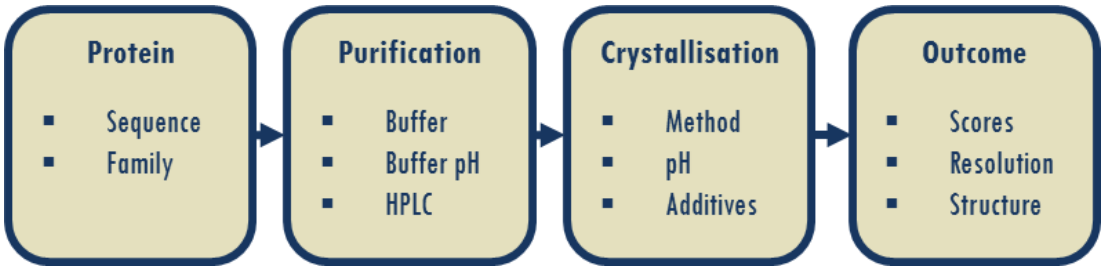


Figure 7: Entity relationship diagram for SGC data.

Table headers indicate the stage of the structure determination process; bullet points show examples of field names for which there is a record.

Figure 7 shows examples of the data fields in the SGC dataset for each phase of the structural determination process. Data analysis was restricted to one screen, *SGC JCSG +4*, which is a sparse (random) matrix screen used to identify regions of parameter space under which a protein is likely to crystallise. Successful regions can then be searched more rigorously using follow up screens. Analysis was performed on this screen because it was the most prevalent screen in the SGC database with 9,608 plate barcodes associated with 4,154 purification IDs and 2,553 construct IDs and. In total, 69 structures have been deposited in the PDB using crystals produced in this screen with at least a further 32 from follow-up screens. Each successful experiment was annotated as crystal (possibly salt), a protein crystal (as determined by X-ray), any crystal diffracting to more than 3.6 Å or as the highest quality-structure determined.

2.4. Crystallisation Conditions from the PDB

The final dataset used was a snapshot of the PDB. Each PDB ID in the standardised PDB snapshot (obtained from Fazio et al. (2014)) has an associated protein sequence and components of the crystallisation solution. After removing a number of potentially malformed entries, the number of PDB IDs was reduced to 60,999 (97% of the original data) to form a dataset referred to as PDB-RAW. Some proteins have been deposited in the PDB many times with different ligands, modifications or crystallisation space groups. For example, hen egg white lysozyme (*Gallus gallus* lysozyme) alone currently has an associated 460 X-ray structures. We also considered the data at different levels of redundancy. The similarity of protein

sequences was assessed using BLAST (Madden, 2012) and only one representative sequence from any sequence grouped as similar (using a p-value of 10^{7}) was included in the PDB-UNIQUE dataset (Figure 8). This dataset comprises 37,249 non-redundant PDB entries. A second dataset, referred to as PDB-BLAST, consists of entries from PDB-RAW grouped according the BLAST analysis. This dataset has 8,958 groups with between one and 2115 (for kinase) similar sequences in each group, giving 59,734 entries in total. After removing duplicate entries with the same experimental conditions for the same protein (but keeping all entries for the same protein where experimental conditions differ), the PDB-BLAST dataset was reduced to 44,063 PDB entries.

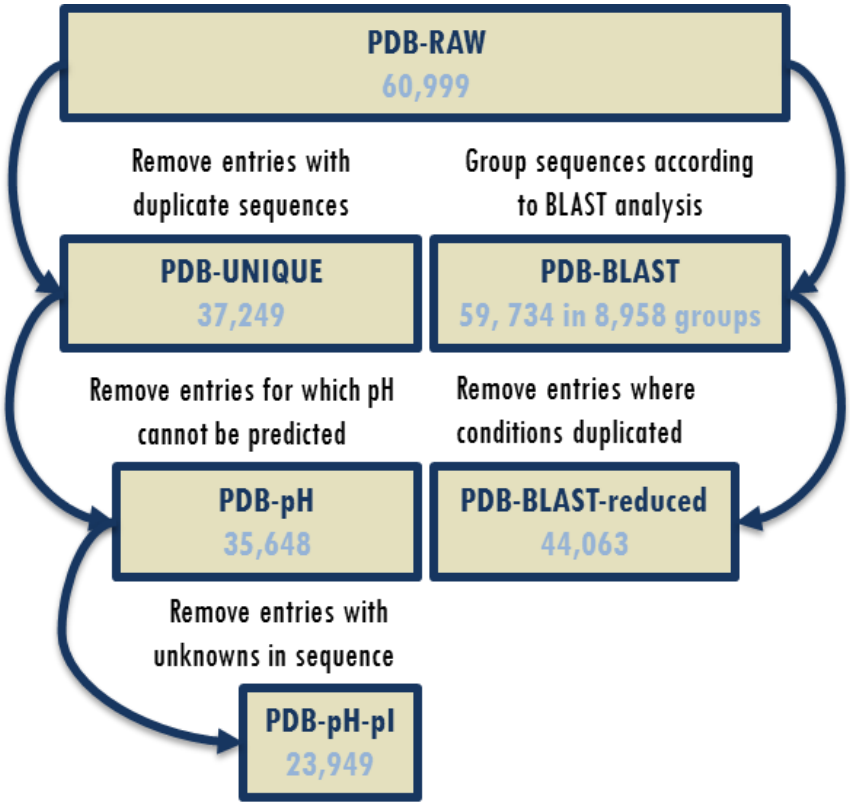


Figure 8: Data structure of the PDB snapshot.

The structure of data used for different types of analysis showing the number of PDB entries in the various data subsets.

2.5. Custom Crystallisation Experiments

Protein	Source*	Concentration (mg/ml^{-1})	Buffer Solution
Protease K212A	h	13	1
Protease K234A	h	13.4	1
Protease K249A	h	12.1	1
ProteaseE171A	h	13.2	1
Concanavalin A	s	15	2
Bovine Catalase	s	15	2
Porcine Trypsin	s	31	2
Thaumatin	s	50	2
α- Chymo A	s	15	2
Galine Lysozyme	s	20	2
Glycolytic A	h	30.1	3
Glycolytic D	h	21.9	3
Glycolytic wt	h	9.8	3
Kinase 1	h	12.2	4

(a) Proteins used for crystallisation.

1	20 mM MES, 5mM calcium chloride, 5 mM DTT, 100mM sodium chloride, 300 mM AHA, pH 6.5
2	10mM PCTP, 100 mM sodium chloride, 0.5mM TCEP, pH 7.6
3	20 mM TRIS HCL , 150 sodium chloride, 2mM TCEP, pH 7.5
4	10 mM TRIS, 50 mM sodium chloride, 1mM DTT, 50 µM zinc acetate, pH 7.5

(b) Buffer solutions used with the proteins named in (a).

h	In-house
s	Sigma

Table 3: Custom protein solution details.

Protein solution details for the commercially available and in-house protein targets that were screened are shown in (a) with buffer solution details in (b).

Commercial proteins were obtained from Sigma-Aldrich and were buffered at pH 7.6. In-house proteins were also buffered at near neutral pH (either pH 6.5 or pH 7.5). For each

experiment, 80µl of crystallisation solution was dispensed using a Thermo Scientific Matrix Hydra II robot. Frozen protein samples were defrosted to room temperature before using a Mosquito pipetting robot (TTP Labtech) to dispense 1µl protein with 1µl of the mother liquor in MRC Wilden crystallisation trays. Trays were sealed manually using transparent, pressure-sensitive adhesive tape (Hampton) and stored in a Formulatrix Rock Imager hotel at 20° C. All images were assessed for crystallisation after 21 days.

2.6. Discussion

One problem with crystallisation data is that it is the result of a high-throughput process and the determination of a protein structure is its goal. This means that scores assigned to images of crystallisation wells, manually, are likely to be targeted at the most promising wells. If the first well scored contains a perfect crystal with good diffraction it is of little interest to the crystallographer what happens in the rest of the wells. This then creates the illusion that protein x is one that is a poor crystalliser (in terms of the number of wells) or that the protein was only tested in a certain set of wells (range of conditions). This latter event is observable in the AZ dataset. Thus, for certain analyses only projects (proteins) that have a score recorded for every well of a screen are included. If a crystallographer scores every image, with each image being of a well of a crystallisation screen, they introduce their own opinion and bias on whether the precipitate is light or heavy or whether the well contains a crystal or just something that shines (such as skin (denatured protein) or cellophane). In an attempt to reduce this bias, we reduced the number of annotations in the AZ data to create the York Scale. Upon creation this was an incremental scale, with 6 suggesting a better crystal than 5 and so on. However, in the analysis that followed, most of the results are reported as if the York Scale was binary, either crystal or non-crystal. For SGC data and PDB data the results of diffraction provided objective evidence that crystals were formed in certain conditions. In our own custom dataset we used the binary system of crystal or non-crystal. These were scored images and the images were taken after 21-days from creation of the experiment. In this time it is possible that crystals were formed and dissolved, however, this aspect of crystallisation was not explored.

3. Methods

Machine learning and data mining methods are used to simplify data, recognise patterns and provide statistical evidence in support of hypotheses. A general overview of the methods used within this work is given here along with an indication of which methods were used in the specific analyses described.

3.1. Cluster Analysis

Clustering is a method of grouping similar objects together based on characteristic properties. In Chapter 9 the objects are crystallisation conditions and their similarity is determined by which proteins they crystallise.

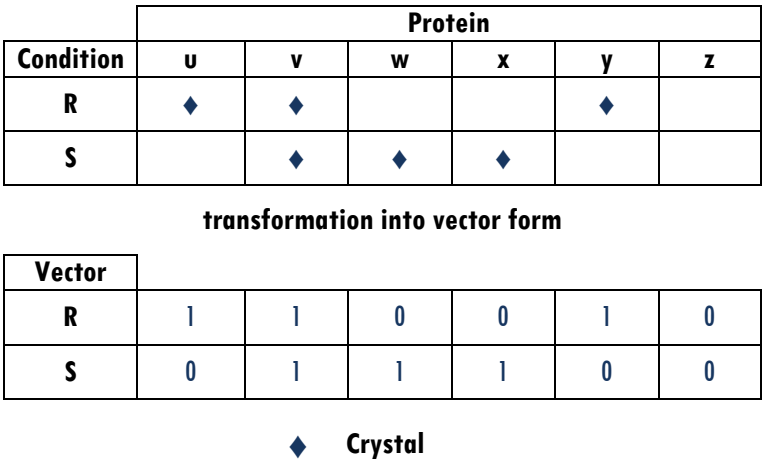


Figure 9: Transforming from experiments to vector form.

The top table shows the results for six proteins (u, v, w, x, y and z) in two different crystallisation conditions (R and S), where the diamonds indicate successful crystallisation. The results are transformed to give two binary vectors, in which a 1 indicates successful crystallisation and 0 indicates a failed experiment.

Figure 9 shows how the results of a crystallisation experiment are transcribed into vector form for clustering. A binary vector of length n is obtained for each crystallisation condition tested, where n is the number of proteins trialled. The ith element of the jth vector is populated with a one if the ith protein crystallised in the jth condition and a zero otherwise. In the example shown in Figure 9, condition R crystallised proteins *u*, *v* and *y* and is used to define the object R in the form of a

vector [1, 1, 0, 0, 1, 0]. Once such vectors have been generated it is then possible to measure their similarity.

There are many different distance metrics available for quantitative data and at least 12 specific to binary data (Cox & Cox, 2010). It is, therefore, an informed trial and error process to decide which measure best reflects the data. In this thesis, the Hamming distance, the Jaccard distance and the Euclidean distance are used. The first two are measures developed for binary data.

3.1.1. Binary Distance Measures

For objects that are defined by binary vectors, the distance between two objects is calculated by combining the differences between corresponding elements in the two vectors. Figure 10 shows how the number of matched and mismatched elements can be counted. Distance metrics for binary data differ in the weight given to the matches and mismatches, for example the number of 1 - 1 matches may be considered more important than the number of 0 - 0 matches.

		Object S: 1	Object S: 0	
Object R	1	a	b	a + b
	0	c	d	c + d
		a + c	a + d	N = a + b + c + d

Figure 10: The relationship between two objects in binary form.

The elements of the objects, *R* and *S*, can either take the form 1 or 0, as previously shown in Figure 9. The value *a* is a count of where there is a 1 in the same position in R and S. Similarly, *b* is are a count of where there is 1 in R and 0 is S, *c* is a count of 0 in R and 1 in S, and *d* is a count of where there is a 0 in the same position in both objects (Cox & Cox, 2010). N is the sum of all the terms, which is the same as the number of elements in the objects. This requires both objects to be defined by a vector of the same number of elements.

With *a*, *b*, *c* and *d* defined as in Figure 11, the Hamming distance, *H*, between R and S is defined as:

$$H(R,S) = \frac{b+c}{a+b+c+d} = \frac{b+c}{N} \qquad \mathbf{1}$$

The Hamming distance counts the number of mismatches, i.e. the elements that are 1 in object R and 0 in the same position in object S, or 0 for R and 1 for S. If all elements for both objects are identical then the Hamming distance will be 0, whereas if all elements of S are different in R then the distance will be 1. Figure 11 shows the terms that each element pair contributes to for an example in which $a = 1$, $b = 2$, $c = 2$ and $d = 1$ so that $N = a + b + c + d = 6$ and the Hamming distance is 0.66.

Object						
R	1	1	0	0	1	0
S	0	1	1	1	0	0
Hamming Term	b	a	c	c	b	d

$$H(R,S) = \frac{2+2}{1+2+2+1} = 0.66\dot{}$$

Figure 11: Example Hamming distance calculation.

The Hamming distance between objects R and S is 0.66. This is based on a count of 2 for term b plus a count of 2 for term c, divided by the total number of elements in the objects, 6.

As protein crystallisation experiments are much more likely to fail than to result in crystals, the data is negatively biased and therefore comparison of conditions using the Hamming distance shows two conditions to be highly similar if both fail to crystallise the same proteins, which will often happen. Although this provides information, it obscures the desired identification of conditions that crystallise the same proteins. Use of the Jaccard distance compensates for the effect of negative bias.

The Jaccard distance, *J*, between objects R and S is defined as:

$$J\,(R,S) = \frac{b+c}{a+b+c} \qquad \mathbf{2}$$

Equation 2 shows that the Jaccard distance, in which two failed experiments (zero outcomes) *d* are not considered and avoids problems caused by negative bias (Cox & Cox, 2010, Teknomo, 2006). An example of the Jaccard distance is shown in Figure 12 using the crystallisation results from Figure 9.

Object						
R	1	1	0	0	1	0
S	0	1	1	1	0	0
Jaccard Term	b	a	c	c	b	d

$$J(R,S) = \frac{2+2}{1+2+2} = 0.8$$

Figure 12: Example Jaccard distance calculation.

The Jaccard distance between objects R and S is 0.8. This is based on a count of 2 for term b plus a count of 2 for term c, divided by the total number of pairs which contain a 1.

3.1.2. The Euclidean Distance

The Euclidean distance is perhaps the most widely used distance metric and measures the distance between two objects as a straight line 'as the crow flies'. The Euclidean distance, *E*, between objects R and S is defined as:

$$E(R,S) = |R-S| = \sqrt{\sum_{i=1}^{n} |R_i - S_i|^2} \qquad \mathbf{3}$$

where R_i, S_i are the *i*th elements of the *n*-dimensional feature vectors **R** and **S,** which may be real numbers but can also binary variables. In Chapter 5 quantifiable features

are calculated from protein sequences and the use of clustering to predict crystal quality is investigated. An example of the Euclidean distance is shown in Figure 13 using the crystallisation results from Figure 9.

Object						
R	1	1	0	0	1	0
S	0	1	1	1	0	0
$\lvert R_i - S_i \rvert^2$	1	0	1	1	1	0

$$E(R,S) = \sqrt{4} = 2$$

Figure 13: Example Euclidean distance calculation.
The Euclidean distance between objects R and S is 2. This is based on the square root of the sum of distances between each element of the objects

3.1.3. K-means Clustering

The aforementioned metrics allow the distance between objects to be quantified. *K*-means clustering provides a method to group *n* objects into *k* clusters based on their similarity, where *k* is a number to be specified by the user. The process for this grouping is as follows (MacQueen, 1967):

1. Randomly place *k* points throughout parameter space, to represent cluster centres.
2. Assign each object to its nearest cluster centre.
3. Redefine the cluster centre as the average of all the objects in that cluster.
4. Repeat steps 2 and 3 until convergence.

In Chapter 5 *k*-means clustering, implemented in the R programming environment (R Core Team, 2012), is used to determine whether four groups of proteins cluster, based on 13 properties calculated from their sequence. Here, the Euclidean distance metric is used with $k = 4$.

3.1.4. Hierarchical Clustering

Hierarchical clustering is another method of grouping *n* objects based on their distance from each other. Hierarchical clustering can be agglomerative or divisive. In agglomerative clustering every object is initially considered as an individual cluster and clusters are gradually combined according to their similarity until all objects belong to a single cluster. Divisive clustering on the other hand begins with all objects in a single cluster and continues to divide clusters until all objects are considered as separate clusters. The process of agglomerative clustering is as follows (Johnson, 1967):

1. Assign each object to an individual cluster, giving *n* clusters.
2. Calculate the distance matrix between all pairs of clusters.
3. Group the two clusters with the smallest distance, giving one less cluster.
4. Calculate the distance between the new cluster (step 3) and all other clusters.
5. Repeat steps 2 to 4 until all objects belong to the same cluster.

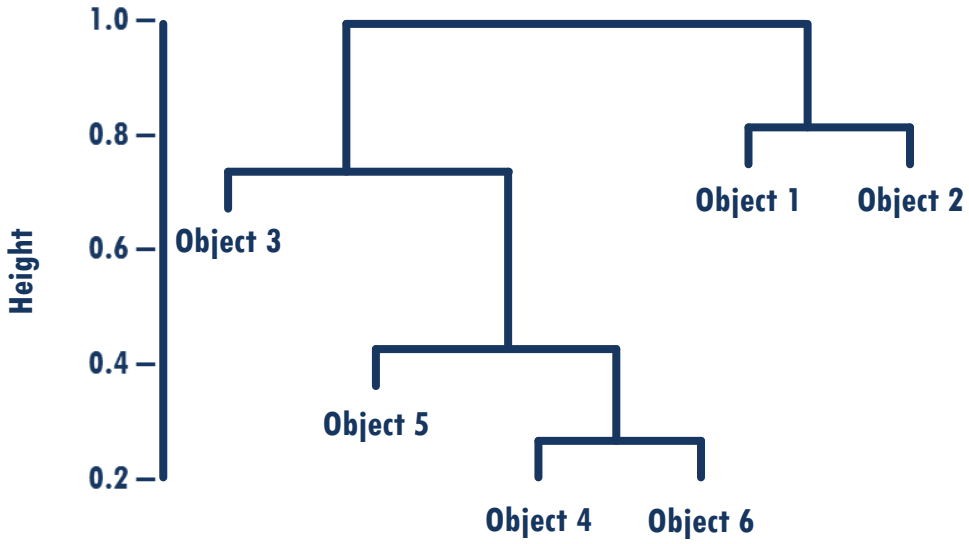


Figure 14: An example dendrogram.

The dendrogram shows the hierarchical clustering of six objects. Here, objects 4 and 6 were the first two to be clustered; object 5 then joined this cluster, followed by object 3. Objects 1 and 2 formed a new cluster that finally joined the cluster with the other four objects.

Hierarchical clustering can be viewed graphically by means of a dendrogram, in which the heights at which clusters are joined are proportional to the distances between the clusters, as shown in Figure 14. Hierarchical clustering and dendrograms are used Chapter 9, to compare the representation of differences between crystallisation conditions given by C6 distance metric against experimental data.

Both *k*-means and hierarchical clustering can be varied by choosing different distance metrics and which point in a cluster is considered to be representative of that cluster to define the distance between clusters (the linkage criteria). The average linkage method used in this thesis calculates the distance between two clusters as the average distance between each object in one cluster to every object in the other cluster.

3.2. Eigenpairs

A number of multivariate methods, including Principal Components Analysis and Linear Discriminant Analysis, described in the next sections, rely on the use of eigenvectors and eigenvalues. Eigenvectors are the vectors that undergo no transformation, other than scaling, when multiplied by a matrix. They are defined in the following way: let **M** be an $n \times n$ square matrix, and let **e** be a column vector of length n. Then the constant λ is an eigenvalue of **M**, with corresponding eigenvector, **e**, if $\mathbf{Me} = \lambda\mathbf{e}$. The eigenvector and corresponding eigenvalue together are referred to as an eigenpair. Eigenpairs are obtained from square matrices and are typically found using iterative computational methods. A matrix of orthonormal vectors (unit vectors that are perpendicular to one another) is representative of a rotation in Euclidean space and therefore can be used as a data transformation matrix

3.2.1. Principal Components Analysis

Principal Components Analysis (PCA) is a technique that uses eigenpairs for data reduction and visualisation of a feature matrix, **X**. A new coordinate system, **P**, is obtained by a rotation that maximises that variance in the first few dimensions and is achieved by finding the eigenvectors, **A**, of the data covariance matrix. The eigenvector with the largest eigenvalue is referred to as the first principal component

and gives the direction in the data with the largest variation. The second principal component is orthogonal to the first and is the eigenvector with the second highest eigenvalue, giving the direction in which there is most variation not already accounted for by the first eigenvector and so on (Rajaraman & Ullman, 2012). For dimensionality reduction only those eigenvectors with eigenvalues of 'significant' size are used. The number of eigenvectors may be chosen according the proportion of the total variance accounted for, but the choice is subjective and a number of different rules of thumb exist (Valle *et al.*, 1999). While removing eigenvalues and their respective eigenvectors results in a loss of some information, ϵ, this is minimised by ensuring that most of the variance in the data is in the first few components.

A data matrix **X** with *n* observations with *m* features, transformed by *k* principal components can be written as

$$\boldsymbol{X} = \boldsymbol{P}\boldsymbol{A}^T + \epsilon \qquad \mathbf{4}$$

where, **P** is a $n \times k$ matrix of the transformed data, or scores, **A** is a $k \times m$ matrix of the eigenvectors, or loadings and ϵ is the $n \times m$ matrix of residuals when $k < n$. In Chapter 5, the PCA scores are used for visualisation to determine any grouping of proteins according to 13 calculated protein features, $m = 13$ and $k = 2$. Protein properties are also used in Chapter 6 in the classification of proteins that can and cannot be crystallised. PCA was employed in order to reduce the set of properties, $m = 87$, before use in for machine learning algorithms. A review of dimensionality reduction techniques in 2009, found that PCA could not be outperformed by non-linear techniques (Van der Maaten *et al.*, 2009). PCA was implemented in the R programming environment using the *prcomp* function (Zurich, 2012, R Core Team, 2012).

3.2.2. Linear Discriminant Analysis

Linear Discriminant Analysis (LDA) is a linear classification method that attempts to maximise the separation between classes. Good separation is found when the

difference between the class mean and the grand mean (mean over all classes) is large compared to the standard deviation of a class.

For a two-class system, the Fisher linear discriminant function achieves this by maximising the function

$$J(\boldsymbol{\alpha}) = \frac{(\tilde{u}_1 - \tilde{u}_2)^2}{\tilde{\sigma}_1^2 + \tilde{\sigma}_2^2} \quad \mathbf{5}$$

where $\tilde{u_i}$ and $\tilde{\sigma}_i^2$ denote the projected mean and variation of class *i* and $\boldsymbol{\alpha}$ is the vector of coefficients, or loadings, in the linear discriminant function. The function $J(\boldsymbol{\alpha})$ is maximised by maximising the difference between the projected group means whilst minimising the projected within group variance. The vector of coefficients $\boldsymbol{\alpha}$ determines the first linear discriminant function and is one-dimensional projection that gives maximal separation between groups. A second discriminant function can be obtained that separate the groups in a way that has not already been exploited by the first discriminant function. In general, the *k*th discriminant function D_k is chosen so that the within-groups covariance between this and each of D_1,..., D_{k-1} is zero. Test data is transformed using the loadings derived from the training data and the transformed points are assigned to the class with the mean closest to them (Balakrishnama & Ganapathiraju, 1998).

LDA was employed in Chapter 6 to separate crystallisable and non-crystallisable proteins, based on calculated properties. The LDA loadings allow the features that are important for class separation to be identified. LDA was implemented in R using the *lda* (Zurich).

3.3. Feed Forward Neural Network

An artificial neural network (ANN), such as that indicated by the schematic in Figure 15, is a machine learning technique inspired by the neurons of the brain. They work by being trained to associate particular outputs with particular inputs. Weighted input features are combined and the output determined by a transfer function in a threshold unit (layer 1). The output values from one layer are then input to the next layer (layer

2 or 3) or output for the user (layer 3). During training if the output is incorrect the weights on the network are amended to improve the output. For example, the new weight, w_{new} for a single neuron is defined as:

$$w_{new} = w_{old} + (t - a)x \quad \textbf{6}$$

where w_{old} is the weight prior to output, t is the target output, a is the actual output and x is the feature associated with that neuron.

There are many types of neural network, with different numbers of layers, training functions and transfer functions. The type of network employed in this thesis is a feed-forward perceptron network, an example of which is as shown in Figure 15, in which each layer passes information to the next. The threshold units in the hidden layers are tangent sigmoid functions, where the weighted input is converted by a tangent function limited between $y = \pm 1$ and in the final layer is a linear output, which, for classification, determines the class. The training method used is the Levenberg-Marquardt backpropagation method. This method is designed to train a network in a time efficient manner (MathWorks, 2011) by reducing the weights on the nodes in proportion to the size of the error (Beale & Jackson, 1990, Hagan & Menhaj, 1994).

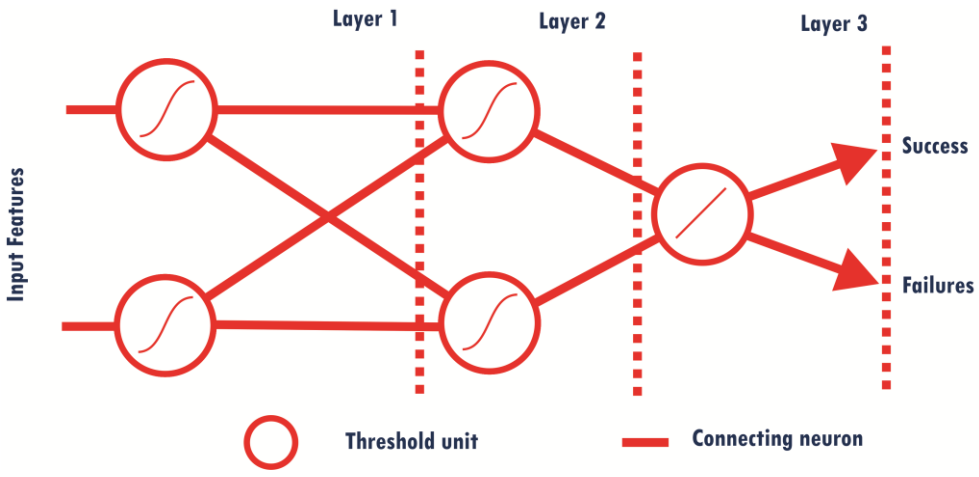


Figure 15: A schematic of an artificial neural network.

An example of a multi-layered perceptron for classification with three layers: two layers each with two tangent-sigmoid threshold units and a third layer with a single node giving the output.

In Chapter 6 a three-layered perceptron is used to classify proteins into crystallisable and non-crystallisable, whereas in Chapter 5 a simpler network is used with just one layer consisting of 5 nodes. This network is used to predict the pH of crystallisation conditions. Neural networks have been shown to be as good as other machine learning techniques for classification (Nookala *et al.*, 2013). Neural networks were implemented using the Matlab neural network toolbox (MathWorks, 2011).

3.4. Measuring the Performance of Classifiers

Determining the success of a classifier requires a metric to measure its performance. For Chapter 6 the accuracy is based on the true positives (TP) and the true negatives (TN), which here are the number of proteins correctly predicted as crystallisable and non-crystallisable respectively, and the false positives (FP) and false negatives (FN), here the number of proteins incorrectly predicted as crystallisable and non-crystallisable respectively. These terms are combined, as follows, to give a measure of accuracy as the percentage of correct classifications:

$$Accuracy\ (\%) = \frac{TP + TN}{TP + TN + FP + FN} \times 100 \qquad \textbf{7}$$

3.5. Pearson's Product-Moment Correlation

The correlation coefficient measures the linear relationship between pairs of observations (coordinates) within two variables (x and y). It is obtained by using a measure of the deviation of all the points away from the straight line $y = mx + c$. The coefficient gives a value between 1 and -1, where 1 indicates perfect positive correlation; as variable x increases, so does variable y; and -1 indicates perfect negative correlation, as variable x increases, variable y decreases. A value of zero indicates no linear link between variable x and y. Pearson's Product-Moment Correlation, used in this thesis, can be computed using the following equation:

$$r = \frac{\sum(x - \bar{x})(y - \bar{y})}{\sqrt{\sum(x - \bar{x})^2 (y - \bar{y})^2}} \qquad \textbf{8}$$

where $\bar{x}$, $\bar{y}$ are the means of the populations x and y.

The significance of the correlation coefficient can be determined to show whether the coefficient is different enough from zero to suggest a relationship. The T statistic

$$T = r\sqrt{\frac{n-2}{1-r^2}} \quad \textbf{9}$$

where r is the correlation coefficient (Equation 8) and n is the number of pairs of observations. T is then compared to critical values for a t- distribution (with n-2 degrees of freedom) in order to determine whether the null hypothesis (there is no relationship) should be rejected. In Chapter 4 the correlation coefficient is used to indicate the relationship between pH measured with a meter and that obtained by spectrophotometry.

3.6. Regression Analysis

A regression model shows the relationship of a dependent variable with one or more predictor variables. The simplest regression model is a linear relationship between one input (predictor variable) and one output (the dependent variable) and is of the general form:

$$\hat{y} = \beta_0 + \beta_1 X_1 \quad \textbf{10}$$

where $\hat{y}$ is the modelled dependent variable, β_0 is a regression coefficient with no predictor variable, β_1 is the regression coefficient effecting the predictor variable, X_1. The model can be extended to include numerous predictor variables, with the general form:

$$\hat{y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n \quad \textbf{11}$$

where the values for β_n are obtained so that sum of the squared error between the y (observed) and $\hat{y}$ (modelled) is minimised. Analysis of the residuals (y-yˆ) can show whether the relationship between the predictor variables and the dependent variables is nonlinear. The model should have residuals that are independent and form a

normal distribution, suggesting that the model is not biased towards particular points. If this is not the case, it may be possible to transform the data, by algebraic manipulation, so that it is in a linear form.

3.7. Normality and Significance

3.7.1. Determining Normality

Probability plots are used to compare the distributions of two samples. Observed data is compared to a theoretical normal distribution that has been generated using the observed mean and standard deviation. The quantiles of these two distributions are plotted against each other and the closer to the line y = x the distribution is, the more similar the distributions. The assessment of this plot is performed manually. Probability plotting is performed in R using the functions *qqnorm* and *qqplot* from the statistics package (R Core Team, 2012).

An extension of this is the Kolmogorov-Smirnov (KS) test, which compares the observed distribution to a theoretical distribution with the same mean and standard deviation. The largest vertical distance (y-axis) between the two distributions (supremum) for any given input (x-axis) provides the statistic, D_{max}. This can be compared to D_{crit}, given by

$$D_{crit} = \frac{1.36}{\sqrt{n}} \qquad \textbf{12}$$

where n is the average sample size of the two distributions, to test the null hypothesis that the two distributions are the same. In this thesis n is the same size for both the observed and the theoretical normal distribution. KS tests are performed in R, using *ks.test* to compare observed data to a normal distribution (R Core Team, 2012).

3.7.2. Mann-Whitney-Wilcoxon Test

The Mann-Whitney-Wilcoxon (MWW) test is a non-parametric test for the comparison of two samples, an analogue of the parametric unpaired t-test in which

no assumptions about the distribution of the data are required. The MWW ranks the observations in the two samples and then calculates the statistic

$$U = n_1 n_2 + \frac{n_1(n_1 + 1)}{2} - R_1 \quad \textbf{13}$$

where n_1, n_2 are the sample sizes and R_1 is the sum of the ranks for the sample with the greatest rank sum. U is compared to a critical value to determine whether the null hypothesis, that there is no difference between the samples, should be rejected. MWW tests are performed in R using the command *wilcox.test (R Core Team, 2012)*

3.7.3. Binomial Distribution

To determine the significance of a binary system, a binomial distribution is required. In Chapter 6 a neural network is used to classify protein sequences into crystallisable or non-crystallisable. As there are only two options, randomly classifying them would result an accuracy of ~50%. A binomial distribution provides a guide to what percentage accuracy could have occurred by random chance, depending on the sample size, *n*.

The mean of the binomial distribution can be defined as:

$$\mu = np \quad \textbf{14}$$

and the standard deviation as

$$\sigma = \sqrt{(np(1-p)} \quad \textbf{15}$$

where *p* is the probability of an event being successful.

A sample of 100 sequences with each having probability of 0.5 of being classified correctly, would create a distribution (near normal) with a mean accuracy of 50%, with a standard deviation of 5%. Assuming that 95% of the data is within 2 standard deviations either side of the mean, it is possible to say that any classification with 40% - 60% could have occurred by chance alone.

3.8. Proportional Error

Proportional error is a measure of the deviation from the true value of a proportional success rate, as defined in Equation 16 .

$$\varepsilon = \sqrt{\frac{p(1-p)}{n}} \qquad \textbf{16}$$

The error is dependent on the probability of success, p divided by the number of observations, n, therefore, the greater the number of observations the smaller the error. It is used in Chapter 7 to provide a margin of error on the success of chemical species relative to the number of times they had been trialled.

3.9. Cross Validation

Cross validation is a method used in the training and testing of datasets to ensure that by chance the random selection originally made does not, in some way, bias the results. Here, one quarter of the data was used for training, with the remaining three-quarters reserved for testing. The training set was then replaced and the process repeated four times, as shown in Figure 16.

1	Training	Testing	Training	Testing
2	Testing	Training	Training	Testing
3	Testing	Testing	Training	Testing
4	Testing	Testing	Testing	Training

Figure 16: Venetian blinds cross validation.

Over 4 iterations 25% of the data is removed for training, the remaining 75% is used for testing. The data is then replaced and the next 25% removed and the process is repeated.

4. Determination of pH Using Spectrophotometry

In protein crystallisation, important parameters include the chemical type and concentration along with its ionic strength, the inclusion of heavy metals, the purity of the reagents and the pH. The pH of the experiment is often a critical parameter. Proteins are solubilised, stabilised and crystallised in a specific range of pH (McPherson, 1989a, Newman, Sayle, *et al.*, 2012). Crystallisation screens are designed to sample pH as well as other parameters such as salts, precipitants and other additives in order to find conditions giving initial crystallisation hits. Optimisation of the conditions is achieved by finer sampling of the parameter space around these initial hits (Jancarik & Kim, 1991, Luft *et al.*, 2003, Luft *et al.*, 2011). For successful optimisation it is essential that the properties of the original conditions are accurately known and reproduced. The pH of a particular solution is often quoted as the pH of the buffer used but this can be highly inaccurate due to the effect of other components in the mixture. This is particularly true for high concentrations of the salts of weak acids and to a lesser extent any molecule which affects the hydrogen ion activity (Kohlmann, 2003). Furthermore, the pH of stock chemicals is known to change over time due to chemical decomposition (Bukrinsky & Poulsen, 2001). As a consequence simply knowing the components of a solution does not mean that the resultant pH will always be the same.It has been shown that the actual pH of crystallisation conditions can be as much as four pH units away from that of the buffer (Newman, Sayle, *et al.*, 2012, Wooh *et al.*, 2003). Figure 17 shows pH measurements from 84 crystallisation solutions (conditions) of a custom sparse matrix screen, NPCF_4, used at the Collaborative Crystallisation Centre (C3), Australia (Newman, Sayle, *et al.*, 2012). The pH of each set of conditions was recorded both as the pH of the buffer and as determined using a pH meter. It can be seen that most solutions, 73% (61/84), have an actual pH within 0.5 pH units of the buffer pH, although 18% (15/84) are between 0.5 and 1 pH units away from the buffer pH and 10% (8/84) are over one unit away.

Accurate measurement of the properties of conditions is particularly important for crystallographers making up their own crystallisation screens. Stock chemicals that are prepared or labelled incorrectly or simply placed in an incorrect location in the robotic system will be incorporated into screens unnoticed. This can be particularly damaging if the chemical is a buffer stock that is included in multiple conditions. Although a well-calibrated and well-maintained pH meter can be used to measure acidity accurately, it is time consuming and impractical as the solution may also require reformatting to accommodate the probe.

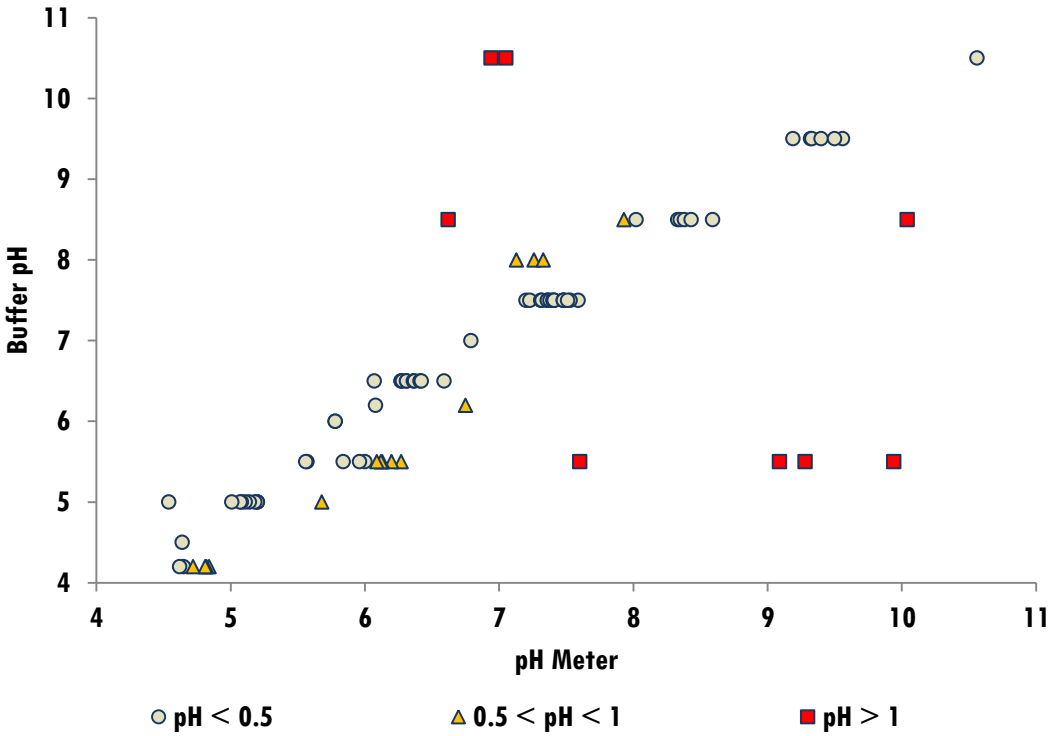


Figure 17: Measured pH in relation to recorded pH of buffer.

The buffer pH in comparison to pH measured with a meter for 84 conditions from the NPCF_4 sparse matrix crystallisation screen. The data was obtained from supplementary information in Newman *et al*. (2012). Differences between buffer pH and meter pH of <0.5, between 0.5 and 1; and >1 are indicated by circles, triangles and squares respectively.

Newman, Sayle, *et al.* (2012) describe a method for high-throughput measurement of pH using the indicator dye Yamada Universal Indicator together with automated imaging. The colour information of a dyed crystallisation solution was recorded as a single hue obtained from an image of a region of the well. This hue value is compared to those obtained for standards prepared from broad-range buffer systems

to provide an estimate for the true pH of the solution. For structural genomics centres and other laboratories with automated imaging systems in place, the method provides a fast, low-cost pH assay with a strong correlation to measurements obtained with a pH meter over the pH range 4.0 to 10.5. The need for a suitable imaging system that will provide consistent reproducible results, however, makes the method unfeasible for many laboratories. Furthermore, Newman, Sayle, *et al.* (2012) found little variation in colour within pH 5.5 to 7.0, a range common in crystallisation trials. Although recognising the limitation of UI, they point out the difficulty in producing dyes with good discriminatory power over a wide pH range.

In this chapter a method to estimate the final pH of a crystallisation solution is described that does not require an imaging system, but instead measures the absorbance of solutions using spectrophotometry. We show that the indicator dye bromothymol blue gives greater discrimination than UI and other dye systems over the pH range 5.5 to 7.5.

4.1. Material and Methods

4.1.1. Preparation of pH Gradients

A 96-point pH gradient (referred to as the 96-point screen) was produced using the two part broad range buffer system PCTP (Newman, 2004) supplied by Molecular Dimensions. The buffer is made from three chemicals: propionate, cacodylate and bis-tris propane in the proportions 2:1:2. One part of the system is created by adding hydrochloric acid, to the aforementioned 2:1:2 solution, until a pH of 4 is reached. Similarly, the second part is made by adding sodium hydroxide until a pH of 9.5 is reached. Ideally the pH would be linearly proportional to the two components of the buffer system. It is in fact sigmoidal as shown in Figure 18. This distribution, also found by the team who developed the buffer (Newman, 2004), shows that a solution containing 70% of the pH 9.5 component (and 30% pH 4 component) should have pH of 7.7, whereas it is actually pH 7.5.

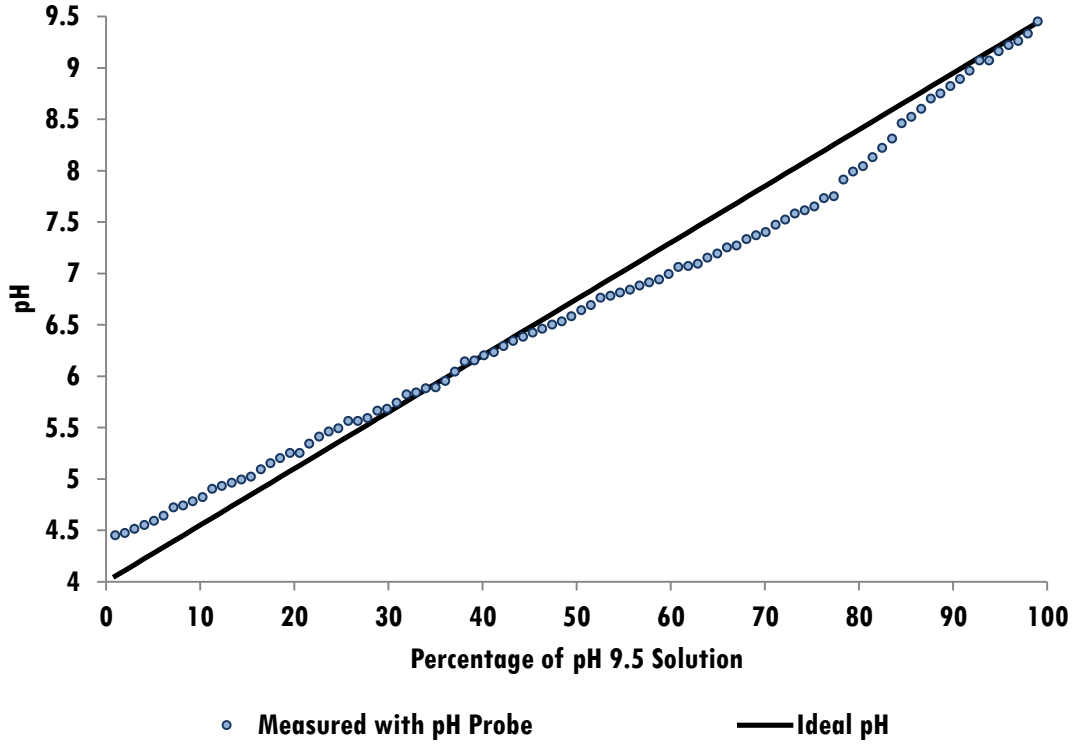


Figure 18: Measured pH of PCTP.

The pH of the broad range buffer PCTP measured with a pH meter in relation to an ideal linear pH for the proportion of chemicals used.

The 96-point screen of PCTP was dispensed into a 96 deep well block using an Emerald Bioscience Matrix Maker at final concentration of 100mM. A second 96 deep well block (referred to as the short screen) was produced where each row (A1-A12, B1-B12 etc.) was composed of a 12 point linear pH gradient 4.0-9.5 (PCTP, 100mM). In order to assess the performance of the spectrophotometric method against common crystallisation buffers a third screen was dispensed (referred to as the "buffer screen") containing buffers in a 12 point range spanning ± 1 of their respective pK_a values with a final concentration of 100mM. The contents of the buffer screen were as follows (rows A-H): sodium acetate (pK_a 4.75), sodium citrate (pK_{a3} 5.40), MES (pK_a 6. 10), sodium cacodylate (pK_a 6.27), sodium HEPES (pK_a 7.50) and Tris-HCl (pK_a 8.30), PCTP pH 4.0-9.5. Row H contained only water which was included as a control. The pH of all three screens was measured using a well maintained and calibrated Jenway 4330 pH meter (with Jenway probe Catalogue Number 924005) calibrated using standards: Fisher phthalate, pH 4.00; phosphate, pH 7.00; and borate, pH 10.00.

4.1.2. Measuring Absorbance

Into a 96-well flat-bottomed Costar 3635 UV/vis assay plate, 20µl of stock solution (Sigma) was dispensed using a Robbins Hydra 96 robot. To this was added 150µl of the 96-point screen using a Thermo Scientific Matrix Hydra II robot and the plate mixed briefly using an orbital plate mixer. The plate was then read using a Bio-Tek powerwave XS UV/visible plate reader programmed to scan across the visible light range from 400nm to 700nm in 5nm increments generating a 61-point absorption spectrum for each well, which was exported to Excel (Microsoft) for data processing.

4.1.3. Curve Normalisation

It has been observed that different chemicals absorb different amounts of light (Silverstein & Webster, 2006, Reusch, 2013). The structure of acid-base indicators is modified by excess hydrogen or hydroxyl ions so that they absorb different wavelengths of the electromagnetic spectrum. This occurs because of the conjugated bonds present in the benzene rings of acid-base indicators. These bonds allow for electrons in π orbitals to be moved by photons from the light wave into anti-bonding orbitals, therefore, removing energy from light wave - essentially it has been absorbed. For example, phenolphthalein when exposed to basic pH has its structure modified such that the energy required to move an electron is reduced and this in turn means that longer wavelengths are absorbed, allowing only those wavelengths, associated with the colour violet (around 400nm) to be seen. In Figure 23**,** row E shows the colour change of phenolphthalein when dissolved in a basic solution. Conversely, when the structure is in a neutral or acid solution the energy required to move electrons is much greater and so wavelengths associated with ultraviolet are absorbed. Variation in volume and concentration can also affect absorbance- as defined by the Beer-Lambert Law in Equation **17** (Crouch & Ingle, 1988).

$$A = \varepsilon \cdot l \cdot c \qquad \textbf{17}$$

The Beer-Lambert Law states that the absorbance of light, *A*, is equal to the molar absorptivity of the solution, ε, multiplied by the concentration, *c* and the length of the solution, *l*, which the light travels through. In short, the further the light has to travel

or the higher the concentration of the sample the more light is absorbed, as represented in Figure 19.

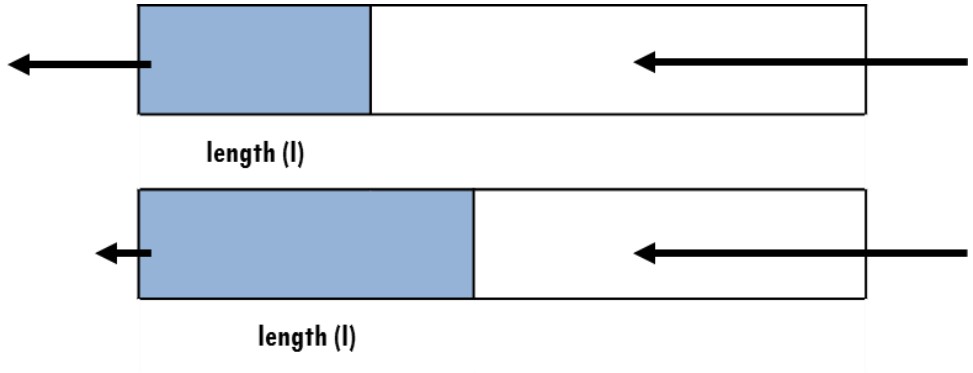


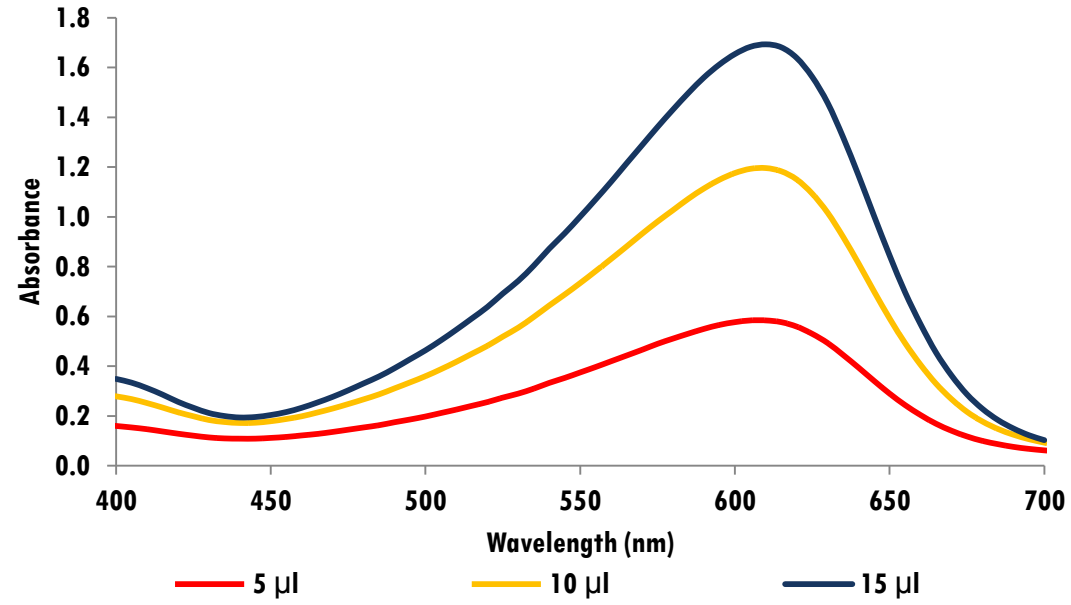
Figure 19: Effect of length on the absorbance of light.

Light is represented as an arrow, where the length of the arrow indicates the amount of energy. The further that light has to travel through a solution; the more of it is absorbed.

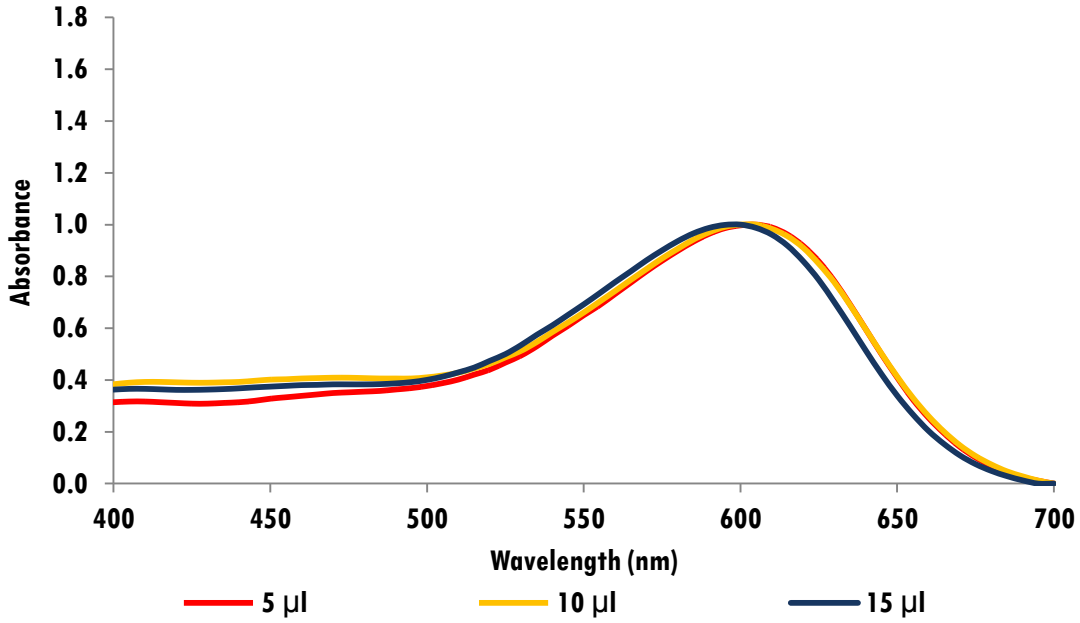
In order to compensate for variation in measured absorbance due to pipetting and mixing errors, Min-Max data normalisation was used. The normalised absorbance at wavelength *x* is given by

$$\hat{a}_x = \frac{a_x - \min(a)}{\max(a) - \min(a)} \qquad \textbf{18}$$

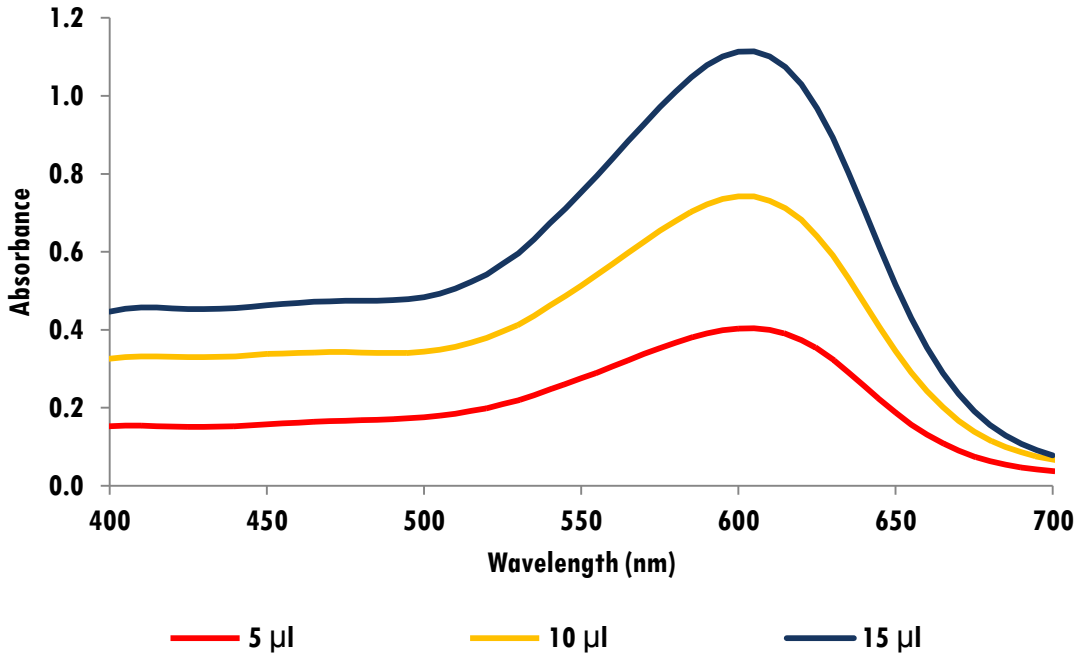
where min(*a*) and max(*a*), are the minimum and maximum absorbance over the range 400nm to 700nm. Figure 20 shows three spectra obtained for PCTP buffer at pH 4.5 and pH 7.5 using different volumes of indicator dye before and after normalisation. It can be seen that normalisation preserves the overall curve shape. In Figure 20(d) the slight translation of the maximum with the addition of 15µl of indicator dye corresponds to a difference of just 0.03 pH units.



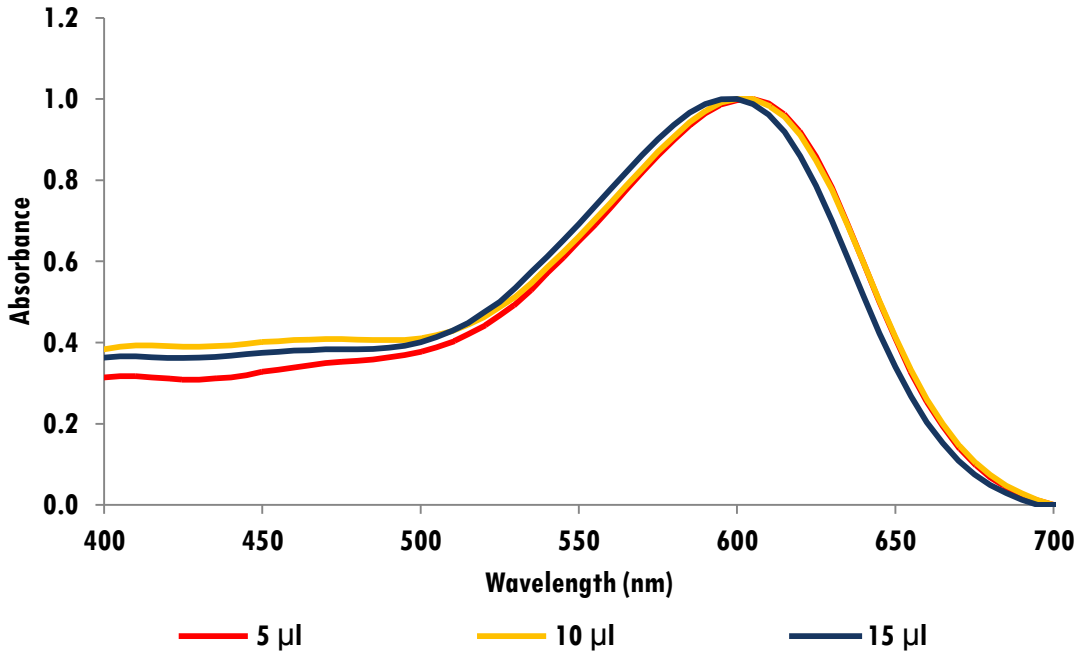
(a) The absorption spectra obtained for PCTP buffer at pH 4.5 with three different volumes of indicator dye.



(b) The three spectra from (a) after Min-Max normalisation.



(c) The absorption spectra obtained for PCTP buffer at pH 7.5 with three different volumes of indicator dye.



(d) The same three spectra shown in (c) but after Min-Max normalisation.

Figure 20: Normalisation of absorbance spectra.

The effects of Min-Max normalisation on absorbance values for light passing through an acid-base indicator solution of pH 4.5 (a) and pH 7.5 (c).

4.1.4. Curve Matching

In order to assign a pH value to a solution using the spectrophotometric method, the normalised spectrum obtained for the unknown solution is compared with normalized spectra obtained for standard curve solutions of known pH. The best match is determined using the smallest Mean Absolute Deviation (MAD) as a distance metric. The MAD between two vectors, *x* and *y*, of length *n* is defined by

$$MAD(x, y) = \frac{1}{n}\sum_{i=1}^{n}|x_i - y_i|. \quad \textbf{19}$$

The pH corresponding to the best match is assigned to the solution of unknown pH. However, when the MAD value to the best match is above a certain threshold, the pH value is still assigned but a warning is given. This threshold was determined as follows.

For each column in the 96-well plate, an artificial absorption curve was produced by randomly generating 61 numbers to represent values from 400nm to 700nm in 5nm increments. This was repeated 100,000 times to represent 960,000 wells in total.

The random absorption values were compared to those obtained for ten 96-point screens (used in the results shown in Figure 25). The lowest MAD for each of the 960,000 *in silica* wells was recorded and a histogram was produced (Figure 21). The distribution for random MAD values was of a normal distribution (confirmed by QQ plot and KS test) around a mean of 0.31 with a standard deviation of 0.03. A threshold was imposed at 3 standard deviations from the calculated mean of 0.31 (theoretically covering 99.7% of all data values that could be derived randomly). This meant the threshold was imposed at the lower limit of 0.23 and therefore any value above this could potentially be obtained from a random distribution of absorbance values.

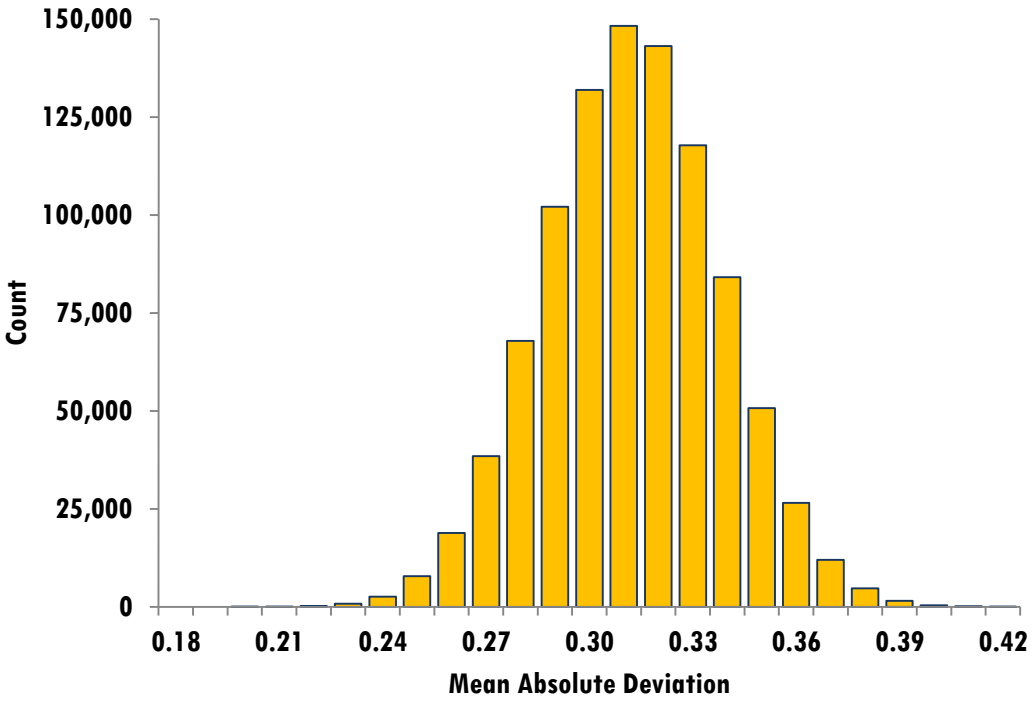
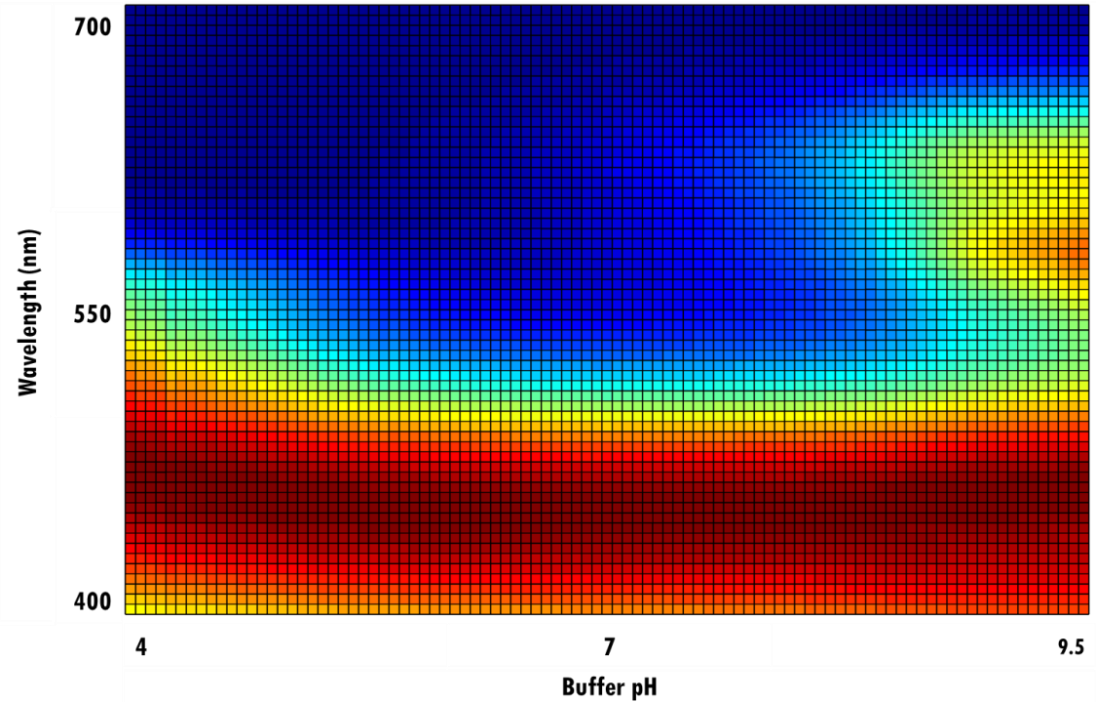


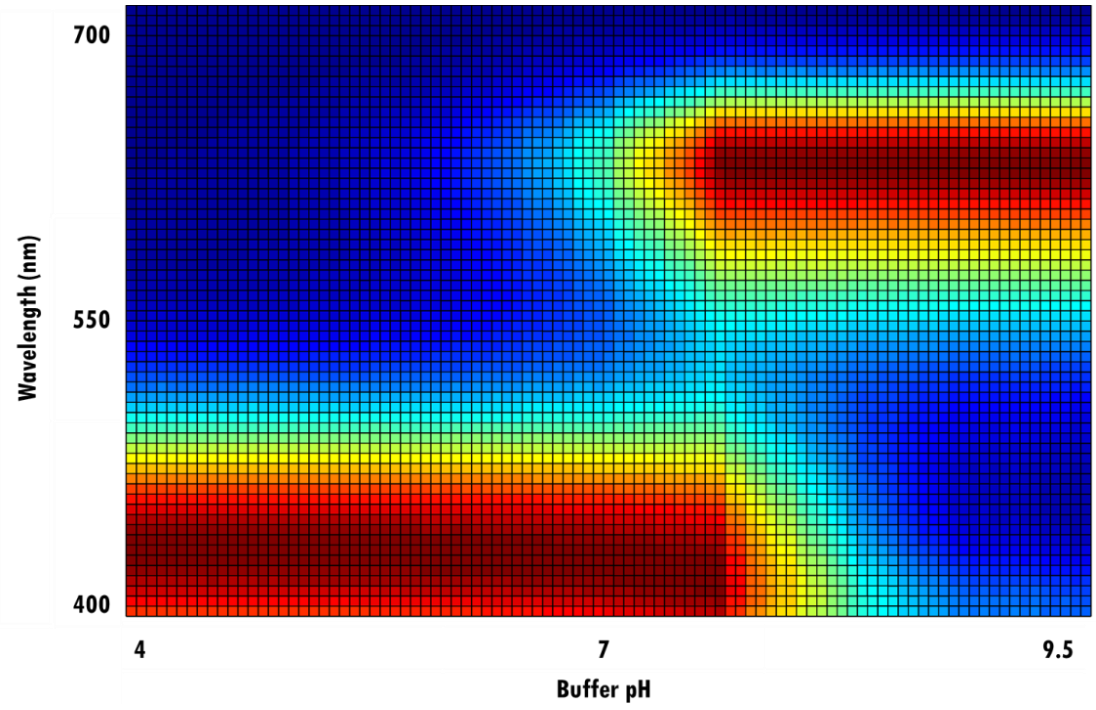
Figure 21: Binned MAD for random absorbance curves.

The distribution of MAD in 0.1 bins for 960,000 artificially generated absorbance curves.

MAD was also used to assess the usefulness of other dyes. A good indicator dye system should have a MAD value representative of the pH change between spectra. The heat plot in Figure 22a shows the absorbance spectrum obtained for the standard solutions using 20 µl of UI. The pH increases from pH 4.5 in well 1 to pH 9.5 in well 96. The span between pH 5.5 and pH 7.0, shows very little difference between the spectral curves, echoing the work of Newman, Sayle, *et al.* (2012) who found the response for UI determined from RGB values to be poor for this range of pH, which is important for protein crystallisation (Kantardjieff & Rupp, 2004). Conversely, Figure 22b shows that bromothymol blue has large MAD values in the range pH 5.5 to pH 7.0 and is able to discriminate between similar pH values. However, Figure 22b also shows that the discrimination between pH values is poorer for the most basic (> pH 7.5) and acidic values (<pH 5.5).



(a) Universal Indicator.

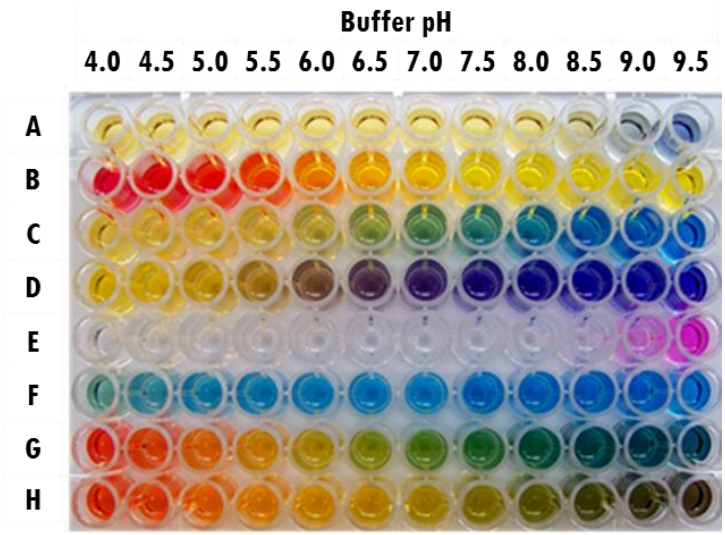


(b) Bromothymol blue.

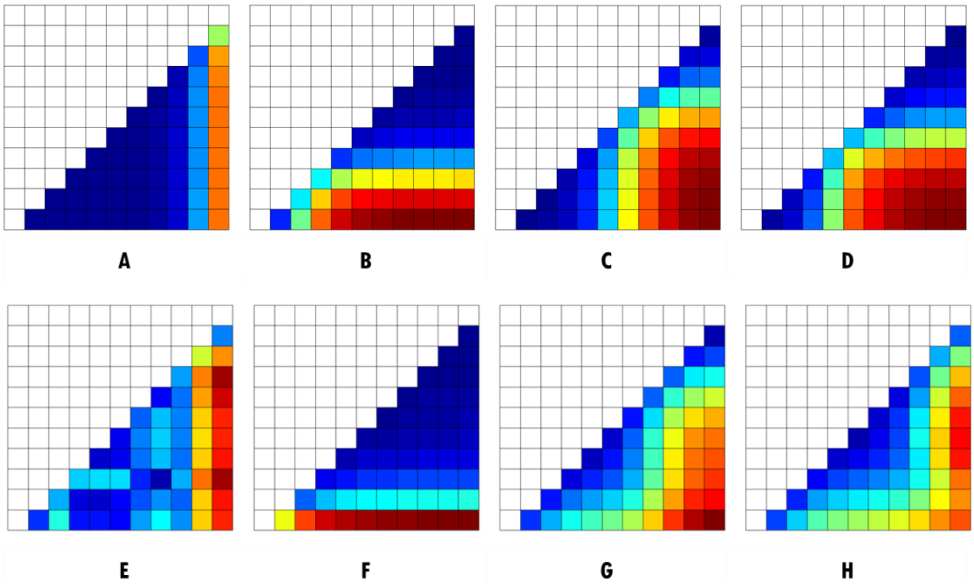
Figure 22: Heat plots of absorbance for different indicators.

Heat plots giving a bird's eye view of the normalised absorbance spectra obtained for the 96 standard curve solutions using (a) Universal Indicator and (b) bromothymol blue.

4.1.5. Testing Other Dyes



(a) Photograph of different indicator dyes in flat-bottomed Costar 3635 UV/vis assay plate.



(b) Heat plots showing the difference in MAD values for different indicator dyes as pH is incremented.

Figure 23: Discrimination between pH values for 8 indicators.

(a) A photograph of the short screen buffer gradient plate with various indicator dyes. The dyes shown are: A- thymol blue; B- methyl red; C-bromothymol blue; D- nitrazine yellow; E- phenolphthalein; F- bromocresol green, G- Universal Indicator minus phenolphthalein; H- Universal Indicator. (b) Heat plots of the MAD between the absorbance spectra obtained the screen shown in (a). Red indicates the highest MAD values (good discrimination) through to blue indicating low MAD values.

Owing to the poor performance of UI it was decided to test other indicator dyes using MAD analysis. The component dyes of UI along with nitrazine yellow and bromocresol green were tested. The component dyes of UI were made up in 100% dimethyl sulfoxide (DMSO) at the concentration ratios they are generally used at in the indicator solution (thymol blue, 1.1mM; methyl red, 4.6mM; bromothymol blue, 8.0mM; phenolphthalein, 31.4mM). Nitrazine yellow and bromocresol green were made up at 2mM in 100% DMSO. A mixture of the UI dyes excluding phenolphthalein was also made by combining the stocks at a 1:1:1 ratio (equivalent to UI without phenolphthalein and referred to as UI-p). It was hypothesised that UI-p may have a better response over the pH 4 to 9.5 range under investigation as phenolphthalein has a sharp colour transition (colourless to fuschia red) above pH 8 and the colour differs from the other components which are of a blue hue. Using a multichannel pipette, 10µl of each dye (20 µl for UI) was dispensed into a separate row of a Costar 3635 UV/vis assay plate, after which 150µl of the short screen was added.

Figure 23a shows the results for the comparison of indicator dyes with the short screen buffer gradient. It can be seen from the photograph of the plate that different indicator dyes change colour at different pH values according to the protonation state of the dye molecule governed by the pK_a of the dye. No single dye covers the entire pH range tested (pH 4.0-9.5) and some dyes have a very narrow transition range. UI (row H) is a combination of thymol blue, methyl red, bromothymol blue and phenolphthalein which capitalises on the complementarity of the dye pK_as and colour transitions (Foster & Gruntfest, 1937).

Calculation of the MAD values for the eight indicator dyes correlates with the observed pattern of colour changes and is shown as heat plots in Figure 23b. The ideal indicator dye would discriminate between pH values across the full range from pH 4 to pH 9.5. Thymol blue (row A), phenolphthalein (row E) and bromocresol green (row F) have narrow response ranges, only changing colour over a small pH range with negligible MAD values between the standard curve spectra for most pH values. Both thymol blue and phenolphthalein only show a response at our most basic pH, giving insignificant MAD values between wells at lower pH. Similarly,

methyl red (row B) and bromocresol green only respond to the most acidic pH and cannot discriminate between wells of higher pH.

Both bromothymol blue (row C) and nitrazine yellow (row D) show a response across a range of pH values with significant differences between the absorbance curves indicated by large MAD values. Figure 23a shows that both indicator dyes are able to discriminate between wells representing the range pH 5.5 to pH 7.5. Notably though, both dyes have very small MAD values at the extremes although bromothymol blue changes more across the basic pH range whereas nitrazine yellow changes more with acidic pH. UI-p only marginally improved the sensitivity of the dye system over the mid-range of pH. Based on these findings it was decided to continue experimentation with the simple bromothymol blue dye system.

Phenolphthalein has a colour transition from colourless to fuchsia at a basic pH and therefore has a very limited range relevant to crystallisation solutions. Figure 23 shows that the indicator phenolphthalein is colourless from pH 4.5 to pH 8.5. If phenolphthalein is used as an indicator that is only assessed by the human eye the only differences detectable would be for crystallisation experiments with pH > 8.5, of which there are very few (Kantardjieff & Rupp, 2004). As the spectrophotometer used to perform absorbance readings is able to provide ultraviolet (UV) light readings too, we are not constrained to those waves of the electromagnetic spectrum that we are able to see (400 − 700nm) and can record absorbance for UV wavelengths from 100 − 400nm. In a study to determine pk_a values using spectrophotometric methods, Tarn and Takács-Novák (1999) show that nicotinic acid and p-aminosalicylic acid have good discrimination between UV light absorbance curves for some of the pH range suitable for protein crystallisation. Although we have not performed any tests with indicators based on UV light, there is the potential to improve upon accuracy and discrimination for certain pH values.

4.1.6. Effects of Protein Buffering

In order to test the effect of protein buffer and protein on the final pH of a standard crystallisation experiment, 10ml of lysozyme solution (Sigma) was prepared at 50mg/ml in 10mM PCTP, 100mM sodium chloride at pHs 5.0, 7.0 and 9.0. Here,

the concentration of the buffer is low so that the buffering capacity of the protein (if any) is not seriously compromised. The pH of each protein solution was adjusted using 10mM sodium hydroxide after the addition of the lysozyme before making up the final volume.

4.1.7. Efficient pH Determination

To improve the applicability of the method, we also investigated miniaturisation of the pH assay using a 384 well Greiner UV plate. For each of the 96-point standard screen solutions, 25µl was pipetted in quadruplicate with 2ul of bromothymol blue indicator dye. The plate was read using the scan function on the plate reader which improved the overall turnaround time from 40 minutes for a 96-well plate to less than 20 minutes for the 384-well plate.

4.2. Results

In order to test the spectrophotometric pH assay with a wider range of crystallisation buffers, bromothymol blue was used in conjunction with the buffer screen as described previously. It was clear from initial results that the row containing only water consistently gave acidic values (Figure 24), possibly due to carbon dioxide from air being dissolved into the water. We therefore only consider our method suitable for determining the pH of buffered solutions.

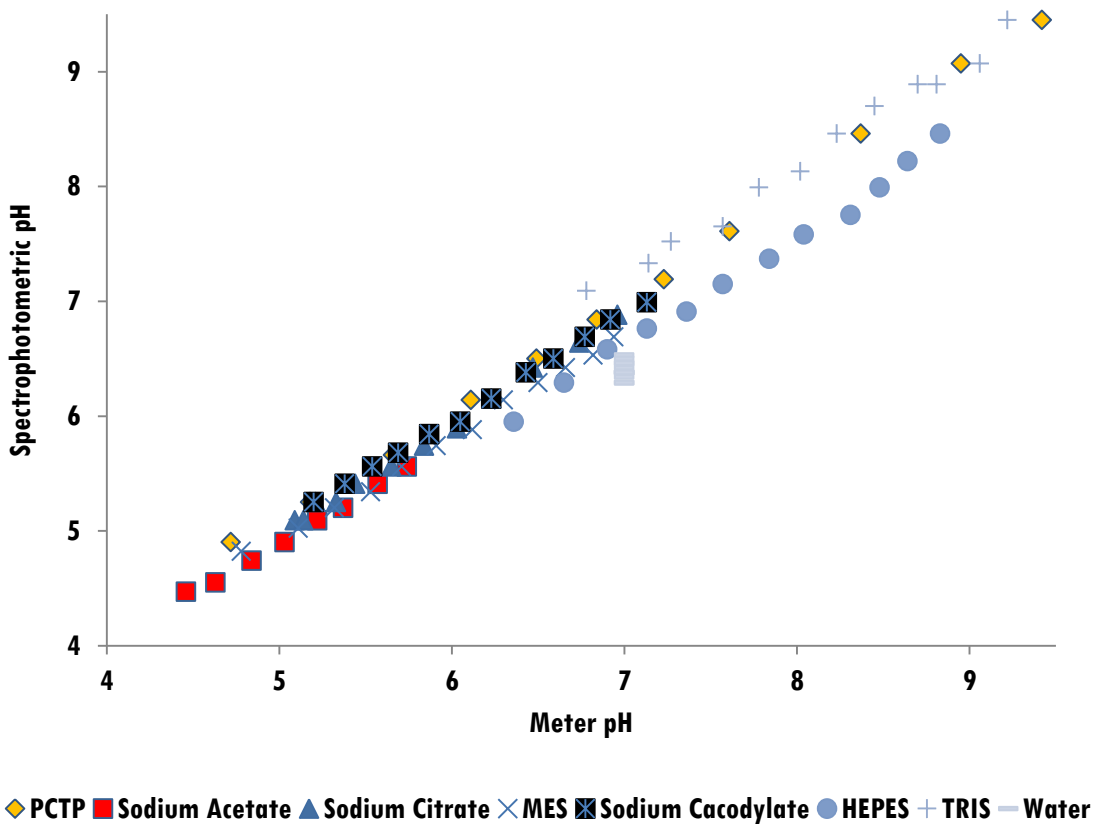


Figure 24: Testing bromothymol blue.

Scatter plot showing the spectrophotometric pH in relation to meter pH for eight chemical species. The correlation between both methods is 0.99.

Figure 24 shows the spectrophotometric pH values for the 96-point buffer screen plotted against the measurements obtained using a pH meter. Only 91 points are shown as five points were measured with pH meter to be outside the pH range of our system. For the buffers there is a very strong correlation of 0.998 between the spectrophotometric and measured pH values. The distribution of deviation is positively skewed, with a mean value of 0.16 for the buffered observations.

In order to test the reproducibility, which is more important than accuracy in crystallisation trials, 7 trays of the buffer screen were dispensed, measured spectrophotometrically and compared with the absorbance values from 10 separate 96-point buffer screens. Figure 25 shows the reproducibility of the system. Correlations of between 0.987 and 0.989 were obtained, with regression slopes between 0.90 (intercept 0.43) and 0.94 (intercept 0.26). As five observations were

removed due to being outside the range (pH 4.5 – 9.5) of the 96-point screen, the graph represents 5,530 observations from 79 buffers and 840 from water. For the best and worst models, the distribution of error was positively skewed, with mean values of 0.17 and 0.27 respectively.

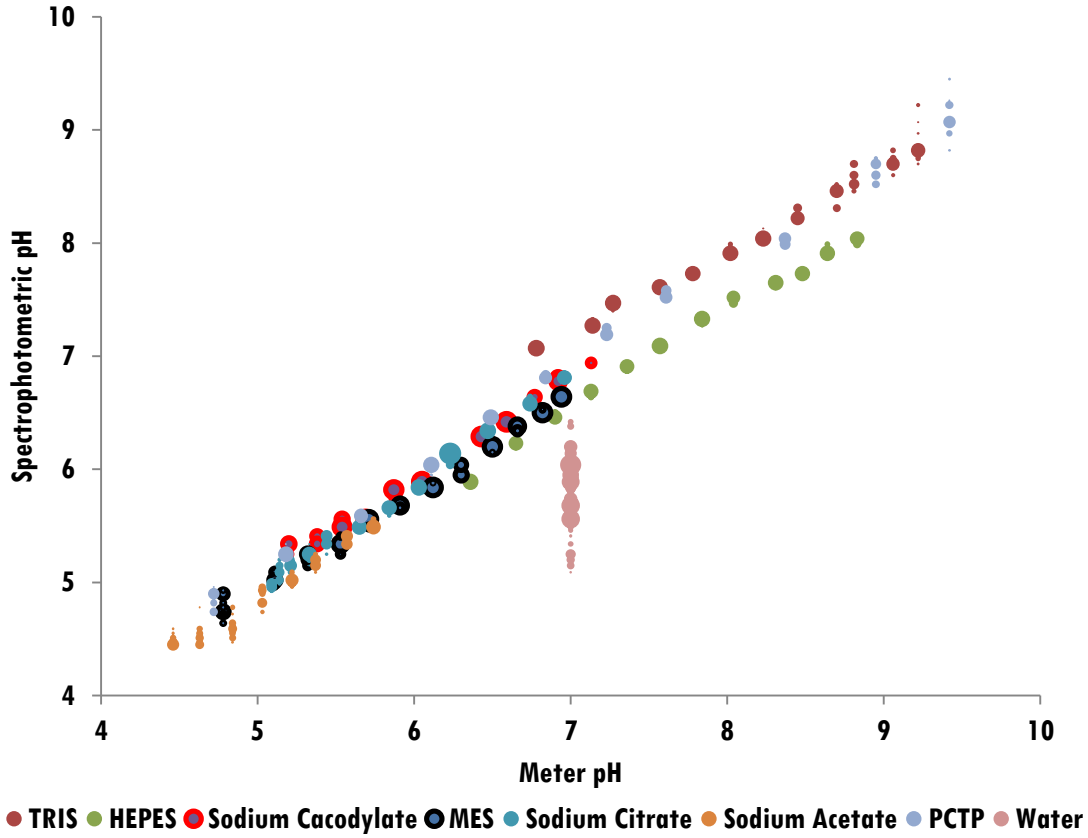
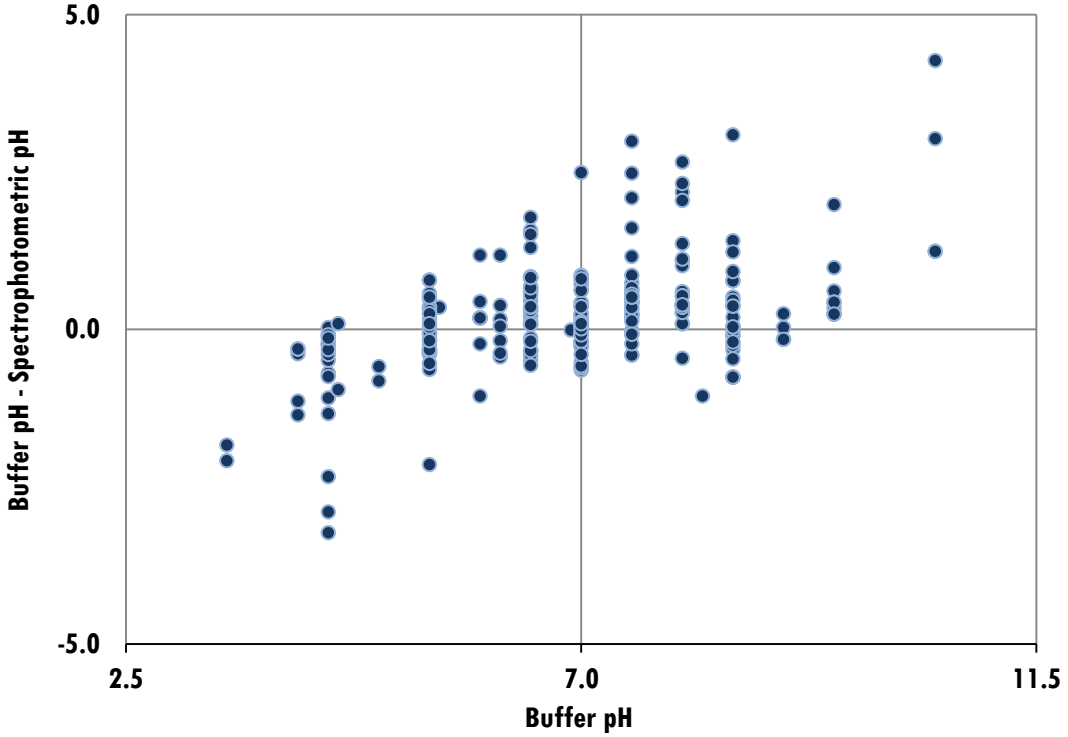


Figure 25: Repetition of experiments with bromothymol blue.

Bubble plot showing the pH values obtained for a set of 79 in-house buffer solutions and 12 containing only water. Bubble size is positively correlated to the number of times a value is repeated.

Bukrinsky and Poulsen (2001) tested the pH of the solutions in the Crystal Screen (Jancarik & Kim, 1991) and found several differed by more than one unit from the pH of the buffer system, with two conditions differing by more than three units. We used our method to test three common crystallisation screens: Hampton IndexHT, Rigaku Wizard and Molecular Dimensions JCSG-Plus. A total of 247 conditions remained after the removal of data points corresponding to wells without buffer and those with a spectrophotometric pH of 4.5 or pH 9.5. Data associated with this latter group was removed due to being assigned a pH at the edges of the 96-point screen,

therefore, there is an increased chance that the true pH could potentially lie outside this range.



a) Distribution of buffer pH in relation to distance from spectrophotometric pH.

Initial pH	pH Change: −	=	+
<7	32	27	41
7	36	45	19
>7	69	19	12

b) Percentage summary of the pH change shown in (a).

Figure 26: Spectrophotometric and buffer pH of commercial screens.

Differences between the buffer pH values and the values obtained by spectrophotometry for the 247 conditions in three commercial screens. b) shows whether the pH increased, decreased or stayed the same for acidic (<7), neutral (7) and basic (>7) buffered solutions.

Figure 26 shows the differences between the buffer pH values and the values obtained by spectrophotometry for the 247 conditions in the three screens. For buffer pH values less than pH 7.0, 27% differ by less than 0.2 pH units (the estimated error in our method) the determined values are higher for 41% and lower for 32%. The greatest differences are for the more acidic buffers, some of which differ by more than two pH units, being more neutral than the buffer pH would suggest. For buffer pH values greater than 7.0, 69% are determined to be more neutral than the buffer pH with even more extreme differences. Only 12% had calculated values more basic than the buffer pH and 19% differed by less than 0.2 pH units. For solutions with a buffer pH of 7.0, 36% were calculated to be more neutral, 19% less neutral and 45% differed by less than 0.2 pH units. Overall, we found that the spectrophotometrically determined values are often more neutral than buffer values. This is particularly true for the most extreme buffer pH values.

Figure 27 shows a histogram for various pH differences with the number of wells in each bin. We found 18 conditions with pH values measured by spectrophotometry were more than two units away from the pH of the buffer (2 for Index, 10 for Wizard and 5 for JCSG-Plus). In the Wizard screen, we determined the pH of a well containing 1.2M of sodium phosphate and 0.2M of potassium phosphate to be 6.23, 4.27 pH units away from the buffer pH of 10.5. In total, 74% of conditions were found to differ from the pH of the buffer by more than 0.2 pH units. Other conditions with a large disparity between our measured pH and the buffer pH included those containing PEGS and ammonium. It is known that PEGs undergo degradation overtime (Jurnak, 1986, Ray Jr & Puvathingal, 1985) and that ammonium compounds slowly release ammonia (Newman, Sayle, *et al.*, 2012, Mikol *et al.*, 1989) and could therefore create problems with reproducibility. Our analysis shows that screens may not search pH parameter space as systematically or specifically as the design intended.

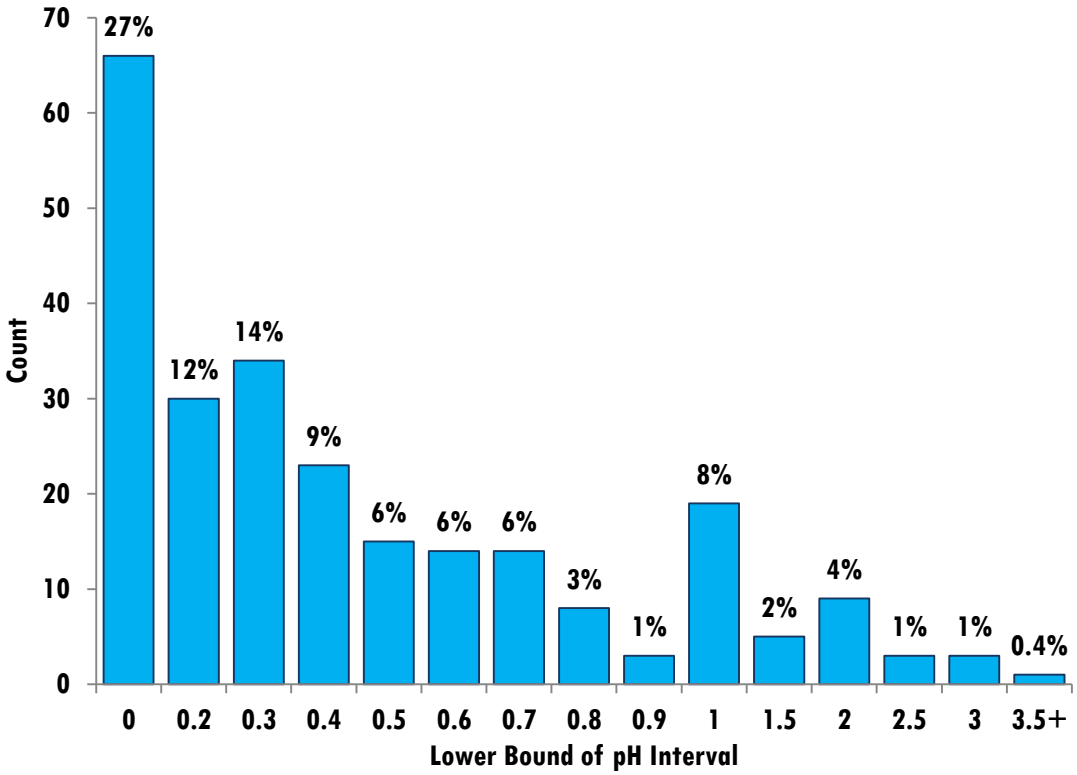
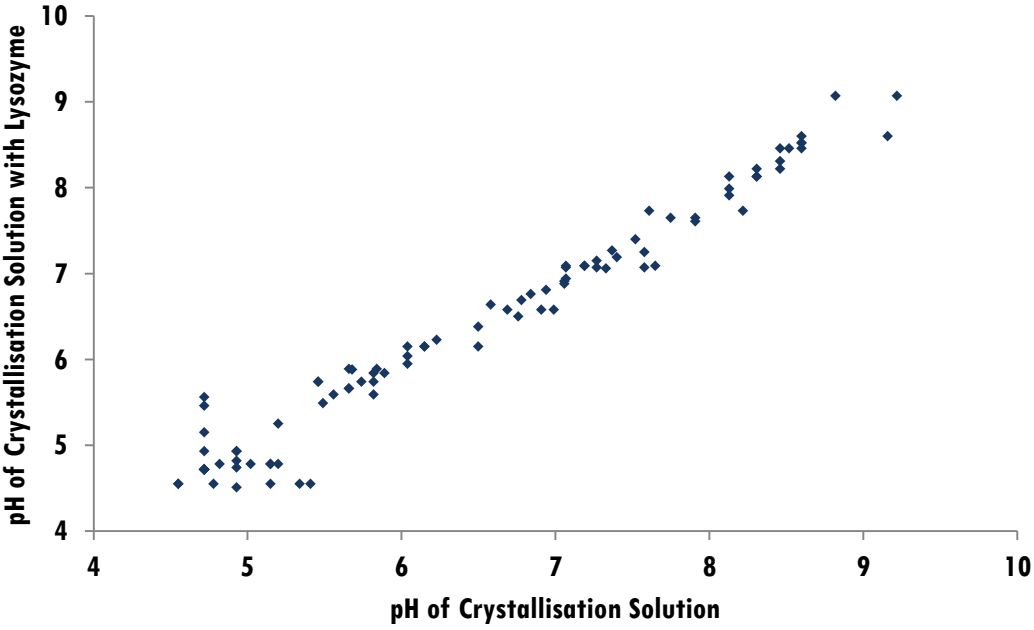


Figure 27: Errors in recorded pH for commercial screens.

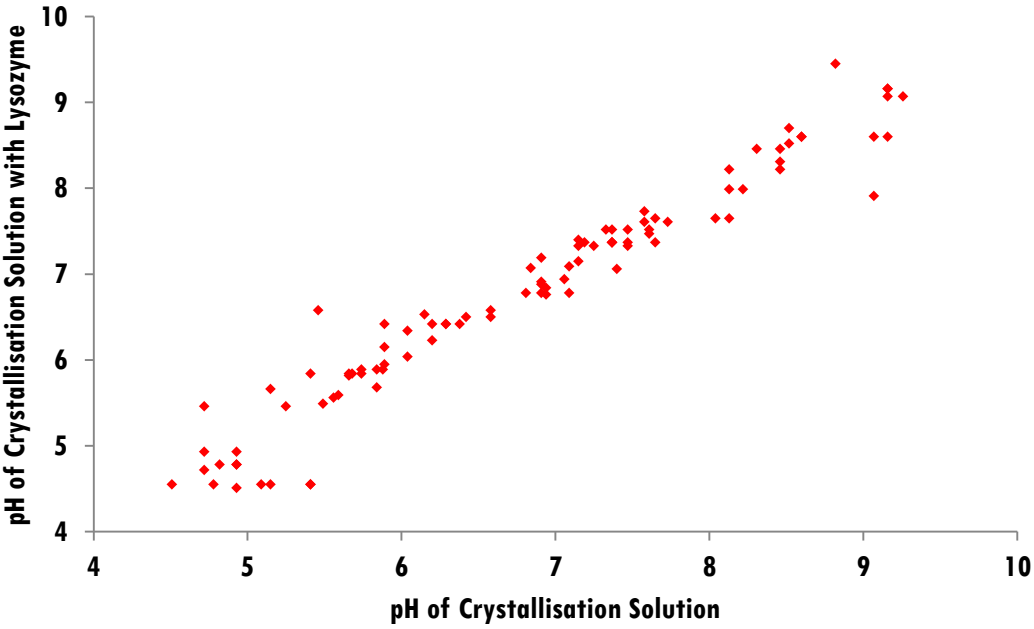
Histogram of pH differences between buffer pH and spectrophotometric pH for 247 solutions used in commercial crystallisation screens. The percentage of the 247 solutions for each bin are shown above the bars.

The results described so far relate to the pH of the crystallisation solution, or mother liquor, rather than mother liquor mixed with a buffered protein solution. Crystallisation occurs at the pH of this mixture, which could differ from that of the crystallisation solution due to the effects of any salts in the protein solution, the pH of the buffer or the protein itself. The effect of protein buffer and protein on the final pH of a standard crystallisation experiment was investigated using lysozyme buffered at pH 5.0, 7.0 and 9.0. It was noted that the addition of the lysozyme shifted the pH considerably; giving values of 3.8, 4.34 and 4.87 before final adjustment for PCTP buffers 5, 7, and 9 respectively. In addition the three buffers were prepared without lysozyme to test the effect of the buffer without protein. A selection of standard crystallisation conditions was dispensed and the pH was determined by the spectrophotometric method. The procedure was then repeated substituting 75µl of the screen for 75µl of water, buffer only and buffer with lysozyme at the pH stated. Figure 28 shows that there is little change in the pH of a solution after the inclusion

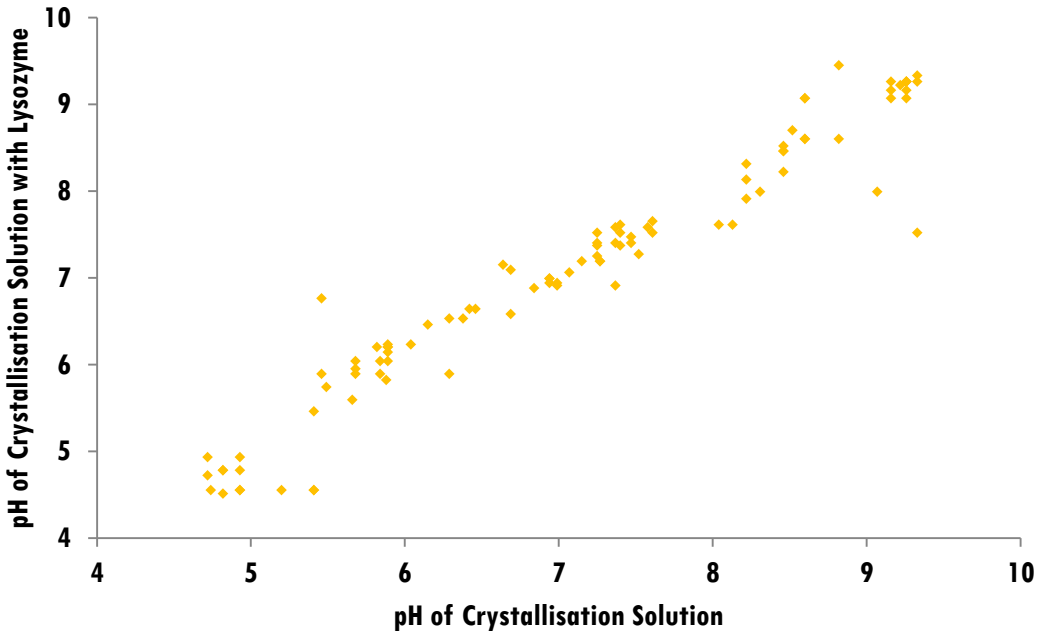
of a buffered protein. All three pH levels of buffered lysozyme have a strong correlation between the pH before and after the inclusion of the lysozyme. The correlation coefficients are 0.98, 0.97 and 0.97, with mean absolute deviations of 0.23, 0.20 and 0.18 for pH 5, 7 and 9 respectively. As these deviations are within the expected error of the method, it is assumed that these differences are caused predominantly by the spectrophotometric system and not the buffered lysozyme.



(a) Protein buffered at pH 5.



(b) Protein buffered at pH 7.



(c) Protein buffered at pH 9.

Figure 28: The effects of protein on buffer pH.

Scatter plots showing the pH of the crystallisation solution plotted against the pH of a 50:50 mixture of crystallisation solution and protein in buffer for lysozyme buffered at three different pH levels: (a) pH 5, (b) pH 7 and (c) pH 9.

Analysis of the data obtained from the miniaturised assay, using a 384 well Greiner UV plate, showed it to be of comparable accuracy to that of the normal volume assay, with a correlation of 0.94 and a MAD of 0.35. The unusual value (6.6, 8.9) corresponds to the buffer MES, for which 11 other measurements fit the expected pattern. When this outlier is removed the correlation increases to 0.97, with a MAD of 0.35, the results are shown below in Figure 29.

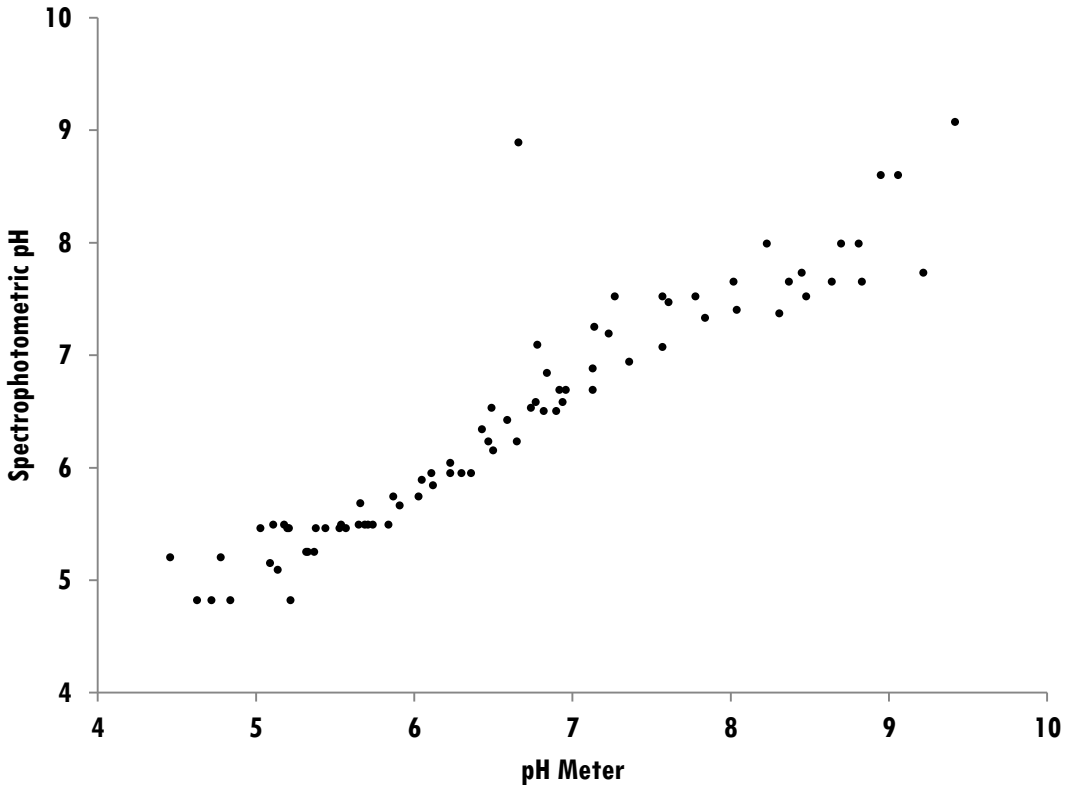


Figure 29: Reduced volume spectrophotometric pH analysis.

Results for the 384-well buffer screen using reduced volumes. The scatter plot shows pH values calculated spectrophotometrically plotted against pH meter measurements, for which the correlation is 0.94.

4.3. Discussion and Conclusions

While the colour based pH assay of Newman, Sayle, *et al.* (2012) is suitable for use in a high-throughput crystallisation facility where automated imaging is already in place, the authors recognised the need for a colour imager as a drawback of their method. They suggested that spectrophotometry could provide a more accessible assay; however, they found using a UV–Vis spectrophotometer to measure absorbance curves unreliable and concluded that the method was not viable. We have demonstrated that the use of spectrophotometry via the visible light plate reader together with the indicator dye bromothymol blue can be used to determine pH with an average absolute deviation of ~0.2 pH units from the pH measured using a pH meter. The comparison makes the pH meter the "benchmark" for a pH reading, although is well known that pH meters can be inaccurate (Illingworth, 1981). We tested the variation between pH meters using three different meters Table 4 and

found the overall average error to be 0.09 pH units. Sodium ion interference at high pH, acid errors at very low pH and temperature can cause measured values to differ from the theoretical pH (Kohlmann, 2003, Beynon & Easterby, 1996). These factors are likely to affect pH in crystallisation trials too as many conditions contain very high concentrations of salts contributing to changes in the activity co-efficient of hydrogen ions and crowding effects. These complex phenomena make relying on the buffer pH in crystallisation experiments inaccurate.

	pH meter			
	Jenway 4330 (1)	**Jenway 4330 (2)**	**Corning 240**	**Average error**
pH 4	4.10	4.02	3.96	0.05
pH 6	6.07	6.02	6.03	0.04
pH 7	7.07	7.05	7.00	0.04
pH 9	9.29	9.13	9.20	0.21
			Overall average error	**0.09**

Table 4: Variance of pH measurement from different meters.

Variation between pH meters was tested using a Corning 240 pH meter and two different Jenway 4330 pH meters. All three meters were equipped with a Jenway pH probe (catalogue number 924005). The solutions tested were phthalate, pH 4.00; phosphate, pH 7.00; and borate, pH 10.00 bought from Fisher Scientific. The readings for the three meters are shown together with the average (absolute) errors for each standard solution.

The indicator dye bromothymol blue gives good discrimination between absorbance spectra in the pH range 5.5 to 7.5, where UI shows a flat response. Bromothymol blue is less reliable, however, at lower pH and above pH 8.0. The vast majority of proteins crystallise within the mid pH range, where bromothymol blue can be used reliably and the use of a single dye avoids the potential impact on reproducibility that would result from a mixture of components. For other uses, for example the quality control of stock solutions where pH falls outside the pH 5.5 to 7.0 range, combinations of dyes are likely to be convenient and effective. Rather than mixing the components in an attempt to provide an indicator dye that covers the full pH range required for protein crystallisation, multiple standard curves could be used. For

example, separate standard curves could be produced for different dyes and the conditions within a screen checked using the appropriate dye and standard curve.

Currently we are only able to provide accurate readings for those solutions with a buffer, as the unbuffered solution have their colour modified due the by the solution becoming more acidic after absorbing carbon dioxide from the air. However, the solution will become saturated with carbon dioxide at a certain point and, therefore, cannot become any more acidic. As we know the pH of water and the spectrophotometric pH of water it should be possible to compensate for this carbon dioxide effect on unbuffered solutions. Even in instances where the compensation is too great, the estimate of the pH should still be more reliable than that of the buffer.

We have developed a fast method that is easy to implement and can provide pH values with a high correlation (0.98) to the measurement made with a pH meter. The pH of crystallisation solutions has been shown to change over time (Jurnak, 1986) and the spectrophotometric method can be used provide a simple check on screens used repeatedly. The method compares favourably with the RGB method to determine pH (Newman, Sayle, *et al.*, 2012) and could be more accessible in that it requires a UV–Vis plate-reader to measure absorbance curves rather than an integrated imaging system. The time required to dispense and read a 96-well plate and calculate the pH values in Excel is approximately 40 minutes, but this was reduced to less than 20 minutes for the 384-well plate using the scan function on the plate reader. Tailoring the wavelength to specific dyes could increase the speed of data acquisition further. For example, it is not necessary to read methyl red at lower wavelengths as the dye absorbs in the higher wavelength region. It may also be possible to make use of different universal indicators provided they contain methoxy reds and phthalein, as these have been shown to be integral to providing good colour discrimination in both acidic and basic solutions(Woods & Mellon, 1941). While this method is fast and accurate it can only be used at the onset of a crystallisation trial and can improve the accuracy of recorded pH going forward. The majority of crystallisation data to date is likely to have the pH recorded as that of the buffer, however, as we know this is likely to be inaccurate, this data is misleading and should be made redundant.

5. The Prediction and Use of pH in Crystallisation

We have shown in the previous chapter that accurate measurements of pH can be obtained quickly for crystallisation solutions using spectrophotometry. This coupled with the knowledge that the buffer pH is sometimes over several pH units away from the measured pH suggests that new crystallisation trials could and should have a closer estimate of their true pH determined. A more accurate pH would assist in accurate reproduction of experimental conditions and provide more meaningful results from the data mining that occurs on structural genomics data (Rupp & Wang, 2004, Hennessy *et al.*, 2000). Conclusions made using data with potentially inaccurate pH values could be misleading. In order to make use of previously generated data, a better estimate of pH than that of the buffer is required. Here, we use the spectrophotometric pH values obtained from numerous experiments to train a neural network to assign pH values to crystallisation conditions. These values are shown to provide accurate estimates of the pH that can be used, for example, when mining databases such as the Protein Data Bank (PDB). Using data obtained from a custom experiment, the SGC and the PDB we investigate predicted pH distributions and attempt to provide evidence for a link between the isoelectric point of a protein and the pH at which it crystallises.

5.1. Prediction of pH for Buffered Solutions

An estimate of pH, for a single chemical species solution, can be determined using its acid dissociation constant (pKa). These constants are obtained from published tables or collected experimentally by chemical titration. The constant is then transformed into a pH using the Henderson-Hasslebalch equation, which defines the number of hydrogens in an equilibrium, from which a pH can be obtained. However, published values are limited to certain chemicals and chemical titration of all chemicals used in crystallisation experiments would be extremely time consuming. Furthermore, there are limitations to the accuracy of the Henderson-Hasslebalch equation (Po & Senozan, 2001) and the constants vary significantly between authors.

It is possible that an accurate estimate of pH can be obtained through the use of regression modelling. The method described in Chapter 4 to determine pH using spectrophotometry was used to determine pH values for buffered conditions in a variety of crystallisation screens, including the JCSG-Plus, the Rigaku Wizard, the Hampton Index and the JCSG +6. Following the removal of those pHs for solutions without a buffer and those at the limit of our system (pH 4.5 and pH 9.5) a total of 5,161 spectrophotometric pH values were obtained.

The concentrations of the chemical species involved were divided into a training set consisting of those conditions with only one chemical species in addition to the buffer and a test set of the conditions with multiple chemical species.

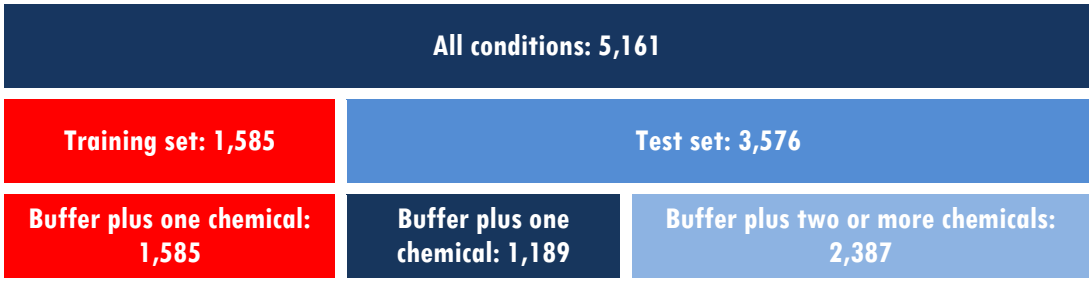


Figure 30: **Organisation of data used for regression modelling.**
The table shows the division of spectrophotometric pH values between training and testing sets used in regression modelling.

It was found that a linear regression model of the form:

$$\widehat{pH_S} = \beta_0 + \beta_1 B + \beta_2 \log_{10} C + \beta_3 B \log_{10} C \qquad \textbf{20}$$

where $\widehat{pH_S}$ is the predicted pH, B is the buffer pH, C is the concentration and the β terms are the regression coefficients, was suitable for each chemical species. Inspection of the regression coefficients for individual chemical species revealed patterns, such as modifying the buffer pH. Chemicals which we assumed to be behaviourally similar (in terms of crystallisation) also had similar regression coefficients and the same predictor variables shown to be insignificant**.**

	Dihydrogen Salts	Ammonias	Hydroxide Salts	Organics
	ammonium dihydrogen phosphate	ammonium acetate	potassium phosphate dibasic	1,2-propanediol
	potassium dihydrogen phosphate	ammonium citrate tribasic	sodium citrate tribasic	glycerol
β_0	1.74	0.74	-6.55	1.67
β_1	0.80	0.92	1.83	0.71
β_2	0.71	1.06	4.03	0.00
β_3	-0.21	-0.16	-0.48	0.00

	PEGs	Salts	Salts of Weak Acids
	jeffamine ed-2001	cadmium chloride	calcium acetate
	pegs of various molecular weights	lithium sulfate	sodium formate
β_0	1.91	1.18	0.20
β_1	0.72	0.87	1.01
β_2	0.00	0.00	1.00
β_3	-0.03	0.00	-0.13

Table 5: Regression models for different types of chemicals.

Each of the seven groups of chemicals is shown with two example chemical species that have been assigned to this group. The lower part of each table shows the coefficients for the linear regression models.

Table 5 shows the final regression models for the seven groups of chemicals: salts, salts of weak acids, organics, polyethylene glycols (PEGs) of different molecular weights and different functional groups, compounds containing ammonia, hydroxide and di-hydrogen salts. Regression models were calculated for each group, after removing 10% of the data from each group for validation. This grouping of chemicals not only provides a more reliable predictive model due to the increased sample size but it also allows new chemicals that are not present in the training set to be assigned to a group and an estimate of pH obtained.

For solutions containing multiple chemical species, pH values were obtained by combining the predicted pH values for each individual chemical at the appropriate concentration using the formula:

$$\widehat{pH}_A = -\log_{10}\left(\frac{\sum_{i=1}^{n} 10^{-\widehat{pH}_{S_i}}}{n}\right) \quad \textbf{21}$$

where $\widehat{pH}_A$ is the predicted pH for the solution containing all elements, *n* is the number of chemical species in the solution and $\widehat{pH_{S_i}}$ is the predicted pH the individual species, S_i. The formula effectively determines the pH value by averaging the number of hydrogen atoms for each chemical in the solution. The ten-fold increase in hydrogen ions per pH unit decrease shows that the pH of the solution is dominated by the most acidic species, which is modified slightly by more basic species. The model requires no weighting of the parameters as the concentration of individual chemicals has already been accounted for.

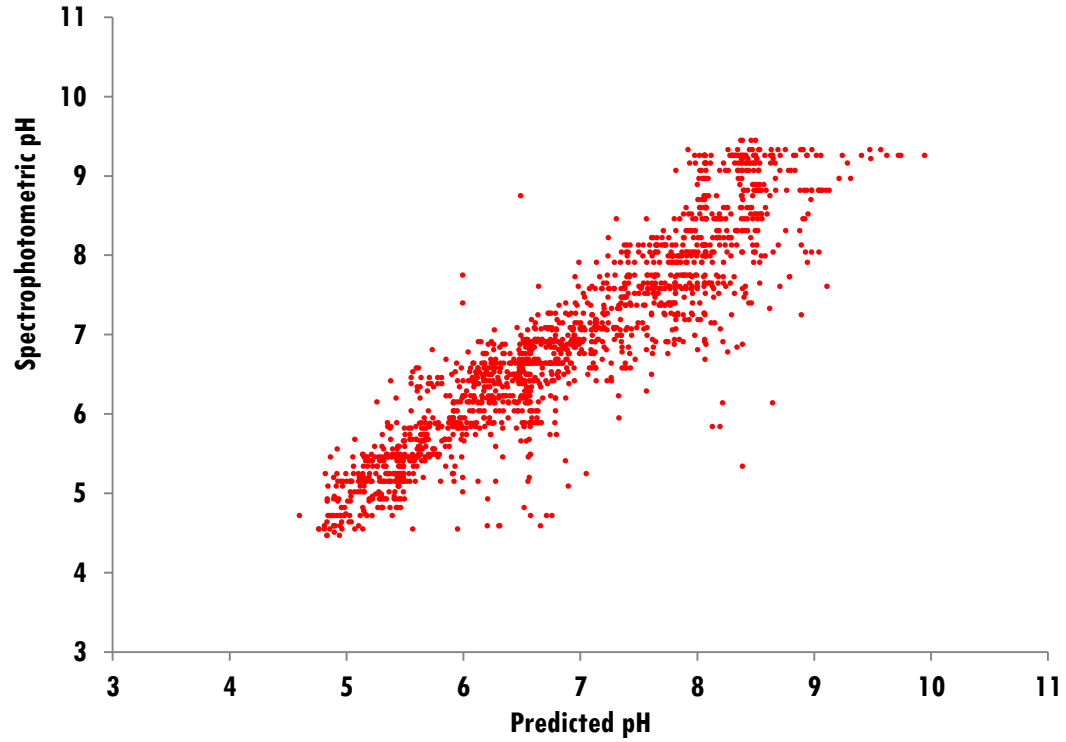
The mean squared error (MSE) between the spectrophotometric and predicted pH values is 0.28 in comparison to 0.8 between the values measured by spectrophotometry and the buffer pH values. The correlation with the measured values is 0.89 for the predicted pH in comparison to 0.77 for the buffer pH.

Linear regression showed that four chemical groups- ammonia, dihydrogen salts, hydroxide salts and salts of acids, require the full model including the interaction term relating both the buffer pH and the additional chemical concentration to the pH of the experiment. The model for PEGs does not include the chemical concentration as a separate term, but does include the interaction between chemical concentration and buffer pH. Organics and salts have the simplest models, only involving the buffer pH as a variable.

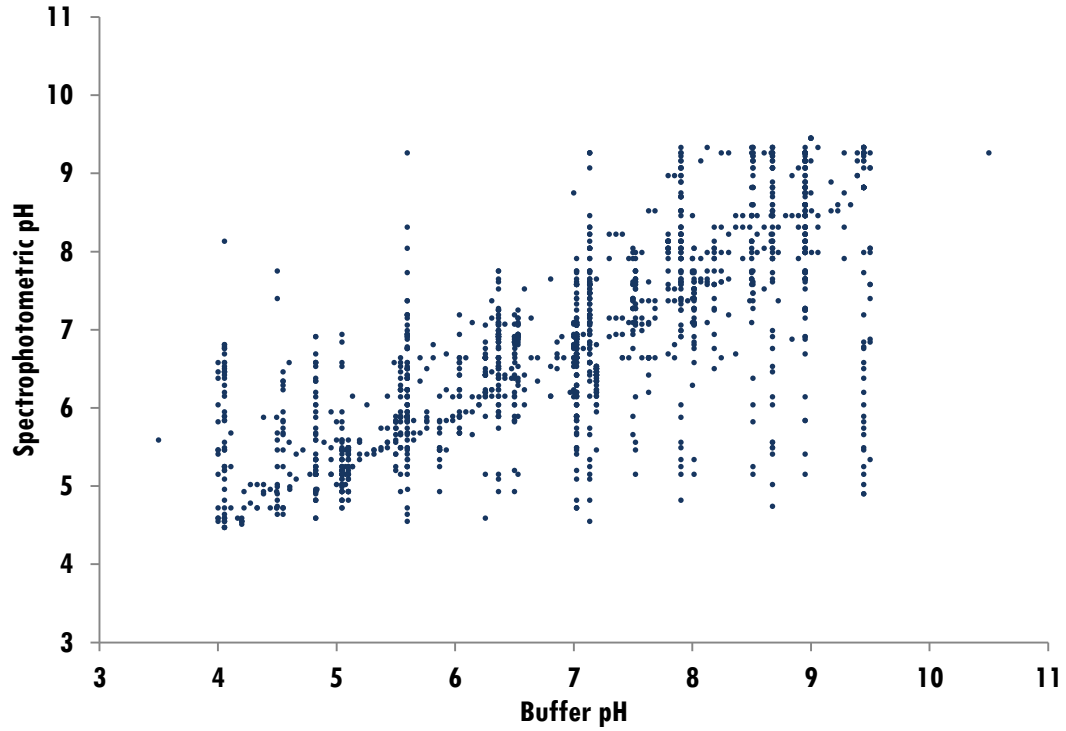
5.1.1. Modelling pH Using Machine Learning

An artificial neural network (ANN) implemented in Matlab (MathWorks, 2011) was trained to assign a pH value to crystallisation solutions. An ANN was employed as they are able quickly to create a richer, non-linear model than that of regression. Approximately two thirds of the data for the 5,161 conditions for which pH values could be determined by spectrophotometry were used to train a single hidden layer network using the Levenberg-Marquardt back-propagation method (Beale & Jackson, 1990, MathWorks, 2013). The other third was reserved as an independent test set. Chemicals were broadly grouped as suggested by the linear regression analysis (Table 5) and stratified sampling used to divide the chemical groups evenly between the training and test sets (3524: 1637). The concentration of chemicals in each group was calculated for each condition and these values, together with the buffer pH, used as inputs to the neural network. We chose a network with a single hidden layer of 5 nodes as this was the simplest network that gave a low mean squared error between the output pH and the spectrophotometric pH during training without over fitting (as assessed by the independent test set).

Figure 31a shows the pH values measured by spectrophotometry plotted against those predicted by the neural network for the independent test set. The linear relationship between measured and predicted pH can be shown to have an intercept close to zero and a gradient close to one suggesting a strong relationship between the two methods of obtaining pH. For the same test data the spread of values obtained by spectrophotometry for any particular buffer pH is much greater than for the corresponding predicted pH, as can be seen in Figure 31b. The correlation of the spectrophotometric pH with the predicted pH is 0.92 (MSE 0.25) in comparison to 0.75 with the buffer pH (MSE 0.97).



(a) Measured v predicted pH values.



(b) Measured v assumed buffer pH.

Figure 31: Accuracy of different pH values.

(a) shows pH predicted using a neural network in relation to the spectrophotometric pH and (b) shows the buffer pH in relation to the spectrophotometric pH.

Figure 30 shows the distribution of differences between the pH values obtained by spectrophotometry and those assigned by the neural network as well as those provided by the buffer pH. The histograms show the absolute deviations in 0.1 pH unit bins. Whilst 75% of predicted pH values are within 0.5 units of the measured pH (i.e. ±0.5 pH units) and 95% are within one unit, only 53% of the buffer pH values are within 0.5 units and just 80% are within one unit.

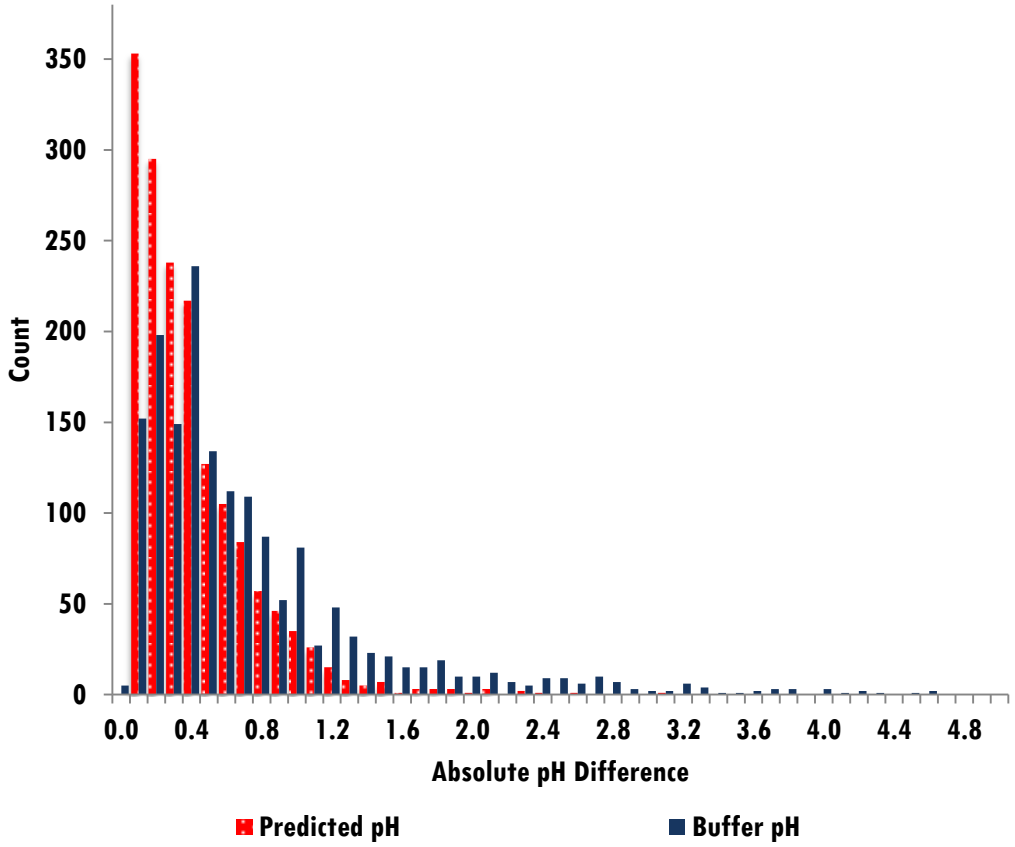


Figure 32: Histogram of errors for different methods of estimating pH.

Histogram showing errors in predicted (dotted bar) and buffer pH (solid bar) values in relation to the spectrophotometric pH.

Closer inspection of the predicted values reveals that six of the 66 individual chemicals were involved in the conditions where the deviation from the spectrophotometric pH values was unusually high. One of these chemicals, PEG 2000 DME, should be neutral but spectrophotometry suggested a pH of just over 4.5, at the limit of the method's reliability. It is known, however, that PEGs degrade becoming more acidic over time (Hampton, 2012, Jurnak, 1986, Ray Jr & Puvathingal, 1985). Indeed, when checked with a Jenway 4330 pH meter, the

solution was found to have a pH of just 2.6. The other five chemicals that were associated with large errors (jeffamine ed-2003, ammonium phosphate dibasic, dl-malic acid, sodium malonate, magnesium chloride) were not well-represented in the training data. Re-training the network with a larger dataset could therefore improve the results further.

Buffer \ Protein	pH 5	pH 7	pH 9
pH 5		6.3	7.1
		6.3	7.1
		6.2	7.1
pH 7	5.8		8.0
	5.8		8.0
	5.9		8.0
pH 9	6.5	7.8	
	6.4	7.8	
	6.4	7.8	

(a) Ratio of buffer to protein 1: 1.

Buffer \ Protein	pH 5	pH 7	pH 9
pH 5		6.5	7.6
		6.53	7.6
		6.53	7.6
pH 7	5.4		8.4
	5.5		8.4
	5.5		8.4
pH 9	5.8	7.5	
	5.9	7.5	
	5.9	7.5	

(b) Ratio of buffer to protein 1: 2.

Table 6: The pH within the crystallisation drop.

The table shows the measured pH of the components of the crystallisation drop for two different ratios of crystallisation cocktail: protein solution. The buffer used was 50mM PCTP, the protein solution consisted of 40mg/ml lysozyme, 100mM sodium chloride and 50mM PCTP. The experiments were repeated three times and the pH values measured are given in the tables.

Here we have used models to predict the pH of the crystallisation solution although only a proportion of this is contained within the drop containing the protein. Using a typical lysozyme solution at 40 mg/ml with 100mM sodium chloride and the buffer PCTP at 50mM, we have shown that, when mixed with 50mM PCTP at pH 5, 7 and 9, the final pH could be predicted from the two buffering components with neither the salt nor the lysozyme having a noticeable effect. For example, protein solution at pH 5 to crystallisation solution pH 7 in the ratio 2:1 gives a predicted pH of 5.66, which compares to an average measured pH of 5.46 (Table 6). Only when the ratio of

protein solution to: crystallisation solution was increased to 3:1 did we find that the lysozyme affected the pH.

5.2. Isoelectric Point

It is possible that the use of protein sequence information could assist in crystallisation. In 1992, Samudzi and co-workers attempted to answer the question 'Under which conditions will my protein crystallise?' They studied the BMCD for crystallisation trends, which at the time contained 820 macromolecules. Their motive was to move from an experimenter's own experience, which could be based on anecdotal evidence to a more scientific approach. Cluster analysis using several properties relating to each experiment (such as temperature, precipitant concentration and crystallisation method) allowed eight clusters to be determined, each with its own features. Using the characteristics of these clusters Samudzi *et al.* were able to provide a recommended strategy for crystallising new proteins. For example, they recommend that a protein of relatively high molecular weight should be tried in a screen with properties of a typical of cluster 2, 3 or 5. Cluster 2 is dominated by entries containing alcohols and high concentration of protein solution, cluster 3 contains entries that include the use of PEGs and low concentration of protein solution and cluster 5 contains entries where the precipitant is ammonium sulfate. They also recommend that proteins of low molecular weight should be crystallised in conditions similar to those of cluster 2 and 5, suggesting that molecular weight might not be useful in determining how to crystallise a new protein. A similar analysis by Farr Jr *et al.* (1998) on 1,500 macromolecules showed less of an overlap in the clusters and recommended strategies, but with both low and high weights crystallising broadly in the same temperature and pH conditions and with several shared chemical species. Despite the fuzzy clustering, the conclusion that protein properties can be helpful in determining the conditions under which proteins crystallise has since been supported by further studies. Using data from the BMCD, Hennessy *et al.* (2000) were able to provide software that would design a chemical screen for use on a specific class of macromolecule. Entering the class of macromolecule (enzyme, virus etc.) along with a choice of buffer, temperature, precipitating agent and other additives the program uses a Bayesian approach to calculate the combinations of parameters that have been most successful for

crystallising similar macromolecules. The most successful combinations of parameters are output as a design for a crystallisation screen. Hennessy and co-workers concluded that there are correlations between families of proteins and their crystallisation conditions. For example, ligand-binding proteins and enzymes have a significantly different distribution of pH values under which they crystallise, whilst immunoglobulin-like proteins and enzymes have a significantly different distribution of temperatures. It should be noted that these classes are not based on any structural classification system such as the Structural Classification of Proteins (SCOP) (Hadley & Jones, 1999) and as such, they may have unknowingly introduced their own bias. In the most recent study of the BMCD, using 12,765 proteins, Lu *et al.* (2012) were also concluded that different families of proteins have their own particular crystallisation conditions.

It has been postulated for some time that the best pH at which to initialise crystallisation experiments is one that matches the isoelectric point (pI) of the protein (McPherson, 1982). The pI of a protein is the pH at which its overall net charge is 0 and it determines a protein's minimum solubility level due to protein-protein interactions being favoured over protein-water interactions (Gilliland, 1988, Luft *et al.*, 2011). It should therefore follow that a solution with a pH matching the isoelectric point would be ideal for crystallisation, although this has never been confirmed. One possible reason for this is that the recorded pH is that of the buffer in the crystallisation solution rather than the final pH of the crystallisation cocktail (Zhang *et al.*, 2013).

An analysis of 9,596 structures obtained from the PDB suggested a link between a protein's pI and the pH at which it would crystallise. It was found that acidic proteins tended to crystallise 0 to 2.5 pH units above their pI, whereas basic proteins crystallised 0.5 to 3 pH units below their pI (Kantardjieff & Rupp, 2004). The authors reported a correlation between pI and pH-pI, that was challenged with claims that the predictive statements had been made using a misinterpretation of the data (Huber & Kobe, 2004). As a form of data normalisation, there will always be a link between pI and pH-pI, but it was also highlighted that no correlation between pI and pH had been found previously (Page *et al.*, 2003, Wooh *et al.*, 2003). In defence of their work the authors of the original study showed a correlation between the pI of

acidic proteins and the pH of successful crystallisation and that a linear model could be used to predict the optimal pH for such proteins. They concluded, however, that a similar model could not be created for basic proteins because no significant correlation was found (Kantardjieff *et al.*, 2004). Since the original study, similar relationships between the pI of proteins and the buffer pH of successful crystallisation experiments have been noted for both acidic and basic proteins (Charles *et al.*, 2006).

Proteins can become more positively or negatively charged by gaining or losing protons due to the pH of their environment. The isoelectric point (pI), the pH at which a protein has a net charge of zero can be calculated using the charges for the specific amino acids in the protein sequence. Estimated values for the charges are called acid dissociation constants or pK_a values. In the following analysis the pK_a values used are those used in the EMBOSS software suite (Rice *et al.*, 2000) as shown in Table 7.

Amino Acid	pK_a	Charge
Amine Group	8.6	Positive
Carboxyl Group	3.6	Negative
Cysteine (C)	8.5	Negative
Aspartic Acid (D)	3.9	Negative
Glutamic Acid (E)	4.1	Negative
Histidine (H)	6.5	Positive
Lysine (K)	10.8	Positive
Arginine (R)	12.5	Positive
Tyrosine (Y)	10.1	Negative

Table 7: EMBOSS acid dissociation constants.

For a protein with $n-$ negatively charged amino acids and $n+$ positively charged amino acids, the pI can be determined as the pH for which the net charge given by equation 22 is zero.

$$net\ charge = \sum_{i=1}^{n_-} \frac{-1}{1 + 10^{pK_n - pH}} + \sum_{i=1}^{n_+} \frac{1}{1 + 10^{pH - pK_p}} \qquad \textbf{22}$$

where pK_n and pK_p are the pK values for negatively charged and positively charged amino acids respectively. As an example, consider the small amino acid sequence, CRV, with one cysteine (pK_n = 8.5), one arginine (pK_p = 12.5) and one valine (no charge). Including the N-terminal amine group (pK_p = 8.6) and C-terminal carboxyl group (pK_n = 3.6) the net charge for an initial pH of 0 is given by equation 23:

$$\begin{aligned} net\ charge = & \frac{-1}{1 + 10^{3.6-0}} + \frac{-1}{1 + 10^{8.5-0}} + \frac{1}{1 + 10^{0-12.5}} \\ & + \frac{1}{1 + 10^{0-8.6}} \\ & \approx 0 + 0 + 1 + 1 \approx 2 \end{aligned} \qquad \textbf{23}$$

The charge for the sequence CRV, therefore, is approximately 23 at pH 0. By gradually increasing the theoretical pH, to make the net charge in equation 22 equal to zero, it is found that the isoelectric point of CRV is 8.555, as shown in Figure 33.

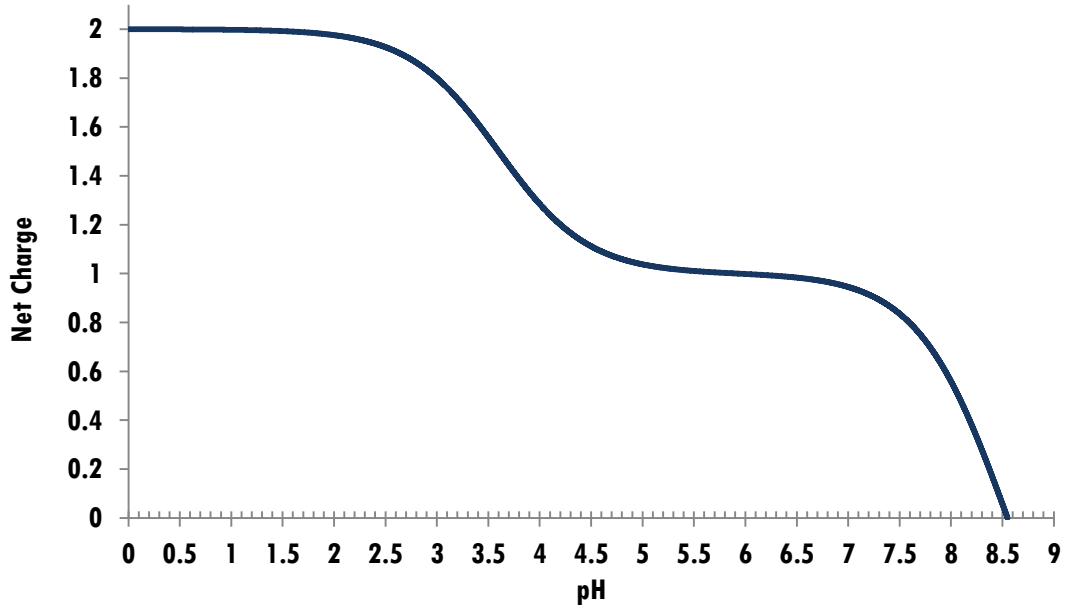


Figure 33: Net charge of the sequence CRV with varying pH.

In the calculation of a protein's pI from sequence, an assumption is made that all residues have the potential to affect its overall charge. This assumption does not account for partially buried residues that may not have their pK_a modified by the

environment. Fortunately, the approximation of pI accounting for buried residues (referred to as 3D isoelectric point) is strongly correlated to that of the linear pI. Using the PROPKA server (Rostkowski *et al.*, 2011) it has been possible to calculate both forms of pI 21,045 different proteins obtained from the PDB in December 2013. Figure 34 shows a scatter plot of the isoelectric points calculated using the two methods. The correlation between the two groups is 0.9 with 85% of proteins being within one pH unit of each other and 94% being within two units.

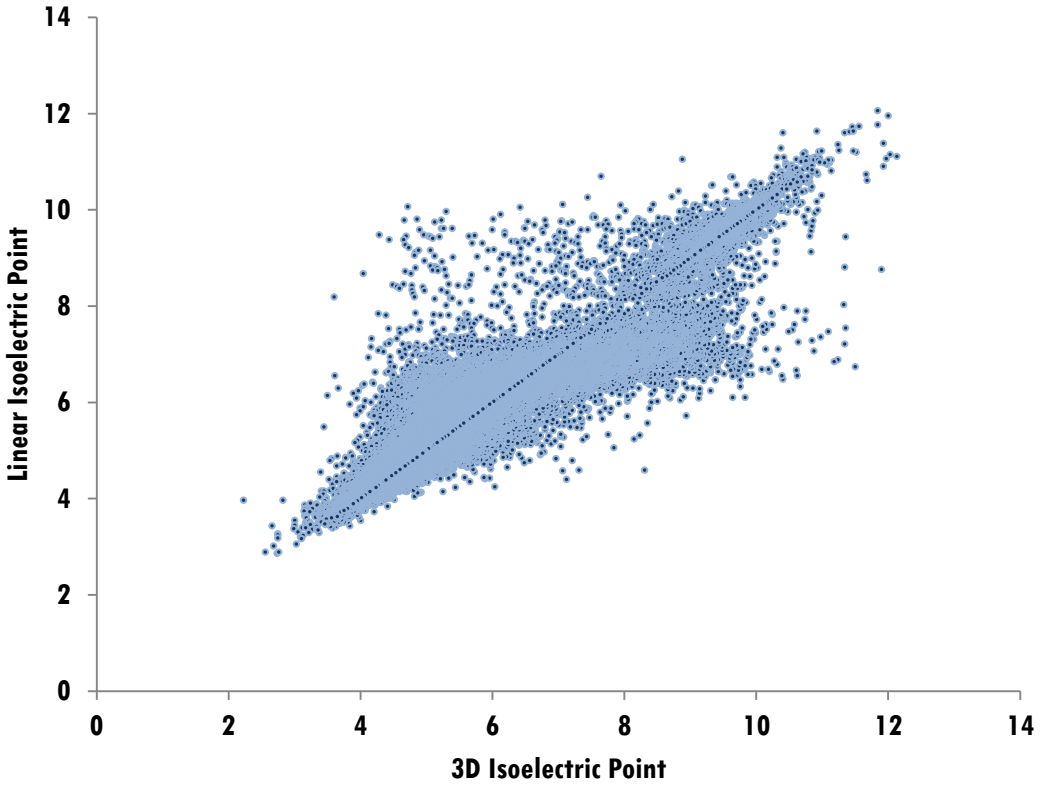


Figure 34: Primary v tertiary isoelectric point.

The isoelectric point for 21,045 proteins obtained from the PDB is calculated using linear sequence (y) and from assumed 3D structure (x).

The PROPKA server calculates the linear and 3D isoelectric point of protein sequences using a custom set of pK_a values. There is no global agreement on such values and several sets exist. However, we found a very strong correlation (~0.99) between six sets tested, suggesting they can be used interchangeably.

5.3. Relationship between pI and pH

The investigation into the link between a protein's isoelectric point at the point at which it crystallises was explored using three datasets; the first was a custom experiment performed at AstraZeneca, the second used data from crystallisation experiments at the SGC and the third utilised data from the PDB.

For the custom dataset, the pI was either obtained from Zhang *et al.* (2013) or calculated as above and was confirmed using isoelectric focusing. The isoelectric point for each sequence in the SGC and the PDB datasets was determined in the same manner using an Excel spreadsheet with visual basic for applications (Microsoft VBA).

5.3.1. Custom Crystallisation Experiment

In order to determine the conditions for crystallisation, fourteen proteins (described in Chapter 2) were initially screened using sitting-drop vapour diffusion with a 96-condition sparse matrix screen buffered at 6 different pHs using the multi-component buffer PCTP (Newman, 2004, Zhang *et al.*, 2013). This gave a total of 576 conditions with the buffer pH fixed between pH 4.5 and pH 9.5. The best crystallisation conditions were selected for each protein and a finer sampling of pH was then performed with these conditions in a 96-well plate buffered between pH 4.5 and pH 9.5 with PCTP as shown in Figure 35.

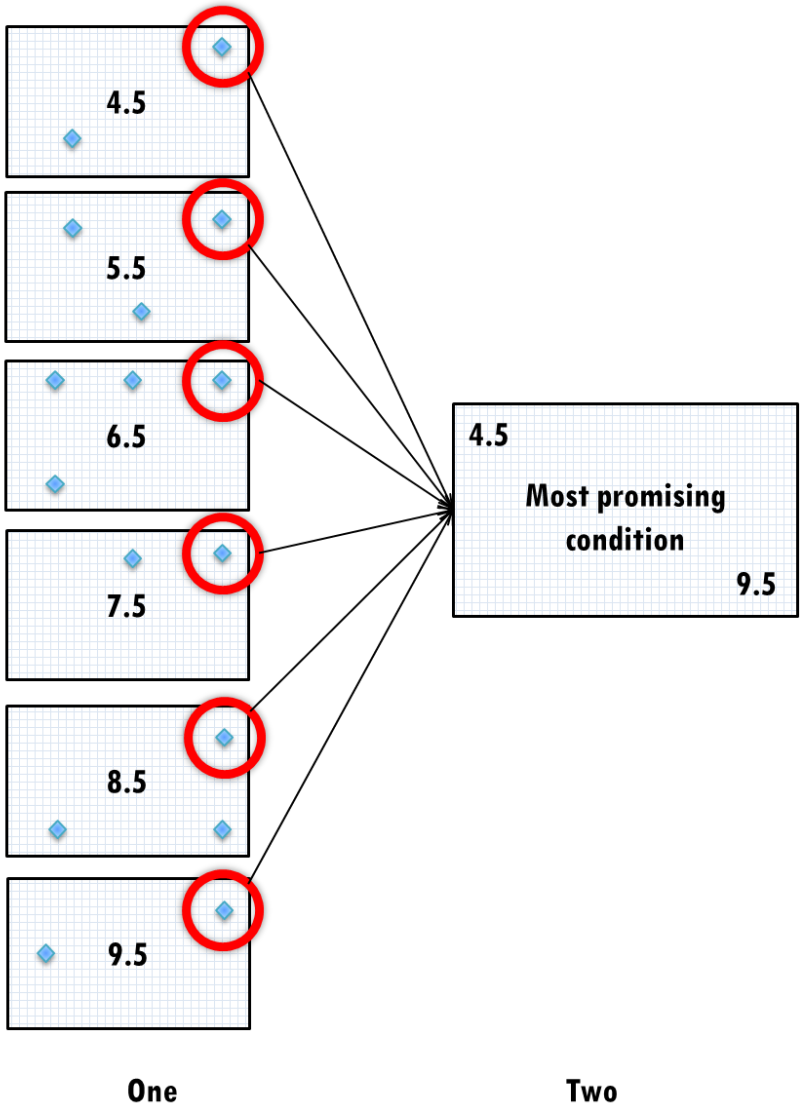


Figure 35: Optimisation of conditions in the custom experiment.

Fourteen proteins were each screened in 6 sparse matrix screens, each screen with a different buffer pH. The condition that consistently crystallised each protein across the range of pH was then used as the single condition for another screen. This screen had an incremental increase of pH from 4.5 to 9.5, which was then measured using a spectrophotometer together with the acid-base indicator bromothymol blue.

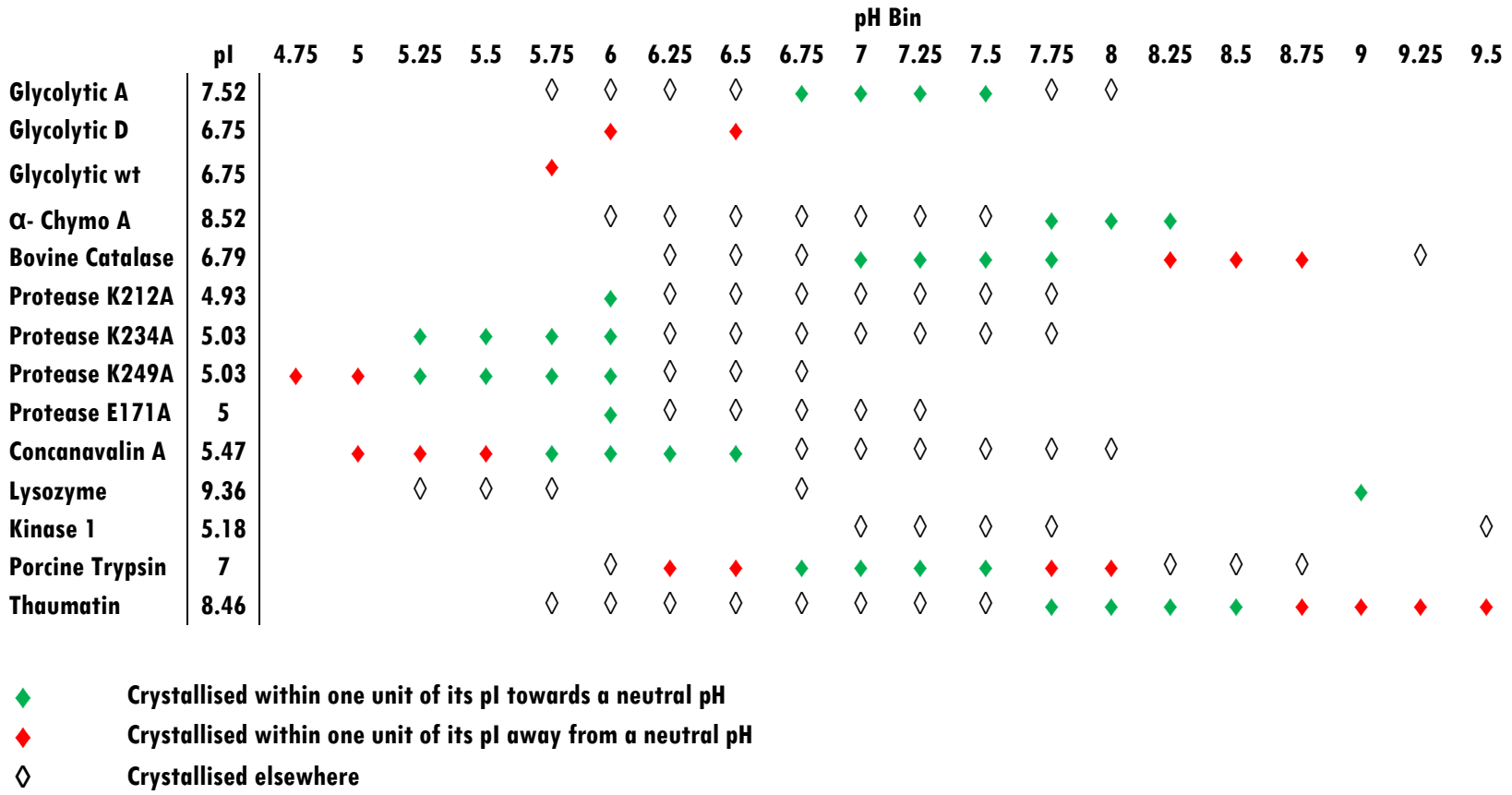


Figure 36: Distribution of crystals in the custom experiment.

The 0.25 pH bin in which each of the fourteen proteins crystallised is indicated by a diamond. Crystals obtained within one unit of the pI towards a neutral pH are shown in green. Similarly, crystals within one unit of the pI but away from a neutral pH are shown in red.

The fourteen proteins in the custom dataset were used to further test the relationship between pI and the pH of successful crystallisation. Once the best crystallisation components had been determined for a particular protein, a fine sampling of pH was performed in a 96-well plate with the chosen components buffered between pH 4.5 and pH 9.5. Figure 36 shows crystals were obtained within one pH unit towards neutral from their pI for 11 of the 14 proteins and 13 out of 14 crystallise within one pH unit either side of their pI. The glycolytic enzymes D and wt crystallised within one unit of their pI but away from a neutral pH. Only one protein, Kinase 1, with a pI of 5.18, did not crystallise within two pH units of its pI. The stochastic nature of protein crystallisation compounds the difficulties of pattern recognition. Figure 36 shows that, whilst several proteins crystallise across a wide range of pH values, crystals are not seen in every 0.25 bin within that range. Reproducibility in screening has been investigated and the results suggest that replication could improve success rates in crystallisation experiments (Newman, et al., 2007).

5.3.2. Structural Genomics Data

The second dataset, obtained from the SGC (described in Chapter 2), comprised of the experimental conditions for 1,039 different protein sequences. Experimental results were assessed using the score given by a crystallographer, together with the resolution of the diffraction data and whether or not the structure was solved. For crystals that were not of diffraction-quality no estimated resolution is given and it was assumed that the structure was not determined. In instances where crystals were identified as salt, the associated data were removed.

The remaining data were grouped according to the final stage reached in the structure determination pipeline as follows:

Group 1	58 sequences that resulted in structure determination;
Group 2	48 sequences that resulted in a crystal that diffracted to at least 3.6Å;
Group 3	210 sequences that result in a least one protein crystal;
Group 4	723 sequences that were annotated as ‘crystal - to be followed up’.

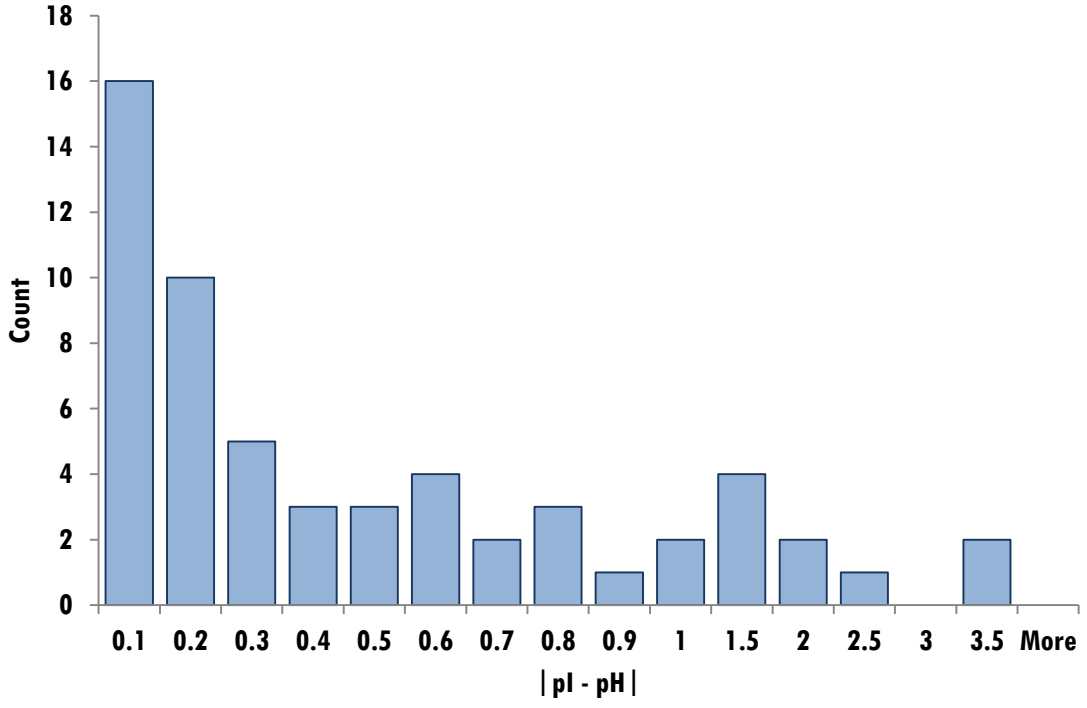
It should be noted that sequences in the final group may not relate to diffraction quality crystals or could be salt crystals that had not yet been identified as such. Conversely, it is possible that perfectly good crystals may have been overlooked.

These data, all screened using the SGC JCSG +4 sparse matrix screen, were selected from the full SGC database and assigned to chemical groups in order to predict pH. Although a spectrophotometric pH value was available for some of the conditions in the SGC JCSG +4 screen (used either for training the neural network or reserved to test the accuracy of the assignments), the pH used here for all conditions was that assigned using the trained neural network. In addition to the chemical concentrations, the pH of the crystallisation buffer is also input to the network. For those wells without a buffer solution (21/96), the pH of the purification buffer was used instead. Data for any wells where neither buffer pH nor purification pH were available were removed. In Chapter 4 we showed that the buffering capacity of the protein itself is negligible in vitro and this has also be demonstrated in vivo (Poznanski *et al.*, 2013).

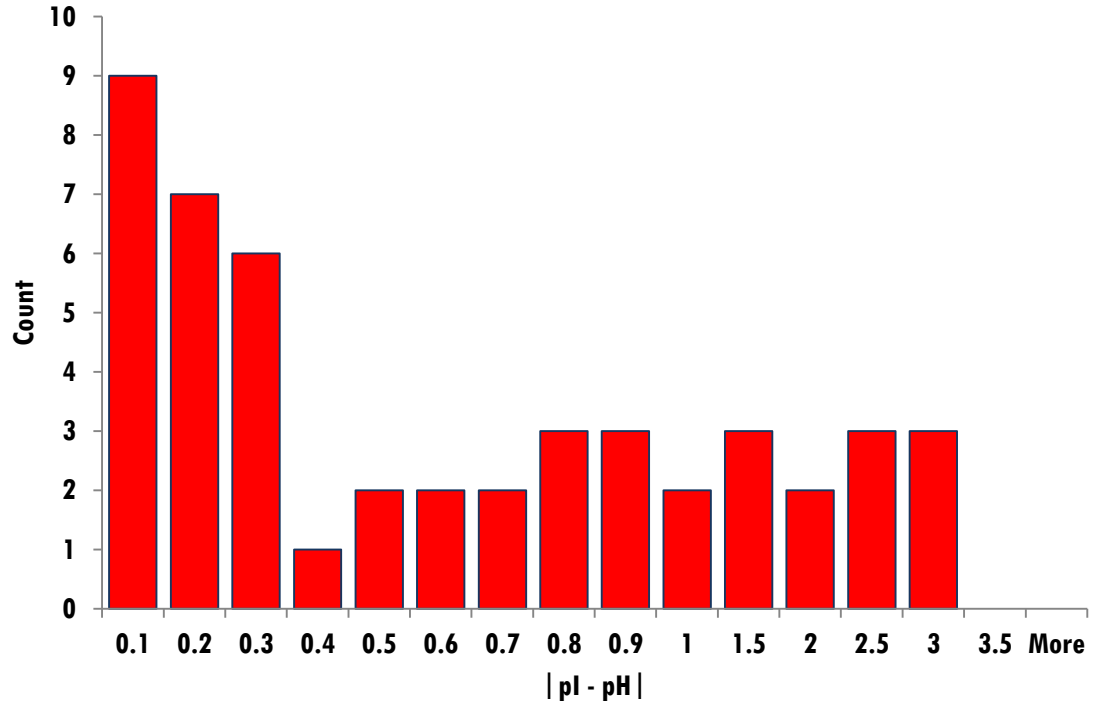
For each well in which a crystal was observed, the calculated pI was then compared to the assigned pH. The protein sequences were considered in groups, as defined in section previously, reflecting the maximum stage in the structure determination pipeline that was attained. The protein structure was determined and deposited in the PDB for the 58 protein sequences in Group 1. In addition to the conditions that led to the final structure, we also have information about other conditions that produced crystals. Analysis shows that crystals are only obtained in conditions with a pH within one unit of the pI for 9 of the 58 sequences. A total of 28 sequences only result in crystals within two pH units of the pI, 45 sequences only result in crystals within three pH units, 57 sequences only result in crystals within four pH units and the final 4 proteins crystallise up to five pH units away from the pI. Thus, for over 70% of these protein sequences, crystals are only obtained in experiments buffered within 3 pH units of the pI.

Particularly in cases when available protein is limited, it is important to identify suitable conditions in as few trials as possible and restricting screening to a particular pH range would reduce the number required. Promising initial conditions (including the pH), could then be optimised to obtain crystals suitable for crystallographic

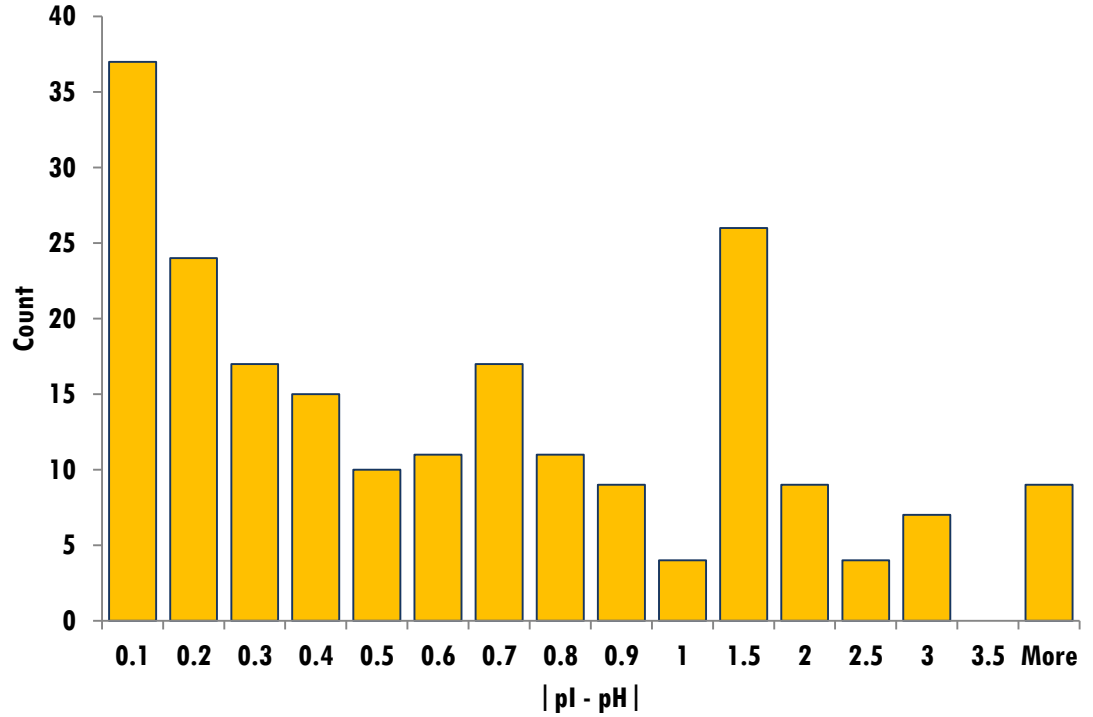
studies (Jancarik & Kim, 1991). For the 58 proteins in the SGC dataset that resulted in a structure deposited in the PDB, we found a correlation of 0.8 between the pH of any crystalline result and the pH at which the final structure was obtained. We therefore investigated the differences between a protein's isoelectric point and the closest pH value for any conditions producing crystals. Again the proteins were considered in the four groups according to the stage reached in the crystallisation pipeline. For those proteins in group 1, 84% crystallised within one pH unit of their pI and 95% crystallised within two pH units of their pI. Crystals were found within one pH unit of their pI for 78% of proteins in group 2 and within two pH units for 88%. In group 3, 74% of proteins crystallised within one pH unit of their pI and 90% within two pH units and for group 4 proteins, 55% produced crystals within one pH unit of their pI and 82% within two pH units. Overall, 85% of proteins produced crystals within two pH units of their pI. Histograms showing the distribution of the absolute difference between the pI and the closest pH at which crystals were obtained for each group are given in Figure 37. It is worth noting that those proteins for which no crystals were found within three pH units of their pI (6% of all protein sequences here) tended to have more extreme isoelectric points. Of the 64 such proteins, 46 had a pI outside the range 5 to 9 and of the 18 protein sequences with a pI in this range, only one with a pI of 7.9 is within the range 6 to 8.



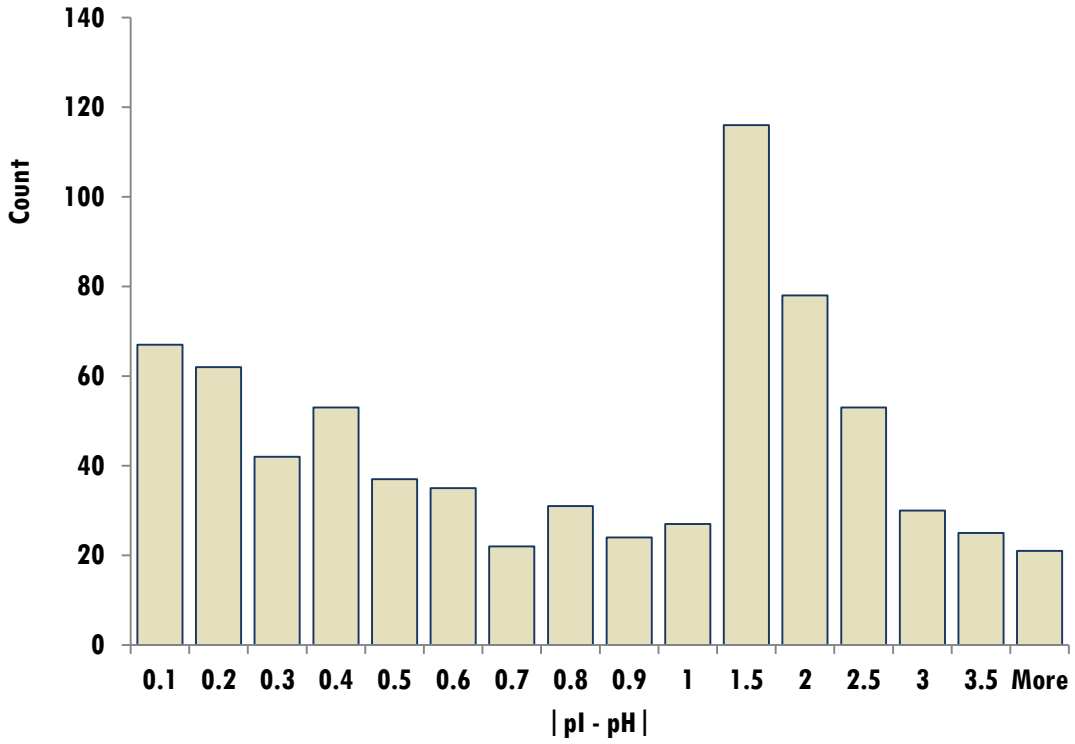
(a) Group 1- structure determined.



(b) Group 2- diffraction to at least 3.6Å.



(c) Group 3- at least one protein crystal.



(d) Group 4- crystal to be followed up.

Figure 37: Distribution of differences between pI and pH.

Histograms showing the absolute difference between the pI and the closest pH at which crystals were obtained for proteins in the SGC dataset.

5.3.3. Protein Data Bank Snapshot

Using a snapshot of the PDB with standardised crystallisation conditions courtesy of Fazio *et al.* (2014) we were able to calculate for 23,949 proteins their pI and the pH of the crystallisation solution in which is crystallised. A full overview of the structure of the data obtained from the PDB is described in Chapter 2.

Figure 38 shows a pattern in the relationship between the pI of proteins and the pH at which they have been crystallised. Acidic proteins, i.e. those with a pI below 7, tend to crystallise about one pH unit above their pI, whereas basic proteins tend to crystallise below their pI by around 1.5-3 pH. These results support those found in our custom experiment and those derived from the SGC data along with those of other studies (Kantardjieff *et al.*, 2004, Kantardjieff & Rupp, 2004, Charles *et al.*, 2006).

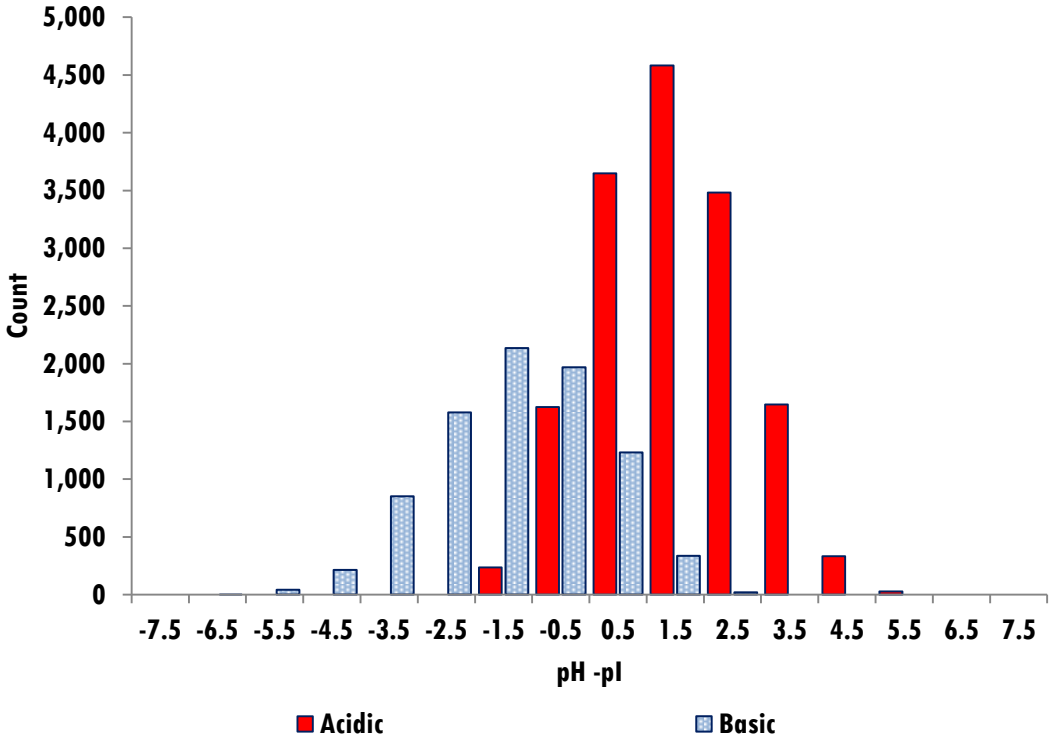


Figure 38: The relationship between pH and pI for PDB proteins.

The distribution of the difference between the pH from which a structure was obtained and the isoelectric point for 23,949 proteins in the PDB-UNIQUE dataset. The distributions are shown separately for proteins with a pI less than 7 (acidic) and those with a pI greater than 7 (basic). Those with a pI of precisely 7 (of which there were 4) were grouped with the basic proteins.

5.4. Discussion and Conclusions

Linear regression modelling revealed groups of chemicals with similar effects on the pH of a crystallisation experiments. The simplest models were obtained for salts with no hydrogen ions and neutral organic compounds. Although a simple linear regression model can be used to relate the pH of the experiment to the buffer pH for both of these chemical groups, the model is different for each group, with the constant offset larger for organics than that for salts. For other groups the effect of the additional chemical on the buffer pH depends on the concentration of that chemical. In the case of PEGs, the chemical concentration does not appear as a separate variable, but the interaction term between buffer pH and the chemical concentration is significant. It is known and we have shown (Appendix C) that PEGs degrade over time (Hampton, 2012, Jurnak, 1986, Ray Jr & Puvathingal, 1985),

increasing the acidity of the solution. Similarly, ammonia-containing compounds slowly release the ammonia and affect the pH of a condition (Mikol *et al.*, 1989, Newman, Sayle, *et al.*, 2012). Ammonia containing compounds are more acidic than PEGs, which when fresh and correctly stored are close to neutral pH, and like the final two groups (acids and basic) require the full linear regression model including the interaction term to represent the pH of the experiment. The last two groups either contain hydrogen ions that have a large impact on pH or contain a hydroxide group, with a large but opposite effect on pH. The largest errors in prediction are due to chemicals that undergo degradation. The deterioration of chemicals, such as PEGs, cannot be predicted but should be considered and storage conditions such as light exposure and temperature could perhaps be controlled.

The grouping of chemicals according to their effect on the pH of a solution means that individual models are not required for each chemical and the effect of chemicals for which there are no examples in the training set can be predicted from the model for the appropriate group. Moreover, the increase in the number of examples available for each model reduces the possibility of over-fitting of the training data and provides more robust models for prediction.

Using the chemical grouping suggested by linear regression modelling, the most accurate results were obtained using a single-layer neural network with five nodes but the method is less intuitive, and similar results were obtained using the regression equations.

The ability to predict the effect of different combinations of chemicals on the pH of an experiment allows information in databases such as the PDB to be used in data mining studies that aim to reduce the number of crystallisation trials required. Over the last decade a number of investigations have considered a possible link between the pI of a protein and the pH at which it will crystallise (Charles *et al.*, 2006). Such a link has also been disputed, with Zhang *et al.* (2013) suggesting that "the pI value of a protein should be avoided when choosing the pH for a protein solution". Zhang and co-workers also discuss the issue of the recorded pH not necessarily being the pH of the experimental conditions. Previous findings have been based on the pH of the buffer solution, which can differ from the actual pH by more than 3 pH units

(Bukrinsky & Poulsen, 2001, Newman, Sayle, *et al.*, 2012). Using more accurate pH values that take into account how the concentrations of the various chemicals in the crystallisation cocktail affect the pH of the buffer solution, we have shown that a relationship between a protein's pI and the pH under which it will crystallise does exist. In addition to data for the conditions leading to protein structure solution we have considered the pH of experiments producing crystals that may not have been confirmed as diffraction quality. We found that proteins frequently crystallise within one pH unit of their pI and that 85% of the proteins produced crystals within two pH units of their pI. In most cases, proteins tended to crystallise at a more neutral pH with acidic proteins crystallising above their pI and basic proteins below their pI, confirming those results found previously (Charles *et al.*, 2006, Kantardjieff & Rupp, 2004). As the majority of proteins that crystallise are of an acidic pI (Figure 39), we therefore suggest that a useful initial pH for crystallisation trials can be obtained from the pI of the protein in question, but this pH should not simply be taken as that of the buffer solution but, if not measured, should be adjusted to take into account the effect of any additional chemicals.

It may also be possible to distinguish between whether the pH of the solution is imperative for crystal nucleation, crystal growth or both. This could be performed with two experiments, one where a crystal/seed is added to a solution of a desired pH and the growth, maintenance or degradation of the crystal is monitored and compared to the same set of conditions but without the seed. This strategy could then be used to improve microcrystals or use artificial seeds to grow protein crystals around.

Distribution of Isoelectric Points

Figure 39 shows a trimodal distribution for isoelectric points with modes of approximately 4.8, 6.6 and 9 and the majority of crystallisable proteins having an acidic isoelectric point. This distribution contradicts the findings of others who show, for smaller sample sizes, that the isoelectric points of proteins are bimodally distributed with one peak representing acidic pIs and another basic pIs (Canaves *et al.*, 2004, Kantardjieff & Rupp, 2004). Analysis of the BMCD shows that the most successful buffer pH is normally distributed around pH 7 (Samudzi *et al.*, 1992) which is supported by a later study of the PDB (Fazio *et al.*, 2014). These are reports of the buffer pH and not of predicted or measured pH and therefore it is difficult to

untangle which of these pH values are accurate and which proteins (and isoelectric point) they are associated with.

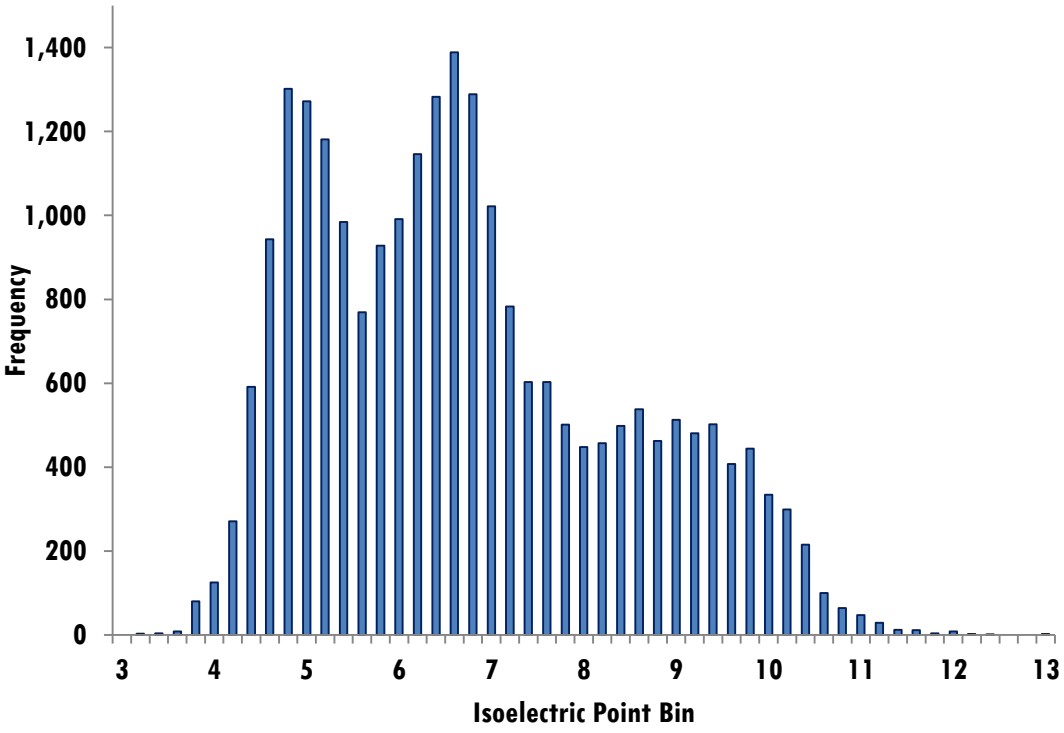


Figure 39: Histogram of isoelectric point for PDB proteins.

The distribution of pI for 23,949 significantly different proteins obtained from the PDB.

5.4.1. Prediction of Crystallisation Group

As some proteins do not crystallise close to their pI, we investigated protein properties to determine whether or not such proteins could be predicted. In addition to pI, the GRAVY (Kyte & Doolittle, 1982) and the number of D, C, G, H, M, F, P, S, T, W, Y residues (Overton *et al.*, 2008) were calculated for each sequence in the SGC dataset. A Euclidean distance matrix was created between each sequence based on the difference in their scaled features. This distance matrix was then used for multidimensional scaling k-means clustering, which is described in Chapter 3. PCA was also performed, a plot of which is shown in Figure 40, it can be seen that there is no discrimination between the groups.

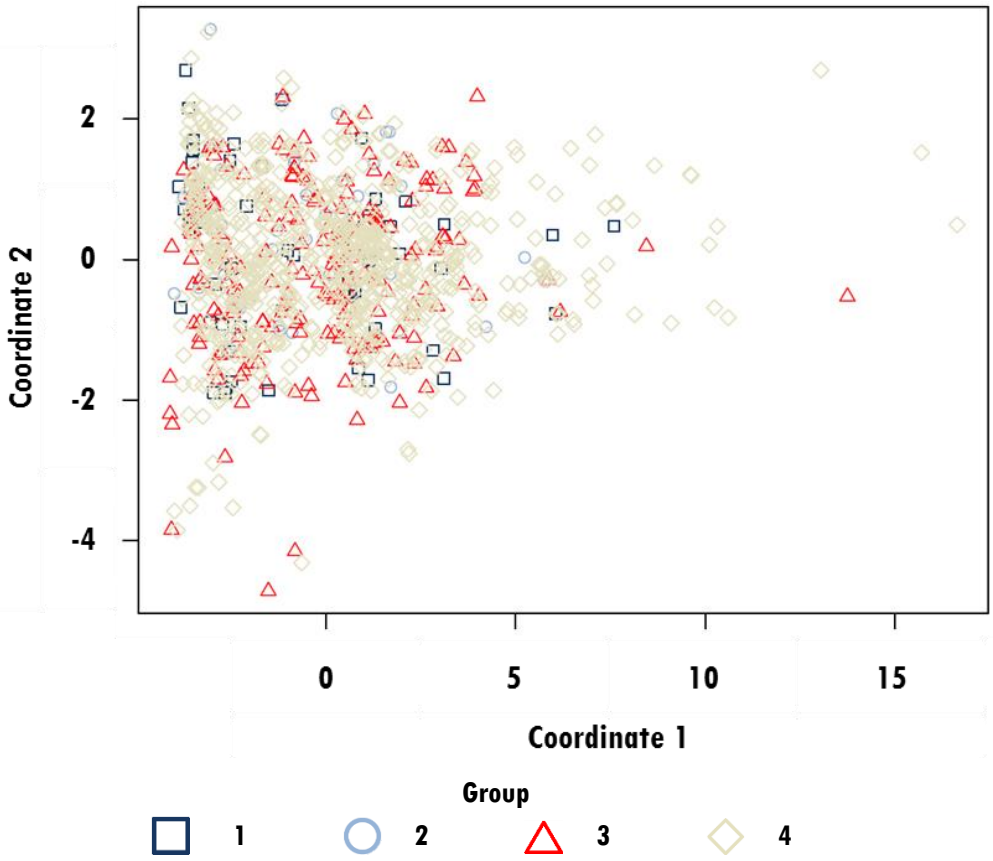


Figure 40: PCA of 1,039 sequences represented by 13 features.

The plot shows the four groups (1- structure determined, 2- diffracted to at least 3.6 Å, 3- protein crystal, 4- annotated by eye as crystal) of the SGC data plotted with respect to their first and second principal components.

MacQueen k-means clustering was implemented, with k=4 (MacQueen, 1967) and Figure 41 shows the confusion matrix obtained. Each group had around 25% of its targets classified correctly overall due to the differences in group's sizes the accuracy is 25%. Without clustering the overall accuracy increases to 30%. These results suggest that experimentation may be the only way of determining whether a sequence will result in a determined structure. A possible explanation for this is that the properties used do not give any indication of the complex intra- and intermolecular interactions. Differences in physical crystallisation conditions such as temperature, use of ligands and whether the protein sample was frozen were not taken into account.

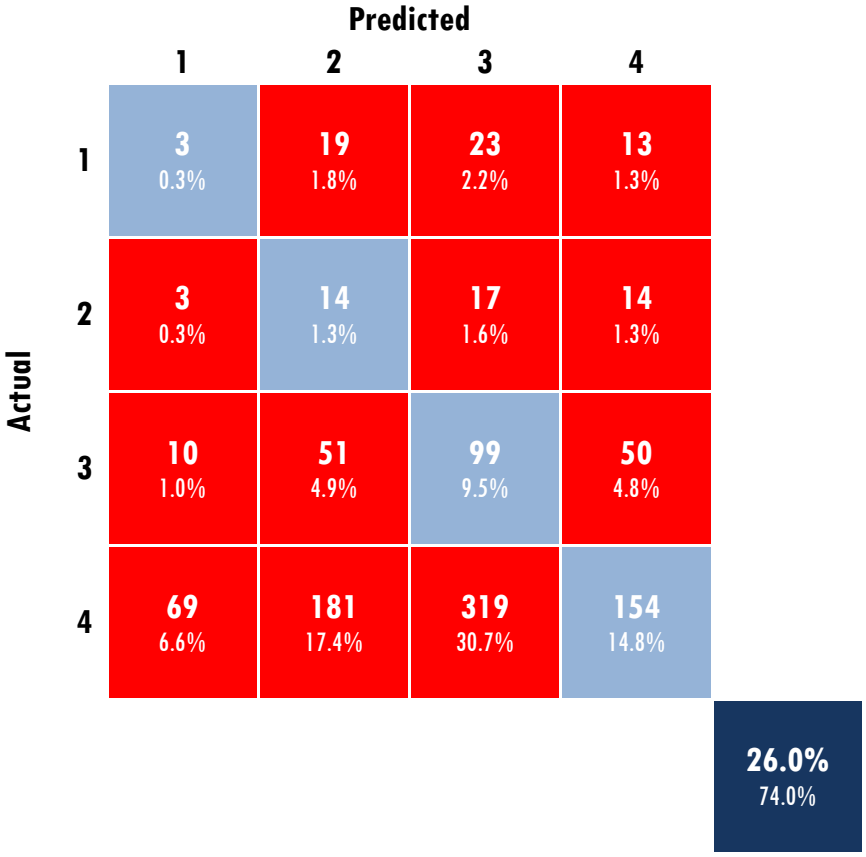


Figure 41: Confusion matrix for k-means clustering.

These are the results of k-means clustering for the SGC data. The overall accuracy of the clustering was just 19%.

.

6. Predicting a Protein's Propensity to Crystallise

Using parameters derived from the amino acid sequence, a number of researchers have tried to predict whether a protein is suitable for structure determination by X-ray crystallography. In 2006, a SEquence-based CRystallisability EvaluaTor (SECRET) created by Smialowski *et al.* (2006) was developed using two classes of protein sequences from the PDB. The first class, proteins solved by X-ray crystallography and the second, proteins solved by NMR. As proteins solved by NMR are small, the classifier is limited to sequences 40 to 200 amino acids in length. This ensures that the separation of the classes is not dependent on length. They calculated properties such as single amino acid frequency, dipeptide frequency and hydrophobicity of the amino acids (using 3 hydropathy scales (Rose *et al.*, 1985, Kyte & Doolittle, 1982, Engelman *et al.*, 1986)). After reducing the number of properties using wrapper feature selection (Kohavi & John, 1997), Smialowski *et al.* (2006) were able to accurately classify 62.7% of proteins in their sample with support vector machines. When Overton *et al.* (2008) assessed a new dataset with this classifier, they reported an accuracy level of 58.1%.

The work of the team who developed SECRET was challenged by Chen *et al.* (2007) who claimed that SECRET used many features and yet has relatively low prediction rates, as 50% accuracy should be achieved by the flip of a coin. They proposed a new classifier, named CRYSTALP, which uses features such as a count of all the individual amino acids in a sequence divided by the length of the sequence as well as the collocation of amino acid pairs. Their final predictor with 46 features is also limited by the size of the proteins that can be classified as to make it directly comparable to SECRET. When tested with the same dataset as SECRET an accuracy rate of 77.51% was reported (Chen *et al.*, 2007), although Overton *et al.* (2008) report that, on one of their restricted length datasets, CRYSTALP only achieved an accuracy level of 46.5%, a percentage that would be expected by a random guess. Jahandideh and Mahdavi (2012) reported the accuracy of CRYSTALP to be 68.40% and 75.69% in two separate trials. An improvement on the classifiers came with

CRYSTALP2, which had no upper length restriction (Kurgan *et al.*, 2009). New features include the use of collocated tripeptides, pI and Grand Average of Hydropathy (GRAVY). In total 1,103 features were used. Results show a classification accuracy of between 69.3% and 77.5% depending on the dataset (Kurgan *et al.*, 2009).

In 2006, the same year as the publication of SECRET, the OB-Score was published. The OB-Score "ranks potential targets by their predicted propensity to produce diffraction-quality crystals". A high OB-Score suggests that a protein is likely to be successfully crystallised; a low one suggests it is unlikely (Overton & Barton, 2006). The OB-Score was trained using the predicted isoelectric point and the GRAVY of the 5,545 amino acid sequences from the PDB with a diffraction quality of <3.0Å. The accuracy of the OB-score predictor has been reported as 69% by Kurgan *et al.* (2009) and 73% by Jahandideh and Mahdavi (2012). In a similar manner to the OB-Score, XtalPred provides a guide on how likely a protein is to crystallise, using protein properties derived from the sequence such as molecular weight and GRAVY. XtalPred was developed on the back of comments from as early as 1984 suggesting a "crystallisation feasibility score". From the 2007 publication it is unclear how the score is derived. Like SECRET, OB-Score and CRYSTALP2, XtalPred is freely available online and has accuracy levels of 76% and 72.40% that have been published by Kurgan *et al.* (2009) and Jahandideh and Mahdavi (2012) respectively.

There have since been several other classifiers and predictors all using properties derived from the protein sequence with statistical pattern recognition methods. ParCrys uses Parzen Window probability density estimators with a measure of randomness of the sequence (Wan & Wootton, 2000), pI and hydropathy values; an accuracy of 79.1% was reported (Overton *et al.*, 2008). RFCRYS uses the machine learning method of random forests to predict crystallisability. In their own tests they report 80.4% accuracy (Jahandideh & Mahdavi, 2012). The CRYSpred predictor, uses a set of sequence derived properties that are described in the Amino Acid Index Database (Kawashima *et al.*, 2008). They include several methods for calculating disorder, hydrophobicity, disorder and charge. In total they use 15 features, achieving an accuracy of 73.4% on a test set of 2,000 proteins (Mizianty & Kurgan, 2012).

Name	Year	Prediction Method	TEST Accuracy (%)	TEST-RL Accuracy (%)	
SECRET **RL**	2006	Support Vector Machine	-	58.1	[1]
The OB-Score	2006	Z-score Matrix	64.6	69.8	[1]
CRYSTALP **RL**	2007	Naïve Bayes	-	46.5	[3]
Xtalpred	2007	Logarithmic Opinion Poll Method	79.2	76.7	[1]
ParCrys	2008	Parzen Window	71.5	79.1	[1]
CRYSTALP2	2009	Radial Basis Function Network	75.7	69.8	[1]
Metappcp	2009	Logistic Model Tree	81.0	-	[2]
MCSG	2010	Support Vector Machine	-	-	[4]
Hyxg-1	2010	Regression Partitioning	-	-	[4]
Xannpred	2010	Neural Network	-	-	[4]
SVMCRYS	2010	Support Vector Machine	86.8	89.53	[3]
RFCRYS	2012	Random Forest	81.25	-	[2]
CRYSpred	2012	Support Vector Machine	79.9	80.2	[1]

-	test set was not tried with the named predictor
RL	the predictor was trained on sequences of restricted length

Table 8: The accuracy of different predictors.

Accuracy rates are shown for the two independent test data sets, TEST and TEST-RL.

[1] Figures for accuracy were obtained from the CRYSPred paper (Mizianty & Kurgan, 2012).

[2] Figures obtained from the RFCYRS paper (Jahandideh & Mahdavi, 2012).

[3] Figures obtained from the SVMCRYS paper (Kandaswamy *et al.*, 2010).

[4] These predictors were not evaluated on the named test sets.

This accuracy rate was surpassed by Kandaswamy *et al.* (2010) using a support vector machine, but this classifier had a large discrepancy between the accuracy rates on training and test data sets, which suggests that their results might be unreliable due to over fitting (Mizianty & Kurgan, 2012).

6.1. Datasets

Two particular datasets, TEST and TEST-RL, originally introduced by the developers of the ParCrys predictor (Overton *et al.*, 2008) have since been used by a number of authors to allow comparisons to be made. The TEST dataset contains 144 sequences obtained from TargetDB, 72 of which had been given the annotation 'diffraction quality crystal' and the other 72 had been given the annotation 'work stopped'. TEST-RL contains 86 sequences of proteins that are less than 200 amino acids in length. The 43 crystallisable sequences in TEST-RL are a subset of those in TEST that have been filtered by length. The 43 non-crystallisable sequences were selected at random from a larger dataset, which again had a length restriction, and had the TargetDB status of 'work stopped'. TEST-RL was introduced to compare the performance of those predictors with a length restriction to those without. A summary of the prediction accuracies for the various classifiers is shown in Table 8.

Other researchers used the FEAT or TEST-NEW dataset. The FEAT dataset, again introduced by the developers of the ParCrys predictor, contains entries from the TargetDB: 728 entries with status 'diffraction quality crystal' and 728 entries with status 'work stopped' (Overton & Barton, 2006, Overton *et al.*, 2008). The TEST-NEW dataset, introduced by (Kurgan *et al.*, 2009), was also obtained from TargetDB and contained 1000 entries with status 'diffraction quality crystals' and 1000 entries with status 'work stopped' (TargetDB, 2010, Kurgan *et al.*, 2009).

Using data obtained from the Structural Genomics Consortium (SGC) Oxford we derived our own list of sequences with a crystallisable or non-crystallisable outcome. A positive data set was obtained from the 69 protein structures that had been determined from a single sparse matrix screen, the SGC JSCG +4. Some of these structures were obtained from the same sequence, for example, the structures with PDB IDs 2IZR, 2IZS and 2IZU are all from the same sequence. After removing the

repeated sequences, the positive data set, SGC61POS, comprised 61 entries. To ensure that proteins that were not successfully crystallised in the SGC JCSG+4 screen were unsuccessful due to properties intrinsic to the sequence and not the screening conditions, it was necessary to determine whether they had been crystallised in any other screen. If they had been crystallised in other screens at the SGC they were not included in the negative data set, SGC382NEG, which finally comprised 382 sequence entries. Table 9 shows the number of entries in the commonly used data sets together with the custom data set introduced here. As our datasets are different sizes they are reported separately.

Dataset Name	Number of Successful Sequences	Number of Unsuccessful Sequences
TEST	72	72
TEST-RL	43	43
FEAT	728	728
TEST-NEW	1000	1000
SGC61POS	61	-
SGC382NEG	-	382

Table 9: Datasets used for predicting a protein's crystallisability.

The number of successful and unsuccessful sequences in various datasets used for predicting the crystallisability of proteins.

6.2. Protein Sequence Properties

During the development of the various predictors described above, at least 34,500 features have been calculated from protein sequences in order to determine whether a protein would indeed crystallise. These features include counts of di- and tripeptides, separated by up to four other amino acids (Chen *et al.*, 2007, Kurgan *et al.*, 2009), features given in the AAIndex (Kawashima *et al.*, 2008) and many others. It is difficult to determine which features harness the most predictive power as every predictor uses a different set. We investigated this using features previously used by others as well as new features.

The core features calculated for our data matrix are defined on ExPASy's ProtParam web tool. This online Bioinformatics Resource Portal from the Swiss Institute of Bioinformatics provides access to scientific databases and software tools (Artimo *et al.*, 2012). One tool available on the ExPASy server is ProtParam created by Gasteiger *et al.* (2005). This tool computes physical and chemical parameters of a protein from its amino acid sequence.

Proteins are compositions of 20 amino acids in various frequencies with an amino group at one end (the N-terminal) and a carboxyl group (the C-terminal) at the other, as shown in Figure 42. Each amino acid has different properties that can be combined to provide a feature for the whole sequence. The following section describes the various features used in our analysis. The number in parentheses following the feature type indicates the number of parameters calculated for this feature.

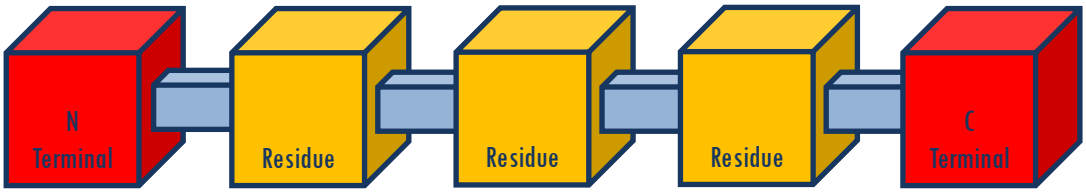


Figure 42: Standard protein sequence structure.
A standard amino acid sequence has an N-terminal and a C-terminal (shown in red). Between these two terminals are the amino acid residues (orange) connected by peptide bonds (blue).

The *Molecular Weight* (1), *M*, is the sum of the molecular mass of each atom making up the protein. This can be calculated by summing the molecular masses of the amino acids (*aa*) in the sequence after adjusting for the dehydration reaction, as shown in Equation 24. For each peptide bond formed between amino acids one water molecule (~18Da) is lost. The water is lost as a combination of a hydroxide (OH) from the carboxyl group of one amino acid and hydrogen (H) from the amine group of another amino acid. If the mass of one water molecule is subtracted for each amino acid in the sequence, then the mass of one water molecule should be added to

account for the remaining hydroxide on the C-terminal and the remaining hydrogen on the N-terminal.

$$M = \left(\sum_{all\ aa} (aa - H_2O) \right) + H_2O \qquad \textbf{24}$$

The *Net Charge* (15) is a summation of the individual charges of certain amino acids in the sequence. The amino acids arginine, histidine and lysine are known to have a positively charged side chain and the amino acids aspartic acid and glutamic acid are known to have a negatively charged side chain. The other 15 amino acids have a side chain with neutral charge. In calculations for net charge the charge of the N-terminal and the C-terminal is also included. The N-terminal is affected by pH in the same way as positively charged side chains and the C-terminal acts in the same way as negatively charged side chains. Charged side chains are affected by pH. Amino acids with a positively charged side chain remain positively charged while their pK_a value is above the pH of the solution. If the pH is greater than the pK_a value, then the side chain becomes neutral. Similarly, negatively charged side chains remain negatively charged while the pH of the solution is of greater value than their associated pK_a value. If the pK_a value is greater than the pH they become neutral.

The *Isoelectric Point (pI*) (1) is the pH at which the net charge of the protein is zero and its calculation from seven key amino acids (Kozlowski, 2012) is described in detail in Chapter 5. We implemented the computational algorithm of Sillero and Maldonado (2006) in VBA (Microsoft Excel) to calculate pI.

The *Sequence Length* (1) is simply the total number of each of amino acids in a sequence.

The *Amino Acid Composition* (20) is given by the number of each of the 20 different amino acids contained within a sequence.

The *Amino Acid Frequency* (20) is the number of the 20 different amino acids within a sequence divided by the sequence length.

The *Atomic Composition* (6) *is* the number of each atom type within the sequence. This can be calculated from the number of carbon, oxygen, nitrogen, hydrogen and sulphur atoms in each amino acid multiplied by the number of each amino acid, accounting for the loss of water through the dehydration reaction. The total number of atoms in a sequence is also used.

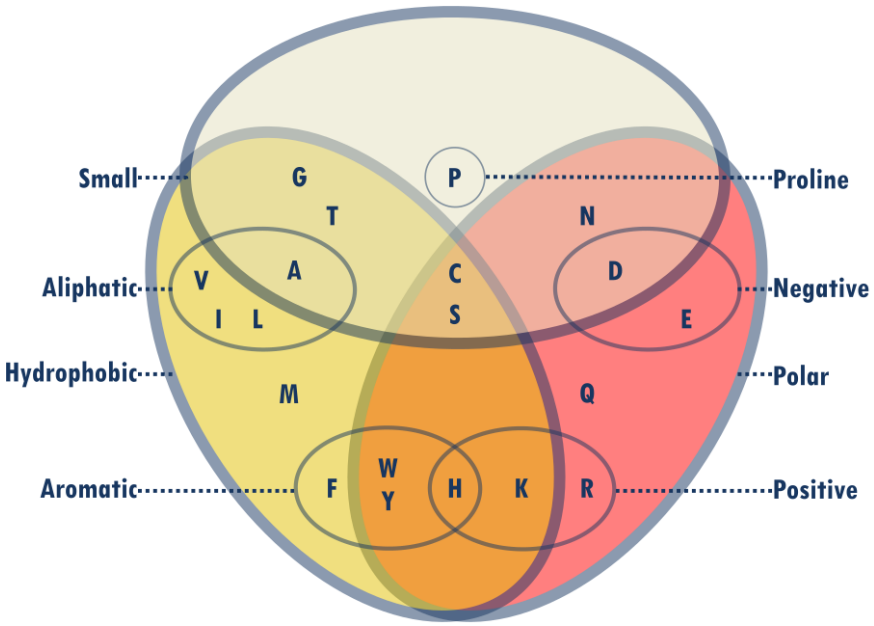


Figure 43: A Venn diagram of amino acid types.

The 20 amino acids, indicated by their single letter code, are divided into groups with other amino acids that share the same properties.

Amino Acid Types (8) are different properties of amino acids due to by the particular side chain. There are eight different types, as shown in Figure 43. The eight types are small, aliphatic, hydrophobic, aromatic, negative, polar, positive and proline. For each sequence, a count of the number of amino acids with this property provides a feature.

The *Extinction Coefficient* (4) can be used to determine protein concentration. Using the method provided by Pace *et al.* (1995), the extinction coefficient for proteins in water is calculated as 5,500 times the number of tryptophan residues plus 1,490 times the number of tyrosines plus 125 times the number of cysteine pairs.

The *Half-Life* (3) of a protein is the length of time it takes for half of the protein to disappear within a cell. The method to determine the half-life is referred to as the 'N-end rule', which refers to which amino acid is on the N-terminal of the protein sequence (Varshavsky, 1997). The ProtParam documentation provides a list of half-lives for mammalian, yeast and e.coli cells for each amino acid (Gasteiger *et al.*, 2005). For example, for a protein with an alanine as N-terminal amino acid, the half-life would be 4.4 hours, >20 hours and >10 hours for mammalian, yeast and E.coli cells respectively.

The *Instability Index* (1) is a value assigned to a protein sequence based on dipeptide combinations. It has been reported that proteins containing certain proportions of some dipeptides undergo rapid degradation and that proteins containing high frequencies of proline, glutamic acid and serine can be unstable (Guruprasad *et al.* (1990). The latter was also found by Rogers *et al.* (1986), who reported a similar affect for methionine and glutamine. On the other hand, asparagine, lysine and glycine are reported to occur in high frequencies in stable proteins (Guruprasad *et al.* (1990). Guruprasad and coworkers have provided a table of instability values for each dipeptide within a sequence. These can be summed to provide the instability index for a protein sequence, where an instability index > 40 suggests the protein is unstable.

The *Aliphatic Index* (1), suggested by Atsushi (1980), is a metric for a given protein sequence based on the quantity of four specific amino acids: alanine, valine, isoleucine and leucine. Atsushi reported that proteins derived from thermophilic bacteria are known to have significantly higher frequencies of these aliphatic amino acids. A protein with a high index may be regarded as having high thermostability.

The *Grand Average of Hydropathy (GRAVY)* (1) is the average hydropathy value for an amino acid sequence. The sum of the hydropathy values for each individual amino acid, provided by Kyte and Doolittle (1982) is divided by the number of amino acids in the sequence to give the average hydropathy value. A positive GRAVY value suggests a hydrophobic sequence, whereas a negative GRAVY value indicates a hydrophilic sequence. Rose *et al.* (1985) and Engelman *et al.* (1986) have also defined a set of amino acids that they consider to be hydrophobic or hydrophilic

although they do not provide a method to calculate a score. Note that hydropathy and hydrophobicity are used interchangeably.

The *Mean Side-Chain Entropy* (3) is the average amount of entropy for each protein sequence, based on estimations of entropy provided by Creamer (2000). In general, entropy is a measure of the amount of energy in a system that is unavailable in a particular state. In folded proteins this energy is unavailable due to protein folding.

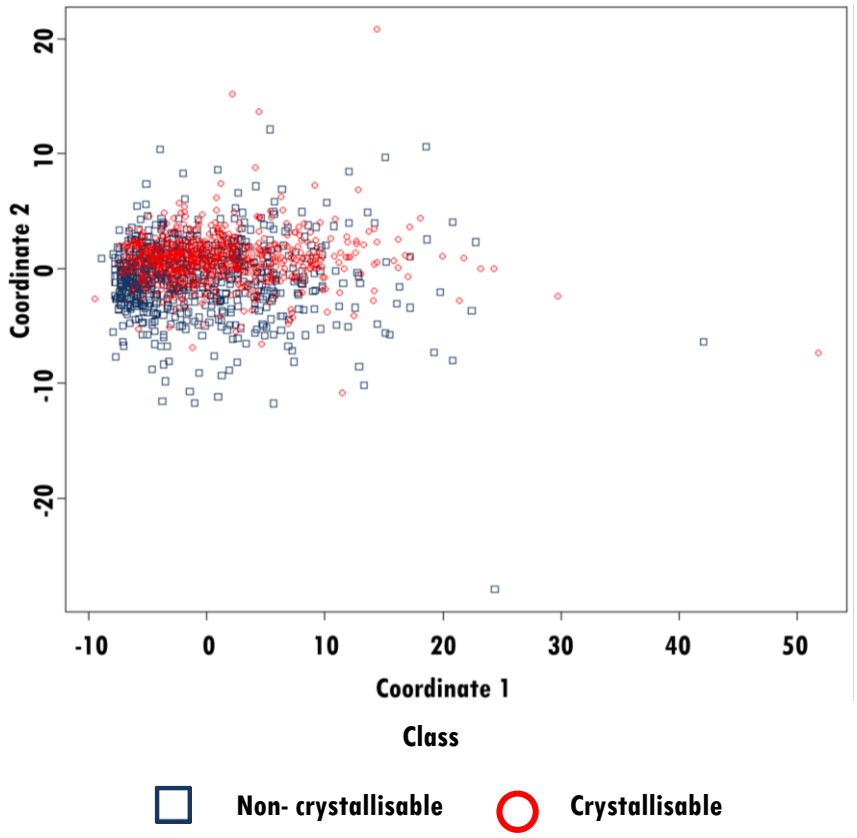


Figure 44: Scores plot for first two principal components.

The fsOur87 features were calculated for the FEAT dataset and then scaled. Principal components analysis was performed and the scores obtained within respect to the first two principal components are shown here.

Here we refer to the set of 87 features described above as *fsOur87*. This set of properties was calculated for the sequences in the FEAT dataset, which we use as training data. Principal components analysis (PCA) was implemented for data reduction in R on the FEAT dataset (Zurich, 2012). The data were scaled to prevent

large variables dominating the analysis. The top 5 principal components only account for 67% of the variance, with 40% of this in the first principal component. The scores plot in Figure 44 that this variance is not due to a difference between groups. Each subsequent component adds little to the cumulative variance with 22 principal components being required for 95% of the variance. As PCA showed that most of the variance in the data was not related to any difference between the two groups and the method did not offer effective data reduction, it was not pursued further. Instead a feature selection method, which does not require any transformation of the variables was used.

It has been shown that the removal of highly correlated features (correlation >0.9) can improve the performance of neural networks (Wendemuth *et al.*, 1993, Hall & Smith, 1997) and therefore we produced a second feature set, referred to as *fsUncorrelated,* consisting of 54 features. We also used the feature sets used to test previous predictors, which we refer to by the name of the original predictor preceded by *fs* for feature set. For example, *fsCRYSTALP* is the set of features used in the *CRYSTALP* predictor. A summary of the size of the feature sets and their source is shown in Table 10. A full list of the features in each of the named sets is listed in Appendix A.

Feature Set	Source	Feature Count
fsOur87	Original Features	87
fsUncorrelated	Original without correlated features	54
fsOB	The OB-Score	2
fsCRYSTALP	CRYSTALP	46
fsParCrys	ParCrys	13
fsCRYSTALP2	CRYSTALP2	88

Table 10: The number of features in the various datasets.

The feature sets fsOur87 and fsUncorrelated were compiled specifically for this study, whereas the other features sets correspond to those used by other researchers to develop prediction tools for determining the propensity of a protein to crystallise.

6.3. Classification

Various machine-learning algorithms have been used to predict a protein's propensity to crystallise, but as Table 8 shows, no particular method stands out as the most successful. Using data on gene expression in cancers, it has also been shown that many of the machine learning methods can be equally successful in classification (Nookala *et al.*, 2013, Caruana & Niculescu-Mizil, 2006). Here we chose an artificial neural network (ANN).

The ANN model is inspired by the neurons in the brain and is trained to associate a particular output with certain input features. A weighted combination of the input features is passed through transfer function in a threshold unit to determine the output of a neuron, as described in detail in Chapter 3. This output is then sent to the next layer or output to provide the class. Here the input features are protein properties and the final output is given as 01 representing uncrystallisable (failure) and 10 for crystallisable (success).

6.3.1. Validation

Figure 45 shows a binomial distribution for a sample of 382 experiments, with each experiment having a 0.5 probability of being classified correctly. Only the range from 40% to 60% accuracy is shown, as this is the range of accuracies most likely to occur if each sequence had a 0.5 probability of being classified accurately. The probability of accurately classifying precisely half (191/382) of the sequences by random chance would be 0.04. The probability of classifying more than 207/382 sequences (54.2%) by random assignment is less than 5%. For most purposes, this would be the accuracy at which the null hypothesis (the distributions are the same) would be rejected in favour of the alternative hypothesis (the distributions are different). Similarly, the probability of classifying fewer than 162/382 (42.4%) of sequences by random assignment is less than 5%. The value of 5% is the probability of making a type 1 error, rejecting the null hypothesis when it is true. Similar random distributions were created for different numbers of experiments. In each instance the distribution was normally distribution around a mean of 50% accuracy with a varying standard deviation (increasing for smaller samples and vice versa).

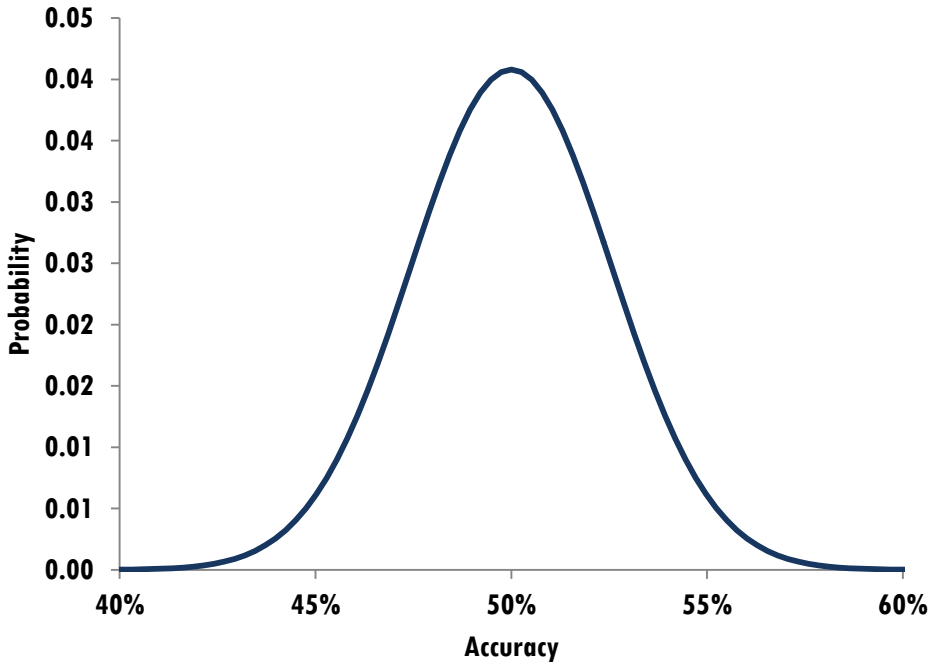


Figure 45: Accuracy for random probabilities.

The binomial distribution for 382 experiments with the probability of success of 0.5 gives a mean accuracy of 50% with a standard deviation of 2.5%. There is 99.9% probability that the accuracy rate for 382 random experiments, with a 50% chance of success for each experiment, lies between 40% and 60%.

6.3.2. Training

Feature Set	15% Testing Set Accuracy (%)
fsOur87	75.7
fsUncorrelated	79.4
fsOB	69.7
fsCRYSTALP	74.3
fsParCrys	75.2
fsCRYSTALP2	74.3

Table 11: The accuracy during training for different feature sets.

The features from the named feature sets were determined for each sequence in the FEAT dataset and used to train a neural network. During training, 15% was used as an internal test set and the results for each feature set are shown.

In order to compare the results of the neural network with reported results using other machine learning algorithms, we trained the network with different feature sets. In each case, the FEAT dataset was used and was split into training, validation and test data sets in the ratio 70:15:15. Using a feed-forward network with the Levenberg-Marquardt training method, various architectures were trialled for each feature set using the Neural Network Toolbox in Matlab (MathWorks, 2011).The optimal architecture was used in each case, where the simplest model with greatest accuracy on the test set was considered as optimal. The difference between training and testing accuracies was used to check for over-fitting. For all feature sets the optimal architecture was found to have two nodes on each of two hidden layers. The accuracy of the training test set is shown in Table 11.

6.3.3. Testing

In order to compare the results from our neural network with those of others, we used our trained network with the previously used independent test data sets, TEST-RL, TEST and TEST-NEW. The results for each set of features from each test data set are shown in Table 12. The results for each test set, using for our own feature sets are also shown Table 13.

Feature Set	TEST-RL	TEST	TEST-NEW	SGC61POS	SGC382NEG
fsOur87	77.9	78.5	68.8	29.5	85.8
fsUncorrelated	74.4	77.1	71.0	57.3	68.3
fsOB	70.9	67.4	68.3	62.3	39.5
fsCRYSTALP	60.5	55.6	60.9	57.4	56.8
fsParCrys	**75.6**	**76.4**	**73.9**	**49.2**	**68.8**
fsCRYSTALP2	59.3	60.4	63.2	39.3	60.5

Table 12: Accuracy of testing sets.

A comparison of the accuracy achieved using different feature sets on five different test data sets. Overall the most successful feature sets are fsParCrys (bold) and our own fsUncorrelated.

Table 12 shows the accuracy for different feature sets ranges from 29.5 to 85.8%. Using publicly available datasets, the features used in the ParCrys predictor provide the best results with our own uncorrelated feature set giving comparable results. The 46 features from CRYSTALP perform worse with the neural network than the OB-Score features, of which there are only two. This shows that the success of the predictor does not depend on the number of features used.

Predictor	Number of Features	TEST-RL	TEST	TEST-NEW
fsParCrys	**13**	**75.6**	**76.4**	**73.9**
CRYSpred	**15**	**80.2**	**79.9**	**73.4**
XTALPred	9	76.7	79.2	70.0
CRYSTALP2	88	69.8	75.7	69.3
ParCrys	13	79.1	71.5	70.6
OB-Score	2	69.8	64.6	Unreported

Table 13: Comparison of results.

The results are shown for the standard test data sets obtained using a neural network trained the fsParCrys feature set in comparison with the results for other predictors (obtained from Mizianty and Kurgan (2012)).

Table 13 shows the results for the neural network trained with the fsParCrys feature set in comparison to other published predictors. Although CRYSPred performs best for two datasets, TEST-RL and TEST, both contain duplicate entries and are smaller than the TEST-NEW dataset. In fact our network trained with the fsParCrys feature set outperforms CRYSPred on the larger TEST-NEW dataset and uses fewer and simpler features. Again, there appears to be no connection between the number of features and the accuracy of the predictor.

We also used the trained networks to predict the crystallisability of the SGC data and obtained mixed results. For most feature sets the results seem to be biased towards either positive or negative outcomes and are generally lower than those obtained for the publicly available data. In some instances, fsOur87 for example, the result is

worse than randomly choosing between success and fail. This suggests that the data used for training is not representative of the data being tested.

6.4. Biochemical parameters

In the previous analysis it has been demonstrated that a selection of features derived from a protein sequence can, to some extent, be used to predict whether a protein can be crystallised. The nature of the neural network with two hidden layers, each with two nodes, makes it difficult to determine which features are most important for the separation of the two groups. The features (pI; GRAVY; counts of the amino acids: D, C, G, H, M, F, P, S, T, W and Y) from the most successful feature set, ParCrys, were used to further explore the properties than can be used to determine crystallisability.

6.4.1. Individual Parameters

The thirteen features in the ParCrys feature set were used separately in linear discriminant analysis (LDA) to determine the discriminatory power of individual features. The LDA was created in R using the entire FEAT dataset for training.

LDA results show the features pI and GRAVY to be the most *powerful* when it comes to predicting the outcome of a crystallisation experiment. This might be expected as these features have been used with several predictors (Overton & Barton, 2006, Overton *et al.*, 2008, Kurgan *et al.*, 2009), Isoelectric point shows notably different rates of accuracy across the different datasets (Table 14), showing a bias towards positives on the SGC data. It might be expected that features involving the amino acids used in pI calculation (D, C, H and Y) would have better discriminatory power than the other amino acids. Similarly, it might be expected that the hydrophobic phenylalanine (F) and hydrophilic aspartic acid (D) would have greater discriminatory power than demonstrated because of their close link to average hydrophobicity. However, this is not the case. The use of the different amino acid counts as a feature varies across the different test sets and no single amino acid count stands out as particularly useful for classification. For the SGC datasets some of the amino acid features give results that are worse than random guessing.

Feature	TEST-RL	TEST	TEST-NEW	SGC61POS	SGC382NEG
pI	66.3	66.0	69.9	57.3	32.9
GRAVY	53.5	52.1	57.7	52.4	66.8
D	53.5	47.9	64.6	50.8	57.6
C	53.5	62.5	56.2	44.3	72.0
G	53.5	43.1	58.8	42.6	67.5
H	50	43.1	55.1	59.0	30.1
M	50	45.1	57.3	50.8	56.0
F	50	38.9	57.5	45.9	59.7
P	51.2	35.4	53.0	41.0	65.2
S	54.7	64.6	46.5	59.0	41.6
T	53.5	38.2	59.5	44.2	61.3
W	53.5	47.2	52.7	39.3	66.5
Y	47.7	46.5	60.7	47.5	48.4

Table 14: Accuracy of individual features for prediction.

The percentage accuracy is shown for the different test sets when using each feature individually in LDA to classify as either crystallisable or non-crystallisable. The FEAT dataset was used for training and TEST-RL, TEST and TEST-NEW for testing.

6.4.2. Combinations of Parameters

The thirteen features from ParCrys were also used together in LDA. Discriminant functions were identified using the FEAT dataset and four fold venetian blind cross-validation. In each fold 25% of the data was used as the training set and the remaining 75% was used as the test set. The results from each of the four (non-overlapping) subsets are shown in Figure 46. On average the LDA achieved 70% accuracy across the four test sets, with only marginal differences between them.

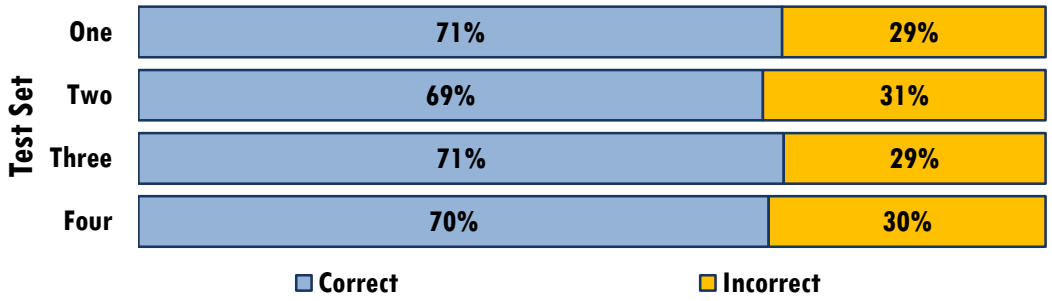


Figure 46: Summary of cross-validation results.

LDA was performed using the ParCrys features on the FEAT dataset with four fold cross-validation. The figure shows the percentages of correct (dark) and incorrect (light) classifications on the test data.

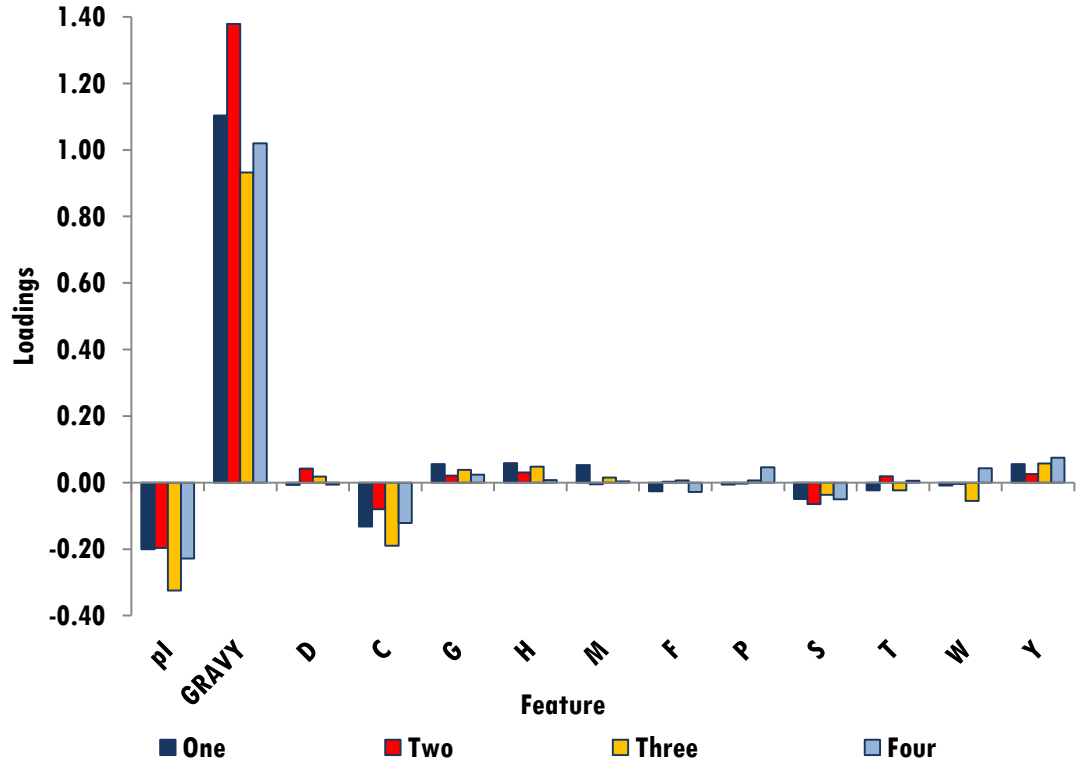


Figure 47: LDA loadings for the FEAT dataset with ParCrys features.

The loadings of the four linear discriminant functions obtained for the FEAT dataset show that GRAVY, pI and the number of cysteines are the dominant features in determining whether a protein will crystallise for each of the four testing sets.

Not only were the results similar for each of the four test sets, but in each model the linear discriminant created from 25% of the data was dominated by GRAVY, followed by pI and the number of cysteines as indicated by the loadings in Figure 45. The other ten features had loadings of 0.1 or less showing that they contribute little to the discriminant functions. In the analysis of individual features, C (the cysteine count) did not appear to be useful for discrimination, but when combined with GRAVY and pI, it does seem to have some discriminatory power. This is also true to some extent of Y (the tyrosine count), it has the fourth highest loading when used in combination with other variables. Although GRAVY, the sum of hydropathy values scaled by residue count, has the greatest discriminatory power, the features F (phenylalanine count) and D (aspartic acid count), which are closely associated with hydrophobicity, do not appear to be useful. It may be that these features cannot add to the discrimination due to their high correlation with GRAVY.

When the three parameters, GRAVY, pI and the number of cysteines, are plotted against each other, some separate areas corresponding to crystallisable and non-crystallisable sequences can be seen (Figure 48**)**. It can be seen that the crystallisable proteins are clustered in an acid to neutral pI, with a slightly negative GRAVY value and a low cysteine count. The number of crystallisable proteins outside of this zone decreases to a handful, although the opposite is not true. The non-crystallisable proteins are spread across a large area in all three parameters.

The model was used to categorise the sequences in the TEST-NEW set and achieved a correct classification rate of 73%. However, none of the neural network models, trained using the fsParCrys or other feature sets, were able to classify both the positive and negative sequences in the SGC data well.

Figure 49 shows the three parameters, GRAVY, pI and the number of cysteines, plotted against each other, for the SCG data. As with the TEST-NEW data, the uncrystallisable proteins are spread across a large area of this parameter space. Although many sequences corresponding to proteins that crystallise are in the zone identified for the TEST-NEW data (acid to neutral pI, with a slightly negative GRAVY value and a low cysteine count), they are also spread across a larger area, making the overlap greater for the SGC data than for the TEST-NEW data.

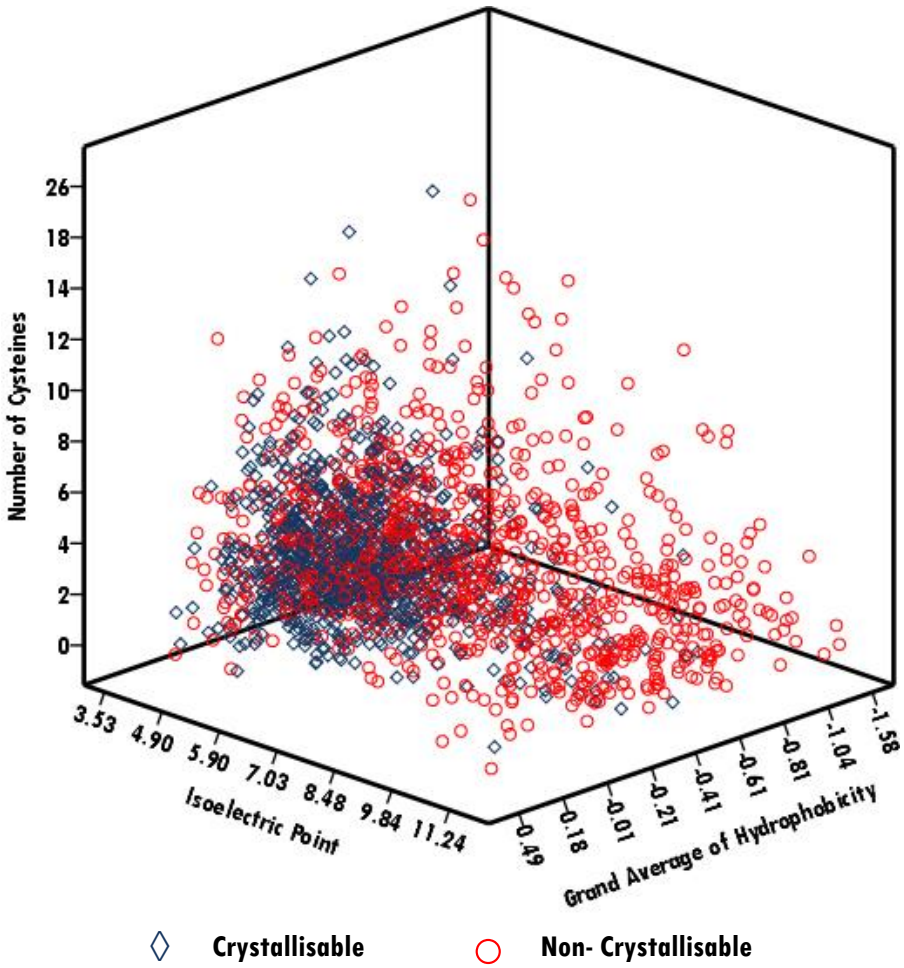


Figure 48: TEST-NEW data based on the most discriminatory variables.

The TEST-NEW data is plotted for the variables GRAVY, pI and the number of cysteines. Crystallisable proteins are represented by blue diamonds and non-crystallisable proteins by red circles.

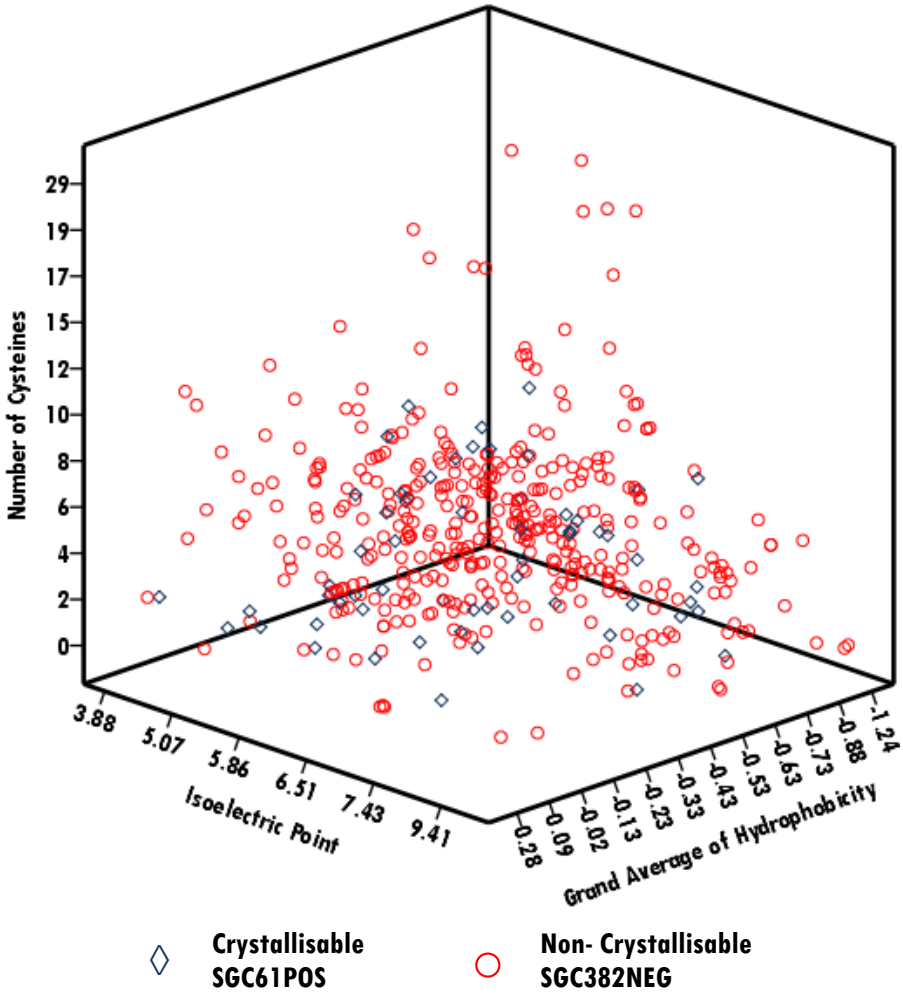


Figure 49: SGC data based on the most discriminatory variables.

The two SGC datasets plotted for the variables GRAVY, pI and the number of cysteines. Crystallisable proteins (SGC61POS) are represented by blue diamonds and non-crystallisable proteins (SGC382NEG) by red circles.

As the data for each of the three properties pI, GRAVY and cysteine count was not normally distributed for the FEAT dataset, the non-parametric Mann-Whitney-Wilcoxon test (MWW) was performed and revealed differences in the populations between the positive and negative sequences of the FEAT data for all three properties. However, the difference between positive and negative populations for SGC data was not so well defined. For the isoelectric point and cysteine count the null hypothesis (there is no difference between positive and negative populations), could not be rejected with p-values of 0.23 and 0.95 respectively. A p-value of 0.001 provided evidence against the null hypothesis for the variable GRAVY, suggesting a difference between the distributions of the hydrophobicities of the positive sequences

and the negative sequences. The problem in classification occurs because, although the negative SGC dataset has a similar distribution to the negative FEAT dataset used for training, the positive SGC data has a different distribution to the positive FEAT dataset. Furthermore the positive SGC data has a similar distribution the negative FEAT dataset. The distributions are shown in Figure 50.

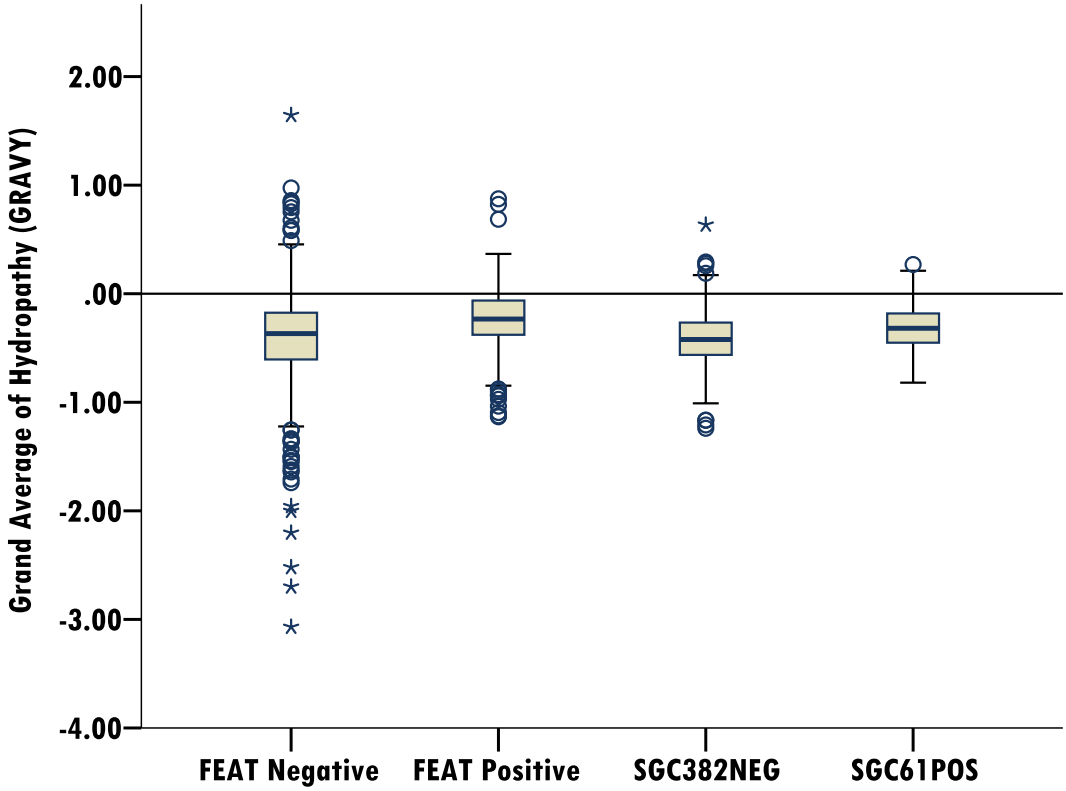


Figure 50: Boxplots of GRAVY values for the FEAT and SGC datasets.

The four boxplots show the distribution of the GRAVY values for the positive and negative sequences in the FEAT and SGC datasets datasets. The line in the centre of the box represents the median, the lower and upper bounds to the box represent the first (25%) and third (75%) quartiles. Each whisker is drawn to the most extreme value within 1.5 box lengths of its respective box boundary. Circles are representative of data points more than 1.5 box lengths away from the closest box quartile and stars are 3 box lengths away. It can be seen that the SGC61POS and SGC382NEG populations both have median values closer to that of FEAT negative than that of FEAT positive.

The overall homology of the sequences was also inspected, as strong homology would compound the number of incorrect classifications. The correlation between the

properties of each pair of sequences was found to be non-significant (at 5%). The sequences were also separated into their families (determined by the SGC) and the standard deviation calculated for each feature, this again revealed that there was no similarity between sequences, as suggested by standard deviations that were large with respect to the mean. To further test the classification, a much larger set from the PDB was used. The ParCrys features were calculated for 25,316 sequences from the PDB. The neural network using these features, that had previously achieved an average of 75% accuracy on three test datasets, was used to classify this PDB data and only achieved 55% accuracy. Restricting the PDB data to sequences submitted between July 2006 and December 2008 to reflect the dates of the TEST-NEW data did little to improve the accuracy with just 58% (3180/5453) correctly predicted as crystallisable. As shorter sequences are not well represented in the FEAT dataset, we also tried restricting the PDB data to sequences more than 99 amino acids in length. Again an accuracy of just 58% (13,233/22,829) was achieved. To ensure our methodology was not the cause of the low prediction rate we used another predictor. Taking a random sample of 1,000 sequences from the PDB with length between 100 and 1,000 residues we were able to use the CRYSTALP2 online predictor (Kurgan *et al.*, 2009). Again the accuracy was low, with only 60% of sequences correctly classified as crystallisable'

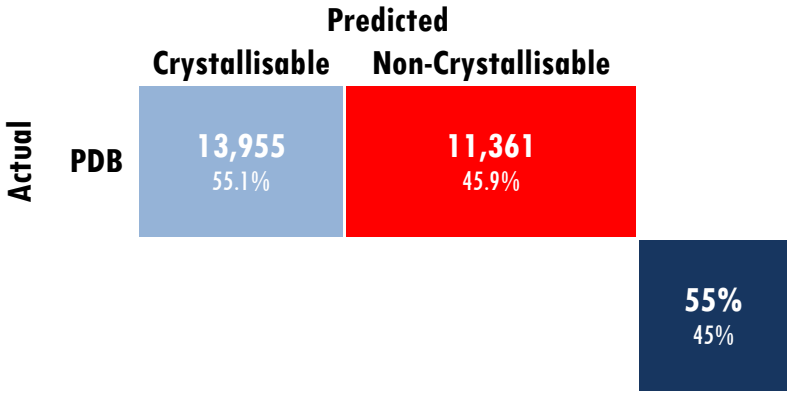


Figure 51: Confusion matrix for the prediction of PDB sequences.

The classification results are shown for 25,316 sequences from the PDB, of which the neural network correctly predicted just 55%.

The original training and test datasets were both obtained from TargetDB and no PBD data was included in either (FEAT or TEST-NEW). Kurgan *et al.* (2009)

specifically state that crystallisable proteins in the TEST-NEW dataset were chosen if they were annotated as having "Diffraction-quality Crystals", but not annotated with In PDB in the Status field. No motivation for excluding sequences resulting in PBD structures is given. The reason for sequence differences between proteins designated as producing diffraction-quality crystals in TargetDB and those that result in a structure deposited in the PDB is not clear. One possible explanation is the fact that only structural genomics targets are included in TargetDB and may be restricted, for example by particular medical interests, whereas structures deposited in the PDB are from a wider, and potentially more difficult to crystallise, range of proteins. On the other hand, proteins for which diffraction data is collected, but the structure is not solved are presumably the most different from known protein structures. Diffraction data is collected for about a third of the structural genomics targets for which crystals are obtained and only two-thirds of these result in a protein structure in the PDB (Westbrook *et al.*, 2003). Sequences from these proteins with diffraction-quality crystals, but no PDB entry are precisely those included in the training and test datasets producing models that do not generalise to PDB data.

6.5. Discussion and Conclusions

It has been stated that the use of several predictors together could allow accurate classification of 90% of sequences (Mizianty & Kurgan, 2009). The datasets used are sufficiently large to overcome any difficulties with accuracy caused by data size (Cunningham, 2000) and the number of features searched is large, 34,618 (including 34,000 from *n* collocated amino acids). However, the number of features that could possibly be calculated from a protein sequence is unknown. For example, the user could potentially determine that the oligomeric state (determined through PDB search for protein similarity) or the amount of *n* collocated amino acids could determine the difference between the positive and negative datasets.

Several values of *n*, for *n* collocated amino acids have been used in previous predictors to determine a protein's crystallisability (Chen *et al.*, 2007, Kurgan *et al.*, 2009, Charoenkwan *et al.*, 2013). In our search, however, we find like others that isoelectric point and grand average of hydrophobicity are the properties that hold the most predictive power (Goh *et al.*, 2004, Overton & Barton, 2006, Mizianty &

Kurgan, 2009). A search for further features is likely to involve complex properties and be based on predictions such as secondary structure or molecular interactions, which are currently limited to 80% accuracy (Dor & Zhou, 2007), but are already included in some predictors (Mizianty & Kurgan, 2012, Smialowski *et al.*, 2006).

We aimed to find out, not only 'will my protein crystallise?', but 'why will my protein crystallise?' and the information provided from the black box method of a neural network cannot answer this. We therefore also used linear discriminant analysis (LDA), which is easier to interpret, but was unable to classify as accurately. From this, we were able to show that just three features, pI, GRAVY and cysteine count, were providing the majority of the discrimination between classes.

The use of pI and GRAVY has been shown before to have predictive power when it comes to determining the crystallisability of a sequence. In a study of 500 proteins from the *Thermotoga maritima* proteome, Canaves *et al.* (2004) were able to show that that crystallisable proteins are located in a cluster, similar to ours, in the GRAVY-pI parameter space. The results of other predictors also demonstrate the usefulness of GRAVY and pI (Overton & Barton, 2006, Kurgan *et al.*, 2009).

Artificial neural networks have been shown to be as good as any other classifier (Nookala *et al.*, 2013) and indeed we were able to produce a classifier that is comparable to the best of those already available when applied to the same datasets. Our results show that neural networks can be used to predict whether a sequence will crystallise, at least as successfully as any other machine learning method. Over three publically available test sets our neural network successfully classified more sequences than any other predictor. Although the percentage accuracy is a marginal improvement over the previous best classifier, CRYSpred (Mizianty & Kurgan, 2012), the model is simpler. Our classifier uses two calculated values: isoelectric point and grand average of hydrophobicity along with 11 counts of amino acid frequency. In comparison, the features used in CRYSpred include the sum of predicted disorder scores and the distribution of amino acids in alpha helices in thermophilic and mesophilic proteins. As CRYSpred requires features obtained from a predicted secondary structure, it follows that if this prediction is wrong any classification using this information is also likely to be wrong. Occam's razor

suggests that a simpler model should be selected over a more complicated model without strong evidence to support its use. In classification, a simpler model also helps to prevent over-fitting.

The use of sequence-derived variables as useful indicators of a proteins propensity to crystallise must be questioned, given that these are optimised to identify the most promising crystallisation targets from particular protein families. The problems in classification of the SGC and PDB data must be caused by properties intrinsic to the proteins in these datasets. Difficulties in predicting whether a protein will crystallise may arise due to the purification process, or to chemical and physical parameters which are not considered in sequence-based predictors (Smialowski *et al.*, 2006). In order to re-train classification algorithms, data on unsuccessful experiments would be needed as well as data on successful experiments, such as can be obtained from the PDB. Information on failed experiments is also necessary to investigate the relationship between protein properties and the conditions that result in crystals (Hennessy *et al.*, 2000). This could potentially allow properties that can be measured or calculated before crystallisation trials begin to be used to predict the best initial conditions to try.

PPCpred (Mizianty & Kurgan, 2011) and PredPPCrys (Wang *et al.*, 2014), both available online, have not only determined protein crystallisability but also whether it will pass certain stages of the structure determination pipeline. Once the decision has been made to progress with crystallisation it is then necessary to determine under which conditions a protein will crystallise.

7. The Propensity of Chemicals to Promote Crystallisation

Once a crystallographer has decided to attempt to crystallise a protein, whether or not persuaded by the outcome of a predictor, it is necessary to determine suitable conditions. In 1962 Max Perutz and John Kendrew were awarded the Nobel Prize for Chemistry for being the first people to solve the structure of a complex protein, specifically equine haemoglobin (MFPL, 2012). Perutz once said *"crystallisation is a little like hunting, requiring knowledge of your prey and a certain low cunning" (Fink et al., 2009),* which caused Hennessy *et al.* (2000) to pose the question *"are there good hunting grounds?",* or less poetically, are there favourable regions of crystallisation parameter space? The search for successful regions of parameter space using complete factorial sampling is not possible. This problem is further compounded by physical properties such as pressure, gravity and biochemical properties such as the protein itself and any ligands.

One method for optimal searching of crystallisation parameter space uses regions of known success. For example, Jancarik and Kim (1991) describe the design of a sparse matrix screen which used the most popular conditions found in the literature. However, determining the success of individual chemicals from frequency counts in the literature and online repositories can be problematic. Repositories such as the PDB (Berman *et al.*, 2000) and the BMCD (Tung & Gallagher, 2008) provide no negative examples and therefore give no indication of relative success rates, i.e. how many times a chemical is used before successful crystallisation (Newman, Bolton, *et al.*, 2012). Consideration of the properties of the proteins to be crystallised, together with the success rates of the chemicals for particular types, would allow greater optimisation in crystallisation. This may be achieved using anecdotal evidence (Samudzi *et al.,* 1992) or the use of statistical analysis.

In 2004 Rupp and Wang provided a broad overview of some of the issues in crystallisation and suggested generic techniques that could be used to improve the attrition rates of protein crystallisation. They suggest a method of normalised

frequency analysis known as crystallisation propensity to provide relative success rates. This takes into account both positive (crystalline) and negative (non-crystalline) outcomes and may provide support for the reduced use, or even removal from screens, of less successful chemicals and allow focus on those with higher success rates.

Propensity is an intra-property comparison, for example, pH 6.5 can be compared with pH 7.5, and provides a statistic that ranges from 0 upwards. It provides an indication of how a specific property of parameter space (chemical, pH or protein) relates to the average property. For example, a propensity value of 2 suggests that protein *x* is twice as likely to crystallise as the average protein, whereas a value of 0.5 suggests that protein *y* is only half as likely to crystallise as the average. Throughout this chapter *a trial* refers to the data associated with one particular well from one specific screen.

The crystallisation propensity for a parameter is defined as follows:

$$Propensity = \frac{Rate\ of\ Success\ (RS)}{Average\ Success\ Rate\ (AS)}, \quad \textbf{25}$$

where

$$RS = \frac{Occurence\ in\ successful\ trials}{Occurence\ in\ all\ trials}, \quad \textbf{26}$$

and

$$AS = \frac{All\ successful\ trials}{All\ trials}. \quad \textbf{27}$$

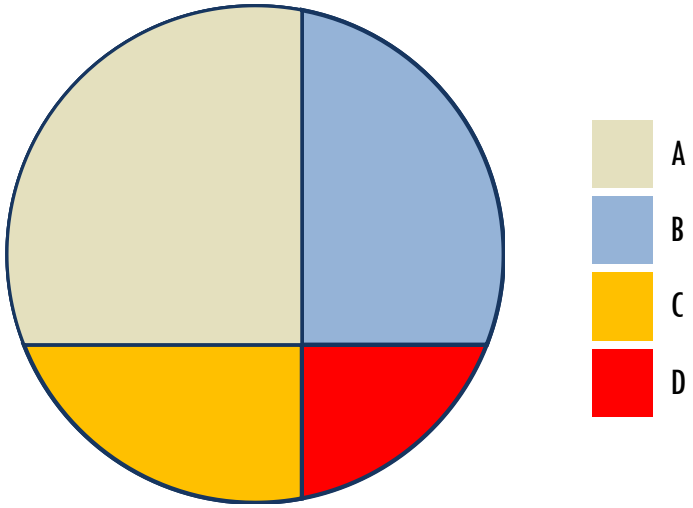


Figure 52: Visualisation of propensity.

As a whole, the circle (A + B + C + D) represents all trials. The section beneath the horizontal chord (C + D) represents those trials that resulted in crystalline material and the section to the right of the vertical chord (B + D) represents trials with the property for which the propensity is being calculated. A represents those trials without the property or result in crystalline material; B represents trials with the property but do not give crystalline material; C represents trials without the property but do result in crystalline material; and D represents those trials with the property resulting in crystalline material.

Re-writing equations 26 and 27 to correspond with Figure 52 gives:

$$RS = \frac{D}{B + D}, \qquad \textbf{28}$$

and

$$AS = \frac{C + D}{A + B + C + D}. \qquad \textbf{29}$$

In demonstrating the usefulness of propensity, Rupp and Wang analysed data from the TB structural genomics consortium on 230,000 crystallisation trials, sampling 55 chemical species across 5 unit intervals of pH from pH 4.5. They found that the propensity of all chemical species in their data was normally distributed which allowed them to define 'supercrystallisers' - the chemical species in the top 5% of the

distribution. At the other end of the spectrum, although not discussed by Rupp and Wang, this allows 'awfulcrystallisers' to be defined- the chemical species with a propensity in the bottom 5% of chemicals.

The results of Rupp and Wang show that just under half (45%) of their chemicals have a propensity greater than one, with the top ten chemicals, the supercrystallisers, including PEG 2000 MME, PEG 5000 MME, PEG 2000, PEG 6000, PEG 4000 PEG 400, calcium chloride, sodium formate, potassium sodium tartrate and MES. The eight awfulcrystallisers are 2-butanol, isopropanol, MPD, EDTA, ammonium phosphate and acetate, ethanol and DMSO.

7.1. Results

We have investigated crystallisation propensity with new data and compared our results with those in the literature. The AstraZeneca dataset has information on 573,786 crystallisation trials. This includes 13,550 (2.4%) successful trials, where a trial with any crystalline result is deemed a success. In this analysis we only consider the chemicals in the crystallisation screens as variables and do not consider temperature, purification method, or the use of ligands due to the extent of missing data in these fields. The proteins include human, bacterial, virus and other mammalian targets. Although propensity cannot be calculated for data in the PDB, due to the lack of negative results, it is still possible to perform frequency analysis to determine the chemicals that are used most often. We show that the chemicals contained within a well are interdependent and that the propensity of one chemical is affected by another, but that propensity does provide more information than simple frequency counts.

Propensity analysis for the 37 chemicals in nine screens at AstraZeneca (Figure 53) suggests that slightly more than half (55%) of the chemicals perform better than average. The propensities were not normally distributed (as verified with a QQ plot and a KS test) so it was not possible to define supercrystallisers. The top 10 includes four salts (lithium sulfate, magnesium chloride, calcium acetate and ammonium sulfate), two buffers (TRIS and PCTP) and one organic (ethanol).

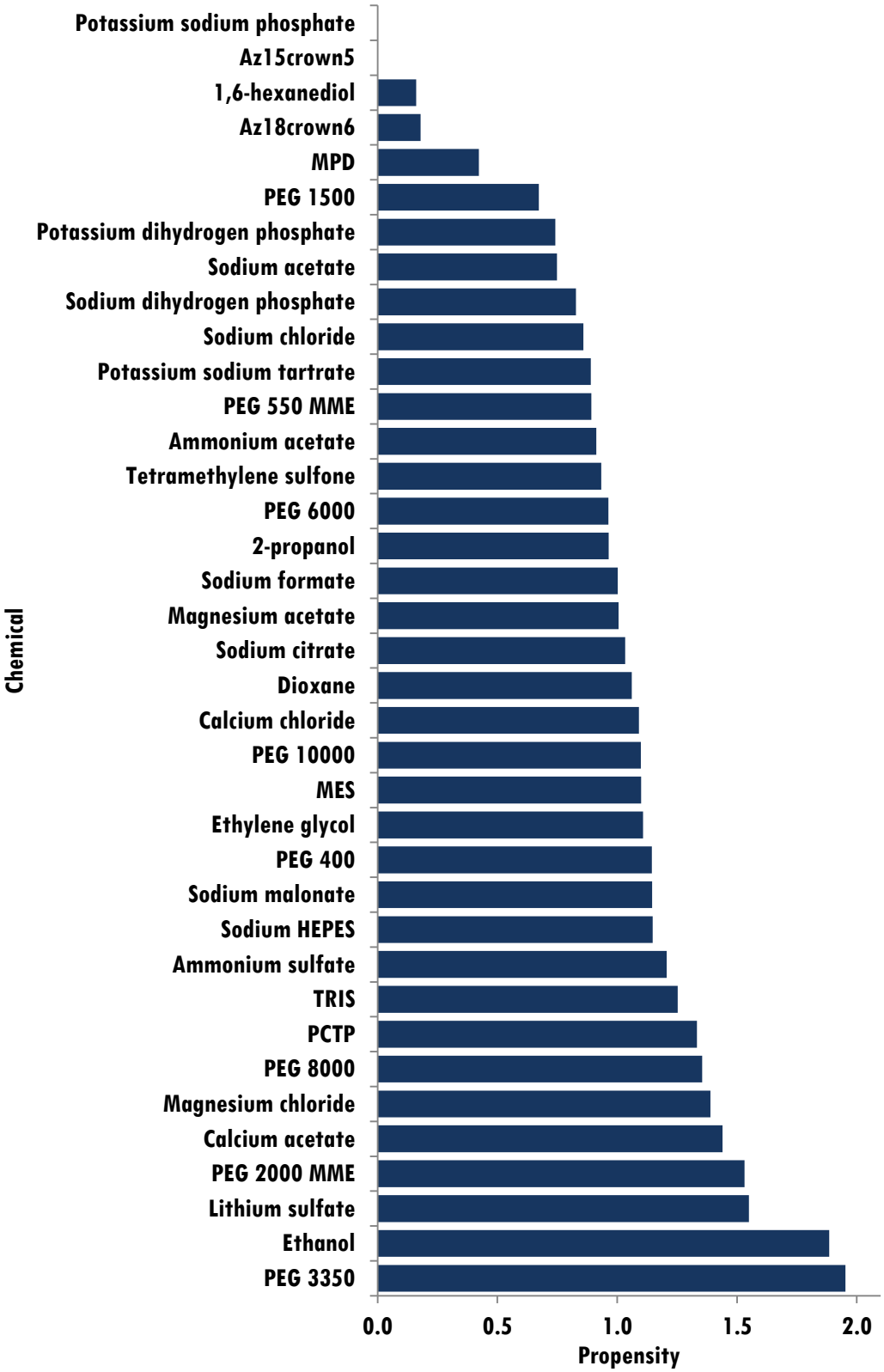


Figure 53: The propensity of chemicals in AstraZeneca screens.

The propensity of 37 chemicals used in the nine screens of the AZ dataset. Propensities were calculated using 13,550 crystalline results from 573,786 trials. Error bars are not shown, as they are negligible after normalisation due to the large sample size.

The top ten chemicals with the highest propensity also includes three PEGs of varying weights (3350, 2000 MME, 8000). To the best of our knowledge, PEG (6000) was first used to crystallise alcohol oxidase in 1968 (Janssen & Ruelius). It was not until ten years later that PEGs became the reagent of choice, following an endorsement from McPherson Jr (1976) who concluded that

'[PEG] may be the best initial trial reagent for crystallisation'.

A summary of 44,063 crystallisation conditions from the PDB dataset also suggests PEG to be a successful crystallisation reagent, appearing three times in a list of the top ten chemical species (Table 15). Like Figure 53, this also includes the buffer TRIS and the salts magnesium chloride and ammonium sulfate along with five other chemicals.

Rank	Chemical	Count
1	PEG 3350	9,264
2	TRIS	8,375
3	Ammonium sulfate	8,225
4	HEPES	5,795
5	PEG 4000	5,637
6	Sodium chloride	5,248
7	Sodium acetate	5,194
8	PEG 8000	4,095
9	Magnesium chloride	3,845
10	MES	3,664

Table 15: The ten most prevalent chemicals reported in the PDB.

The ten most prevalent chemical species are shown together with the number of entries in the PDB-BLAST-reduced dataset, consisting of 44,063 PDB entries.

Subsequent studies of crystallisation data have provided evidence to support McPherson's claim (Hui & Edwards, 2003, McPherson, 1999). In 1984 PEG was ranked second in a list of species used to induce crystallisation (Gilliland & Davies, 1984) and in 1991 PEGs were included in half (25/50) of the wells of Jancarik and

Kim's popular sparse matrix screen. PEGs have also been previously reported to be amongst the most prevalent chemicals in the PDB (Fazio *et al.*, 2014, Peat *et al.*, 2005) and have shown to be more successful in crystallising protein-protein complexes than ammonium sulfate (Radaev & Sun, 2002). Although the mechanism that allows PEGs to be successful crystallization reagents is not well understood, it appears that they compete with water molecules to interact with the protein, forcing it out of solution (McPherson, 1989a, Lee & Lee, 1981). The varying weights and lengths enable a steric exclusion mechanism to occur that excludes protein from zones of the solution and increases local activity and solubility (Laurent, 1963, Ward *et al.*, 1975). A further advantage is that since they are of neutral pH they do not require large concentrations of buffer, however, we have shown previously that they become acidic over time (Ray Jr & Puvathingal, 1985) and as a result it might not be possible to reproduce certain crystallisation experiments.

A study of one protein, *Aspergillus flavus* urate, showed that modification of the concentration and the weight of PEG included in a crystallisation solution can modify chemical parameter space in such a way that the thermodynamic parameter A_2 is changed and moved into the 'crystallisation slot', where crystallisation is more favourable. A_2, the second virial coefficient, is used to provide corrections to the ideal gas law. In practice A_2 is a number obtained by interpreting Static Light Scattering (SLS) output (Kratochvíl, 1987) or through self-association chromatography (Tessier & Lenhoff, 2003). It has been shown by George and Wilson (1994) that proteins which successfully crystallise have an A_2 value between -1×10^{-4} and -8×10^{-4}. With this knowledge, Vivares and Bonneté (2002) were able to show that different crystallisation parameters such as pH, temperature and the volume and weight of PEG in the crystallisation solution could affect the experiment such that A_2 was in the range in which crystallisation had been shown to occur.

Other successful chemicals are either salts or buffers used to control pH and assumed to be otherwise chemically inert with respect to crystallisation (although this is contestable (McPherson, 1995). The salts have a similar effect to PEGs in crystallisation solutions, by competing with the protein for water molecules (McPherson, 1989a). In some instances metal cations (and anions) from salts, such as those of calcium and magnesium, bind to proteins and can stabilise the crystal lattice

(Kretsinger, 1976, Jayachandran *et al.*, 2007). Whilst sodium chloride and ammonium sulfate increase solubility, magnesium chloride has been suggested to decrease solubility by binding to the protein and freeing up water (Arakawa & Timasheff, 1985, Kretsinger, 1976, Jayachandran *et al.*, 2007). Many of these salts have been identified in successful crystallisation conditions before, using data from the PDB (Peat *et al.*, 2005) and the BMCD (Lu *et al.*, 2012), but the literature is inconclusive as to whether they do encourage crystallisation. Two studies reported sodium chloride to be a poor crystallisation reagent, but suggested ammonium sulfate to be successful (McPherson, 2001, Rupp & Wang, 2004). In their analysis of crystallisation propensity, Rupp & Wang also found that magnesium chloride was less likely to produce crystals than the average chemical.

Surprisingly, analysis of the AZ data shows ethanol and lithium sulfate to be successful crystallisation reagents. The success rate of lithium sulfate was previously shown to be average (Rupp & Wang, 2004, McPherson, 2001) and ethanol has been reported amongst the least successful groups of chemicals in crystallisation (Hosfield *et al.*, 2003, Page & Stevens, 2004, Rupp & Wang, 2004). Ethanol is included as a cryoprotectant and has been shown to have to no ill effect (Tran *et al.*, 2004, Farley & Juers, 2014). As it occurs in many successful solutions in the AZ data, it does appear that ethanol does not adversely affect crystallisation. Organic chemicals tend to evaporate from the crystallisation solution to the sitting drop, causing the concentration in the drop to fall rather than to increase to a supersaturated state (Kimber *et al.*, 2003). To compound this, certain concentrations of some organic chemicals denature the protein by acting like a detergent (where part of the molecule binds with the protein and the other part binds with water) (McPherson, 1989a).

In contrast to the highly successful chemicals, some additives appear in very few successful crystallisation solutions. In analysis of the PDB data, 268 chemicals have been used fewer than five times, with 108 leading to a single protein structure. For 83 of these 108 chemicals (76%), a protein structure was obtained for the same BLAST group using alternative conditions. The other 25, of which 8 are ligands, are chemicals that did lead to a unique protein structure and might be considered a last resort list. The chemicals with the lowest propensities in our data contained the set of custom chemicals, referred to as AZ crowns, which were designed to contribute

similar effects to crystallisation as PEGS. As these chemicals were not used for many projects and do not appear in the literature, it is not possible to give a reason for their low propensity. Another group of chemicals with low propensity are pH amending chemicals, including potassium and sodium dihydrogen phosphates and some PEGs (possibly degraded). It is possible that, due to their structure, these chemicals have an effect on pH in addition to their effect on crystallisation as a salt.

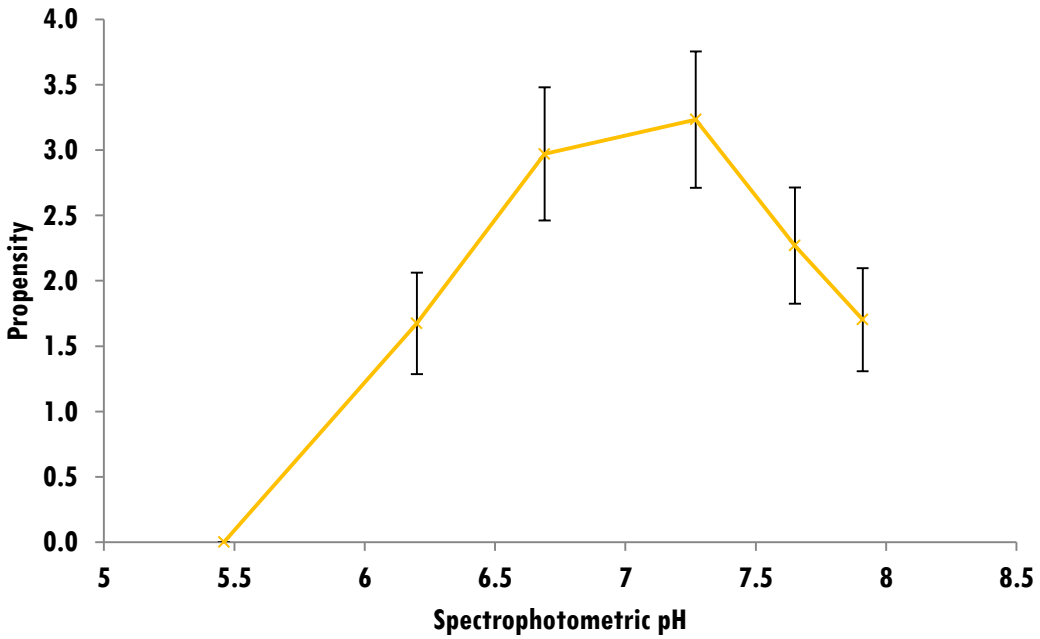


Figure 54: The effect of pH on propensity.

The propensity of 25% v/v PEG 3350 with 0.2 M magnesium chloride is shown for different pH values (as determined using spectrophotometry). Data was obtained from one filter screen with 809 crystalline results in 33,828 trials. Errors shown are normalised proportional errors.

Figure 54 shows that the propensity of PEG 3350 varies with pH, with values between 0 and ~2.75 and potentially up to 3.3 when allowing for error. Thus 15% v/v PEG 3350 could be considered more than twice as effective in crystallisation as the average chemical, but the variation in propensity means that it could also be viewed as a chemical that does not result in crystalline material. Propensity can therefore only be useful if pH is taken into account in the calculation. Furthermore, the results shown here are for magnesium chloride and PEG 3350 together, so these trials count

positively towards the propensity of both chemicals, even though one may have had no effect on crystallisation.

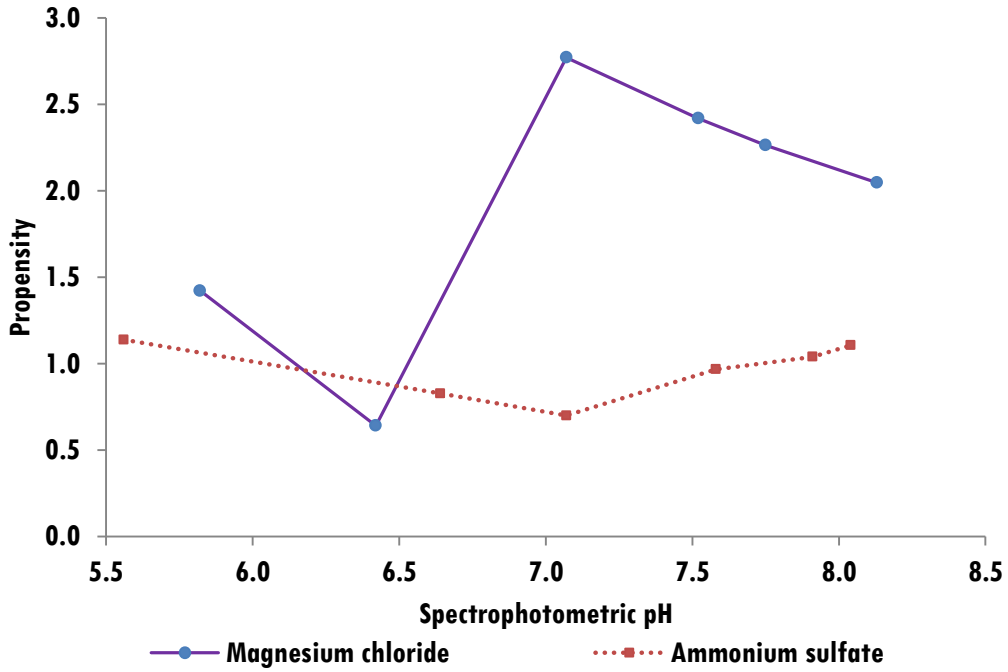


Figure 55: The propensity of PEG 3350 with different chemicals.

The propensity for 15% v/v PEG 3350 with 0.1 M ammonium sulfate is compared to 15% v/v PEG 3350 with 0.1 M magnesium chloride. Data was obtained from the Filter 2 screen with 497 crystalline results out of 22,712 trials.

Figure 55 shows how varying the chemical combination within a screen can affect the propensity. The two graphs show that propensity changes when PEG 3350 is combined with different chemicals. Near pH 5.5 the propensity of PEG 3350 with either chemical is roughly similar, but near pH 7, PEG 3350 with ammonium sulfate has a propensity of ~0.75 (less than average) whereas PEG 3350 with magnesium chloride has a propensity of ~2.75 (almost 3 times the average). Thus, we must consider combinations of chemicals rather than individual components within a crystallisation solution.

The C6 distance metric (Newman *et al.*, 2010) provides a similarity measure between crystallisation conditions and gives the distance between PEG 3350 with ammonium sulfate and PEG 3350 with magnesium chloride as 0.66, provided the pH is the same

for both solutions. However, the observed difference between these two solutions does not show the same result.

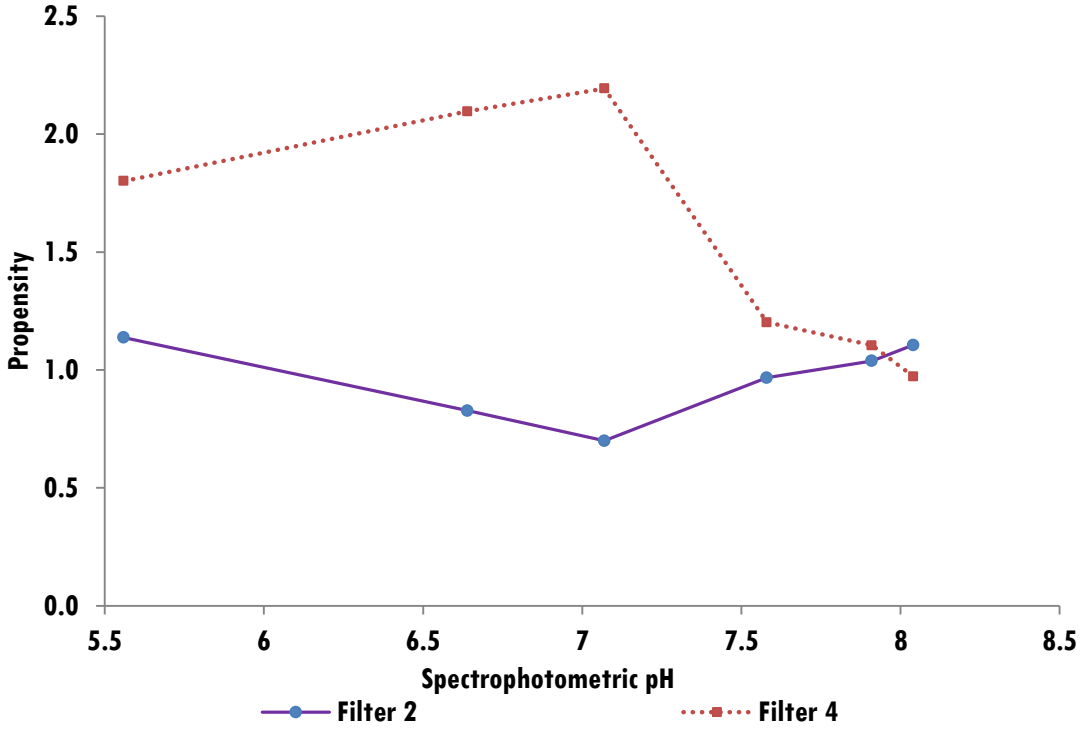


Figure 56: The effect of proteins on propensity.

Propensity was calculated for 15% v/v PEG 3350 with 0.1 M ammonium sulfate across a range of pH for two screens. The propensity for the Filter 2 screen was calculated based on 497 crystalline results out of 22,712 trials and the propensity for the Filter 4 screen was calculated from 701 crystalline results out of 33,048 trials.

We must also consider the proteins involved in the propensity calculations. Figure 56 shows data for the same experimental condition, 15% v/v PEG 3350 with 0.1 M ammonium sulfate, calculated from two different screens, Filter 2 and Filter 4. As the projects being screened are different, the two sets of proteins are mutually exclusive. The variation shown in Figure 56 is predictable; if all proteins crystallised in the same conditions, then parameter space could be reduced to a set of core conditions and resource-consuming exploration would no longer be required. It also shows that all proteins that crystallised in the Filter 2 screen have a better than average propensity to crystallise in these conditions whereas proteins that crystallised in the Filter 4 screen only ever perform below or close to average. For example, at pH 6.5 it

is possible to see a propensity of ~2.5 for Filter 2 and ~0.7 for Filter 4. As the conditions are the same, the difference must be due to the proteins being screened. This shows that quite different results can be obtained, depending on the proteins involved. In fact the results here contradict those displayed in Figure 51 which shows pH 4.5 to have a propensity well below average, whereas here it can be seen that for some proteins pH 4.5 has twice the average propensity.

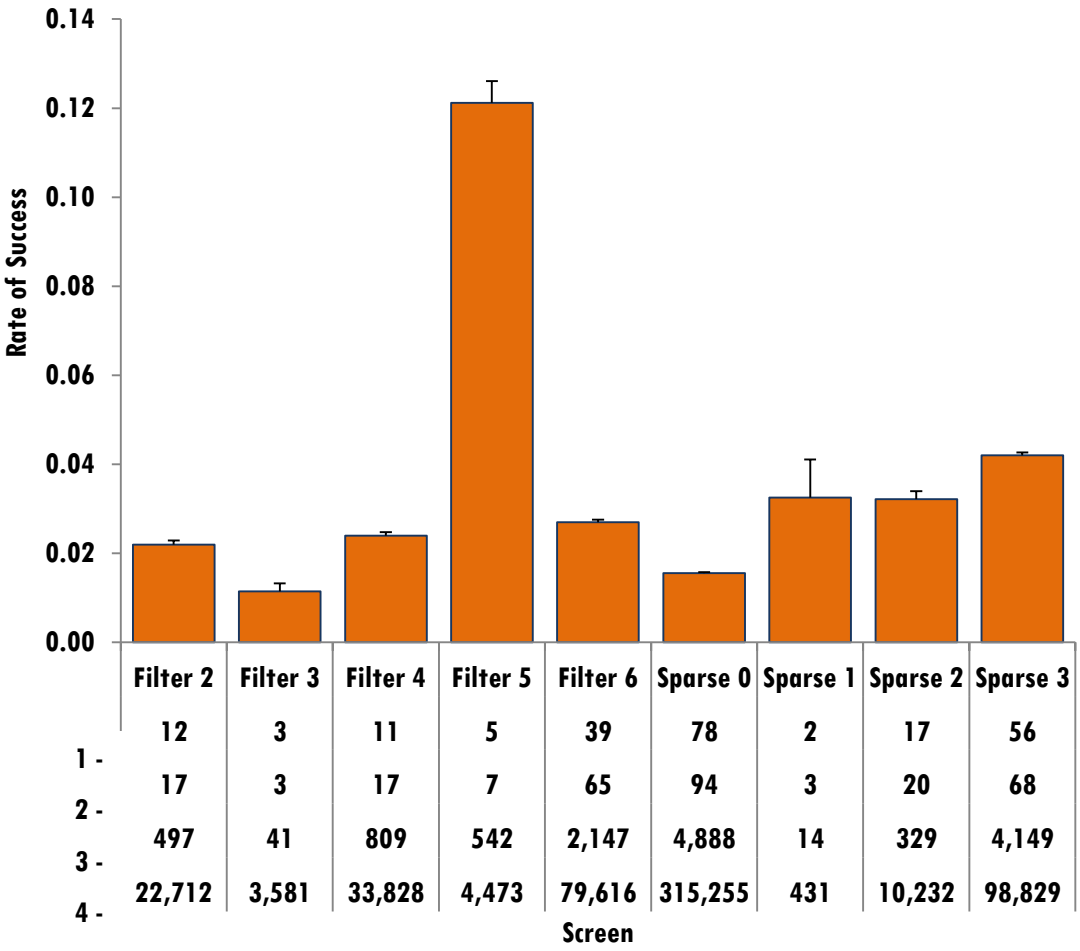


Figure 57: Success rates for AZ screens.

The rate of success for five evolutions of a filter screen and four evolutions of a sparse matrix screen used in initial screening at AstraZeneca are shown. The figures shown are as follows: (1) the number of projects crystallised; (2) the number of projects attempted; (3) the number of wells containing crystalline material; and (4) the number of wells in total. Proportional error bars are shown.

The screens at AstraZeneca have evolved over time as poor chemical combinations have been identified and removed. This should mean that the more recent the screen

is, the more successful it should be. Screen evolution is indicated by the number following the screen type, so that, for example, Filter 3 has evolved from Filter 2. The success rates for evolutions of the filter screen are shown in Figure 57. It can be seen that, in most cases, later screens are indeed more successful and the slight modifications to the screening conditions in the filter screen have been beneficial overall. The rate of success for Filter 2 is ~2% which falls to 1% for Filter 3. However, Filter 3 was only used with three projects so the sample size is too small to determine whether the modification to the screen was really detrimental or whether the result is just due to the proteins involved. Conversely, it is known that Filter 5 has an artificially high success rate of ~12% due to a particular project involving proteins that tended to crystallise easily. A similar pattern can be seen for the sparse screen, with the initial Sparse 0 having a success rate of 1.5% and this being gradually improved to 4% for the latest version, Sparse 3.

7.2. Discussion and Conclusions

7.2.1. Average Success Rate

One issue with the use of the propensity formula is how to calculate the average success rate (AS) and from what data. The problem is that AS is described as the number of successful trials divided by the total number of trials. The total number of trials could include screens that have had no success and failed projects or just include those that have been successful on at least one protein. If a specific screen is being considered, then the total number of trials may mean the number of trials for that particular screen.

The AstraZeneca screens selected for analysis were those that were commonly used. The removal of screens that were only used once or had no positive (successful) results means that we have, therefore, modified our AS.

7.2.2. Hypersensitivity

As most crystallisations trials are unsuccessful, the number of successful trials for each set of conditions is very small. This affects the reliability of the calculated

propensity, which is sensitive to small changes. For example, 2 successful trials out of 15 would give a RS of 0.133 (3 sig. fig.), divided by the average success rate of 0.02 giving a propensity of 6.67 (3 sig. fig.). Having 3 successful trials out of 15 gives a RS of 0.2, divided by the average success rate (0.02) gives a propensity of 10. By our definition of propensity this means that something that was 6 times more likely to crystallise than average is now rated 10 times as likely due to one crystalline result. It is also important to note that the rate of success is calculated from an observed rate and therefore changes from screen to screen.

7.2.3. Protein Dependent Success Rate

The average rate of success calculated depends on the number of trials with specific proteins and can be artificially raised by repeated experiments with those that crystallise easily in many different conditions. For example, one screen involved 7 projects in a total of 4,473 trials, 5 of the projects were successful in 542 of the trials. The rate of success for this screen is therefore ~0.12 (542/4473), which is significantly higher than the ~0.02 obtained for all other screens. Further investigation revealed that one of the 5 successful projects crystallised in 432 of 646 trials. This project has such a high success rate (~0.66) that the overall rate of success for the screen is artificially high. If data related to this project is left out of the calculations, the screen has an average ~0.02 success rate.

A higher rate of success for a particular screen, therefore, does not necessarily mean that it has a significantly higher chance of crystallising a new protein, but can simply mean that a particular project was crystallised numerous times in it.

7.2.4. Conclusion

A comparison of the results obtained for the propensity of pH and chemical species in our study shows some differences with those obtained by Rupp and Wang. For example, PEG 2000 MME is their most successful chemical with a propensity of ~ 2.4, but our results show a propensity of ~ 1.5. Of the 10 chemicals with the lowest propensities in our dataset, only one, MPD, is in the 10 lowest propensities of Rupp and Wang. The top 10 are broadly similar groups, highlighting the fact that PEGs are

highly successful chemicals. However, their presence in a solution can artificially boost the propensity other chemicals in the crystallisation solution.

The complex nature of crystallisation means that changes in one parameter affect the results of another in a nonlinear way. As our data is limited to only specific combinations of chemicals/pH/protein it is difficult, using the propensity metric, to make any solid conclusions about revising the AstraZeneca screens to make them more successful. In order to provide a reliable measure, propensity calculations would need to involve combinations of chemicals, requiring much more data. As well as understanding the properties of the parameter space, it is clear that the properties of the protein also require analysis. As this data was not available to us it was not possible to link specific protein properties with patterns in crystallisation parameter space.

The use of a statistic such as propensity provides an insight into success rates of pH and chemicals within a screen. It can provide statistical evidence that certain chemicals with anecdotal evidence of success, do not actually work. Conversely, it can provide support for chemicals with known success. The AZ crowns were removed from screen because they were observed as poor crystallisation reagents. However, due to the interdependence between chemicals and pH, the poor performance of chemical *x* might simply be because more buffer is required.

8. Minimal Set of Conditions

Crystallisation screens cover a large range of crystallisation parameter space using different methods of sampling, many of which are described in the section below. Some are a systematic sampling of the space and some are designed with a target in mind. Here, we describe our method for a multi-target initial screen, which samples space in a non-systematic manner. We are able apply our algorithm to data from AstraZeneca and the PDB to design screens that, if used from outset, would have crystallised the maximum number of proteins while using the minimal number of conditions. In 1937, Laufberger crystallised ferritin by adding cadmium salts directly onto slices of the horse spleen (Laufberger, 1937, McPherson, 1991). Unfortunately, not all proteins crystallise so readily and are usually screened against various combinations of chemical species at different concentrations and pH in order to identify suitable crystallisation conditions. Carter and Carter (1979) used a factorial approach (Fisher, 1942) to rationalise the process of protein crystallisation screening. A complete factorial design is the systematic sampling of every combination of parameters. For protein crystallisation the number of possible salts, polymers, organic and non-organic solvents, detergents and other additives, at different concentrations, pH and temperatures, makes such a search impossible. Therefore, a method of incomplete factorial design was devised to sample a subset of parameter combinations. Carter and Carter explored 6 parameters: precipitating chemical, pH, temperature, divalent cation, counter anion and counter cation, further broken down into specific chemicals or units. For example, pH was investigated at just 4 levels: pH 4.5, 5.5, 6.5 and 7.5. In total, 35 combinations were used whereas a complete factorial design would have required 4,032. The volume of ethanol is the only difference between the two combinations in the example below and, therefore, only one of them would be included in the *incomplete factorial screen*:

Combination 1

pH 5, 289K, 15% PEG 3350, 200mM sodium chloride, 100mM ammonium sulfate, 5% ethanol.

Combination 2

```
pH 5, 289K, 15% PEG 3350, 200mM sodium chloride,
100mM ammonium sulfate, 10% ethanol.
```

Carter and Carter found polyethylene glycol (PEG) to be favourable for crystal growth, whereas they found sodium and ammonium salts detrimental. This latter conclusion was reached based on one protein, for which the best results were reported as 'single, three-dimensional crystal showing little or no diffraction'. A later, more comprehensive study showed that the best crystals were actually obtained in ammonium sulfate with the previous apparent negative effect on crystallisation explained as misidentification of crystalline precipitate (Carter *et al.*, 1988).

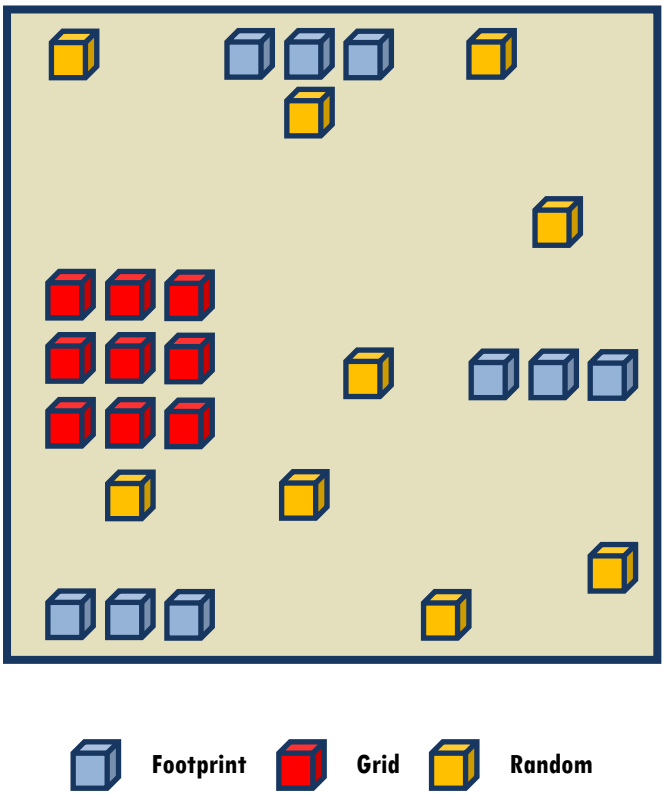


Figure 58: Crystallisation parameter space sampling.
Representations of crystallisation screens that use different sampling methods are shown. The footprint screen is represented by blocks of three rows of three blue cubes spread across parameter space, the grid screen is shown as a block of nine red cubes on the centre left and the random screen is represented by nine yellow cubes spread randomly across parameter space.

The *sparse matrix screen* was designed to sample those conditions known to be favourable for crystallisation rather than provide a systematic sampling of

crystallisation parameter space. Jancarik and Kim (1991) compiled a list of 50 favourable crystallisation conditions, as determined from their own and others' experience. This sampling of crystallisation parameter space is similar to random (non-systematic) sampling as shown in Figure 58. The rationale for the screen design is that, provided approximate crystallisation conditions are found for a particular protein (i.e. result in some form of crystalline material), it is then 'relatively easy to optimise the conditions' to obtain a diffraction quality crystal (Jancarik & Kim, 1991). As more crystallisation experiments are performed, more information becomes available about conditions that can then be used to update the screen. The original screen of Jancarik and Kim included the use of ammonium and sodium salts and half of the conditions contained PEG of various weights. A total of 46 proteins were trialled in this initial screen and all 15 proteins that had previously been crystallised with other screens produced crystals. Of the 31 proteins that had not been crystallised before, 26 produced crystals, giving a success rate of 84%. However, no crystal quality is reported and it is not known whether structures could have been determined from the crystals obtained.

The *grid screen* is described by McPherson (1989b) as a 24-well screen in which the concentration of precipitant is varied across six columns, with the pH varied down four rows (centre left grid, Figure 58). A pH range between 3.5 and 9.0 is suggested, to be reduced or extended as appropriate, with ammonium sulfate or PEG 4000 as the initial precipitant. If sufficient protein is available it is recommended that an organic solvent, specifically ethanol or MPD, also be tested. They also recommend that all conditions be tested at both 4°C and 25°C for comparison. Should a protein still not crystallise the use of complexes, ligands or alternative forms of the same protein could be tried. The screen can be used to determine how precipitant concentration together with pH affects the growth of protein crystals. In contrast to the incomplete factorial screen, which aims to sample crystallisation space as widely as possible, the grid screen involves a very detailed and systematic search of particular regions. A combination of the two methods, i.e. the identification of a region followed by an in-depth search, is thought to be the best approach (Jancarik & Kim, 1991, McPherson, 1992). To achieve this the PACT screen was developed (Newman *et al.*, 2005) as part of a two-screen strategy. The first screen, a sparse matrix, was followed by the PACT screen - a systematic sampling of pH, anions, cations and PEG. Such

sampling allows for insights into regions of parameter space which are favourable to crystal growth, for specific proteins, without obtaining a crystal. The results serve to prove this strategy is successful, crystallising 20/34 (58%) proteins that had never been crystallised before.

Another type of screen, the *footprint screen* (Stura *et al.*, 1992), is a 24-well screen utilising six different precipitants across the columns with pH varied over the four rows (represented by three separate blocks of three blue cubes in Figure 58). Although the choice of precipitants includes PEG 4000 and ammonium sulfate, the authors note that ammonium sulfate is usually avoided at high pH as the release of ammonia can change the pH. The precipitants chosen are those used 'successfully in the crystallisation of many proteins' (Stura *et al.*, 1992).

In the years following the publication of the sparse matrix screen several complementary screens were designed to sample the parameter space not covered by the sparse matrix screen. The screen of Cudney *et al.* (1994) is one such screen designed to use novel chemicals, suggested by their own experience, to 'uncover new additional leads for further optimisation'. In trials with the novel chemicals, crystals were obtained for previously uncrystallised proteins as well as those that had been crystallised before.

The MORPHEUS screen developed by Gorrec (2009) is intended to be used as an alternative initial screen. Gorrec reports that even with 40 commercial screening kits covering over 1,500 conditions many proteins still do not produce diffraction-quality crystals. Even so, many of the conditions are repeated across commercial screening kits (Wooh *et al.*, 2003). The MORPHEUS screen contains ligands and additives such as amino acids in the crystallisation solution, although Gorrec also notes that the inclusion of ligands and additives can sometimes have a detrimental effect. With successful crystallisation trials reported for previously uncrystallised proteins, the MORPHEUS screen shows that there is still potential for alternative initial screens.

Others have tried to improve the efficiency of crystallisation screens particularly for situations where limited protein is available. Brzozowski and Walton (2001) created a pair of screens both containing just 24 wells, a reduction of 75% of the standard 96-

well plate. Their first screen was designed to improve crystallisation success for enzymes, the main protein group studied in their laboratory. Using this screen, they crystallised proteins that had not previously crystallised in commercially available screens. In their second screen they developed a complementary set of conditions with different sections of the screen containing a specific chemical species. The screens are designed to allow the user to incorporate any available information about the protein or to include particular preferred conditions.

8.1. The Most Efficient Screening Method

The rate of success, that is the number of proteins crystallised divided by the number trialled, varies from screen to screen. A comparison of the three different major screening methods was undertaken by Segelke (2001) who tested the random, footprint and grid screens with five proteins. The results of five proteins (four of which are commercially available) suggest that the most successful screen type is the random screen. Results from the TB SGC show that random screening can also be used to obtain diffraction quality crystals without optimisation (Rupp, 2003).

Here a similar analysis of the AstraZeneca dataset was performed to investigate the efficiency of their different screen types. Using data from two screen types - filter (a mix of footprint and grid screens) and sparse (a sparse matrix screen) we show that sparse matrix sampling is also the most efficient method for the proteins studied at AstraZeneca (Table 16).

	Screen	Projects Trialled	Projects Resulted in a Score of 4, 5 or 6	Projects Resulted in a Score of 6	RS Score of 4, 5 or 6 (%)	RS Score of 6 (%)
C1	Sparse	151	131	80	86.75	52.98
	Filter	88	60	37	68.18	42.05
C2	Sparse	87	77	53	88.51	60.92
	Filter	87	58	37	66.67	42.53

Table 16: Comparison of screens' success.

Here, we show the rates of success (RS) for each screen type, filter and sparse. Comparison one (C1) accounts for the projects used in the named screening type; comparison two (C2) only takes into account projects used within both screen types. The rates of success are calculated by the number of projects with a particular crystalline annotation (4, 5, 6) divided by the number of projects trialled in that screen type.

The first comparison (C1) involved any project that had been trialled in the named screen type, whereas in the second comparison (C2), only projects tried in both screen types was included in the analysis. Both comparisons show that the sparse screens produce crystalline results for about 20% more proteins than the filter screens. Furthermore, high quality crystals are obtained for between 10% and 20% (for C1 and C2 respectively) more proteins with the sparse screens.

3	3	3	3	3	3	3	4	4	3	3	3
2	3	2	1	0	2	0	1	3	0	1	0
2	5	6	13	9	5	1	2	5	7	4	4
2	4	10	10	6	7	2	3	6	6	2	2
1	3	3	2	2	2	3	3	3	2	0	2
4	8	11	16	13	9	4	8	9	9	7	4
6	9	9	9	7	6	5	5	10	9	7	5
1	7	12	14	12	8	2	7	11	10	11	8

Figure 59: Heat plot of proteins crystallised in Filter 6 (59).

The figure shows the number of projects crystallised by each condition in the Filter 6 screen, for which the mean number of projects crystallised is 5.28. The most successful (shown in yellow) crystallised 16 projects and the least successful conditions (shown in white) did not crystallise any projects.

As sampling in a non-systematic manner is the most efficient method for initial protein screens, the question becomes how to select such conditions. Jancarik and Kim (1991) observed that many proteins would crystallise in several conditions and so it follows that many conditions could be used to crystallise several proteins. Analysis of a set of 59 projects that were trialled in every condition of the Filter 6

screen also shows this (Figure 59). The mean, median and mode number of projects crystallised in each condition are 5.28, 4.5 and 2 respectively. A total of 16 projects were crystallised in the most successful conditions whereas some conditions did not produce crystals for any projects. It comes as no surprise, following propensity analysis, that the most successful condition contains PEG 3350 and ethanol buffered at pH 8.5. Generally the most successful rows of the screen (shown in red) are rows 6 to 8, all of which contain PEG. Conversely, the least successful conditions contain either MPD buffered at pH 9.5 or tetramethylene sulfone at varying pH. Rows 2 and 5 containing these chemicals are shown in the palest colours in Figure 59.

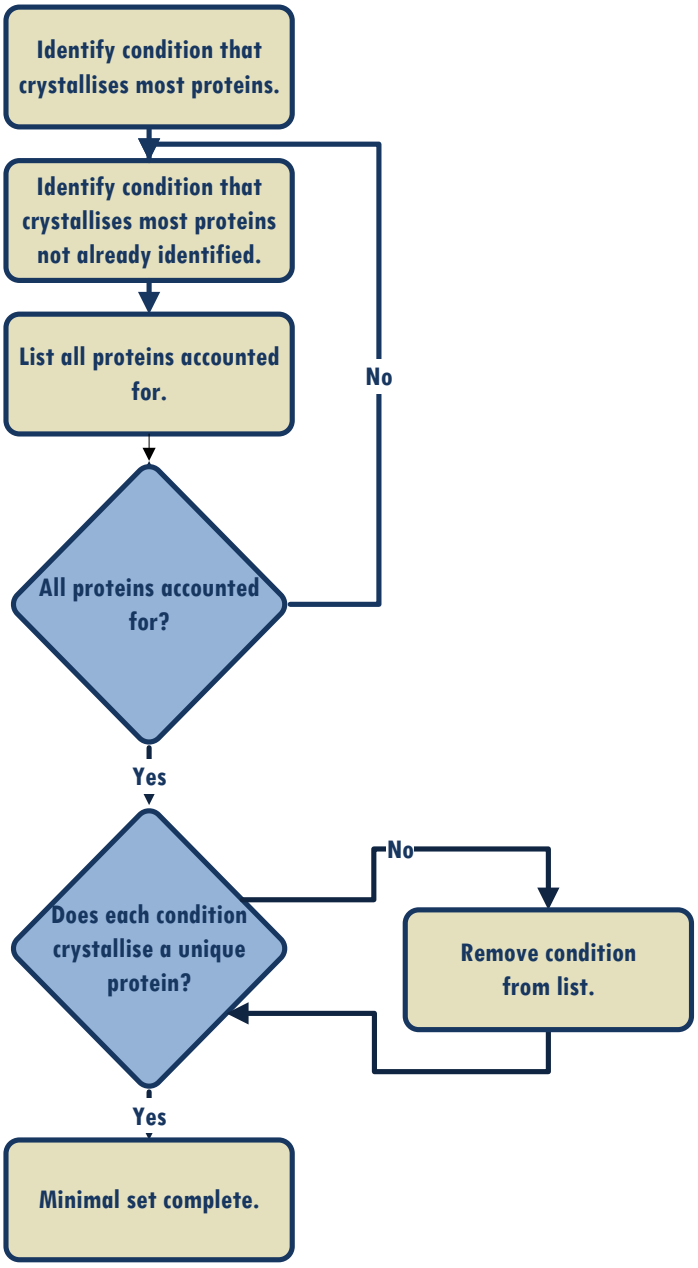


Figure 60: Flowchart for implementation of the minimal set algorithm.

As some conditions crystallise numerous proteins, the minimal number of conditions in which all the successful projects have crystallised can be calculated. This NP-hard problem can be solved using mathematical minimal set theory (Karp, 1972). A problem is NP-hard if the algorithm for solving it can be translated into one that could solve an NP-problem. NP refers to nondeterministic polynomial, which means is can be solved by a nondeterministic Turing machine in polynomial time (the time taken to provide a solution is a polynomial function of the number of inputs).

	Project ID								
Well	A	B	C	D	E	F	G	H	Sum
1	♦	♦	♦	♦					4
2	♦					♦			2
3		♦			♦			♦	3
4				♦					1
5			♦	♦			♦		3

♦ | Crystal

Figure 61: Example of a minimal set algorithm.

The figure shows which of the five conditions (wells 1-5) the eight projects (A-H) produced crystals. The algorithm to find the minimal set of conditions is as follows:

1. Identify the well in which most projects crystallise: this is well 1 for projects A, B, C and D.
2. Find the well in which most projects not already crystallised in well 1 crystallise: this is well 3 for projects E and H.

   Six of the eight projects are now accounted for (A, B, C, D, E and H). Conditions only need to be found for project F and G.
3. Add well 2 in which project F crystallises and well 5 in which project G crystallises and the set is complete.
4. Remove any members of the set that do not contribute something unique: remove well 1 as projects A, B, C and D also crystallise in wells 2, 3 and 5.

The minimal set required for crystallisation of the 8 projects consists of wells 2, 3 and 5.

A greedy algorithm, for which the flowchart is shown Figure 60, was employed to ensure a solution in a reasonable amount of time. A greedy optimisation algorithm chooses the optimal solution at each step but may not result in the overall optimal solution, as this may not be the sum of the optimal parts. The example in Figure 61 shows the implementation of a minimum spanning set across conditions in which proteins crystallise. The first step in the algorithm is to identify the condition that crystallises the most proteins and save this to a set MSET. The next step identifies the condition that crystallises the most proteins not already accounted for by MSET and is repeated until (a) either there are enough conditions to fill a crystallisation screen or (b) all possible proteins are accounted for. At this stage, each condition in MSET is checked to ensure it crystallises at least one protein that another condition in the set does not.

8.2. AstraZeneca Minimal Sets

For each screen (96 conditions/wells) used to produce the AstraZeneca data, a minimal number of conditions that gave crystals for all projects that could be crystallised was found.

Table 17(a) shows that the minimal set that gave crystals for the most projects per condition, is that of Sparse 0 with an average of 4.33 projects crystallising in each condition and one condition (containing PEG 8000 and calcium acetate) giving crystals for 33 different projects. Conversely, just one project crystallised in each condition in the Sparse 1 screen. Table 17(b) shows the results for minimal sets calculated over those projects that were trialled in every condition. This also shows that, on average, the Sparse 0 screen minimal spanning set gives crystals for the most projects per condition (4.8) and again Sparse 1 is shown to be the least efficient with one condition per project in the minimal set. This result could be expected as there are fewer projects and therefore a smaller minimal set with higher redundancy. For each screen, at least one condition produced crystals for ~40% of the projects trialled. In the most extreme case of Filter 4, one condition gave crystals for 7 out of the 11 proteins (~64%). Kimber et al (2003) also showed that a single condition could give crystals for a high percentage of proteins, with 99 successes out of 338 (~29%). The redundancy rate is the percentage of wells not included in the minimal

and ranged from 81% for Sparse 0 up to 98% for Filter 3. Page *et al.* (2003) also reported high redundancy rates of up to 77%, with 23% of conditions giving crystals for all proteins trialled (Page *et al.*, 2003).

Screen	Minimal Set Size	Maximum*	Redundancy (%)
Filter 2	4	8	96
Filter 3	2	2	98
Filter 4	3	7	97
Filter 5	3	3	97
Filter 6	10	16	90
Sparse 0	18	33	81
Sparse 1	2	1	98
Sparse 2	6	6	94
Sparse 3	16	20	83

(a) The minimum number of conditions required for crystallisation of all projects.

Screen	Minimal Set Size	Maximum*	Redundancy (%)
Filter 2	3	7	97
Filter 3	2	3	98
Filter 4	3	8	97
Filter 5	2	4	98
Filter 6	8	36	92
Sparse 0	15	73	84
Sparse 1	2	2	98
Sparse 2	6	16	94
Sparse 3	16	56	83

(b) The minimum number of conditions required for crystallisation of all projects that were trialled in every condition of the named screen.

Table 17: Minimal sets for AstraZeneca projects.

Tables (a) and (b) show the size of the minimal sets and redundancy rates for AZ screens. The results for Table (a) included any project that had been trialled in that screen, whereas

the results for Table (b) were calculated using projects that had been trialled in all 96 wells of the screen.

* This is the maximum number of projects that could be crystallised in one condition.

To improve efficiency, the number of conditions trialled within the screen could be reduced by only including a minimal set of conditions. Replacing the other conditions with new ones could improve the chances of crystallising more proteins. The number of projects crystallised has a strong correlation with the size of the minimal set. As the number of projects trialled with a screen increases, the number of conditions included within the minimal set also increases (redundancy decreases). To investigate how successful the conditions within the minimal spanning might be with a new project, it is necessary to test the stability of the size of the minimal spanning set. Stability is defined as the percentage of projects required for the minimal set size to stop changing i.e. the fewer proteins required for the size of the minimal set to stabilise, the greater its reliability. It is likely that *project n* and *project n+1* are proteins which are chemically similar, as typically each protein modification is assigned a new project number and this may bias the stability of the set. To reduce any such bias, the order in which the projects were included was randomised.

For the first project on the project list we randomly select the condition in which it crystallised and form the list of conditions, *MSET*. The remaining projects for the same screen are checked to see if they crystallise in this condition. If they do, they are removed from the project list. Another project is then randomly selected from the project list and the conditions in which this project crystallised are added to MSET and checked to see if projects already removed from the project list crystallise in these conditions. If this is not the case then the project is simply removed from the project list. Otherwise, the minimal set algorithm is run on MSET, to determine the smallest subset of MSET that will crystallise all projects removed from project list. This process is repeated until all projects are removed from the project list. The size of MSET might stop increasing before all projects are removed from the project list. The percentage of projects that have been removed at this point is indicative of MSET stability.

Table 18 shows that a high percentage of projects is required for the size of the minimal set to stabilise. The percentage of projects required ranges from 75% for

Filter 5 up to 100% for Filters 2 and 3 and Sparse 1. It is highly probable that the order of testing affected the percentage of projects required for the size of the minimal set to stabilise. This can be demonstrated by the following example: consider ten projects of which nine crystallise in one condition and one project requires a separate condition. Starting the minimal set derivation with the project that requires a unique condition gives an initial of one condition. The other nine projects do not crystallise in this condition and so a project from the remaining nine is added and the size of the minimal set is increased to two. It would then be found that the remaining eight projects crystallise in these two conditions so that the percentage of projects required for the size of the minimal set to stabilise is two out of ten (20%). However, if the order was changed so that each one of the nine projects that crystallise in the same conditions, were added before the project that requires unique conditions, all ten projects would need to be tested before the minimal stabilised and produced crystals all projects.

Screen	Projects Required	Projects Crystallised	Projects Required (%)
Filter 2	7	7	100
Filter 3	3	3	100
Filter 4	7	8	88
Filter 5	3	4	75
Filter 6	35	36	97
Sparse 0	60	73	82
Sparse 1	2	2	100
Sparse 2	15	16	94
Sparse 3	54	56	96

Table 18: Number of projects required for minimal set size to stabilise.

The minimal set of conditions changes depending on which projects are sampled. It is possible that a minimal set to crystallise all projects is found before all projects have been trialled. Here, the number of projects required for the minimal spanning set to stop changing size is shown in Projects Required.

Further analysis was undertaken to determine the number of projects that could not be crystallised with a minimal set of conditions derived from the other projects

within the same screen. For every screen, each project was removed in turn and a minimal set of conditions found for the remaining projects. It was then determined whether or not the removed project could be crystallised in this minimal set. Table 19 shows the number of projects, *S*, for each screen that when removed could be crystallised in the minimal set of conditions derived from the remaining projects.

To establish whether it would be more efficient to adopt a minimal set for all projects, then use an extra screen to crystallise those that will not crystallise in the minimal set, requires calculation of how many conditions would be needed to use this method in comparison to screening everything with the standard 96 conditions (the standard screening protocol).

For the minimal set to be the most efficient method, the following condition has to be met:

$$PM + 96(P - S) < 96P \qquad \textbf{30}$$

where *P* is the number of projects, *S* is the number of projects that can be crystallises in a minimal set derived from other projects and *M* is the number of conditions within the minimal set. Rearranging equation 30 gives:

$$\frac{\boldsymbol{M}}{\boldsymbol{96}} < \frac{\boldsymbol{S}}{\boldsymbol{P}} \qquad \textbf{31}$$

which shows that it is more efficient to use the minimal set provided its size, when divided by 96, is less than the number of successes divided by the number of projects.

Table 19 shows that, for eight out of nine screens, it is more efficient to use a minimal set of conditions as the crystallisation screening protocol. Sparse 1, with only two projects trialled, is the only screen where this is not the case, no single condition gave crystals for both projects.

Screen	Successes (S)	Projects (P)	P-S	(%)	Minimal Set Size	Apt*
Filter 2	3	7	4	57	3	Yes
Filter 3	1	3	2	67	2	Yes
Filter 4	4	8	4	50	3	Yes
Filter 5	2	4	2	50	2	Yes
Filter 6	30	36	6	17	8	Yes
Sparse 0	58	73	15	21	15	Yes
Sparse 1	0	2	2	100	2	No
Sparse 2	11	16	5	31	6	Yes
Sparse 3	43	56	13	23	16	Yes

Table 19: Minimal set of conditions efficiency.

The number of projects that can be crystallised in a minimal set derived from the other projects trialled with the same screen. If a project would not crystallise in a minimal set derived from other projects, then standard screening protocol is resumed. *Apt means is it more appropriate to use the minimal set?

8.2.1. Minimal Set for Combined Screens

At AstraZeneca, nine crystallisation screens, covering 281 conditions were used with 152 projects. Of these 152 projects, 134 were successfully crystallised. Using a greedy algorithm, as described previously, we were able to obtain a set of 27 conditions in which the 134 projects could be crystallised. Where two or more conditions contribute the same number of projects, the algorithm chooses the one it finds first. To examine the effect this might have the algorithm was run 1,000 times. For 583 runs the minimal set was of size 27, for 382 runs the size was 28 and for 35 runs the size was 29.

The minimal set with conditions in which the most projects could be crystallised was taken as the optimal minimal set. As Figure 62 shows, two different minimal sets (Set One and Set Two) can be used to crystallise all six projects with three wells. The sum of the parts for Set One is 2+4+3=9 and the sum of the parts for Set Two is 2+2+3=7. In this instance Set One would be chosen as the sum of the parts is greater.

		Project Number						
	Well	1	2	3	4	5	6	Total Crystallised
Set One	1	♦	♦					2
	2		♦	♦	♦	♦		4
	3				♦	♦	♦	3
Set Two	4	♦	♦					2
	5		♦	♦				2
	6				♦	♦	♦	3

♦ | Crystal

Figure 62: Comparison of two minimal sets.

An example of two minimal sets, both consisting of 3 wells where one set (Set One) gives more crystals overall than the other.

For the *all projects minimal set,* the sum of parts was 640 projects, with two very successful conditions each giving crystals for 48 projects and the least successful condition producing crystals for eight projects. Over 50% of projects can be crystallised using just two conditions (one containing PEG 8000 and calcium acetate and the other PEG 3350 and ethanol), 75% in six conditions, 85% in ten conditions and 95% in 21 conditions. Of the 27 conditions, 15 only contributed one project. Of the 134 projects, 14 crystallised in only one condition, and 10 of these 14 are the only project associated with this condition in the minimal set. Similarly Kimber *et al.* (2003) found that 27% of their proteins would crystallise in 6 conditions, 36% in 12 and 42% in 24.

Some conditions are found to have components that are over-represented in the minimal set or under-represented in the original 281 conditions from which the minimal spanning set is derived. For example: PEG 10000 is used in 16 of the 281 conditions and therefore, in a minimal set of 16 conditions, one might expect PEG 10000 in two (16/(281/27)), but it is only observed once. Conversely, sodium chloride is found in 21 of the 281 conditions and might be expected to be in 2

conditions of the minimal set. It is, however, found in 4 conditions of the minimal set.

Chemicals included more frequently in the minimal set than might be expected are PEGs-400,500 MME, 2000MME and 8000, ethanol and isopropanol and calcium acetate among others. In Chapter 7 the properties of such chemicals was discussed. Their inclusion in successful conditions is due, for example, to the precipitating effects of PEGs, the bonding of calcium-based compounds and the inclusion of additives such as ethanol as a cryoprotectant. Some of the chemicals found to be under-represented are: MPD, PEGs 10000, 6000 and 1500, magnesium acetate, ethylene glycol and tetramethylene sulfone.

A new screening protocol at AZ might allow crystallisation space to be sampled in a manner that maintains the same ratios of chemicals in the minimal set. For example, it is known that sodium chloride is associated with conditions that crystallise numerous projects or ones that do not crystallise elsewhere, therefore, the inclusion of more conditions containing sodium chloride might prove beneficial.

8.2.2. Filling Up the Screen

As crystallisation plates are typically made up of 96 wells, the use of a 27-condition screen leaves 69 wells empty. These wells could be used to repeat the 27 conditions twice more as it is known that repetition is beneficial in crystallisation as nucleation is a stochastic process (Newman *et al.*, 2007) or used to further explore chemical parameter space. Previously we showed that the more sparsely spread across parameter space the chemical conditions are, the more successful the screen is. To fill the remaining 69 positions of the screen it would be ideal to use conditions that are as chemically different as possible, in terms of their influence on crystallisation.

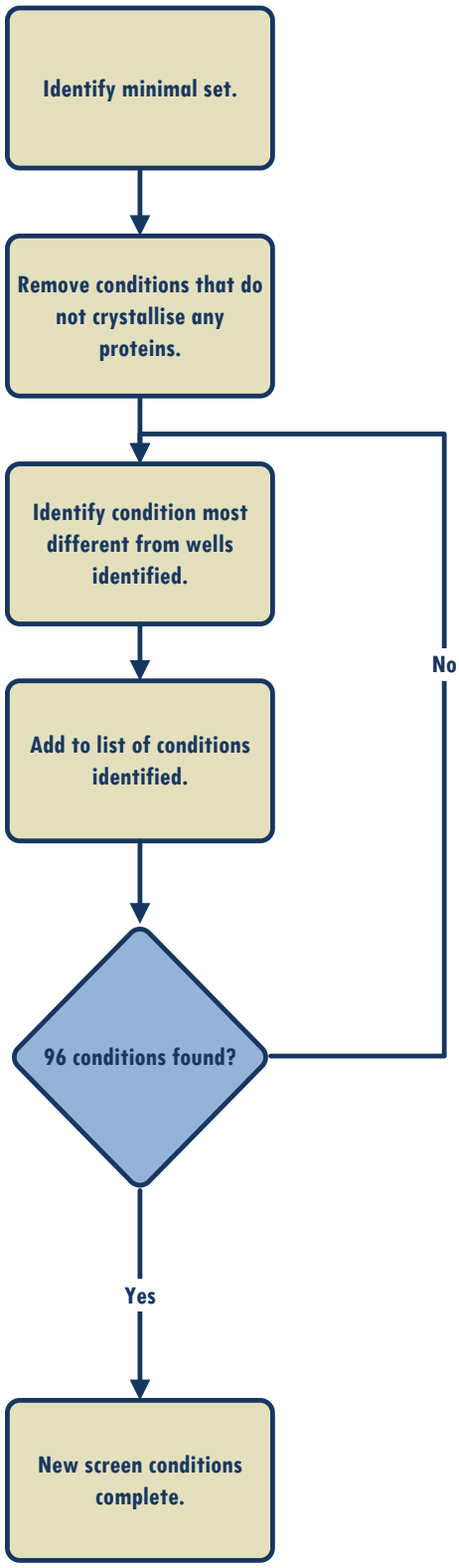


Figure 63: Identifying new crystallisation conditions.

The flowchart shows the process to identify new conditions to improve the sampling of crystallisation space.

The C6 metric (Newman *et al.*, 2010) and CD*coeff* (Bruno *et al.*, 2014) are two measures that provide a similarity metric between the contents of a crystallisation solution. We can use such a metric to determine the most diverse conditions in the 281 conditions used at AstraZeneca. However, this might identify two solutions that are chemically different, but they may not be useful as crystallisation reagents, used only where the outcome was unsuccessful. Our best measure of how similar conditions were in terms of their propensity to crystallise is given by which projects crystallise in which conditions.

A data matrix $\boldsymbol{X}$ was formed in which $x_{ij} = 1$ if project *i* crystallised in condition *j,* and $x_{ij} = 0$ otherwise. This allowed the Hamming distance between conditions to be calculated, giving the most similar conditions a score closer to zero and the most different a score closer to one. With a core screen made up of 27 conditions obtained through the minimal set algorithm, the other 69 conditions could be chosen as follows. The next condition would be the one that (a) crystallised a project and (b) was most different, according to the sum of Hamming distances, to the 27 conditions already identified. The process of identifying the new conditions for the screen is, again a greedy process, shown in Figure 63. After obtaining 96 conditions, the C6 metric can be used to analyse the diversity of the new screen in comparison to previous screens. The 'internal diversity score' on the C6 web tool can be used to determine the spread of conditions (Newman *et al.*, 2010). This score is obtained from the average of the pairwise C6 distance scores for conditions within a screen and ranges from 0 (identical) to 1 (completely different). The new screen scored 0.75. To put this in perspective, the internal diversity scores for commercially available sparse matrix screens are around 0.9. It is, therefore, possible that drawing from a limited sample of conditions and with one quarter (27/96) of the screen being fixed, contains bias towards highly similar conditions.

8.3. Minimal Set for the Protein Data Bank

Analysis of the PDB data shows that one crystallisation solution (30% PEG 4000, 0.1M tris pH 8.5, 0.2M magnesium chloride,) occurs 90 times, with the successful crystallisation of 69 different proteins (i.e. from different groups in the PDB-BLAST dataset). Minimal set analysis of the PDB dataset shows that only 5,683 unique

conditions are required to crystallise all proteins, whereas 35,389 different conditions were actually used, giving a redundancy rate of 84%.

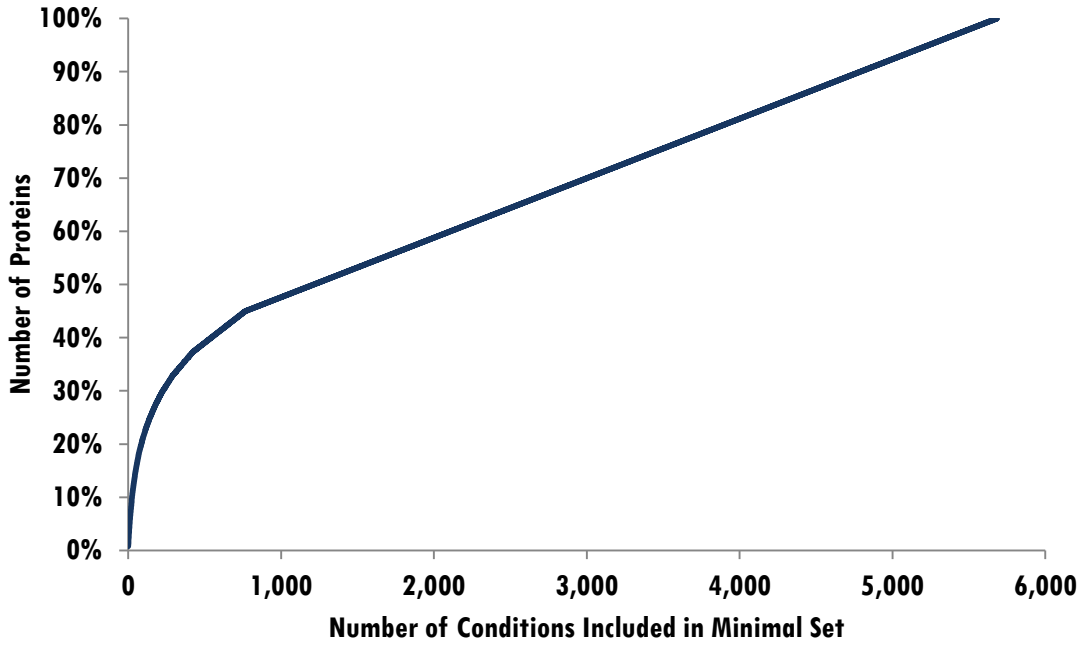


Figure 64: The number of proteins required for minimal set.

The number of proteins in the PDB is plotted against the numbers of conditions used to crystallise them. The growth of the minimal set becomes linear after 768 conditions have been included.

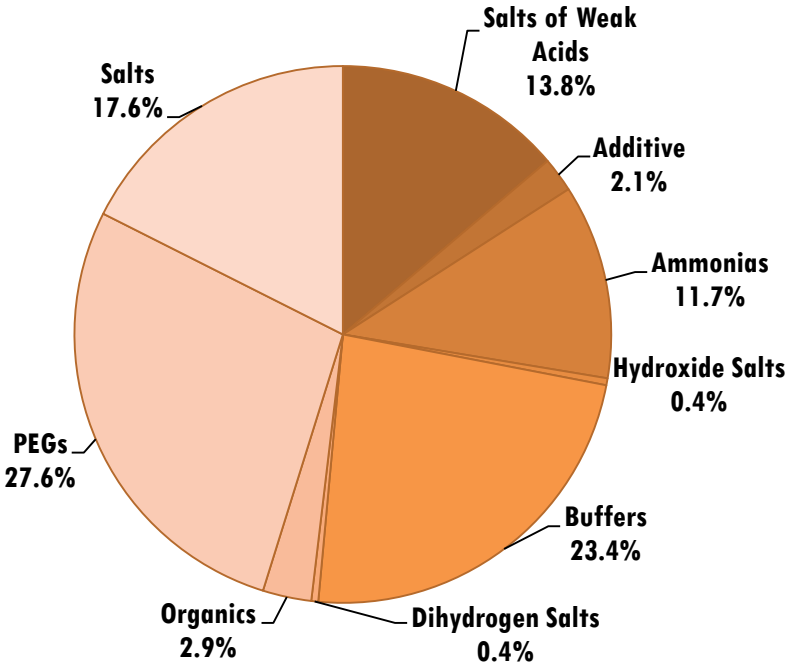
Figure 64 shows the number of proteins that can be crystallised by different numbers of conditions. After 768 conditions are included, each protein not already accounted for requires its own conditions (according to the data in the PDB) so that the minimal set grows linearly with the number of proteins. The solutions in the minimal spanning set of 96 conditions (one screen) determined from the PDB data are listed in Appendix B. These conditions have been used to crystallise 1,905 of the 8,937 proteins in the PDB (~21%). The similarity of sets of conditions can be assessed using the C6 metric (Newman *et al.*, 2010). The C6 score has been used to show that 1,795 entries in the PDB have similar conditions to those in the MCSG_1 screen, making this the most successful commercial screen. The 96 conditions in the minimal set derived here together match 2,929 entries in the PDB. The minimal set screen obtained from the PDB data has an internal diversity score of 0.93, suggesting

a very good sampling of crystallisation parameter space. This compares with the Hampton Index, MCSG_1 and Rigaku Wizard screens with diversity scores of 0.91, 0.91 and 0.94 respectively. Currently (Spring 2015), the screen designed using data from the PDB is in trials in the York Structural Biology Laboratory at and AZ.

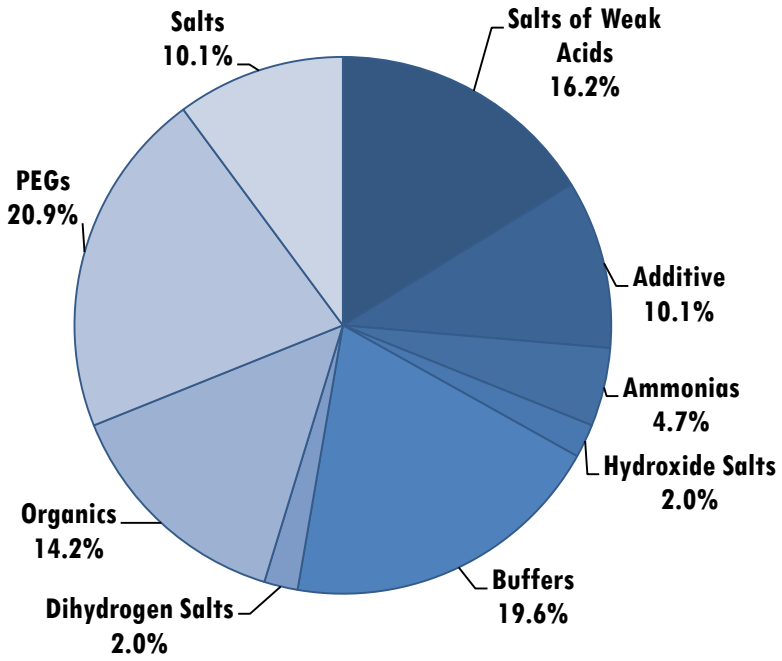
8.3.1. Minimal Sets for Acidic and Basic Proteins

Separate minimal sets were developed for proteins with acidic and basic isoelectric points. Proteins associated with BLAST groups that contained both acidic and basic proteins were removed, leaving 4,695 acidic and 2,125 basic proteins. Of these, the 96 condition minimal set for the acidic proteins could crystallise 19% (915/4,695) and the basic, 16% (350/2,125). The average predicted pH for the minimal sets were 6.73 and 6.67 for acidic and basic respectively. The most productive solution for acidic proteins contained magnesium chloride, PEG 3350 and bis-tris with a predicted pH of 5.5, whereas for basic proteins, ammonium sulfate, PEG 400 and HEPES with a predicted pH of 7.35 was the most productive. The most productive condition in the acidic minimal set is the eleventh in the basic minimal set; the most productive condition in the basic minimal set is the seventeenth in the acidic minimal set.

Figure 65 shows the composition of the minimal sets when divided into the chemical groups described in Chapter 5. The type most sampled group for both sets is PEGs (27.6% for acidic and 20.9% for basic) and the least sampled group is the dihydrogen salts (0.4% for acidic and 2% for basic). The largest difference between the groups sampled in the two minimal sets is for the organics, with 14.2% of chemicals sampled in the basic minimal set being organic but only 2.9% in the acidic minimal set. Conversely, salts account for 17.6% of the species used in the acidic minimal set but only 10.1% in the basic. The acidic minimal set samples 56 different chemical species and the basic, 49. For the chemical species that the two sets have in common, 54 are sampled at the same concentration, the acidic set samples 40 concentrations that the basic does not and the basic set samples 54 concentrations that are not in the acidic set.



(a) Types of chemicals for acidic minimal spanning set.



(b) Types of chemicals for basic minimal spanning set.

Figure 65: Types of chemicals occurring in different minimal sets.

The chemical species found in two minimal sets (a) acidic and (b) basic are shown grouped as described in Chapter 5.

8.4. Discussion and Conclusions

We have been able to confirm that non-systematic sampling of protein crystallisation chemicals is the most efficient method of initial screening. With this information, new initial screens were designed using minimal set theory, one for data at AstraZeneca and three for the PDB.

There are, however, alternative methods of obtaining a minimal spanning set. Some of these provide exact solutions. One such solution for was obtained using SCIP (Achterberg, 2009) for the AstraZeneca data. This solution was 26 conditions, whereas the custom algorithm suggested 27 conditions. Of 26 conditions, 20 were identical to those in the custom solution and another 3 contained the same species but a different buffer pH. For larger datasets, if the problem cannot be solved exactly there exist other approaches that approximate the solution (Paschos, 1997).

Minimal set screens could be particularly useful when the amount of protein purified is limited. The use of conditions that crystallise most proteins first should maximise the chances of crystallisation success. Using a minimal screen followed by a more specific screen for any proteins that do not produce crystals, could provide an efficient method of screening. Since the number of conditions required never stops growing, minimal sets can only be developed post-crystallisation with the hope they will work for a new protein.

It has been observed that conditions are sometimes missed due to dispensation errors and those on the edge of a plate are vulnerable to desiccation if they are not sealed properly before being stored in a temperature controlled unit. Using repeated experiments it may be possible to determine any long term effects of these physical parameters. Minimal sets should also be updated perhaps every couple of years to account for changing tastes in crystallisation reagents (Jancarik & Kim, 1991) due to the requirement for more and more complex proteins (Aloy & Russell, 2006). It is possible to determine bespoke minimal sets for particular subsets of proteins, for example, minimal sets for acidic and basic proteins. There is the potential to create different minimal sets for specific protein families as it has been reported previously

that certain families prefer certain conditions (Samudzi *et al.*, 1992, Hennessy *et al.*, 2000).

9. Shrinking Crystallisation Parameter Space

The “instant physician” is a neural network that was trained using data from patient records, including their symptoms, diagnosis and treatment. When new patient symptoms are entered, the network returns the diagnosis and recommended treatment (Maithili *et al.*, 2011). The symptoms are analogous to protein features and the diagnosis represents the conditions in which a protein will crystallise. Many of the protein features are redundant through being highly correlated with others or having little-to-no discriminatory power. It has been shown that a set of 13 features can provide good discrimination between crystallisable and non-crystallisable proteins. The assumption that the protein features and the list of conditions are all inputs and that the output is either successful or unsuccessful crystallisation creates a problem. As most crystallisation data (98% of the AstraZeneca data) is associated with experiments where the outcome is unsuccessful, a large bias is introduced. A classifier could perform well in terms of accuracy simply by predicting that every combination of protein and crystallisation features would fail crystallisation but this would not be very useful. Ideally, a classifier would be trained using equal class sizes.

By defining which chemical combinations we consider to be chemically similar, meaning that they give similar experimental outcomes, it is possible to cluster conditions. Provided one condition in the cluster crystallises a protein we can assume the other ones would if the experiment was repeated. Using custom crystallisation experiments we are able to investigate the theoretical distance, C6 metric, for defining similarity in crystallisation chemical parameter space. Using data from AZ and the SGC we test the C6 against a modification, C8.

9.1. The C6 Metric

The C6 web tool (Newman *et al.*, 2010) includes a distance metric, referred to as C6, to compare the contents of crystallisation conditions *i* and *j*, providing a theoretical distance, D_{ij}, between the two in chemical parameter space. Where D_{ij} = 0 if conditions are identical and D_{ij} = 1 if the two conditions have no chemical species in common. In all other cases it is defined by:

$$D_{ij} = \frac{1}{T+3}(\alpha + \beta + \gamma + \delta) \quad \textbf{32}$$

where,

$$\alpha = \sum_{t=1}^{T} \frac{\left|[s_{ti}] - [s_{tj}]\right|}{\max[s_t]} \quad \textbf{33}$$

$$\beta = \min\left(1, \frac{\left|PEG_i - PEG_j\right|}{\frac{\max[PEG_i] + \max[PEG_j]}{2}} + 0.2\right) \quad \textbf{34}$$

$$\gamma = \min\left(1, \frac{\left|ion_i - ion_j\right|}{\frac{\max[ion_i] + \max[ion_j]}{2}} + 0.3\right) \quad \textbf{35}$$

$$\delta = \frac{|pH_i - pH_j|}{pH\ range} \quad \textbf{36}$$

In every applicable term, max[*x*], is the maximum concentration of *x* obtained from a collection of commercial crystallisation screens. In the normalising factor, (1/T+3), in Equation 32, *T* is the number of distinct chemical species in conditions *i* and *j*. The first term, α (Equation 33), compares the concentrations of matching species, where $[s_{ti}]$ is the concentration of chemical *t* in condition *i*. The molecular weights of any PEGs in both conditions are taken into account in the second term, β (Equation 34). If the molecular weights are within a factor of two, they are considered similar. For example, PEG 400 and PEG 600 are considered similar, but PEG 400 and PEG 4000

are not. Here, $[PEG_i]$ is the PEG concentration of PEG in condition *i*. The penalty, 0.2 is added to compensate for the fact the distant metric is qualitative. If this term is greater than one then, then β is set to one. If the two conditions contain an anion or cation in common, they are considered to be more similar. This is accounted for by the third term, γ (Equation 35). For example, sodium chloride and lithium chloride are considered similar as they both contain chloride. Similarly, ammonium sulfate and ammonium acetate both contain ammonium. In this term, $[ion_i]$ is the concentration of the ion in condition *i* and 0.3 is the added penalty. Again, if $\gamma > 1$ the term is set to one. The final term, ∂ (Equation 36) is the absolute difference between the pH of both solutions (taken from the buffer pH or that of the predominant component) divided by the pH range for crystallisation.

Manually entering every pair within 96 conditions into the C6 online web tool was impractical. Moreover, some of the distances obtained were not as expected given the equation. After obtaining the code from the authors, it was possible to see that the implemented distance measure differed somewhat from the published equation (Newman *et al.*, 2010). Along with the four terms discussed above, the measure implements further penalties for differences between conditions.

9.1.1. Investigation of the C6 Terms

To investigate any similarities between PEG weights and various ions used in crystallisation, custom crystallisation screens were created. Each screen was trialled with 11 of the proteins described in Chapter 2 (excluding the glycolytic proteins). All of the proteins were obtained from the same batch, which had been frozen and were defrosted to room temperature immediately before use. The conditions were buffered using PCTP at a pH close to that of the protein's isoelectric point. The crystallisation method was vapour diffusion, sitting drop, stored at room temperature. The screens were manually assessed after 21 days.

To investigate similarities between PEGs, a custom screen was produced with PEGs of various molecular weights, various functional groups (mono-, di- methyl ethers), ethylene glycol and tetramethylene sulfone, each in various concentrations. Ignoring pH and molecular weight, the most successful concentrations were found to be 14%,

20% and 26%, each crystallising 9 of the 11 proteins. The same nine proteins crystallised at both 20% and 26%, and eight of these also crystallised at 14%. Conversely, the most extreme concentrations were the least successful, with just one protein being crystallised in any condition with PEG concentrations of 1%, 52%, 77% or 90%. Grouping PEGs according to the maximum concentration at which they are soluble, shows that PEGs with a weight of 10,000Da or 20,000Da, soluble to a concentration of 27% (w/v) are more successful at concentrations of 8%, 12% and 15%. Those PEGs that have a molecular weight less than 10,000Da, soluble up to 45% (w/v), are more successful at 20% and those that can be used up to concentrations of 90% are more successful at 26%. In all instances no single concentration could be used to crystallise all proteins but a selection of 14%, 20% and 39% would ensure that all proteins were crystallised.

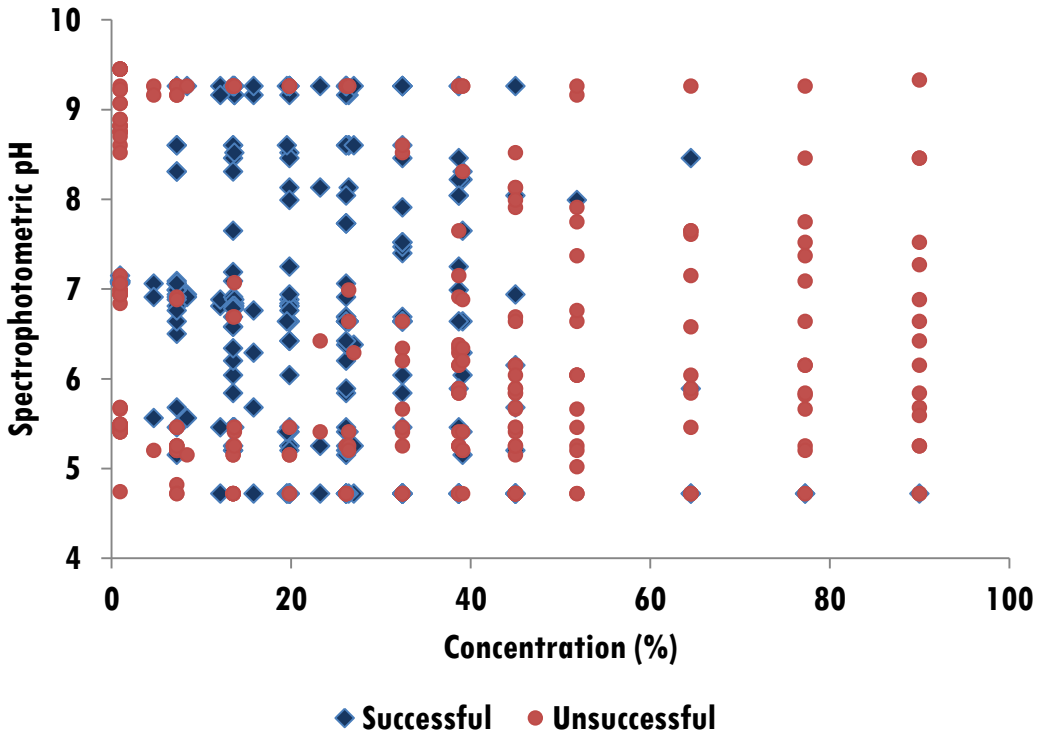


Figure 66: Success of PEG conditions in relation to their pH and concentration. Crystallisation results for 11 proteins across 552 conditions containing buffered PEG solutions. Successful crystallisation is indicated by blue diamonds and unsuccessful experiments by red circles.

Hierarchical clustering of the PEG results suggests that concentration and pH are the dominant parameters defining the difference between success and failure. Linear discriminant analysis and a one-layered ANN gave similar results.

Figure 66 shows the distribution of the conditions from custom screening using their spectrophotometric pH, PEG concentration and whether they crystallised. Most crystals are obtained between pH 5.5 and pH 9, and at concentrations between 5% and 35%. This zone contains PEGs of different molecular weights and functional groups. Of the 24 PEG-like chemicals trialled, 19 are found at least once within this zone. The other five chemicals are two non-PEG chemicals, tetramethyl sulfone and ethylene glycol, the highly acidic PEG 2000 dimethyl ether (pH ~2), and the low molecular weighted PEG 200 and PEG 400. These latter two have been found previously to crystallise a different set of proteins to other PEGs, (Kimber *et al.*, 2003). We have also observed that the pH of PEG 400 is not stable under any storage conditions.

This distribution suggests that the different weights of PEGs does not affect their similarity in terms of which proteins they crystallisation. These findings are also supported by those of Zhu *et al.* (2006) who describe the phase diagram for PEG 3350 at 21^{o}C (the temperature at which we crystallised our proteins). They state that the concentration range in which nucleation would occur being from 18%w/v to 30%w/v.

A potential modification to C6 would take the concentration of PEGs into account rather than molecular weight. In this modification, the β term is only included where PEGs are within 5% to 40% concentration. For any concentration outside this range, the β term is not used. So for example: 15% PEG 400 and 39% PEG 600 are considered similar and the β term would be used, but for 4% PEG 400 and 39% PEG 600 or 12% PEG 400 and 41% PEG 600, the β term would not be used.

A set of similar experiments to that of the PEGs was performed for salts. A total of 30 salts with anion sand cations of different valences, were trialled at different concentrations with 11 proteins.

Unlike the results of the PEGs, however, there was no obvious pattern in crystallisation results. Even ignoring variables such as pH and concentration did not suggest that salts with shared ions were the same in terms of what they crystallised. The results of the three buffer levels 5, 7 and 9 were combined as there were few successful crystallisation results. The most successful chemical was potassium bromide, crystallising 9 out the 11 proteins. The least successful were potassium phosphate, ammonium fluoride and sodium bromide all crystallising just one protein, thaumatin. The fact that thaumatin crystallised in 25 of the 30 salts, means that it is a property of the protein rather than the salt that allows crystallisation in these three salts. The average salt crystallised four proteins and the average protein crystallised in 11 salts. All proteins crystallised in three salts, potassium bromide, calcium chloride and sodium phosphate. Interestingly, these three salts all have a different cation and anion group which might suggest that different protein properties have different interactions with different chemicals in the crystallisation solution. Using the results of McPherson (2001) study of salts, we performed cluster analysis and found that few chemical species giving similar crystallisation results had matching ions. As a consequence, the ion term, γ, of the formula can be removed.

The final term in the C6 metric is related to pH. We modified this term by limiting the normalisation range to two pH units so that solutions that are not within two pH units of each other are considered to be significantly different. We showed earlier that when pH is carefully controlled and measured, most crystals for a particular protein are found within a narrow pH range.

In the modified metric we used the more accurate pH predicted by a regression model from the concentrations of the chemicals and the buffer pH to obtain a more accurate value for the true pH, rather than simply using the buffer pH. In instances where the buffer pH is not stated the pH term is not used, as in the original C6.

We refer to the modified C6 metric as C8.

9.2. Comparison of the C6 and C8 Metrics

In order to compare the results of experimental data to the theoretical C6 and C8 distances, we initially used the *Hamming distance* to provide experimental distances. A 39 × 96 matrix of 1's and 0's was created from 39 projects that had been trialled in all 96 conditions of the Filter 6 screen. The element on the *i*th of the *j*th column was set to 1 if the project *i* crystallised for in condition *j* and 0 otherwise. The Hamming distance was then calculated from this matrix for each pair of conditions. It became apparent that the clustering (not shown) was predominantly grouping together conditions that were similar due to being highly unsuccessful. To overcome this, the use of a different distance metric was employed, the *Jaccard distance*. However, a dendrogram of the results showed very little structure in the clusters, with many clusters consisting of just one pair of conditions.

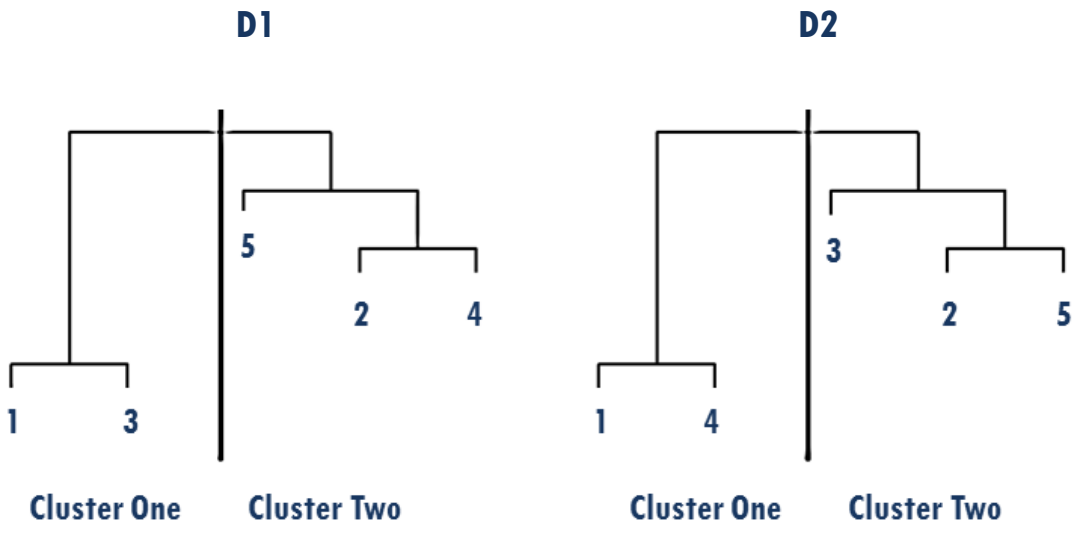


Figure 67: Two example dendrograms for comparison.

To determine objectively how well the C6 and C8 metrics reflect the observed data we used the B_k measure of dendrogram similarity (Fowlkes & Mallows, 1983). This allows two dendrograms to be compared at each cluster level. The dendrograms are cut to give *k = 2, 3…, n-1*clusters, where *n* is the total number of items. A matrix, ***m*** is defined in which the element m_{ij} is the number of objects that occur in the same cluster in both dendrograms for any given *k*.

Therefore,

$$B_k = \frac{T_k}{\sqrt{P_k Q_k}}, \quad \textbf{37}$$

where,

$$T_k = \left(\sum_{i=1}^{k} \sum_{j=1}^{k} {m_{ij}}^2 \right) - n, \quad \textbf{38}$$

$$P_k = \left(\sum_{i=1}^{k} \left(\sum_{j=1}^{k} m_{ij} \right)^2 \right) - n, \quad \textbf{39}$$

$$Q_k = \left(\sum_{j=1}^{k} \left(\sum_{i=1}^{k} m_{ij} \right)^2 \right) - n. \quad \textbf{40}$$

B_k is calculated for every value of *k* to create a plot of B_k versus *k*. A B_k value of 1 indicates the two dendrograms are identical, whereas a value of 0 indicates they share no common element.

For dendrogram D1 with objects 1 and 3 in cluster one (C1) and objects 2, 4 and 5 in cluster two (C2) and dendrogram D2 with objects 1 and 4 in C1 and objects 2, 3 and 5 in C2. The matrix, ***m*** is:

$[m_{ij}] =$

		D_2		
	C	1	2	
D_1	1	1	1	2
	2	1	2	3
		2	3	5

From Equations 38 to 40, we have

$$P_2 = (2^2 + 3^2) - 5 = 8,$$
$$Q_2 = (2^2 + 3^2) - 5 = 8,$$
$$T_2 = (1^2 + 1^2 + 1^2 + 2^2) - 5 = 7 - 5 = 2,$$

therefore,

$$\boldsymbol{B_2} = \frac{2}{\sqrt{8 \times 8}} = \frac{2}{8} = \mathbf{0.25}.$$

For small dendrograms it is possible to interpret the B_k by visual inspection. Simulations were performed in order to interpret B_k values for large numbers of objects. For our data with 96 different conditions (objects), $k \leq 95$.

For each *k*, simulations of the matrix ***m*** were produced to model varying similarity, where between 0 and 100% of the objects were in the same cluster (the leading diagonal) and the rest evenly spread across the off diagonal terms. For example, when $B_k = 1$, 100% of the data lies on the leading diagonal.

$[m_{ij}] =$

A_1 \ A_2	1	2	3	
1	24	4	4	32
2	4	24	4	32
3	4	4	24	32
	32	32	32	96

Figure 68: Example of B_k modelling.

A matrix obtained from three clusters is shown where 75% of the 96 objects are in the same cluster.

Figure 66 shows a matrix corresponding to $k = 3$, with 75% of the data on the leading diagonal, for 96 objects. This means 72 objects occur in the same clusters in each dendrogram, with 24 in each. The remaining 24 objects are found in other clusters and are evenly spread over the other 6 elements of the matrix. For this example B_3 would be calculated from:

$$P_3 = (32^2 + 32^2 + 32^2) - 96 = 2976,$$
$$Q_3 = (32^2 + 32^2 + 32^2) - 96 = 2976,$$
$$T_3 = (3 \times 24^2 + 6 \times 4^2) - 96 = 1728 + 96 - 96 = 1728,$$

Therefore,

$$\boldsymbol{B_3} = \frac{1728}{\sqrt{2976 \times 2976}} = \frac{1728}{2976} = \mathbf{0.58}.$$

A score of 0.58 for 3 clusters, therefore, corresponds to 75% of the data being in the same cluster in both dendrograms. Repeated simulations produce the surface shown in Figure 69.

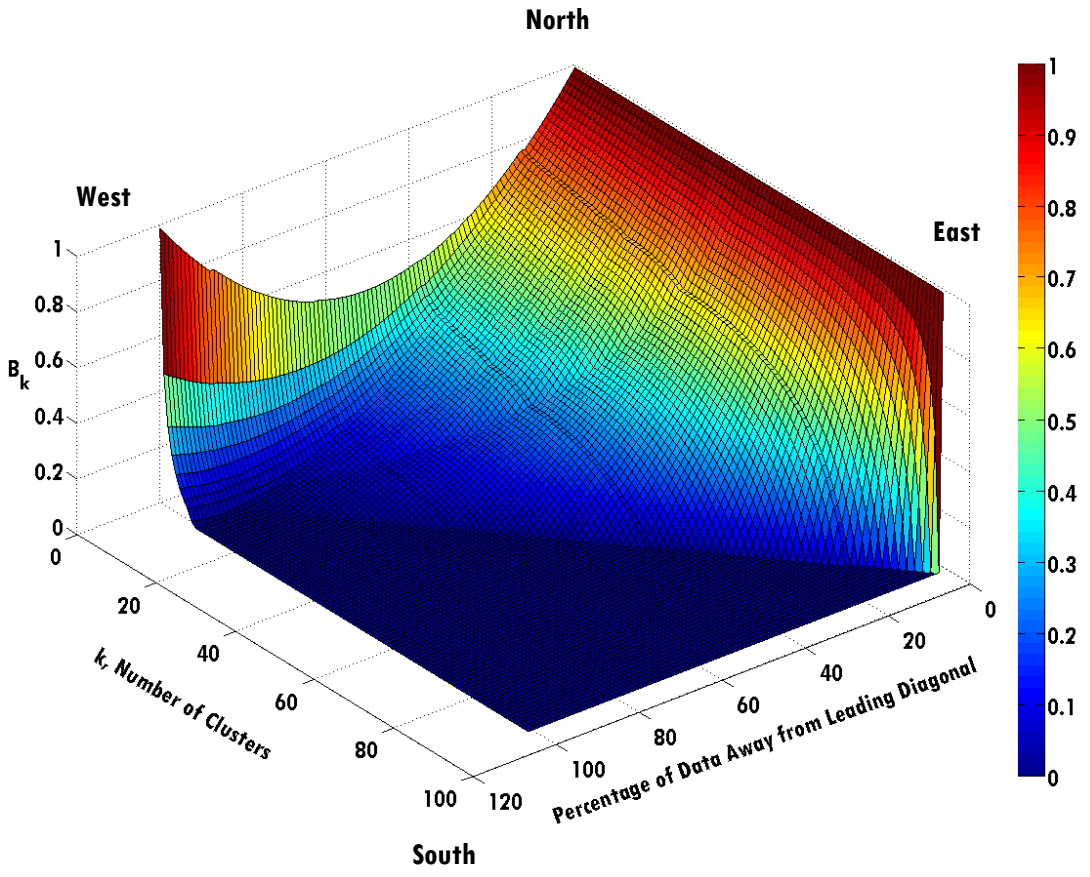


Figure 69: Surface of B_k values for 96 data points.

The surface shown is obtained from the B_k values for 96 objects with varying numbers of objects in the same cluster. Cardinal directions are given for ease of description.

Figure 69 shows how the B_k values change with different numbers of clusters and distribution of objects within these clusters. For two clusters (running from north to west) the value falls steeply from 1, where 100% of the objects are in the same cluster, to a minimum of 0.49 when the objects are evenly distributed. This value then increases until 100% of the objects are once again in the same cluster. For all *k*, when 100% of the objects are in the same cluster (from north to east) the B_k value is 1. The value of B_k decreases from north to south as the number of clusters increases and the number of objects are found in different clusters. Provided that at least 90% of the objects are away from the leading diagonal, from $k = 11$ onwards, the value of B_k is 0. This percentage drops by 1%, per increase of 1 in the value of *k* until at $k = 95$, when 99% of the objects are in different clusters giving a B_k value of 0. This can be seen by a plateau of value 0 running from west to east for the majority of the southern part of the surface, where south is defined in relation to a bisection of the

objects running from west to east. It is possible to obtain B_k values between two clusters and find their corresponding value on this surface to determine how many objects are in the same cluster.

	Jaccard	C6	C8
Jaccard	1.00		
C6	0.40	1.00	
C8	0.46	0.63	1.00

Figure 70: Correlations between methods assessing the similarity of conditions.
The correlation matrix is shown for three methods of determining similarity between conditions in the Filter 6 screen. The Jaccard distance is obtained from observed experimental results, while C6 and C8 are theoretical distances.

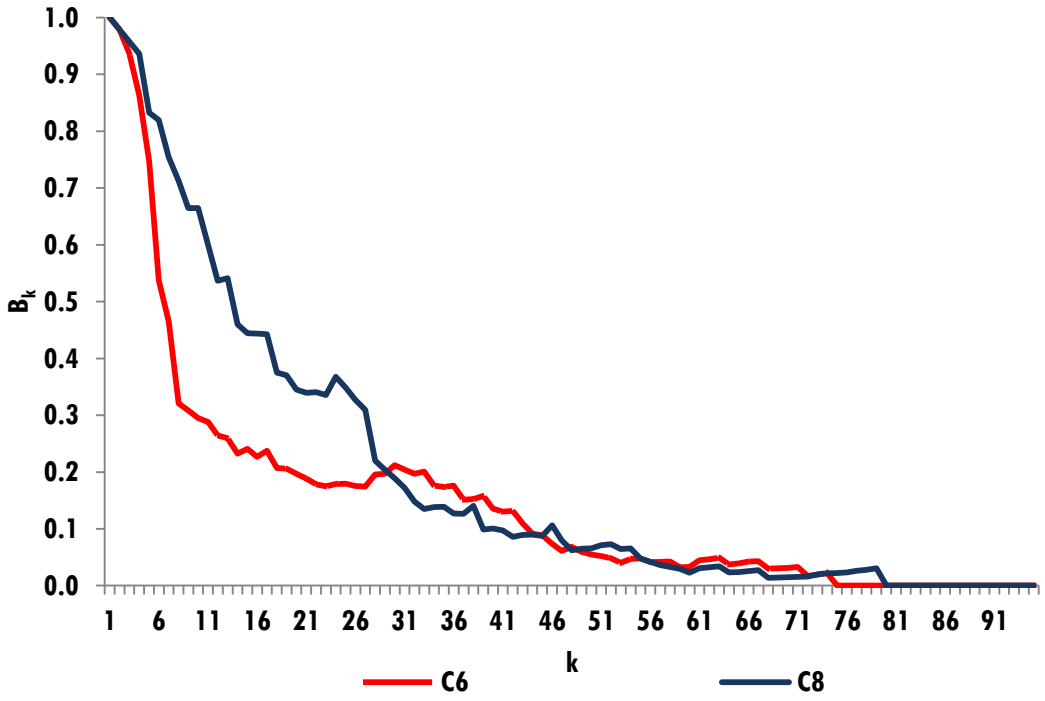
Figure 70 shows the correlation for the two methods, C6 and C8, of measuring distance in crystallisation chemical parameter space, in comparison to the Jaccard distances calculated from the crystallisation results for 39 projects in the Filter 6 screen. C6 and C8 have a similar structure, as might be expected, which is reflected in the 0.63 correlation coefficient between them. Neither C6 nor C8 have a strong correlation with the observed Jaccard distance as neither model can reflect the stochastic nature of crystallisation experiments (Newman *et al.*, 2007). It should be noted that the Filter 6 screen (a grid footprint hybrid) provides a systemic sampling of pH, chemical species and concentration.

Data from the SGC was introduced to provide further test data. This data was obtained from 1,039 proteins crystallised in the JCSG +4, a sparse matrix screen. Figure 71 shows the correlation between Jaccard and C6 is 0.63 for this data and between Jaccard and C8 is 0.71. Although both metrics have higher correlation with the observed data than seen for the Filter 6 screen, they are not as highly correlated with each other. Figure 71 also shows the B_k metric used to determine the similarity of the clustering. The B_k values can be interpreted by comparison with the surface shown in Figure 69. For both metrics, when two clusters are used, 99% of the objects

are in the same cluster as obtained using the Jaccard distance. The C8 metric gives more similar clusters to the observed data until k = 28. With 24 clusters, 68% of the conditions are in the same cluster for C8 and the Jaccard metrics whereas Jaccard and C6 had 53% of conditions within the same clusters.

	Jaccard	C6	C8
Jaccard	1.00		
C6	0.63	1.00	
C8	0.71	0.46	1.00

(a) Correlation matrix between C6, C8 and Jaccard distances.



(b) B_k value between C6 and Jaccard clusters and C8 and Jaccard clusters.

	k			
	2	6	12	24
Bk Value				
C6	0.97	0.53	0.26	0.17

C8	0.97	0.81	0.63	0.36

Percentage in same cluster

C6	99	74	55	53
C8	99	91	76	68

(c) Summary of B_k values derived from the graph shown in (b).

Figure 71: Comparison of C6 to C8 for the JCSG +4 screen.

The similarity of the C6 and C8 clustering is compared to patterns observed (Jaccard) for 1,039 proteins trialled in the JCSG+4 sparse matrix screen. (a) shows the correlation between Jaccard, C6 and C8. (b) shows the B_k measure for comparing dendrograms, between the clustering of the distances C6 (red dots) and C8 (blue line), with the observed distance (Jaccard). (c) shows the percentage of conditions that have been clustered together for each metric using different numbers of clusters.

9.3. Conclusions

Crystallisation chemical parameter space is large and often populated with regions that are chemically similar to other regions but do not always crystallise the same proteins. The C6 distance metric (Newman *et al.*, 2010) provides a method of grouping regions by assuming that certain chemical species and pH are similar, however, this may not be correlated to the patterns obtained by experimental data due the stochastic nature of crystallisation. The C6 metric was deliberately designed without the use of empirical data due to difficulties in obtaining a globally representative sample of proteins. Here, we have investigated the accuracy of this metric with real datasets obtained from different screen types (custom, filter one and sparse matrix). After comparing the clustering of conditions obtained from experimental data with the theoretical clusters obtained from C6 and the new C8 metric, we show that the new metric provides a more accurate reflection of the observed patterns. This new metric, obtained via modifications to the terms in C6 involving PEGs, ions and pH, allows efficient design and assessment of

crystallisation screens and conditions. The repeated analysis of crystallisation data should allow this metric to be updated to more accurately reflect crystallisation parameter space.

10. Conclusions and the Future

Crystallisation is essential for the determination of protein structures by X-ray crystallisography. In this thesis some of the problems associated with protein crystallisation are discussed along with methods to potentially reduce the high attrition rates for this process.

Protein crystallisation solutions are typically recorded in terms of severable variables, the chemicals species, the chemical concentration and the pH of the buffer or a component that has had its pH modified. The pH is known to be an important variable but it is well known that the recorded buffer pH can be inaccurate (Bukrinsky & Poulsen, 2001, Newman, Sayle, *et al.*, 2012). Using spectrophotometry and the acid-base indicator bromothymol blue, we have shown that the pH of crystallisation solutions can be determined accurately, efficiently and quickly. This allows conditions to be replicated without the requirement for a pH meter.

The inaccurate recording of pH in databases has meant that any analysis of such data is also likely to be erroneous. Inaccurate entries were recorded in data at AstraZeneca, the Structural Genomics Consortium, Oxford and the world's largest repository of successful crystallisation data - the Protein Data Bank. Many studies have reported the distribution of buffer pH (Samudzi *et al.*, 1992, Rupp & Wang, 2004, Bonneté, 2007), however, drawing conclusions from these distributions could be misleading. Kantardjieff and Rupp (2004) also reported a link between the pH at which a protein would crystallise and its isoelectric point (pI). We were able to develop a method to predict a pH as accurate as that of the spectrophotometric pH from the buffer pH using linear regression and machine learning. Using neural networks to associate input combinations of chemical with output pH values, it has been possible to achieve a more accurate prediction of the true pH than that of the buffer. We were then able to reinvestigate some of the results that have been previously reported.

Using the newly modelled pH, with custom experiments, data from the SGC and the PDB it has been possible to examine the fiercely contested link between a protein's

pI and the pH of crystallisation. A lack of correlation between pI and pH was confirmed, however, some patterns were observed. Such information could aid the identification of suitable initial conditions and, therefore, could help reduce attrition rates.

Proteins can be grouped into distinct categories according to the stage of crystallisation they reached. Some protein crystals were of sufficient quality to be used for structure determination while others crystallised and were never followed up. Although attempts to classify proteins into these categories were not successful, it was possible to classify proteins as crystallisable or non-crystallisable based on properties intrinsic to their sequence. Three properties were deemed to have the most predictive power: pI, GRAVY and the number of cysteines. This analysis involved the use of data from TargetDB (TargetDB, 2010), which provided both positive and negative data. Using a neural network we achieved a correct classification rate of around 70%, comparable to other published results. However, when the trained classifier was using data from the PDB we found that the classification algorithm only returned the correct result for 55%, a percentage that could have been achieved through guessing. We suggest that this is because the data used to train the network included proteins that were annotated as having "Diffraction-quality Crystals", but not annotated with "In PDB" in the "Status" field. The structural genomics targets in TargetDB may be restricted, for example due to interest in particular medical interests (human proteins, for example, which do not over-express in bacteria), whereas structures deposited in the PDB are from a wider, and potentially more difficult to crystallise, range of proteins.

Chemical parameters were also explored and, using data from AstraZeneca and the PDB, it was possible to confirm that PEGs, especially PEG 3350, were the crystallisation reagents that were the most successful. Other successful crystallisation reagents include ammonium sulfate and buffers (assumed to be chemically inert). We obtained these results through the use of a previously proposed metric, propensity analysis (Rupp & Wang, 2004) and minimal set analysis. The minimal set of conditions, obtained by mining the Protein Data Bank, builds on the work of Jancarik and Kim and could help increase the number of proteins crystallised.

Using minimal spanning analysis of a set of approximately 9,000 proteins and their crystallisation conditions in the PDB it was possible to create a screen that if used from the onset of time would have crystallised over 2,000 of these proteins. The conditions for this screen were analysed using the C6 metric, giving an internal diversity score of 0.9, which is comparable to other commercially available screens. Initial trials have proven successful at AstraZeneca and YSBL, where crystalline material has been obtained for 23/31 (74%) of the proteins trialled. There is the potential for it to become an integral part of their initial screening protocol. The conditions of the screen are hosted on the C6 web tool (Newman *et al.*, 2010).

We also explored the possibility to reduce the number of variables in crystallisation parameter space by determining which chemical species are similar when crystallising biomolecules. This would help in the development of screens. It was possible to develop further the C6 metric, which compares crystallisation conditions and provides a distance between 0 and 1 (Newman *et al.*, 2010). Using custom experiments and data from AstraZeneca, analysis was performed on the similarity of PEGs and ions. It was possible to develop the C6 metric, using this analysis, to fit the patterns observed in crystallisation screens more accurately.

The development of a high-throughput method of measuring pH highlighted problems with a common acid-base indicator, Universal Indicator. This indicator covers a large pH range, but has very little colour change over neutral pHs- those most used in crystallisation. We found that the indicator Bromothymol Blue does show differences over this range, but has no observable colour change for pH values below 5 or above 8. Further experiments with combinations of indicators or the use of a multi-well tray to test several indicators simultaneously would allow an even more accurate method of determining pH.

Using models to predict the effect of various additives within a solution provides a more useful estimate of the measured pH than that of the buffer. This work could be extended to further groups of chemicals to build more models and allow subtler effects to be taken into account. With regards to buffers, it might also be possible to determine which buffers maintain pH when placed in solution with an increasing

concentration of strong acid or base. It may also be possible to determine the different effects that buffers have on crystallisation other than controlling pH.

The difference between the proteins and their properties at Structural Genomics (SG) centres and the Protein Data Bank needs to be explored further in order to explain why some proteins can be classified as crystallisable or non-crystallisable whereas some cannot. There have been several studies using SG data (Page & Stevens, 2004, Chen *et al.*, 2004, Kimber *et al.*, 2003) and it seems that machine learning algorithms trained on data from these centres does not generalise to other more varied proteins being explored elsewhere. Although the use of such properties has been shown to provide information on a proteins propensity to crystallise, it is known that even slight modifications to a protein sequence can affect crystallisability.

The standardised PDB facilitates data mining studies and could be used to investigate further indicators of a proteins ability to crystallise including, for example, molecular weight and domain structure. Is low molecular weight better than high molecular weight, are single domain proteins more likely to crystallise than multi-domain proteins and is an oligomeric state multimer better than a monomer? We have seen that the most widely used crystallisation agents include both ‘salting-in’ and ‘salting-out’ chemicals and further investigations could explore any links between protein properties and salt types. Where similar proteins have been crystallised in multiple conditions, potentially in different crystal forms, any link between the resolution of diffraction and the crystallisation conditions could be investigated.

Appendix A

Features for Predicting a Protein's Propensity to Crystallise

Feature Set	Features
fsOur87	Length, molecular weight, isoelectric point, instability index, aliphatic index, Grand Average of Hydropathy (GRAVY), count of 20 standard amino acids, count of 20 standard amino acids normalised by sequence length, mean entropy (3 features), total number of charged residues (2 features), number of different types of amino acids (8 features), number of atom counts (6 features), extinction coefficient (4 features), half-life (3 features), net charge (15 features).
fsUncorrelated	Isoelectric point, instability index, aliphatic index, GRAVY, the amino acids A, R, N, D, C, Q, G, H, I, K, M, F, S, T, Y, the number of sulfurs each sequence had, the number of small, aromatic, aliphatic and proline amino acids, extinction coefficient (1 feature), half-life (3 features), count of 20 standard amino acids normalised by sequence length, mean entropy (1 feature), the net charge at pH, 4,6,10,12 and 14.
fsOB	Isoelectric point, GRAVY.
fsParCrys	Isoelectric point, GRAVY, count of amino acids S, C, G, F, Y, M, T, H, D, W, P.

fsCRYSTALP	The following amino acids, where - is indicative of another unnamed amino acid. Y, DL, EH, LR, PD, RI, RT, SS, WC, YT, H-H, I-C, L-E, Q-L, T-E, T-T, Y-F, E--C, F--Q, I--P, L--E, Q--S, S--L, T--G, W--V, Y--N, A---G, C---L, E---L, E---Q, H---S, L---D, M---A, N---I, N---Q, C----S, D----N, F----T, G----R, I----G, M----A, M----Y, N----H, T----G, T----Y, V----T.
fsCRYSTALP2	The following amino acids, where - is indicative of another unnamed amino acid. Isoelectric point, GRAVY, L, Y, RI, DL, QG, QM, ES, GL, HH, IR, LF, LS, PP, SS, SV, WC, WM, WW, WV, YI, YT, R-S, D-L, C-A, Q-L, H-R, H-G, H-H, I-R, L-E, F-S, T-K, T-S, T-T, D--M, H--C, H--H, L--N, K--W, S--L, T--G, W--W, Y--N, R---D, Q---C, E---Q, E---S, G---H, L---D, L---L, F---T, Y---I, V---Y, C----E, C----H, C----S, E----Q, E----F, G----R, I----E, L----L, M----Y, M----V, S----H, W----H, W----M, V----T, EFV, IVV, TKV, R-PS, Q-QQ, K-TV, M-DS, F-TK, P-PE, DP-V, LR-F, MG-S, SA-D, YV-E, VT-G, N-P-G, K-I-R, F-E-F,S-T-S.

Appendix B

List of conditions for PDB Minimal Spanning Screen

A	Number of proteins contributed
B	Predicted pH
C	Buffer pH

A	B	C	Conditions
69	8.0	8.5	30% (w/v) polyethylene glycol 4000; 0.1 M tris chloride; 0.2 M magnesium chloride
56	7.9	8.5	0.1 M tris chloride; 0.2 M sodium acetate; 30% (w/v) polyethylene glycol 4000
52	6.7	7.5	20% (w/v) polyethylene glycol 4000; 10% (v/v) 2-propanol; 0.1 M hepes
45	5.5	5.5	25% (w/v) polyethylene glycol 3350; 0.1 M bis-tris
43	5.5	5.5	0.2 M magnesium chloride; 0.1 M bis-tris; 25% (w/v) polyethylene glycol 3350
43	7.7	8.5	2 M ammonium sulfate; 0.1 M tris chloride
42	6.5	6.5	20% (w/v) polyethylene glycol 8000; 0.1 M sodium cacodylate; 0.2 M magnesium acetate
39	5.2	4.6	30% (w/v) polyethylene glycol 4000; 0.1 M sodium acetate; 0.2 M ammonium acetate
38	6.5	6.5	1.6 M ammonium sulfate; 0.1 M mes; 10% (v/v) dioxane
37	5.7	5.5	0.2 M ammonium acetate; 0.1 M bis-tris; 25% (w/v) polyethylene glycol 3350
35	5.5	5.5	25% (w/v) polyethylene glycol 3350; 0.2 M sodium chloride; 0.1 M bis-tris
34	5.5	4.6	0.1 M sodium acetate; 2 M ammonium sulfate
32	5.5	5.5	0.1 M bis-tris; 0.2 M lithium sulfate; 25% (w/v) polyethylene glycol 3350
29	6.5	6.5	12% (w/v) polyethylene glycol 20000; 0.1 M mes
28	6.3	6.5	0.2 M magnesium chloride; 0.1 M bis-tris; 25% (w/v) polyethylene glycol 3350

28	7.4	7.5	2 M ammonium sulfate; 2% (v/v) polyethylene glycol 400; 0.1 M hepes
27			2 M ammonium sulfate
27	5.2	5.6	20% (w/v) polyethylene glycol 4000; 20% (v/v) 2-propanol; 0.1 M sodium citrate
27	5.2	4.6	0.2 M ammonium sulfate; 0.1 M sodium acetate; 25% (w/v) polyethylene glycol 4000
27	6.4	6.5	30% (w/v) polyethylene glycol 8000; 0.1 M sodium cacodylate; 0.2 M ammonium sulfate
27	5.7	5.5	0.2 M ammonium sulfate; 0.1 M bis-tris; 25% (w/v) polyethylene glycol 3350
26	6.3	6.5	0.1 M bis-tris; 0.2 M ammonium sulfate; 25% (w/v) polyethylene glycol 3350
26	8.0	8.5	30% (w/v) polyethylene glycol 4000; 0.2 M magnesium chloride; 0.1 M tris
25	5.7	5.6	0.2 M ammonium acetate; 0.1 M sodium citrate; 30% (w/v) polyethylene glycol 4000
24	7.2	7.5	0.2 M magnesium chloride; 0.1 M hepes; 25% (w/v) polyethylene glycol 3350
24	6.6	6.5	30% (w/v) polyethylene glycol monomethyl ether 5000; 0.2 M ammonium sulfate; 0.1 M mes
23	6.3	4.6	0.1 M sodium acetate; 2 M sodium formate
22	7.1	7.5	8% (v/v) ethylene glycol; 10% (w/v) polyethylene glycol 8000; 0.1 M hepes
22	7.3	7.5	20% (w/v) polyethylene glycol 10000; 0.1 M hepes
21	6.6	6.5	20% (w/v) polyethylene glycol monomethyl ether 5000; 0.1 M bis-tris
20	6.1	6	20% (w/v) polyethylene glycol 8000; 0.1 M mes; 0.2 M calcium acetate
20	5.2	4.6	30% (v/v) 2-methyl-2,4-pentanediol; 0.1 M sodium acetate; 0.02 M calcium chloride
20	8.0	8.5	0.2 M magnesium chloride; 0.1 M tris chloride; 25% (w/v) polyethylene glycol 3350
20	7.9	7.5	0.1 M hepes; 1.4 M trisodium citrate
20	5.1	4.6	0.1 M sodium acetate; 30% (w/v) polyethylene glycol monomethyl ether 2000; 0.2 M ammonium sulfate
20	8.0	8.5	0.2 M lithium sulfate; 0.1 M tris chloride; 30% (w/v) polyethylene glycol 4000

19	7.1	7.5	0.1 M hepes; 25% (w/v) polyethylene glycol 3350
18	7.2	7.5	28% (v/v) polyethylene glycol 400; 0.2 M calcium chloride; 0.1 M hepes
18	7.0	7	2.4 M sodium malonate
18	7.9	7.5	0.1 M hepes; 1.4 M sodium citrate
17	5.9	5.6	2 M ammonium sulfate; 0.1 M sodium citrate; 0.2 M potassium sodium tartrate
17	6.9	7.5	0.2 M ammonium sulfate; 25% (w/v) polyethylene glycol 3350; 0.1 M hepes
16	7.0	6.5	1 M sodium citrate; 0.1 M sodium cacodylate
16			0.1 M sodium chloride; 0.005 M dithiothreitol; 0.02% (v/v) sodium azide; 0.01 M tris chloride
16	8.2	8.5	20% (w/v) polyethylene glycol 8000; 0.2 M magnesium chloride; 0.1 M tris
16	5.9	5.5	0.1 M sodium citrate; 20% (w/v) polyethylene glycol 3000
16	5.3	4.6	0.1 M sodium acetate; 8% (w/v) polyethylene glycol 4000
16	7.2	7.5	10% (w/v) polyethylene glycol 6000; 5% (v/v) 2-methyl-2,4-pentanediol; 0.1 M hepes
16	5.7	5.5	0.1 M bis-tris; 2 M ammonium sulfate
15	6.2	5.5	1 M ammonium sulfate; 0.1 M bis-tris; 1% (w/v) polyethylene glycol 3350
15	7.7	7.5	1.5 M lithium sulfate; 0.1 M hepes
15	6.9	7.5	25% (w/v) polyethylene glycol 3350; 0.1 M hepes; 0.2 M ammonium acetate
15	7.3	7.5	20% (w/v) polyethylene glycol 8000; 0.1 M hepes
15	6.3	6.5	25% (w/v) polyethylene glycol 3350; 0.2 M lithium sulfate; 0.1 M bis-tris
15	7.2	7.5	0.2 M lithium sulfate; 0.1 M hepes; 25% (w/v) polyethylene glycol 3350
14	7.6	9	2.4 M ammonium sulfate; 0.1 M bicine
14	6.2	6.5	0.1 M bis-tris; 0.05 M calcium chloride; 30% (v/v) polyethylene glycol monomethyl ether 550

14	7.2	7.5	30% (v/v) polyethylene glycol 400; 0.2 M magnesium chloride; 0.1 M hepes
14	8.0	8.5	30% (w/v) polyethylene glycol 4000; 0.2 M lithium sulfate; 0.1 M tris
14	6.7	6.5	1.6 M magnesium sulfate; 0.1 M mes
14	6.5	6.5	18% (w/v) polyethylene glycol 8000; 0.1 M sodium cacodylate; 0.2 M calcium acetate
14	6.8	6.5	2 M ammonium sulfate; 0.1 M bis-tris
14	6.4	6.5	30% (w/v) polyethylene glycol 8000; 0.1 M sodium cacodylate; 0.2 M sodium acetate
13	6.1	6.5	0.1 M bis-tris; 28% (w/v) polyethylene glycol monomethyl ether 2000
13	7.8	8	10% (w/v) polyethylene glycol 8000; 0.1 M imidazole; 0.2 M calcium acetate
13			60% (v/v) tacsimate
12			2.1 M dl-malic acid
12			0.15 M dl-malic acid; 20% (w/v) polyethylene glycol 3350
12	6.8	6.5	2 M ammonium sulfate; 0.2 M sodium chloride; 0.1 M sodium cacodylate
12			4 M sodium formate
11	7.1	6.5	1.4 M sodium acetate; 0.1 M sodium cacodylate
11	7.2	7.5	0.2 M sodium chloride; 0.1 M hepes; 25% (w/v) polyethylene glycol 3350
11	7.4	7.5	4.3 M sodium chloride; 0.1 M hepes
11			20% (w/v) polyethylene glycol 3350; 0.2 M ammonium chloride
11			20% (w/v) polyethylene glycol 3350; 0.2 M sodium formate
10	6.4	6.6	0.2 M ammonium formate; 20% (w/v) polyethylene glycol 3350
10	7.0	7.3	0.2 M calcium acetate; 20% (w/v) polyethylene glycol 3350
10	5.9	5.6	0.2 M potassium sodium tartrate; 0.1 M trisodium citrate; 2 M ammonium sulfate

10	6.4	6.5	1.8 M ammonium sulfate; 0.01 M cobalt chloride; 0.1 M mes
10	6.7	10.5	1.2 M sodium dihydrogen phosphate; 0.8 M dipotassium hydrogen phosphate; 0.1 M caps; 0.2 M lithium sulfate
10	6.3	6.5	40% (v/v) polyethylene glycol 300; 0.1 M sodium cacodylate; 0.2 M calcium acetate
10	5.5	4.5	0.1 M sodium acetate; 2 M ammonium sulfate
10	6.5	7.5	0.1 M hepes; 70% (v/v) 2-methyl-2,4-pentanediol
10	6.7	7.5	20% (w/v) polyethylene glycol 4000; 10% (v/v) 2-propanol; 0.1 M sodium hepes
10	8.0	8.5	25% (w/v) polyethylene glycol 3350; 0.2 M magnesium chloride; 0.1 M tris
10	7.4	7.5	0.1 M sodium hepes; 2% (v/v) polyethylene glycol 400; 2 M ammonium sulfate
10	6.3	6.5	0.1 M bis-tris; 0.2 M sodium chloride; 25% (w/v) polyethylene glycol 3350
10	7.7	8.5	2 M ammonium sulfate; 0.1 M tris
9	5.1	4.5	30% (w/v) polyethylene glycol 8000; 0.1 M sodium acetate; 0.2 M lithium sulfate
9	6.1	6	20% (w/v) polyethylene glycol 6000; 1 M lithium chloride; 0.1 M mes
9	6.9	4.6	3.5 M sodium formate; 0.1 M sodium acetate
9	8.6	9.5	0.1 M ches; 20% (w/v) polyethylene glycol 8000
9	9.0	9.5	1 M sodium citrate; 0.1 M ches
9	4.8	4.2	20% (w/v) polyethylene glycol 1000; 0.2 M lithium sulfate; 0.1 M phosphate-citrate
9	8.0	7	2.8 M sodium acetate
9	7.2	6.5	1.6 M sodium citrate

Appendix C

The Changing pH of PEGs Overtime

It is known that the pH of a crystallisation solution can be imperative to the success of trial. Many crystallisation solutions are unbuffered or contain a buffer at a concentration that is not capable of stabilising the pH with some chemicals. Chemicals such as dihydrogen-, hydroxide-, and weak acid salts modify the pH most, but ammonia-containing compounds and PEGs undergo degradation overtime and therefore modify the pH in an unpredictable manner. PEGs are the most successful crystallisation reagents, their inclusion in screens is common and so their potential to modify pH is of particular interest. Here, we assess the effects of storage conditions on PEGs of various molecular weights that were purchased from four different suppliers as shown in Table 20.

Aldrich (A)	Fluka (F)	Hampton Research (HR)	Molecular Dimensions (MD)
2000 (S)	-	-	2000
2000 MME (S)	-	2000 MME	2000 MME
-	4000 (S)	4000 (S)	4000
10000 (S)	10000 (S)	10000	10000

-	Unavailable
(S)	Solid form

Table 20: The different weight PEGs purchased from various suppliers.

Those PEGs that were purchased in a solid form (waxy granules) were made up to a 50% weight per volume solution by dissolving the granules in warmed ultrapure water and allowing cooling to room temperature. PEGs purchased as solutions were also 50% (w/v). For each available molecular weight from each manufacturer, solutions were dispensed into twelve 5ml polystyrene containers with a plastic screw cap and the pH measured immediately from two aliquots using a pH meter. The remaining ten containers (for each molecular weight and manufacturer) were stored in pairs in the following conditions:

- FD: -18°C and dark (freezer)
- DC: 6°C and dark (cold room cupboard)
- LC: 6°C and light (cold room shelf)
- DR: 20°C and dark (laboratory cupboard)
- LR: 20°C and light (laboratory shelf)

After 115 days, two containers (for each molecular weight and manufacturer) were removed from storage and the pH measured. This was repeated after a further 78 days (193 days from dispensation).

Results

t	**0**		**115**		**193**	
	1	**2**	**1**	**2**	**1**	**2**
A-2000-DC	7.7	7.8	7.9	7.9	8.2	8.3
A-2000 MME-DC	8.3	8.3	6.8	6.9	8.3	8.4
A-10000-DC	6.2	6.3	6.0	5.9	6.0	6.0

(a) The pH measurements for PEGs purchased from Aldrich stored in the cold and dark.

t	**0**	**115**	**193**	
A-2000-DC	7.8	7.9	8.2	0.4
A-2000 MME-DC	8.3	6.8	8.4	0
A-10000-DC	6.2	5.9	6	0.2

(b) The pH measurements for each initial and last time interval averaged, with the maximum absolute difference in measurements shown in red.

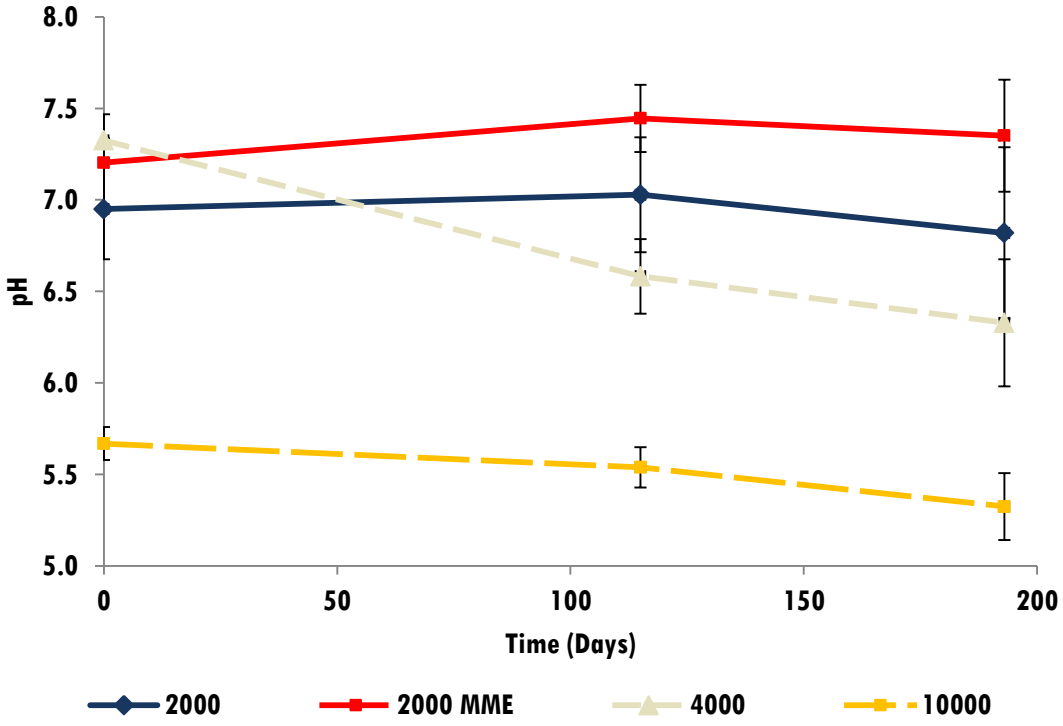
Table 21: pH measurements for PEGs purchased from Aldrich.

The pH measurements for PEG 2000 purchased from Aldrich and stored in the fridge in a cupboard (dark cold). Table (a) shows the raw data and table (b) shows how it was averaged.

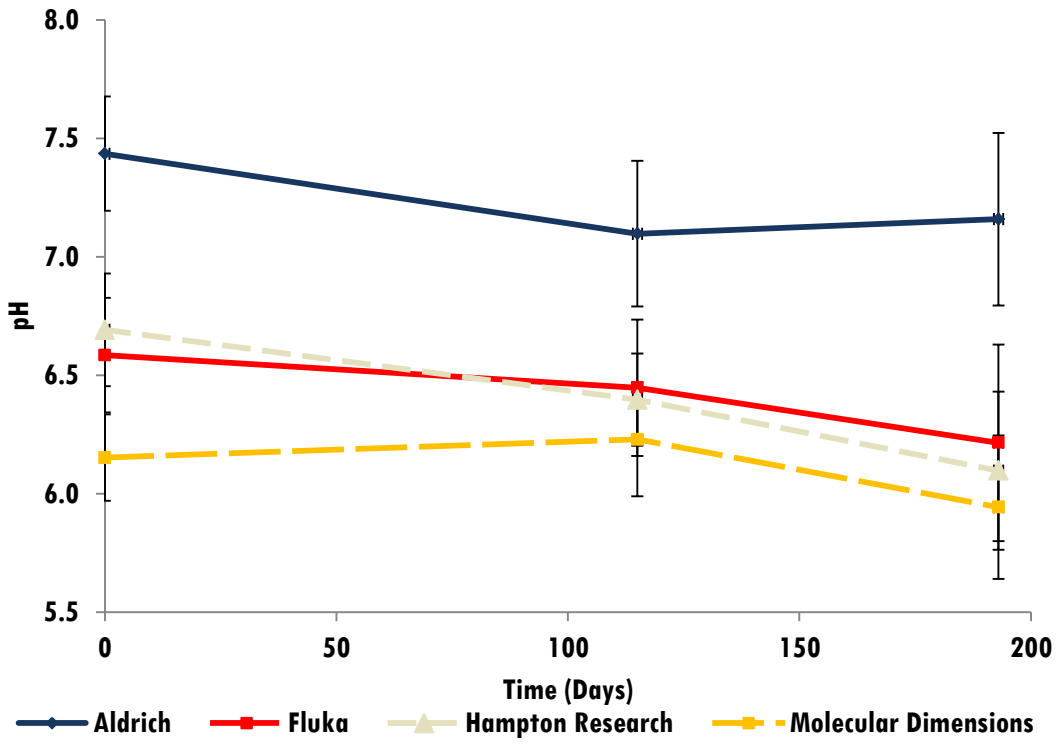
In addition to the PEGs shown in Table 21, PEG 400 was also purchased from each supplier. However, we found that it was not possible to obtain a stable pH measurement at this molecular weight and therefore, the pH of PEG 400 is not included here.

Table 21 shows an example set of recordings and how they were averaged and analysed. Whilst the pH for PEG 2000 gradually increased over the 193 days from an average of pH 7.8 to 8.2, increasing 0.4 pH units overtime, whereas PEG 2000 MME only changed by 0.04 pH units, well within the error expected for a pH meter.

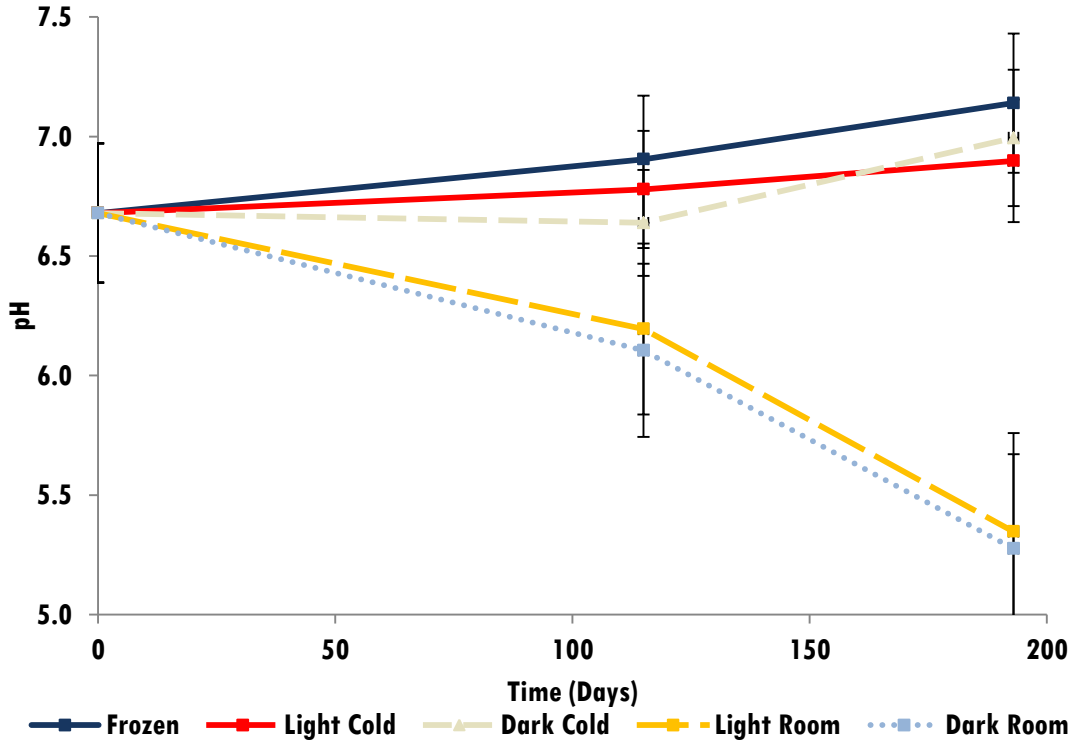
Figure 72a shows the pH measurement for each time point averaged over all storage conditions and all manufacturers supplying the different molecular weights. The results suggest that, on average, the pH of PEG 4000 changes most from a basic pH of 7.3 to an acidic one of 6.3. The smallest change in pH is for PEG 10000 which remains constant around the already acidic pH of 5.5.



(a) Each molecular weight averaged over manufacturer and storage conditions.



(b) Each manufacturer averaged over molecular weight and storage conditions.



(c) Each storage condition averaged over molecular weight and manufacturer.

Figure 72: The change in pH over time associated with different parameters.

The pH measurements were averaged over combinations of the three major parameters: molecular weight, manufacturer and storage conditions. (a) shows the results for each molecular weight averaged over manufacturer and storage conditions. (b) shows the results for each manufacturer averaged over molecular weight and storage conditions. (c) shows the results for each storage condition averaged over molecular weight and manufacturer. The error bars shown are standard error, defined as the standard deviation over the square root of the number of observations.

The measurements were also averaged over all molecular weights and storage conditions to determine any patterns due to manufacturer (Figure 1b) and over all molecular weights and manufacturers to examine differences between storage conditions. Figure 72b shows that the chemicals obtained from Fluka (F) and Hampton Research (HR) become more acidic, whereas those from Molecular Dimensions (MD) and Aldrich (A) seem to remain constant. For the storage conditions, the largest change in pH is seen for PEGs stored at room temperature, with averages for both light and dark falling from around pH 6.7 to 5.3. On the other hand the chemicals stored in the cold became slightly more basic from an average around 6.7 increasing to a more neutral pH.

Taking the mean absolute difference (MAD) between the time of dispensation and the final measurement provides summary statistics for the PEGs by molecular weight, manufacturer and storage method. The PEG weights which changed the least across all storage methods and all manufacturers were PEG 2000 and 10000, with a MAD of 0.73 for both. The largest MAD of 1.3 was for PEG 4000, as shown in Figure 72. However, the largest MADs for the manufacturers were Molecular Dimensions (1.3) and the smallest Aldrich (0.45). On closer inspection we found that the MD solutions change in both an acidic and a basic direction creating an illusion of little change. For example, MD 2000 which was frozen or in the fridge became more basic by at least half a unit and MD 2000 which was at room temperature became more acidic by over 1.5 units. The storage method with the lowest MAD was dark and cold with a value of 0.5, similarly low was frozen (0.6) and light cold (0.5) and the largest MAD was for light room of 1.6 and dark room 1.4 as suggest by the graph in Figure 72. Those PEGs that were made up from solid had a MAD of 0.7 and those already in solution of 1.1. On an individual basic, MD-4000-DR changed 3.2

units from pH 7.3 to 4.1, whereas HR-2000 MME-DC only changed in the hundredths of the measurement.

Conclusions

Manufacturers (MolecularDimensions, 2015) suggest that light can affect PEG solutions:

> *"PEG solutions are light sensitive and can degrade over time if kept in the light. Therefore, we recommend keeping them in the dark."*

Our results suggest that it is temperature rather than light that causes the largest change in the pH of PEGs. Storage in both light and dark gave similar results for a given temperature. Certain molecular weights, PEG 4000 and PEG 2000 MME have a pH that is susceptible to change, whereas PEG 10000 appears to be more stable and, therefore, requires less monitoring. It might be possible to conclude that the purchasing of chemicals in their dry form from Aldrich and Fluka (where available) and making them up might results in a more consistent PEG pH. It is possible, as shown in Figure 73, that PEGs in solution are already undergoing change, as suggested supported by the average initial pH for liquid (on purchased) PEGs being 6.2, whereas those solid PEGs (we made into solutions) having an initial average pH of 7.1, which is closer to the expected neutral. Through continued monitoring it will be possible to determine any longer term effects due to differences in molecular weight, manufacturer and storage method.

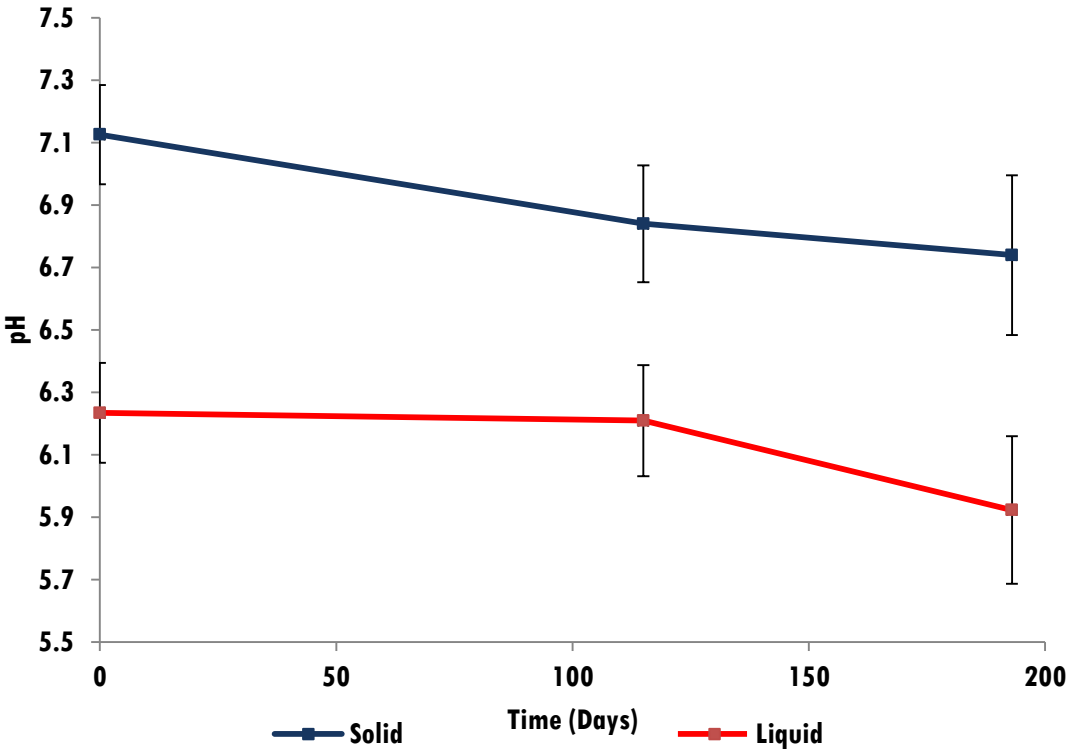


Figure 73: The pH of PEGs for the two different forms purchased.

PEGs were purchased in either granular (solid) form or liquid form (premade solution). The solid PEGS that we made into solution have a more neutral pH than those premade. Initially, the standard deviation of both types is 0.87. Standard error bars are shown.

List of References

Achterberg, T. (2009). *Mathematical Programming Computation* **1**, 1-41.

Aloy, P. & Russell, R. B. (2006). *Nat Rev Mol Cell Biol* **7**, 188-197.

Anfinsen, C. B., Haber, E., Sela, M. & White Jr, F. (1961). *Proceedings of the National Academy of Sciences of the United States of America* **47**, 1309.

Arakawa, T. & Timasheff, S. N. (1985). *Methods in enzymology* **114**, 49-77.

Artimo, P., Jonnalagedda, M., Arnold, K., Baratin, D., Csardi, D., de Castro, E., Duvaud, S., Flegel, V., Fortier, A., Gasteiger, E., Grosdidier, A., Hernandez, C., Ioannidis, V., Kuznetsov, D., Liechti, R., Moretti, S., Mostaguir, K., Redaschi, N., Rossier, G., Xenarios, I. & Stockinger, H. (2012). *ExPASy: SIB bioinformatics resource portal*, http://www.expasy.org/.

Asherie, N. (2004). *Methods* **34**, 266-272.

Atsushi, I. (1980). *Journal of biochemistry* **88**, 1895-1898.

Baker, D. & Sali, A. (2001). *Science's STKE* **294**, 93.

Balakrishnama, S. & Ganapathiraju, A. (1998). *Institute for Signal and information Processing*.

Beale, R. & Jackson, T. (1990). *Neural computing: an introduction*. Taylor & Francis.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. & Bourne, P. E. (2000). *Nucleic Acids Research* **28**, 235-242.

Beynon, R. J. & Easterby, J. S. (1996). *Buffer solutions*. IRL Press at Oxford University Press Oxford, UK.

Bolanos-Garcia, V. M. & Chayen, N. E. (2009). *Progress in biophysics and molecular biology* **101**, 3-12.

Bolboaca, S. D. & Jantschi, L. (2007). *Bulletin of University of Agricultural Sciences and Veterinary Medicine-Animal Sciences and Biotechnologies* **63**, 311-316.

Bonneté, F. (2007). *Crystal Growth and Design* **7**, 2176-2181.

Bragg, W. & Bragg, W. (1913). *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, 428-438.

Bruno, A. E., Ruby, A. M., Luft, J. R., Grant, T. D., Seetharaman, J., Montelione, G. T., Hunt, J. F. & Snell, E. H. (2014). *PloS one* **9**, e100782.

Brzozowski, A. M. & Walton, J. (2001). *Journal of applied crystallography* **34**, 97-101.

Bukrinsky, J. T. & Poulsen, J. C. N. (2001). *Journal of applied crystallography* **34**, 533-534.

Canaves, J. M., Page, R., Wilson, I. A. & Stevens, R. C. (2004). *Journal of molecular biology* **344**, 977-991.

Carter, C. W., Baldwin, E. T. & Frick, L. (1988). *Journal of crystal growth* **90**, 60-73.

Carter, C. W. & Carter, C. W. (1979). *Journal of Biological Chemistry* **254**, 12219-12223.

Caruana, R. & Niculescu-Mizil, A. (2006). *Proceedings of the 23rd international conference on Machine learning*, pp. 161-168. ACM.

CASIS (2013). *CASIS Announces Latest Funded Project for Protein Crystallization in Microgravity*, http://www.iss-casis.org/NewsEvents/PressReleases/tabid/111/ArticleID/66/ArtMID/586/CA

SIS-Announces-Latest-Funded-Project-for-Protein-Crystallization-in-Microgravity.aspx.
Charles, M., Veesler, S. & Bonneté, F. (2006). *Acta Crystallographica Section D: Biological Crystallography* **62**, 1311-1318.
Charoenkwan, P., Shoombuatong, W., Lee, H.-C., Chaijaruwanich, J., Huang, H.-L. & Ho, S.-Y. (2013). *PloS one* **8**, e72368.
Chayen, N. E. (2003). *Journal of structural and functional genomics* **4**, 115-120.
Chayen, N. E. (2004). *Current Opinion in Structural Biology* **14**, 577-583.
Chayen, N. E. & Saridakis, E. (2008). *Nature Methods* **5**, 147-153.
Chen, K., Kurgan, L. & Rahbari, M. (2007). *Biochemical and biophysical research communications* **355**, 764-769.
Chen, L., Oughtred, R., Berman, H. M. & Westbrook, J. (2004). *Bioinformatics* **20**, 2860-2862.
Chirgadze, D. (2001). *Protein Crystallisation in Action* Cambridge.
Chou, P. Y. & Fasman, G. D. (1977). *Trends in Biochemical Sciences* **2**, 128-131.
Christopher, G., Phipps, A. & Gray, R. (1998). *Journal of crystal growth* **191**, 820-826.
Coates, L., Tomanicek, S., Schrader, T. E., Weiss, K. L., Ng, J. D., Juettmer, P. & Ostermann, A. (2014). *Applied Crystallography* **47**.
Cox, T. F. & Cox, M. A. (2010). *Multidimensional scaling*. CRC Press.
Creamer, T. P. (2000). *Proteins: Structure, Function, and Bioinformatics* **40**, 443-450.
Crouch, S. R. & Ingle, J. D. (1988). *Spectrochemical analysis*. Prentice Hall.
Cudney, R., Patel, S., Weisgraber, K., Newhouse, Y. & McPherson, A. (1994). *Acta Crystallographica Section D: Biological Crystallography* **50**, 414-423.
Cunningham, P. (2000). Trinity College Dublin, Department of Computer Science.
D'Arcy, A. (1994). *Acta Crystallographica Section D: Biological Crystallography* **50**, 469-471.
Dale, G. E., Oefner, C. & D'Arcy, A. (2003). *Journal of structural biology* **142**, 88-97.
DeLucas, L. J., Bray, T. L., Nagy, L., McCombs, D., Chernov, N., Hamrick, D., Cosenza, L., Belgovskiy, A., Stoops, B. & Chait, A. (2003). *Journal of structural biology* **142**, 188-206.
Dor, O. & Zhou, Y. (2007). *Proteins: Structure, Function, and Bioinformatics* **66**, 838-845.
Ducruix, A. & Giegé, R. (1992). *Crystallization of nucleic acids and proteins: a practical approach*. IRL Press at Oxford University Press.
Eaton, W. A., Henry, E. R., Hofrichter, J. & Mozzarelli, A. (1999). *Nature Structural & Molecular Biology* **6**, 351-358.
Elands, J. & Hax, W. (2004). CryoEM as a complement to current techniques in protein structural analysis, Netherlands: FEI Company.
Engelman, D., Steitz, T. & Goldman, A. (1986). *Annual review of biophysics and biophysical chemistry* **15**, 321-353.
Farley, C. & Juers, D. H. (2014). *Journal of structural biology* **188**, 102-106.
Farr Jr, R. G., Perryman, A. L. & Samudzi, C. T. (1998). *Journal of crystal growth* **183**, 653-668.
Fazio, V. J., Peat, T. S. & Newman, J. (2014). *Structural Biology and Crystallization Communications* **70**, 1303-1311.
Fink, J. L., Weissig, H. & Bourne, P. E. (2009). *Structural Bioinformatics* **44**, 321.
Fisher, R. A. (1942). *The design of experiments.*

Fleissner, M. R., Cascio, D. & Hubbell, W. L. (2009). *Protein Science* **18**, 893-908.
Foster, L. S. & Gruntfest, I. J. (1937). *Journal of Chemical Education* **14**, 274.
Fowlkes, E. B. & Mallows, C. L. (1983). *Journal of the American Statistical Association* **78**, 553-569.
Garnier, J., Gibrat, J. F. & Robson, B. (1996). *Methods in enzymology* **266**, 540.
Gasteiger, E., Hoogland, C., Gattiker, A., Duvaud, S., Wilkins, M. R., Appel, R. D. & Bairoch, A. (2005). *Protein Identification and Analysis Tools on the ExPASy Server,*
George, A. & Wilson, W. W. (1994). *Acta Crystallographica Section D: Biological Crystallography* **50**, 361-365.
Gilliland, G. L. (1988). *Journal of crystal growth* **90**, 51-59.
Gilliland, G. L. & Davies, D. R. (1984). *Methods in enzymology* **104**, 370-381.
Gilliland, G. L., Tung, M. & Ladner, J. (1996). *Journal of Research-National Institute of Standards and Technology* **101**, 309-320.
Goh, C. S., Lan, N., Douglas, S. M., Wu, B., Echols, N., Smith, A., Milburn, D., Montelione, G. T. & Zhao, H. (2004). *Journal of molecular biology* **336**, 115-130.
Gong, C.-X., Liu, F., Grundke-Iqbal, I. & Iqbal, K. (2005). *Journal of neural transmission* **112**, 813-838.
Gorrec, F. (2009). *Journal of applied crystallography* **42**, 1035-1042.
Guruprasad, K., Reddy, B. V. B. & Pandit, M. W. (1990). *Protein engineering* **4**, 155-161.
Hadley, C. & Jones, D. T. (1999). *Structure* **7**, 1099-1112.
Hagan, M. T. & Menhaj, M. B. (1994). *Neural Networks, IEEE Transactions on* **5**, 989-993.
Hall, M. A. & Smith, L. A. (1997).
Hampton (2012). PEG Stability: A Look at pH and Conductivity Changes over Time in Polyethylene Glycols Hampton Research Corporation.
Hennessy, D., Buchanan, B., Subramanian, D., Wilkosz, P. A. & Rosenberg, J. M. (2000). *Acta Crystallographica Section D: Biological Crystallography* **56**, 817-827.
Hosfield, D., Palan, J., Hilgers, M., Scheibe, D., McRee, D. E. & Stevens, R. C. (2003). *Journal of structural biology* **142**, 207-217.
Howard, E. I., Blakeley, M. P., Haertlein, M., Haertlein, I. P., Mitschler, A., Fisher, S. J., Siah, A. C., Salvay, A. G., Popov, A. & Dieckmann, C. M. (2011). *Journal of Molecular Recognition* **24**, 724-732.
Huber, T. & Kobe, B. (2004). *Bioinformatics*.
Hui, R. & Edwards, A. (2003). *Journal of structural biology* **142**, 154-161.
Illingworth, J. A. (1981). *Biochem. J* **195**, 259-262.
Jahandideh, S. & Mahdavi, A. (2012). *Journal of Theoretical Biology*.
Jahnke, W. & Widmer, H. (2004). *Cellular and Molecular Life Sciences CMLS* **61**, 580-599.
Jancarik, J. & Kim, S. H. (1991). *Journal of applied crystallography* **24**, 409-411.
Janssen, F. W. & Ruelius, H. W. (1968). *Biochimica et Biophysica Acta (BBA)-Enzymology* **151**, 330-342.
Jaskolski, M., Dauter, Z. & Wlodawer, A. (2014). *FEBS Journal*.
Jayachandran, R., Sundaramurthy, V., Combaluzier, B., Mueller, P., Korf, H., Huygen, K., Miyazaki, T., Albrecht, I., Massner, J. & Pieters, J. (2007). *Cell* **130**, 37-50.
Johnson, S. C. (1967). *Psychometrika* **32**, 241-254.

Jurisica, I., Rogers, P., Glasgow, J. I., Fortier, S., Luft, J. R., Wolfley, J. R., Bianca, M. A., Weeks, D. R. & DeTitta, G. T. (2001). *IBM Systems Journal* **40**, 394-409.
Jurnak, F. (1986). *Journal of crystal growth* **76**, 577-582.
Kam, Z., Shore, H. & Feher, G. (1978). *Journal of molecular biology* **123**, 539-555.
Kandaswamy, K. K., Pugalenthi, G., Suganthan, P. & Gangal, R. (2010). *Protein and peptide letters* **17**, 423-430.
Kantardjieff, K., Jamshidian, M. & Rupp, B. (2004). *Published in: Bioinformatics, vol. 20, no. 14, September 1, 2004, pp. 2172-2174* **20**.
Kantardjieff, K. A. & Rupp, B. (2004). *Bioinformatics* **20**, 2162-2168.
Kawashima, S., Pokarowski, P., Pokarowska, M., Kolinski, A., Katayama, T. & Kanehisa, M. (2008). *Nucleic Acids Research* **36**, D202-D205.
Kimber, M. S., Vallee, F., Houston, S., Nečakov, A., Skarina, T., Evdokimova, E., Beasley, S., Christendat, D., Savchenko, A. & Arrowsmith, C. H. (2003). *Proteins: Structure, Function, and Bioinformatics* **51**, 562-568.
Kohavi, R. & John, G. H. (1997). *Artificial intelligence* **97**, 273-324.
Kohlmann, F. (2003). What is pH and How is it Measured?, edited by H. Company, p. 7.
Kozlowski, L. (2012). *Calculation of protein isoelectric point*, http://isoelectric.ovh.org/.
Kratochvíl, P. (1987).
Kretsinger, R. H. (1976). *Annual review of biochemistry* **45**, 239-266.
Kurgan, L., Razib, A. A., Aghakhani, S., Dick, S., Mizianty, M. & Jahandideh, S. (2009). *BMC structural biology* **9**, 50.
Kyte, J. & Doolittle, R. F. (1982). *Journal of molecular biology* **157**, 105-132.
Laufberger, V. (1937). *Bull. Soc. chim. biol* **19**, 4582.
Laurent, T. (1963). *Biochemical Journal* **89**, 253.
Lee, J. C. & Lee, L. (1981). *Journal of Biological Chemistry* **256**, 625-631.
Longenecker, K. L., Garrard, S. M., Sheffield, P. J. & Derewenda, Z. S. (2001). *Acta Crystallographica Section D: Biological Crystallography* **57**, 679-688.
Lorber, B., Skouri, M., Munch, J. P. & Giegé, R. (1993). *Journal of crystal growth* **128**, 1203-1211.
Lu, H.-M., Yin, D.-C., Liu, Y.-M., Guo, W.-H. & Zhou, R.-B. (2012). *International Journal of Molecular Sciences* **13**, 9514-9526.
Luft, J. R., Collins, R. J., Fehrman, N. A., Lauricella, A. M., Veatch, C. K. & DeTitta, G. T. (2003). *Journal of structural biology* **142**, 170-179.
Luft, J. R., Wolfley, J. R. & Snell, E. H. (2011). *Crystal Growth & Design* **11**, 651-663.
MacQueen, J. (1967). *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, p. 14. California, USA.
Madden, T. (2012). Chapter 16. The BLAST Sequence Analysis Tool. The NCBI Handbook; 2002.
Maithili, A., Kumari, R. V. & Rajamanickam, M. S. (2011). *International Journal of Modern Engineering Research* **1**, 55-64.
MathWorks (2011). *MATLAB R2011a.*
MathWorks (2013). *trainlm*, http://www.mathworks.co.uk/help/nnet/ref/trainlm.html.
McDermott, A. E. (2004). *Current opinion in structural biology* **14**, 554-561.
McPherson, A. (1982). *Preparation and analysis of protein crystals*. John Wiley & Sons.
McPherson, A. (1989a). *Preparation and analysis of protein crystals*. Krieger.

McPherson, A. (1989b). *European Journal of Biochemistry* **189**, 1-23.
McPherson, A. (1991). *Journal of crystal growth* **110**, 1-10.
McPherson, A. (1992). *Journal of crystal growth* **122**, 161-167.
McPherson, A. (1995). *Journal of applied crystallography* **28**, 362-365.
McPherson, A. (1999). *Crystallization of biological macromolecules*. Cold Spring Harbor Laboratory Press Cold Spring Harbor, NY.
McPherson, A. (2001). *Protein Science* **10**, 418-422.
McPherson Jr, A. (1976). *Journal of Biological Chemistry* **251**, 6300-6303.
MFPL (2012). *The Max F. Perutz International PhD Program*, http://www.mfpl.ac.at/phd-program/max-f-perutz/.
Mikol, V., Rodeau, J.-L. & Giegé, R. (1989). *Journal of applied crystallography* **22**, 155-161.
Milne, J. L., Borgnia, M. J., Bartesaghi, A., Tran, E. E., Earl, L. A., Schauder, D. M., Lengyel, J., Pierson, J., Patwardhan, A. & Subramaniam, S. (2013). *FEBS Journal* **280**, 28-45.
Mizianty, M. J. & Kurgan, L. (2009). *Biochemical and biophysical research communications* **390**, 10-15.
Mizianty, M. J. & Kurgan, L. (2011). *Bioinformatics* **27**, i24-i33.
Mizianty, M. J. & Kurgan, L. A. (2012). *Protein and peptide letters* **19**, 40-49.
MolecularDimensions (2015). *PEGs*, http://www.moleculardimensions.com/shopdisplayproducts.asp?id=324&cat=PEGs).
Morikawa, M., Yajima, H., Nitta, R., Inoue, S., Ogura, T., Sato, C. & Hirokawa, N. (2015). *The EMBO journal*.
Navia, M. A. & Murcko, M. A. (1992). *Current Opinion in Structural Biology* **2**, 202-210.
Newman, J. (2004). *Acta Crystallographica Section D: Biological Crystallography* **60**, 610-612.
Newman, J., Bolton, E., Muller-Dieckmann, J., Fazio, V., Gallagher, D., Lovell, D., Luft, J., Peat, T., Ratcliffe, D. & Sayle, R. (2012). *Acta Crystallographica Section F: Structural Biology and Crystallization Communications* **68**, 0-0.
Newman, J., Egan, D., Walter, T. S., Meged, R., Berry, I., Ben Jelloul, M., Sussman, J. L., Stuart, D. I. & Perrakis, A. (2005). *Acta Crystallographica Section D: Biological Crystallography* **61**, 1426-1431.
Newman, J., Fazio, V. J., Lawson, B. & Peat, T. S. (2010). *Crystal Growth & Design* **10**, 2785-2792.
Newman, J., Sayle, R. A. & Fazio, V. J. (2012). *Acta Crystallographica Section D: Biological Crystallography* **68**, 1003-1009.
Newman, J., Xu, J. & Willis, M. C. (2007). *Acta Crystallographica Section D: Biological Crystallography* **63**, 826-832.
NHS (2013). *Creutzfeldt-Jakob disease* http://www.nhs.uk/Conditions/Creutzfeldt-Jakob-disease/Pages/Introduction.aspx.
Nookala, G. K. M., Orsu, N., Pottumuthu, B. K. & Mudunuri, S. B. (2013). *International Journal*.
Overton, I. M. & Barton, G. J. (2006). *FEBS letters* **580**, 4005-4009.
Overton, I. M., Padovani, G., Girolami, M. A. & Barton, G. J. (2008). *Bioinformatics* **24**, 901-907.
Pace, C. N., Vajdos, F., Fee, L., Grimsley, G. & Gray, T. (1995). *Protein Science* **4**, 2411-2423.

Page, R., Grzechnik, S. K., Canaves, J. M., Spraggon, G., Kreusch, A., Kuhn, P., Stevens, R. C. & Lesley, S. A. (2003). *Acta Crystallographica Section D: Biological Crystallography* **59**, 1028-1037.
Page, R. & Stevens, R. C. (2004). *Methods* **34**, 373-389.
Paschos, V. T. (1997). *ACM Computing Surveys (CSUR)* **29**, 171-209.
PDB (2015). *Yearly Growth of Total Structures*, http://www.rcsb.org/pdb/statistics/contentGrowthChart.do?content=total&seqid=100.
Peat, T. S., Christopher, J. A. & Newman, J. (2005). *Acta Crystallographica Section D: Biological Crystallography* **61**, 1662-1669.
Perutz, M. F., Rossmann, M. G., Cullis, A. F., Muirhead, H. & Will, G. (1960). *Nature* **185**, 416-422.
Pietzsch, J. (2002). *Nature*.
Po, H. N. & Senozan, N. (2001). *Journal of Chemical Education* **78**, 1499.
Poznanski, J., Szczesny, P., Ruszczyńska, K., Zielenkiewicz, P. & Paczek, L. (2013). *Biochemical and biophysical research communications* **430**, 741-744.
R Core Team (2012). *R: A Language and Environment for Statistical Computing*.
Radaev, S. & Sun, P. D. (2002). *Journal of applied crystallography* **35**, 674-676.
Rajaraman, A. & Ullman, J. D. (2012). *Mining of massive datasets*. Cambridge University Press.
Ray Jr, W. J. & Puvathingal, J. M. (1985). *Analytical biochemistry* **146**, 307-312.
Reusch, W. (2013). *Visible and Ultraviolet Spectroscopy*, http://www2.chemistry.msu.edu/faculty/reusch/VirtTxtJml/Spectrpy/UV-Vis/spectrum.htm.
Rice, P., Longden, I. & Bleasby, A. (2000). *Trends in genetics* **16**, 276-277.
Rogers, S., Wells, R. & Rechsteiner, M. (1986). *Science (New York, NY)* **234**, 364.
Rose, G. D., Geselowitz, A. R., Lesser, G. J., Lee, R. H. & Zehfus, M. H. (1985). *Science* **229**, 834-838.
Rostkowski, M., Olsson, M. H., Søndergaard, C. R. & Jensen, J. H. (2011). *BMC structural biology* **11**, 6.
Rupp, B. (2003). *Journal of structural biology* **142**, 162-169.
Rupp, B. & Wang, J. (2004). *Methods* **34**, 390-407.
Samudzi, C. T., Fivash, M. J. & Rosenberg, J. M. (1992). *Journal of crystal growth* **123**, 47-58.
Segelke, B. W. (2001). *Journal of crystal growth* **232**, 553-562.
Sillero, A. & Maldonado, A. (2006). *Computers in biology and medicine* **36**, 157-166.
Silverstein, R. & Webster, F. (2006). *Spectrometric identification of organic compounds*. John Wiley & Sons.
Smialowski, P., Schmidt, T., Cox, J., Kirschner, A. & Frishman, D. (2006). *Proteins: Structure, Function, and Bioinformatics* **62**, 343-355.
Stevens, R. C. (2000). *Current Opinion in Structural Biology* **10**, 558-563.
Stura, E. A., Nemerow, G. R. & Wilson, I. A. (1992). *Journal of crystal growth* **122**, 273-285.
TargetDB (2010). *TargetDB Statistics Summary Report*, http://targetdb-dev.rutgers.edu/statistics/TargetStatistics.html.
Tarn, K. & Takács-Novák, K. (1999). *Pharmaceutical research* **16**, 374-381.
Teknomo, K. (2006). *Jaccard's Coefficient*, http://people.revoledu.com/kardi/tutorial/Similarity/Jaccard.html.

Tessier, P. M. & Lenhoff, A. M. (2003). *Current opinion in biotechnology* **14**, 512-516.
Tewary, S. K., Oda, T., Kendall, A., Bian, W., Stubbs, G., Wong, S.-M. & Swaminathan, K. (2011). *Journal of molecular biology* **406**, 516-526.
Tickle, I. J., Sibanda, B. L., Pearl, L. H., Hemmings, A. M. & Blundell, T. L. (1984). *X-ray crystallography and drug action*, pp. 427-440. Oxford: Clarendon Press.
Tran, T. T., Sorel, I. & Lewit-Bentley, A. (2004). *Acta Crystallographica Section D: Biological Crystallography* **60**, 1562-1568.
Tung, M. & Gallagher, D. (2008). *Acta Crystallographica Section D: Biological Crystallography* **65**, 18-23.
Valle, S., Li, W. & Qin, S. J. (1999). *Industrial & Engineering Chemistry Research* **38**, 4389-4401.
Van der Maaten, L., Postma, E. & Van Den Herik, H. (2009). *Journal of Machine Learning Research* **10**, 1-41.
Varshavsky, A. (1997). *Genes to Cells* **2**, 13-28.
Vivares, D. & Bonneté, F. (2002). *Acta Crystallographica Section D: Biological Crystallography* **58**, 472-479.
Von Dreele, R., Stephens, P., Smith, G. & Blessing, R. (2000). *Acta Crystallographica Section D: Biological Crystallography* **56**, 1549-1553.
Wan, H. & Wootton, J. C. (2000). *Computers & chemistry* **24**, 71-94.
Wang, H., Wang, M., Tan, H., Li, Y., Zhang, Z. & Song, J. (2014). *PloS one* **9**, e105902.
Ward, K., Wishner, B., Lattman, E. & Love, W. (1975). *Journal of molecular biology* **98**, 161-177.
Weber, P. C. (1997). *Methods in enzymology* **276**, 13-22.
Wendemuth, A., Opper, M. & Kinzel, W. (1993). *Journal of Physics A: Mathematical and General* **26**, 3165.
Westbrook, J., Feng, Z., Chen, L., Yang, H. & Berman, H. M. (2003). *Nucleic Acids Research* **31**, 489-491.
Woods, J. & Mellon, M. (1941). *The Journal of Physical Chemistry* **45**, 313-321.
Wooh, J. W., Kidd, R. D., Martin, J. L. & Kobe, B. (2003). *Acta Crystallographica Section D: Biological Crystallography* **59**, 769-772.
Wright, P. E. & Dyson, H. J. (1999). *Journal of molecular biology* **293**, 321-331.
Wu, Z., Gogonea, V., Lee, X., Wagner, M. A., Li, X.-M., Huang, Y., Undurti, A., May, R. P., Haertlein, M. & Moulin, M. (2009). *Journal of Biological Chemistry* **284**, 36605-36619.
Xie, X. Z., Chen, R. Q., Wu, Z. Q., Cheng, Q. D., Shang, P. & Yin, D. C. (2012). *Journal of applied crystallography* **45**, 758-765.
Yin, D.-C., Chen, R.-Q., Xie, S.-X., Liu, Y.-M., Zhang, X.-F., Zhu, L., Liu, Z.-T. & Shang, P. (2010). *Journal of applied crystallography* **43**, 1021-1026.
Yonekura, K., Kato, K., Ogasawara, M., Tomita, M. & Toyoshima, C. (2015). *Proceedings of the National Academy of Sciences* **112**, 3368-3373.
Yoshizaki, I., Nakamura, H., Fukuyama, S., Komatsu, H. & Yoda, S. (2004). *Annals of the New York Academy of Sciences* **1027**, 28-47.
Zhang, C.-Y., Wu, Z.-Q., Yin, D.-C., Zhou, B.-R., Guo, Y.-Z., Lu, H.-M., Zhou, R.-B. & Shang, P. (2013). *Acta Crystallographica Section F: Structural Biology and Crystallization Communications* **69**, 821-826.
Zhu, D. W., Garneau, A., Mazumdar, M., Zhou, M., Xu, G. J. & Lin, S. X. (2006). *Journal of structural biology* **154**, 297-302.

Zurich, S. F. I. o. T. *Linear Discriminant Analysis*, http://stat.ethz.ch/R-manual/R-devel/library/MASS/html/lda.html.
Zurich, S. F. I. o. T. (2012). *Principal Components Analysis*, http://stat.ethz.ch/R-manual/R-patched/library/stats/html/princomp.html.