# An *in vivo* assay for the high-throughput analysis and directed evolution of biopharmaceuticals



## **Romany Jane McLure**

School of Molecular and Cellular Biology Astbury Centre for Structural Molecular Biology University of Leeds

Submitted in accordance with the requirements for the degree of Doctor of Philosophy

March 2023

For Steven

### Declaration

The candidate confirms that the work submitted is her own. The candidate confirms that appropriate credit has been given within the thesis where reference has been made to the work of others. This copy has been supplied on the understanding that it is copyright material and that no quotation from this thesis may be published without proper acknowledgement.

Throughout this thesis the work directly attributable to the candidate is as follows: (i) Literature research and compilation of the manuscript stated above.

(ii) The candidate performed all the experimental work and data analysis unless otherwise stated.

Chapter 1 includes work from the following publication: McLure, R. J., Radford, S. E., and Brockwell, D. J. (2022). High-Throughput Directed Evolution: A Golden Era for Protein Science. *Trends in Chemistry*, 4(5).

Romany Jane McLure March 2023

### Acknowledgements

Thanks must of course go to my supervisors, Professors Sheena Radford and David Brockwell, for their support and guidance over the last four years. Thank you for your helpful suggestions, encouragement, and words of motivation when things were not going so well.

I wish to thank the BBSRC Doctoral Training Program and UCB Biopharma UK for funding me playing around with *E. coli* and proteins for the last 4 years. Particularly Dr John O'Hara, Dr Oliver Durrant, Mr Mark Ellis, and Dr Michael Knight (UCB) for providing additional funding, materials, experimental support and advice over the course of my PhD and particularly during my placement.

I am very grateful to have been able to work in the diverse and supportive environment that is the Radford-Brockwell lab. I am thankful for all the people I have crossed paths with in my 4 years at Leeds. Particularly G Nasir Khan, who was always there to order all the kits and reagents I could ever want, listen to any of my stresses of the day, and just for generally being a good friend. Also for keeping me and Jim well-fed with the most delicious spicy curries, we would be much skinnier without you. To my fabulous cohort, James Whitehouse, Samantha Lawrence and Sam Haysom, who have struggled alongside me and always been there for advice and support, both sciencey and not. Particularly Sammy, who has been a massive support throughout my PhD in so many ways, and was the absolute best Maid of Honour! To all the Radford-Brockwell members, old and new, who have made coming into work every day fun. Particular thanks go to Dr Mike Davies (Mike who?), not only for all your help with the coding in this thesis but also for all the nights climbing and all the stolen cats. To Dr Nicolas Guthertz, for teaching me crystallography and putting up with my terrible attempts at French. To Katy Dewison (all hail the worm queen), Emily Byrd, Conor McKay and Ailsa MacRae, for helping me get through the long days and cope with the never ending pile of MIC assays and PCRs.

As always, I have to thank my family for always being there supporting and encouraging me in everything that I do. To my grandparents, Angus and Barbara, who made me into the

person I am today and always encouraged me to push myself further. To my mum, Debbie, who always believes in me even when I don't.

I will forever be grateful for the kindness, guidance, and support of Dr Steven James Lee. I miss you every day. You were the greatest friend I ever had the privilege of knowing, you encouraged me to take this PhD and I hope you would be proud.

Finally, the most thanks have to go to my wonderful husband Jim. I was lucky enough to meet you in this lab in the first month of my PhD and your support since then has been endless, both science and otherwise. My PhD and this thesis would never have been so successful without you, your kindness and your patience and your never-ending fountain of knowledge, so this is for you. I look forward to a lifetime of confusing people when they ask for Dr Horne.

### Abstract

Directed evolution is a robust and powerful tool for engineering new and/or improved functions in biomolecules for therapeutic and industrial applications, as well as to uncover fundamental insights into protein behaviour. It works by exploiting the principles of natural evolution and accelerating it through multiple rounds of gene diversification and selection. In order to evolve the desired property, an appropriate assay for the property of interest must be chosen. However, improving proteins often proves challenging as most mutations are destabilising.

An *in vivo* TriPartite  $\beta$ -Lactamase Assay (TPBLA) has been shown to rapidly and easily identify misfolded, unstable or aggregation-prone sequences without the need for protein purification. Furthermore, TPBLA has successfully been utilised as a directed evolution screen to evolve thermodynamic stability and aggregation-resistance in a protein of interest. However, the methodology was limited in throughput, and due to the use of first-generation sequencing techniques identification of improved variants was laborious and high-cost. Chapter 3 in this thesis develops a new methodology for combining TPBLA with next-generation sequencing to enable high-throughput identification of hotspot regions and improved variants. This new approach has the potential to assess hundreds to thousands of variants in a single experiment and give a more comprehensive and extensive overview of a proteins' fitness landscape. In Chapter 4, this new high-throughput methodology is applied to biopharmaceutically-relevant targets to improve their aggregation behaviour. Furthermore, the ability of TPBLA to screen and rank a panel of clinically-relevant antibody therapeutics based on their developability is assessed. This demonstrated the potential of TPBLA for identifying poorly developable candidates, as well as potential late-stage clinical failures early in development prior to protein purification.

A common challenge for directed evolution studies is that often there is a trade-off between particular properties, such as stability and function, and by selecting for one you can negatively impact the other. Therefore, in the absence of a selection for function, evolving biopharmaceutical test proteins using TPBLA to improve their aggregation resistance could result in evolved variants that no longer bind to their target. Therefore, the work in Chapter 5 develops a novel assay, Solubility 'n' Affinity Coselection (SnAC), which introduces a selection for binding into TPBLA evolution experiments to enable evolution of biologics for both stability and function.

Overall, the work presented in this thesis details a novel and powerful approach for the analysis and directed evolution of stability and binding in biopharmaceutically-relevant proteins.

# **Table of contents**

L	ist of f	igures		xix
L	ist of 1	tables	2	XXV
N	omen	clature	x	xix
1	Intr	oductio	n	1
	1.1	Protein	n folding, misfolding and aggregation	1
		1.1.1	Principles of protein folding	1
		1.1.2	Mechanisms of aggregation	3
		1.1.3	Chemically induced protein aggregation	4
		1.1.4	Factors affecting protein aggregation	6
		1.1.5	Structure and morphology of protein aggregates	7
	1.2	Antibo	ody fragments and derivatives	8
		1.2.1	Antibodies	8
		1.2.2	Antibody fragments	10
	1.3	Bioph	armaceuticals and the aggregation problem	12
		1.3.1	History of biopharmaceuticals	12

	1.3.2	Monoclonal antibodies and their generation	13
		1.3.2.1 Hybridoma technology	13
		1.3.2.2 Phage display	15
		1.3.2.3 Yeast surface display	18
		1.3.2.4 Ribosome display	20
		1.3.2.5 Mammalian display	21
	1.3.3	Aggregation in biopharmaceuticals	21
1.4	Metho	ds to assess biopharmaceuticals	24
	1.4.1	Analytical techniques	24
		1.4.1.1 Assessing protein aggregation	24
		1.4.1.2 Assessing binding affinity	30
	1.4.2	Computational approaches	32
1.5	Preven	tion and inhibition of protein aggregates	36
	1.5.1	Promotion of protein refolding	36
1.6	Proteir	n engineering	36
	1.6.1	Rational design	37
	1.6.2	Directed evolution	40
		1.6.2.1 Creating DNA libraries	40
		1.6.2.2 Screening technologies	45
	1.6.3	Deep mutational scanning (DMS)	51
1.7	Tripart	ite $\beta$ -Lactamase Assay (TPBLA)	55
	1.7.1	Peptidoglycan biosynthesis inhibition by $\beta\mbox{-lactam}$ antibiotics	55
	1.7.2	$\beta$ -lactamase enzyme	55

		1.7.3	$\beta$ -lactamase as a reporter protein	57
		1.7.4	Tripartite $\beta$ -Lactamase Assay (TPBLA)	59
	1.8	Aims o	of this thesis	63
2	Mat	erials a	nd methods	65
	2.1	Materi	als	65
		2.1.1	Chemicals and kits	65
		2.1.2	Enzymes for molecular biology	69
		2.1.3	Media	69
		2.1.4	Buffers	70
		2.1.5	Antibiotics	71
	2.2	Molec	ular biology methods	71
		2.2.1	Bacterial strains	71
		2.2.2	<i>E. coli</i> transformation	72
		2.2.3	Polymerase chain reaction	72
		2.2.4	Agarose gel electrophoresis	74
		2.2.5	Restriction digest of plasmid DNA	74
		2.2.6	Dephosphorylation of restriction digests	75
		2.2.7	Ligation of DNA	75
		2.2.8	Q5 site-directed mutagenesis	76
			2.2.8.1 Amplification of DNA	76
			2.2.8.2 Kinase, ligase, Dpnl (KLD) treatment	76
		2.2.9	Sequencing and storage of plasmid DNA	77

	2.2.10	Plasmids	and primers	77
2.3	Protein	expression	on and purification	79
	2.3.1	Purificati	on of MBP constructs	79
		2.3.1.1	Small-scale expression trials	79
		2.3.1.2	Large-scale expression of protein constructs	79
		2.3.1.3	Refolding of protein from inclusion bodies	80
		2.3.1.4	HisTrap purification	80
		2.3.1.5	Sodium dodecyl sulfate polyacrylamide gel electrophore- sis	80
		2.3.1.6	TEV protease treatment	81
		2.3.1.7	Gel filtration chromatography	81
		2.3.1.8	Mass spectrometry	82
	2.3.2		ion of IgG and Fab proteins from chinese hamster ovary ells	82
2.4	In vitro	e technique	es	83
	2.4.1		usion chromatography multi angle light scattering (SEC-	83
	2.4.2	Circular	dichroism (CD) spectroscopy	83
	2.4.3	Urea den	aturation	83
	2.4.4		onapthalene-1-sulphonic acid (ANS) fluorescence spec-	84
	2.4.5	X-ray cry	vstallography	84
	2.4.6	Hydroph	obic interaction chromatography (HIC)	85
	2.4.7	-	capture self-interaction nanoparticle copy (AC-SINS)	85

	2.4.8	Differential scanning fluorimetry (DSF)	86
	2.4.9	Dot blot analysis	87
2.5	Tripart	tite $\beta$ -lactamase assay (TPBLA)	88
	2.5.1	Preparation of 48-well agar plates	88
	2.5.2	Culture inoculation and induction	88
2.6	β-lacta	mase evolution bioassay	89
	2.6.1	Vector design	89
	2.6.2	Library creation	89
		2.6.2.1 Error-prone PCR	89
		2.6.2.2 Electroporation of TG1 cells	91
	2.6.3	Directed evolution	92
2.7	Next-g	generation sequencing	93
	2.7.1	Illumina and Pacbio Sequencing	93
	2.7.2	EZ-Amplicon sequencing	93
	2.7.3	Illumina fragment and Pacbio sequencing analysis	93
	2.7.4	Illumina EZ-Amplicon sequencing analysis	94
2.8	Hierar	chical clustering	95
2.9	Multip	le linear regression analysis for predicting TPBLA score	95
	2.9.1	AMSCI mAbs	95
	2.9.2	Jain mAbs	99
2.10		ds an assay for the simultaneous evolution of aggregation resistance nding affinity	00
	2.10.1	Western blot analysis	00

		2.10.2	Fluorescence spectroscopy	101
		2.10.3	Plate reader fluorescence	102
		2.10.4	Fluorescence activated cell sorting	102
		2.10.5	Dual selection screening	103
			2.10.5.1 TPBLA screen for solubility	103
			2.10.5.2 FACS screen for binding affinity	103
3	Com	bining	deep sequencing with TPBLA for directed evolution	105
	3.1	Introdu	ction	105
		3.1.1	Maltose binding protein	107
		3.1.2	Directed evolution	108
		3.1.3	Aims of the study	108
	3.2	Results	3	109
		3.2.1	TPBLA can be used to assess different sequence liabilities	109
		3.2.2	Golden gate assembly can be used to robustly create large libraries	120
		3.2.3	Illumina shotgun libraries allow identification of hotspots and single point mutations enriched due to selection	125
		3.2.4	Pacbio sequencing allows analysis of co-evolution and identifica- tion of sequences with enhanced properties	128
	3.3	Discus	sion	136
4	App	lying TI	PBLA to assess and evolve therapeutically relevant proteins	139
	4.1	Introdu	iction	139
		4.1.1	Aims of this chapter	140
	4.2	Results	8	140

		4.2.1	TPBLA can be used to screen and rank biotherapeutics	140
		4.2.2	TPBLA does not correlate to one single biophysical parameter	148
		4.2.3	Multiple regression can be used to rationalize performance in TPBLA	153
		4.2.4	TPBLA can be used to evolve therapeutic scaffolds to improve their biophysical properties	157
		4.2.5	Evolved single point mutations of AMS134 improve aggregation behaviours	165
		4.2.6	Analysis of 35 clinical late stage therapeutics demonstrates no correlation between MIC and any single developability assay	173
		4.2.7	Multiple regression models can be used to rationalise TPBLA score	177
		4.2.8	Multiple regression models can predict TPBLA score based on performance in other developability assays	181
	4.3	Discus	sion	185
5	Tow TPB		nultaneous improvement of both aggregation and binding with	189
	5.1	Introdu	action	189
		5.1.1	Aims of the study	190
	5.2	Results	s	190
		5.2.1	Split fluorescent proteins as sensors	190
		5.2.2	Split mNeonGreen2 combined with TPBLA is not able to detect binding affinity <i>in vivo</i>	191
		5.2.3	CadC periplasmic sensor for screening binding affinity	195
		5.2.4	Adapting the <i>E. coli</i> CadC transmembrane transcriptional activator to include a selection for binding affinity into TPBLA	197

		5.2.5	Fusion of caffeine-inducible dimerising nanobody (VHH) to TP-BLA does not inhibit $\beta$ -lactamase activity $\ldots \ldots \ldots \ldots$	199
		5.2.6	CadC can be used to measure cognate binding in the periplasm	200
		5.2.7	Flow cytometry can be used to identify binders	205
		5.2.8	SnAC can be used to screen a library of variants to identify the most stable and highest affinity variant	207
	5.3	Discuss	sion	211
6	Fina	l conclu	isions	215
	6.1	Overall	l conclusion of results	215
	6.2	Future	work	220
	6.3	Final re	emarks	222
Re	feren	ces		223
Ар	pend	ix A P	rimers used in this study	253
Ар	pend	ix B Fi	irst derivatives of DSF and SLS data for AMSCI mAbs	261
Ар	pend	ix C M	Iultiple regression model statistics for AMSCI mAbs	267
Ар	pend	ix D M	Iultiple regression model statistics for Jain mAbs	269

# List of figures

1.1	Protein folding pathway and energy landscape	2
1.2	Mechanisms of aggregation.	5
1.3	Classes of human immunoglobulins.	9
1.4	Structures of mAbs and antibody fragments	11
1.5	Hybridoma technology for producing mAbs	14
1.6	Humanisation of mAbs	15
1.7	Phage display for producing mAbs	17
1.8	Yeast surface display for producing mAbs	19
1.9	Ribosome display for producing mAbs	20
1.10	Overview of mAb purification process.	23
1.11	Overview of methods to detect protein aggregation.	25
1.12	Continued overview of methods to detect protein aggregation	27
1.13	Surface Plasmon Resonance (SPR) to measure antibody-antigen binding affinities.	31
1.14	Direct and Indirect Enzyme-Linked Immunosorbent Assays (ELISAs)	32
1.15	Overview of protein engineering techniques	38
1.16	Directed evolution and <i>in vitro</i> mutagenesis	41

1.17	Schematic overview of <i>in vivo</i> diversification techniques
1.18	Overview of biosensors for evolving protein stability
1.19	Schematic of phage-assisted continuous evolution (PACE) and its related screens
1.20	Overview of Illumina sequencing technologies
1.21	Overview of Pacific biosciences SMRT sequencing technologies 54
1.22	Biosynthesis of peptidoglycan and its inhibition by $\beta$ -lactam antibiotics. 56
1.23	Structure of <i>E. coli</i> TEM-1 $\beta$ -lactamase
1.24	<i>In vivo</i> tripartite $\beta$ -lactamase assay (TPBLA)
1.25	Assessing aggregation using TPBLA
2.1	Mean sum of squares regression and error
3.1	Sequence alignment of MBP <sup>WT</sup> , MBP <sup>Y283D</sup> , and MBP <sup>4A</sup> 106
3.2	TPBLA analysis of MBP <sup>WT</sup> and variants
3.3	Purification of MBP <sup>WT</sup> in autoinduction media.
3.4	Purification of MBP <sup>Y283D</sup> in autoinduction media
3.5	Refolding and purification of MBP <sup>4A</sup> in autoinduction media
3.6	Biochemical characterisation of MBP <sup>WT</sup> , MBP <sup>Y283D</sup> , and MBP <sup>4A</sup> 114
3.7	Comparison of MBP <sup>WT</sup> , MBP <sup>Y283D</sup> and MBP <sup>4A</sup> crystal structures 116
3.8	MBP <sup>4A</sup> crystal structure
3.9	MBP <sup>4A</sup> forms a dimer in the crystal structure
3.10	Golden gate library preparation for directed evolution
3.11	1.5% ( $w/v$ ) agarose gel showing restriction digestion and golden gate reactions to assess the success of the golden gate library method 122

	3.12	1.5% ( <i>w/v</i> ) agarose gel colony PCR of blaMBP <sup>4A</sup> and blaMBP <sup>Y283D</sup> libraries.12	23
	3.13	Overview of directed evolution and data analysis of Illumina sequencing.	24
	3.14	Evolution of $MBP^{Y283D}$ and $MBP^{4A}$ analysed by Illumina sequencing 12	26
	3.15	<i>In vivo</i> screen of evolved MBP <sup>4A</sup> point mutants identified by Illumina sequencing.	27
	3.16	Evolution of MBP <sup>Y283D</sup> and MBP <sup>4A</sup> analysed by Pacbio sequencing 12	29
	3.17	Analysis of evolved MBP <sup>4A</sup> point mutants identified by Illumina and Pacbio sequencing.	32
	3.18	Expression trial of evolved MBP <sup>4A</sup> selected point mutants identified by Pacbio sequencing	33
	3.19	Analysis of evolved MBP <sup>4A</sup> selected point mutants identified by Illumina and Pacbio sequencing	34
	3.20	Dot blot against $\beta$ -lactamase for MBP <sup>WT</sup> , MBP <sup>Y283D</sup> , MBP <sup>4A</sup> and evolved variants.	35
2	4.1	Structures of mAbs and antibody fragments	42
2	4.2	Thermal stability and aggregation behaviour characterisation of 11 AMSCI mAbs	44
2	4.3	Biophysical characterisation of 11 AMSCI mAbs	45
2	4.4	TPBLA analysis of AMSCI mAbs in scFv format	45
2	4.5	Biophysical characterisation of 11 chosen IgGs	47
2	4.6	TPBLA plotted against biophysical characterisation of 11 AMSCI mAbs. 15	50
2	4.7	TPBLA rank plotted against biophysical characterisation of 11 AMSCImAbs.15	52
۷	4.8	Multiple regression models to predict TPBLA score from biophysical parameters.	55

4.9	Multiple regression models to predict TPBLA score from biophysical parameters.	156
4.10	Evolution of AMS134 analysed by Illumina sequencing	159
4.11	Evolution of AMS134 analysed by Illumina sequencing	160
4.12	AMS134 hotspots identified in Illumina sequencing mapped onto the scFv structure	161
4.13	Evolution of AMS197 analysed by Illumina sequencing	162
4.14	Evolution of AMS197 analysed by Illumina sequencing	163
4.15	AMS197 hotspots identified in Illumina sequencing mapped onto the scFv structure	164
4.16	<i>In vivo</i> screen of evolved AMS134 and AMS197 point mutants identified by Illumina sequencing	166
4.17	Biophysical characterisation of evolved AMS134 variants	168
4.18	Dynamic light scattering (DLS) of AMS134 evolved point mutants mea- sured before and after thermal melt.	169
4.19	Thermal stability and aggregation behaviour of evolved AMS134 variants.	170
4.20	AC-SINS absorbtion spectra and wavelength shifts of AMS134 evolved point mutants.	172
4.21	TPBLA analysis of 35 chosen Jain mAbs as scFvs	176
4.22	TPBLA plotted against biophysical characterisation of 35 Jain mAbs	179
4.23	Ranked TPBLA plotted against HEK titre, Tm1 by DSF, and ELISA for 35 Jain mAbs	180
4.24	Spearmans rank correlation and hierachical clustering of developability assays compared with TPBLA for 35 Jain mAbs.	181
4.25	Multiple regression models to predict TPBLA score from biophysical parameters.	182

4.26	Multiple regression models to predict TPBLA score from biophysical parameters.	33
4.27	Testing the regression model to predict TPBLA score for 6 test mAbs 18	34
5.1	Crystal structure of the HA4 monobody bound to the SH2 domain of human Ab1 kinase	<del>)</del> 2
5.2	Overview of TPBLA combined with a split fluorescent protein assay (Solubility 'n' Affinity Coselection (SnAC) 1.0) to assess binding affinity. 19	<del>)</del> 3
5.3	Western blot against SH2 showing SH2-mNG2 $_{1-10}$ localisation in the cell. 19	€
5.4	SnAC 1.0 using a split fluorescent protein is not able to measure binding affinity <i>in vivo</i>	€
5.5	Overview of Solubility 'n' Affinity Coselection (SnAC) 2.0 assay com- bined with TPBLA	€
5.6	TPBLA screen of cells co-transformed with CadC-SH2 and either blaHA4 <sup>WT</sup> -VHH (strong binder), blaHA4 <sup>Y87A</sup> -VHH (weak binder), blaHA4 <sup>2A</sup> -VHH (destabilised), blaHA4 <sup>Y87A 2A</sup> -VHH (destabilised), or blaGG <sub>STOP</sub> -VHH (negative control)	00
5.7	Fluorescence intensity endpoints of cells co-transformed with CadC-SH2 and either blaHA4 <sup>WT</sup> -VHH (strong binder) or blaHA4 <sup>Y87A</sup> -VHH (weak binder), or transformed with only CadC-SH2 (negative control) or CadC- VHH (positive control) alone	02
5.8	Fluorescence intensity over time of cells co-transformed with CadC-SH2 and either blaHA4 <sup>WT</sup> -VHH (strong binder) or blaHA4 <sup>Y87A</sup> -VHH (weak binder) at varying arabinose concentrations	)3
5.9	Fluorescence intensity over time of cells co-transformed with CadC-SH2 and either blaHA4 <sup>WT</sup> (strong binder) or blaHA4 <sup>Y87A</sup> (weak binder), or transformed with only CadC-SH2 (negative control) or CadC-VHH (posi- tive control) alone	)4
5.10	CadC-based sensor can measure binding in the periplasm and be used to sort positive cells using FACS	)6

5.11	SnAC 2.0 can identify positive binders in the periplasm and be used to sort positive cells using FACS
5.12	Proof of principle screening using SnAC and analysed using Illumina amplicon sequencing
5.13	CadC-based sensor can measure binding in the periplasm and be used to sort positive cells using FACS
<b>B</b> .1	Thermal stability of 11 AMSCI mAbs
B.2	Aggregation behaviour of 11 AMSCI mAbs
B.3	Thermal stability of AMS134 <sup>WT</sup> and evolved variants
B.4	Aggregation behaviour of AMS134 <sup>WT</sup> and evolved variants

## List of tables

2.1	Materials	65
2.2	Molecular biology enzymes	69
2.3	Media	69
2.4	Buffers	70
2.5	Antibiotics used in this study	71
2.6	Vent PCR components	73
2.7	Vent PCR thermocycling conditions	73
2.8	Q5 PCR components	73
2.9	Q5 PCR thermocycling conditions	74
2.10	Double restriction digest components	75
2.11	Dephosphorylation components	75
2.12	Ligation components	76
2.13	Q5 mutagenesis components	76
2.14	KLD treatment components	77
2.15	List of plasmids	78
2.16	SDS-PAGE recipe	81
2.17	AKTA program for gel filtration chromatography	82

2.18	Components for 48-well agar plates for TPBLA	88
2.19	epPCR mutation frequency	90
2.20	Components for epPCR	90
2.21	epPCR thermocycling conditions	91
2.22	Golden gate components	91
2.23	Golden gate thermocycling conditions	91
3.1	Crystallographic data collection and refinement statistics	19
3.2	Top mutations from $MBP^{Y283D}$ evolution	28
3.3	Top mutations from MBP <sup>4A</sup> evolution	30
3.4	Thermal and thermodynamic stabilities of MBP variants	33
3.5	Dot blot densitometry of soluble protein expression of MBP variants 13	35
4.1	Stability and aggregation behaviours of AMSCI mAbs	44
4.2	Stability and aggregation behaviours of AMS134 mAbs	67
4.3	Hydrodynamic radii of AMS134 IgGs measured by DLS before and after temperature ramp	67
4.4	Developability assays used to characterise the Jain mAbs	74
A.1	Primers used in Q5 mutagenesis and restriction digest cloning 25	53
A.2	Primers used in golden gate cloning	57
A.3	Primers used in sequencing	58
A.4	Primers used to amplify libraries for NGS	59
C.1	Linear regression statistics for AMSCI mAbs regression model with 5 parameters	60

C.2	Linear regression statistics for AMSCI mAbs regression model with 4 parameters
C.3	Linear regression statistics for AMSCI mAbs regression model with 3 parameters
C.4	Linear regression statistics for AMSCI mAbs regression model with 2 parameters
D.1	Linear regression statistics for Jain mAbs regression model with 7 parameters270
D.2	Linear regression statistics for Jain mAbs regression model with 6 parameters270
D.3	Linear regression statistics for Jain mAbs regression model with 5 parameters271
D.4	Linear regression statistics for 29 Jain mAbs regression model with 5 parameters

## Nomenclature

#### **Greek Symbols**

- βLa β-lactamase
- µg Microgram
- µM Micromolar

#### **Other Symbols**

- Ka Association constant
- K<sub>d</sub> Dissociation constant
- T<sub>a</sub> Annealing temperature
- T<sub>m</sub> Melting temperature

#### Acronyms / Abbreviations

- A<sub>260</sub> Absorbance at 260 nm
- A<sub>280</sub> Absorbance at 280 nm
- Ab Antibody
- AC-SINS Affinity capture self-interaction nanoparticle spectroscopy
- ADCC Antibody-dependent cellular toxicity
- Amp Ampicillin
- ANS 8-Anilino-1-napthalenesulfonic acid
- APR Aggregation-prone region

bp	Base pair	
CD	Circular dichroism	
CDR	Complementary determining region	
dAb	Single domain antibody	
Da	Dalton	
DF	Diafiltration	
DLS	Dynamic light scattering	
DMS	Deep Mutational Scanning	
DMSO Dimethyl sulfoxide		
DNA	Deoxyribonucleic acid	
DTT	Dithiothreitol	
EDTA	Ethylenediaminetetraacetic acid	
EM	Electron microscopy	
epPCR Error-prone polymerase chain reaction		
ESI-IM-MS Electrospray ionisation ion-mobility mass spectrometry		
EtOH	Ethanol	
Fab	Antigen binding fragment	
Fc	Crystalisable fragment	
FcRn	Fc receptor cycling	
FDA	Food and drug administration	
FP	Fluorescent protein	
Fv	Variable fragment	
GOI	Gene of interest	
GS linker Glycine-serine linker		

- HPLC High performance liquid chromatography
- hr Hour
- Ig Immunoglobulin
- kb Kilobases
- kDa Kilodalton
- LB Luria-Bertani
- mAb Monoclonal antibody
- MBP Maltose binding protein
- MCD<sub>GROWTH</sub> Maximum cell dilution allowing growth
- mCh mCherry
- MIC Minimal inhibitory concentration
- min Minute
- mNG mNeonGreen
- mol Mole
- MS/MS Mass spectrometry / mass spectrometry (Tandem mass spectrometry)
- mSc mScarlet
- MS Mass spectrometry
- MWCO Molecular weight cutoff
- MWM Molecular weight marker
- NMR Nuclear magnetic resonance
- OD600 Optical density at 600 nm
- PACE Phage Assisted Continuous Evolution
- PCR Polymerase chain reaction

- PDB Protein data bank POI Protein of interest psi Pounds per square inch RNA Ribonucleic acid rpm Rotations per minute scFv Single chain variable fragment SDM Site directed mutagenesis SDS-PAGE Sodium dodecyl sulfate polyacrylamide gel electrophoresis SEC-MALS Size exclusion chromatography multi angle light scattering SEC Size exclusion chromatography sfGFP Superfolder GFP sFP Split fluorescent protein SLIC Sequence and ligation independent cloning SnAC Solubility 'n' Affinity Coselection SOC Super Optimal broth with catabolite repression TAE Tris-acetate-EDTA TEM Transmission electron microscopy Tet Tetracycline TPBLA Tripartite β-Lactamase Assay UV Ultra-violet V<sub>H</sub> Variable heavy domain VL. Variable light domain v/v Volume:volume ratio WT Wild-type w/v Weight:volume ratio
- XL-MS Crosslinking mass-spectrometry

## Chapter 1

## Introduction

### **1.1** Protein folding, misfolding and aggregation

#### **1.1.1** Principles of protein folding

A protein's three-dimensional structure is determined by its primary sequence, its amino acid monomer chain linked together by peptide bonds. Proteins are thought to initially fold by hydrophobic collapse and burial of non-polar amino acid residues within the protein core, termed the hydrophobic effect, a stabilising and thermodynamically favourable process (Hartl et al., 2011). The structure they adopt, termed the native fold, is vital for a protein's function, and deviations from this can result in potentially catastrophic consequences (Wang and Roberts, 2018; Hartl et al., 2011). Anfinsen's pioneering research showing ribonuclease A can refold after denaturation demonstrated that folding is reversible, occurs without outside energy input, and proteins adopt the structure with the lowest free energy (Anfinsen et al., 1961; Anfinsen, 1973). Furthermore, it was concluded that all the information for a protein to adopt its final native state was contained within its primary sequence. However, a question remained that how does a protein sample the astronomically large number of potential conformations available to an unfolded polypeptide chain within a biologically relevant timescale? No protein, no matter how small, could sample all of these conformations within this time, a notion known as Levinthal's paradox (Levinthal, 1968). It was therefore proposed that the path to the native state takes more of a guided search, where it follows predetermined folding pathways using specific and controlled mechanisms and the formation of rapid local interactions (Figure 1.1A) (Levinthal, 1968). Levinthal's paradox led to a search for folding pathways, and

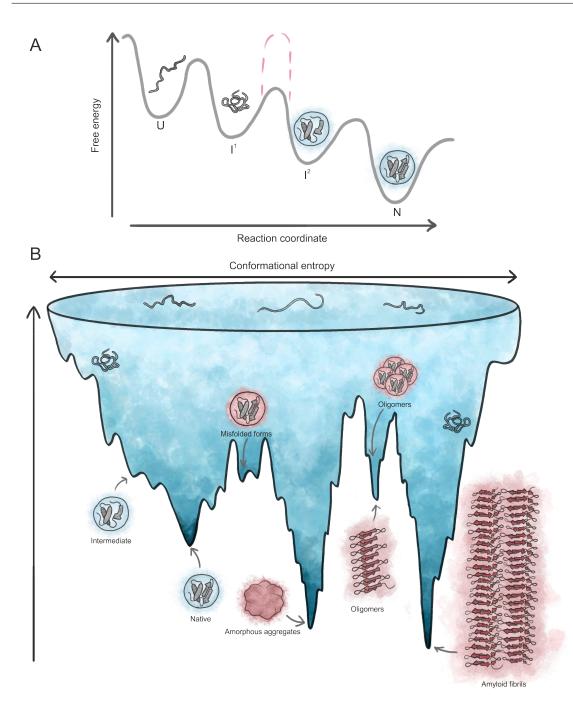


Fig. 1.1 **Protein folding pathway and energy landscape.** A) Levinthal's proposed protein folding theory whereby proteins fold to the native state via specific intermediates through a defined pathway. Intermediates can become trapped in 'kinetic wells' (dotted lines). B) Folding funnel theory whereby proteins fold energetically downhill making favourable intramolecular contacts and decreasing conformational entropy to the native state at a local energy minima. Amorphous aggregates and amyloid fibrils often exist in the global energy minima (they are more stable than the native fold). Unfolded, misfolded, intermediate and native-state proteins can aggregate via different pathways. Figure redrawn and adapted from (Jahn and Radford, 2005; Englander and Mayne, 2014).

the simplistic folding pathway theory whereby proteins fold via distinct intermediate states through a distinct pathway eventually evolved into one of a funnel-shaped energy landscape (Dill and Chan, 1997; Englander and Mayne, 2014). In this theory, the unfolded polypeptide chain folds by forming energetically favourable intramolecular contacts and reducing the conformational entropy as the protein is funnelled towards the local energy minima, or native state (Figure 1.1B) (Dill and Chan, 1997; Englander and Mayne, 2014). There is not one single folding pathway, but proteins may take various routes as they move towards the native state. However, the pathway to the native state is often 'rugged', meaning intermediate states must often pass so called 'kinetic barriers' during folding whereby proteins must input energy to overcome these barriers (Hartl et al., 2011). This can result in transient populations of partially folded states, as they are unable to overcome this high free-energy barrier, which have the potential to lead to protein aggregation (Jahn and Radford, 2005).

#### **1.1.2** Mechanisms of aggregation

An aggregate can broadly be defined as any self-associated protein with a quaternary structure different to that of the native fold (Ratanji et al., 2014). While oligomerisation can be desirable for some proteins as it is required for function, for others this is detrimental. Proteins are fundamentally aggregation-prone when unfolded or partially-folded, which can lead to the formation of extremely stable and long-lived aggregates (Roberts, 2014b). While various mathematical models exist to explain the phenomenon, there is no single mechanism of aggregation as it depends on both the protein and its environment and the same protein may aggregate via a variety of different mechanisms depending on this environment and the particular stresses it is subjected to (Roberts, 2014a). The formation of aggregates can be either reversible or irreversible depending on the strength and amount of interactions involved, however, it is important to note that no aggregate is technically completely irreversible, but it is irreversible within a biologically relevant timescale (Cromwell et al., 2006). Furthermore, the stage of the aggregation process often dictates reversibility of aggregate formation, as initial formation of aggregates is often reversible whereas latter stage aggregates are often irreversible (Wang, 2005).

Proteins are inherently dynamic by nature and experience constant global structural fluctuations in solution as well as local structural pertubations, which can result in exposure of otherwise buried aggregation-prone regions (APRs) (Krause et al., 2012; Abbas et al., 2013). These APR are often short hydrophobic stretches which when exposed can cause proteins to self-associate and form oligomers, a process generally thought of as the initial

aggregation event (Wang and Roberts, 2018). At this point aggregation can be easily reversed by dilution or alteration of solution pH or salt concentration. These 'nuclei' can grow into larger aggregates via monomer addition or aggregate association to form ordered or disordered (amorphous) aggregates and eventually grow into amyloid fibrils or insoluble macroscopic particles (Figure 1.2) (Amin et al., 2014). These aggregates are often termed irreversible as they are extremely thermodynamically stable and can often have a lower energy minima than the native state, meaning they are not easily reversible without the use of high temperatures and pressures or highly concentrated chemical denaturants (Roberts, 2014a). A protein's conformational stability can influence its aggregation propensity, as low conformational stability of a protein can result in partial unfolding which can expose APRs and result in aggregate formation. Furthermore, proteins can unfold as a result of interactions at air-water and water-surface interfaces, a process that can be particularly problematic for biopharmaceuticals (Amin et al., 2014). As previously mentioned, there are a wide variety of different complex mechanisms whereby a nucleus might grow into a larger aggregate such as a fibril or amorphous cluster, a detailed review of which is beyond the scope of this thesis. Furthermore, different models have been developed to explain the aggregation process but no universal model has yet been identified, partially due to the dependence on the particular protein and its environment as well as the occurrence of multiple different aggregation mechanisms in the same protein (Wang and Roberts, 2018).

While it is true that aggregation can occur as a result of unfolding, some proteins are capable of forming aggregates from native-like conformations without undergoing unfolding. Amyloid fibrils can be formed by self-association of monomeric peptides that oligomerise to form a nucleus that rapidly elongates (Zapadka et al., 2017). The homotetrameric transport protein transthyretin (TTR) can dissociate into monomers which can initiate aggregation (Garcia-Pardo et al., 2014). Furthermore, proteins containing transthyretin-like domains have been experimentally shown to form amyloid aggregates without extensive unfolding (Garcia-Pardo et al., 2014). Intrinsically disordered proteins lack a single stable three-dimensional structure and exist in a wide variety of conformations, typically compacted forms due to the formation of hydrogen bonds and salt bridges (Kumari et al., 2018). Abnormal regulation of these IDPs within the cell can result in aggregation, which can have implications in health and disease. An example of this is  $\alpha$ -synuclein, a human IDP with a function in remodelling lipid vesicles where it binds to lipid membranes and adopts an  $\alpha$ -helical structure (Doherty et al., 2020). However, due to its aggregation behaviour α-synuclein has been implicated in Parkinsons disease where it is thought to form fibrilar aggregates that disrupt cellular homeostasis and cause massive neuronal death (Stefanis, 2012).

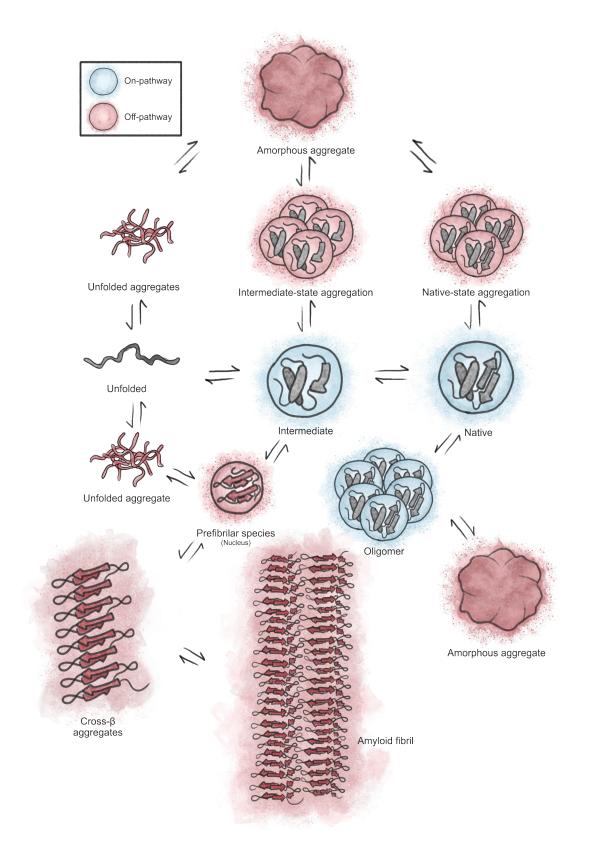


Fig. 1.2 **Mechanisms of aggregation.** Protein aggregation can occur by a variety of different mechanisms, making it difficult to predict. As a protein folds from an unfolded polypeptide chain via intermediate states to the final native fold it may aggregate at any step, either forming unstructured amorphous aggregates or structured amyloid fibrils. Redrawn and adapted from Ebo et al. (2020a).

# 1.1.3 Chemically induced protein aggregation

Chemical reactions between amino acid side chains can directly crosslink proteins and alter their behaviour. Deviation from the native state can increase the rate of these reactions, and so influence aggregation rate. Covalently bonded aggregates can typically arise via chemical reactions between monomers (Cromwell et al., 2006). Common covalent links include disulphide bridges that can be initiated between oxidising reduced cysteine residues, resulting in exposure of hydrophobic residues and subsequent oligomerisation and aggregation of proteins (Wang, 2005; Trivedi et al., 2009). The affect of disulfide formation on protein aggregation is generally protein-specific, however non-native disulfide bonds have been shown to result in precipitation of insoluble aggregates (Wang, 2005; Ciaccio and Laurence, 2009). Additionally, oxidation of residues can alter the primary sequence and result in the formation of protein oligomers via covalent bonds. For example, oxidised tyrosine residues in  $\alpha$ -synuclein amyloid fibrils have been shown to covalently form dityrosine which contributes to the high stability of the complex (Al-Hilaly et al., 2016).

A major cause of chemical degradation during storage and manufacture of monoclonal antibodies (mAbs) is deamidation of neutrally charged asparagine (Asn) residues to negatively charged aspartate (Asp) residues (Yan et al., 2018). The introduction of an unfavourable negative charge can influence a protein's structure as well as its biophysical properties, resulting in increased charge-mediated interactions which can lead to aggregation. For this reason, low levels of deamidation impurities (less than 5% of the total sample) have been shown to result in increased mAb aggregation (Nilsson et al., 2002).

# 1.1.4 Factors affecting protein aggregation

Broadly speaking, factors influencing protein aggregation can be categorised into structural (internal factors) or environmental (external factors). Structural factors include the primary sequence, which is widely accepted to modulate a protein's aggregation propensity along with the external environmental factors. Even single amino acid substitutions have been shown to drastically alter a protein's ability to aggregate (Ventura, 2005). Particularly, alteration of the primary sequence to increase the amount of non-polar (hydrophobic) amino acids has been shown to increase aggregation propensity (Kim and Hecht, 2006). Individual residues have been shown to be important when determining a protein's aggregation propensity as stabilisation of non-native interactions can be regulated by key 'gatekeeper' residues (proline, arginine, lysine, aspartic acid and glutamic acid) (Rousseau et al.,

2006). These are electrostatically charged residues (and proline) that specifically aid in proper folding and oppose aggregation by blocking misfolding reactions via electrostatic repulsion (Rousseau et al., 2006). Mutations of these residues can dramatically affect protein aggregation, with a single mutation of a non-polar residue for a lysine in an amyloid-forming *de novo*  $\beta$ -sheet protein being sufficient to instead result in the formation of monomeric  $\beta$ -sheet proteins (Wang et al., 2002; Frokjaer and Otzen, 2005).

A protein's secondary structure can influence its aggregation propensity, as proteins rich in  $\beta$ -sheets are more prone to aggregation than those rich in  $\alpha$ -helices (Shifman, 2008). Moreover, transition of  $\alpha$ -helical structures to  $\beta$ -sheet rich aggregates has been described as a mechanism of amyloid fibril formation (Mudedla et al., 2018). The high propensity of  $\beta$ -sheet rich proteins to aggregate makes rational design of soluble and monomeric  $\beta$ -sheet proteins challenging (Shifman, 2008). However, advances in computational modelling have enabled *de novo* design of soluble proteins of increasing complexity (Langan et al., 2019; Ng et al., 2019; Silva et al., 2019), including the first examples of *de novo* design of two functional soluble  $\beta$ -barrel proteins (Dou et al., 2018; Marcos et al., 2018). Design of membrane proteins is more challenging again, although there have been various examples of *de novo* designed  $\alpha$ -helical membrane proteins (Lu et al., 2018; Joh et al., 2017). The first example of *de novo* designed transmembrane  $\beta$ -barrel proteins that fold spontaneously and reversibly into synthetic lipid membranes was in 2021, paving the way for design of custom protein nanopores that have wide potential in biotechnology (Vorobieva et al., 2021).

Various environmental factors can initiate protein aggregation. High temperatures induce protein unfolding, exposing hydrophobic residues as well as increasing the frequency of molecular collisions resulting in aggregation (Wang, 2005). pH and ionic strength of a solution can influence aggregation behaviours by altering the strength of electrostatic interactions between proteins (Zapadka et al., 2017). The type and distribution of surface charges on a protein is governed by the solution pH, influencing intra- and inter-molecular protein interactions and so the aggregation propensity (Wang et al., 2010). Furthermore, high protein concentrations can result in increased aggregation as a result of macromolecular crowding as excluded volume effects and high concentrations of macromolecules restrict the volume of accessible solvent (White et al., 2010). This leads to limited entropic freedom resulting in compact non-native forms of proteins being favoured, which can lead to aggregation (Hong and Gierasch, 2010). Hydrodynamic forces including extensional and shear flow have also been linked to biopharmaceutical aggregation, particularly as these proteins undergo such stresses during manufacture (Willis and Chin, 2018; Willis et al., 2020). Additional environmental factors include: shaking, increased pressure, addition of organic solvent, freeze-thawing, freeze drying, spray drying, spray freeze drying, or reconstitution of lyophilised powder (Wang, 2005).

# **1.1.5** Structure and morphology of protein aggregates

Protein aggregate morphology can generally be categorised as amorphous or fibrillar. Amorphous aggregates have no regular interactions, whereas fibrils are structurally ordered aggregates comprised mainly of  $\beta$ -sheets. Fibrillar aggregates are commonly associated with amyloid diseases caused by the formation of insoluble proteinaceous deposits resistant to degradation (Seuma et al., 2021). Furthermore, aggregate size can vary from soluble submicron to insoluble macroscopic particles depending on the type of aggregate (Wang, 2005). The morphology of aggregates is pathway dependent and can be influenced by a variety of external factors (Wang et al., 2010). In some instances, the same protein can form both amorphous or fibrillar aggregates depending on the environmental conditions, demonstrating the primary structure is not the single determining factor of aggregate morphology (Chaturvedi et al., 2016).

# **1.2** Antibody fragments and derivatives

# 1.2.1 Antibodies

Antibodies, or immunoglobulins, are 'Y-shaped' molecules functionally separated into two regions: the variable (V) region which is involved in antigen binding, and the constant (C) region which interacts with effector cells. There are five different classes of immunoglobulins that differ in their C regions - IgA, IgD, IgE, IgM and IgG (Figure 1.3) (Wang et al., 2007). The most abundant immunoglobulin in human serum, as well as the most widely used for therapeutic purposes, is IgG which can be split into four subclasses (IgG<sub>1</sub>, IgG<sub>2</sub>, IgG<sub>3</sub>, and IgG<sub>4</sub>) (Zhang et al., 2009). These subclasses differ in hinge region length as well as number and location of interchain disulphide bonds (Wang et al., 2007).

IgGs are comprised of two heavy (H) and two light (L) chains joined by disulphide bonds (Figure 1.4) (Safarnejad et al., 2011). Each of these chains consists of immunoglobulin (Ig) domains with a structure consisting of 7-9 antiparallel  $\beta$ -strands organised into a  $\beta$ -sandwich with an intersheet disulphide bond between  $\beta$ -strands B and F; each H chain is comprised of four Ig domains (one variable, V<sub>H</sub>; three constant, C<sub>H</sub>1, C<sub>H</sub>2, and C<sub>H</sub>3) and

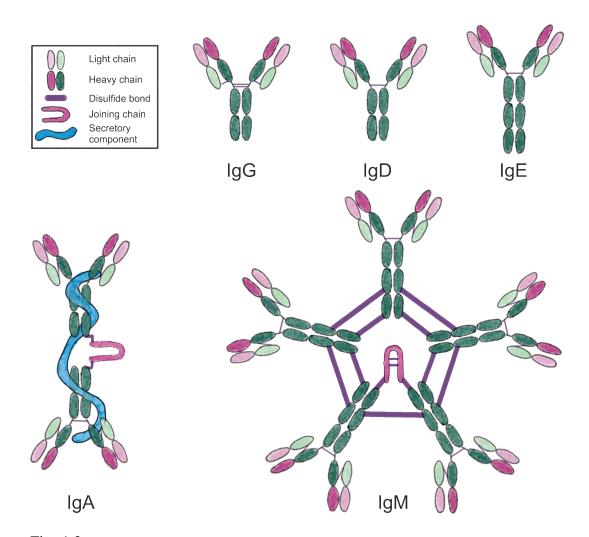


Fig. 1.3 **Classes of human immunoglobulins.** The most common in human serum, IgG, is the most predominant antibody in the secondary immune response and the most widely used for therapeutic purposes. IgE has two additional constant domains and is involved in allergic response. IgA is involved in immune function of mucosal areas, such as the gut or respiratory tract. It exists as a dimer and has additional secretory component. IgM is the largest antibody (pentameric) and is involved in the primary immune response. The function of IgD is not fully known, however it is thought to have evolved soon after IgM after the emergence of the adaptive immune system and it may have a function in regulating mucosal membranes (Ohta and Flajnik, 2006; Schroeder and Cavacini, 2010; Gutzeit et al., 2018)

each L chain is comprised of two Ig domains (one variable,  $V_L$ ; one constant,  $C_L$ ) (Wang et al., 2007; Bodelón et al., 2013). These chains combine to form the full IgG molecule made of one crystallisable fragment (Fc) and two antigen binding domains (Fab) joined at a hinge region (Figure 1.4). Within the Fab, each V domain contains three complementarity determining regions (CDRs) - hypervariable regions which form the antigen binding site (Schroeder and Cavacini, 2010). IgGs are N-glycosylated at Asn297 in CH2 domains of the Fc domain (Jennewein and Alter, 2017). Glycosylation of the Fc has been shown to influence stability, with deglycosylated antibodies exhibiting lower thermal stability and higher aggregation rates compared with their glycosylated counterparts (Zheng et al., 2011). Furthermore, the glycosylated Fc region plays a role in downstream immune responses through binding to Fc receptors (Kennedy et al., 2017).

### **1.2.2** Antibody fragments

Immunoglobulins are modular in nature, a characteristic that has the potential to be exploited allowing engineering of therapeutics optimised for specific targets. Antibody fragments include antigen-binding fragments (Fab), single chain variable fragments (scFv), miniaturised antibodies such as nanobodies as well as bispecific antibodies. These are taken from the antigen-binding part of antibodies and produced using recombinant processes (Figure 1.4). These smaller fragments provide higher tissue penetration in comparison to full sized mAbs and are less costly to produce as they can be expressed in prokaryotes due to the lack of glycosylation sites associated with the Fc domain (Roopenian and Akilesh, 2007; Nelson, 2010). Fab fragments were the first class of antibody fragments to be developed, as well as the most successful as they represent around half of all antibody fragments that have entered clinical trials (Nelson, 2010; Bates and Power, 2019). However, to date only four Fab fragments have been approved by the FDA (https://www.antibodysociety.org/resources/approved-antibodies/ accessed 10th February 2023). The lack of Fc domain means the fragments have short circulating half-lives as there is no interaction with the neonatal Fc receptor allowing FcR-mediated recycling, a process that extends half-life by recycling IgGs and reducing lysosomal degradation (Roopenian and Akilesh, 2007). This can therefore lead to needing larger and more frequent doses (Bates and Power, 2019). Furthermore, as the presence of the Fc domain increases the thermodynamic stability of antibodies, fragments lacking this domain have reduced thermodynamic stability which can cause increased aggregation risk and hence immunogenicity (Nelson, 2010). Attempts to increase half-life include chemical conjugation to proteins such as albumin and PEGylation, which in itself can raise more issues due to technical challenges and expense (Nelson, 2010).

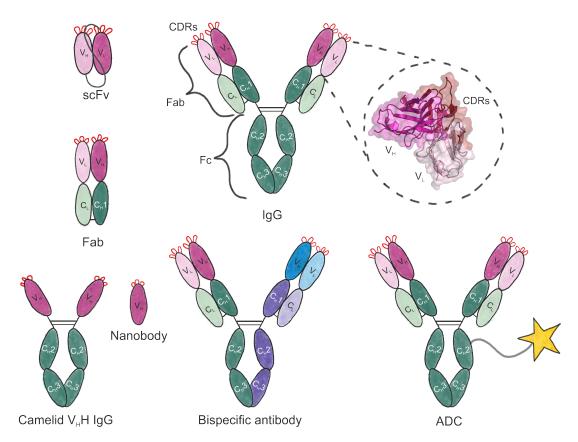


Fig. 1.4 Structures of mAbs and antibody fragments. IgG, immunoglobulin; Fab, antigen binding fragment; scFv, single chain variable fragment; Fc, crystallisable fragment; V<sub>L</sub>, variable light chain; V<sub>H</sub>, variable heavy chain; C<sub>H</sub>, constant heavy chain; C<sub>L</sub>, constant light chain; CDR, complementarity determining region. scFv fragment showing CDRs is from PDB 5JZ7.

scFvs are fusion proteins consisting of the V<sub>H</sub> and V<sub>L</sub> chains joined via a short linker (Figure 1.4). Multiple variants of these fragments can be produced in order to optimise both stability and affinity for the target (Nelson, 2010). However, their reduced half-life due to the lack of Fc domain limits their application potential as therapeutics and the only clinically approved molecule to date is Brolucizumab, a humanised scFv targeting vascular endothelial growth factor-A to treat neovascular age-related macular degeneration (Kaplon et al., 2020). Bispecific antibodies can be created with two different antigen-binding domains exhibiting binding specificity to two different targets (Nelson, 2010). Many bispecifics are currently undergoing clinical trials or are awaiting approval, such as Faricimab, targetting anti-vascular endothelial growth factor-A (VEGF-A) and anti-angiopoietin-2 (Ang-2) for the treatment of ophthalmic disorders (Kaplon et al., 2022). Heavy-chain only antibodies were first isolated from camelids in the early 1990s (Figure 1.4) (Hamers-Casterman et al., 1993). This enabled development of single domain antibodies, or "nanobodies", that comprise of a single V<sub>H</sub> domain. These molecules exhibit high levels of tissue penetration, high stability and solubility as well as low immunogenicity, making them potentially powerful therapeutic agents (Hu et al., 2017). Caplacizumab is a nanobody targeting anti-von Willebrand factor for the treatment of acquired thrombotic thrombocytopenic purpura (aTTP), a rare blood clotting disorder characterised by low platelet number that can lead to anaemia and organ failure of varying severity (Peyvandi et al., 2016). Caplacizumab was FDA approved in early 2019 and, in combination with plasma exchange and immunosuppressive therapy, is currently being used to treat adult patients with aTTP (Kaplon et al., 2020).

The advancement of genetic engineering techniques has enabled generation of combinations of these modular antibody fragments to create novel complexes (Khatib and Salla, 2022). These include fragments with a wide range of specificities and valencies, expressed either as a single chain, assembled in multimeric forms or stringed in tandem (Khatib and Salla, 2022).

Finally, mAbs can be used to 'deliver' highly potent cytotoxic small molecule drugs to the target using so called antibody-drug conjugates (ADCs). ADCs can reduce off-target effects by exploiting the antibodies' specificity in order to direct the drug conjugate to a specific site. Drugs are joined to an antibody via a conjugating linker, typically via lysine residues on the antibodies surface (Chudasama et al., 2016). Cysteine residues can also be used for conjugation by reduction of existing disulphide bridges or introduction of free cysteine residues via protein engineering. However, this poses the risk of increased aggregation during the conjugation step by formation of disulphide bridges (Chudasama et al., 2016).

# **1.3** Biopharmaceuticals and the aggregation problem

# **1.3.1** History of biopharmaceuticals

Biopharmaceuticals are defined as therapeutics produced from biological sources (Rader, 2008). Generally, biopharmaceuticals are more effective at lower concentrations and result in fewer side effects compared with their small molecule counterparts (Wang et al., 2007). From the approval of Humulin (human insulin) as the first recombinant protein therapeutic in 1982, the biopharmaceutical market has grown dramatically with total sales reaching \$188 billion in 2017 (Johnson, 1983; Walsh, 2018). Monoclonal antibodies (mAbs) represent an important sector of biopharmaceuticals, with sales exceeding \$123 billion in 2017 they represent almost two thirds of the biopharmaceutical market (Walsh, 2018).

## **1.3.2** Monoclonal antibodies and their generation

#### 1.3.2.1 Hybridoma technology

Generally mAbs with a high affinity for the epitope are generated by two approaches, both of which were recognised by award of a Nobel Prize. In 1975 Köhler and Milstein developed mouse hybridoma technology to produce mAbs where immortalised myeloma cells were fused with spleen cells (B-cells) from a mouse immunised against a particular antigen (Kohler and Milstein, 1975). This generates a stable line of immortalised cells (due to the myeloma) producing an antibody of interest (due to the spleen cells). Prior to electrofusion, myeloma cells lacking the hypoxanthine-guanine phosphoribosyltransferase (HGPRT) enzyme are selected for. After electrofusion cells are grown on HAT medium (hypoxanthine-aminopterin-thymidine medium). This selection works by combining aminopterin, a drug inhibiting de novo DNA synthesis, with thymidine and hypoxanthine, which provide the cells with the tools to use the nucleotide salvage pathway (Ribatti, 2014). This pathway is only available to cells that have the right enzymes, including HGPRT (Ribatti, 2014). B-cells are unable to survive long on their own, and since myeloma cells lack the HGPRT gene they can only survive if they fuse into hybridomas. This allows selection for hybridoma cells that secrete antibodies against the particular antigen of interest representing a stable source of mAbs (Figure 1.5) (Rodgers and Chou, 2016).

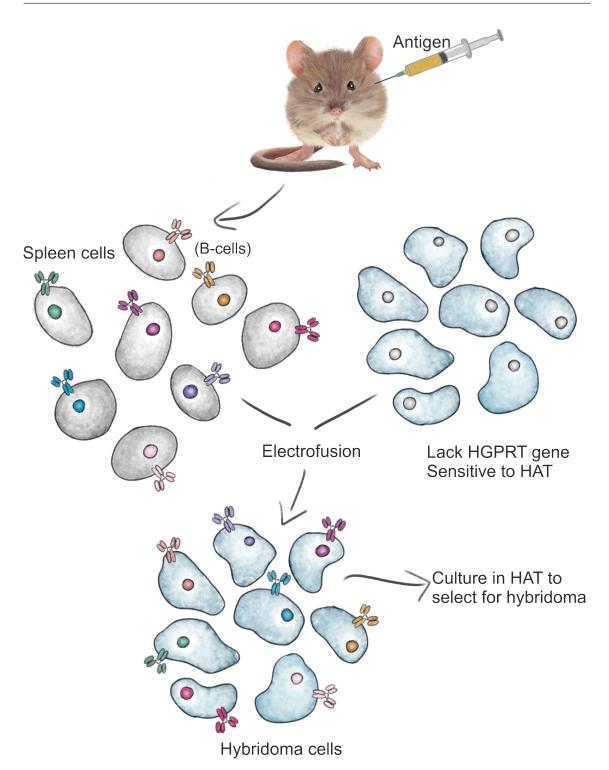


Fig. 1.5 Köhler and Milstein's mouse hybridoma technique for producing mAbs. Mice are immunised with antigen of interest and spleen lymphocytes (B-cells) are isolated. These are fused with mouse myeloma cells that lack the hypoxanthine-guanine phosphoribosyltransferase (HGPRT) enzyme required to grow on HAT medium. The resulting hybridoma cells are selected for on HAT medium, as the myelomas fusing with the spleen lymphocytes (B-cells) allows growth on HAT. B-cells do not survive long on their own and only survive if they fuse into a hybridoma. These hybridomas have the antibody producing ability of a B-cell and the immortality of a myeloma, representing a stable source of mAbs.

This work led to Köhler and Milstein being awarded the Nobel Prize for Physiology or Medicine in 1984 (Alkan, 2004). Building on this discovery, the first monoclonal antibody therapy (Orthoclone OKT3) was approved in 1986 to prevent kidney organ transplant rejection (Ecker et al., 2015). Since then the FDA have approved over 100 monoclonal antibodies globally and there are currently over 570 in various stages of clinical trials (Cai, 2018; Kaplon et al., 2022). A drawback of mAbs produced via the hybridoma method is that due to the murine lineage the use of such therapeutics can result in an immune response in the patient, therefore various gene technologies to 'humanise' mAbs have been developed (De Groot and Scott, 2007; Rodgers and Chou, 2016). Partially humanised mAbs formed of mouse variable domains and human constant regions or fully humanised mAbs formed of mouse CDRs and a human mAb scaffold were developed to reduce immunogenicity (Figure 1.6) (Morrison et al., 1984; Jones et al., 1986). Furthermore, fully human mAbs can be produced by means of genetic engineering to further reduce the risk of immunogenic response (Lonberg et al., 1994; Mahler et al., 1997). UK based biopharmaceutical company Kymab have created a fully human antibody system using mice by replacing the mouse variable genes with human variable genes within the mouse genome, resulting in production of high-affinity antibodies with human-like CDRs (Lee et al., 2014). Mice are fertile and able to elicit typical immune responses to generate high-affinity therapeutic human mAbs without the need for extensive optimisation. While these methods are effective at minimising immunogenicity in patients as a result of mAbs, even fully human mAbs have been shown to induce an immune response in patients (Kay et al., 2008; Lee et al., 2014). This risk is increased by the presence of contaminants and aggregates in the final formulated mAb. A recent study suggested the immunogenic response associated with aggregation in biotherapeutics is due to specific epitopes in mAbs that can be exposed as a result of aggregation and recognised by the immune system (Eyes et al., 2019). This study used molecular dynamics simulations to show biophysical stress as a result of aggregation can exacerbate epitope exposure. Other studies have suggested a mechanism whereby aggregation leads to immunogenicity is a result of ordered oligomerised antigens that may resemble structures of viruses or foreign microorganisms, hypothesising the immune system has evolved to recognise and respond to these repetitive epitopes (Hermeling et al., 2004; Kessler et al., 2006; Ratanji et al., 2014; Kuriakose et al., 2016). Furthermore, these epitopes have been shown to specifically activate B cell responses by cross-linking of antigen receptors (Kuriakose et al., 2016). Interestingly, the route of mAb administration has been shown to contribute to aggregation, with subcutaneous injection being the most immunogenic and intravenous being the least, as subcutaneous injection results in prolonged and localised exposure of the mAb in close proximity to the lymph nodes which are major sites of immune cells (Ratanji et al., 2014; Kuriakose et al., 2016).

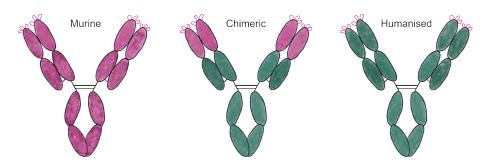


Fig. 1.6 Humanisation of mAbs. To reduce the immunogenicity of murine mAbs (pink), various technologies have been developed to 'humanise' them. Partially humanised 'chimeric' mAbs consist of murine variable domains and human constant regions. Fully humanised mAbs consist of murine CDRs grafted onto human frameworks. Techniques exist to produce fully human mAbs, but these can still elicit an immmunogenic effect.

#### 1.3.2.2 Phage display

In 2018 the Nobel Prize in Chemistry was awarded to Frances Arnold, George Smith and Greg Winter for their pioneering work on directed evolution methods to design biological molecules, namely enzymes and antibodies (Gibney et al., 2018). Arnold developed directed protein evolution methods using random mutagenesis to rapidly evolve enzymes with a desired characteristic (Chen and Arnold, 1993). In 1985, Smith reported that genes inserted in the middle of the bacteriophage filamentous phage gene III (pIII) are displayed on the surface of the bacteriophage (Smith, 1985). Winter and Smith developed this technology, named phage display, as a way to develop antibodies with high affinity and selectivity for a specific target by fusing  $V_H$  and  $V_L$  genes, generating an single chain Fv (scFv), with pIII (Clackson et al., 1991; Winter et al., 1994). A library of phage expressing various scFvs on their surface are screened against an immobilised antigen of interest, unbound phage are washed away and bound phage are eluted. Various techniques exist to elute bound phage, including low pH buffers, high ionic strength, reductants such as DTT, or ultrasound (Vodnik et al., 2011). Isolated DNA from bound phage is extracted and subjected to random mutagenesis to create a mutated library, which in turn is used to produce more phage; repeated cycles are carried out to improve affinity for the target (Figure 1.7) (Frei and Lai, 2016). The combination of phage display with Arnold's directed evolution methods led to the approval of the first fully human mAb Humira (adalimumab) for the treatment of rheumatoid arthritis in 2002, which has gone on to be the top-selling biopharmaceutical in the world with 2021 sales exceeding \$20 billion (Mullard, 2022).

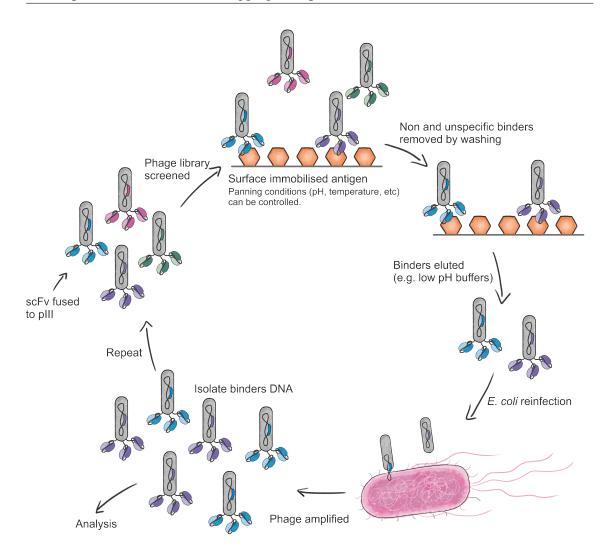


Fig. 1.7 **Phage display for producing mAbs.** scFv gene is inserted between pIII so fragment is displayed on the surface of filamentous M13 bacteriophage. Phage are screened for binding against an immobilised antigen. During this step conditions can be modified (pH, temperature, etc) to alter the selection pressure. Unbound phage are washed away, bound phage are eluted and used to re-infect *E. coli.* DNA for binders is isolated and used in a second round of selection to improve the affinity. Alternatively, mutated libraries of binders can be created to isolate unique binders.

### 1.3.2.3 Yeast surface display

Yeast surface display was first published in 1997 as an alternative method of producing high affinity antibodies (Boder and Wittrup, 1997). A scFv library is incorporated into Saccharomyces cerevisiae as a fusion protein with cell wall protein Aga2p mating adhesion receptor which results in the scFv being presented on the surface of yeast (Boder and Wittrup, 1997). Conjugation of the antigen of interest to magnetic beads or a fluorophore allows sorting by magnetism-assisted cell sorting (MACS) or fluorescence-assisted cell sorting (FACS), respectively (Cherf and Cochran, 2015). DNA of binders is isolated and used to create a mutated library, which is used in subsequent cycles to optimise affinity for the target (Figure 1.8) (Cherf and Cochran, 2015). Yeast surface display also has the potential for displaying Fab libraries to isolate high-affinity binders (Rosowski et al., 2018). Yeast surface display has been modified to include a selection for stability as well as affinity by addition of a conformational ligand (Protein A) that is specific for folded V<sub>H</sub> domains, meaning it recognises folded V<sub>H</sub> domains so can be used to select for variants that are both folded and have affinity for the target (Julian et al., 2015). Using selection for protein A binding alone without the additional selection for antigen binding allowed isolation of variants that have enhanced stability but reduced affinity, showing the inevitable trade-off between properties and how it is often the case that improving one property negatively impacts the other as well as highlighting the importance of co-selection (Julian et al., 2015). Yeast surface display has been widely utilised in various directed evolution studies for improving both stability and affinity in antibody-based proteins (Julian et al., 2015, 2017, 2019; Tiller et al., 2017b,a).

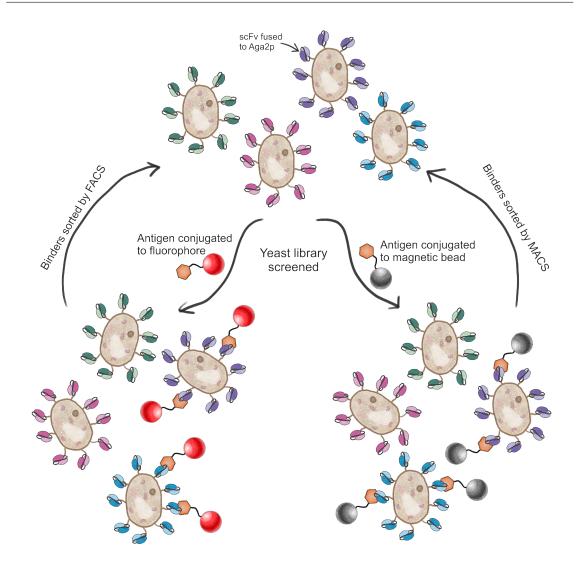


Fig. 1.8 Yeast surface display technique for producing mAbs. scFvs are presented on the surface of *Saccharomyces cerevisiae* by fusion with the Aga2p protein. Cells are screened for binding by incubating with antigens conjugated to magnetic beads or a fluorophore, allowing FACS or MACS, respectively. Binders DNA is isolated and repeated rounds of mutation and cell sorting can be used to isolate variants with improved binding affinities.

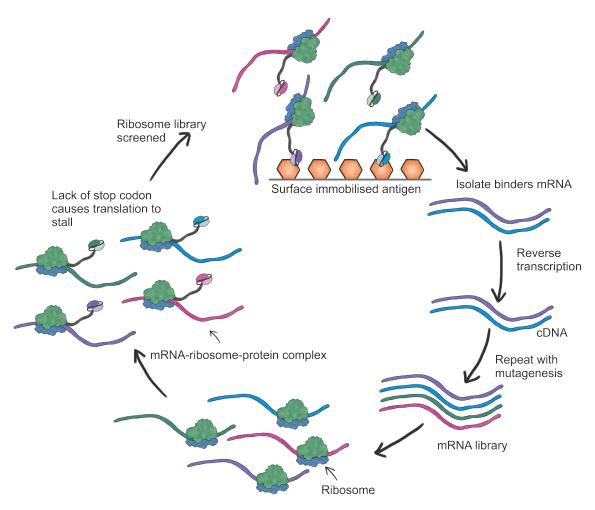


Fig. 1.9 **Ribosome display technique for producing mAbs.** scFv DNA library lacking a stop codon is PCR amplified and transcribed to mRNA *in vitro*. The lack of a stop codon causes the mRNA transcript to stall on the ribosome forming an mRNA-ribosome-protein complex. These are screened against an immobilised antigen and binders are isolated. mRNA of binders is isolated and translated back to cDNA for analysis. Repeated cycles using mutated libraries can improve the affinity for the target.

### 1.3.2.4 Ribosome display

Also published in 1997, ribosome display represents a cell-free method for antibody development, overcoming ethical implications associated with the use of animals in mouse hybridoma technology (Hanes et al., 1997). In this method, a scFv library is amplified by PCR and transcribed to mRNA *in vitro*. The transcript lacks a stop codon causing translation to stall and the protein and its encoding mRNA remain attached to the ribosome (Hanes et al., 1997). mRNA-ribosome-protein complexes are screened against immobilised antigens; unspecific complexes are washed away then bound complexes are dissociated. RNA of binders is isolated and transcribed back to cDNA for amplification using reverse transcription PCR (RT-PCR) (Figure 1.9). This is used to create a mutated library and the process repeated to increase antigen binding affinity (Hanes et al., 1997). Similar to phage and yeast-surface display, the screening conditions can be modified to coevolve beneficial biophysical properties (such as stability or solubility) alongside binding affinity. However, the instability of mRNA restricts the potential selection conditions (e.g. temperature, pH) that can be applied to screen libraries (Galán et al., 2016). Ribosome has the highest capacity of any of the display technologies, due to its cell-free nature it can screen up to  $10^{12-15}$  antibody variants in a single reaction (Kunamneni et al., 2020), and it has been widely exploited to evolve antibody fragments including scFvs (Zhao et al., 2009; Kunamneni et al., 2018) and single domain antibodies (Bencurova et al., 2015) with high-affinity for their targets.

#### 1.3.2.5 Mammalian display

A more recent approach for antibody display is mammalian display, in which the scFv (Ho et al., 2006) or full-length IgG (Akamatsu et al., 2007) is displayed on the surface of a mammalian cell via fusion to the transmembrane domain of human platelet-derived growth factor receptor (PDGFR) and screened for binding to an antigen of interest (Ho and Pastan, 2009). This method has been demonstrated to not only enable isolation of high-affinity binders, but also to isolate variants with improved solubility, reduced aggregation, and reduced immunogenicity (Dyson et al., 2020).

# **1.3.3** Aggregation in biopharmaceuticals

Throughout their lifetime, biopharmaceuticals are exposed to a multitude of physical, chemical and mechanical stresses that can result in formation of aggregates. These aggregates pose a significant risk to patients as they can invoke an immune response, causing side effects ranging from intolerance to adverse reactions and death (Jiskoot et al., 2012). Typically, bioreactors house mammalian cell cultures secreting mAbs into the growth medium (Cromwell et al., 2006). The protein is harvested and undergoes various purification steps from centrifugation to Protein A chromatography, ion-exchange chromatography and ultrafiltration/diafiltration during formulation before transportation to patient for administration (Figure 1.10). The formulated protein can be frozen in order to maintain stability for a period of time before filling and finishing into its vials or pre-filled syringes (Cromwell et al., 2006). The high concentrations of mAbs required for administration can pose an aggregation risk, as the maximum volume per dose is usually 1

mL and can require over 200 mg of product (Roberts, 2014b). All these processing steps expose mAbs to a host of aggregation-inducing stresses that can hinder or completely derail commercialisation of the product, and can therefore contribute to the over 96 % of drugs which fail development (Hingorani et al., 2019). Low pH conditions are used to elute mAbs during Protein A chromatography and can induce protein aggregation, whereas filtration and centrifugation exposes cells to harsh flow conditions which can cause protein unfolding and formation of aggregates (Cromwell et al., 2006). Physical stresses that may induce aggregation include changes in temperatures, such as high temperatures in the bioreactor, or freeze-thaw processes for storage and transport. Furthermore, slower rates of freezing can result in protein cryoconcentration leading to zones of higher protein concentration that can harm proteins during thawing (Rayfield et al., 2017).

After the final product is filled into syringes or vials there are no further purification steps (Cromwell et al., 2006). This process places mAbs under high hydrodynamic stress that can induce the formation of aggregates; these aggregates can be injected into the patient and may cause adverse reactions (Willis et al., 2020). Furthermore, the presence of aggregates will increase the viscosity of the solution, making administration of the drug slow and painful (Tomar et al., 2016). Studies have investigated the effectiveness of using an inline filter during administration of intravenous drugs at reducing the amount of particles injected into the patient (Pollo et al., 2019). While these filters were shown to be effective at significantly reducing the concentration of particles >2 microns, particles between 1 and 2 microns were still detected indicating these filters are not capable of removing all particles. Limits of acceptable levels of soluble aggregates are determined on a case-by-case basis as there are no general predefined acceptable levels in biopharmaceuticals (Mahler et al., 2009). However, levels are commonly kept below 5% of the protein mass in order to reduce any adverse reactions in patients (van Reis and Zydney, 2007). It is important to detect aggregation-prone mAbs as early as possible in the manufacturing process in order to reduce unnecessary time and expense. Effort to decrease aggregation levels by investigation into a mAbs' aggregation behaviour is vital to minimise cost and time to market as well as immunogenicity of therapeutics (Roberts, 2014a).

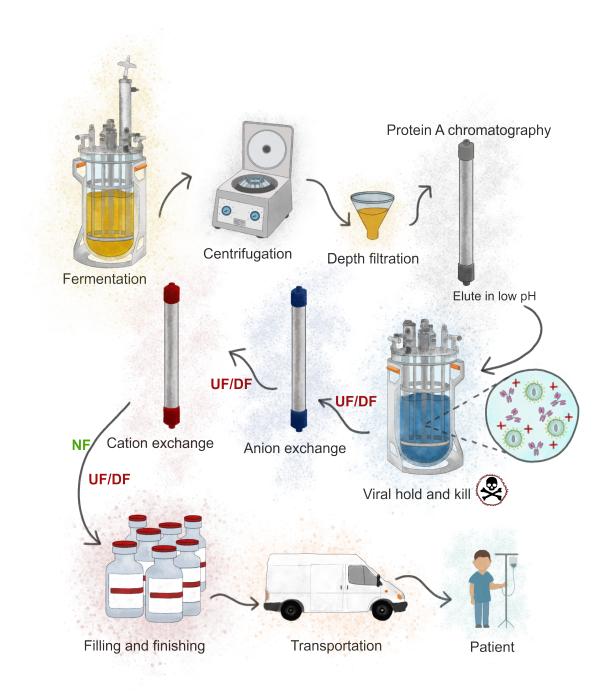


Fig. 1.10 **Overview of mAb purification process.** A biopharmaceutical undergoes upstream and downstream processing before administration into patients, many of which can impart stresses resulting in aggregation. Downstream processing starts with inoculation of cell culture for large scale fermentation through to centrifugation and depth filtration. Upstream processing involves chromatography steps, Protein A chromatography, anion- and cation- exchange chromatography, to remove residual impurities as well as nanofiltration (NF) and Ultra-/Dia-filtration (UF/DF) which concentrates and buffer-exchanges into the formulation buffer.

# **1.4** Methods to assess biopharmaceuticals

## **1.4.1** Analytical techniques

### 1.4.1.1 Assessing protein aggregation

Various analytical techniques can be employed to detect biopharmaceutical aggregation. The most commonly used is size-exclusion chromatography (SEC) which separates macromolecules based on their hydrodynamic radius and produces an elution profile that can be used to estimate molecular weight. The sample passes through a column packed with porous beads; molecules with a large hydrodynamic radius (such as aggregates) do not enter the pores and elute before molecules with a smaller hydrodynamic radius (such as the natively folded protein) which are retarded by the pores and have a longer retention time (Figure 1.11A) (Mahler et al., 2009). SEC analysis is often performed using high performance liquid chromatography (HPLC) to provide a relatively easy and high throughput method to rapidly analyse therapeutic samples. However, using this method may be misleading as the conditions therapeutic proteins are analysed in are not necessarily representative of the final formulation. The occurrence of protein adsorption to the media can affect elution profiles, also indicating a poor IgG, although this can be minimised by increasing salt concentration in the mobile phase or by the addition of arginine (Carpenter et al., 2010). SEC is ineffective at analysing large, insoluble aggregates as they would clog the column and are often removed by filtration of the sample prior to SEC (Mahler et al., 2009). Various chromatography-based assays exist to probe antibody self-association, such as self-interaction chromatography (SIC) (Patro and Przybycien, 2000) and crossinteraction chromatography (CIC) (Jacobs et al., 2010), as well as antibody colloidal stability, such as standup monolayer adsorption chromatography (SMAC) (Kohli et al., 2015) and hydrophobicity, such as hydrophobic interaction chromatography (HIC) (Haverick et al., 2014). SIC involves measuring an antibodies retention time as it passes through a column conjugated with the same antibody, therefore a longer retention time correlates to higher levels of self-interaction (Patro and Przybycien, 2000). Similarly, CIC measures the retention time of an antibody as it passes through a column where the resin is coupled to polyclonal IgGs from human serum, and studies have inversely correlated retention times to the solubility of the candidate antibody (Jacobs et al., 2010). SMAC is a highthroughput HPLC method whereby the retention time of an antibody is measured as it passes through a column with a hydrophobic self-assembled monolayer in a 'standing-up phase' (or a 'standup monolayer') covering terminal hydrophobic groups (Kohli et al., 2015). The setup is hypothesised to mimic a protein exterior, and therefore proteins that

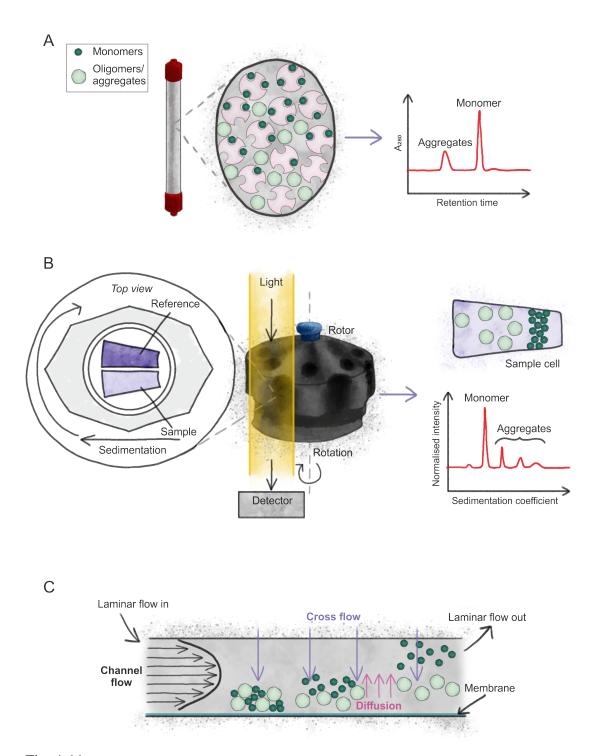


Fig. 1.11 Overview of methods to detect protein aggregation. A) Size-exclusion chromatography separates proteins based on hydrodynamic radius using porous beads. Proteins with a smaller hydrodynamic radius (e.g. monomers) are trapped by the pores in the beads and therefore have a longer retention time (take longer to pass through the column) than those with a larger hydrodynamic radius (aggregates). B) Analytical Ultracentrifugation (AUC) separates proteins of different sizes by using a centrifugal force to redistribute them across the sample cell. C) Asymmetrical Flow Field-Flow Fractionation (AF4) uses flow to separate molecules of different sizes. Proteins pass through a thin flow channel with a perpendicular cross flow pushes them towards an inpermeable membrane. Larger molecules take longer to diffuse back into the laminar flow and so are separated from smaller molecules. AF4 diagram adapted from Cho and Hackley (2010).

have low colloidal stability would be more prone to interact non-specifically with the column leading to longer retention times. Similarly, HIC involves separating mAbs based on their hydrophobicity using a column with a hydrophobic resin so that candidates with increased surface hydrophobicity have an increased retention time (Haverick et al., 2014). HIC has been employed to assess various post-translational modifications in mAbs such as tryptophan oxidation and aspartic acid isomerization as well as protein hydrophobicity which has been linked to increased aggregation due to association of hydrophobic patches (Haverick et al., 2014). While chromatography-based assays are useful and widely used for predicting protein solubility and colloidal stability, with SEC generally classed as the gold standard in industry, they are still limited for assessing the many hundreds of candidates that are generally identified during antibody discovery and affinity as they are particularly low-throughput (Elgundi et al., 2017).

Analytical Ultracentrifugation (AUC) can be employed to characterise protein aggregates and overcome some of the limitations of SEC. The differing masses, sedimentation equilibrium and sedimentation velocity of monomers, oligomers and aggregates causes them to redistribute when subjected to a centrifugal field (Figure 1.11B). This allows molecular weight calculation using the sedimentation and diffusion coefficients as well as quantification of aggregate level using optical detection (absorbance, interference or fluorescence) (Cole et al., 2008). Results obtained by AUC can be representative of the final formulated mAb as it can be carried out in the original buffer (Berkowitz, 2006). However, AUC is a low-throughput/high cost method and sample concentrations are limited to less than 50 mg/mL and at high concentrations require dilution prior to analysis therefore it may not be fully representative of the final formulated mAb (Shah, 2018).

Asymmetrical Flow Field-Flow Fractionation (AF4) allows separation of particles with diameters ranging from 1 nm to ~1000 nm (Fraunhofer and Winter, 2004). Samples are separated using a thin flow channel with a perpendicular cross-flow that pushes proteins towards an impermeable membrane; smaller molecules (monomers) diffuse back to the channels' laminar flow more rapidly than larger molecules (oligomers, aggregates) (Figure 1.11C) (Den Engelsman et al., 2011). An important method to reinforce data from SEC, AF4 lacks the capability to accurately quantify aggregation in comparison to SEC (Berkowitz et al., 2013).

Dynamic Light Scattering (DLS) detects scattered light that arises from diffusion of proteins and aggregates to determine the size distribution profile. The presence of aggregates will result in detection of two or more populations, monomer and aggregates (Figure 1.12A) (Li, 2011). DLS is a high-throughput, non-destructive method that is highly sensitive to large particles; this sensitivity means DLS can even detect tiny contaminants

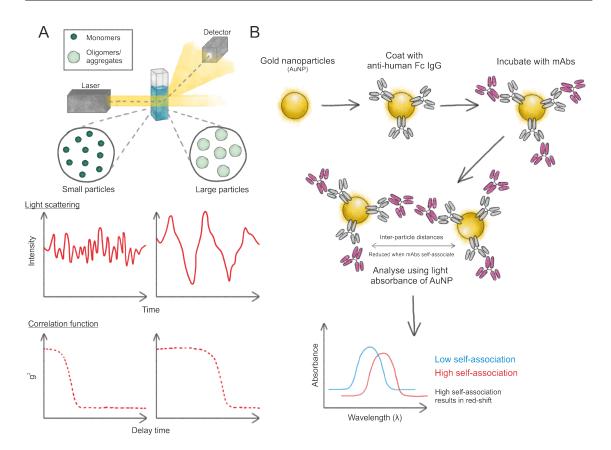


Fig. 1.12 **Overview of methods to detect protein aggregation.** A) Dynamic Light Scattering (DLS) allows analysis of protein sizes (hydrodynamic radii) by assessing how much light the sample scatters. Smaller molecules move faster than larger molecules and show faster fluctuations of scattered light over time, whereas larger molecules result in larger differences between the minima and maxima intensities. Using the intensity trace a correlation function can be determined by analysing the length of time a single particle is in the same place. As smaller molecules move faster, this correlation function rapidly shows an exponential decay. As larger molecules move slower, the exponential decay of the correlation function is delayed. B) Affinity-capture selfinteraction nanoparticle spectroscopy (AC-SINS) uses gold nanoparticles (AuNPs) coated with polyclonal antibodies with a high affinity for human Fc regions. These bind to mAbs and the coated AuNPs are used to monitor mAb self-association. As the mAbs self-associate, the interparticle distances between the AuNPs is decreased. This affects their absorbance properties, resulting in a red shift of the maximum wavelength of absorbance. AC-SINS diagram redrawn from (Wu et al., 2015).

such as dust, therefore the sample must be filtered before analysis which may remove aggregates (Den Engelsman et al., 2011). Additionally, highly concentrated samples (100 g/L), such as some mAbs, cannot be accurately analysed using this method (Shah, 2018).

Various spectroscopic and spectrometric technologies exist to detect and characterise aggregation of biopharmaceuticals. Spectroscopic methods represent easy to perform techniques that are not destructive to the sample (Den Engelsman et al., 2011). Circular dichroism (CD) spectroscopy uses the difference in absorption of left- and right- handed circularly polarised light in chiral molecules to determine secondary and tertiary structure by UV analysis. CD requires low amounts of protein but is subject to interference from excipients and solvents and the requirement for sample dilutions mean analysis is not representative of the final formulated mAb (Den Engelsman et al., 2011).

Mass spectrometry (MS) is a powerful technique employed to determine the molecular weight of an analyte. More accurately, MS analyses the mass to charge ratio (m/z) of ions in a sample. For mAb analysis, MS can be used to determine amino acid sequence, post-translational modifications (such as glycosylation) and chemical modifications, as well as higher order structure and location of disulphide bridges (Zhang et al., 2009). Typically the sample is ionised and transferred into a gas phase where ions are separated based on their m/z ratio. So called "soft" ionisation sources such as electrospray ionisation (ESI) and matrix-assisted laser desorption ionisation (MALDI) developed in the 1980s allowed analysis of large intact biomolecules, widening the applications of such methods for analysis of biologics (Tian and Ruotolo, 2018). MS provides precise, accurate, high resolution analysis of proteins up to the MDa range. However, volatile buffers are required as well as expensive equipment and expert personnel (Den Engelsman et al., 2011). While proteins analysed under denaturing conditions provide greater sequence coverage and higher mass accuracy, native MS exists to analyse proteins in their native fold. This can be a powerful tool for detecting and analysing misfolding and aggregation of biopharmaceuticals (Tian and Ruotolo, 2018). Various chemical labelling techniques can be combined with MS to study protein structure and dynamics. Hydrogen/deuterium exchange (HDX) can be used to label surface exposed areas of a protein by using the principle that deuterium will exchange with the backbone amide hydrogens (and exchangeable protons on side chains), resulting in an increase in mass that can be detected by MS (Cornwell et al., 2018). Fast Photochemical Oxidation of Proteins (FPOP) uses a laser to split hydrogen peroxide into hydroxyl radicals, which can covalently bond with surface exposed residues, again producing an increase in mass that can be detected by MS (Cornwell et al., 2018). These techniques can be used to study changes in a protein's solvent exposure as a result of its

environment, the addition of a stress, or just over time, and so aid in understanding how this can influence the protein's aggregation behaviour.

After affinity maturation many hundreds of candidates are identified, making it necessary to screen these using quality control (QC) or "developability" assays which will generally be highly robust and high-throughput techniques (Den Engelsman et al., 2011). Often plate reader based assays are employed to quickly analyse large numbers of samples (Bhirde et al., 2018). However, when it comes to assessing aggregation what has become clear is that no single assay has the power to assess aggregation-prone sequences in isolation, presumably due to the various different complex pathways whereby a protein might aggregate, therefore multiple assays are generally deployed in combination (Willis et al., 2020). While SEC is the industry standard for analysing therapeutic protein aggregates, AUC and plate reader-based DLS are also used (Bhirde et al., 2018). Affinity-capture self-interaction nanoparticle spectroscopy (AC-SINS) is a high-throughput method capable of screening large quantities of mAbs for their propensity to self-associate (Liu et al., 2014). Gold nanoparticles (AuNP) are coated with polyclonal antibodies with high specificity for the Fc region of human mAbs. These coated AuNP are then incubated with a mAb of interest. If the mAb of interest (now immobilised on the AuNP) self-associates or aggregates with other mAbs on other nanoparticles, this results in reduced inter-particle distances between AuNP which changes the absorbance spectra of the AuNP (Figure 1.12B) (Liu et al., 2014). Increased antibody self-association results in an increase of the wavelength at maximum absorbance (plasmon wavelength) (Liu et al., 2014). The absorbance spectra of AuNP incubated with the mAb of interest are compared with AuNP alone to calculate a plasmon wavelength shift. This is a useful technique during early mAb development stages as it can use low concentrations of unpurified mAbs (Liu et al., 2014). Capillary electrophoresis (CE-SDS) is often used to characterise protein aggregates by separating a sample in a capillary by size when subjected to an electric field (Den Engelsman et al., 2011). This technique is usually combined with SEC in QC assays. CE-SDS is a rapid, high resolution technique that only requires low amounts of sample and, unlike SDS-PAGE, does not require staining for quantification as it uses UV absorption (Den Engelsman et al., 2011). However, this technique is unable to detect noncovalent aggregates and results may be impacted by the sample interacting with the capillary (Den Engelsman et al., 2011). Antibody clone self-interaction by bio-layer interferometry (CSI-BLI) involves immobilising an antibody of interest onto a bio-layer interferometry tip (Sun et al., 2013). The white light is reflected from the tip of the biosensor as well as internal reference layer, and the waves reflected from both these layers are combined to form the interference pattern (Ciesielski et al., 2016). Self-association of antibodies increases the thickness of the bio-layer, affecting the optical properties and resulting in a wavelength shift (Sun et al.,

2013). CSI-BLI is a high-throughput, label-free technique where a 96 well plate can be tested in as little as 2 hours, making it an attractive approach for screening early-stage discovery candidates and strong self-interactions as detected by CSI-BLI have been correlated to delayed retention times in SIC and CIC (Sun et al., 2013).

#### 1.4.1.2 Assessing binding affinity

When developing a biopharmaceutical, there are a number of 'drug-like properties' to be considered that contribute to that molecule being successful. This includes high affinity, specificity, and solubility as well as low toxicity, low immunogenicity, and slow clearance rates (Starr and Tessier, 2019). More than 20 methods to measure protein binding affinities and kinetics have been described in the literature (Vuignier et al., 2010). However, the most commonly used are isothermal titration calorimetry (ITC) (Ladbury and Chowdhry, 1996; Duff et al., 2011), surface plasmon resonance (SPR) (Willander and Al-Hilli, 2009; Kastritis and Bonvin, 2013), as well as fluorescence-based methods which correlate binding to a fluorescent output (Bee et al., 2013).

ITC measures heat uptake or release as a result of binding (Kastritis and Bonvin, 2013). Two cells are held in a microcalorimeter, one a reference cell containing buffer and the other containing the sample. An analyte is added by a series of injections. Binding of the analyte to the sample will result in a small change in temperature, and this change of temperature in the sample cell is compared with the reference cell (Vuignier et al., 2010). Each injection of the analyte will lead to a specific amount of protein complex, which is dictated by the binding affinity (Kastritis and Bonvin, 2013).

SPR measures the amount of protein complex formed between two molecules without using fluorescent or radioisotopic labels (Myszka and Rich, 2000). It works by immobilising one binding partner onto a sensor surface. The second binding partner (the analyte) is passed over the immobilised partner in a flow cell. Binding of the analyte to the immobilised partner results in a change in the refractive index at the sensor surface, which is measured over time (Figure 1.13) (Vuignier et al., 2010). Measuring this change of refractive index over time once the analyte is added and then removed can also give the equilibrium dissociation constant (Kd), or the propensity of the two molecules to dissociate, therefore giving information about the strength of the interaction (Kastritis and Bonvin, 2013).

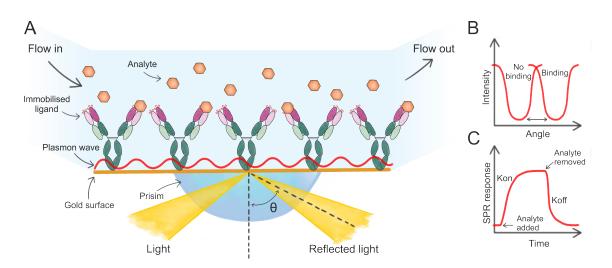


Fig. 1.13 Surface Plasmon Resonance (SPR) to measure antibody-antigen binding affinities. A) In SPR, one binding partner (here the mAb) is immobilised on the sensor surface. The second binding partner (the analyte) is washed over the immobilised partner through a flow cell. Binding of the analyte to the immobilised partner results in a change in refractive index at the sensor surface, which can be detected. B) The change in the angle of the refracted light is indicitive of binding. C) By measuring this over time once the analyte is added then removed, binding paramaters such as the association (Kon) and dissociation (Koff) constants can be calculated to determine the strength of the interactions. Redrawn and adapted from Myszka and Rich (2000).

Enzyme-linked immunosorbent assays (ELISAs) are also commonly used to measure protein-protein interactions for biopharmaceuticals, however they are more often employed to measure the presence of antibody in fluid or to look for promiscuous binding/polyspecificity rather than to determine specific binding affinities (Figure 1.14) (Syedbasha et al., 2016).

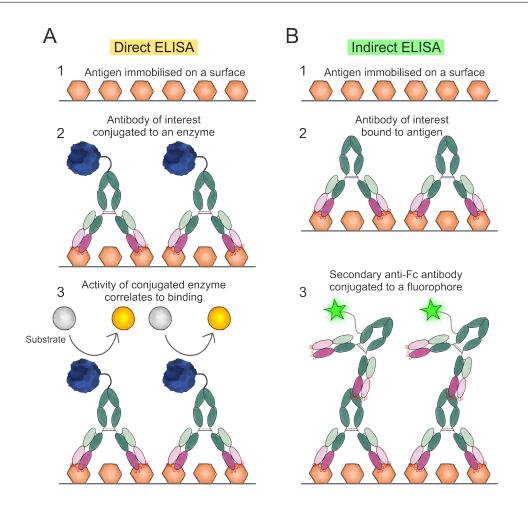


Fig. 1.14 **Direct and indirect Enzyme-Linked Immunosorbent Assays (ELISAs).** A) In direct ELISA, an antigen of interest is immobilised on a surface. The antibody of interest, conjugated to an enzyme (e.g. horseradish peroxidase) is incubated with the immobilised antigen. Unbound antibody is washed off, and the presence of binders is detected using a substrate for the conjugated enzyme that gives a detectable response, such as a colourmetric change or releasing a fluorophore. B) In indirect ELISA, the antibody of interst has no conjugated enzyme. Binding is detected using an anti-Fc antibody conjugated to something detectable, such as a fluorophore.

# **1.4.2** Computational approaches

Aggregation-prone regions (APRs) of proteins in the primary sequence can be predicted as they tend to be 5-15 amino acid long stretches of hydrophobic residues. However, depending on the tertiary structure of the protein these sequences may be packed into the core of the protein so may have less of an impact on aggregation compared with a solvent accessible stretch of hydrophobic residues on the surface (Wang and Roberts, 2018). Conversely, partial unfolding can expose otherwise buried APRs resulting in an enhanced probability of self-association. Various algorithms exist to predict APRs *in silico*, either sequencebased or structure-based. Sequence-based methods rely solely on analysis of the primary

sequence whereas structure-based methods take into account 3D structure of a protein and the solvent accessibility of residues (Elgundi et al., 2017). Examples of sequence-based methods include TANGO (Fernandez-Escamilla et al., 2004), PASTA (Trovato et al., 2007), Waltz (Maurer-Stroh et al., 2010; Louros et al., 2020) and AGGRESCAN (Conchillo-Solé et al., 2007). TANGO considers the physicochemical properties, such as pH or ionic strength, of a 5 residue segment to determine secondary structure formation propensity by considering different competing conformations (Fernandez-Escamilla et al., 2004; Buck et al., 2012). The probability of each residue in the segment occupying the  $\beta$ -aggregate conformation gives the protein a  $\beta$ -sheet propensity score which is used to estimate the probability of aggregation (Fernandez-Escamilla et al., 2004). Solubis is a combination of TANGO and the FoldX force field (Stricher et al., 2005), an online server that analyses the effect of mutations on a protein's thermodynamic stability by calculating the free energy based on structure (Rousseau et al., 2015). This algorithm mutates residues to gatekeeper residues (proline, arginine, lysine, aspartic acid and glutamic acid) and evaluates the impact on the TANGO score to identify APRs. Gatekeeper residues are electrostatically charged residues (and proline) that specifically aid in proper folding and oppose aggregation by blocking misfolding reactions and promoting proper folding (Rousseau et al., 2006). Prediction of amyloid structure aggregation (PASTA) considers the likelihood of sequences forming the cross- $\beta$  core that stabilises amyloid fibrils (Elgundi et al., 2017). Waltz predicts amyloid propensity based on a dataset of experimentally determined amyloid-forming hexapeptide sequences (Maurer-Stroh et al., 2010). The dataset has recently been updated to include 229 hexapeptides, some of which have been determined to be amyloidogenic using electron microscopy and Thioflavin-T binding assays, others were amyloidogenic peptides curated from the literature (Louros et al., 2020). Similar to TANGO, Waltz considers the physicochemical properties of the amino acids and their influence on amyloid propensity to identify APRs (Louros et al., 2020; Navarro and Ventura, 2022). AGGRESCAN calculates the aggregation propensity of an amino acid taking into account the neighbouring residues as well as the "hot spot" threshold (average aggregation propensity of all 20 amino acids scaled by their % composition in the Swiss-Prot data bank) to predict APRs or "hot spots" (Conchillo-Solé et al., 2007).

To overcome the limitations of sequence-based algorithms APRs can be mapped onto a protein's 3D structure in structure-based methods. AGGRESCAN 3D (A3D) is based on the original AGGRESCAN algorithm but takes into account the protein's 3D structure (Elgundi et al., 2017). This significantly improves the algorithms' ability to accurately predict the aggregation propensity of globular proteins (Zambrano et al., 2015). Additionally, AGGRESCAN 3D can incorporate molecular dynamics simulations to assess the influence of dynamic structural fluctuations on aggregation propensity, to increase the accuracy of

the algorithm (Zambrano et al., 2015). In 2019 an update for the AGGRESCAN 3D web server was released (AGGRESCAN 3D 2.0) (Kuriata et al., 2019). This update included an extension to the dynamic simulations to allow analysis of larger and multimeric proteins, as well as to analyse the effect of mutations on the stability and solubility in parallel, further improving the predictive ability of the algorithm. These dynamics simulations allow modelling of the flexibility of the protein and analyses the influence of this on the aggregation propensity, allowing detection of aggregation-prone regions that are exposed as a result of dynamic fluctuations. Furthermore, AGGRESCAN 3D 2.0 can virtually mutate all residues to charged residues (gatekeepers) and simultaneously evaluate their effect on the solubility and stability of the molecule to suggest a variety of beneficial variants. This can be particularly useful when studying scFvs, as the algorithm allows the user to select any residues they want to omit from the screen (such as CDRs) so mutations do not effect binding affinities (Kuriata et al., 2019). Spatial Aggregation Propensity (SAP) is a structure-based algorithm to identify solvent accessible patches of hydrophobic residues. SAP considers dynamic protein fluctuations that occur under physiological conditions and the effect on the size of hydrophobic patches by performing molecular dynamics simulations on the protein of interest (POI) (Buck et al., 2012). CamSol is a computational method developed to aid in the prediction of protein solubility and the rational design of therapeutics with improved solubility (Sormanni et al., 2015a). Unlike previously described methods that consider aggregation, CamSol calculates intrinsic solubility from the primary sequence of a protein then uses the solvent accessibility of residues to assess their impact on the overall solubility. The algorithm takes into account the physicochemical properties of amino acids to identify those that have the greatest impact on solubility. This is used to provide a solubility score - a solubility score +1 indicates a highly soluble region, whereas a score of -1 indicates a poorly soluble region.

The therapeutic antibody profiler (TAP) is a computational algorithm designed specifically to assess antibodies, or more accurately scFv fragments (Raybould et al., 2019). TAP was developed to assess five metrics that were thought to be related to poor developability in antibody therapeutics, with some potentially affecting a candidates aggregation propensity - the total length of the CDRs, levels of surface hydrophobicity, positive and negative charges in the CDRs and asymmetry in the net charges in the heavy and light chains. Using a large set of antibodies that were post phase-I in their development stage TAP was trained to identify threshold values for each metric, allowing identification of metrics where the test candidate differs significantly from the clinical stage therapeutics used to train the algorithm. This works under the assumption that clinical stage therapeutics posess desirable properties that enhance their 'developability', and having properties that differ from these may have detrimental effects. To assess an scFv, TAP requires the amino acid sequence of the  $V_H$  and  $V_L$  domains which it uses to build a structural model using ABodyBuilder (also developed by the same group) (Leem et al., 2016). Candidates are then compared to the database of clinical stage therapeutics to identify any significant differences from this database, as judged by the threshold values. The original algorithm was built using a database of 242 clinical stage therapeutics, however this is constantly being updated to enhance the accuracy of the predictions; currently the model is tracking 591 post Phase-I therapeutics (https://opig.stats.ox.ac.uk/webapps/newsabdab/sabpred/tap accessed 27th June 2022).

Another antibody-specific predictor algorithm is AbLIFT, an automated webserver that takes a V<sub>H</sub> and V<sub>L</sub> amino acid sequence and predicts affinity and stability enhancing mutations between the V<sub>H</sub> and V<sub>L</sub> interface (Warszawski et al., 2019). The model was built using deep mutational scanning (DMS) data of an anti-lysozyme antibody D44.1 whereby 135 positions along the scFv were subjected to saturation mutagenesis to give a library with every single amino acid at these positions. The library was transformed into yeast cells and used to select for mutants with enhanced affinity and those variants were subjected to deep sequencing using Illumina. Mutations that were enriched due to the selection were identified, with the majority being within the CDRs. However, a cluster of mutations was identified to be at the V<sub>H</sub> and V<sub>L</sub> interface, leading to the hypothesis that mutations at this region can improve affinity while simultaneously enhancing stability as mutations at this region have the potential to improve Fv assembly (Warszawski et al., 2019). Using this dataset and Rosetta based design methods, AbLIFT was built through a series of design-built-test cycles to predict potentially affinity enhancing and stabilising mutations in scFv sequences. While not specifically an algorithm to detect protein aggregation, suboptimal stability can often lead to increased levels of aggregation and mutations predicted using this algorithm have been shown to improve thermal stabilities and aggregation resistance (Warszawski et al., 2019). A similar algorithm has been developed by the same group based on experimental data using an enzyme (human acetylcholinesterase) has also been used to predict mutations that enhance protein expression and stability in bacteria (Goldenzweig et al., 2016). Furthermore, when assessing antibody based therapeutics *in silico* it may be more useful to use predictor algorithms that have been developed specifically for antibodies, and use experimental data from antibody-based drugs as these may be more accurate. In addition to algorithms for the prediction of antibody aggregation propensity and stability, various machine learning algorithms to predict performance in analytical developability assays such as hydrophobic interaction chromatography have been developed (Jain et al., 2017).

A complete description of all computational tools specifically developed for antibodies is beyond the scope of this thesis, and an overview of these has already been described elsewhere (Santos et al., 2020; Navarro and Ventura, 2022). These computational tools are useful for providing an insight into the underlying aggregation propensities and stabilities of proteins, however they have not yet surpassed *in vitro* techniques and so are still only advisory. They should therefore be combined with experimental data to get a complete understanding of a protein's aggregation behaviour.

# **1.5** Prevention and inhibition of protein aggregates

# **1.5.1** Promotion of protein refolding

Controlling aggregation is imperative to ensure biopharmaceuticals retain their activity while minimising adverse reactions in patients. This can be achieved by promotion of protein refolding; methods include reduction of temperature, reduction of protein concentration, or alteration of formulation (Wang, 2005). High temperatures and protein concentrations can increase aggregation rates by increasing intermolecular interactions. Furthermore, macromolecular crowding due to high protein concentrations present in final formulations of biopharmaceuticals can favour aggregation. Levels of reversible aggregates have been shown to decrease following dilution (Mahler et al., 2009). However, since many mAbs are required to be at high concentrations for intravenous delivery, dilution to reduce aggregation is often not an option.

Optimisation of the formulation buffer can be achieved using additives to promote protein refolding. Denaturants impact protein solubility at different concentrations; high denaturant concentrations weaken protein-protein interactions in water suppressing aggregation and increasing solubility (Ho et al., 2003; Wang, 2005). Addition of L-arginine has been shown to suppress protein aggregation and enhance solubility by blocking unfavourable intermolecular hydrophobic interactions between mAbs (De Bernardez Clark et al., 1999). Furthermore, the combination of L-arginine with L-glutamate has been shown to be even more effective at suppressing aggregation of mAbs (Kheddo et al., 2014). Other additives include; surfactants, cyclodextrins, PEG, Na<sub>2</sub>SO<sub>4</sub>, organic solvents, glycerol, and sucrose (Frokjaer and Otzen, 2005; Wang, 2005). It is important to understand how different additives impact the aggregation behaviour of different biologics in order to design the best formulation buffer; this can prove to be vital at ensuring new therapeutics are safe as well as reducing the time they take to get to market.

# **1.6** Protein engineering

It is known that proteins are only marginally stable in their folded states, which greatly limits their use in industrial and therapeutic applications (Taverna and Goldstein, 2002). Many industrial applications require proteins to be stable and functional at extreme conditions (such as high pH or temperature), functions that natural proteins rarely possess (Littlechild, 2015; Walia et al., 2017; Stimple et al., 2020). For a biopharmaceutical it is desirable to maintain a low aggregation propensity and high stability, as well as low viscosity at high concentrations and low off-target binding, amongst many other properties (Starr and Tessier, 2019). New advances in protein design and genetic engineering technologies allow structural modification of therapeutics to reduce aggregation propensity (Roberts, 2014a). However, it is not as simple as identifying and removing hydrophobic patches as this may impact folding and activity. Furthermore, when considering mAbs the most inherently aggregation-prone regions that are commonly identified in computational algorithms are often in functionally active regions such as the CDRs that are responsible for antigen binding as constant domains have more evolutionary conserved residues making them less prone to aggregation (Wang et al., 2009). Modification of these regions can affect binding affinity and reduce the mAbs activity. Therefore, various in vivo techniques combined with computational algorithms are employed to screen large libraries and detect aggregationprone "hot spots" within the primary sequence. These "hot spots" can be mapped onto the protein structure to identify functionally active regions and used to inform rational design of therapeutics (Roberts, 2014a) (Figure 1.15).

# **1.6.1** Rational design

Rational design of antibodies is generally used to engineer a number of drug-like properties, such as aggregation resistance or affinity, and usually involves using either experimental or computational tools (or a combination of both) to design variants with improved properties. For aggregation resistance, commonly these approaches are used to identify aggregation-prone regions and mutations designed to rectify this aggregation behaviour (Figure 1.15). For rational design, it is required to have specific structural information about the protein that is being designed, as well as information about their aggregation and stability behaviours (if these are the properties being designed). The computational tools described in Section 1.4.2 have been applied to specifically design improved variants of an antibody of interest. Solubis has been exploited to aid in the rational design of mAbs with reduced aggregation propensity (van der Kant et al., 2017). Although the presence of APRs

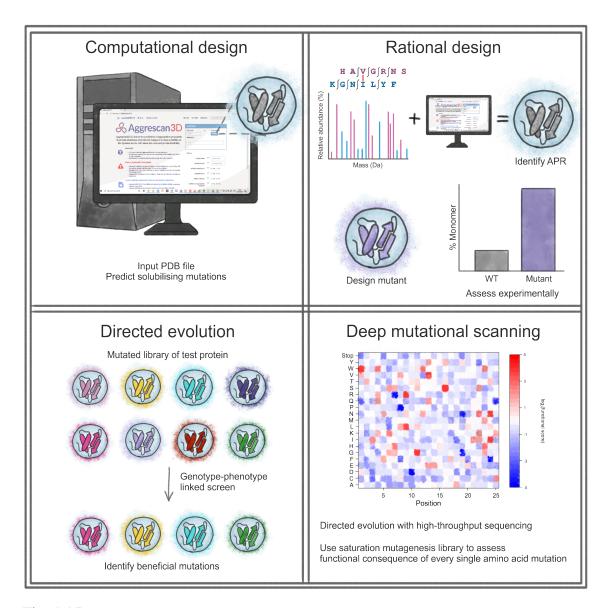


Fig. 1.15 **Overview of protein engineering techniques.** *In silico* tools can be used to computationally mutate a test protein to assess the affect of these mutations on the aggregation propensity and produce a list of potentially solubilising mutations. Rational design involves identifing aggregation-prone regions, often utilising a combination of computational (using *in silico* predictor softwares such as Aggrescan 3D 2.0) and experimental tools (e.g using NMR (spin labelling) or mass-spectrometry (XL-MS or HDX)), and designing mutations to resolve the APR. Mutations are then assessed using experimental biophysical methods. Directed evolution uses a mutated library of a test protein and selects for beneficial mutations using a genotype-phenotype coupled screen. Deep mutational scanning (DMS) is similar to directed evolution but uses high-throughput sequencing to allow massively-parallel analysis of the functional effect of mutations. It usually requires systematic libraries with every single amino acid mutation as single point mutations at some or all of the residues in the test protein. Also, like directed evolution, requires a genotype-phenotype coupled screen for selection. Redrawn and adapted from Ebo et al. (2020a).

was detected in CDRs, variants with reduced aggregation propensity while still retaining antigen binding affinity were designed by artificially mutating aggregation-prone residues in other APRs to gatekeeper residues to identify stabilising mutants. Similarly SAP has been used to reduce the aggregation propensity of bevacizumab, a therapeutic mAb used for the treatment of various cancers (Courtois et al., 2016). SAP was used to identify APRs and single point mutations were introduced by site-directed mutagenesis. Additionally, four glycosylation sites were engineered on the Fab domain to mask residues in APRs with a carbohydrate. These approaches led to significant decreases in aggregation without compromising binding affinity (Courtois et al., 2016). AGGRESCAN3D 2.0 (A3D 2.0) has been used to design a V<sub>H</sub> antibody with increased aggregation resistance (Gil-Garcia et al., 2018). A V<sub>H</sub> segment of the human germline antibody DP47 was analysed using the A3D 2.0 algorithm that virtually mutated the strongest aggregation prone regions to gatekeeper residues and analysed the impact of these mutations on the stability (FoldX) and solubility (A3D) of the protein. Three APRs were detected and mutating a single residue within each of these APRs to lysine was predicted to be both stabilising and solubilising. A triple lysine mutant was engineered and analysis using light scattering showed this variant was 3-fold more resistant to aggregation compared with wild type (Gil-Garcia et al., 2018; Kuriata et al., 2019).

Nanobodies specific to intrinsically-disordered proteins (IDPs) have been designed using a computational approach. This approach seeks to identify pairs of short peptides which interact via interfaces which are contiguous in their primary sequence. To do this, they use the Protein Data Bank (PDB) to identify short peptides in  $\beta$ -strand to  $\beta$ -strand interactions as unlike, for example,  $\alpha$ -helical interactions each residue along the chain interacts with a complementary residue in the opposing strand (Sormanni et al., 2015b). This is used to generate a dataset of interacting peptides (complementary peptides), which can be used outside of the structural context of the original protein in the PDB to design binding peptides against a region of interest in an IDP of therapeutic interest. These peptides are then grafted onto the CDR3 of a human V<sub>H</sub> domain and screened for affinity to the target. Nanobodies specific for  $\alpha$ -synuclein, A $\beta$ 42 and IAPP were successfully generated and shown to have high affinities and specificities, as well as inhibit aggregation of their targets at substoichiometric concentrations (Sormanni et al., 2015b). An alternative approach to rationally design nanobodies against amyloid domain proteins involved grafting short peptide segments from the target amyloid protein into CDR3 of a V<sub>H</sub> domain creating so-called "grafted amyloid-motif antibodies", or gammabodies (Julian et al., 2015; Lee et al., 2016). The resulting gammabodies bind to the amyloid proteins with high specificity via homotypic interactions between the peptide fragment in HCDR3 and the corresponding peptide within fibrils (Lee et al., 2016). In other words, these

nanobodies work like poisoned monomers which presumably terminate the elongation of the fibril. Rational design of gammabodies was combined with directed evolution approaches to develop nanobodies with conformational specificity for fibrils, using a library design method that introduces mutations based on the relative frequency of specific amino acids at specific positions within HCDR3 in human antibodies and exploiting yeast surface display combined with MACS to identify variants that specifically bind to AB fibrils (Julian et al., 2019). Combined computational analysis and directed evolution methods have been used to engineer proteins with enhanced expression, solubility, and stability. One study evolved two human proteins with known drug developmental issues using ribosome display (Buchanan et al., 2012). Granulocyte colony-stimulating factor (G-CSF) has solubility issues when expressed in *Escherichia coli* and erythropoietin (EPO) is thermodynamically unstable and prone to aggregation at increased temperatures (Buchanan et al., 2012). These were evolved using ribosome display, resulting in expression levels of G-CSF improving 1000-fold and aggregation of EPO reducing from 80% to undetectable levels. The predicted impact of mutations that arose as a result of directed evolution using ribosome display were analysed using in silico methods to understand how these mutations were improving the properties limiting production of these proteins.

# **1.6.2** Directed evolution

Over the last 3.5 billion years life on Earth has been adapting and evolving, facilitated by proteins developing innovative and creative solutions to enable organisms to grow and survive across a diverse range of environments. Since the advent of recombinant DNA technology, protein engineers have been working to exploit and expedite Nature's evolutionary processes to evolve and improve various protein functions. Directed evolution utilises the principles of Darwinian evolution whereby genetic diversity is introduced into the test protein which is then subjected to a selective pressure (Figure 1.16A). Compared with natural evolution, directed evolution has higher mutation rates to accelerate the process. By using an appropriate genotype-phenotype screen, rare beneficial mutations are enriched and can be identified (Figure 1.15) (Foit et al., 2009; Julian et al., 2017; Wang et al., 2018). However, improving protein stability often proves challenging as most mutations are destabilising. Furthermore, a common challenge for directed evolution studies is that often there is a trade-off between particular properties, such as stability and function, and by selecting for one you can negatively impact the other (Julian et al., 2017). Nevertheless, directed evolution has proven to be invaluable for engineering proteins from developing monoclonal antibodies that treat cancers to enzymes that produce biofuels (Smith, 1985; Winter et al., 1994; Heater et al., 2019).

#### 1.6.2.1 Creating DNA libraries

#### 1.6.2.1.1 In vitro mutagenesis

The first step in any directed evolution experiment is the creation of genetic diversity upon which selection pressures can be applied. Early work developing directed evolution techniques for engineering enzymes in the 1990s used random mutagenesis technologies to create genetic diversity (Chen and Arnold, 1993, 1991; Currin et al., 2021). Error-prone PCR (epPCR) is by far the most popular of these techniques owing to its ease of use. It works by using an error-prone DNA polymerase (DNAP) to randomly generate mutations during PCR amplification, or by modifying the buffer components to decrease the fidelity of standard DNAP (Figure 1.16B) (Leung et al., 1989; Wang et al., 2021). Reaction components can be modified to increase the mutation rate, such as by using unbalanced dNTP concentrations, increasing the concentration of magnesium ions, increasing the number of PCR cycles, or adding manganese ions (Wang et al., 2021). However, epPCR has limitations: often the DNAP has a bias for certain nucleotide substitutions over others which can affect the amino acids available for a particular codon. Additionally, consecutive mutation of two bases is rare which can further reduce the possible amino acids available; it requires large amounts of screening in order to sample the entire library; and can result in stop codons as well as insertions and deletions (Leung et al., 1989; Wang et al., 2021). Despite these limitations, epPCR is still widely employed and has been used successfully to engineer the properties of proteins, such as to increase aggregation-resistance in biopharmaceuticals (Ebo et al., 2020b), to increase enzyme activity (Nearmnala et al., 2021), and to determine protein fitness landscapes using DMS (Seuma et al., 2021; Ren et al., 2021). Another often-used method for in vitro gene diversification is DNA shuffling, wherein libraries are created by random fragmentation and recombination of homologous DNA sequences (Figure 1.16C) (Stemmer, 1994). This approach is especially useful for mixing and combining a library of mutants that have already been evolved and selected as beneficial, in order to combine advantageous characteristics and improve them further. Since its invention, DNA shuffling has been widely used and adapted to engineer a wide range of properties, including improved thermostability (Hao and Berry, 2004), improved catalytic activity (Nearmnala et al., 2021), and to develop chemogenetic fluorescent reporters with tuneable fluorescent properties (Benaissa et al., 2021).

Targeted gene mutagenesis methods have been developed to overcome the limitations of classic random mutagenesis methods, and have been reviewed at length elsewhere (Currin et al., 2021). In short, recent advances in solid-phase DNA synthesis methods allows tight control over designed libraries, allowing the effect of defined sets of amino acid

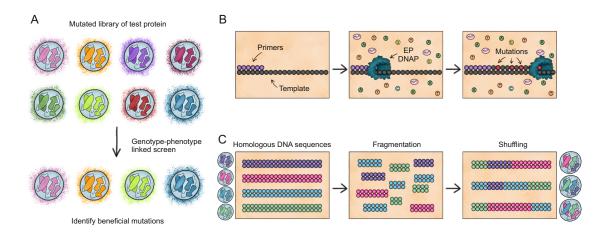


Fig. 1.16 Directed evolution and *in vitro* mutagenesis. A) A directed evolution experiment works by creating a library of gene variants of a protein-of-interest and subjecting them to a selective pressure to identify beneficial mutations. B) Error-prone PCR (epPCR) uses an error-prone DNA polymerase (EP DNAP) to amplify a gene of interest and introduce mutations. Alternatively, the buffer conditions can be modified to increase the mutation rate of a standard DNAP, such as by adding magnesium ions (pink) or having unbalanced dNTP concentrations (A, green; T, orange; G, yellow; C, blue). C) DNA shuffling allows mixing of homologous sequences, such as variants of the same protein with single point mutants, to create hybrid genes combining different mutations. Libraries are created by random fragmentation of genes, which are then joined together using primer-free PCR.

substitutions in focussed regions of interest (e.g. antibody CDRs) or the entire primary sequence to be determined (Currin et al., 2021). Such libraries are particularly useful for DMS experiments as they allow understanding of protein functional landscapes and can be used to uncover the contribution of the identity (e.g., amino acid side chain chemistry) and individual residues to protein function, stability and/or aggregation (Fowler and Fields, 2014).

Natural Diversity Mutagenesis is a more rational approach specifically designed to create mutated libraries for antibody variable domains (Tiller et al., 2017b). Here a computational alanine scan is utilised to identify permissive sites in the CDRs. Degenerate codons are used at these permissive sites to introduce the most frequently occuring residues at these positions in human antibodies, as determined by the abYss database (Swindells et al., 2017; Tiller et al., 2017a). Libraries created using this method and screened by yeast surface display were shown to result in >5-fold gains in affinity as a result of four to six CDR mutations, highlighting both the success of this rational library design apporach as well as the importance of assessing multiple amino acid mutations in combination to gain large improvements in affinity (Tiller et al., 2017a).

#### 1.6.2.1.2 In vivo mutagenesis

*In vivo* mutagenesis approaches involve altering the genome sequence of an organism only, or altering the sequence of genetic material within an organism (i.e. plasmids) via the addition of mutagens (such as chemicals or UV light), or the use of hypermutator strains that contain deletions or modifications in genes for enzymes involved in proofreading, mismatch-repair, and base-excision (such as XL1-Red) (Badran and Liu, 2015; Greener et al., 1997). Alternatively, various examples of mutagenic plasmids expressing different mutagenic enzymes involved in mismatch repair, translesion synthesis, and proof-reading have been developed with a wide range of induced mutagenic potency to globally increase the mutation rate in *E. coli* (Badran and Liu, 2015). These strategies have the potential to yield high mutation rates (up to 322 000-fold over wild type *E. coli*). Such methods can be problematic as the accumulation of mutations throughout the *E. coli* genome can result in toxic mutations if they occur within essential regions of the genome. Alternatively, these mutations accumulating outside of the gene-of-interest (GOI) could allow the bacteria to circumvent the selection pressure.

To overcome these limitations, targeted *in vivo* mutagenesis strategies have been developed. An early example of this strategy is the use of a mutated *E. coli* polymerase I (pol I) that selectively mutates genes on a ColE1 plasmid (although mutations are limited to within a few kb of the ColE1 origin) (Allen et al., 2011; Camps et al., 2003). Furthermore, pol I still replicates parts of the genome, which can result in off-target mutations (Allen et al., 2011).

A popular method of *in vivo* mutagenesis is fusing specific DNA binding proteins to DNA-mutating enzymes. An example of this is MutaT7, wherein a cytidine deaminase is fused to T7 RNA polymerase (RNAP) to continuously direct mutations to specific, well-defined, DNA regions of any length in *E. coli* (Moore et al., 2018). This allows targeted mutagenesis of genes under the control of the T7 promoter (Figure 1.17A). However, this approach has the potential to accumulate off-target effects, which can be problematic, particularly in the promoter regions. For example, they can potentially inhibit expression of the GOI, or lead to escape mutations, which allow the cells to evade the selection pressure applied without evolving the GOI. Furthermore, as this method utilises cytidine deaminases, their specific activity is limited to C>T and G>A mutations. Alternative cytidine deaminases have been employed to increase the mutation rate and expand the applicability of this method (Park and Kim, 2021), and MutaT7 has also been adapted for use in eukaryotic cells (TRACE; T7 polymerase-driven continuous editing) (Chen et al., 2020).

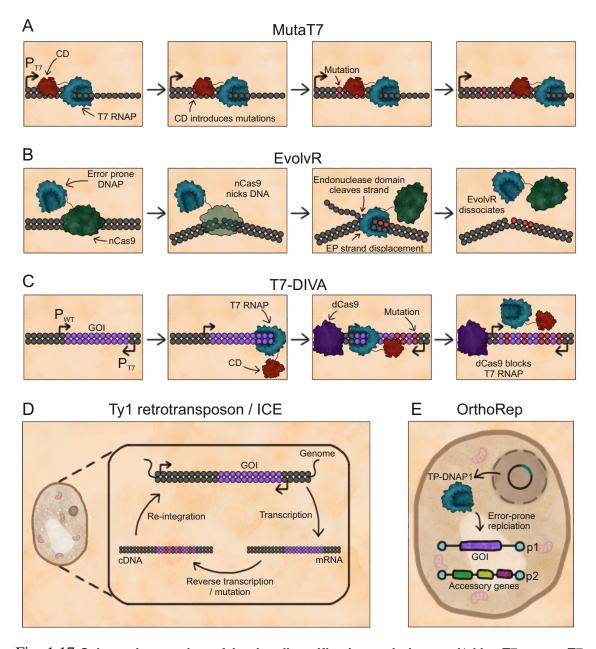


Fig. 1.17 Schematic overview of *in vivo* diversification techniques. A) MutaT7 uses a T7 RNA polymerase (RNAP) fused to a cytidine deaminase (CD), which enables targeted mutations to genes under the T7 promoter (PT7). B) EvolvR uses an error-prone DNA polymerase (DNAP) fused to a nickase Cas9 (nCas9), which enables targeted mutation within regions adjacent to the nick site via error-prone (EP) strand displacement. C) T7-targeted dCas9-limited *in vivo* mutagenesis (T7-DIVA) enables targeted mutagenesis of genes without altering their genomic promoter (PWT). By introducing an antisense PT7, a T7 RNAP fused to a cytidine deaminase (CD) is able to introduce mutations. A catalytically dead Cas9 (dCas9) is used as a 'roadblock' to demarcate the boundaries of mutagenesis. D) Ty1 retrotransposon mutagenesis, or *in vivo* continuous evolution (ICE), uses native yeast retrotransposon Ty1. The replication cycle of Ty1 is error-prone, so by introducing an inducible gene-of-interest (GOI), each time Ty1 is replicated mutations will accumulate. E) OrthoRep uses an orthogonal plasmid/polymerase pair (TP-DNAP1/p1) whereby the error-prone TP-DNAP1 (expressed from a nuclear plasmid) replicates p1 and introduces mutations. All accessory genes required for the replication of p1 are encoded on a second plasmid (p2) to spare them from mutagenesis.

A similar method (EvolvR), developed for use in both yeast and bacteria, utilises a fusion between an error-prone DNA polymerase (DNAP) and a nickase-Cas9 (nCas9), which allows mutations within a region adjacent to the Cas9 nick site (Figure 1.17B) (Halperin et al., 2018; Tou et al., 2020). The mutation rate can be tuned by using polymerases with different fidelities ( $\sim 1$  in  $10^7 - 10^3$ ) and this method enables all possible nucleotide substitutions, unlike those utilising cytidine deaminases. The approach is limited due to elevated off-target mutation rates ( $\sim 1$  in  $10^{11} - 10^8$ ) and the narrow mutation window within the sequence (most mutations occur within 50 bp of the nick site).

T7-targeted dCas9-limited *in vivo* mutagenesis (T7-DIVA) utilises a similar method whereby T7 RNAP fused to a cytidine deaminase is used to introduce mutations (Figure 1.17C) (Álvarez et al., 2020). The GOI can remain under the control of its genomic promoter and a T7 promoter is inserted downstream of the GOI on the antisense strand. This allows the T7 RNAP to translocate along the GOI and to introduce mutations without altering the endogenous 5' promoter. A catalytically dead Cas9 (dCas9) is used as a 'roadblock' demarcating the boundaries of the mutagenesis, enabling targeted *in vivo* mutagenesis of specific genes. However, as this method requires introduction of a downstream T7 promoter, it is unable to mutate specific regions of a GOI.

Error-prone DNA replication utilising the native yeast retrotransposon Ty1 has been developed for selective mutation of genes inserted between long terminal repeats (Figure 1.17D) (Crook et al., 2016). The replication cycle of Ty1 occurs via an RNA intermediate that is converted into complementary DNA through an encoded reverse transcriptase and re-integrated back into the genome. Heterologous gene expression from Ty1 has previously been demonstrated and the replication cycle has been shown to be error-prone (Crook et al., 2016). This enables random mutations to accumulate within a GOI expressed off Ty1, without any bias towards transitions or transversions over lengths of 5 kb (Crook et al., 2016). However, as the diversification occurs across the whole length of the Ty1 retrotransposon leading to escape mutations to evade the selection pressure (Crook et al., 2016). Nonetheless, its large mutagenesis window makes this approach a powerful tool for *in vivo* continuous evolution of entire biosynthetic pathways (Crook et al., 2016).

OrthoRep is an extranuclear replication system in *Saccharomyces cerevisiae* consisting of an orthogonal DNA polymerase-DNA plasmid (TP-DNAP1/p1) pair (Ravikumar et al., 2014). It involves an engineered error-prone DNA polymerase (TP-DNAP1) that selectively replicates a specific plasmid (p1) encoding the GOI and introduces mutations (Figure 1.17E). TP-DNAP1 is expressed in trans from a yeast nuclear plasmid and a second polymerase/plasmid (TP-DNAP2/p2) pair encodes all the essential accessory genes for

replication, transcription, and maintenance of p1 and p2, sparing them from error-prone replication and reducing off-target effects (Javanpour and Liu, 2019). This method was developed further to adapt the TP-DNAP2/p2 pair for error-prone replication, which would allow two mutationally orthogonal DNA replication systems within the same cell, each with different custom mutation rates (Arzumanyan et al., 2018). However, this method still does not enable targeted mutagenesis of specific regions of a GOI, as the polymerase replicates the entire plasmid. Nevertheless, OrthoRep has been used to evolve a wide range of proteins, including enzymes with promiscuous activities (Rix et al., 2020), small molecule biosensors (Javanpour and Liu, 2021), and antibody fragments (Wellner et al., 2021).

#### **1.6.2.2** Screening technologies

#### 1.6.2.2.1 Display technologies

Phage (Wojcik et al., 2010), yeast surface (Julian et al., 2019) and ribosome (Buchanan et al., 2012) display techinques, as described in Section 1.3.2, have all been utilised to evolve affinity in antibody based drugs. These methods can also be utilised to evolve thermodynamic stability and aggregation resistance by modifying the protein folding conditions, such as carrying out display experiments at increased temperatures (Jespers et al., 2004; Park et al., 2006; Jones et al., 2011; Pavoor et al., 2012). However, each have their own negative aspects and so various modifications on these classic assays have been developed for enhancing their evolutionary properties. Ribosome display has been enhanced by combining it with next-generation sequencing (NGS) to allow the high-throughput identification of antibody-specific peptide ligands to aid in analytical identification of antibodies in human serum (Heyduk and Heyduk, 2014). Phage display has been improved in a similar way by exploiting NGS techniques (Christiansen et al., 2015). In a separate study phage display was enhanced by modifying various paramaters within the display system; modifying the signal peptide that translocates the P3-scFv fusion protein to the periplasm from the posttranslational OmpT signal sequence to the co-translational DsbA signal sequence, as well as modifying the culture conditions to use baffled flasks was shown to improve display levels over 1000 fold (Wojcik et al., 2010). Furthermore, phage display has been modified to add in a selection for protein stability by displaying nanobodies on the surface of phage, heating to induce unfolding followed by cooling then screening displayed nanobodies against a coformational ligand specific for folded V<sub>H</sub> domains (protein A), resulting in identification of soluble, aggregation-resistant

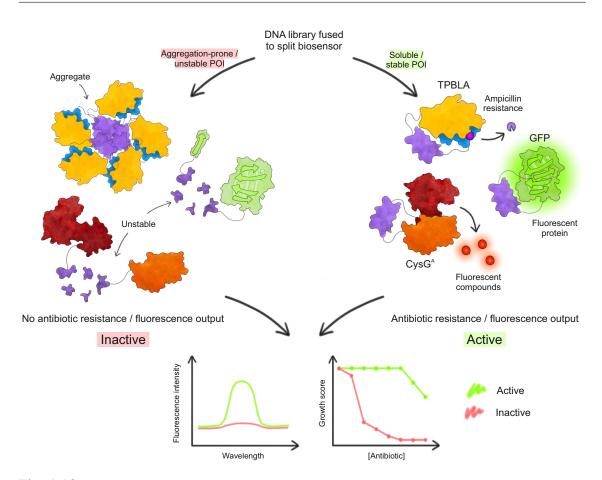
nanobodies (Jespers et al., 2004). Yeast surface display has been modified in a similar way, by utilising protein A to include a selection for protein stability (Julian et al., 2015).

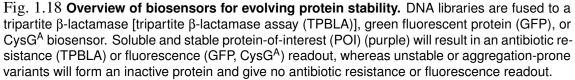
#### 1.6.2.2.2 Protein reporter biosensors

Green fluorescent protein (GFP) is able to be split into two halves that can generate fluorescence via their noncovalent reassembly (Baird et al., 1999; Ghosh et al., 2000). Making use of this property of GFP, an array of different systems has been developed enabling fluorescence to be correlated with protein stability, solubility, or the ability of a protein of interest (POI) to interact with a target protein (Figure 1.18) (Ghosh et al., 2000; Lindman et al., 2010; Magliery et al., 2005; Golinski et al., 2021). However, the main issue with using fluorescent proteins as reporters is that GFP itself, which is added to the POI via a short linker, may alter the properties of the POI. Issues can arise because the fluorescent protein itself can remain fluorescent even when the POI aggregates if the rates of aggregation of the POI are slower than the rate of formation of the chromophore (Kothawala et al., 2012). Early examples of using a split GFP assay for evolving stability were relatively low throughput as they involved individually picking colonies of E. coli displaying increased fluorescence levels (Lindman et al., 2010). More recently, the split GFP assay has been expanded into a high-throughput assay by utilising FACS and deep sequencing to identify soluble variants of Gp2 (an affibody) (Golinski et al., 2021).

A number of alternative split fluorescent proteins have been developed to expand the usefulness of these systems as they encompass a range of excitation and emission wavelengths (Feng et al., 2017, 2019; Tamura et al., 2021). Recently, a split luciferasebased biosensor was developed for detecting SARS-CoV-2 (anti-severe acute respiratory syndrome coronavirus 2) antibodies in patient sera (Elledge et al., 2021; Yao et al., 2021). This method has not yet been used for directed evolution, but it is similar in concept to the split GFP assay for evolving binding affinity and has the potential to be utilised in the same way (Magliery et al., 2005; Rozbeh and Forchhammer, 2021).

Recently a tripartite biosensor using *E. coli* uroporphyrinogen-III methyltransferase CysG<sup>A</sup> was developed in which the POI is inserted into a loop of CysG<sup>A</sup> and used to evolve protein stability (Figure 1.18) (Ren et al., 2021). CysG<sup>A</sup> catalyses the formation of fluorescent compounds, therefore by inserting a POI within a permissive site in CysG<sup>A</sup> protein stability can be correlated with a fluorescence readout (Ren et al., 2021). The assay was first evaluated using variants of the *E. coli* immunity protein 7 (Im7), along





with maltose binding protein and acylphosphatase. It was then used in a deep mutational scan to unpick the contribution of individual residues of the catalytic domain of a histone H3K4 methyltransferase to understand its stability landscape (Ren et al., 2021). As a result of the reducing environment of the *E. coli* cytoplasm, proteins that require disulfide bonds, such as antibody fragments and many enzymes, cannot be analysed using this system. In the first report of the CysG<sup>A</sup> system the authors used manual inspection to select bacteria with increased fluorescence. Combining the assay with FACS followed by deep sequencing, however, has the potential to expand its capabilities so that variants with improved properties can be selected in a high-throughput manner.

A number of groups have demonstrated that  $\beta$ -lactamase can be used as a selectable reporter to assess stability and protein-protein interactions, as well as to engineer these properties into POIs (Galarneau et al., 2002; Guntas and Ostermeier, 2004; Guntas et al., 2005; Edwards et al., 2008, 2010; Foit et al., 2009; Saunders et al., 2016; Ebo et al., 2020b). This is discussed in more detail in Section 1.7.

#### 1.6.2.2.3 In vivo continuous evolution

Phage Assisted Continuous Evolution (PACE) uses filamentous bacteriophage (selection phage, SP) to enable fully *in vivo* continuous evolution of a wide range of protein properties (Esvelt et al., 2011; Wang et al., 2018; Blum et al., 2021; Morrison et al., 2021; DeBenedictis et al., 2022). In this method, a population of SP is continuously diluted in a fixed volume of E. coli, known as the 'lagoon'. The gene of interest (GOI) replaces the gene (gIII) for the minor coat protein (pIII) within these SP, a protein which is required for a phage to be infectious. Therefore, the gene for pIII is supplemented on an accessory plasmid (AP), where its expression is dependent on the property of the GOI being evolved. Variants are only able to persist if they are able to produce enough pIII before being diluted out of the lagoon (Figure 1.19A) (Wang et al., 2018). To allow continuous directed evolution, an arabinose inducible mutagenesis plasmid (MP) is used which increases the E. *coli* global mutagenesis rate by expressing mutagenic enzymes involved in mismatch repair, translesion synthesis and proof-reading (Section 1.6.2.1.1) (Badran and Liu, 2015). This enables mutations to accumulate within a GOI. Since its conception, PACE has been used to evolve a diverse range of proteins; including polymerases with new recognition sites (Esvelt et al., 2011), dehydrogenases with improved activity (Roth et al., 2019), proteases with novel specificities (Packer et al., 2017; Blum et al., 2021), biosynthetic pathways (Johnston et al., 2020), antibody fragments (Wang et al., 2018; Morrison et al., 2021), proteins for DNA binding and manipulation (Miller et al., 2020; Thuronyi et al., 2019; Richter et al.,

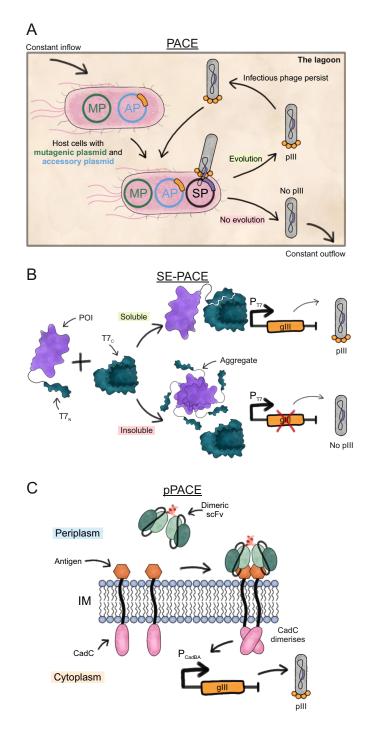


Fig. 1.19 Schematic of phage-assisted continuous evolution (PACE) and its related screens. A) In PACE, expression of phage protein pIII (via the gene gIII) is linked to the property of the protein-of-interest (POI) being evolved. Successful variants will be able to express pIII, and the resulting phage will be able to infect a new host. Unsuccessful variants will not be able to express pIII and the resulting phage will be diluted out of the fixed volume 'lagoon'. B) Soluble expression PACE (SE-PACE) splits T7 RNAP into two parts (T7N and T7C) and fuses a POI to T7N. By expressing gIII under a T7 promoter (PT7), only the expression of a soluble POI will result in gIII expression and infectious phage. C) Periplasmic PACE (pPACE) replaces the periplasmic sensor domain of CadC with an antigen and puts gIII under the control of the PCadBA promoter, which is switched on in response to dimerisation of CadC. By expressing a dimerising scFv, gIII will be expressed if the dimerising scFv binds to the antigen. This figure is redrawn and adapted from Wang et al. (2018); Esvelt et al. (2011); Morrison et al. (2021). Abbreviations: AP, accessory plasmid; IM, inner membrane; MP, mutagenesis plasmid; SP, selection phage.

2020), and novel quadruplet tRNAs for genetic code expansion (DeBenedictis et al., 2021, 2022).

As well as evolving protein function, PACE has been utilised to evolve protein solubility (soluble expression PACE; SE-PACE) by linking pIII expression to the soluble expression levels of a POI (Wang et al., 2018). By utilising a split intein pIII, the method has included a selection for binding affinity of antibody fragments (scFvs) to ensure this property is not lost when evolving solubility (Figure 1.19B). SE-PACE has been successfully utilised to evolve scFvs with soluble expression yields improved up to 5-fold with comparable binding affinities to the wild type. However, its incredibly complex nature makes it difficult to use as it requires technical expertise and equipment, and it is difficult to design the genetic circuits to link pIII expression to the property being evolved. Furthermore, the fact that screening for binding occurs in the cytoplasm could be problematic for assessing antibody-based drugs as they contain disulphide bonds that may not form properly in this oxidising environment.

Consequently, PACE has recently been adapted to carry out evolution in the oxidising environment of the *E. coli* periplasm, termed periplasmic PACE (pPACE) (Morrison et al., 2021). This approach uses the natural *E. coli* transmembrane transcriptional activator CadC, which is part of a two-component sensor that transduces signals in the periplasm to the cytoplasm. CadC senses acidic pH and high lysine levels in the periplasm, causing the periplasmic sensor domain to dimerise and bind two motifs on the CadBA promoter and initiate gene transcription(Figure 1.19C) (Kuper and Jung, 2005). By replacing the periplasmic sensor domains of CadC with antigens and expressing a dimerising scFv, binding of these two proteins in the periplasm can be linked to gene expression of pIII which is under the control of CadBA promoter (Figure 1.19C) (Morrison et al., 2021). pPACE has been used to evolve novel protein-protein interactions and restore binding between subunits of the homodimeric YibK; to restore binding affinity of a non-binding mutant of an anti-GCN4  $\Omega$ -graft antibody, as well as improve its soluble expression levels ~8-fold; and to evolve a ~2-fold improvement in binding affinity and ~5-fold improvement in soluble expression of the scFv fragment Trastuzumab (Morrison et al., 2021).

Despite its advantages, a number of challenges remain to be overcome with PACE: experiments have a high failure rate whereby phage expressing the evolving protein frequently "wash out" meaning the selection pressure is too high, and experiments are difficult to multiplex (DeBenedictis et al., 2022). To overcome this, PACE has been as miniaturised and extended as Phage-and-Robotics-Assisted Near-Continuous Evolution (PRANCE), which automates the process of continuous evolution utilising a liquid handling robot and a 96-well plate format to enable multiplexing (DeBenedictis et al., 2022). To

reduce the failure rate, PRANCE uses real-time monitoring of phage activity by expressing luciferase alongside pIII to give a read-out of phage propagation and to trigger a feedback control whereby selection pressure is modified depending on luminescence. PRANCE was used recently to characterise the evolutionary fitness landscape of T7 RNAP to recognise the foreign T3 promoter by conducting 90 simultaneous evolutions (DeBenedictis et al., 2022). As small volumes are required, PRANCE also allows the evolution of aminoacyl-tRNA synthetases to incorporate non-natural amino acids, as well as allowing multiplexed evolution of quadruplet tRNAs (DeBenedictis et al., 2022), both of which require expensive reagents.

#### **1.6.3** Deep mutational scanning (DMS)

Next-generation sequencing (NGS) technologies have revolutionised the genomics field due to their ability to provide sequencing analysis at a massively high-throughput scale, from enabling de novo sequencing of genomes, allowing high-throughput analysis of microbiomes, as well as assessing transcriptomes and enabling quantification of translation levels (Wheeler et al., 2008; Huang et al., 2009; Qin et al., 2010; Schadt et al., 2010). Furthermore, these technologies have subsequently facilitated huge advances in the protein engineering field by enabling comprehensive analysis of the functional consequence of thousands of variants when combined with a phenotypic screen (Fowler and Fields, 2014). In contrast, first-generation sequencing is massively limiting due to its low-throughput and high cost. NGS has been widely combined with directed evolution using systematic libraries containing every single amino acid substitution at every position in a POI, which enables comprehensive characterisation of the stability or functional consequence of every possible substitution, a method known as deep mutational scanning (DMS). These libraries will either be homemade using saturation mutagenesis or synthesised, meaning they are either very time consuming or very expensive. For example, a synthesised saturation mutagenesis library for a 300 amino acid protein would be around \$15,000 (Twist Bioscience, San Francisco, CA). Nevertheless, DMS using these saturation mutagenesis libraries has proven to be incredibly useful in understanding health and disease, for example when investigating amyloidogenic proteins  $\alpha$ -synuclein (Newberry et al., 2020) and TDP-43 (Bolognesi et al., 2019), involved in Parkinson's disease and amyotrophic lateral sclerosis (ALS), respectively. Furthermore, DMS has been successfully exploited to assess therapeutic proteins such as for identifying regions of single-chain variable fragments (scFvs) involved in aggregation, instability or affinity and driving development of predictor algorithms (Warszawski et al., 2019). While DMS has proven to be a powerful method for a number of applications, most notably gaining a broader understanding of sequence-function

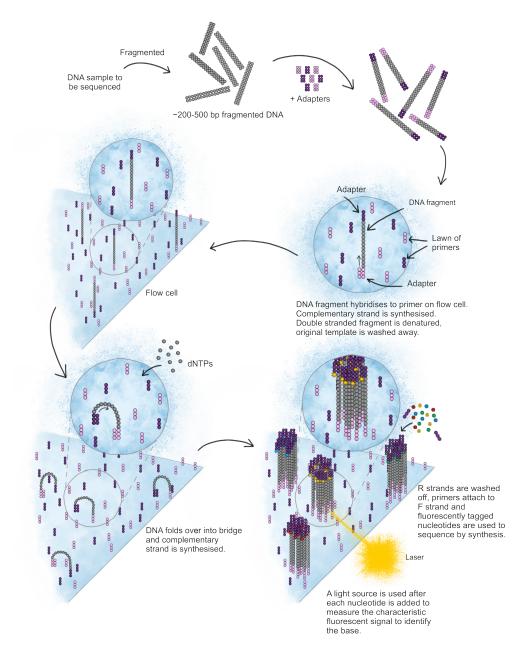


Fig. 1.20 Overview of Illumina sequencing technologies. The DNA to be sequenced is fragmented, using enzymatic approaches or sonication. Adapters are added onto the 5' and 3', allowing hybridisation to complementary primers immobilised on a flow cell. A complementary strand is synthesised, the double stranded DNA fragment is denatured and the original template is washed away, leaving the fragment to be sequenced immobilised on the flow cell by its adapter. These fragments are clonally amplified using bridge amplification, where the immobilised DNA fragment folds over and hybridises with a complementary nearby primer and the complementary strand is synthesised, resulting in clusters of the same fragment. Sequencing then is carried out by first cleaving reverse strands and blocking the 3' ends to avoid unwanted priming before supplementing the reaction with forward primers and fluorescently-tagged nucleotides. After the addition of each nucleotide the flow cell is excited using a light source to measure the characteristic fluorescent signal and identify the base. As a cluster represents many copies of the same fragment, each cluster is read simultaneously to identify the consensus sequence and improve signal-tonoise. After sequencing the synthesised fragment is cleaved and washed away. For paired-end sequencing the reverse strand is sequenced after the forward strand by deprotecting the 3' end to allow the fragment to fold over and hybridise with a complementary primer on the flow cell for a single round of bridge amplification. Forward strands are then cleaved and washed away, the 5' ends are deprotected to avoid unwanted priming and the reverse strand is sequenced in the same way as the forward strand. Figure adapted from Strausberg et al. (2008).

or stability relationships, for simply evolving a protein to reduce its aggregation propensity or increase its stability, for example a biotherapeutic, it is not necessary nor feasible for an industrial lab to understand the functional consequence of each amino acid substitution in an antibody. Moreover, it has been shown that multiple mutations are often necessary to achieve large enough improvements in the property being enhanced, particularly when evolving affinity or stability, and while single mutations identified using DMS could be combined the collective effects are often not additive and very complex (Mateu et al., 1992; Daugherty et al., 2000; Marvin and Lowman, 2003; Lippow et al., 2007; Goldenzweig et al., 2016). Furthermore, often DMS experiments are limited by the length of the Illumina fragment size, as the longest fragment currently available is 250 bp meaning experiments are unable to sequence test proteins in one run that are longer than this. Illumina library preparation can include longer amplicons being fragmented into shorter reads that can be sequenced, and the short reads can then be mapped back onto a reference, a method generally used for sequencing genomes. However this 'shotgun library' approach is not currently supported by available packages for visualising and analysing DMS data such as Enrich2 (Rubin et al., 2017). An approach for the directed evolution of biotherapeurics that includes aspects from DMS would involve exploiting the power of NGS, but assessing multiple point mutations using a genotype-phenotype screen to gain vast improvements in various drug-like properties, such as stability, aggregation resistance or affinity. The high-throughput nature of NGS would allow massively parallel analysis of a wide sequencespace, and assessment of libraries before and after selection would allow identification of sequences that have been enriched as a result of the selection. Utilising a suitable screen which allows simultaneous selection of stability and function could enable isolation of stable, high-affinity therapeutics against any target in a high-throughput manner.

With the reducing cost of next-generation sequencing technologies such as Illumina (Figure 1.20), deep sequencing techniques are becoming more accessible to researchers and an attractive alternative to low-throughput, high-cost first-generation sequencing. While the Illumina method uses short read sequencing and therefore is unable to assess multiple mutations within proteins larger than the read length, it represents a cheap, quick and easy method for assessing the success of the experiment and for deciding whether to move forward with the more costly Pacbio sequencing (Figure 1.21), that provides the long-reads necessary for assessing co-evolution of larger proteins (Rhoads and Au, 2015).

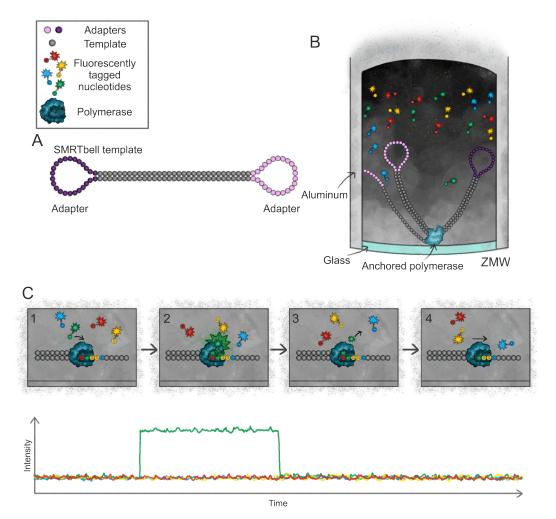


Fig. 1.21 Overview of Pacific biosciences (Pacbio) single-molecule real-time (SMRT) sequencing technologies. A) A template is prepared for Pacbio sequencing by the addition of hairpin adapters to create the 'SMRTbell', a closed, circular piece of DNA. B) Sequencing is carried out in a SMRT cell, a chip containing hundreds of thousands of sequencing units called zero-mode waveguides (ZMW). At the bottom of each ZMW is an immobilised  $\varphi$ 29 polymerase with a short bound DNA fragment complementary to the SMRTbell adapters. A single SMRTbell template diffuses into a ZMW and the adapter hybridises to the corresponding fragment bound on the immobilised polymerase to allow sequencing by the addition of fluorescently-tagged nucleotides. C) (1) The immobilised polymerase translocates along the SMRTbell template adding the corresponding fluorescently-tagged nucleotide as it goes. (2) The incorporation of the nucleotide results in an increase of the particular fluorescence signal linked to that nucleotide, which is used to determine the base at that position. (3) The phosphodiester bond linking the nucleotide and the fluorescence signal is broken, allowing the dye to diffuse out of the ZMW. (4) The polymerase translocates to the next position and the sequencing cycle continues by the addition of the next fluorescently-tagged nucleotide. Figure adapted from Eid et al. (2009).

# **1.7** Tripartite $\beta$ -Lactamase Assay (TPBLA)

Within this thesis we make use of another protein reporter biosensor, the Tripartite  $\beta$ -Lactamase Assay (TPBLA). This section examines in more detail the structure and function of the  $\beta$ -lactamase enzyme, before giving an overview of the TPBLA and its previous applications.

#### **1.7.1** Peptidoglycan biosynthesis inhibition by $\beta$ -lactam antibiotics

β-lactam antibiotics are, to date, the most prescribed antibiotic drugs in the world, mainly due to their low toxicity and broad-spectrum activity against both Gram-negative and Gram-positive bacteria (Bozcal and Dagdeviren, 2017; Kaderabkova et al., 2022). They contain a four-membered β-lactam ring which is key for their function as a competitive inhibitor of the penicillin binding proteins (PBPs) which are involved in the final step of cross-linking disaccharide components of peptidoglycan during cell wall biosynthesis (Tooke et al., 2019). Bacteria surround their cell membrane with a net-like peptidoglycan layer, which is comprised of repeating units of *N*-acetylglucosamine (GlcNAc) and *N*-acetylmuramic acid (MurNAc), which are added to a growing glycan chain by transglycosylation (Figure 1.22A). Glycan chains are cross-linked by a peptide chain of 5 amino acids by penicillin binding proteins in a process known as transpeptidation (Egan et al., 2020). β-lactam antibiotics bind to PBPs as competitive inhibitors as they are structurally similar to the cross-linked C-terminal dipeptide D-Ala-D-Ala (Figure 1.22B, C) (Tipper and Strominger, 1965; Tooke et al., 2019). β-lactamase hydrolyses the amide bond in the β-lactam ring of the antibiotic, thereby providing bacterial resistance (Kaderabkova et al., 2022).

## **1.7.2** $\beta$ -lactamase enzyme

The most widespread route of resistance to  $\beta$ -lactam antibiotics is via the hydrolysis of their central  $\beta$ -lactam ring by a group of enzymes known as  $\beta$ -lactamases. More than 6500 unique enzymes capable of degrading  $\beta$ -lactam antibiotics and thus providing antibiotic resistance have been identified to date (Furniss et al., 2022). The most well-studied and highly encountered of these is TEM-1  $\beta$ -lactamase, named "TEM" after the patient it was isolated from (Temoniera), was first isolated from penicillin-resistant bacteria in Athens 1963 (Datta and Kontomichalou, 1965; Turner, 2005; Salverda et al., 2010).

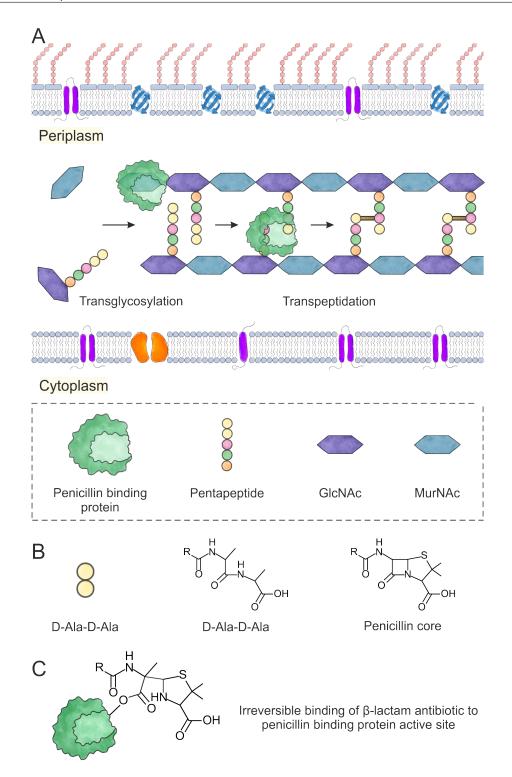


Fig. 1.22 Biosynthesis of peptidoglycan and its inhibition by  $\beta$ -lactam antibiotics. A) Peptidoglycan is comprised of repeating units of *N*-acetylglucosamine (GlcNAc) and *N*-acetylmuramic acid (MurNAc), which are added to a growing glycan chain by transglycosylation. Glycan chains are cross-linked by a peptide chain of 5 amino acids (pentapeptide) by penicillin binding proteins in a process known as transpeptidation. (B) The C-terminus of the peptide cross-link is D-Ala-D-Ala, which bears structural similarity to  $\beta$ -lactam antibiotics. (C) This structural similarity enables  $\beta$ -lactam antibiotics to competitively inhibit penicillin binding proteins, hindering the maturation of the bacterial cell wall leading to loss of cellular integrity and cell death (Kaderabkova et al., 2022). Figure redrawn from Saunders (2014).

TEM-1  $\beta$ -lactamase is a 29 kDa, monomeric protein comprised of two globular domains; an  $\alpha/\beta$  domain (three  $\alpha$  helices and five  $\beta$  sheets), and an  $\alpha$  domain (eight  $\alpha$  helices) (Figure 1.23) (Fonzé et al., 1995). The  $\alpha/\beta$  domain is formed by the 36 N-terminal and the 76 C-terminal residues of the protein chain (residues 26-62 and 215-290), whereas the  $\alpha$ domain is formed by residues 63-214 (Vandenameele et al., 2010). The catalytic site is located at the cleft between these two domains and contains the active serine (Ser70), which serves as the nucleophile for attack on the  $\beta$ -lactam carbonyl of the amide bond (Palzkill, 2018; Tooke et al., 2019). The isomerization of the Glu166-Pro167 bond from a *trans* to a *cis* conformation has been shown to be the rate limiting step for enzyme folding (Vandenameele et al., 2010). The position of Glu166 is key for the function of  $\beta$ -lactamase, as along with Ser70 and Lys73 it forms the catalytic site (He et al., 2020).

#### **1.7.3** β-lactamase as a reporter protein

In a similar way to the protein reporters described in Section 1.6.2.2,  $\beta$ -lactamase has been exploited as a protein reporter biosensor making use of its enzymatic readout of antibiotic resistance, or with colourmetric assays using chromogenic substrates. Opposite the catalytic site between Gly196 and Leu198 was proposed as a site to dissect the protein in half, so that each domain would fold but be inactive on its own (Figure 1.23B) (Galarneau et al., 2002).  $\beta$ -lactamase can be split at this site to form two fragments, these can be fused to two POIs and used to assess protein-protein interactions (Galarneau et al., 2002). If the two POIs interact, this will bring the two domains of  $\beta$ -lactamase into close proximity enabling them to form the active site via their non-covalent assembly. Antibiotic resistance is only generated if these two domains of  $\beta$ -lactamase are fused to interacting proteins; if they are expressed independently or fused to proteins which do not interact the resulting *E. coli* do not retain  $\beta$ -lactamase activity (Galarneau et al., 2002). This first example of using  $\beta$ -lactamase in a protein complementation assay was successfully used to assess protein-protein interactions of the homodimeric GCN4 leucine zipper, apoptotic proteins Bcl2 and Bad, and homodimeric Smad3 (Galarneau et al., 2002). This method used cephalosporin nitrocefin as a substrate for  $\beta$ -lactamase as it undergoes a colourmetric change following hydrolysis of its  $\beta$ -lactam ring from yellow (380 nm) to red (492 nm), which can be detected visually or using the change in absorbance.

Since the development of this protein complementation assay,  $\beta$ -lactamase has been widely exploited to assess protein-protein interactions *in vivo* (Wehrman et al., 2002; Spotts et al., 2002; Cavrois et al., 2002), as well as to screen libraries of scFvs (Secco et al., 2009). This method has also been used in more novel approaches including screening for

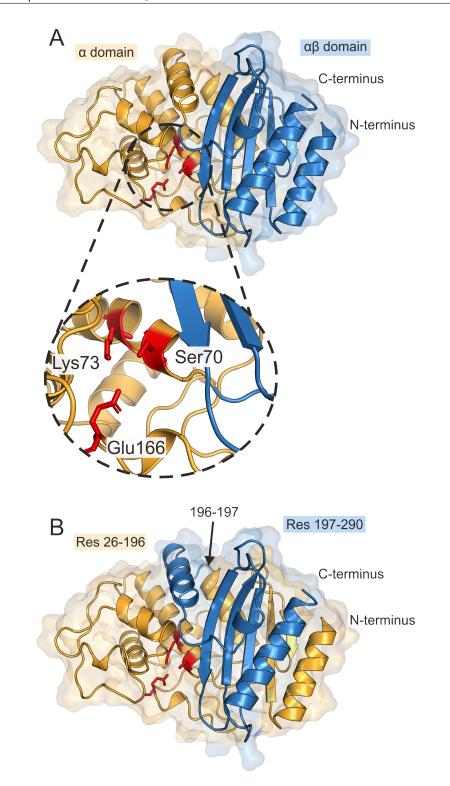


Fig. 1.23 **Structure of** *E. coli* **TEM-1**  $\beta$ -lactamase. A)  $\beta$ -lactamase is comprised of an  $\alpha\beta$  domain (blue, residues 26-62 and 215-290) and an  $\alpha$  domain (orange, residues 63-214). The catalytic site residues Ser70, Lys73 and Glu166 are highlighted in red. B) To use  $\beta$ -lactamase as a protein biosensor, the protein is split between residues 196 and 197 to allow insertion of a POI. PDB ID 1BTL (Jelsch et al., 1992). Figure created with PyMOL 2.5.2 (Schrödinger)

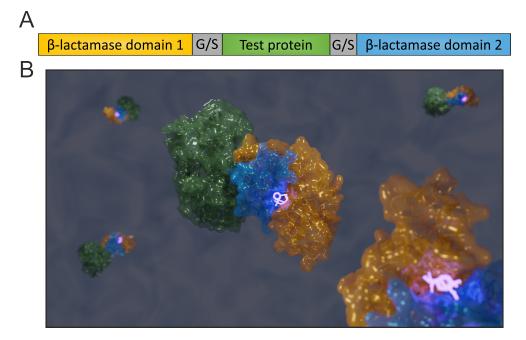


Fig. 1.24 *In vivo* tripartite  $\beta$ -lactamase assay (TPBLA). A) TPBLA construct. Test protein (green) is inserted between two domains of genetically separated TEM-1  $\beta$ -lactamase via a 28 residue glycine/serine linker inserted between residues 196/197. B) Correct folding of the test protein results in association of the two  $\beta$ -lactamase domains to form an active enzyme able to hydrolyse  $\beta$ -lactam antibiotics (purple). Figure created in PyMol 2.5.2 and Blender 3.4. Maltose Binding Protein (MBP) is shown as the test protein (PDB 1ANF) between TEM-1  $\beta$ -lactamase (PDB 1BTL) with bound ampicillin.

open reading frames (ORFs) by cloning a library between the signal sequence and mature sequence of  $\beta$ -lactamase so that only those containg ORFs can translate and express the  $\beta$ -lactamase, resulting in antibiotic resistance (D'Angelo et al., 2011).

### **1.7.4** Tripartite $\beta$ -Lactamase Assay (TPBLA)

A TPBLA was developed to assess protein folding and stability by inserting a POI between residues 196 and 197, joined via a 28 residue glycine/serine linker (Figure 1.24A, Figure 1.25A) (Foit et al., 2009). Previous studies have assessed test proteins with linkers of 33 and 68 residues in length, but the 28-residue linker was demonstrated to be the most broadly applicable to different sized test proteins (Foit et al., 2009; Ebo et al., 2020b; Saunders et al., 2016; Saunders, 2014). Correct folding of the test protein allows association of the two domains of TEM-1  $\beta$ -lactamase to form a functional enzyme that provides resistance to  $\beta$ -lactam antibiotics by hydrolysis of the  $\beta$ -lactam ring (Figure 1.24B, Figure 1.25B). However, misfolding, aggregation or instability of the test protein blocks association of TEM-1  $\beta$ -lactamase as the misfolded test proteins associate and form aggregates and/or

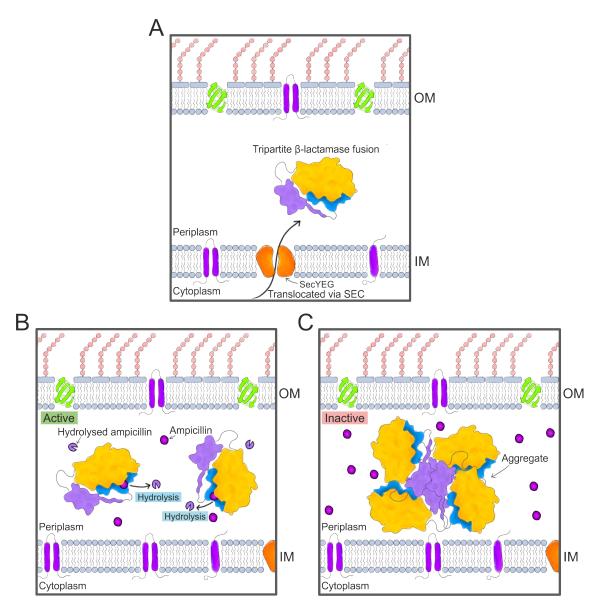


Fig. 1.25 Assessing aggregation using TPBLA. A) The fusion protein consisting of a test protein (purple) inserted between two domains of genetically separated TEM-1  $\beta$ -lactamase (yellow and blue) is translocated to the periplasm via SecYEG using the  $\beta$ -lactamase signal sequence. B) If the test protein folds properly and is not aggregation-prone,  $\beta$ -lactamase is able to hydrolyse ampicillin and *E. coli* resistance to the antibiotic. C) Misfolding, aggregation or instability of the test protein blocks association of  $\beta$ -lactamase, inhibiting formation of the catalytic site. Therefore,  $\beta$ -lactamase will be inactive and unable to provide antibiotic resistance.

the construct is degraded by cellular machinery, causing *E. coli* to lose its resistance to  $\beta$ -lactam antibiotics (Figure 1.25C). This methodology was able to correlate antibiotic resistance with thermodynamic stability by assessing variants of immunity protein 7 (Im7), granulocyte colony-stimulating factor (GCSF), maltose binding protein (MBP) and cytochrome b<sub>562</sub> (Foit et al., 2009). Furthermore, by introducing variance into the POI by epPCR and screening for improved antibiotic resistance, TPBLA was successfully used to evolve thermodynamic stability in Im7 (Foit et al., 2009). This evolution demonstrated the stability:function trade-off in Im7, as many of the stabilising mutations map predominantly to the surface used to bind its cognate toxin colicin E7. The TPBLA for measuring protein folding has been applied to assess the thermodynamic stability and proper folding of *de novo* designed proteins, and to evolve these proteins to improve their thermodynamic stability by more than 20 °C following six rounds of directed evolution (Xiong et al., 2014; Wang et al., 2017).

Throughout their lifetime, biopharmaceuticals are exposed to a multitude of physical, chemical and mechanical stresses that can result in aggregation from expression through processing to finally delivery to patients. These aggregates pose a significant risk to patients as they can invoke an immune response, causing side effects ranging from intolerance to adverse reactions and death (Jiskoot et al., 2012). Therefore, the main challenge for the biopharmaceutical industry is to identify aggregation-prone proteins early in development as well as the environmental factors that trigger aggregation in order to minimise unnecessary effort and expense. TPBLA has been adapted to assess proteins based on their aggregation propensity, including Aβ40 and Aβ42, wild type/D76N β2-microglobulin (\(\beta 2m)\), and human/rat islet amyloid polypeptide (hIAPP/rIAPP) (Saunders et al., 2016; Guthertz et al., 2022), as well as antibody fragments (single domains and scFv) relevant to the biopharmaceutical industry (Ebo et al., 2020b). Making use of the porosity of the E. coli outer membrane to small molecules (<600 Da), the assay has also been used as a screening method for identifying excipients (Hailu et al., 2013) and small molecules (Saunders et al., 2016) that inhibit protein aggregation. As this assay is carried out in the oxidising environment of the E. coli periplasm, it permits the proper formation of disulfide bonds, allowing analysis and evolution of proteins such as peptide hormones (e.g., hIAPP), immunoglobulin domains (i.e.,  $\beta$ 2m), or antibody fragments (such as scFvs). In contrast to the split  $\beta$ -lactamase protein complementation assay described in Section 1.7.3 analysing protein-protein interactions, TPBLA enables analysis of test proteins based on their biophysical properties and therefore has great potential for assessing the developability of candidate biotherapeutics.

As well as evolving thermodynamic stability into a POI, TPBLA has been successfully exploited to evolve aggregation-resistance in a model scFv as modification of the antibiotic concentration allows tight control over the level of selective pressure (Ebo et al., 2020b). Using deep sequencing to identify fitter variants that enable bacterial growth at increasingly high antibiotic concentrations, the TPBLA has the potential to unpick the complex relationship between sequence, thermodynamic stability, and aggregation for intrinsically disordered proteins, as well as globular POIs. However, as TPBLA selects for correct protein folding and/or aggregation-resistance it neglects to select for function. This could be problematic when evolving a protein for enhanced biophysical properties that needs to maintain a function, such as an antibody fragment or an enzyme, as it could result in stability: function trade-off costs. Indeed, functional residues were found most often to be mutated when evolving protein stability using TPBLA, consistent with the concept of protein frustration (stability:function trade-off) (Foit et al., 2009). This highlights the importance of choosing an appropriate selective assay for the system under investigation. Consequently, it can be necessary to develop orthogonal selection platforms with the ability to simultaneously evolve function and biophysical properties.

TPBLA represents a powerful tool for assessing innate aggregation propensity with the potential application as a developability assay for detecting aggregation-prone biopharmaceutical candidates early, without the need for purified protein. Furthermore, the use of TPBLA as a directed evolution screen has great potential for evolving stability and aggregation-resistance in a POI, importantly without requiring any prior structural knowledge of the protein or its mechanism of aggregation (Ebo et al., 2020a). However, previous work using this assay for directed evolution studies was limited by first-generation sequencing techniques, making the process laborious, low-throughput and high cost.

## **1.8** Aims of this thesis

Biopharmaceutical aggregation can occur at any stage in the developmental pipeline, and despite a range of developability screens available these are commonly employed late in the developmental pipeline as they require large amounts of purified protein. If aggregation-prone candidates could be identified early, this would significantly minimise unnecessary time and expense. TPBLA could be a powerful tool for identifying these aggregation-prone candidates quickly and easilly, without the need for purified protein.

With the reducing cost of NGS technologies such as Illumina, deep sequencing techniques are becoming more accessible to researchers and an attractive alternative to lowthroughput, high-cost first-generation sequencing. TPBLA has previously been used as a directed evolution screen, but was limited in throughput by first-generation sequencing technologies. Combining TPBLA with NGS for directed evolution has the potential to assess hundreds to thousands of variants in a single experiment and give a more comprehensive and extensive overview of a protein's fitness landscape.

The golden rule of directed evolution is 'you get what you screen for' (You and Arnold, 1996), and using the appropriate screen is paramount. This can be difficult as selecting for one property, such as stability, can have a negative effect on another, such as function. Often affinity-matured antibodies have decreased stability or increased aggregation, due to this trade-off between different properties (Rabia et al., 2018). In the absence of a selection for function, evolving test proteins using TPBLA to improve their aggregation resistance could result in evolved variants that no longer bind to their target. Therefore, including a selection for binding into TPBLA evolution experiments could enable evolution of biologics for both stability and function.

Therefore, the overall objectives of this thesis are to:

- Design a robust methodology to create large error-prone PCR libraries for TPBLA
- Adapt TPBLA into a high-throuhgput directed evoution assay using NGS
- Apply this to therapeutically-relevant proteins to improve their developability
- Investigate the use of TPBLA as a developability screen
- Correlate TPBLA with other developability assays and biophysical properties to better understand how it assesses and evolves target proteins
- Develop a novel assay to enable dual selection of stability and binding with TPBLA

# Chapter 2

# **Materials and methods**

# 2.1 Materials

## 2.1.1 Chemicals and kits

Deionised 18 M $\Omega$  water used in all methods.

Α	Supplier	Catalogue Number
Acetic acid, glacial	Fisher Scientific, Loughborough, UK	A/0400/PB17
Acrylamide 30 % ( <i>w/v</i> ): <i>bis</i> -acrylamide 0.8 % ( <i>w/v</i> )	Severn Biotech, Kidderminster, UK	20-2100-10
AffiniPure goat anti-human IgG Fc <sub>Y</sub> Fragment specific	Jackson ImmunoResearch, PA, USA	109-005-008
Agar	Melford Laboratories, Suffolk, UK	A20250-500.0
	Fisher Scientific, Loughborough, UK	BP-1423-500
Agarose	Melford Laboratories, Suffolk, UK	MB1200
L-(+)-Arabinose	Sigma Life Sciences, MO, USA	A3256
Ammonium persulfate (APS)	Sigma Life Sciences, MO, USA	A7460
Ampicillin sodium salt	Formedium, Norfolk, UK	AMP25
		CSB-
Anti-β-Lactamase IgG	Cusabio, TX, USA	PA352353YA01ENL
Anti-mouse IgG horseradish peroxidase conjugate	Cell Signaling Technology	7076S
Anti-rabbit goat IgG horseradish peroxidase conjugate	New England Biolabs, MA, USA	7074

## $Table \ 2.1 \ \textbf{Materials}$

В		
Bromophenol blue	Sigma Life Sciences, MO, USA	B0126
С		
Caffeine	Sigma Life Sciences, MO, USA	W222402
Carbenicillin disodium	Formedium, Norfolk, UK	CAR0025
ChromePure Goat IgG, whole molecule	Jackson ImmunoResearch, PA, USA	005-000-003
Citrate-stabilized 20nm gold nanoparticles	Expedeon, UK	741965-25ML
D		
Deoxynucleotide (dNTP) Solution Mix	New England Biolabs, MA, USA	N0447S
Dithiothreitol (DTT)	Formedium, Norfolk, UK	DTT025
Dimethyl sulfoxide (DMSO)	Sigma Life Sciences, MO, USA	P841
	Fisher Scientific (Invitrogen), Loughborough, UK	D12345
DNA ladders	New England Biolabs, MA, USA	N0552G
	Promega, WI, USA	G5711
Ε		
Ethanol	Sigma Life Sciences, MO, USA	E/0650DF/17
Ethidium bromide (EtBr)	Sigma Life Sciences, MO, USA	E-8751
Ethylenediaminetetraacetic acid (EDTA)	Acros Organics, Geel, Belgium	409930010
Mini, EDTA-free protease inhibitor cocktail	Roche Applied Science	11836170001
tablets	Koene Applied Science	11850170001
G		
Gel loading dye, purple (6x)	New England Biolabs, MA, USA	B7024S
Glycerol	Fisher Scientific, Loughborough,	G/0650/17
	UK Fisher Scientific, Loughhorough	
Glycine	Fisher Scientific, Loughborough, UK	G/0800/60
н	UK	
	Fisher Scientific, Loughborough,	
Hydrochloric acid (HCl)	UK	H/1100/PB17
Ι	-	
Imidizole	Sigma Life Sciences, MO, USA	I202
Instant Blue Coomassie Blue Stain	Expedeon, CA, USA	ISB1LUK
Isopropanol	Honeywell Research Chemicals, Seelze, Germany	190764
Isopropyl β-D-1-thiogalactopyranoside (IPTG)	Formedium, Norfolk, UK	IPTG100
K	i officulum, Noriolk, OK	n 10100
Kanamycin	Formedium, Norfolk, UK	KAN0025
L		11 11 10020
α-Lactose	Sigma Life Sciences, MO, USA	L8783
	Fisher Scientific, Loughborough,	
LB Broth	UK	1289-1650
M		

Magnesium sulphate	Fisher Scientific, Loughborough, UK	7487-88-9
Methanol	Fisher Scientific, Loughborough, UK	M/4000/17
Molecular weight marker (Precision Plus Dual Xtra Standards)	Bio-Rad Laboratories, CA, USA	161-0377
MOPS	Sigma Life Sciences, MO, USA	69947
N		
NEB Golden Gate Assembly Kit	New England Biolabs, MA, USA GE Healthcare, Buckinghamshire,	E1601L
Nickel sepharose	UK	17531802
Nickel(II) sulfate heptahydrate	Fluorochem, Hadfield, UK	510236
Nickel nitrilotriacetic acid (Ni-NTA)	QIAGEN, Crawley, UK	30210
NucleoSpin Gel and PCR Clean-up Kit	Macherey-Nagel, Düren, Germany	740609.50
Р		
Phosphate buffered saline (PBS) tablets	Fisher Scientific, Loughborough, UK	BR0014G
Potassium chloride (KCl)	Fisher Scientific, Loughborough, UK	P/4200/60
Potassium hydroxide (KOH)	Fisher Scientific, Loughborough, UK	P/5600/53
Precision Plus Protein Dual Xtra Prestained Protein Standard	Bio-Rad, CA, USA	1610377
Q		
Q5 Site-directed mutagenesis kit	New England Biolabs, MA, USA	E0554
QIAquick PCR Purification Kit	QIAGEN, Crawley, UK	28106
QIAquick Spin Miniprep Kit	QIAGEN, Crawley, UK	27106X4
Qubit dsDNA Assay Kit	Invitrogen, Carlsbad, California, USA	Q32851Q32851
S		
SnakeSkin dialysis tubing, 3.5K MWCO	Fisher Scientific, Loughborough, UK	68035
Sodium chloride (NaCl)	Fisher Scientific, Loughborough, UK	S/3160/60
Sodium dodecyl sulphate (SDS)	Fisher Scientific, Loughborough, UK	S/P530/53
	Severn Biotech, Kidderminster, UK	20-4000-01
	Sigma Life Sciences, MO, USA	L4509
Sodium hydroxide (NaOH)	Fisher Scientific, Loughborough, UK	S/4920/60
Syringe filter (nylon) (0.22 µm)	Camlab Ltd., Cambridge, UK	1181466
Syringe filter (PES) (0.22 $\mu m$ & 0.45 $\mu m)$	Jet Biofil, Guangzhou, China	FPE-204-025, FPE-404-025
Sucrose	Fisher Scientific, Loughborough, UK	S/8600/53

Suman Ontimal Catabalita (SOC)	New England Dieleha MA USA	D00205
Super Optimal Catabolite (SOC) SuperSignal <sup>™</sup> West Pico PLUS	New England Biolabs, MA, USA	B90205
Chemiluminescent Substrate	Thermo Scientific, MA, USA	34580
SYBR safe DNA gel stain	Invitrogen, Carlsbad, California, USA	\$33102
Т		
Tetracycline Triton X-100	Sigma Life Sciences, MO, USA Sigma Life Sciences, MO, USA Calbiochem, CA, USA	87128 X100-500 648463-50
Tris	Fisher Scientific, Loughborough, UK	BP152-1
Tetramethylethylenediamine (TEMED)	Sigma Life Sciences, MO, USA	T9281
Tris-tricine SDS running buffer 10X, cathode buffer, pH 8.3	Alfa Aesar, Heysham, UK	J60992
Tryptone	Fisher Scientific, Loughborough, UK	1285-1660
U		
Urea	MP Biomedicals, Loughborough , UK	04821527
	Fisher Scientific, Loughborough, UK	29700
V		
Vivaspin 20 centrifugal concentrators (10K MWCO)	Sartorius, Göttingen, Germany	S2002
W		
Wizard Plus SV Minipreps DNA purification systems	Promega, WI, USA	A1460
Y		
Yeast Extract	Melford Laboratories, Suffolk, UK	Y20025-2000.0
Z		
ZebaSpin Desalting Column 7K MWCO (0.5 ml)	Fisher Scientific, Loughborough, UK	89883

# 2.1.2 Enzymes for molecular biology

Ezyme	Supplier	Catalogue Number
Antarctic phosphatase	New England Biolabs, Hitchin, UK	A/0400/PB17
BamHI-HF restriction endonuclease	New England Biolabs, Hitchin, UK	20-2100-10
Q5 High-Fidelity DNA polymerase	New England Biolabs, Hitchin, UK	A20250-500.0
T4 Quick ligase	New England Biolabs, Hitchin, UK	MB1200
Vent DNA polymerase	New England Biolabs, Hitchin, UK	A3256
XhoI restriction endonuclease	New England Biolabs, Hitchin, UK	A7460

## Table 2.2 Enzymes for molecular biology.

# 2.1.3 Media

Medium	Components	
	10 g Tryptone	
	5 g Yeast extract	
Lysogeny broth (LB) medium	10 g NaCl	
	Up to 1L in purite 18 M $\Omega$ H <sub>2</sub> O	
	Autoclave 20 min at 121°C, 15 psi	
	10 g Bactotryptone	
	5 g Yeast Extract	
	Up to 500 ml in purite 18 M $\Omega$ H <sub>2</sub> O	
Autoinduction (AI) medium	Autoclave 20 min at 121°C, 15 psi	
	1 ml 1M MgSO <sub>4</sub>	
	$25 \text{ ml } 20 \times \text{NPSC}$	
	10 ml 50× LAC	
	125 g Glycerol	
	12.5 g Glucose	
50× LAC	50 g Lactose	
	Up to 500 ml in purite 18 M $\Omega$ H <sub>2</sub> O	
	Filter sterilise using 0.22 µM filter	
	53.52 g NH <sub>4</sub> Cl	
	32.2 g Na <sub>2</sub> SO <sub>4</sub>	
20 VIDSC	68 g KH <sub>2</sub> PO <sub>4</sub>	
20× NPSC	$70 \text{ g Na}_2 \text{HPO}_4$	
	Up to 1L in purite 18 M $\Omega$ H <sub>2</sub> O	
	Autoclave 20 min at 121°C, 15 psi	

## Table 2.3 Media used in this study.

# 2.1.4 Buffers

Buffer	Components	
Lysis buffer	20 mM Tris.HCl 300 mM NaCl 5 mM Imidazole 0.5% Triton X-100 1 mM PMSF (Stock = 200 mM) 2 mM Benzamidine pH 8	
Wash buffer	20 mM Tris.HCl 300 mM NaCl 10 mM Imidazole pH 8	
Elution buffer	20 mM Tris.HCl 300 mM NaCl 250 mM Imidazole pH 8	
Equilibration buffer	50 mM Tris.HCl 300 mM NaCl pH 8.0	
Phosphate-buffered saline (PBS)	137 mM NaCl 2.7 mM KCl 10 mM Na <sub>2</sub> HPO <sub>4</sub> 1.8 mM KH <sub>2</sub> PO <sub>4</sub> pH 7.4	
Electrophoresis cathode buffer	100 mM Tris.HCl 100 mM tricine 0.1 % (w/v) SDS pH 8.25	
Electrophoresis anode buffer	200 mM Tris.HCl pH 8.9	
2× SDS-PAGE loading dye	50 mM Tris.HCl 100 mM DTT 2 % (w/v) SDS 0.1 % (w/v) bromophenol blue 10 % (v/v) glycerol pH 6.8	

#### $Table \ 2.4 \ \text{Buffers used in this study.}$

Buffer	Components
1× Tris-acetate-EDTA (TAE)	40 mM Tris.HCl 20 mM acetic acid 1 mM EDTA pH 8.0
1× Lithium-acetate-borate (LAB)	10 mM lithium acetate 10 mM boric acid

## 2.1.5 Antibiotics

Antibiotic	Solvent	Stock solution (mg/mL)	Working concentration (µg/mL)	Sterilisation
Ampicillin	Purite 18 MΩ H2O	100	100	Filter sterilised
Kanamycin	Purite 18 MΩ H2O	50	50	
Tetracycline	100 % (w/v) ethanol	3	10	through 0.22 μm
	100 % (w/v) ethanol	25	25	filter

Table 2.5 A	Antibiotics used	in thi	s study.
-------------	------------------	--------	----------

# 2.2 Molecular biology methods

### 2.2.1 Bacterial strains

*E. coli* DH5 $\alpha$  derivative strain NEB5 $\alpha$  (New England Biolabs, Cat#: C2987H) F<sup>-</sup> fhuA2  $\Delta$ (argF-lacZ)U169 phoA glnV44  $\Phi$ 80  $\Delta$ (lacZ)M15 gyrA96 recA1 relA1 endA1 thi-1 hsdR17 *E. coli* NEB10 $\beta$  (New England Biolabs, Cat#: C3019H)  $\Delta$ (ara-leu) 7697 araD139 fhuA  $\Delta$ lacX74 galK16 galE15 e14-  $\Phi$ 80dlacZ $\Delta$ M15 recA1 relA1 endA1 nupG rpsL (Str<sup>R</sup>) rph spoT1  $\Delta$ (mrr-hsdRMS-mcrBC) *E. coli* SCS1 (Agilent, Cat# 200231) recA1 endA1 gyrA96 thi-1 hsdR17 (r<sub>K</sub>- m<sub>K</sub>+) supE44 relA1 *E. coli* TG1 (Lucigen, Cat# 60502-2) supE thi-1  $\Delta$ (lac-proAB)  $\Delta$ (mcrB-hsdSM)5 (r<sub>K</sub>- m<sub>K</sub>-) [F' traD36 proAB lacIq Z $\Delta$ M15]

#### 2.2.2 E. coli transformation

Chemically competent *E. coli* cells (10 µl SCS1, 50 µl DH5 $\alpha$  or NEB10 $\beta$ ) were thawed on ice for 10 minutes. Cells were transformed with 50-200 ng DNA and incubated on ice for 30 minutes. Cells were heat shocked at 42°C for 30 seconds (DH5 $\alpha$ /NEB5 $\alpha$ , NEB10 $\beta$ ) or 45 seconds (SCS1) then returned to ice for 5 minutes before the addition of 950 µL (DH5 $\alpha$ ) or 100 µL (SCS1) of SOC medium or 950 µL NEB10 $\beta$  Stable Outgrowth Medium (NEB B9035S, NEB10 $\beta$ ). These were cultured at 37°C, 200 rpm for 1 hr. Cultures of laboratory generated competent cells (DH5 $\alpha$ ) were centrifuged (3000xg, 3 min) and the resulting pellet was resuspended in 100 µL supernatant before plating on LB agar containing the appropriate antibiotic (Table 2.5) and grown overnight at 37°C. 100 µL commercial cells (NEB5 $\alpha$ , NEB10 $\beta$ , SCS1) were plated on LB agar containing the appropriate antibiotic and grown overnight at 37°C.

#### 2.2.3 Polymerase chain reaction

Polymerase chain reaction (PCR) was carried out for selective amplification of DNA sequences. All enzymes and buffers used were from New England Biolabs (NEB). The components for a typical 50  $\mu$ L reaction using Vent polymerase are detailed in Table 2.6 and the thermocycling conditions used detailed in Table 2.7. Components for a typical 50  $\mu$ L reaction using Q5 High-Fidelity polymerase are detailed in Table 2.8 and the thermocycling conditions used detailed in Table 2.9. PCR reactions were set up on ice and thermocycling was carried out using a BioRad T100 thermal cycler. For each PCR reaction, a control without the template DNA was carried out to ensure specificity of amplification. The New England Biolabs T<sub>a</sub> calculator tool was used for accurate calculation of annealing temperature as it incorporates information from both sequence and buffer composition.

To clone into the  $\beta$ -lactamase vector, the genes of test proteins were amplified using Vent polymerase using primers with 5' overhangs to add XhoI and BamHI restriction sites onto the 5' and 3' ends for ligation. Primers used are detailed in Appendix A, Table A.1. MBP variants were amplified from pMal-c5x, other genes were synthesised using Twist bioscience (HA4 and SH2) or UCB (scFvs).

 $5 \,\mu\text{L}$  of PCR product was visualised using agarose gel electrophoresis (Section 2.2.4), and the remaining PCR product was purified using QIAquick PCR Purification Kit (QIA-GEN, Crawley, UK) according to the manufacturer's instructions.

Component	Volume (µL)	Final concentration
10× Thermopol reaction buffer	5	1×
10 mM dNTPs	1	200 µM
10 µM Forward primer	1	0.2 μM
10 µM Reverse primer	1	0.2 μM
1-25 ng/µL Template DNA	1	1-25 ng
Vent DNA polymerase	0.5	1 unit
MgSO <sub>4</sub>	Optional	1-6 mM
Nuclease-free water	Up to 50	

Table 2.6 Components for a typical Vent PCR.

 $Table \ 2.7 \ \text{Thermocycling conditions for a typical Vent PCR.}$ 

Step		Temperature (°C)	Time
Initial denaturation	on	95	2-5 minutes
	Denaturation	95	15-30 seconds
25-30 cycles	Annealing	Tm of primer	15-30 seconds
	Elongation	72	1 minute/kb
Final extension		72	5 minutes
Hold		4	

Component	Volume (µL)	Final concentration
5× Q5 reaction buffer	10	1×
10 mM dNTPs	1	200 µM
10 µM Forward primer	2.5	0.5 μΜ
10 µM Reverse primer	2.5	0.5 µM
1-25 ng/µL Template DNA	1	1-25 ng
Q5 High-Fidelity DNA polymerase	0.5	1 unit
5×Q5 high GC enhancer (optional)	(10)	(1x)
Nuclease-free water	Up to 50	

Step		Temperature (°C)	Time
Initial denaturat	ion	98	30 seconds
	Denaturation	98	5-10 seconds
25-30 cycles	Annealing	Tm of primer	10-30 seconds
	Elongation	72	20-30 seconds/kb
Final extension		72	2 minutes
Hold		4	

Table 2.9 Thermocycling conditions for a typical Q5 PCR or site-directed mutagenesis.

## 2.2.4 Agarose gel electrophoresis

1.5 % (*w/v*) agarose gels were created by dissolving agarose in 1× Tris-acetate-EDTA (TAE) or Lithium-acetate-borate (LAB) buffer (Table 2.4) and heating. Once this had cooled < 50 °C, 0.001 % (*v/v*) SYBR safe was added and the solution mixed before pouring into a gel tray (12 × 15 cm) with a comb and allowed to set. Samples were diluted in 6× Purple gel loading dye and 5  $\mu$ L 100 bp and 1 kb Quick-load purple DNA ladders (NEB) were used to provide a size standard (Table 2.1). Agarose gel electrophoresis was carried out in 1× TAE or LAB buffer at 100V until fragments were resolved. Gels were visualised under ultraviolet (UV) light and imaged using UVItec Q9 Alliance Gel Doc.

## 2.2.5 Restriction digest of plasmid DNA

Restriction digests were carried out using enzymes and buffers from NEB. Components for a 50  $\mu$ L double digest are detailed in Table 2.10. Control reactions were set up containing either one or no enzymes. Digests were incubated at 37 °C for 1 hr. 5  $\mu$ L of digest was visualised using agarose gel electrophoresis (Section 2.2.4), and the remaining digest was purified using QIAquick PCR Purification Kit according to the manufacturer's instructions.

Component	Volume (µL)	Final amount
Plasmid DNA or PCR product	Variable	1 µg
20 U/µL XhoI restriction endonuclease	1	20 U
20 U/µL BamHI-HF restriction endonuclease	1	20 U
10× BSA	5	1×
10× Cutsmart buffer	5 μL	1×
Nuclease-free water	Up to 50	1 unit

 $Table \ 2.10$  Components for a double digest using XhoI and BamHI restriction endonucle-ases.

#### 2.2.6 Dephosphorylation of restriction digests

The 5'- ends of the restriction digested vector were dephosphorylated using Antarctic phosphatase to prevent re-ligation of the plasmid. Components for a typical dephosphorylation reaction are detailed in Table 2.11. Reactions were incubated at 37 °C for 15 minutes, followed by 65 °C for 5 minutes for enzyme inactivation. Following dephosphorylation, DNA was purified using QIAquick PCR Purification Kit according to the manufacturer's instructions and the concentrations of the restriction digested insert and restriction digested and dephosphorylated vector were quantified using NanoDrop 2000 UV-VIS spectrophotometer (Thermo Scientific).

 $Table \ 2.11 \ \mbox{Components for dephosphorylation of restriction digested vector using Antarctic phosphatase.}$ 

Component	Volume (µL)
PCR purified restriction digested vector	30
5 U/µL Antarctic phosphatase	1
10× Antarctic phosphatase reaction buffer	3.5
Nuclease-free water	Up to 50

### 2.2.7 Ligation of DNA

Ligation of DNA fragments was carried out using T4 quick ligase from NEB. Components for a typical ligation are detailed in Table 2.12. A control reaction was set up containing no insert. Reactions were set up on ice then incubated for 5 minutes at room temperature before immediate transformation into *E. coli* DH5 $\alpha$  cells (Section 2.2.2).

Component	Volume (µL)	Final amount
Digested and dephosphorylated vector DNA (6kb)	Variable	50 ng
Digested insert DNA (1kb)	Variable	25 ng
2× Quick ligase buffer	10	1×
T4 Quick ligase	1	
Nuclease-free water	Up to 20	

Table 2.12 Components for ligation of vector and insert using T4 quick ligase.

## 2.2.8 Q5 site-directed mutagenesis

#### 2.2.8.1 Amplification of DNA

Components for a typical 25 µL Q5 mutagenesis using Q5 Hot Start High Fidelity DNA polymerase are detailed in Table 2.13. Reactions were set up on ice and thermocycling was carried out using a BioRad T100 thermal cycler. Thermocycling conditions for a typical Q5 mutagenesis reaction are detailed in Table 2.9. Mutagenic primers used are detailed in Appendix A, Table A.1.

Table 2.13 Components	for a typical	Q5 mutagenesis.
-----------------------	---------------	-----------------

Component	Volume (µL)	Final amount
Q5 Hot Start High-Fidelity 2× Master Mix	12.5	1×
10 µM Forward primer	1.25	0.5 μM
10 µM Reverse primer	1.25	0.5 μΜ
1-25 ng/µL Template DNA	1.0	1-25 ng
Nuclease-free water	9.0	

#### 2.2.8.2 Kinase, ligase, Dpnl (KLD) treatment

Following PCR amplification using mutagenic primers, the PCR product was treated with kinase, ligase, and DpnI enzymes. Kinase phosphorylates the 5' end to allow intramolecular ligation by ligase, whereas DpnI digests methylated template DNA. Components for a typical KLD reaction are detailed in Table 2.14. Reactions were incubated for 5 minutes at room temperature before immediate transformation into *E. coli* DH5 $\alpha$  cells (Section 2.2.2).

Component	Volume (µL)	Final amount
PCR product	1	1 μL
2× KLD reaction buffer	5	1×
10× KLD enzyme mix	1	1×
Nuclease-free water	3	

Table 2.14 Components for kinase, ligase, Dpnl (KLD) treatment for Q5 mutagenesis.

#### 2.2.9 Sequencing and storage of plasmid DNA

10 mL 2.5 % (*w/v*) LB with the appropriate antibiotic was inoculated with single colonies from transformation plates and grown overnight at 37°C, 200 rpm. DNA was extracted using QIAquick Spin Miniprep Kit (QIAGEN, Crawley, UK) according to the manufacturer's instructions. DNA to be used in molecular biology experiments was eluted in the elution buffer provided in the kit. DNA for long term storage was eluted in TE buffer (Table 2.1) and stored at -80°C. DNA concentrations were measured using NanoDrop 2000 UV-VIS spectrophotometer by measuring absorbance at 260 nm (A<sub>260</sub>) using the relationship that an A<sub>260</sub> of 1.0 = 50 µg/mL pure dsDNA. 15 µL of DNA was sent with the required primers (Appendix A, Table A.3) for sequencing by Eurofins Genomics.

#### 2.2.10 Plasmids and primers

Plasmids used in this study are summarised in Table 2.15. The vector used for the TPBLA contained  $\beta$ -lactamase ( $\beta$ La) with a 28-residue glycine-serine (GS) linker in a pBR322 derivative encoding a pBAD promoter and was provided by Jim Bardwell, University of Michigan. *E. coli* MBP in a pMAL-c5x vector containing a tac promoter under control of the lac operon was obtained from New England Biolabs. All subsequent mutant MBP plasmids were derived from this and the genes cloned into the  $\beta$ La vector for analysis using TPBLA. Genes encoding HA4 and SH2 were synthesised by Twist bioscience using the sequences from Wang et al. (2018). Split sfCherry2 was cloned from a pETDuet\_sfCherry2(1-10)\_sfCherry2(11)-SpyCatcher gifted from Bo Huang (Addgene plasmid # 117656 ; http://n2t.net/addgene:117656 ; RRID:Addgene\_117656) (Feng et al., 2019). Split mNG2 was cloned from a pET\_mNG2(1-10)\_32aalinker\_mNG2(11) gifted from Bo Huang (Addgene plasmid # 82611 ; http://n2t.net/addgene:82611 ; RRID:Addgene\_82611) (Feng et al., 2017). mScarlet-I was cloned from a pEB2-mScarlet-I gifted from Philippe Cluzel (Addgene plasmid # 104007 ; http://n2t.net/addgene:104007 ; RRID:Addgene\_104007

(Balleza et al., 2018). pET28a-sfGFP was a gift from Ryan Mehl (Addgene plasmid # 85492 ; http://n2t.net/addgene:85492 ; RRID:Addgene\_85492) (Peeler and Mehl, 2012). pHJ12-CadC-VHH-Caffeine (No linker) was a gift from Jerome Bonnet (Addgene plasmid # 108244 ; http://n2t.net/addgene:108244 ; RRID:Addgene\_108244). Primers used in this study are detailed in Appendix A.

Plasmid name and vector	Description of sequence	Resistance marker
pBR322-βLa-28GS	$\beta$ -lactamase with a 28-residue GS linker under a pBAD promoter used for TPBLA.	Tetracycline
pMAL-c5x MBP <sup>WT</sup>	pMAL-c5x vector with <i>E. coli</i> wild type maltose binding protien (MBP) under a tac promoter under the control of the lac operon.	Ampicillin
pMAL-c5x MBP <sup>4A</sup>	As above but with four point mutations (L160A+I161A+L192A+L195A).	Ampicillin
pMAL-c5x MBP <sup>Y283D</sup>	As pMAL-c5x MBP but with single point mutation Y283D.	Ampicillin
βLa-MBP <sup>WT</sup>	β-lactamase with wild type MBP insert between 28-residue GS linker. For TPBLA.	Tetracycline
βLa-MBP <sup>4A</sup>	As above but with MBP <sup>4A</sup> insert.	Tetracycline
βLa-MBP <sup>Y283D</sup>	As above but with MBP <sup>Y283D</sup> insert.	Tetracycline
βLa-AMSXXX	As above but with AMSCI scFv insert where XXX is 106, 122, 132, 134, 137, 147, 148, 155, 197, 198, or 214.	Tetracycline
βLa-HA4 <sup>WT</sup>	As above but with monobody HA4 insert.	Tetracycline
βLa-HA4 <sup>Y87A</sup>	As above but with Y87A point mutation in HA4 gene to create a non-binding mutant.	Tetracycline
βLa-HA4 <sup>WT</sup> -VHH	$\beta$ -lactamase with wild type HA4 insert and caffeine inducible nanobody fused to the C-terminus. For SnAC.	Tetracycline
βLa-HA4 <sup>Y87A</sup> -VHH	As above but with Y87A point mutation in HA4 gene. For SnAC.	Tetracycline
pETDuet_sfCherry2(1- 10)_sfCherry2(11)- SpyCatcher	pET vector with split sfCherry2 used to clone into $\beta$ La vector.	Ampicillin
pET_mNG2(1-10) 32aalinker_mNG2(11)	pET vector with split mNeonGreen2 used to clone into $\beta$ La vector.	Kanamycin

 $Table\ 2.15$  List of plasmids used in this thesis and description of gene insert and mutations.

Plasmid name and vector	Description of sequence	Resistance marker
pEB2-mScarlet-I	pEB2 plasmid with mScarlet-I gene used to clone into $\beta$ La vector.	Kanamycin
pET28a-sfGFP	pET28a vector with sfGFP gene used to clone into $\beta$ La.	Kanamycin
pHJ12-CadC-VHH- Caffeine (No linker)	pHJ12 vector with CadC gene and VHH caffeine.	Kanamycin

# 2.3 Protein expression and purification

#### 2.3.1 Purification of MBP constructs

#### 2.3.1.1 Small-scale expression trials

Small scale expression trials were carried out to assess the amount of soluble protein expressed for each of the MBP constructs. *E. coli* BL21(DE3) cells were transformed with the relevant plasmid (Table 2.15) as described in Section 2.2.2. 100 mL autoinduction medium containing 100 µg/mL carbenicillin was inoculated with 250 µL overnight culture and incubated ( $37^{\circ}$ C, 220 rpm) for 30 h, taking two 1 mL samples at 2 h intervals between 24 and 30 hrs. The OD<sub>600</sub> was corrected to 0.5 and cells were harvested by centrifugation (13000 rpm, 10 min). 50 µL lysis buffer (Table 2.4) was added to the cell pellet and vortexed. The soluble and insoluble fractions were separated by centrifugation (13000 rpm, 10 min) before addition of 2× SDS-PAGE loading dye (Table 2.4) and samples were analysed by SDS-PAGE (Section 2.3.1.5).

#### 2.3.1.2 Large-scale expression of protein constructs

Expression plasmids were transformed into *E. coli* BL21 (DE3) cells by heat shock (Section 2.2.2). Successful transformants were selected on LB agar containing 100 µg/mL carbenicillin after growth overnight at 37°C. Single colonies were used to innoculate 100 mL LB containing 100 µg/mL carbenicillin and incubated overnight (37°C, 220 rpm). 2 mL starter culture was used to innoculate  $2 \times 500$  mL autoinduction medium (Table 2.3) prepared in 2 L conical flasks and incubated for 48 hours (37°C, 220 rpm). Cells were harvested by centrifugation at 8000 rpm at 4°C and the pellet resuspended in lysis buffer

(Table 2.4) before homogenesis, addition of DNase and incubation with roller agitation at 4°C for 1 hr. Lysate was passed through a cell disruptor at 30 kpsi, 25°C and centrifuged to separate out the soluble and insoluble fractions (15000 rpm, 4°C, 30 mins).

#### 2.3.1.3 Refolding of protein from inclusion bodies

MBP<sup>4A</sup> was expressed in the insoluble fraction so the inclusion body pellet was first washed three times in 20 mM Tris.HCl, 300 mM NaCl (pH 8) and centrifuged (16000 rpm, 4°C, 20 mins). The pellet was then dissolved in 100 mL 20 mM Tris.HCl, 300 mM NaCl (pH 8) 8M Urea and centrifuged (16000 rpm, 4°C, 30 mins). The supernatant was diluted 1:5 with 0.7 M arginine (pH 8) and dialysed overnight at 4°C into 20 mM Tris, 300 mM NaCl (pH 8) to remove all the urea. The refolded MBP<sup>4A</sup> was centrifuged and filtered before adding imidazole to give a final concentration of 5 mM and loaded onto the column in the same way as MBP<sup>WT</sup> and MBP<sup>Y283D</sup> (Section 2.3.1.4).

#### 2.3.1.4 HisTrap purification

Samples were filtered (0.45  $\mu$ m) prior to loading onto the column. MBP<sup>WT</sup> and MBP<sup>Y283D</sup> soluble fractions and refolded MBP<sup>4A</sup> were loaded peristaltically overnight at 2 mL/minute onto 20 mL Ni-NTA resin (HisTrap) pre-equilibrated with lysis buffer. The resin was washed with 5 column volumes wash buffer (Table 2.4). Protein was eluted with 5 column volumes elution buffer (Table 2.4) and the fractions containing protein were determined using SDS-PAGE (Section 2.3.1.5). These fractions were pooled, concentrated using VivaSpin 10 kDa molecular weight cut-off (MWCO) concentrator (GE Healthcare).

#### 2.3.1.5 Sodium dodecyl sulfate polyacrylamide gel electrophoresis

Tris-tricine buffered SDS-PAGE gels were made using components in Table 2.16 in 8× 10 cm casts using a 1.5 mm spacer. Samples were diluted in 2× SDS-PAGE loading dye (Table 2.4), boiled for 10 mins and centrifuged before loading 15  $\mu$ L into the well. 5  $\mu$ L protein standard was loaded into the first well of the gel for estimation of molecular weight (Table 2.1). The inner reservoir was filled with cathode buffer (100 mM Tris.HCl, 100 mM tricine, 0.1 % (*w/v*) SDS, pH 8.25) and the outer reservoir was filled with anode buffer (200 mM Tris.HCl, pH 8.9), both diluted from a 10× stock (Table 2.1). Gels were run at 35 mA until samples entered the resolving gel, when the current was increased to 65 mA

until the dye front reached the bottom of the gel. Gels were then stained for 15-60 mins using Coomassie Instant Blue Stain (Table 2.1), washed, and visualised using a white light transilluminator.

Solution component	Resolving gel (ml)	Stacking gel (mL)
30 % w/v acrylamide:0.8 % w/v bis-acrylamide	6.25	0.83
3 M Tris.HCl, 0.3 % (w/v) SDS, pH 8.45	5	1.55
H <sub>2</sub> O	1.64	3.72
Glycerol	2	0
10 % (w/v) ammonium persulfate	0.1	0.2
TEMED	0.01	0.01

Table 2.16 Recipe for two 12.5% Tris-tricine SDS-PAGE gels.

#### 2.3.1.6 **TEV protease treatment**

5 mg TEV protease kindly provided by Sophie Cussons, University of Leeds, was added and the protein isolated in Section 2.3.1.4 was dialysed into 50 mM Tris, 0.5 mM EDTA and 1 mM DTT (pH 8.0) overnight at 4°C. Digested protein was then loaded onto 5 mL Ni-NTA resin pre-equilibrated with 50 mM Tris, 300 mM NaCl (pH 8) and the flow through collected. Resin was washed with 1 column volume 50 mM Tris, 300 mM NaCl (pH 8) then 2 column volumes elution buffer. Presence of the digested protein in the elution was determined by SDS-PAGE (Section 2.3.1.5).

#### 2.3.1.7 Gel filtration chromatography

The purified TEV digested protein (Section 2.3.1.6) was concentrated using VivaSpin 10 kDa MWCO concentrator to a final volume of 10 mL and filtered (0.22  $\mu$ m) before loading onto a HiLoad 26/600 Superdex 75 gel filtration column (GE Life Sciences) pre-equilibrated with 20 mM Tris (pH 8.0). The protein was eluted from the column at a flow rate of 1 mL/min (Table 2.17). Protein elution was monitored by absorbance at 280 nm. Fractions corresponding to the monomeric protein were analysed by SDS-PAGE (Section 2.3.1.5) and dialysed into purite 18 M $\Omega$  H<sub>2</sub>O before lyophilising.

Breakpoint (mL)	Flow rate (mL/min)	Fraction size (mL)	Injection valve position	Auto zero
0	1	0	Load	No
10	1	0	Inject	Yes
20	1	0	Load	No
90	1	1.5	Load	No
320	1	0	Load	No

 $Table\ 2.17$  AKTA program for purification of MBP and derivatives by size exclusion chromatography.

#### 2.3.1.8 Mass spectrometry

The molecular mass of purified MBP proteins was measured using electrospray ionisation mass spectrometry (ESI-MS) carried out by Samantha Lawrence (University of Leeds). The molecular masses of purified IgG and Fab proteins was measured using ESI-MS carried out by Adam Long (UCB).

# 2.3.2 Purification of IgG and Fab proteins from chinese hamster ovary (CHO) cells

Expression and purification of IgG and Fab antibodies was undertaken at UCB (Slough). Variants were expressed in proprietary UCB CHO cell line (CHO SXE cells). Sequences were eukaryote codon-optimised and the heavy and light chains cloned into a proprietary UCB mammalian expression vector with a human serum albumin (HSA) signal sequence. The plasmids were co-transfected into CHO SXE cells for expression and grown for 7 days. The proteins were harvested by centrifugation  $(4,000 \times g, 30 \min, 4 \text{ °C})$  and filtered using a 0.22 µm filter. The supernatent was loaded onto Protein A (IgG) or Protein L (Fab) agarose (MabSelect SuRe; GE Healthcare Life Sciences) pre-equilibrated in PBS. Columns were washed in 6 column volumes PBS before eluting in 100 mM citric acid pH 3.5 (Protein A) or pH 2.5 (Protein L) in 2 mL fractions. Elutions were immediately neutralised using 250 µL 1.5 M Tris to give pH 6-7. Fractions were buffer exchanged into PBS and presence of the purified protein in the elution was determined by SDS-PAGE (Section 2.3.1.5) and Mass Spectrometry (Section 2.3.1.8).

### 2.4 In vitro techniques

# 2.4.1 Size exclusion chromatography multi angle light scattering (SEC-MALS)

The oligometric states of MBP<sup>WT</sup> and MBP<sup>4A</sup> were analysed by size exclusion chromatography (SEC) coupled to a multi-angle light scattering detector (MALS) (miniDAWN TREOS, Wyatt) using a TSKgel G3000 SWxL column (Tosoh Bioscience) equilibrated with PBS at room temperature. 50  $\mu$ L of sample at 51.9  $\mu$ M (MBP<sup>WT</sup>) or 51.7  $\mu$ M (MBP<sup>4A</sup>) was loaded onto the column at 0.75 mL/min. Samples were analysed using UV absorbance at 280 nm, with 3-angle static light scattering and refractive index (Wyatt Optilab T-rEX detector) also measured. Together they can be analysed to assess the absolute molecular weight and concentration. Data were collected and analysed using the software ASTRA version 6.1 (Wyatt), using the Debye model to fit the data.

#### 2.4.2 Circular dichroism (CD) spectroscopy

Far- and near- UV CD were used to assess the secondary and tertiary structures of MBP<sup>WT</sup> and variants. To measure secondary structure using far-UV CD, 0.2 mg/mL sample in 10 mM potassium phosphate (pH 7.4) was analysed in a 1 mm cuvette and measurements were taken from 250 to 180 nm at 25°C. To measure tertiary structure using near-UV CD, 0.6 mg/mL sample in 10 mM potassium phosphate (pH 7.4) was analysed in a 0.1 mm cuvette and measurements were taken from 350 to 250 nm at 25°C. For both far- and near-UV CD, empty cuvette and buffer samples were measured as blanks. To assess the thermal stabilities, the far-UV CD spectra of the proteins at 0.2 mg/mL in a 1 mm cuvette was measured from 20°C to 90°C in 1°C steps.

#### 2.4.3 Urea denaturation

All urea denaturation experiments were carried out in 10 mM potassium phosphate (pH 7.4) at 25°C. Solutions of buffer and concentrated urea in buffer were dispensed into Corning 96-well, half-area, black polystyrene plates (3881) using a Microlab ML510B dispenser in 0.2 M denaturant steps. The protein was then dispensed from a 10× concentrated stock to give a final well volume of 150  $\mu$ L and final protein concentration of 0.1  $\mu$ M. All plate measurements were carried out on a CLARIOstar Plate Reader (BMG Labtech) using an

excitation wavelength of 280 nm and collecting emission spectra between 335 and 345 nm. Plates were covered with a Corning 96-well microplate aluminium sealing tape to minimise evaporation (Perez-Riba and Itzhaki, 2017). Urea denaturation curves were fitted using IgorPro.

# 2.4.4 8-Anilinonapthalene-1-sulphonic acid (ANS) fluorescence spectroscopy

A 1 mM ANS stock was prepared in 1 mL M $\Omega$  H<sub>2</sub>O. The stock concentration was determined using  $\varepsilon_{350} = 4900 \text{ M}^{-1} \text{ cm}^{-1}$  (Azzi, 1974). Samples for fluorescence emission spectroscopy were prepared in 10 mM potassium phosphate (pH 7.4) with a final concentration of 1 µM protein and 100 µM ANS. Fluorescence experiments were carried out using an excitation wavelength of 380 nm and collecting emission spectra between 400 nm and 600 nm with 1 nm slit widths. Three spectra were recorded and averaged for each sample. A control was carried out with the dye in buffer.

#### 2.4.5 X-ray crystallography

A protein stock solution of MBP<sup>WT</sup> (18.4 mg/mL), MBP<sup>Y283D</sup> (20.1 mg/mL) and MBP<sup>4A</sup> (21.7 mg/mL) was prepared in 20 mM MES pH 6.2. Maltose was added to the MBP<sup>WT</sup> and MBP<sup>Y283D</sup> stocks at a molar ratio of 1:1. Crystals were grown by mixing 0.1  $\mu$ L (MBP<sup>4A</sup>) or 0.2  $\mu$ L (MBP<sup>WT</sup>, MBP<sup>Y283D</sup>) of the protein sample and 0.1  $\mu$ L (MBP<sup>4A</sup>) or 0.2  $\mu$ L (MBP<sup>WT</sup>, MBP<sup>Y283D</sup>) of the crystallization solution in sitting drop plates at 293 K. The crystallization solution for MBP<sup>WT</sup> (40% (*w/v*) PEG 300, 0.1 M Phosphate-Citrate pH 4.2) and MBP<sup>Y283D</sup> (39-48% (*w/v*) PEG 300, 0.1 M Phosphate-Citrate pH 4.2) and MBP<sup>Y283D</sup> (39-48% (*w/v*) PEG 300, 0.1 M Phosphate-Citrate pH 4.11) was prepared from stock solutions. The crystallization solution for MBP<sup>4A</sup> (50 mM HEPES, 50 mM MOPS, pH 7.5, 7.03% (*v/v*) MPD, 7.03% (*w/v*) PEG 1000, 7.03% (*w/v*) PEG 3350, and 27 mM each of sodium nitrate, sodium dibasic, and ammonium sulfate) was prepared from Morpheus Buffer System 2 at pH 7.5, Morpheus Precipitant Mix 4, and Morpheus NPS Mix (all from Molecular Dimensions).

After 2 weeks, crystals were fished and flash-frozen in liquid nitrogen. The diffraction data were collected on beamline I24 at Diamond Light Source (U.K.). The data were processed using the xia2 (Winter, 2010) bundle, with DIALS (Winter et al., 2018) for integration and using Pointless/Aimless (Evans, 2006; Evans and Murshudov, 2013) for scaling and merging. The data were processed using CC1/2 and completeness as

cutoff criteria (Karplus and Diederichs, 2012). The structures were solved by molecular replacement, using apo MBP<sup>WT</sup> (PDB 10MP (Sharff et al., 1992)) as the search model in PHASER (McCoy et al., 2007). COOT (Emsley et al., 2010) and REFMAC5 (Murshudov et al., 2011) were used for refinement. The quality of the final structure was assessed using MolProbity (Chen et al., 2010). Data collection and refinement statistics are shown in Table 3.1. Figures were prepared using PyMOL (version 2.7, Schrödinger).

#### 2.4.6 Hydrophobic interaction chromatography (HIC)

 $5 \ \mu g \ IgG \ samples (1 \ mg/mL)$  were spiked in with a mobile phase A solution (1.8 M ammonium sulfate and 0.1 M sodium phosphate at pH 6.5) to achieve a final ammonium sulfate concentration of about 1 M before analysis. A Sepax Proteomix HIC butyl-NP5 4.6x 35mm column was used with a linear gradient of mobile phase A and mobile phase B solution (0.1 M sodium phosphate, pH 6.5) over 26 min at a flow rate of 1 mL/min with UV absorbance monitoring at 280 nm.

# 2.4.7 Affinity-capture self-interaction nanoparticle spectroscopy (AC-SINS)

AC-SINS experiments were undertaken at UCB (Slough). AffiniPure goat anti-human IgG Fc $\gamma$  fragment specific (IgG $\alpha$ -Fc) antibodies were buffer exchanged into 20 mM potassium acetate, pH 4.3 and diluted to 0.4 mg/mL. 50 µL IgGα-Fc was added to 450 µL 20 nm gold nanoparticles (BBI solutions) and briefly vortexed before being incubated overnight at room temperature. 55.5  $\mu$ L 1  $\mu$ M thiolyated PEG 2000 was added and the particles incubated at room temperature for 1 hr. Nanoparticles were centrifuged (15,000 rpm, 6 min) and the supernatent carefully removed without disturbing the pellet. Nanoparticles were resuspended in 120 µL 20 mM potasisum acetate, pH 4.3 and used the same day. Polyclonal goat IgG (pol-IgG) was buffer exchanged into PBS and diluted to 222 µg/mL. 200 µL pol-IgG was combined with 20 µL mock supernatent and 180 µL 22 µg/mL test antibody before brief vortexing. 72  $\mu$ L of this mix was combined with 8  $\mu$ L nanoparticle solution in a 384-well polystyrene plate. Samples were incubated at room temperature for 2hrs before measuring the absorbance on a FLUOstar Omega Microplate Reader (BMG Labech) from 500 nm to 600 nm in 1 nm increments. The maximum absorbance ( $\lambda_{max}$ , plasmon wavelength) was determined and the relative shift compared to nanoparticles with the test antibody replaced with PBS was calculated.

#### **2.4.8** Differential scanning fluorimetry (DSF)

NanoDSF experiments were undertaken at UCB (Slough). 9  $\mu$ L of 1 mg/mL IgG was heated from 15-95 °C at a rate of 0.4 °C/minute while monitoring fluorescence (UNcle, Unchained Labs). Protein unfolding was measured using intrinsic protein fluorescence by exciting with a 266 nm laser and measuring emission from 315-430 nm. Static light scattering (SLS) was measured at each temperature to delineate unfolding and aggregation. Dynamic light scattering (DLS) reads were 4 acquisitions of 5 seconds each and were taken at the beginning and end of the thermal ramp. Tm and Tagg were determined by the UNcle Analysis software by using the first derivative (for Tm determination). The barycentric mean (BCM) of the fluorescence intensity curves from 315-430 nm was used to plot against temperature, which is defined by Equation 2.1.

$$\lambda_{\rm BCM} = \frac{\sum_{\lambda} \lambda I(\lambda)}{\sum_{\lambda} I(\lambda)}$$
(2.1)

Each wavelength value ( $\lambda$ ) between 315-430 nm is multiplied by the tryptophan fluorescence intensity (I) at that wavelength, and the sum of that value for all wavelengths between 315 to 430 nm is divided by the sum of the fluorescence intensities at those wavelengths. This results in an 'averaged' peak wavelength ( $\lambda$ BCM) for a given spectrum which eliminates noise and accommodates for changes in the shape of the spectrum.

The transition mid-point temperatures (Tm) were calculated using the first derivative of the fluorescence raw data. This was fitted to one or more gaussians using Origin Pro 2020 version 9.7.0.118. First derivative data displayed in Appendix B.

The temperature onset of aggregation (Tonset) was calculated from the first derivative of the static light scattering raw data. This was fitted to one or more gaussians using Origin Pro 2020 version 9.7.0.118. The fit from the first gaussian was normalised between 0 and 1, and the Tonset was defined as the point at which the slope (first derivative) exceeded 0.1 % of the peak value of the first derivative. In other words, a threshold value (0.1 %) was assigned to measure the point at which the static light scattering began to increase above the baseline. First derivative data displayed in Appendix B.

#### 2.4.9 Dot blot analysis

A single colony of fresh E. coli SCS1 cells (transformed with the appropriate plasmid) was used to inoculate 100 mL sterile LB containing 10 µg/mL tetracycline. Cultures were incubated overnight at 37°C with shaking (200 rpm). 1 mL of overnight culture was used to inoculate 100 mL sterile LB containing 10 µg/mL tetracycline and grown at 37°C (shaking at 200 rpm) until an OD<sub>600</sub> of 0.6 was obtained. 10 mL of culture was removed for the uninduced sample and centrifuged at 4,000 g for 10 min (4°C). Expression of  $\beta$ -lactamase MBP fusion constructs were induced by the addition of filter-sterilized arabinose to a final concentration of 0.075 % (w/v) arabinose. Cultures were incubated for 1 h (37°C, 200 rpm) and 10 mL was removed from each (representing the induced sample). The 10 mL cultures were harvested by centrifugation at 4,000 g for 10 min (4°C). The cell pellets (uninduced and induced with arabinose) were resuspended in phosphate buffered saline (PBS, Dulbecco's PBS, Sigma) to obtain an  $OD_{600}$  of 5. For whole cell samples, 300  $\mu$ L of the  $OD_{600} = 5$  sample was combined with 60 µL 6X loading buffer (150 mM Tris pH 6.8, 300 mM DTT, 6% (w/v) SDS). For soluble samples, 400  $\mu$ L of the OD<sub>600</sub> sample was centrifuged (16,000 g for 10 min) and the pellet resuspended in 400 µL bacterial protein extraction reagent (B-PER, ThermoFisher) and incubated with agitation for 10 min. The sample was then centrifuged at 16,000 g for 10 min and the supernatant was carefully pipetted off. 300 µL of soluble sample was combined with 6× loading buffer. The remaining insoluble pellet was resuspended in  $1 \times 1$  loading buffer and the whole cell, soluble and insoluble samples were boiled for 10 min. Samples were then centrifuged at 16,000 g for 3 min.

1 μL of protein samples were applied to nitrocellulose membrane and left to dry. Blots were blocked using 5 % (w/v) milk powder in TBST (Tris-buffered saline Tween; 20 mM Tris.HCl, 150 mM NaCl, 0.2 % (v/v) Tween-20). Membranes were incubated overnight with the anti-β-lactamase antibody (CSB-PA352353YA01ENL, Cusabio) diluted 1:10,000 in 5 % (w/v) milk powder in TBST. The membranes were washed for 3× 10 min in TBST. Membranes were then incubated with goat anti-rabbit IgG horseradish peroxidase conjugate (7074, New England BioLabs) diluted 1:10,000 in TBST. Membranes were then in TBST before incubation with SuperSignalTM western pico chemiluminescent substrate (Thermo Fisher Scientific). The emitted signal was visualised and imaged using UVItec Q9 Alliance Gel Doc.

# **2.5** Tripartite $\beta$ -lactamase assay (TPBLA)

#### 2.5.1 Preparation of 48-well agar plates

LB agar was autoclaved for 20 minutes at 121°C, 15 psi and left to cool to < 50°C. Tetracycline and arabinose were added to give 10 µg/mL and 0.075 % (w/v) final concentrations, respectively. Using a multichannel pipette in a sterile environment, 300 µL of medium was added to the first column of wells in the 48-well plate. The required amount of ampicillin was added to the agar and mixed before 300 µL was added to each well of the next row (Table 2.18). This was repeated so that ampicillin concentration increases in predetermined increments. Plates were left to set in a sterile environment.

Table 2.18 Components required to make four 48-well plates with an ampicillin range of 0-280  $\mu$ g/mL ampicillin increasing in 40  $\mu$ g/mL increments.

Ampicillin concentration (µg/mL)	Agar volume (mL)	100 mg/ml ampicillin required (µL)
0	200	0
40	192.8	77
80	185.6	74
120	178.4	71
160	171.2	68
200	164	66
240	156.8	63
280	149.6	60

#### 2.5.2 Culture inoculation and induction

A single colony of SCS1 cells transformed with the required plasmid was used to inoculate 100 mL LB containing 10 µg/mL tetracycline. This was incubated overnight at 37°C, 200 rpm. 1 mL of this overnight culture was used to inoculate 100 mL LB containing 10 µg/mL tetracycline. Cultures were grown at 37°C, 200 rpm until OD<sub>600</sub> = 0.6. Expression was induced by the addition of arabinose to give a final concentration of 0.075 % (*w/v*) and cultures were grown for 1 hr, 37°C, 200 rpm. Log<sub>10</sub> cell dilutions were performed from this culture into sterile 170 mM NaCl. 3 µL from each dilution was pipetted onto each ampicillin concentration and plates were incubated at 37°C for 18 hrs. Following this the maximal cell dilution allowing growth (MCD<sub>GROWTH</sub>) was determined for each ampicillin concentration. The area under the survival curve is calculated as a sum of the areas of 7

trapezia to give a single value using Equation 2.2 where  $x_i$  and  $y_i$  are the x and y values at any given concentration of ampicillin.

$$A_{\text{curve}} = \sum_{i+1}^{7} \frac{2 + y_i + y_{i+1}}{2} (x_{i+1} - x_i)$$
(2.2)

# **2.6** β-lactamase evolution bioassay

#### 2.6.1 Vector design

First design of an appropriate vector was needed, named  $blaGG_{STOP}$ , whereby all the BsaI sites within the vector were removed and two were introduced within the  $\beta$ La GS linker. As BsaI is a Type IIS restriction enzyme it recognises non-palindromic sites and cleaves outside this site, the vector can be designed so that the final ligated product has no BsaI sites as they cut themselves out. Between the two BsaI sites in  $blaGG_{STOP}$  a premature stop codon was introduced as well as the 7bp Bsu36I restriction site. This prematurely ends translation as well as introducing a frame shift. Consequently, if this template is carried over into selection, only the first domain of  $\beta$ La is translated and so it cannot withstand the ampicillin selection.

#### 2.6.2 Library creation

#### 2.6.2.1 Error-prone PCR

GeneMorph II Random Mutagenesis Kit (Agilent) was used to synthesise an error-prone PCR (epPCR) product of the test protein (estimated error rate of 4.5 mutations per 1000 bp, Table 2.19) using forward and reverse primers that anneal to the GS linker regions upand down-stream of the MBP sequence. The amount of template DNA used in the epPCR amplification varies depending on the required mutation rate and is detailed in Table 2.19. The initial amount of target DNA required to achieve a particular mutation frequency refers to the amount of target DNA to amplify, not the total amount of plasmid DNA template to add to the reaction. For example, to mutate a 1 kb fragment at a low mutation rate Table 2.19 recommends 500 ng, if this insert is in an overall 4 kb vector then 2  $\mu$ g of the plasmid construct should be added to give 500 ng of target DNA. Components for an epPCR reaction are detailed in Table 2.20 and the thermocycling conditions used are outlined in Table 2.21. The epPCR product was purified on a 0.8% (*w/v*) agarose gel (Section 2.2.4) using a QIAquick Gel Extraction Kit (Qiagen) according to the manufacturer's instructions and used as the template in a second round of PCR to introduce BsaI sites onto the 5' and 3' ends using Vent polymerase (Section 2.2.3). For libraries of scFvs AMS134 and AMS197 (Chapter 4), only one round of PCR was carried out where primers that add the 5' and 3' BsaI sites were used in the epPCR step to increase the library diversity and reduce the number of steps (Primers are detailed in Appendix A). This second PCR product was purified in the same way and cloned into blaGG<sub>STOP</sub> using the NEB Golden Gate Assembly Kit (BsaI-HFv2) with an insert to vector molar ratio of 2:1, components are detailed in Table 2.22. This was transferred to a thermocycler and conditions are detailed in Table 2.23. Five golden gate reactions (100 µL total) were carried out. These were pooled and purified using a NucleoSpin Gel and PCR Clean-up Kit (Macherey-Nagel) according to the manufacturer's instructions, eluted in 15 µL purite 18 MΩ H<sub>2</sub>O and the concentration determined using NanoDrop (Thermo).

Table 2.19 Mutation frequency vs. initial target quantity for epPCR using Genmorph II random mutagenesis kit.

Mutation rate	Mutation frequency (mutations/kb)	Initial target amount (ng)
Low	0-4.5	500-1000
Medium	4.5-9	100-500
High	9-16	0.1-100

#### Table 2.20 Components for an epPCR using Genmorph II random mutagenesis kit.

Component	Volume (µL)	Final amount
10× Mutazyme II reaction buffer	5	1×
40 mM dNTP mix	1	200 uM
Primer mix	0.5	250 ng/uL
Mutazyme II	1	2.5 U/uL
Template	Variable	
Nuclease-free water	Up to 50	

Step		Temperature (°C)	Time
Initial denatur	ation	95	2 minutes
	Denaturation	95	1 minute
30 cycles	Annealing	Tm of primer - 5°C	1 minute
	Elongation	72	1 minute/kb
Final extension	n	72	10 minutes

Table 2.21 Thermocycling conditions for an epPCR using Genmorph II random mutagenesis kit.

Table 2.22 Components for library generation using golden gate.
---

Component	Amount
Vector (blaGG <sub>STOP</sub> )	75 ng
Insert (gel extracted)	25 ng
NEB golden gate enzyme mix	2 µL
T4 ligase buffer	2 µL
Nuclease-free water	Up to 20 µL

Table 2.23 Thermocycling conditions for library generation using golden gate.

Step		Temperature (°C)	Time (min)
45 avalas	Digestion	37	1
45 cycles	Ligation	16	1
Denaturation		60	5
Hold		4	

During method development the epPCR step was replaced with a traditional PCR (Vent polymerase, Section 2.2.3) using WT MBP as a test protein to create a 'test library' allowing assessment and identification of any issues with the method and to estimate the potential library size. Restriction digests (Section 2.2.5) of 1  $\mu$ g blaGG<sub>STOP</sub>, blaMBP and the test library using Bsu36I were carried out to assess whether the golden gate reaction had gone to completion. No enzyme controls for the restriction digestion and golden gate reactions were carried out.

#### 2.6.2.2 Electroporation of TG1 cells

A 1.0 mm cuvette (BioRad) and 1.5 mL Eppendorf were cooled on ice. 150  $\mu$ L TG1 cells (Lucigen) in 25  $\mu$ L aliquots were thawed on ice before being transferred into the Eppendorf

carefully avoiding introducing bubbles. 240 ng of library was added to the TG1 cells and the tube carefully stirred to avoid bubbles. The cells were incubated on ice for 1 minute before being transferred to the 1.0 mm cuvette and electroporated using a MicroPulser Electroporator (BioRad) (2.5 kV field strength, 335  $\Omega$  resistance and 15 µF capacitance). Within 10 seconds of the pulse,  $6 \times 1$  mL recovery medium was used to wash the cuvette and transfer the cells to a sterile falcon tube. The 6 mL of culture was incubated at 37°C, 250 rpm, for 1 hr. Serial dilutions were performed and 100 µL of these dilutions was plated on LB agar with 10 µg/mL tetracycline. The remaining culture was centrifuged (3000xg, 3 min) and the pellet resuspended in 1 mL of supernatent before plating on a pre-prepared LB agar plate with 10 µg/mL tetracycline. The plates were incubated overnight at 37°C and the serial dilutions used to estimate the library size. Single colonies were picked for sequence analysis before the remaining colonies were removed from the bioassay plates by addition of 10 mL LB medium and 10 mL 50% (*v*/*v*) glycerol before scraping off. The culture was centrifuged (5000xg, 10 min) and purified using a PureYield Plasmid Midiprep System (ProMega), according to the manufacturer's instructions.

#### 2.6.3 Directed evolution

SCS1 cells were defrosted on ice for 10 minutes. 100 µL was added to a prechilled falcon tube. 4 µL of 100 ng/µL of the prepared plasmid library was added to the cells and incubated on ice for 30 minutes, subjected to 42°C heat shock for 45 seconds followed by 5 minutes on ice. 950 µL SOC medium (NEB) was added to each falcon tube and the cultures were incubated (37°C, 200 rpm) for 1 hr. 3 mL SOC medium and tetracycline (final concentration 10 µg/mL) was added to each falcon tube and the cells were grown until OD<sub>600</sub> = 0.6 (MBP constructs) or for 1 hr (biopharmaceutical derivatives). Expression of the β-lactamase construct was induced with 0.075% (*w/v*) arabinose (final concentration) and grown for a further 1 hour. 1 mL of culture was spread on a bioassay plate containing 2.5% (*w/v*) LB, 1.5% (*w/v*) agar, 10 µg/mL tetracycline, 0.075% (*w/v*) arabinose and three different ampicillin concentrations and incubated overnight at 37°C. The evolved libraries were purified in the same way as the naïve libraries using a PureYield Plasmid Midiprep System (ProMega), according to the manufacturer's instructions.

## 2.7 Next-generation sequencing

#### 2.7.1 Illumina and Pacbio Sequencing

Naïve and evolved libraries were amplified by PCR using primers that bound to the 28residue glycine/serine linker that flank the gene of interest within the  $\beta$ La construct to add ~150 bp on the 5' and 3' ends of the test protein to ensure full coverage by Illumina sequencing (Appendix A, Table A.4). PCR products were separated on 0.5 % (*w/v*) agarose gel (Section 2.2.4), eluted in water and the concentration determined using Qubit (Thermo Fisher). Two technical repeats of each sample were sent for 150 bp paired end Illumina sequencing using the Illumina NextSeq 550 at the Microbial Genome Sequencing Centre (Pittsburgh, PA). One technical repeat of each sample was sent for Pacbio sequencing using the Sequel® system by Genewiz UK.

#### 2.7.2 EZ-Amplicon sequencing

Naïve and evolved libraries from Section 2.10.5 were amplified with primers that added Illumina adapters onto the 5' and 3' ends (Appendix A, Table A.4). The resulting fragment covered 199 bp of the gene of interest (HA4). This was sufficient to sequence all the positions of interest while maximising sequencing depth. 5  $\mu$ L of PCR product was visualised using agarose gel electrophoresis (Section 2.2.4), and the remaining PCR product was purified using NucleoSpin Gel and PCR Clean-up Kit (Macherey-Nagel) according to the manufacturer's instructions, eluted in 30  $\mu$ L purite 18 M $\Omega$  H<sub>2</sub>O and the concentration determined using Qubit (Thermo Fisher). The concentration was corrected to 20 ng/ $\mu$ L and 25  $\mu$ L of each sample was sent for 250 bp paired end Illumina EZ-Amplicon sequencing (Genewiz, UK).

#### 2.7.3 Illumina fragment and Pacbio sequencing analysis

Illumina paired end reads were filtered using cutadapt version 1.18 (Martin, 2011) to remove adapter sequences, low quality reads (average quality of the read below Q30) and reads shorter than 40 bp. Reads were aligned to a reference sequence using breseq version 0.34.1 (Deatherage and Barrick, 2014) using bowtie2 version 2.3.4.3 and R version 3.2.2. The resulting \*.bam file was converted to \*.sam using samtools (Li et al., 2009; Li, 2011), INDELS were filtered out and the remaining aligned fragments were translated in frame

using Biopython (Cock et al., 2009; Chapman and Chang, 2000). Mutation frequencies normalised by coverage and mutated residue counts at each position were calculated using python 3.7 and plotted using Origin Pro 2020 version 9.7.0.118. The mean mutation rate at the unmutated GS linker upstream and downstream of the gene of interest was used as a threshold; a mutation rate below this was classed as zero. This was used to calculate the log<sub>2</sub>(fold change) at each residue. Hotspot residues were identified as those with a log<sub>2</sub>(fold change) of more than 10% of the maximum log<sub>2</sub>(fold change) and being identified in both technical repeats (MBP), or as having a log<sub>2</sub>(fold change) of more than two standard deviations from the mean (> $2\sigma$ ) (scFvs). Scripts were written with the help of Dr Michael Davies, University of Leeds.

Circular consensus sequences (CCS) were generated from raw PacBio reads by Genewiz UK to improve the sequence accuracy. Reads were aligned to a reference sequence using bowtie2 version 2.3.4.3. INDELS were filtered out and the remaining reads were translated using Biopython (Cock et al., 2009; Chapman and Chang, 2000), then reads with premature stop codons were discarded. Mutation frequencies, log<sub>2</sub>(fold change) and mutated residue counts at each position were calculated using python 3.7 and plotted using Origin Pro 2020 version 9.7.0.118. The mean mutation rate at the unmutated GS linker upstream and downstream of the gene of interest was used as a threshold; a mutation rate below this was classed as zero. This was used to calculate the log<sub>2</sub>(fold change) at each residue. Hotspot residues were identified as those with a log<sub>2</sub>(fold change) of more than 10% of the maximum log<sub>2</sub>(fold change) and being identified in both technical repeats.

#### 2.7.4 Illumina EZ-Amplicon sequencing analysis

Paired end reads were merged using BBMerge (Bushnell et al., 2017). Merged reads were filtered using cutadapt version 1.18 to remove adapter sequences, low quality reads (below Q40) and reads shorter than 100 bp. Reads were aligned to a reference sequence using bowtie2 version 2.3.4.3. INDELS were filtered out and the remaining reads were translated using Biopython, then reads with premature stop codons were discarded. Mutation frequencies normalised by coverage and mutated residue counts at each position were calculated using python 3.7 and plotted using Origin Pro 2020 version 9.7.0.118. The mean mutation rate at the unmutated GS linker upstream and downstream of the gene of interest was used as a threshold; a mutation rate below this was classed as zero. This was used to calculate the log<sub>2</sub>(fold change) at each residue.

# 2.8 Hierarchical clustering

To assess correlations and similarites between 13 commonly deployed developability assays and TPBLA, a hierachical clustering analysis was performed on the Jain Abs dataset. First, a Spearman's rank correlation was calculated for each developability assay to find its correlation with each other developability assay (Equation 2.3).

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2 - 1)}$$
(2.3)

Where  $\rho$  is the Spearman's rank correlation coefficient,  $d_i^2$  is the difference between two ranks of each observation, and n is the number of observations. The Spearman's rank correlation coefficient can be anywhere between +1 to -1, where +1 is a perfect positive correlation, 0 is no correlation, and -1 is a perfect negative correlation. Then, the Spearman's rank was used as an input for the hierachical clustering analyis in Origin Pro 2020 version 9.7.0.118.

# 2.9 Multiple linear regression analysis for predicting TP-BLA score

#### 2.9.1 AMSCI mAbs

A multiple regression analysis was conducted in R-studio to examine the correlation between assays and properties which measure thermal stability, aggregation, self-association and hydrophobicity with TPBLA score for 12 IgGs provided by UCB Biopharma UK (Chapter 4). These were; first transition mid-point temperatures (Tm1) calculated using differential scanning fluorimetry (DSF) or differential scanning calorimetry (DSC), temperature onset of aggregation (Tonset) calculated using static light scattering (SLS) or dynamic light scattering (DLS), Affinity-Capture Self-Interaction Nanoparticle Spectroscopy (AC-SINS) (Liu et al., 2014), Hydrophobic Interaction Chromatography (HIC) (Estep et al., 2015), size in kDa, and theoretical pI. A multiple linear regression is defined by Equation 2.4.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \dots + \beta_p x_p$$
(2.4)

Where y is the response variable,  $x_1$ ,  $x_2$  and  $x_p$  are predictor variables, and  $\beta_1$ ,  $\beta_2$  and  $\beta_p$  are coefficients or parameters to be estimated for  $x_1$ ,  $x_2$  and  $x_p$  predictor variables, respectively. p represents the number of predictor variables used in the model.

To assess the statistical significance of a regression model, an F-test is carried out to compare the model with zero predictor variables (the intercept only model, or a straight line), and decide whether the coefficients ( $\beta$ ) for each variable improve the models predictive ability (Figure 2.1). In order for the regression to be statistically significant, the f-statistic needs to be higher than the critical f-statistic which is a set value based on the number of degrees of freedom (the number of independent pieces of information that went into calculating the estimate, or the sample size minus the number of restrictions). The f-statistic is defined in Equation 2.5.

$$f = \frac{MSR}{MSE}$$
(2.5)

Where MSR is the mean sum of squares for regression, and MSE is the mean sum of squares for error. The variance *explained* by the regression model is represented as the mean sum of squares for the model, or sum squares regression (SSR) (Figure 2.1). This is essentially assessing whether the regression model is better at explaining the predictor variable than a straight line. The variance *not explained* by the model is the sum of squares for error (SSE), or the sum of squares for residuals (Figure 2.1). The SSE and SSR are used to calculate the MSE and MSR, respectively. The f-statistic is defined from SSR and SSE by Equation 2.6

$$f = \frac{(SSR/DF_{\rm ssr})}{(SSE/DF_{\rm sse})}$$
(2.6)

Where  $DF_{ssr}$ , or p, is the degrees of freedom for the regression model, or the number of paramaters or coefficients, and  $DF_{sse}$  is the degrees of freedom for error, or the total number of records (N) minus the number of coefficients (p) minus one (Equation 2.7).

$$DF_{\rm sse} = N - p - 1 \tag{2.7}$$

Therefore, the f-statistic can be written as in Equation 2.8.

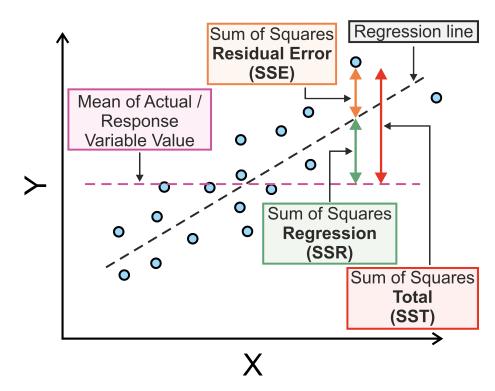


Fig. 2.1 **Mean sum of squares regression and error.** The variance *explained* by the regression model is represented as the mean sum of squares for the model, or sum squares regression (SSR). This is essentially assessing whether the regression model is better at explaining the predictor variable than a straight line. The variance *not explained* by the model is the sum of squares for error (SSE), or the sum of squares for residuals. The f-statistic is defined from SSR and SSE by Equation 2.6. Redrawn with permission from Kumar (2022).

$$f = \frac{(SSR/p)}{(SSE/(N-p-1))}$$
(2.8)

The f-statistic is reported as  $f(DF_{ssr}, DF_{sse})$ . If the f-statistic is above the critical value, which is defined based on the degrees of freedom, the model is assumed to be significant. Furthermore, the absolute value of the f-statistic is used to calculate an exact p-vaue to determine the significance. Here, a p-value of < 0.05 is classed as statistically significant. As a rough rule of thumb, the higher the f-statistic the lower the p-value, and thus the more significant the model.

As well as the f-statistic and p-value, the Pearson correlation coefficient (r) and coefficient of determination ( $R^2$ ) were used to assess the links between predicted TPBLA score using these models, and experimental TPBLA score. r represents the correlation between experimental and predicted TPBLA, whereas  $R^2$  measures the proportion of the variance in predicted TPBLA which can be explained by the predictor variables in the model. In

other words, r is used to identify patterns within the data, whereas  $R^2$  is used to assess the strength of the model.

For each individual variable (assay or property) used in the model to predict TPBLA score, there is a  $\beta$ , t-value, and p-value associated with it to measure the statistical significance. Essentially, to measure how much the individual variable is significantly predicting TPBLA score.  $\beta$  is the value used in the model to multiply that metric by, and essentially the weighting of that metric. The t-value is used to calculate the p-value, which as with before is used to determine significance where a p-value of < 0.05 is classed as statistically significant. As a rough rule of thumb, the further the t-value from 0 the lower the p-value, and thus the more significant the model.

A model including theoretical pI, Tonset by DLS, Tonset by SLS, Tagg by SLS, and Camsol score was statistically significant (f(5, 5) = 5.142,  $R^2 = 0.787$ , r = 0.887, p < 0.05). The fitted model was: TPBLA score = 1490.76 - 398.92(Theoretical pI) + 50.33(Tonset by DLS) - 54.75(Tonset by SLS) + 36.2(Tagg by SLS) + 476.7(Camsol score). Detailed statistics for the model can be found in Appendix C, Table C.1.

By removing Tagg by SLS and including theoretical pI, Tonset by DLS, Tonset by SLS, and Camsol score, the model was statistically significant (f(4, 6) = 5.537,  $R^2 = 0.837$ , r = 0.915, p < 0.05). The fitted model was: TPBLA score = 1093.66 - 282.79(theoretical pI) + 45.8(Tonset by DLS) - 18(Tonset by SLS) + 512.47(Camsol score). Detailed statistics for the model can be found in Appendix C, Table C.2.

A model using only Tonset DLS, Camsol score, and theoretical pI is able to predict TPBLA score reasonably well. The overall regression was statistically significant (f(3, 7) = 7.147,  $R^2 = 0.754$ , r = 0.868, p < 0.05). The fitted model was: 756.443 - 226.442(Theoretical pI) + 25.882(Tonset by DLS) + 493.901(Camsol score). Detailed statistics for the model can be found in Appendix C, Table C.3.

The most parsimonious model was using Tonset by DLS and theoretical pI to give: TPBLA score = 1495.59 + 24.24(Tonset by DLS) - 311.62(theoretical pI). The overall regression was statistically significant (f(2, 8) = 6.192, R<sup>2</sup> = 0.608, r = 0.779, p = 0.0237). Both parameters were significant predictors of TPBLA score (Tonset by DLS:  $\beta$ = 24.24, t = 3.194, p = 0.01; theoretical pI:  $\beta$ = -311.62, t = -3.028, p = 0.02). Detailed statistics for the model can be found in Appendix C, Table C.4.

#### 2.9.2 Jain mAbs

A multiple regression analysis was conducted in R-studio to examine the influence of common developability assays on TPBLA score for 35 clinically relevant therapeutics (Chapter 4). The developability assays were; HEK titer, Thermal midpoint (Tm) determination using differential scanning fluorimetry (DSF) (He et al., 2011), Hydrophobic Interaction Chromatography (HIC) (Estep et al., 2015), Standup Monolayer Adsorption Chromatography (SMAC) (Kohli et al., 2015), Cross Interaction Chromatography (CIC) (Jacobs et al., 2010), Polyspecificity Reagent (PSR) binding (Xu et al., 2013), Accelerated stability (AS), Affinity-Capture Self-Interaction Nanoparticle Spectroscopy (AC-SINS) (Liu et al., 2014), Salt-Gradient AC-SINS (SGAC-SINS) (Estep et al., 2015), Clone Self-Interaction by Biolayer Interferometry (CSI-BLI) (Sun et al., 2013), Enzyme-Linked Immunosorbent Assay (ELISA) (Mouquet et al., 2010), Baculovirus particle (BVP) assay (Hötzel et al., 2012), and Extensional and shear flow device (EFD) (Willis et al., 2020).

An initial model included 7 assays: Tm by DSF, SMAC, AS, PSR binding, CIC, theoretical pI, and scFv molecular weight. The fitted model was: TPBLA score = 2402.4209 + 19.5714(Tm by DSF) - 44.9978(SMAC retention time) + 783.8614(AS) - 362.7396(PSR binding) + 122.7285(CIC retention time) - 122.1716(theoretical pI) - 0.1154(scFv molecular weight). The overall regression was statistically significant (f(7, 27) = 3.968, R<sup>2</sup> = 0.51, r = 0.71, p = 0.004). Detailed statistics for the model can be found in Appendix D, Table D.1.

Individual parameters were removed systematically to find the most parsimonious model. Removing AS gives a statistically significant model: TPBLA score = 2075.57155 + 18.09389(Tm by DSF) - 50.23715(SMAC retention time) - 218.10491(PSR binding) + 115.68934(CIC retention time) - 135.80606(theoretical pI) - 0.08861(scFv molecular weight). The overall regression was statistically significant (f(6, 28) = 3.913,  $R^2 = 0.46$ , r = 0.68, p = 0.005), however using this model PSR and molecular weight do not significantly predict TPBLA score (p > 0.1). Detailed statistics for the model can be found in Appendix D, Table D.2.

Removing molecular weight gives a statistically significant model: TPBLA score = -268.19 + 17.47(Tm by DSF) - 61.70(SMAC retention time) - 263.78(PSR binding) + 134.40(CIC retention time) - 132.78(theoretical pI). The overall regression was statistically significant (f(5, 29) = 4.116, R<sup>2</sup> = 0.42, r = 0.64, p = 0.006) where all parameters significantly predict TPBLA score (p < 0.1). Detailed statistics for the model can be found in Appendix D, Table D.3.

The model with 5 parmeters was re-calculated without 6 mAbs to see if it could be used to predict their TPBLA score. This fitted model was: -158.507 + 16.164(Tm by DSF) - 65.309(SMAC retention time) - 288.447(PSR binding) + 142.548(CIC retention time) - 137.56(theoretical pI). The overall regression was statistically significant (f(5, 23) = 3.375,  $R^2 = 0.42$ , r = 0.64, p = 0.01971) where all parameters significantly predict TPBLA score (p < 0.1). Detailed statistics for the model can be found in Appendix D, Table D.4.

# 2.10 Towards an assay for the simultaneous evolution of aggregation resistance and binding affinity

#### 2.10.1 Western blot analysis

A single colony of fresh *E. coli* SCS1 cells (transformed with the appropriate plasmid) was used to inoculate 100 mL sterile LB containing 10 µg/mL tetracycline. Cultures were incubated overnight at 37°C with shaking (200 rpm). 1 mL of overnight culture was used to inoculate 100 mL sterile LB containing 10 µg/mL tetracycline and grown at 37°C (shaking at 200 rpm) until an OD<sub>600</sub> of 0.6 was obtained. 10 mL of culture was removed for the uninduced sample and centrifuged at 4,000 g for 10 min (4°C). Expression of HA4- $\beta$ -lactamase-mNG2 and SH2-mNG2 fusion constructs were induced by the addition of filter-sterilized arabinose to a final concentration of 0.075 % (*w/v*) arabinose. Cultures were incubated for 1 h (37°C, 200 rpm) and 10 mL was removed from each (representing the induced sample). The 10 mL cultures were harvested by centrifugation at 4,000 g for 10 min (4°C). The cell pellets (uninduced and induced with arabinose) were resuspended in phosphate buffered saline (PBS, Dulbecco's PBS, Sigma) to obtain an OD<sub>600</sub> of 5. For whole cell samples, 20 µL of the OD<sub>600</sub> = 5 sample was combined with 8 µL PBS, 2 µL 1 M DTT and 10 µL of 4× loading dye (200 mM Tris.HCl, pH 6.8, 6 % (*v/v*) SDS, 0.3 % (*w/v*) bromophenol blue, 40 % (*v/v*) glycerol).

For the periplasmic fraction, 1 mL of  $OD_{600} = 5$  culture was pelleted by centrifugation at 4,000 g for 3 min and the supernatant discarded. The pellet was resuspended in 900 µL 0.1 M Tris pH 8.0, 500 mM sucrose, 0.5 mM EDTA pH 8.0 and incubated for 5 min at room temperature. The cells were centrifuged (4,000 g, 3 min) and the supernatant discarded. The pellet was resuspended in 400 µL purite 18 M $\Omega$  H<sub>2</sub>O and incubated on ice for 15 seconds before the addition of 20 µL 20 mM MgSO<sub>4</sub>. The sample was centrifuged (14,000 g, 5 min) and the supernatant (periplasmic fraction) was carefully pipetted off. 28 µL of the periplasmic fraction was combined with 2 µL 1 M DTT and 10 µL 4× loading dye. For the cytoplasmic fraction, the resulting pellet was resuspended in 400  $\mu$ L 400  $\mu$ L bacterial protein extraction reagent (B-PER, ThermoFisher) and incubated with agitation for 10 min. The sample was then centrifuged at 16,000 g for 10 min and the supernatant was carefully pipetted off. 28  $\mu$ L of this was combined with 2  $\mu$ L 1 M DTT and 10  $\mu$ L 4× loading dye. The resulting insoluble pellet was resuspended in 400  $\mu$ L PBS and 28  $\mu$ L of this was combined with 2  $\mu$ L 1 M DTT and 10  $\mu$ L 4× loading dye. The resulting insoluble pellet was resuspended in 400  $\mu$ L PBS and 28  $\mu$ L of this was combined with 2  $\mu$ L 1 M DTT and 10  $\mu$ L 4× loading dye. The whole cell, periplasmic fraction and cytoplasmic fraction, and insoluble samples were then incubated at 90°C for 10 min.

Protein samples were separated on a BIORAD Mini-PROTEAN TGX precast electrophoresis gel and were transferred to a BIORAD 0.2 µm polyvinylidene fluoride membrane using a Trans-Blot Turbo Semi-Dry (Bio-Rad Ltd). Blocking was performed using 5 % (*w*/*v*) milk powder in TBST (Tris-buffered saline Tween; 20 mM Tris.HCl, 150 mM NaCl, 0.2 % (*v*/*v*) Tween-20). Membranes were incubated overnight with the anti- $\beta$ -lactamase antibody (CSB-PA352353YA01ENL, Cusabio) or anti-Abl SH2 domain antibody (06-465, Sigma-Aldrich) diluted 1:10,000 in 5 % (*w*/*v*) milk powder in TBST. The membranes were washed for 3 × 10 mL in TBST. Membranes were then incubated with goat anti-rabbit IgG horseradish peroxidase (HRP) conjugate (7074, New England BioLabs) or anti-mouse IgG HRP conjugate (Cell Signaling Technology, 7076S) diluted 1:10,000 in TBST. Membranes were then washed 3 × 10 mL in TBST before incubation with SuperSignal<sup>TM</sup> western pico chemiluminescent substrate (Thermo Fisher Scientific). The emitted signal was visualised and imaged using UVItec Q9 Alliance Gel Doc.

#### 2.10.2 Fluorescence spectroscopy

A single colony of SCS1 cells transformed with the appropriate plasmid was used to innoculate 5 mL LB containing 10 µg/mL tetracycline. This was incubated overnight at 37°C, 200 rpm. 0.2 mL of this overnight culture was used to inoculate 20 mL LB containing 10 µg/mL tetracycline. Cultures were grown at 37°C, 200 rpm until OD<sub>600</sub> = 0.6. 5 mL of culture was removed for the uninduced sample and centrifuged at 4,000 g for 10 min (4°C). Expression was induced by the addition of arabinose to give a final concentration of 0.075 % (*w/v*) and cultures were grown for a further 3 hr, 37 °C, 200 rpm. 5 mL of culture was harvested by centrifugation at 4,000 g for 10 min (4°C). The cell pellets (uninduced and induced with arabinose) were resuspended in phosphate buffered saline (PBS, Dulbecco's PBS, Sigma) to obtain 1 mL at OD<sub>600</sub> = 1.

To isolate the periplasmic fraction, 1 mL of  $OD_{600} = 1$  culture was pelleted by centrifugation at 4,000 g for 3 min and the supernatant discarded. The pellet was resuspended in 900 µL 0.1 M Tris pH 8.0, 500 mM sucrose, 0.5 mM EDTA pH 8.0 and incubated for 5 min at room temperature. The cells were centrifuged (4,000 g, 3 min) and the supernatant discarded. The pellet was resuspended in 400 µL purite 18 M $\Omega$  H<sub>2</sub>O and incubated on ice for 15 seconds before the addition of 20 µL 20 mM MgSO<sub>4</sub>. The sample was centrifuged (14,000 g, 5 min) and the supernatant was carefully pipetted off. 300 µL of this was mixed with 300 µL PBS to give the periplasmic fraction. To isolate the cytoplasmic fraction, the resulting pellet was resuspended in 400 µL bacterial protein extraction reagent (B-PER, ThermoFisher) and incubated with agitation for 10 min. The sample was then centrifuged at 16,000 g for 10 min and the supernatant was carefully pipetted off. 300 µL of this was mixed with 300 µL PBS to give the cytoplasmic fraction.

mNeonGreen2 fluorescence in the periplasm was measured using an excitation wavelength of 488 nm and collecting emission spectra between 500 nm and 530 nm with 1 nm slit widths. mScarlet-I fluorescence in the cytoplasm was measured using an excitation wavelength of 569 nm and collecting emission spectra between 580 nm and 600 nm with 1 nm slit widths. A control was carried out with PBS only.

#### 2.10.3 Plate reader fluorescence

A single colony of SCS1 cells transformed with the appropriate plasmids were used to innoculate 5 mL LB containing 10 µg/mL tetracycline and 50 µg/mL kanamycin. This was incubated overnight at 37°C, 200 rpm. 0.1 mL of this overnight culture was used to inoculate 10 mL LB containing 10 µg/mL tetracycline and 50 µg/mL kanamycin. Cultures were grown at 37°C, 200 rpm until  $OD_{600} = 0.3$ . Expression was induced by the addition of arabinose ( $\beta$ -lactamase fusion) and IPTG (CadC-SH2). Dimerisation of the VHH domain was induced via the addition of 50 - 100 µM caffeine. 100 µL of culture was added to the wells of a 96-well flat bottom assay plate and sealed with adhesive sealing film. Green fluorescence was monitored in a FLUOstar Omega plate reader at 37 °C with continuous orbital agitation at 200 rpm. The fluorescence of sfGFP was excited at 488 nm and fluorescence emission was monitored at 561 nm.

#### 2.10.4 Fluorescence activated cell sorting

Cells from an overnight culture (Section 2.10.3) were centrifuged and resuspended in PBS to  $OD_{600} = 0.01$ . For visualisation, individual cells were sorted using a Cytoflex S Flow Cytometer (Beckman Coulter). Green fluorescence (mNeonGreen2, sfGFP) was measured using an excitation laser at 488 nm and measuring emission using an excitation band pass filter of 525/40 nm. Side and forward scattering was measured and used to identify whole cells and to filter out lysed or doublet cells. For sorting, individual cells were sorted using a FACS Melody (BD Biosciences). Sorting experiments were carried out at the University of Leeds Bioimaging Facility with Dr Ruth Hughes. Green fluorescence was measured using an excitation laser at 488 nm and measuring emission using an excitation band pass filter of 510/10 nm. Cells with positive fluorescence were sorted into LB. Sorted cells were used to inoculate 20 mL LB containing 10 µg/mL tetracycline and 50 µg/mL kanamycin and grown overnight at 37°C, 200 rpm. The resulting DNA was sequenced using Sanger sequencing (Section 2.2.9) and Illumina EZ-Amplicon sequencing (Section 2.7.2).

#### 2.10.5 Dual selection screening

#### 2.10.5.1 TPBLA screen for solubility

A 'library' was created mixing equal molar amounts of blaHA4WT-VHH, blaHA4Y87A-VHH, blaHA4<sup>2A</sup>-VHH, and blaHA4<sup>Y87A 2A</sup>-VHH (in the pBR322 TPBLA plasmid, Table 2.15). 100 µL of SCS1 cells were transformed with 2 µL of this 'library' and 2 µL of pHJ12 CadC-SH2. Cells were incubated on ice for 30 minutes, subjected to 42°C heat shock for 45 seconds followed by 5 minutes on ice. 950 µL SOC medium (NEB) was added to each falcon tube and the cultures were incubated (37°C, 200 rpm) for 1 hr. 3 mL SOC medium, tetracycline (final concentration 10 µg/mL), and kanamycin (final concentration 50 µg/mL) was added to each falcon tube and the cells were grown for 1 hr. Expression of the  $\beta$ -lactamase-VHH construct was induced with 0.075% (w/v) arabinose (final concentration) and grown for a further 1 hour. 1 mL of culture was spread on a bioassay plate containing 2.5% (w/v) LB, 1.5% (w/v) agar, 10 µg/mL tetracycline, 50 µg/mL kanamycin, 0.075% (w/v) arabinose and three different ampicillin (5, 10, and 15 µg/mL) concentrations and incubated overnight at 37°C. The evolved libraries were scraped and used to inoculate 10 mL LB containing 10 µg/mL tetracycline, 50 µg/mL kanamycin. Cells were grown overnight (37 °C, 200 rpm) and the resulting DNA purified using a QIAquick Spin Miniprep Kit (QIAGEN, Crawley, UK), according to the manufacturer's instructions.

#### 2.10.5.2 FACS screen for binding affinity

20  $\mu$ L of SCS1 cells were transformed with 2  $\mu$ L of the DNA purified in Section 2.10.5.1. A control was set up using 2  $\mu$ L of an equal mix of pBR322 blaHA4-VHH blaHA4<sup>Y87A</sup>-VHH, and 1  $\mu$ L of pHJ12 CadC-SH2 to transform 20  $\mu$ L SCS1 cells. Cells were incubated on ice for 30 minutes, subjected to 42°C heat shock for 45 seconds followed by 5 minutes on ice. 200  $\mu$ L SOC medium (NEB) was added to each falcon tube and the cultures were incubated (37°C, 200 rpm) for 1 hr. These cells were used to inoculate 5 mL LB containing 10  $\mu$ g/mL tetracycline, 50  $\mu$ g/mL kanamycin and grown overnight at 37 °C, 200 rpm.

The overnight grow was diluted 1:1000 into LB containing 10 µg/mL tetracycline and 50 µg/mL kanamycin and grown at 37°C, 200 rpm until  $OD_{600} = 0.3$  (Section 2.10.3). Expression and dimerisation of VHH was induced via the addition of 0.01 mM arabinose, 25 µM IPTG, and 100 µM caffeine (final concentrations). 100 µL of culture was added to the wells of a 96-well flat bottom assay plate and sealed with adhesive sealing film. Green fluorescence was monitored in a FLUOstar Omega plate reader at 37 °C with continuous orbital agitation at 200 rpm. The fluorescence of sfGFP was excited at 488 nm and fluorescence emission was monitored at 561 nm.

Cells were sorted using FACS Melody as described in Section 2.10.4, and the resulting DNA sequenced using Sanger sequencing (Section 2.2.9) and Illumina EZ-Amplicon sequencing (Section 2.7.2).

# **Chapter 3**

# Combining deep sequencing with TPBLA for directed evolution

# 3.1 Introduction

The TPBLA described in Section 1.7.4 has previously been utilised to successfully assess and evolve thermodynamic stability (Foit et al., 2009) and aggregation propensity (Ebo et al., 2020a) in a wide range of proteins, as well as to identify small molecule inhibitors of the aggregation of amyloid proteins (Saunders et al., 2016). However, previous work using TPBLA for directed evolution studies was limited by first-generation sequencing techniques, making the process laborious, low-throughput and high cost. This chapter aimed to combine TPBLA with the power of deep sequencing to enhance the throughput of the assay and create large datasets of mutational data to better understand the mechanisms whereby TPBLA evolves protein behaviour. For a directed evolution experiment a random mutated library must be created, however common approaches to create these libraries have their own limitations, particularly when utilising error-prone PCR (epPCR). Often the rate-limiting step for these libraries is the cloning of the epPCR fragment back into the template vector (Abou-Nader and Benedik, 2010). Therefore, we sought to develop a robust methodology for creating large randomly mutated plasmid libraries by utilising golden gate assembly, and future selection of improved properties using TPBLA.

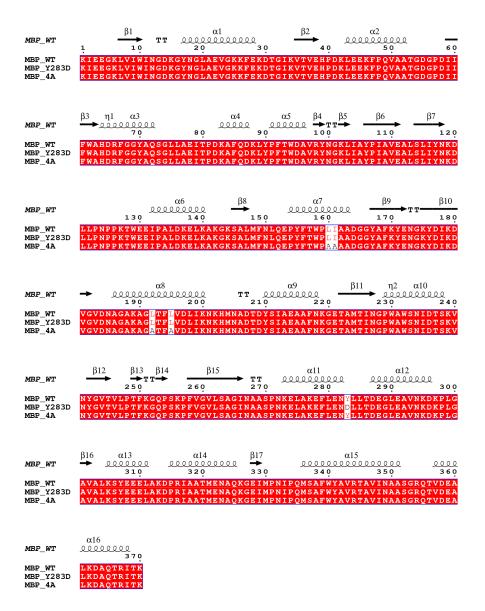


Fig. 3.1 **Sequence alignment of MBP<sup>WT</sup>, MBP<sup>Y283D</sup>, and MBP<sup>4A</sup>** Amino acid sequences of the variants used in this study. MBP<sup>Y283D</sup> has been previously shown to introduce a folding defect (Chun et al., 1993). MBP<sup>4A</sup> was designed by introducing four mutations L160A, I161A, L192A and L195A to reduce the thermodynamic stability by introduction of a cavity. Sequences aligned using MultAlin (Corpet, 1988), figure made using ESPript 3.0 (Robert and Gouet, 2014).

#### **3.1.1** Maltose binding protein

This chapter utilises maltose binding protein (MBP) as a model test protein, providing an example of a well-studied, highly soluble, and highly expressed protein. MBP is a 42 kDa periplasmic protein comprised of two globular domains separated by a groove in which maltose binds (Duan et al., 2001). Upon binding to maltose these two domains rotate to clamp down on the ligand so that it is bound between both domains, forming a "closed" structure (Duan et al., 2001). In the unbound state, both domains rotate and form an "open" structure, exposing the sugar binding cleft (Duan et al., 2001). MBP is often exploited as a solubility tag to aid in folding and solubility of difficult to express proteins (Salema and Fernández, 2013; Costa et al., 2014; Nguyen et al., 2016; Lénon et al., 2021; Raran-Kurussi et al., 2022; Jo, 2022). To evaluate the ability of TPBLA to selectively evolve aggregation resistance and thermodynamic stability, two distinctive sequence liabilities were introduced into MBP (Figure 3.1). A Y283D substitution which has been shown previously to introduce a folding defect, resulting in the accumulation of kinetically trapped folding intermediates, however the extent by which this mutation impacts protein aggregation is unclear (Chun et al., 1993). This variant has been used in folding studies as a "slow-folding species that does not aggregate by fluorescence spectroscopy" (Sparrer et al., 1997), however it has also been exploited in directed evolution studies to evolve solubility as it was shown to have reduced soluble expression in E. coli when compared with wild-type MBP (Wang et al., 2018). Furthermore, a study assessing various MBP point mutants and their effect on their fusion proteins demonstrated that MBP<sup>Y283D</sup> is thermodynamically destabilised with respect to wild-type and has significantly lower expression of soluble protein in E. coli when analysed by SDS-PAGE (~35% soluble MBP<sup>Y283D</sup> compared with ~100% soluble MBP<sup>WT</sup>) (Fox et al., 2001). Moreover, when used as a fusion protein MBP<sup>Y283D</sup> significantly reduced the soluble expression of the fusion construct with <10% soluble protein expressed for 3 different fusion constructs (Fox et al., 2001). The rationale was that the rate of MBP folding is crucial for promoting solubility of the fusion construct. Also in the same study, MBP<sup>Y283D</sup> was shown to have a higher thermodynamic stability than other single point mutants assessed, however these mutants displayed soluble phenotypes. Nevertheless, this mutation was selected to develop our high-throughput evolution methodology using a slow-folding, potentially aggregating variant.

The second MBP variant assessed is MBP<sup>4A</sup>, a variant designed in this study used to analyse the effect of solely reducing the thermodynamic stability without altering the overall global fold. Simple 'large-to-small' substitutions of buried leucine or isoleucine residues to alanine residues (L160A, I161A, L192A, L195A) were introduced within the

hydrophobic core of MBP, as such mutations have been shown to result in destabilisation by introduction of a cavity (Eriksson et al., 1992). Using MBP and these two variants as a model system we develop a next-generation sequencing approach to enhance TPBLA and enable massively parallel analysis of thousands of variants in a high-throughput manner.

#### 3.1.2 Directed evolution

As discussed at length in Section 1.6.2, directed evolution is a widely exploited technique for the selective evolution of desirable properties in a protein. In contrast to rational design (Section 1.6.1), directed evolution does not require prior knowledge about the protein of interest, making it a powerful tool for biopharmaceutical engineering and for the development of novel therapeutics. Often antibodies identified during affinity maturation have reduced thermodynamic stability or an increased aggregation propensity, as proteins are only marginally stable the majority of mutations are likely to be destabilising (Julian et al., 2017). The complex mechanisms governing thermodynamic stability and aggregation propensity will differ between proteins, making it incredibly difficult to predict. Therefore, it is of vital importance to have a robust tool for identifying stabilising and solubilising mutations in a high-throughput manner to streamline the development pipeline and vastly improve the developability of a biotherapeutic.

#### **3.1.3** Aims of the study

This chapter adapts the previously developed TPBLA to increase its throughput and ease of use to evolve variants of maltose binding protein (MBP). Combining TPBLA with Illumina and Pacbio sequencing is shown to have the potential to assess thousands of evolved variants to identify stabilising and solubilising mutations in a high-throughput manner. The relative ease and low-cost of our method make TPBLA a powerful tool for the high-throughput engineering of protein solubility.

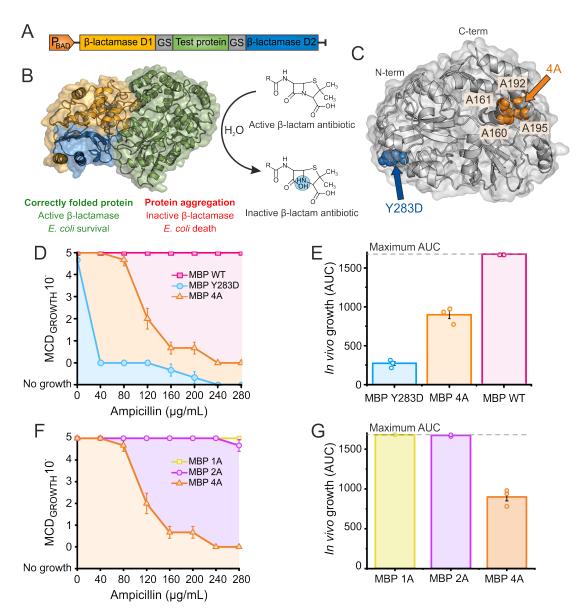
## **3.2 Results**

#### **3.2.1** TPBLA can be used to assess different sequence liabilities

TPBLA has previously been used to assess the aggregation propensity of intrinsically disordered proteins as well as various therapeutic scaffolds and has additionally been correlated to the thermodynamic stability of proteins (Foit et al., 2009; Saunders et al., 2016; Ebo et al., 2020a). The question remains as to which property is the driving force for performance in TPBLA. To test this, we assessed wild-type MBP, as well as a known slow folding mutant (Y283D) that reduces the folding rate and causes the protein to be caught in kinetic traps and form insoluble aggregates (Chun et al., 1993). In addition, a thermodynamically destabilised mutant (4A) which was designed in the current study by mutating leucine or isoleucine to alanine (L160A, I161A, L192A, L195A) within the core of the second domain of the protein was generated. These substitutions should reduce the thermodynamic stability of MBP by introducing a cavity within its hydrophobic core (Eriksson et al., 1992) (Figure 3.1). These proteins are referred here to as MBP<sup>WT</sup>, MBP<sup>Y283D</sup> and MBP<sup>4A</sup>, respectively (Figure 3.2C). MBP<sup>4A</sup> was designed by systematically introducing one (L160A), two (L160A + I161A), then four (L160A, I161A, L192A, L195A) point mutations until the desired destabilising affect had been achieved, as assessed by TPBLA (Figure 3.2D, E).

The *in vivo* growth score of bacteria expressing tripartite  $\beta$ -lactamase with MBP<sup>WT</sup>, MBP<sup>Y283D</sup> or MBP<sup>4A</sup> as the test protein was measured in a 48 well plate format over an ampicillin concentration range of 0-280 µg/mL. Both MBP<sup>Y283D</sup> and MBP<sup>4A</sup> show reduced growth in the assay with respect to MBP<sup>WT</sup>, with MBP<sup>Y283D</sup> having the lowest *in vivo* growth (area under the antibiotic survival curve, Figure 3.2D, E).

MBP<sup>WT</sup>, MBP<sup>Y283D</sup> and MBP<sup>4A</sup> were expressed and purified in *E. coli* for biophysical characterisation. MBP<sup>WT</sup> (Figure 3.3) and MBP<sup>Y283D</sup> (Figure 3.4) were purified from the soluble fraction by Immobilized-metal affinity chromatography (IMAC) using a HisTag. MBP<sup>4A</sup> (Figure 3.5) expressed almost exclusively in the insoluble fraction, so needed to be re-folded prior to HisTag purification. For all three proteins, the HisTag was then cleaved off using TEV protease before a final polishing gel filtration step was carried out to isolate the monomer. The monomeric protein for MBP<sup>WT</sup> (Figure 3.3D), MBP<sup>Y283D</sup> (Figure 3.4D) and MBP<sup>4A</sup> (Figure 3.5E) was analysed using mass spectrometry to confirm its purity. These purified proteins were then characterised based on their structure, stability, and aggregation behaviours to provide a starting point for future directed evolution experiments using TPBLA.



110

Fig. 3.2 In vivo tripartite  $\beta$ -lactamase assay (TPBLA). A) TPBLA construct. Test protein (green) is inserted between two domains (D1, domain 1; D2, domain 2) of genetically separated  $\beta$ -lactamase (orange and blue), joined by a 28-residue glycine/serine linker (grey). B) Correct folding of the test protein results in association of the two β-lactamase domains, forming the active enzyme that can hydrolyse  $\beta$ -lactam antibiotics. Aggregation of the test protein blocks association of the two  $\beta$ -lactamase domains, increasing the *E. coli*'s sensitivity to  $\beta$ -lactams. C) Position of the Y283D and 4A mutations. D) In vivo screen of MBP<sup>WT</sup> and variants. Antibiotic survival curve showing the maximum cell dilution allowing growth (MCD<sub>GROWTH</sub>) over an ampicillin concentration range of 0-280 µg/mL. Error bars show standard error of the mean (S.E.M) from three independent experiments. E) Area under the antibiotic survival curve (AUC) calculated for MBPWT, MBP<sup>Y283D</sup> and MBP<sup>4A</sup>, screened at 0-280 µg/mL ampicillin. Error bars show standard error of the mean (S.E.M) from three independent experiments. Dotted line shows the maximum AUC attainable at this antibiotic concentration range. F) MBP<sup>4A</sup> was designed by systematically introducing one (L160A, MBP 1A), two (L160A + I161A, MBP 2A), then four (L160A, I161A, L192A, L195A, MBP 4A) point mutations until the desired destabilising affect had been achieved. This was measured as a reduction in in vivo growth score. Error bars show standard error of the mean (S.E.M) from three independent experiments. G) Area under the antibiotic survival curve (AUC) calculated for MBP<sup>1A</sup>, MBP<sup>2A</sup> and MBP<sup>4A</sup>. Error bars show standard error of the mean (S.E.M) from three independent experiments. Dotted line shows the maximum AUC attainable at this antibiotic concentration range.

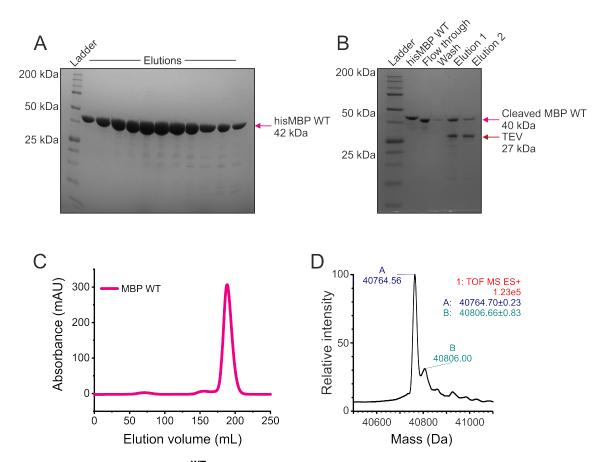


Fig. 3.3 **Purification of MBP**<sup>WT</sup> in autoinduction media. A) HisTrap fractions from purification of MBP<sup>WT</sup>. The eluted protein was cleaved using TEV protease to remove the HisTag. B) Cleaved MBP<sup>WT</sup> was separated from uncleaved his-MBP<sup>WT</sup> by running over a HisTrap column and collecting the unbound protein (flow through). C) The resulting protein was purified on a HiLoad Superdex 75 gel filtration column to isolate the monomeric protein. D) Deconvoluted mass spectra. Peak A corresponds to MBP<sup>WT</sup>. Peak B is one additional species +41 Da, which corresponds to a single molecule of acetonitrile likely picked up during sample preparation as this is the solvent used in the mobile phase. Mass spectrometry experiments and analysis carried out by Samantha Lawrence, University of Leeds.

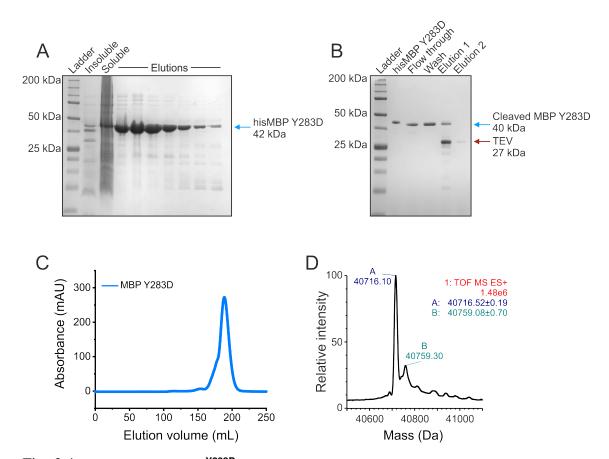


Fig. 3.4 **Purification of MBP**<sup>Y283D</sup> in autoinduction media. A) HisTrap fractions from purification of MBP<sup>Y283D</sup>. The eluted protein was cleaved using TEV protease to remove the HisTag. B) Cleaved MBP<sup>Y283D</sup> was separated from uncleaved his-MBP<sup>Y283D</sup> by running over a HisTrap column and collecting the unbound protein (flow through). C) The resulting protein was purified on a HiLoad Superdex 75 gel filtration column to isolate the monomeric protein. D) Deconvoluted mass spectra. Peak A corresponds to MBP<sup>Y283D</sup>. Peak B is one additional species +41 Da, which corresponds to a single molecule of acetonitrile likely picked up during sample preparation as this is the solvent used in the mobile phase. Mass spectrometry experiments and analysis carried out by Samantha Lawrence, University of Leeds.

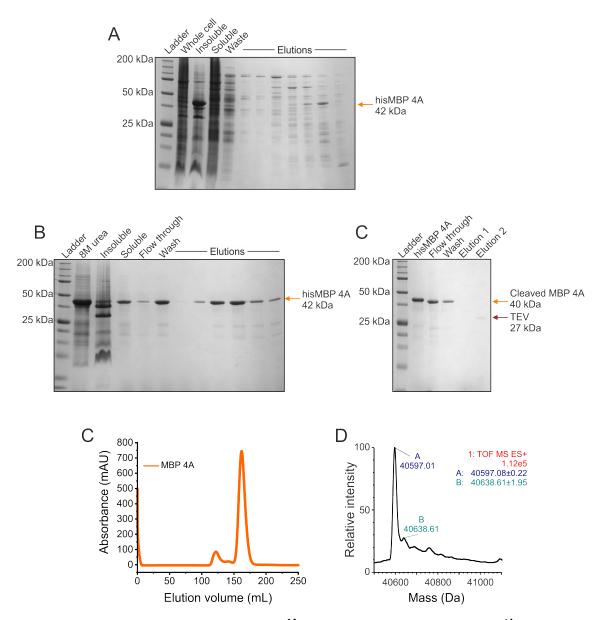


Fig. 3.5 **Refolding and purification of MBP<sup>4A</sup> in autoinduction media.** A) MBP<sup>4A</sup> was not expressed in the soluble fraction so could not be purified in the same way as MBP<sup>WT</sup> or MBP<sup>Y283D</sup>. The insoluble protein was unfolded in 8M Urea before being re-folded (For a more detailed method, see Section 2.3.1.3). B) The refolded his-MBP<sup>4A</sup> was then purified on a HisTrap column. The eluted protein was cleaved using TEV protease to remove the HisTag. C) Cleaved MBP<sup>4A</sup> was separated from uncleaved his-MBP<sup>4A</sup> by running over a HisTrap column and collecting the unbound protein (flow through). D) The resulting protein was purified on a HiLoad Superdex 75 gel filtration column to isolate the monomeric protein. E) Deconvoluted mass spectra. Peak A corresponds to MBP<sup>4A</sup>. Peak B is one additional species +41 Da, which corresponds to a single molecule of acetonitrile likely picked up during sample preparation as this is the solvent used in the mobile phase. Mass spectrometry experiments and analysis carried out by Samantha Lawrence, University of Leeds.

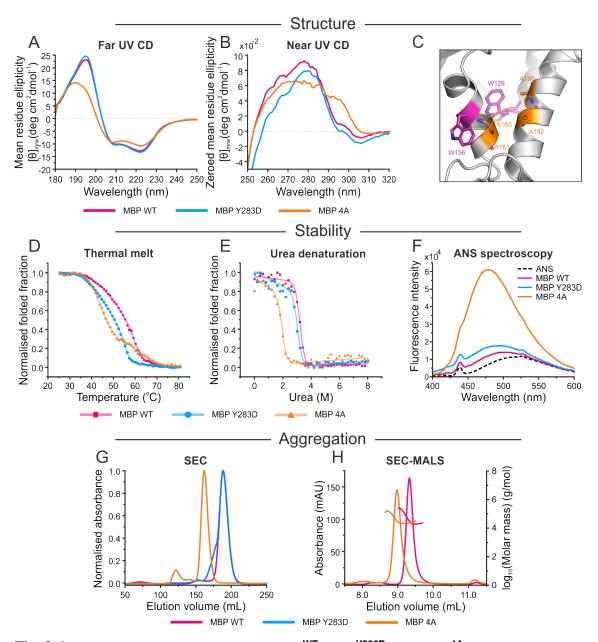


Fig. 3.6 Biochemical characterisation of MBP<sup>WT</sup>, MBP<sup>Y283D</sup>, and MBP<sup>4A</sup>. A) Far and B) near UV CD spectra in 10 mM potassium phosphate pH 7.4. Dotted line shows zero, near UV spectra is zeroed. All proteins show characteristic  $\alpha$  helical spectra. C) Position of MBP<sup>4A</sup> mutations relative to intrinsic tryptophans. These mutations could alter the tryptophan environment, explaining the change in the near UV CD spectra (particularly at 290-300 nm). C) Circular dichroism thermal melt (temperature range 25 °C to 80 °C, protein concentration 0.15 mg/mL) fitted with a sigmoidal curve and D) urea denaturation of MBP<sup>WT</sup>, MBP<sup>Y283D</sup> and MBP<sup>4A</sup> (protein concentration 0.1  $\mu$ M) measured using Trp fluorescence in 10 mM potassium phosphate (pH 7.4) fitted to a two state unfolding curve. E) 8-Anilinonapthalene-1-sulphonic acid (ANS) fluorescence spectroscopy showing increased binding for MBP<sup>4A</sup> indicating more exposure of hydrophobic patches on the surface, or a 'loosening' of the core structure allowing ANS in to bind. F) Gel filtration chromatography of MBP<sup>WT</sup>, MBP<sup>Y283D</sup> and MBP<sup>4A</sup> and a lower elution volume. G) Size exclusion chromatography multi angle light scattering (SEC-MALS) shows MBP<sup>4A</sup> and MBP<sup>WT</sup> are both monomeric (MBP<sup>WT</sup> = 39.8 kDa, MBP<sup>4A</sup> = 39.2 kDa), although MBP<sup>4A</sup> has a smaller elution volume. Dashed lines represent MALS data.

 $MBP^{Y283D}$  retains the same overall global fold as  $MBP^{WT}$ , having comparable secondary and tertiary structures as determined by far and near UV circular dichroism (Figure 3.6A,B).  $MBP^{4A}$  has a reduction in helicity and slight change in tryptophan environment, as assessed by far and near UV circular dichroism, respectively (Figure 3.6A-C). Both variants are thermally destabilised with respect to  $MBP^{WT}$ , as determined by circular dichroism thermal melt (Figure 3.6D).  $MBP^{4A}$  had the lowest thermal melting midpoint (Tm) (45.4 °C), compared with  $MBP^{Y283D}$  (Tm = 48.0 °C) and  $MBP^{WT}$  (Tm = 53.2 °C).

As the pre- and post- transitions from the urea denaturation curves have low signal to noise, the data was used to calculate the midpoint denaturant concentration (Cm) rather than  $\Delta G$  (Figure 3.6E).  $\Delta G$  is very sensitive to changes in the pre- and post- transitions, therefore any calculation of  $\Delta G$  from this data would be inaccurate. As with the thermal stability, MBP<sup>4A</sup> had the lowest Cm (1.87 M), compared with MBP<sup>Y283D</sup> (Cm = 2.97 M) and MBP<sup>WT</sup> (Cm = 3.19 M). This again demonstrates the effect of the destabilising mutations on MBP<sup>4A</sup>.

MBP<sup>WT</sup>, MBP<sup>Y283D</sup> and MBP<sup>4A</sup> were crystallised to determine the affect of the point mutations on the overall structure of MBP. The crystal structure also gives something to map mutational hotspots on to following directed evolution with TPBLA. The crystal structures of MBPWT, MBPY283D, and MBP4A were solved at 1.7 Å, 1.4 Å, and 2.1 Å, respectively. MBP<sup>Y283D</sup> has a similar crystal structure to MBP<sup>WT</sup> (RMSD = 0.159Å) although the Asp283 introduced eliminates a polar interaction between the wild-type Asp30 potentially exposing an otherwise buried patch of hydrophobic residues that may have been protected in the wild-type structure (Figure 3.7). This mutation may slow down folding resulting in aggregation (Figure 3.7). MBP<sup>4A</sup> has undergone significant domain movement in the crystal structure compared with MBP<sup>WT</sup>, and part of the second domain (residues 149-207) is poorly resolved, indicating the protein is more dynamic (Figure 3.8A). Data collection and refinement statistics for the crystal structures are shown in Table 3.1. MBP<sup>4A</sup> forms a dimer in the crystal structure (Figure 3.9) but is monomeric in solution (SEC-MALS, Figure 3.6G-H) suggesting the dimeric state is a crystal packing artefact. Nonetheless, the lack of density in the crystal structure indicates the second domain has been highly destabilised. The 4A mutations may destabilise the protein by removing van der Waals interactions within the core and forming a cavity (Figure 3.8B). This results in the formation of a less packed core that allows 8-Anilinonaphthalene-1-sulfonic acid (ANS) in to bind and creates a more open monomer that has a shorter retention time on a SEC column compared with MBP<sup>WT</sup> (Figure 3.6F).

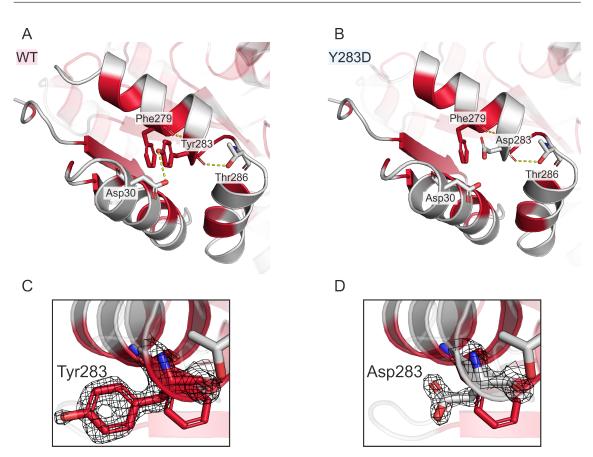


Fig. 3.7 **Comparison of MBP<sup>WT</sup>, MBP<sup>Y283D</sup> and MBP<sup>4A</sup> crystal structures.** A) Tyr283 makes a polar contact with Thr286 and Asp30 in MBP<sup>WT</sup>, and is involved in  $\pi$ - $\pi$  stacking with Phe279. B) The polar contact between Tyr283 and Asp30 and  $\pi$ - $\pi$  interaction between Tyr283 and Phe279 is lost in MBP<sup>Y283D</sup>, potentially exposing otherwise buried hydrophobic aggregation prone residues. Hydrophobic residues are coloured in red. MBP<sup>WT</sup> vs MBP<sup>Y283D</sup> RMSD = 0.159 Å. C) 2Fo-Fc electron density map showing Tyr283 in MBP<sup>WT</sup> contoured at 1.5 $\sigma$  with a 1.6Å carve radius. D) 2Fo-Fc electron density map showing Asp283 in MBP<sup>Y283D</sup> contoured at 1.5 $\sigma$  with a 1.6Å carve radius. Crystallographic data collection and analysis was carried out with the help of Dr Nicolas Guthertz, University of Leeds.

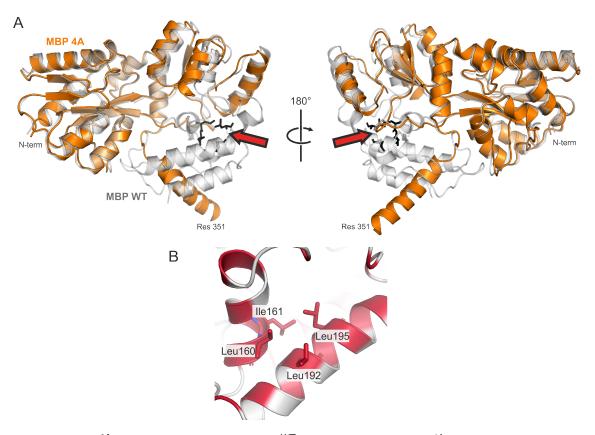


Fig. 3.8 **MBP**<sup>4A</sup> **crystal structure.** A) MBP<sup>WT</sup> (grey) aligned to MBP<sup>4A</sup> (orange). Positions of the original 4A mutations are shown by the red arrow as black sticks on the MBP<sup>WT</sup> structure. Residues 149-207 and 352-370 had no density in the MBP<sup>4A</sup> crystal structure. B) The positions of the original 4A mutations (L160, I161, L192, L195) shown on the MBP<sup>WT</sup> crystal structure. Mutating these residues to alanine may destabilise the protein by removing van der Waals interactions within the core, creating a cavity and potentially exposing an internal APR. Hydrophobic residues are coloured in red. Crystallographic data collection and analysis was carried out with the help of Dr Nicolas Guthertz, University of Leeds.

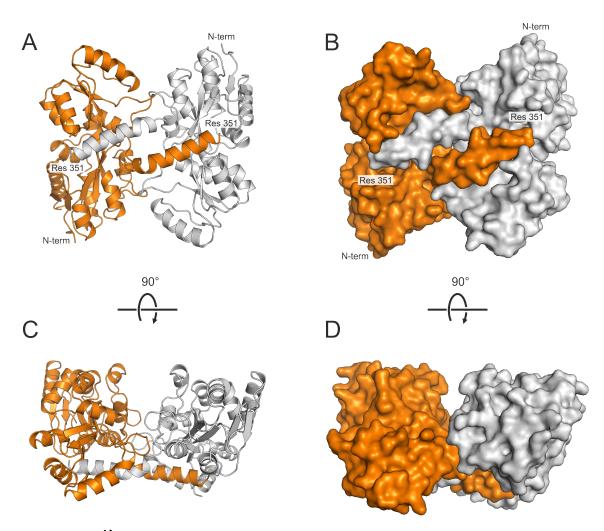


Fig. 3.9 **MBP<sup>4A</sup> forms a dimer in the crystal structure.** A) Ribbon and B) surface representation of the dimer. (C and D) Views similar to those in A and B after a 90 ° rotation. Residues 149-207 and 352-370 had no electron density in the crystal structure. Crystallographic data collection and analysis was carried out with the help of Dr Nicolas Guthertz, University of Leeds.

MBP_WT	MBP_Y283D	MBP_4A
I24	I24	I24
P 1 2 <sub>1</sub> 1	P 1 2 <sub>1</sub> 1	P 4 <sub>3</sub> 2 <sub>1</sub> 2
-	-	-
43.91 64.11 56.78	43.92 64.22 57.66	83.60 83.60 120.25
90.00 100.30 90.00	90.00 100.74 90.00	90.00 90.00 90.00
1.51 - 55.86	1.40 - 43.15	2.10 - 83.60
(1.51 – 1.55)	(1.40 - 1.42)	(2.10 - 2.14)
0.039 (0.668)	0.033 (0.287)	0.071 (3.788)
11.7 (1.2)	20.7 (6.8)	3.8 (0.1)
100.0 (99.9)	100.0 (100.0)	100.0 (99.6)
6.5 (5.7)	12.8 (11.3)	48.4 (40.0)
0.998 (0.316)	0.998 (0.930)	0.998 (0.131)
1.51 - 55.86	1.40 - 43.15	2.10 - 83.60
(1.51 – 1.55)	(1.40 - 1.42)	(2.10 - 2.14)
315012 (20544)	791206 (35101)	1241273 (49635)
48690 (3597)	62019 (3095)	25628 (1240)
0.17 / 0.20	0.16 / 0.19	0.22 / 0.27
2876	2874	2265
0	0	0
203	335	28
0.011	0.014	0.005
1.688	1.870	1.410
24.660	16.990	69.582
	$\begin{array}{c} 124 \\ P \ 1 \ 2_1 \ 1 \\ - \\ 43.91 \ 64.11 \ 56.78 \\ 90.00 \ 100.30 \ 90.00 \\ 1.51 \ - \ 55.86 \\ (1.51 \ - \ 1.55) \\ 0.039 \ (0.668) \\ 11.7 \ (1.2) \\ 100.0 \ (99.9) \\ 6.5 \ (5.7) \\ 0.998 \ (0.316) \\ \end{array}$	$124$ $124$ $P \ 1 \ 2_1 \ 1$ $P \ 1 \ 2_1 \ 1$ $P \ 1 \ 2_1 \ 1$ $  43.91 \ 64.11 \ 56.78$ $43.92 \ 64.22 \ 57.66$ $90.00 \ 100.30 \ 90.00$ $90.00 \ 100.74 \ 90.00$ $1.51 - 55.86$ $1.40 - 43.15$ $(1.51 - 1.55)$ $(1.40 - 1.42)$ $0.039 \ (0.668)$ $0.033 \ (0.287)$ $11.7 \ (1.2)$ $20.7 \ (6.8)$ $100.0 \ (99.9)$ $100.0 \ (100.0)$ $6.5 \ (5.7)$ $12.8 \ (11.3)$ $0.998 \ (0.316)$ $0.998 \ (0.930)$ $1.51 - 55.86$ $1.40 - 43.15$ $(1.51 - 1.55)$ $(1.40 - 1.42)$ $315012 \ (20544)$ $791206 \ (35101)$ $48690 \ (3597)$ $62019 \ (3095)$ $0.17 / 0.20$ $0.16 / 0.19$ $2876$ $2874$ $0$ $0$ $203$ $335$ $0.011$ $0.014$ $1.688$ $1.870$

Table 3.1 Crystallographic data collection and refinement statistics for MBP crystal structures. Values for the highest-resolution shell are shown in parentheses. Crystallographic data collection and analysis was carried out with the help of Dr Nicolas Guthertz, University of Leeds.

# **3.2.2** Golden gate assembly can be used to robustly create large libraries

Error-prone PCR (epPCR) is often the method of choice for creating random mutated plasmid libraries, however the rate-limiting step is often cloning of the epPCR insert into an appropriate vector for screening (Abou-Nader and Benedik, 2010). Here the highly efficient Golden Gate cloning is used to overcome this bottleneck.

First an appropriate vector was designed, blaGG<sub>STOP</sub>, whereby all the BsaI sites within the vector were removed and two new BsaI sites were introduced within the β-lactamase GS linker (Figure 3.10, Section 2.6.2). As BsaI is a Type IIs restriction enzyme it recognises non-palindromic sites and cleaves outside this site, the vector can be designed so that the final ligated product has no BsaI sites as they cut themselves out. Between the two BsaI sites in blaGG<sub>STOP</sub> a premature stop codon was introduced, as well as the 7bp Bsu36I restriction site. This prematurely ends translation as well as introduces a frame shift so that if this template is carried over into selection, only the first domain of  $\beta$ -lactamase is translated and so it cannot withstand the ampicillin selection. The test protein was amplified using epPCR then used in a second round of PCR to introduce corresponding BsaI sites onto the 5' and 3' ends. During initial testing of this method the epPCR step was replaced with traditional PCR to amplify wild-type MBP and clone it into blaGG<sub>STOP</sub> to create a 'test library'. This was to estimate the potential library size and to assess the viability of the method. Using a thermocycler to cycle between 37°C and 16°C for 1 minute each with 45 cycles the ligation reaction was able to go to completion, as after digestion of the test library with Bsu36I there was no product (Figure 3.11, lane 10 =undigested, lane 11 = digested).

This method using epPCR to was used to create libraries for MBP<sup>Y283D</sup> and MBP<sup>4A</sup> of  $2.4 \times 10^9$  and  $4.6 \times 10^{11}$  mutants, respectively, estimated by the number of colony forming units. A colony PCR was performed to confirm the error-prone PCR product was successfully cloned into the  $\beta$ -lactamase vector, and showed that of the 40 colonies assessed 100 % were successfully incorporated (Figure 3.12). The low cost of this technique (~£130 per library) and its ability to consistently produce large libraries makes it an attractive method for directed evolution studies.

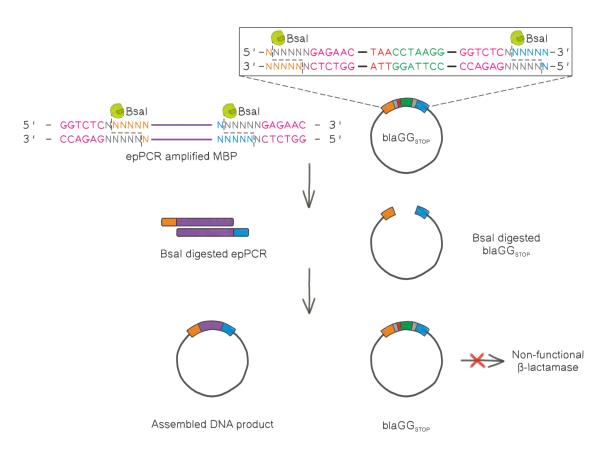


Fig. 3.10 Golden gate library preparation for directed evolution. epPCR of the gene of interest with 5' and 3' Bsal sites is combined with blaGG<sub>STOP</sub>, a variant of the  $\beta$ -lactamase vector for TPBLA with two Bsal sites introduced to allow cloning of the epPCR insert. blaGG<sub>STOP</sub> includes a premature stop codon and the 7 bp Bsu36I restriction site to prematurely stop translation and introduce a frame shift so that any template carried over into the library would produce a non-functional  $\beta$ -lactamase.

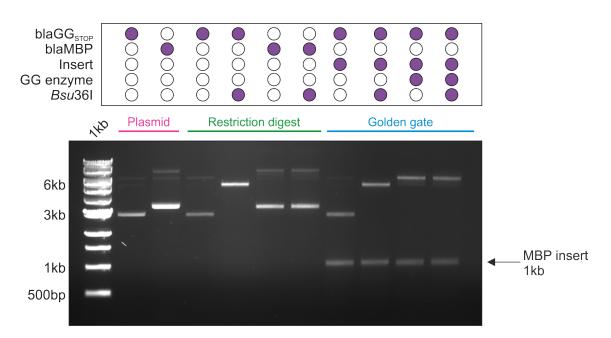


Fig. 3.11 1.5% (*w/v*) agarose gel showing restriction digestion and golden gate reactions to assess the success of the golden gate library method. PCR product of MBP ('Insert') is cloned into the library vector template ( $blaGG_{STOP}$ ) using golden gate.  $blaGG_{STOP}$  has an internal Bsu36l site that is replaced with the MBP insert (which would be an epPCR product when creating a library). A restriction digestion of the final golden gate reaction with Bsu36l shows that the golden gate reaction goes to completion and there is no identifiable  $blaGG_{STOP}$  left over. Filled in purple spots indicate presence of particular component in the reaction. Plasmids  $blaGG_{STOP}$  and blaMBP in wells 1 and 2 are shown as controls. Restriction digestions of these plasmids are shown to indicate that  $blaGG_{STOP}$  (the template used for the library reaction) is cleaved by Bsu36l whereas blaMBP (the product being created by golden gate) is not.

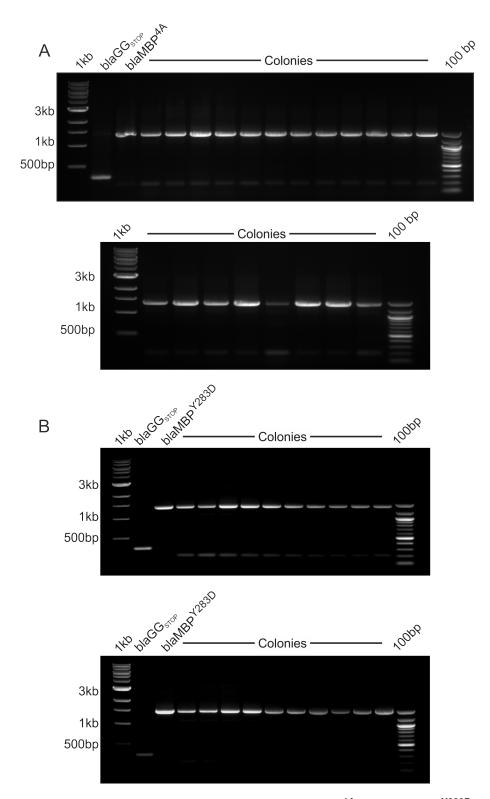
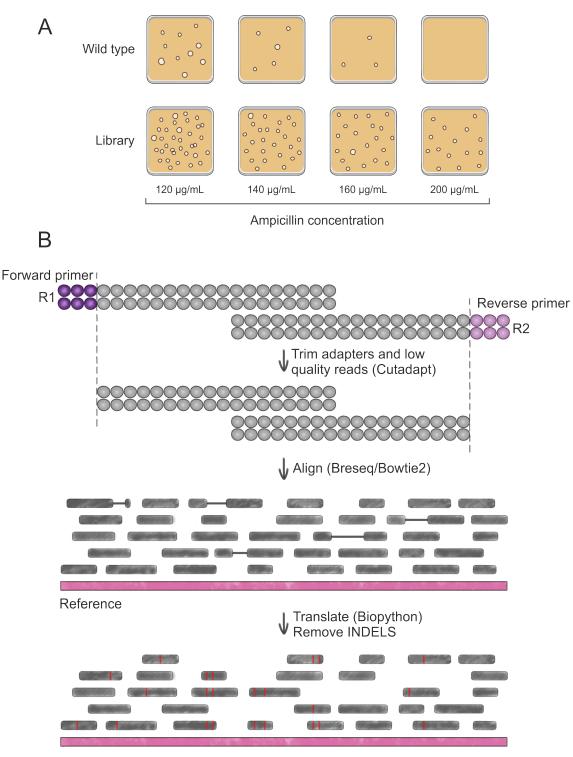


Fig. 3.12 1.5% (*w/v*) agarose gel colony PCR of blaMBP<sup>4A</sup> and blaMBP<sup>Y283D</sup> libraries. Colony PCR using primers that bind within the glycine-serine linker of the TPBLA construct to amplify the test protein. This performed to confirm the error-prone PCR product for (A) MBP<sup>4A</sup> or (B) MBP<sup>Y283D</sup> has been successfully cloned into the  $\beta$ -lactamase vector. A negative control (blaGG<sub>STOP</sub>) was included as a control.



Identify mutations Normalise by coverage

Fig. 3.13 **Overview of directed evolution and data analysis of Illumina sequencing.** A) Cells expressing the mutated library are grown on ampicillin concentrations the wild-type was unable to survive at. B) Successful mutants are analysed by 150 bp paired end Illumina sequencing. Fragments are filtered by trimming adapter sequences and low-quality reads (shorter than 40 bp and average Illumina quality score higher than 30 (Q30)). Fragments are aligned to the wild-type DNA sequence and the fragments in the resulting alignment file are filtered to remove indels before being translated in frame. The mutational frequency at each residue is calculated and normalised by coverage.

## **3.2.3** Illumina shotgun libraries allow identification of hotspots and single point mutations enriched due to selection

MBP<sup>4A</sup> and MBP<sup>Y283D</sup> were evolved using the *in vivo* assay by introducing genetic variation into the respective genes and creating a mutated plasmid library within the  $\beta$ -lactamase vector (Section 2.6.2) to produce  $\beta$ La MBP<sup>4A</sup>\* and  $\beta$ La MBP<sup>Y283D</sup>\*, respectively (Figure 3.10). For screening, the libraries were transformed into E. coli SCS1 cells and plated onto agar containing 280 µg/mL for  $\beta$ La MBP<sup>4A</sup>\* and 50 µg/mL for  $\beta$ La MBP<sup>Y283D</sup>\* (Figure 3.13A). At these concentrations, the 'wild-type' MBP<sup>Y283D</sup> and MBP<sup>4A</sup> sequences were unable to survive (refer to Figure 3.2D). Therefore, variants growing should have beneficial mutations that improve the expression of a folded and soluble fusion protein. The DNA from >700 colonies of MBP<sup>4A</sup> and >1500 colonies of MBP<sup>Y283D</sup> were pooled, purified, and the genes amplified using PCR before two technical repeats were sent for Illumina sequencing along with two technical repeats of the respective unselected (naive) libraries. Paired end fragments were aligned to a reference sequence (the respective 'wildtype' sequence) and the aligned fragments were translated in frame with respect to the reference sequence (Figure 3.13B). By comparing the aligned translated fragments to the original 'wild-type' sequence, mutational frequency at each position was calculated and normalised by read coverage (Figure 3.13B). The number of mapped bases aligned to MBP<sup>Y283D</sup> and MBP<sup>4A</sup> represented an average 353.079 read depth for MBP<sup>Y283D</sup> and 285,112 for MBP<sup>4A</sup>. The mutation frequency was normalised by coverage, and this was used to calculate the log<sub>2</sub>(fold change) at each residue. Hotspot residues were identified as having a  $\log_2(\text{fold change})$  above 10% of the maximum  $\log_2(\text{fold change})$  and being identified in both technical repeats.

For MBP<sup>Y283D</sup> a single obvious hotspot residue was identified, 283, which was reverted to the wild-type Y in 98.3% of cases (Figure 3.14A). Four additional hotspot residues were identified (G5, L7, R367, I368) at the N and C-termini (Figure 3.14B, Figure 3.14C). However, as a result of mapping the Illumina reads onto the reference sequence, the termini have low sequence coverage, meaning there were less reads mapped in these areas, resulting in higher than average scores after normalisation and so should be interpreted with caution.

For MBP<sup>4A</sup> multiple hotspot residues were identified located in and around the core of the second domain and the positions of the original 4A mutations. Three out of the four positions of the 4A mutations were identified as hotspots (A160, A192 and A195 were identified as hotspots; A161 was not), with the most common mutation being back to the wild-type residue validating the ability of the evolution assay to identify more stable

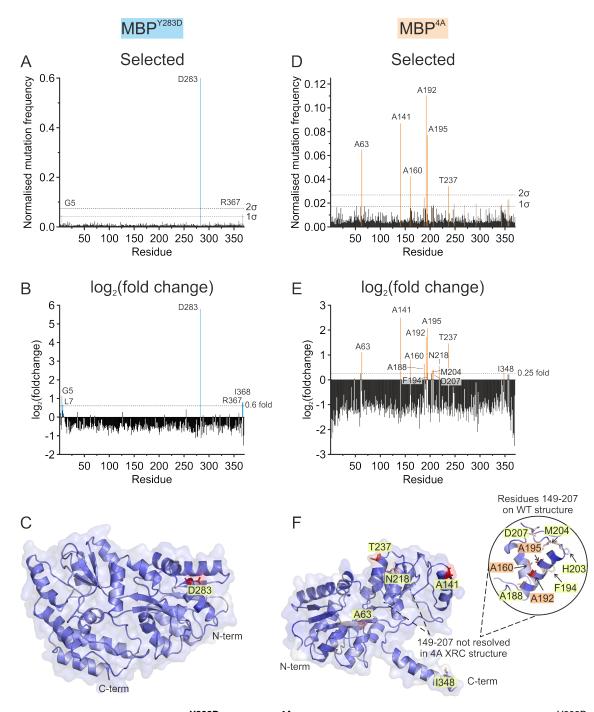


Fig. 3.14 **Evolution of MBP<sup>Y283D</sup> and MBP<sup>4A</sup> analysed by Illumina sequencing.** A) MBP<sup>Y283D</sup> selected library mutational frequency normalised by coverage. B) log<sub>2</sub>(fold change) of the mutational frequency calculated using the naive and selected libraries of MBP<sup>Y283D</sup>. C) hotspots mapped onto the structure of MBP<sup>Y283D</sup> crystal structure, D283 hotspot shown as sticks. D) MBP<sup>4A</sup> selected library mutational frequency normalised by coverage. E) log<sub>2</sub>(fold change) of the mutational frequency calculated using the naive and selected libraries of MBP<sup>4A</sup>. F) hotspots mapped onto the structure of MBP<sup>4A</sup> crystal structure, hotspots shown as sticks. Residues 149-207 and 352-370 had no electron density in the crystal structure, so hotspots are shown mapped onto the MBP<sup>WT</sup> crystal structure. Three of the original 4A mutations that were identified as hotspots are labelled with an orange background, the other residue (A161) was not identified as a hotspot. Sequencing was repeated and showed similar results.

variants (Figure 3.14D). Eight residues were identified as hotspots that form a cluster within the core of this second domain (A160, A188, A192, F194, A195, M204, D207, I348) (Figure 3.14E, Figure 3.14F). Five of these identified hotspots were most commonly mutated to bulkier side chains (A160L, A192L, A195L, I348L), potentially to re-fill the cavity initially introduced by the original 4A mutations and to restore the stabilising Van der Waals forces. To understand the extent by which TPBLA has evolved MBP<sup>4A</sup>, the most frequent residue identified at each of the hotspots was introduced back into MBP4A as a single point mutation and assessed using TPBLA. 11 out of 13 variants showed improved growth compared with MBP<sup>4A</sup>, with one variant (A63T) having a 2.2-fold increase in *in* vivo growth (Figure 3.15). Two variants, D207V and M204V, showed a reduction in in vivo growth compared with MBP<sup>WT</sup> (Figure 3.15). These two hotspots were identified at a low frequency compared with the other hotspots, and had a  $\log_2(\text{fold change})$  close to the 0.25 fold change threshold. It is possible these mutations were carried over with other more beneficial mutations, perhaps picked up early in an epPCR cycle. Together MBP<sup>Y283D</sup> and MBP<sup>4A</sup> evolution analysis using Illumina sequencing demonstrate the ability of TPBLA to assess thousands of sequences in a high-throughput manner and rapidly identify beneficial mutations.

The short fragment length of Illumina sequencing limits this method in assessing co-evolution and whether multiple mutations are selected for in combination. Therefore, libraries were subsequently analysed using long read NGS technique Pacbio.

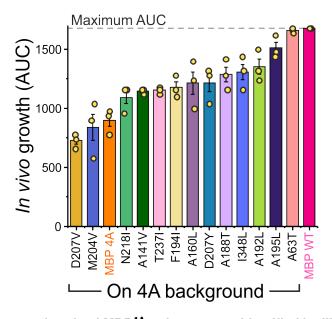


Fig. 3.15 *In vivo* screen of evolved MBP<sup>4A</sup> point mutants identified by Illumina sequencing. Area under the antibiotic survival curve calculated for evolved MBP<sup>4A</sup> point mutants identified by Illumina sequencing, screened at 0-280  $\mu$ g/mL ampicillin. Error bars show standard error of the mean (S.E.M) from three independent experiments.

### **3.2.4** Pacbio sequencing allows analysis of co-evolution and identification of sequences with enhanced properties

The naive and selected libraries from the evolution of  $MBP^{4A}$  and  $MBP^{Y283D}$  were amplified using PCR and sent for Pacbio sequencing (Section 2.7.1). As with the Illumina sequencing, the naive and selected libraries were used to calculate the  $log_2$ (fold change) of mutation frequency at each residue.

For both MBP<sup>Y283D</sup> and MBP<sup>4A</sup>, Pacbio sequencing identified the same hotspot residues as Illumina sequencing by looking at the log<sub>2</sub>(fold change) (Figure 3.16). The additional information Pacbio provides is the ability to distinguish whether mutations are found alone or in combination with others. For MBP<sup>Y283D</sup> in the majority of cases the reversal to wild-type mutation D283Y occurred as a single point mutation (Table 3.2).

Mutation	Count
D283Y	4449
P123S D283Y	85
D283Y A360V	80
P254S D283Y	68
G16D D283Y	63
A71V D283Y	59
D283Y P331S	55
D283Y R367S	52
P159S D283Y	52
A63T D283Y	51
K1Q D283Y	47
D283Y A342V	47
A84V D283Y	46
P248S D283Y	44
G101D D283Y	43
A71T D283Y	41
G174D D283Y	40
G5A D283Y	39
G56D D283Y	38
D283Y A364T	38

 $Table \ 3.2$  Top 20 single- and double-point mutation counts identified in MBP^{Y283D} evolution by Pacbio sequencing

For MBP<sup>4A</sup> the same hotspot residues that were identified in our Illumina dataset were identified using Pacbio sequencing, demonstrating the power of Illumina sequencing at unpicking large mutational datasets at a fraction of the cost of Pacbio (Figure 3.16B). Furthermore, out of the top 100 mutations identified in Pacbio sequencing, 84 % were single point mutations. This therefore demonstrates the relevance of the Illumina sequencing methodology for identifying beneficial mutations.

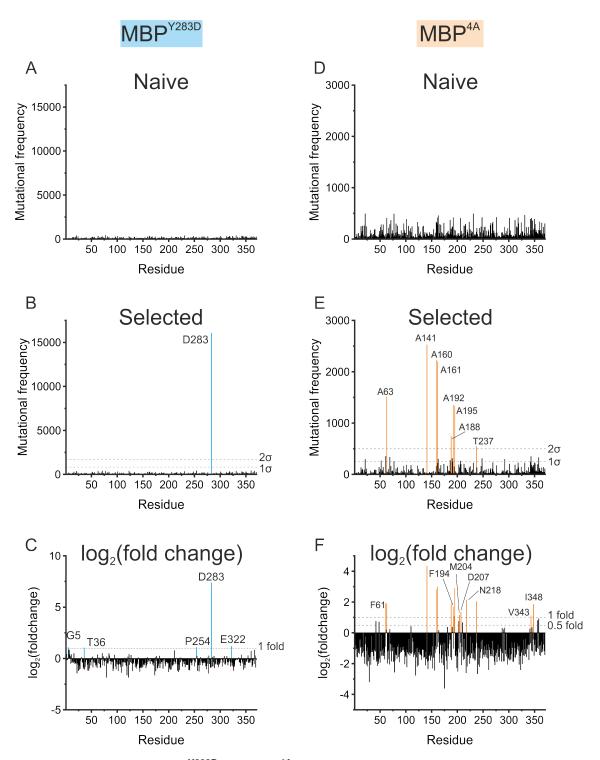


Fig. 3.16 Evolution of MBP<sup>Y283D</sup> and MBP<sup>4A</sup>analysed by Pacbio sequencing. Naive and selected library mutational frequencies, and log<sub>2</sub>(fold change) of the mutational frequency calculated using the naive and selected libraries for (A) MBP<sup>Y283D</sup> and (B) MBP<sup>4A</sup>.

Top single mutations		Top doub	Top double mutations	
Mutation	Count	Mutation	Count	
A141V	1265	A160L A161I	287	
A63T	384	A63T A190S	67	
A188T	336	G69D I348L	37	
F61I	150	H203Y D207V	34	
D358N	136	A63T G191S	33	
V343G	124	A192L A195L	30	
G69D	106	A63T A141V	29	
T237I	89	A52T A63S	27	
A269T	81	A63T V246A	26	
V110I	77	A63T A206V	23	
P248S	69	A63T A292V	22	
A21T	64	A63V A160T	22	
A192V	64	A292V T356S	18	
A163T	63	G69D A141V	18	
V181M	53	T225I V347E	18	
A231T	53	A141V A269T	15	
P331S	53	W62L A192V	14	
G101D	49	A141V T237I	14	
А77Т	49	A186T F194I	14	
I348L	48	A63T F194I	14	

Table 3.3 Top 20 single- and double-point mutation counts identified in MBP <sup>4A</sup> evolution	
by Pacbio sequencing	

Multiple double mutations were identified, with the most frequent variants detailed in Table 3.3. Furthermore, triple and quadruple mutations were identified containing some or all of the original 4A residues mutated back to the wild-type. During library generation, the epPCR was tuned to give an average of one amino acid mutation per gene. Due to this low mutation rate, it is extremely unlikely all of these reversal to wild-type residues came about via epPCR. The most likely explanation is that these are a result of homologous recombination with the E. coli genomic copy of MBP, which has 99 % nucleotide similarity to our MBP sequence. Consistent with this hypothesis, the revertants also contained two silent nucleotide substitutions that matched those in E. coli K12 (G60C and A828G, nucleotide number). Therefore, reversal to wild-type mutations (any combination of single, double, triple or quadruple) at nucleotide or residue level were omitted from further investigations. The top 20 single and double substitutions identified from the MBP<sup>4A</sup> Pacbio sequencing are detailed in Table 3.3. The top 5 single (F61I, A63T, A141V, A188T, D358N) and double (A63T+A141V, A63T+A190S, A63T+G191S, G69D+I348L, H203Y+D207V) substitutions were introduced back into the original MBP<sup>4A</sup> sequence for further analysis. All of the mutations displayed improved growth in TPBLA when compared with MBP<sup>4A</sup> (Figure 3.17A). Again, the best performing variant was A63T as a single point mutation. When A63T was combined with A141V, A190S or G191S as a double point mutant the *in vivo* growth score was lower than A63T as a single point

mutation. The fact that the combined effect of two beneficial mutations (A63T and A141V) is not additive demonstrates the complexity of protein engineering and the inter-residue interactions governing protein stability and solubility. Interestingly, double mutation H203Y+D207V was identified in Pacbio sequencing. These mutations had been identified as single point mutants in Illumina sequencing, and D207V was one of the mutations that resulted in a decrease in *in vivo* growth score in TPBLA. However, the double variant H203Y+D207V shows an increase in *in vivo* growth score in TPBLA that is significantly higher than either H203Y or D207V alone, indicating these residues have been co-evolved to improve stability and/or solubility together. This demonstrates the power of TPBLA for co-evolving mutations that could otherwise be more complex to rationally design.

To understand how TPBLA is evolving these proteins, seven variants (F61I, A63T, A141V, A188T, I348L, D358N, H203Y+D207V) were expressed and purified for further characterisation (Figure 3.18). Binding of the evolved MBP<sup>4A</sup> variants to ANS was measured to identify any conformational changes that might reduce the exposure of hydrophobic patches on the protein. All variants except H203Y+D207V displayed reduced binding to the ANS probe (Figure 3.17B), indicating a reduction in exposed hydrophobic side chains on the surface, or a stabilisation of the core preventing ANS from getting in and binding. This could result in a reduction of aggregation by preventing hydrophobic interactions. H203Y+D207V showed significantly increased binding to the ANS probe, indicating an increase in exposed hydrophobics on the surface or a further destabilisation of the core thereby enabling ANS to get in and bind to hydrophobic residues in the core.

Interestingly, the majority of variants displayed similar thermal melting midpoint (Tm) to those of MBP<sup>4A</sup> (Figure 3.17C, Table 3.4, 45 °C, assessed by circular dichroism, CD). However, only two variants showed an increase in Tm compared with MBP<sup>4A</sup> (F61I, 47 °C; A63T, 48 °C). To assess the effect of these evolved mutations on the thermodynamic stability of MBP<sup>4A</sup>, urea denaturation curves for all variants were measured. As the pre- and post- transitions from the urea denaturation curves have low signal to noise, the data was used to calculate the midpoint denaturant concentration (Cm) rather than  $\Delta$ G (Figure 3.17D, Table 3.4).  $\Delta$ G is very sensitive to changes in the pre- and post- transitions, therefore any calculation of  $\Delta$ G from this data would be inaccurate. As with Tm, only A63T showed a significant increase in midpoint denaturant concentration (Cm) compared with MBP<sup>4A</sup> (Figure 3.17D, Table 3.4). When compared with *in vivo* growth score, Tm and ANS binding showed poor correlation (Spearmans rank = 0.17 and -0.21, respectively), whereas Cm showed moderate correlation (Spearmans rank = 0.47; Figure 3.19A-C).

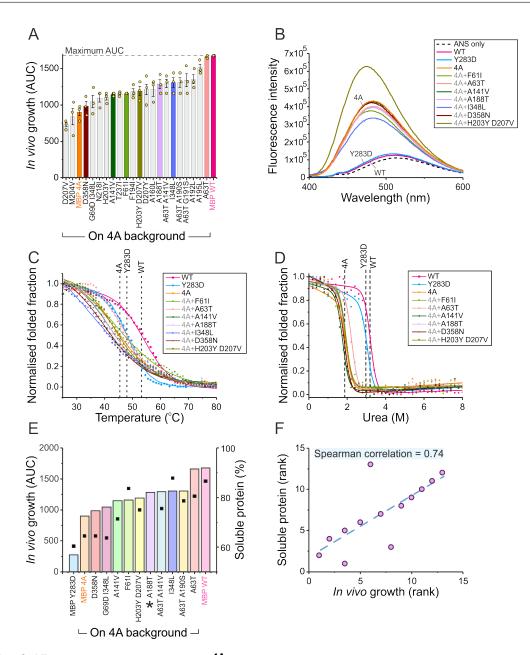


Fig. 3.17 Analysis of evolved MBP<sup>4A</sup> point mutants identified by Illumina and Pacbio sequencing. A) Area under the antibiotic survival curve calculated for evolved MBP<sup>4A</sup> variants identified by Illumina and Pacbio sequencing, screened at 0-280 µg/mL ampicillin. The most freguent amino acid identified at the Illumina hotspots are shown, alongside the top 5 single- and double- point mutants from Pacbio. Error bars show standard error of the mean (S.E.M) from three independent experiments. Highlighted variants (solid coloured bars) were expressed and purified for further analysis. B) 8-Anilinonapthalene-1-sulphonic acid (ANS) fluorescence spectroscopy shows extent of exposed hydrophobic patches on the protein's surface, or a 'loosening' of the core structure allowing ANS in to bind. 1 µM protein and 100 µM ANS final concentration in 10 mM potassium phosphate (pH 7.4). C) Circular dichroism thermal melt (temperature range 25 °C to 80 °C) of MBP<sup>WT</sup>, MBP<sup>Y283D</sup>, MBP<sup>4A</sup> and MBP<sup>4A</sup> evolved variants (protein concentration 0.15 mg/mL) in 10 mM potassium phosphate (pH 7.4). Temperature ramps were fitted to a sigmoidal curve. Dashed lines show Tm for MBP<sup>WT</sup>, MBP<sup>Y283D</sup> and MBP<sup>4A</sup>. D) Urea denaturation (denaturant range 0-8 M urea) of MBPWT, MBP<sup>Y283D</sup>, MBP<sup>4A</sup> and MBP<sup>4A</sup> evolved variants (protein concentration 10 µM) in 10 mM potassium phosphate (pH 7.4). Data points were fitted to a two state unfolding curve using IgorPro 7. Dashed lines show Cm for MBPWT, MBPY283D and MBP<sup>4A</sup>. E) In vivo growth (AUC) plotted against the percentage of soluble protein expression of β-lactamase fusions. Soluble protein expression was calculated from densitometry analysis of a dot blot (see Methods). A188T showed low soluble expression (0.04 %) and is denoted by an asterix. F) In vivo growth rank plotted against the soluble protein expression rank of MBP4A  $\beta$ -lactamase fusions. Spearmans rank correlation = 0.74.

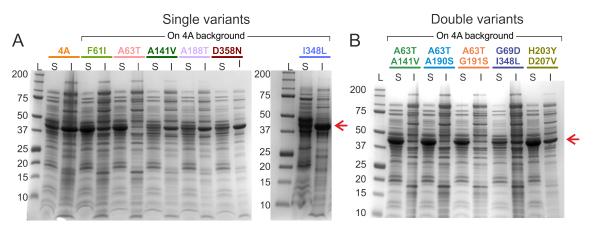


Fig. 3.18 Expression trial of evolved MBP<sup>4A</sup> selected point mutants identified by Pacbio sequencing. Soluble (S) and insoluble (I) protein fractions of MBP<sup>4A</sup> point mutants after 24 hrs growth in autoinduction media. A) Single point mutants and B) double point mutants as identified by Pacbio sequencing.

Table 3.4 **Thermal and thermodynamic stabilities of MBP variants.** Transition mid-point temperatures (Tm) and midpoint denaturant concentration (Cm) calculated using circular dichroism thermal melt and urea denaturation, respectively.

Variant	Tm (°C)	Cm (M)
WT	53.20	3.19
Y283D	47.97	2.97
4A	45.45	1.87
4A+F61I	47.03	1.93
4A+A63T	47.96	2.20
4A+A141V	43.71	1.73
4A+A188T	43.79	1.89
4A+I348L	41.63	N/A
4A+D358N	42.37	1.81
4A+H203Y D207V	45.03	1.92

To assess the affect of these mutations on the *in vivo* solubility of MBP<sup>4A</sup>, a dot blot against the variants as  $\beta$ -lactamase fusions was performed to calculate the fraction of soluble protein expressed (Table 3.5, Figure 3.20). Unlike the measures of ANS binding, Tm, and Cm, the soluble protein fraction showed strong correlation with *in vivo* growth score (Spearmans rank correlation = 0.74; Figure 3.17E-F, Figure 3.19D) and all the evolved variants assessed except A188T displayed elevated levels of soluble protein expressed compared with MBP<sup>4A</sup>.

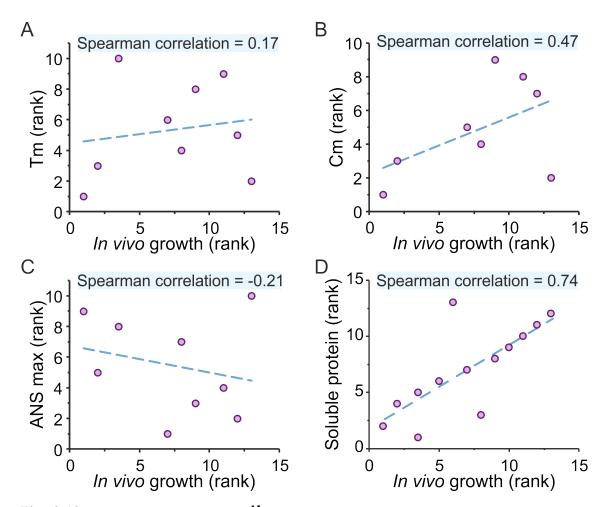


Fig. 3.19 Analysis of evolved MBP<sup>4A</sup> selected point mutants identified by Illumina and Pacbio sequencing. *In vivo* growth of MBP<sup>WT</sup>, MBP<sup>Y283D</sup>, MBP<sup>4A</sup> and evolved variants compared with Tm, Cm, ANS binding and soluble protein expression. *In vivo* growth (AUC) plotted against A) thermal melting midpoint (Tm) as calculated by circular dichroism, B) midpoint denaturant concentration (Cm) as calculated from urea denaturation, C) normalised max fluorescence excitation from 8-Anilinonapthalene-1-sulphonic acid (ANS) fluorescence spectroscopy and D) soluble protein expression calculated using densitometry of a dot blot.

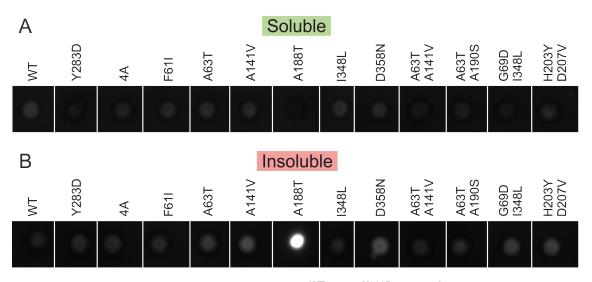


Fig. 3.20 Dot blot against  $\beta$ -lactamase for MBP<sup>WT</sup>, MBP<sup>Y283D</sup>, MBP<sup>4A</sup> and evolved variants. Soluble (A) and insoluble (B) protein expression of  $\beta$ -lactamase fusion MBP variants. Left to right: MBP<sup>WT</sup>, MBP<sup>Y283D</sup>, MBP<sup>4A</sup>, MBP<sup>4A F61I</sup>, MBP<sup>4A A63T</sup>, MBP<sup>4A A141V</sup>, MBP<sup>4A A188T</sup>, MBP<sup>4A I348L</sup>, MBP<sup>4A D358N</sup>, MBP<sup>4A A63T + A141V</sup>, MBP<sup>4A A63T + A190S</sup>, MBP<sup>4A G69D + I348L</sup>, MBP<sup>4A H203Y + D207V</sup>.

Table 3.5 Soluble and insoluble protein expression of  $\beta$ -lactamase fusions with MBP variants as the test protein. Densitometry analysis of dot blot showing soluble and insoluble protein expression for MBP variants. Intensities analysed using ImageJ.

Variant	Soluble	Insoluble	Fraction soluble (%)
WT	4972.569	790.87	86.3
Y283D	1781.284	1305.698	57.7
4A	3517.113	2134.527	62.2
4A+F61I	4114.527	872.87	82.5
4A+A63T	4938.941	1294.406	79.2
4A+A141V	4054.82	1653.234	71.0
4A+A188T	359.062	8023.761	0.04
4A+I348L	3601.991	554.87	86.7
4A+D358N	4736.234	2681.355	63.9
4A+A63T A141V	2776.326	942.456	74.7
4A+A63T A190S	3491.113	970.042	78.3
4A+G69D I348L	2410.284	1351.113	64.1
4A+H203Y D207V	4149.134	1446.87	74.1

### 3.3 Discussion

We demonstrate how the TPBLA can be used to assess different sequence liabilities that might otherwise limit their use in industrial and therapeutic applications. Protein stability is of vital importance for such applications, for example as industrial enzymes are often required to be stable in extreme environments (Littlechild, 2015). Furthermore, protein instability may lead to protein aggregation, another undesirable property in a biopharmaceutical, for example, where it may lead to reduced activity or even immunogenicity and anti-drug antibodies produced in the patient, as well as result in difficulties in manufacture (Jiskoot et al., 2012). As the TPBLA does not use a perturbant to accelerate aggregation, such as increased temperature, pH or addition of a chemical denaturant, it probes the innate aggregation propensity of the test protein which is most likely to reflect the behaviour of a biopharmaceutical during manufacture. This has been demonstrated by previous work correlating performance of various scFv fragments in TPBLA with their aggregation behaviour as full length IgGs (Ebo et al., 2020b). Additionally, the complex nature of aggregation and the variety of different mechanisms it may occur by make it difficult to predict by in silico methods as not one single mechanism drives aggregation (Ebo et al., 2020a). Therefore, the TPBLA represents an attractive alternative for assessing the innate aggregation propensity stability of a protein, and in the case of a biopharmaceutical to flag up any potential developability issues.

TPBLA has previously been used to assess protein thermodynamic stability (Foit et al., 2009) and aggregation propensity (Ebo et al., 2020b), as well as to identify small molecule inhibitors of amyloid formation (Saunders et al., 2016). It has also been used as a directed evolution screen to increase the thermodynamic stability of Im7 (Foit et al., 2009), to reduce the aggregation propensity of aggregation-prone IgG WFL (Ebo et al., 2020b), and through directed evolution to understand the individual residue contribution to aggregation propensity and amyloidogenicity within  $\beta$ -2-microglobulin ( $\beta_2$ m) (Guthertz et al., 2022). Therefore, TPBLA has been widely and successfully utilised to assess and evolve distinct behaviours in different test proteins - thermodynamic stability in Im7, aggregation resistance in WFL, and reduced amyloidogenicity in  $\beta_2$ m. In this chapter we have demonstrated the relationship between TPBLA growth score and solubility for variants of maltose binding protein. Together this demonstrates the complexity of TPBLA, and that the driver for directed evolution is generally protein dependent.

*In vitro* analysis of MBP<sup>WT</sup>, MBP<sup>Y283D</sup> and MBP<sup>4A</sup> indicate MBP<sup>4A</sup> might perform worse in TPBLA. During purification the protein is expressed almost exclusively in the insoluble fraction and needs to be refolded. Furthermore, the presence of high molecular

weight species in SEC, low thermodynamic stability, ability to bind ANS and unresolved patches in the crystal structure indicate that MBP<sup>4A</sup> may have the potential to aggregate from both the native state and via partially unfolding. This could be initiated by the lack of a packed stable core that may increase the dynamic fluctuations resulting in exposure of otherwise buried aggregation-prone regions (APRs) (Eyes et al., 2019; Kuriata et al., 2019; Maas et al., 2007). MBP<sup>Y283D</sup> has a structure very similar to MBP<sup>WT</sup>, whereas MBP<sup>4A</sup> has a less structured and more dynamic second domain when fully folded which has the potential to expose the highly hydrophobic and aggregation-prone core. In the case of MBP<sup>Y283D</sup>, it is known this variant is slow folding compared with MBP<sup>WT</sup> (Chun et al., 1993). MBP<sup>Y283D</sup> loses a polar interaction the wild-type residue makes with Asp30, potentially exposing an otherwise buried hydrophobic APRs that could lead to aggregation during folding (Figure 3.7), resulting in insoluble aggregates that would have been separated out before column purification and so would not be identified as high molecular weight species in SEC. However, once the native state is achieved the potential to aggregate from this fold would theoretically be lower than MBP<sup>4A</sup>, if the mechanism whereby MBP<sup>Y283D</sup> aggregates is not via the native state. Together these demonstrate how the TBPLA is able to assess aggregation from folding intermediates as well as the intrinsic aggregation propensity of the folded state without the addition of any aggregation accelerant. In addition, we demonstrate how TBPLA can rapidly identify developability issues in a high-throughput manner when the alternative would be using a multitude of additional in vitro techniques in combination, an extremely low-throughput and laborious process.

We present here a robust and reproducible method for the high-throughput evolution of protein solubility and aggregation resistance. Our library generation method consistently creates large libraries, overcoming the common bottleneck in random mutated library generation of cloning the epPCR fragment into the selection vector (Alejaldre et al., 2021; Pai et al., 2012). Previous work using TPBLA for protein evolution used the megaprimer method to create libraries of 10<sup>4</sup> - 10<sup>6</sup> variants (Ebo et al., 2020b). In comparison, our simple and robust method was able to make libraries for MBP<sup>Y283D</sup> and MBP<sup>4A</sup> with an estimated 2.4 x 10<sup>9</sup> and 4.6 x 10<sup>11</sup> mutants, respectively. MBP<sup>Y283D</sup> introduced a folding defect resulting in protein aggregation and using TPBLA to evolve this variant we identified a reversion back to the wild-type sequence to resolve this folding defect. MBP<sup>4A</sup> differed from MBP<sup>Y283D</sup> at five residues and had a reduced thermodynamic stability (Figure 3.17D). Evolving MBP<sup>4A</sup> using TPBLA we identified 13 hotspot residues using Illumina sequencing and 15 using Pacbio sequencing, none of which overlapped with hotspots identified in MBP<sup>Y283D</sup> evolution. Assessing the variants identified in our evolution using TPBLA we showed most mutants displayed enhanced *in vivo* growth (3.17A).

Furthermore, most of these variants displayed reduced binding to the ANS probe, indicating a reduction in surface exposed hydrophobic side chains or potentially a stabilisation of the core to prevent ANS from penetrating and binding (Figure 3.17B). This reduction in surface exposed hydrophobic side chains could result in reduced aggregation by preventing hydrophobic interactions. Only F61I and A63T displayed significant improvement in Tm or Cm compared with MBP<sup>4A</sup>. Interestingly, residue A63 is involved in binding maltose (Quiocho et al., 1997). The selection of this residue in our assay could point to a stability-function trade off in MBP. It is commonly observed that regions evolved for function can be more aggregation prone than other solvent exposed regions, and the absence of a selection pressure to bind maltose could have exposed this residue as a region of frustration. Furthermore, the evolved variants displayed higher levels of soluble protein expression compared with MBP<sup>4A</sup>. Soluble protein expression also correlated well with *in vivo* growth (Figure 3.17E, F; Spearmans rank correlation = 0.74), demonstrating the ability of our assay to both assess and evolve protein solubility and therefore *in vivo* aggregation resistance.

Hotspots were initially identified using Illumina sequencing, a fragment-based 'sequence by synthesis' method yielding massive datasets and allowing sequencing of potentially thousands of variants at once. However, the fragment length limits this methods' ability to assess co-evolution and the presence of multiple mutations. Therefore, we sought to combine our screen with Pacbio sequencing to read the entire sequence in one run and identify full-length enriched sequences. However, we showed our evolved library was made up of mostly single point mutants (in the 4A evolved library, 84% of the top 100 mutants were single), demonstrating Illumina short read sequencing was sufficient for accurately assessing mutational profiles within our libraries.

In summary, we demonstrate here a high-throughput methodology to assess and evolve protein solubility. Previous work using TPBLA to evolve aggregation resistance in antibody fragments allowed identification of improved variants (Ebo et al., 2020b). However, the method was labour intensive as it involved manual picking of colonies, low-throughput Sanger sequencing, and manual identification of mutated variants. Our next-generation sequencing methodology improves upon this by using high-throughput next-generation sequencing to identify mutational hotspots, and by automating data analysis and mutant identification using custom scripts. Our improved assay has the ability to rapidly probe thousands to millions of variants of a protein to select for those with increased *in vivo* solubility, making it a powerful tool for assessing and evolving beneficial biophysical properties in a protein of interest. This by extension could be useful for assessing and

evolving biopharmaceutical developability or understanding a protein's mechanism of aggregation.

### **Chapter 4**

## Applying TPBLA to assess and evolve therapeutically relevant proteins

### 4.1 Introduction

In Chapter 3 we developed a high-throughput directed evolution methodology utilising the power of TPBLA combined with next-generation sequencing to engineer improved biophysical characteristics in proteins. This method could be utilised for proteins of interest to biotechnology or biopharmaceuticals in order to create highly developable candidates.

The developability of a biopharmaceutical is its potential to pass through the development process. This is influenced by its biophysical and physicochemical properties including thermal stability, colloidal stability, aggregation propensity, binding affinity, and polyspecificity (Bailly et al., 2020). These properties can be assessed and characterised using a range of 'developability' assays, and improved via protein engineering techniques. Often these developability assays only assess one single property at a time, and require purified protein to do so (Jain et al., 2017). TPBLA could potentially be a powerful technique to assess the developability of a panel of biotherapeutics as (1) it is influenced by multiple properties (thermal stability, aggregation propensity, solubility) that would negatively impact the test protein's developability, and (2) it can be performed in high-throughput without the need for purified protein.

### 4.1.1 Aims of this chapter

The aim of this chapter is to demonstrate the broad applicability of the TPBLA to assess therapeutically relevant proteins. As TPBLA is able to report on a protein's aggregation propensity, solubility, and stability, it could easily be employed as a developability screen early in a biopharmaceuticals developmental pipeline to screen out unviable candidates quickly and easily (Foit et al., 2009; Saunders et al., 2016; Ebo et al., 2020b). In this chapter, we apply TPBLA to various antibody therapeutics to assess their behaviour and correlate TPBLA to their biophysical properties to understand better what the assay is screening for. Secondly, we apply our directed evolution methodology (Chapter 3) to improve the biophysical properties of two selected variants. We then characterise these evolved variants to investigate the extent by which the TPBLA has improved their biophysical properties *in vitro*. Finally, we assess the potential for using TPBLA as a developability screen using the scFv sequences from 35 clinically relevant mAbs, correlating their performance in TPBLA with various currently employed biophysical assays.

### 4.2 Results

#### 4.2.1 TPBLA can be used to screen and rank biotherapeutics

As described in Section 4.1, traditional individual characterisation assays for biopharmaceutials tend to assess a single particular property; e.g. differential scanning fluorimetry (DSF) measures thermal stability, Hydrophobic Interaction Chromatography (HIC) measures hydrophobicity, Affinity-Capture Self-Interaction Nanoparticle Spectroscopy (AC-SINS) measures self-interaction, and PEG precipitation measures solubility (Gibson et al., 2011; Jain et al., 2017). It has been shown that TPBLA can be correlated to a wide range of properties depending on the test protein; thermodynamic stability (Foit et al., 2009), amyloidogenicity (Saunders et al., 2016), mAb aggregation propensity (Ebo et al., 2020b), and soluble protein expression (Chapter 3). It is likely this is because the assay is influenced by multiple factors that can affect the POI's stability and folding (e.g thermal stability, conformational stability, aggregation, colloidal stability, solubility), meaning each individual property will have a different level of impact depending on how strongly it affects the specific test protein.

11 mAbs were provided by UCB Biopharma UK taken from a larger dataset of IgGs designed as a well-characterised set of antibodies whose use is not restricted by intellectual

property (IP) restrictions with a range of sequence liabilities (e.g. methionine oxidation, deamidation). Out of this larger dataset, only some were UCB molecules and therefore had a common IgG1 scaffold. These molecules had been characterised and classed based on their aggregation as having 'high', 'medium' or 'low' aggregation. 3 molecules were chosen from each category, making a total of 9 IgGs. 2 additional approved mAbs were included as controls in this dataset, AMS106 and AMS122. These are the VH and VL domains of infliximab (a known aggregation-prone, highly immunogenic mAb) and trastuzumab (a known aggregation-resistant, low immunogenic mAb), respectively (Jain et al., 2017; Kurki et al., 2021; Mosch and Guchelaar, 2022), grafted onto a common IgG1 scaffold which was shared by all the other 9 mAbs. The original dataset was part of the £11.2 million BioStreamline project, a collaborative project between six industrial partners (Lonza Biologics, UCB Biopharma UK, Sphere Fluidics, Horizon Discovery, Alcyomics Ltd, and CPI) and supported by funding from the UK Government's Advanced Manufacturing Supply Chain Initiative (AMSCI), designed to study the affects of these sequence liabilities on mAb developability (Centre for Process Innovation, 2022). From here on, the 11 chosen mAbs will be referred to as the AMSCI mAbs.

The IgG format of the AMSCI mAbs were assessed using a suite of biophysical characterisation assays (Figure 4.1). They have similar thermal stabilities as measured by differential scanning fluorimetry (DSF), which assesses protein unfolding by measuring intrinsic tryptophan fluorescence (Figure 4.2, Figure 4.3, Table 4.1). Transition midpoint temperature (Tm) and temperature onset of aggregation (Tonset) was calculated using the first derivative (Appendix B, see Methods Section 2.4.8). Most have the first transition mid-point temperature (Tm1) of around 68-70 °C, except for AMS134 (Tm = 65.11 °C), AMS148 (Tm = 60.00 °C), and AMS155 (Tm = 62.45 °C). The highest Tm was that of AMS122 (trastuzumab) at 70.70 °C. The increase in static light scattering while increasing temperature was used to measure the temperature onset of aggregation (Tonset) of the AMSCI mAbs. This combined with the DSF measurements can be used to identify whether the POI is aggregating from the native or the unfolded state. Like Tm, the mAbs had similar Tonset with an average of 67.23 °C (standard deviation = 8.93) (Figure 4.2, Figure 4.3, Table 4.1). The highest Tonset was AMS122 (Tonset = 79.80 °C), followed by AMS197 (Tonset = 79.32 °C), and the lowest was AMS155 (Tonset = 54.48 °C). The scFv segment of the AMSCI mAbs was introduced into the TPBLA construct to assess using the assay (Figure 4.1). By contrast to the lack of differentiation using thermal stability, when analysed using TPBLA across an ampicillin concentration range of 0-140 µg/mL, the AMSCI mAbs show a wide range of responses (Figure 4.4). TPBLA growth score shows no correlation with Tm by DSF alone. Likely this ranking is due to a combination of unwanted biophysical properties (protein stability, solubility, and aggregation propensity), enabling TPBLA to identify unviable candidates based on their performance.

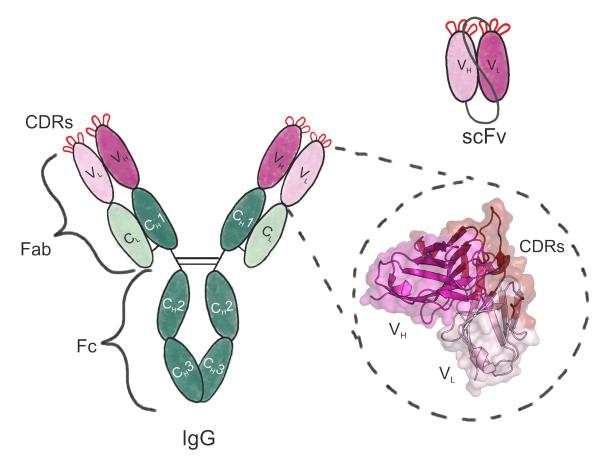
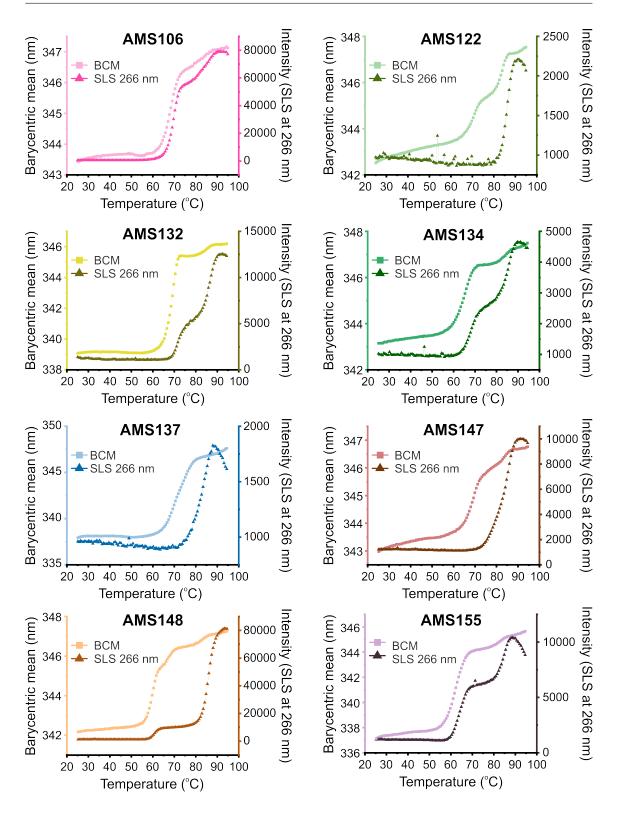


Fig. 4.1 Structures of mAbs and antibody fragments. Biophysical characterisation was carried on full-length mAbs, or IgGs, whereas analysis using TPBLA used the variable domains (V<sub>H</sub> and V<sub>L</sub>) joined by a glycine-serine rich linker to create a single-chain variable fragment, or scFv.



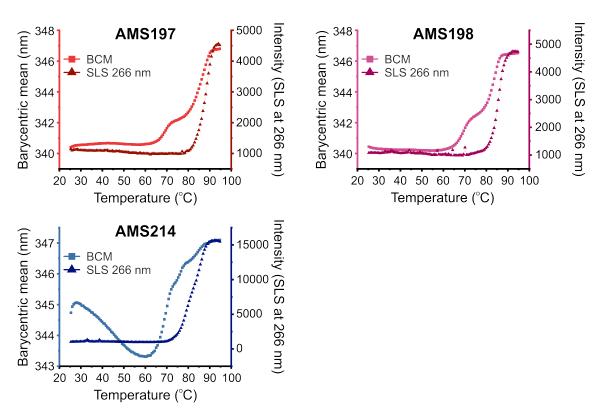
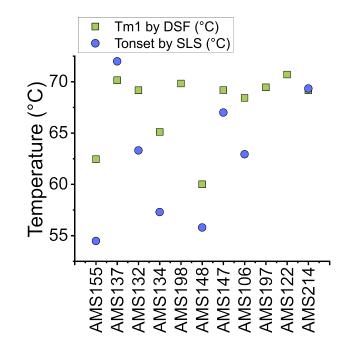
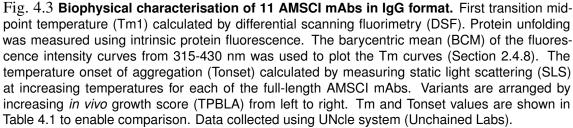


Fig. 4.2 Thermal stability and aggregation behaviour characterisation of 11 AMSCI mAbs in IgG format differential scanning fluormietry (DSF) using intrinsic protein fluorescence by exciting with a 266 nm laser and measuring emission from 315-430 nm. The first transition mid-point temperature (Tm1) was calculated based on the barycentric mean (BCM) of the fluorescence intensity curves from 315-430 nm. Static light scattering (SLS) was measured at each temperature to calculate the temperature onset of aggregation (Tonset) and to delineate unfolding and aggregation. Tm and Tonset values are shown in Figure 4.3 and Table 4.1 to enable comparison. Data collected using UNcle system (Unchained Labs).

Variant	Tm1 by DSF (°C)	Tonset by SLS (°C)
AMS106	68.42	62.94
AMS122	70.70	79.81
AMS132	69.18	63.31
AMS134	65.11	57.30
AMS137	70.14	71.99
AMS147	69.19	67.00
AMS148	60.00	55.79
AMS155	62.45	54.48
AMS197	69.46	79.32
AMS198	69.83	78.25
AMS214	69.16	69.34

Table 4.1 **Stability and aggregation behaviours of AMSCI mAbs in IgG format.** First transition mid-point temperatures (Tm1) and temperature onset of aggregation (Tonset) calculated using differential scanning fluorimetry (DSF) and static light scattering (SLS), respectively.





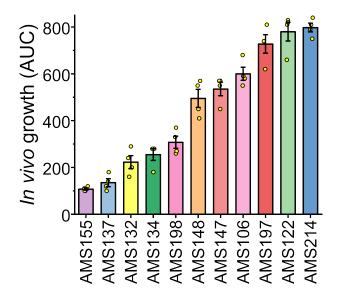


Fig. 4.4 **TPBLA analysis of AMSCI mAbs in scFv format.** Area under the antibiotic survival curve calculated for the scFv regions of the AMSCI mAbs, screened at 0-140  $\mu$ g/mL ampicillin. Data points are from three independent experiments. Error bars show standard error of the mean.

To understand what determines the ranking of the AMSCI mAbs in TPBLA, they were subjected to a panel of developability assays to compare their performance to TPBLA. These assays were chosen based on a landmark study by Jain et al. (2017) assessing the output of 12 commonly employed developability assays on 137 monoclonal antibody therapeutics. Jain et al. (2017) found that many of these developability assays gave similar outputs, and grouped these assays based on their relatedness. Tm by DSF, HIC and AC-SINS (Section 1.4) were chosen to assess the AMSCI mAbs as they each represent one of the 'branches' of relatedness. HIC is a column based method that separates molecules based on their surface hydrophobicity, or their tendency to associate with a column matrix comprised of hydrophobic molecules (Estep et al., 2015). The retention time of the POI is directly related to its association with the column matrix; the least hydrophobic molecules will have a shorter retention time, and the most hydrophobic molecules will have a longer retention time. AMS134 had the longest retention time (18.4 min), indicating it is the most hydrophobic, whereas AMS106 (infliximab) had the shortest retention time (8.3 min), indicating it is the least hydrophobic, despite being the aggregation-prone control (Figure 4.5A, B). The elution profile of AMS148 had two peaks, indicating this variant forms a dimer creating a mixed monomer/dimer population that can interact with the column matrix in different ways. This could potentially be the monomer is interacting via a hydrophobic surface patch, creating a dimer that now has less exposed hydrophobic side chains.

AC-SINS (Section 1.4) is a high-throughput plate-based method designed to assess mAbs based on their propensity to self-associate (Liu et al., 2014). Gold nanoparticles (AuNP) are conjugated to anti-human Fc IgGs and incubated with the mAb of interest. The absorbance spectra of the AuNP are measured and compared with naked AuNP. Selfassociation of the mAb of interest results in a red shift in the wavelength of maximum fluorescence intensity ( $\lambda$ max). A wavelength shift of 5 nm is often used as a threshold, with shifts larger than this classed as high self-association (Liu et al., 2014). The absorbance spectrum for each AMSCI mAb was measured between 500 - 600 nm and compared with two internal controls; CDP850 (aggregation-resistant), and infliximab (aggregationprone). It is important to note that the infliximab used here as an internal control is the therapeutic, which has different constant regions to AMS106 (which contains the infliximab VH and VL domains grafted onto a common AMSCI scaffold). AMS134 has the biggest wavelength shift (mean, M = 20.92 nm; standard deviation, SD = 0.12), followed by AMS137 (M = 16.07 nm, SD = 0.09) and AMS214 (M = 8.89 nm, SD = 0.20), all indicating a high propensity to self-associate (Figure 4.5C, D). The aggregationprone control of infliximab shows a wavelength shift of 35.85 nm (SD = 0.28), whereas the AMSCI variant of infliximab AMS106 only shifts 0.64 nm (SD = 0.14) (Figure 4.5C, D). The aggregation mechanism of therapeutic infliximab has been shown to be

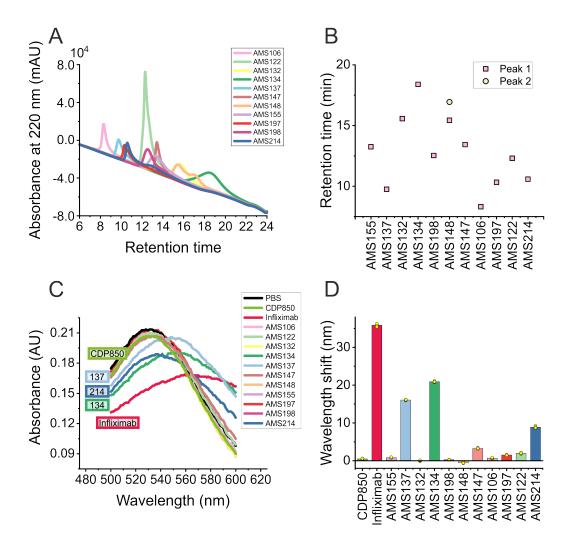
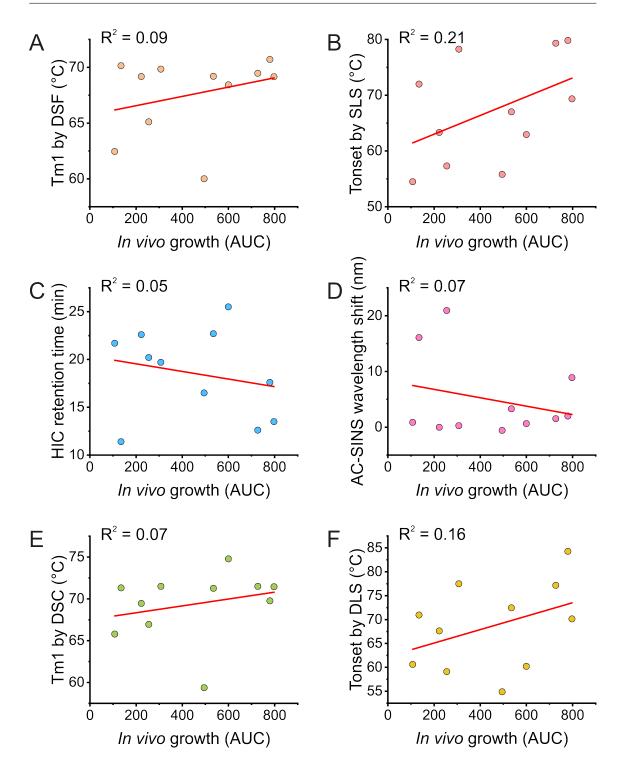


Fig. 4.5 **Biophysical characterisation of 11 chosen IgGs.** A) Hydrophobic Interaction Chromatography (HIC) chromatograms of the 11 full-length AMSCI mAbs (5 μg/mL, PBS). The retention time is shown by an elution peak for each IgG and the corresponding absorbance at 220 nm. B) Retention time values from HIC chromatograms. The chromatogram for AMS148 had two peaks so has two retention times, peak 1 (pink, square) and peak 2 (green, circle). Variants are arranged by increasing *in vivo* growth score (TPBLA) from left to right. HIC experiments carried out with Ailsa MacRae, University of Leeds. C) Affinity-Capture Self-Interaction Nanoparticle Spectroscopy (AC-SINS) of the 11 full-length AMSCI mAbs (1 mg/mL, PBS) alongside two internal controls; CDP850 (aggregation-resistant), and infliximab (aggregation-prone). Larger plasmon wavelength shifts correlate with higher self-association. D) Plasmon wavelength shifts calculated compared with AuNP alone. Variants are arranged by increasing *in vivo* growth score (TPBLA) from left to right. Data points are from five technical repeats, error bars show standard deviation.

as a result of both Fab-Fab interactions in a head-to-tail conformation as well as Fab-Fc interactions (Lerch et al., 2017; Domnowski et al., 2021). As AMS106 only contains the variable domains of therapeutic infliximab with different constant regions, it is possible this disrupted the interfaces involved in these interactions. AMS106, AMS122, AMS132, AMS147, AMS148, AMS155, AMS197, and AMS198 all showed a wavelength shift of below 5 nm, indicating low self-association (Figure 4.5C, D).

#### 4.2.2 TPBLA does not correlate to one single biophysical parameter

Previous studies have correlated aggregation propensity, solubility, and thermodynamic stability to performance in TPBLA. We sought to use the AMSCI mAb dataset to identify the main drivers in performance in TPBLA. The full length IgG fragments were analysed using HIC, AC-SINS, DSF, and SLS to assess hydrophobicity, self-association, thermal stability, and aggregation (Section 4.2.1). The scFv models from ABodyBuilder were assessed using structurally corrected Camsol to assess the predicted solubility (Leem et al., 2016; Sormanni et al., 2015a). These were then plotted against in vivo growth score in TPBLA (Figure 4.6, Figure 4.7). Alternative Tm and Tonset values of the AMSCI mAbs using differential scanning calorimetry (DSC) and dynamic light scattering (DLS), respectively, from the original BioStreamline project was provided by Dr Michael Knight (UCB). Additionally, theoretical pI and IgG size in kDa was included as a metric to plot against TPBLA. TPBLA shows poor correlation with all metrics assessed. Thermal stability (Tm1 by DSF and DSC), aggregation (Tonset by DLS and SLS, Tagg by SLS), hydrophobicity (HIC retention time), self-association (AC-SINS wavelength shift), predicted solubility (Camsol score), and size in kDa all showed no correlation ( $R^2 < 0.25$ ). Looking at the data using Spearmans rank, some metrics show a moderate correlation with TPBLA score (Figure 4.7). The strongest correlation is Tonset by SLS, with a Spearmans rank of 0.51 (Figure 4.7B). Theoretical pI shows a moderate negative correlation, with a Spearmans rank of -0.48 (Figure 4.7G). From these data, it is clear there is no single metric that can individually explain the performance of the AMSCI mAbs in TPBLA.



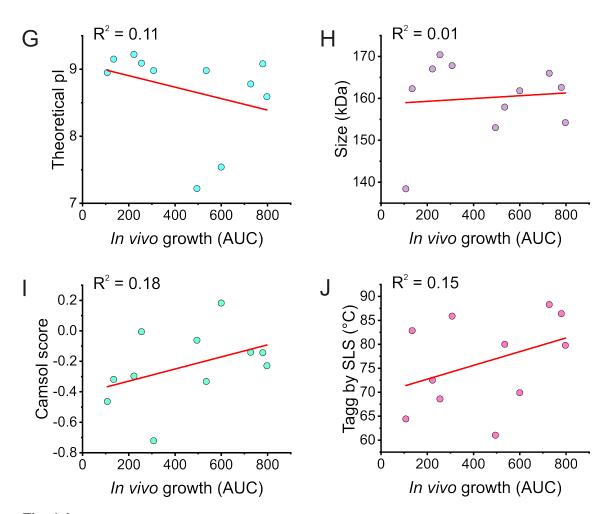
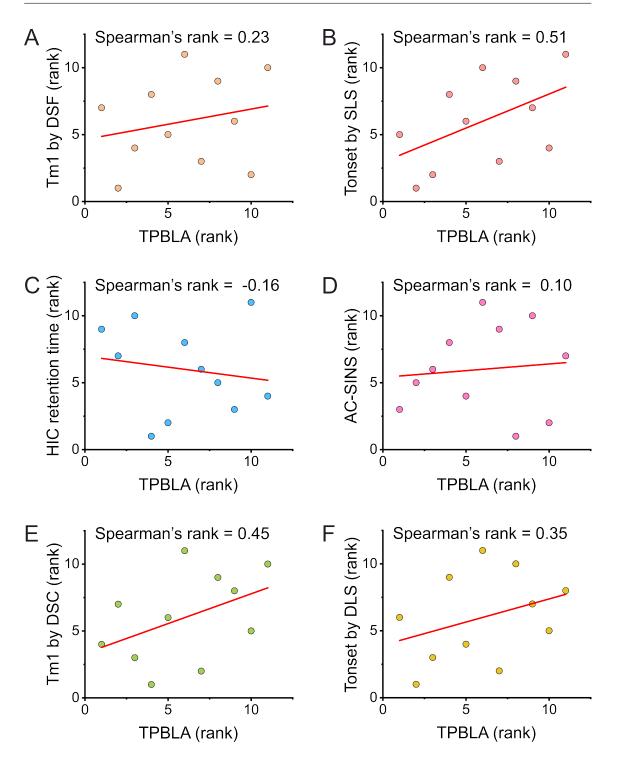


Fig. 4.6 **TPBLA** score shows no strong correlation with any single parameter. *in vivo* growth score from TPBLA plotted against (A) First transition mid-point temperatures (Tm1) by DSF, (B) temperature onset of aggregation (Tonset) by SLS, (C) HIC retention time, (D) AC-SINS plasmon wavelength shift, (E) Tm1 by DSC, (F) Tonset by DLS, (G) Theoretical pl, (H) IgG size in kDa, (I) Camsol solubility score and (J) temperature midpoint of aggregation (Tagg) by SLS. (E) and (F) data provided by Dr Michael Knight (UCB). Linear regression was performed using Orign pro.



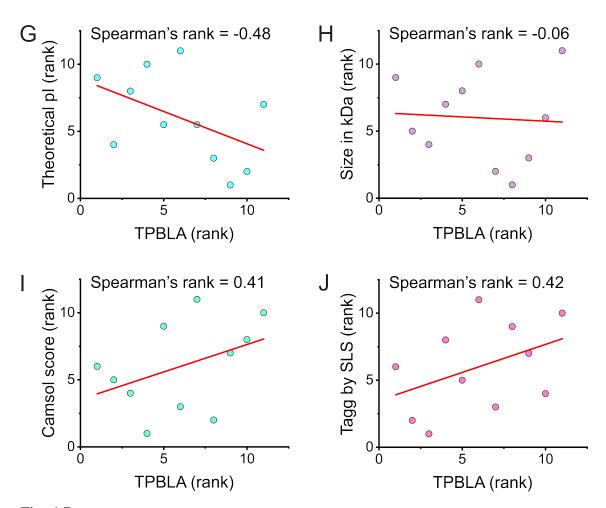


Fig. 4.7 **TPBLA score does not correlate with any single parameter.** Rank of *in vivo* growth score from TPBLA plotted against ranks of (A) First transition mid-point temperatures (Tm1) by DSF, (B) temperature onset of aggregation (Tonset) by SLS, (C) HIC retention time, (D) AC-SINS plasmon wavelength shift, (E) Tm1 by DSC, (F) Tonset by DLS, (G) Theoretical pl, (H) IgG size in kDa, (I) Camsol solubility score and (J) temperature midpoint of aggregation (Tagg) by SLS. (E) and (F) data provided by Dr Michael Knight (UCB). Spearmans rank and linear regression was performed using Orign pro.

## 4.2.3 Multiple regression can be used to rationalize performance in TPBLA

Anything that affects the active concentration or activity of β-lactamase in TPBLA will influence the resulting *in vivo* growth score. This is included but not limited to; soluble expression levels, proper folding, thermodynamic stability, thermal stability, colloidal stability, self-association, aggregation, and degradation by proteases. As these parameters overlap with those that affect a biopharmaceuticals developability, TPBLA has the potential to be used to identify poorly developable candidates. However, it is important to understand which parameters are having the biggest impact on TPBLA score, that is the growth score in the TPBLA (Figure 4.4). In Section 4.2.2, TPBLA score does not correlate with any single parameter assessed. Therefore, we sought to explain TPBLA *in vivo* growth score for the AMSCI mAbs using a combination of the developability assays listed in Figure 4.6 which measure particular biophysical properties. In other words, can the TPBLA score be explained as a function of multiple physicochemical properties of the molecule and what is the minimum number of properties required to do this. This was achieved by using multiple regression models to compare the TPBLA growth score against the output of a combination of the aforementioned developability assays.

A multiple linear regression analysis was used to test associations between TPBLA score and the biophysical characterisation assays used to assess the stability, aggregation, and solubility of the AMSCI mAbs. Essentially, this was to test which parameters can be combined to significantly predict TPBLA score. For a more detailed description of the regression model analysis, see Methods (Section 2.9).

Combinations of parameters from Figure 4.6 were used in a linear regression model, and parameters were systematically removed until a model was found where all the parameters were having a statistically significant influence on predicting TPBLA score (p < 0.05).

A model including theoretical pI, Tonset by DLS, Tonset by SLS, Tagg by SLS, and Camsol score was statistically significant (f(5, 5) = 5.142,  $R^2 = 0.787$ , r = 0.887, p < 0.05) (Figure 4.8A). Detailed statistics for the model can be found in Appendix C, Table C.1. However, the model has almost as many parameters as AMSCI mAb variants which can result in overfitting. Therefore, parameters were systematically removed and to find the most parsimonious model.

By removing Tagg by SLS and including theoretical pI, Tonset by DLS, Tonset by SLS, and Camsol score, the model was statistically significant (f(4, 6) = 5.537,  $R^2 = 0.837$ , r =

0.915, p < 0.05) (Figure 4.8B). Detailed statistics for the model can be found in Appendix C, Table C.2.

A model using only Tonset DLS, Camsol score, and theoretical pI is able to predict TPBLA score reasonably well. The overall regression was statistically significant (f(3, 7) = 7.147,  $R^2 = 0.754$ , r = 0.868, p < 0.05) (Figure 4.8C). Detailed statistics for the model can be found in Appendix C, Table C.3.

The most parsimonious model was using Tonset by DLS and theoretical pI, where the overall regression was statistically significant (f(2, 8) = 6.192, R<sup>2</sup> = 0.608, r = 0.779, p < 0.05) (Figure 4.8D). Both parameters were significant predictors of TPBLA score (Tonset by DLS:  $\beta$ = 24.24, t = 3.194, p < 0.05; theoretical pI:  $\beta$ = -311.62, t = -3.028, p < 0.05). Detailed statistics for the model can be found in Appendix C, Table C.4.

Using a Spearmans rank to assess correlations between the ranked predicted and experimental TPBLA scores for the models gave similar R<sup>2</sup> values as the unranked data (Figure 4.9). While none of these models can be used to accurately predict *de novo* TPBLA scores as the dataset is small and does not have many different antibodies, it can be used to show trends and highlight links within the dataset. Specifically, the final model using Tonset by DLS and theoretical pI suggests that both aggregation and solubility have a strong influence on TPBLA score. This analysis would need to be carried out on a larger dataset to be more confident in this hypothesis.

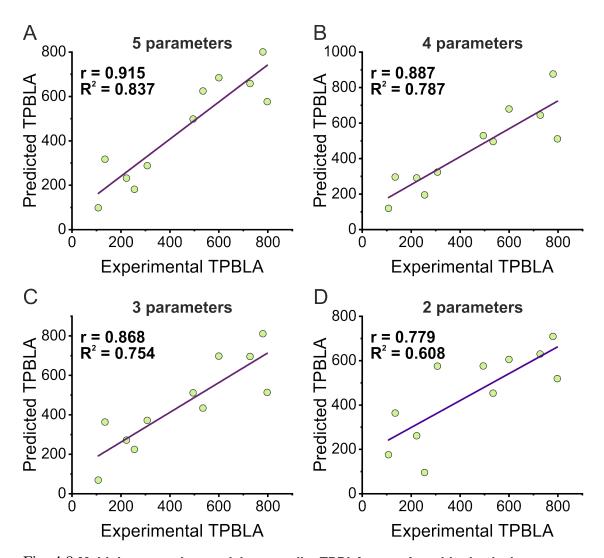


Fig. 4.8 **Multiple regression models to predict TPBLA score from biophysical parameters.** Experimental TPBLA score plotted against predicted TPBLA score using multiple regression models utilising different biophysical parameters. A)Uses theoretical pl, Tonset by DLS, Tonset by SLS, Tagg by SLS, and Camsol score. B) Omits Tagg by SLS and uses only theoretical pl, Tonset by DLS, Tonset by SLS, and Camsol score. C) Omits Tonset by SLS and uses theoretical pl, Tonset by DLS, and Camsol score. D) Is the most parsimonious model, using only Tonset by DLS and theoretical pl.

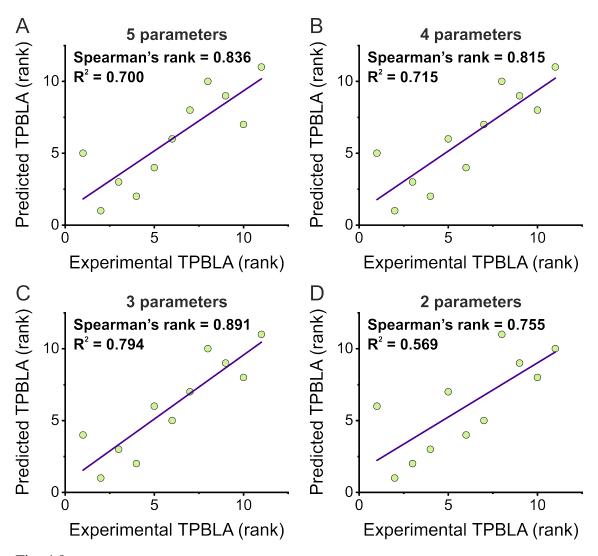


Fig. 4.9 Multiple regression models to predict TPBLA score from biophysical parameters. Experimental TPBLA rank plotted against predicted TPBLA rank using multiple regression models utilising different biophysical parameters. A) Uses theoretical pl, Tonset by DLS, Tonset by SLS, Tagg by SLS, and Camsol score. B) Omits Tagg by SLS and uses only theoretical pl, Tonset by DLS, Tonset by SLS, and Camsol score. C) Omits Tonset by SLS and uses theoretical pl, Tonset by DLS, and Camsol score. D) Is the most parsimonious model, using only Tonset by DLS and theoretical pl.

## 4.2.4 TPBLA can be used to evolve therapeutic scaffolds to improve their biophysical properties

We wanted to apply the high-throughput evolution protocol developed in Chapter 3 to some of the AMSCI mAbs to see if this could be used to evolve poorly developable candidates and improve their developability. AMS134 was chosen as it performs poorly in all the biophysical characterisation assays; it has one of the lowest *in vivo* growth scores in TPBLA, the largest wavelength shift in AC-SINS, longest retention time on a HIC column, and it has one of the lowest Tm1 (DSF) and Tonset (SLS) values of all the AMSCI mAbs. Conversely, AMS197 has one of the highest *in vivo* growth scores in TPBLA apart from the aggregation-resistant therapeutic control (AMS122, trastuzumab). It has no significant wavelength shift in AC-SINS, one of the shortest retention times on the HIC column, one of the highest Tm (DLS) and the highest Tonset (SLS) of the AMSCI mAbs. Therefore, it represents a highly stable and aggregation-resistant variant to test the dynamic range of the evolution methodology.

AMS134 and AMS197 were evolved using the *in vivo* assay by introducing genetic variation into the respective genes and creating a mutated plasmid library within the  $\beta$ -lactamase vector (Section 2.6.2) to produce  $\beta$ La AMS134\* and  $\beta$ La AMS197\*. For screening, the libraries were transformed into E. coli SCS1 cells and plated onto agar containing 40  $\mu$ g/mL for  $\beta$ La AMS134\* and 180  $\mu$ g/mL for  $\beta$ La AMS197\*. At these concentrations, the 'wild-type' AMS134 and AMS197 sequences were unable to survive. Therefore, variants growing should have beneficial mutations that improve the total activity of folded and soluble fusion proteins. The DNA from the resulting colonies were pooled, purified, and the genes amplified using PCR before being sent for Illumina sequencing along with the respective unselected (naive) libraries. Paired end fragments were aligned to a reference sequence (the respective 'wild-type' sequence) and the aligned fragments were translated in frame with respect to the reference sequence. By comparing the aligned translated fragments to the original 'wild-type' sequence, mutational frequency at each position was calculated and normalised by read coverage. At each position the mutation frequency was normalised by coverage. The mean mutation rate at the unmutated GS linker upstream and downstream of the scFv gene was used as a threshold; a mutation rate below this was classed as zero. This was used to calculate the log<sub>2</sub>(fold change) at each residue. Hotspot residues were identified as having a log<sub>2</sub>(fold change) of more than two standard deviations from the mean (> $2\sigma$ ).

For AMS134, a cluster of hotspots was identified in and around HCDR3 (Figure 4.10, Figure 4.11). Three residues were identified within VH CDR3, W113, Y114, F115. This

correlates with the only hotspot identified in Aggrescan3D comprising of G112 - F115, essentially the VH CDR3. These were most commonly mutated to less hydrophobic residues (W113R, Y114H, and F115S), forming a patch of less aggregation-prone residues on the surface (Figure 4.12). Four other hotspot residues were identified in the VH domain, four after CDR3 (G119, G121, V124, I125), and one between CDR1 and CDR2 (P46). M123 was above the  $2\sigma$  threshold in the normalised mutation frequency but just below the threshold for the log<sub>2</sub>(fold change) (Figure 4.10, Figure 4.11). In the ABodyBuilder model, M123 forms a hydrophobic surface patch with P46, G119, G121, V124, and I125. Therefore, it was included as a hotspot residue for further analysis (Figure 4.12). The hotspots were most commonly mutated to P46L, G119D, G121D, M123K, V124D, and I125T. These were often charged residues (G119D, G121D, M123K, V124D), which are widely accepted to oppose protein aggregation via electrostatic repulsion (Sant'Anna et al., 2014). All of these hotspots were mutated to less hydrophobic residues, except for P46 which was mutated to leucine. In the ABodyBuilder model,G119, G121, M123 and I125 seem to form the edge strand of a  $\beta$ -sheet, a region commonly accepted to be involved in mediating protein aggregation via edge-strand interactions (Trinh et al., 2002; Richardson and Richardson, 2002; Siepen et al., 2003). Furthermore, this region was identified as being within an APR using Waltz, a sequence-based algorithm for predicting amyloidogenic peptides (Louros et al., 2020). Therefore, this region could be involved in driving self-association of AMS134, and mutating it could disrupt this association. No hotspots were identified in the VL over the  $2\sigma$  threshold (Figure 4.10, Figure 4.11). The closest was R93, which was most commonly mutated to glycine. To understand the affect of mutations on the VL as well as the VH, R93G was included in future analysis.

Looking at the normalised mutation frequency, AMS197 shows a huge cluster in hotspots between residues 57-86 (Figure 4.13). However, this 'hotspot cluster' is present within the naive library, likely as a result of the error-prone PCR (Figure 4.13). This highlights the importance of looking at the log<sub>2</sub>(fold change), to identify positions whose mutation rate has been enriched as a result of the selection. For AMS197, only 5 hotspots were identified as being above the  $2\sigma$  threshold for the log<sub>2</sub>(fold change) (Figure 4.13, Figure 4.14). Two residues were identified in HCDR1, T29 and D38, which were most commonly mutated to A and N, respectively (Figure 4.15). One hotspot residue was identified within HCDR2, S65, which was most commonly mutated to N (Figure 4.13, Figure 4.14). Unlike in AMS134, the hotspot residues identified in the evolution of AMS197 form no patches on the surface (Figure 4.15).

To assess the impact of the hotspot mutations on the stability, aggregation propensity, and *in vivo* growth in TPBLA, they were introduced as single point mutations back into

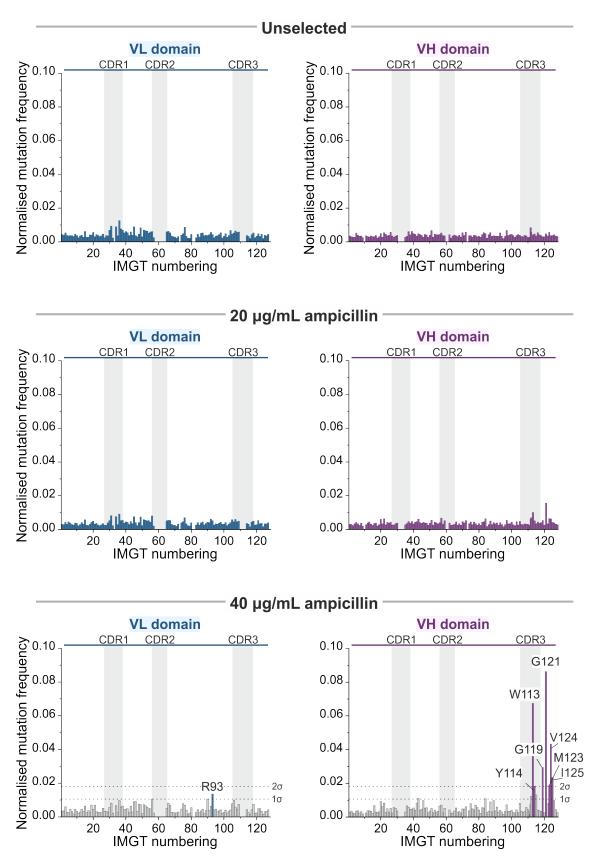


Fig. 4.10 Evolution of AMS134 analysed by Illumina sequencing. Mutational frequency normalised by read coverage for AMS134 showing the naive library, selected library at 20  $\mu$ g/mL ampicillin, and selected library at 40  $\mu$ g/mL ampicillin. Hotspots are identified as having a mutational frequency of more than two standard deviations from the mean (>2 $\sigma$ ). Grey boxes denote residues in CDRs. Evolution experiments carried out with Ailsa MacRae, University of Leeds.

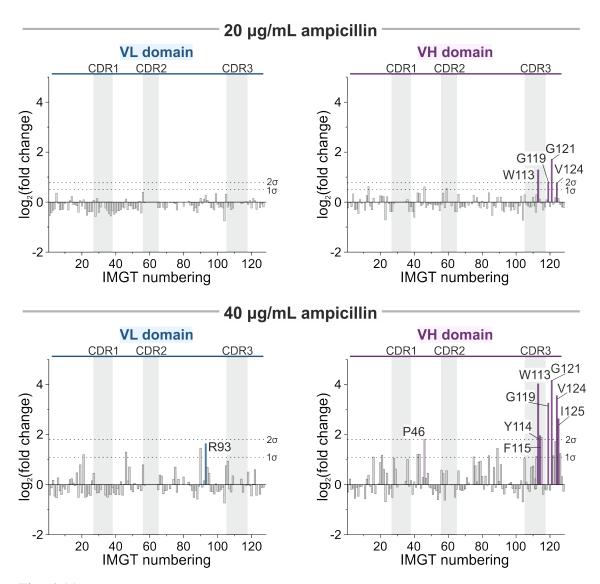


Fig. 4.11 Evolution of AMS134 analysed by Illumina sequencing.  $\log_2(\text{fold change})$  of the mutational frequency calculated using the naive and selected libraries of AMS134 evolved at at 20 µg/mL and 40 µg/mL ampicillin. Hotspots are identified as having a  $\log_2(\text{fold change})$  of more than two standard deviations from the mean (>2 $\sigma$ ). Grey boxes denote residues in CDRs. Evolution experiments carried out with Ailsa MacRae, University of Leeds.

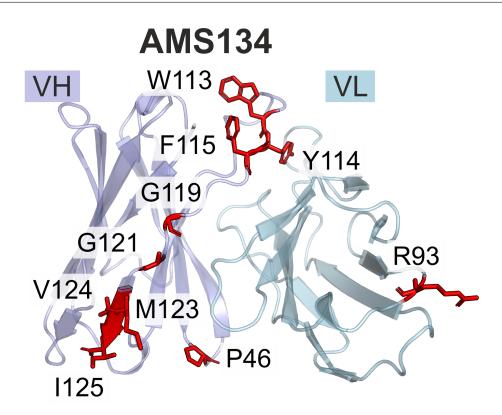


Fig. 4.12 **AMS134** hotspots identified in Illumina sequencing mapped onto the scFv structure. Hotspot residues (red) identified as having a normalised mutational frequency or log<sub>2</sub>(fold change) of more than two standard deviations (>2 $\sigma$ ) from the mean in Illumina sequencing mapped onto the scFv structure of AMS134 predicted by ABodyBuilder (Leem et al., 2016). R93 was included as a hotspot in the VL although it was slightly below the 2 $\sigma$  cutoff as a representative position in the VL. VH is shown in purple and VL is shown in blue.

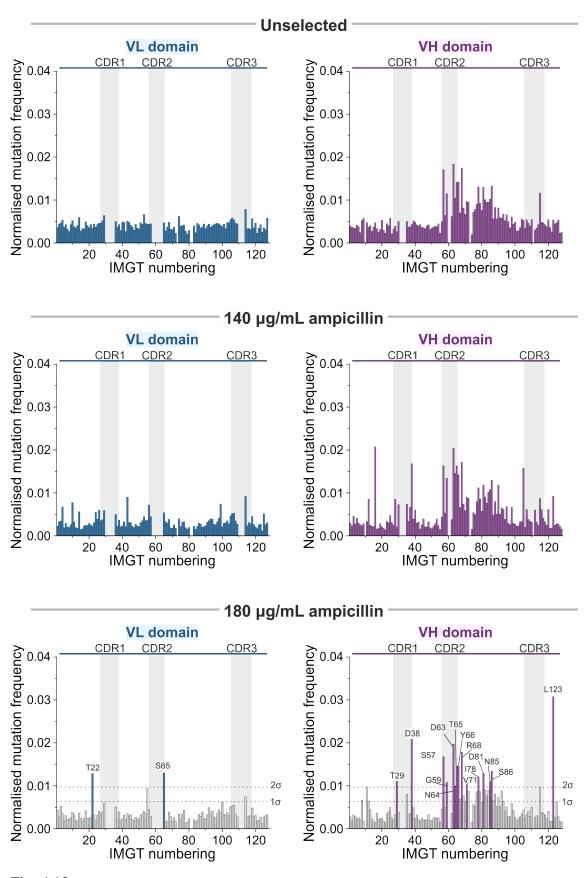


Fig. 4.13 Evolution of AMS197 analysed by Illumina sequencing. Mutational frequency normalised by read coverage for AMS197 showing the naive library, selected library at 140  $\mu$ g/mL ampicillin, and selected library at 180  $\mu$ g/mL ampicillin. Hotspots are identified as having a mutational frequency of more than two standard deviations from the mean (>2 $\sigma$ ). Grey boxes denote residues in CDRs. Evolution experiments carried out with Ailsa MacRae, University of Leeds.

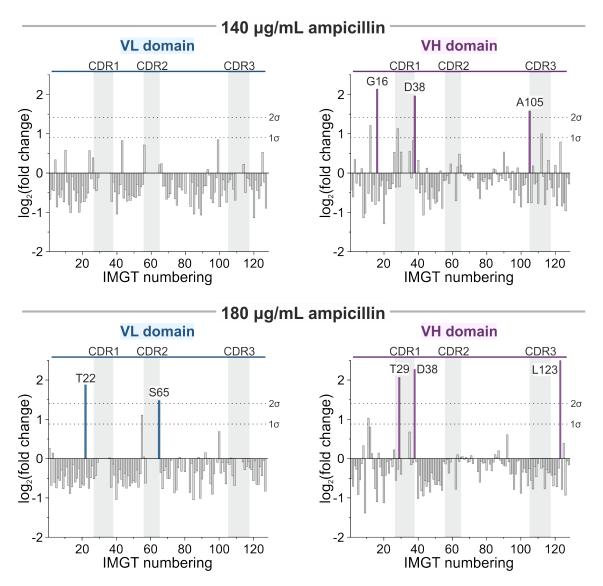


Fig. 4.14 Evolution of AMS197 analysed by Illumina sequencing.  $log_2$ (fold change) of the mutational frequency calculated using the naive and selected libraries of AMS197 evolved at at 140 µg/mL and 180 µg/mL ampicillin. Hotspots are identified as having a log<sub>2</sub>(fold change) of more than two standard deviations from the mean (>2 $\sigma$ ). Grey boxes denote residues in CDRs. Evolution experiments carried out with Ailsa MacRae, University of Leeds.

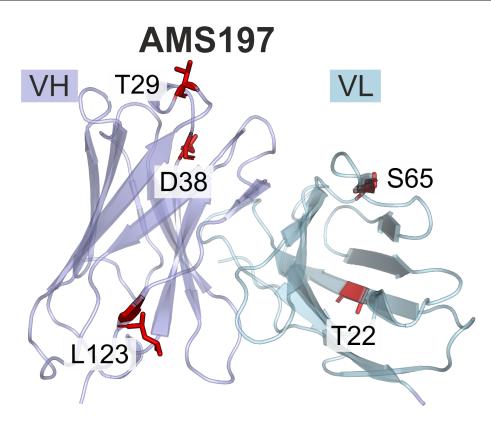


Fig. 4.15 AMS197 hotspots identified in Illumina sequencing mapped onto the scFv structure. Hotspot residues (red) identified as having a normalised mutational frequency or  $log_2$ (fold change) of more than two standard deviations (> $2\sigma$ ) from the mean in Illumina sequencing mapped onto the scFv structure of AMS197 predicted by ABodyBuilder (Leem et al., 2016). VH is shown in purple and VL is shown in blue.

AMS134 and AMS197 and assessed them using TPBLA. All 10 point mutants of AMS134 displayed improved growth in TPBLA compared with AMS134 wild-type (Figure 4.16A). The most improved was V124D, with a 1.5-fold increase in *in vivo* growth (Figure 4.16A). For AMS197, 3 out of the 4 point mutants assessed displayed slightly improved growth in TPBLA compared with wild-type (Figure 4.16B). It is possible the variants identified in the evolution screen came about as multiple point mutants, rather than single point mutants. To test this, the evolved libraries of AMS134 at 40  $\mu$ g/mL ampicillin and AMS197 at 180  $\mu$ g/mL ampicillin were transformed into *E. coli* SCS1 cells and single colonies were picked to be screened by TPBLA. If these single colonies showed significantly improved *in vivo* growth scores compared to the single point mutants assessed, this could be due to them containing multiple mutations. For both libraries, the single colonies showed similar *in vivo* growth scores compared with the single point mutants, suggesting the evolved mutants are likely single point mutants (Figure 4.16).

The single point mutants and single colonies (from the transformation of the evolved library) of AMS197 do not display significantly improved growth in TPBLA relative to the wild-type (Figure 4.16B). For comparison, the best point mutant identified for AMS134 was V124D. This showed an improvement in *in vivo* growth score of 93, which corresponds to a 1.5-fold improvement relative to wild-type AMS134. The best point mutant identified for AMS197 was D38N. This showed an improvement in *in vivo* growth score of 127, which corresponds to only a 1.1-fold improvement relative to wild-type AMS197. This is potentially due to the fact that AMS197 is already a highly stable and aggregation-resistant molecule. Therefore, in order to significantly improve its *in vivo* growth score it may require multiple point mutations in combination, and to be evolved at a higher ampicillin concentration.

# 4.2.5 Evolved single point mutations of AMS134 improve aggregation behaviours

To understand how TPBLA has evolved these antibodies, three variants of AMS134 as well as the wild-type were expressed and purified in Chinese Hamster Ovary (CHO) cells as full-length IgG fragments for further characterisation. R93G in the VL domain was chosen as it is the only residue close to the  $2\sigma$  cutoff in the VL domain. M123K in the VH domain was chosen as it is an example that is near the  $2\sigma$  cutoff. It was over the  $2\sigma$  threshold in the normalised mutation frequency but just below the threshold for the log<sub>2</sub>(fold change), therefore representing a residue that is just on the threshold (Figure 4.10, Figure 4.11).

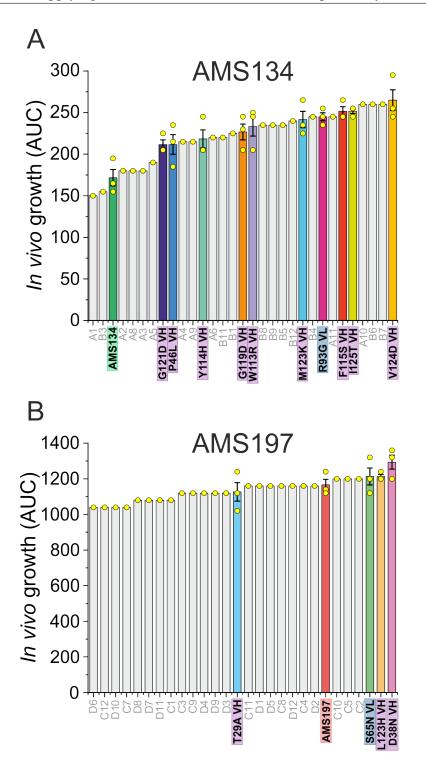


Fig. 4.16 *In vivo* screen of evolved AMS134 and AMS197 point mutants identified by Illumina sequencing. Area under the antibiotic survival curve calculated for evolved (A) AMS134, and (B) AMS197 point mutants identified by Illumina sequencing, screened at (A) 0-70  $\mu$ g/mL ampicillin, and (B) 0-280  $\mu$ g/mL ampicillin. Error bars show standard error of the mean (S.E.M) from three independent experiments. TPBLA carried out with Ailsa MacRae, University of Leeds.

Table 4.2 **Stability and aggregation behaviours of AMS134 mAbs.** First (Tm1) and second (Tm2) transition mid-point temperatures and temperature onset of aggregation (Tonset) calculated using differential scanning fluorimetry (DSF) and static light scattering (SLS), respectively.

Variant	Tm1 (°C)	Tm2 (°C)	Tonset (°C)
AMS134 <sup>WT</sup>	64.26	83.63	54.81
AMS134 <sup>R93G VL</sup>	63.28	83.53	52.29
AMS134 <sup>M123K VH</sup>	65.13	84.05	54.70
AMS134 <sup>V124D VH</sup>	65.33	88.24	

Table 4.3 Hydrodynamic radius (Rh) of AMS134 IgGs measured by dynamic light scattering (DLS) before and after temperature ramp.

	Rh at 25 °C (nm)	Rh at 95 °C (nm)
AMS134 <sup>WT</sup>	9.5	23.8
AMS134 <sup>R93G VL</sup>	9.5	23.8 & 79.9
AMS134 <sup>M123K VH</sup>	9.5	18.6
AMS134 <sup>V124D VH</sup>	8.1	1.4 & 13.5

Conversely, V124D in the VH is above both thresholds and has the highest *in vivo* growth score of the AMS134 point mutants assessed (Figure 4.10, Figure 4.11, Figure 4.16A).

The AMS134 point mutants have similar thermal stabilities as measured by DSF (Figure 4.17A, Table 4.2). AMS134<sup>R93G VL</sup> had the lowest Tm (63.28 °C), whereas with AMS134<sup>V124D VH</sup> which had the highest Tm (65.33 °C) (Table 4.2). At the start (25 °C) and end (95 °C) of the temperature ramp in DSF, the size of the molecule (hydrodynamic radius, Rh) was measured using DLS. AMS134<sup>WT</sup>, AMS134<sup>R93G VL</sup>, and AMS134<sup>M123K VH</sup> showed a shift in hydrodynamic radius following temperature ramp from 9.5 nm to 23.8 nm (WT and R83G) or 18.6 nm (M123K) (Table 4.3, Figure 4.17B, Figure 4.18A, B, C). AMS134<sup>R93G VL</sup> also formed a larger aggregate of 79.9 nm (Table 4.3). However, AMS134<sup>V124D VH</sup> remained around the same size (8.1 nm at 25 °C, 10.6 nm at 95 °C, Table 4.3, Figure 4.17B, Figure 4.18D). Consistent with this, AMS134<sup>V124D VH</sup> abolishes aggregation at high temperatures compared with AMS134<sup>WT</sup>, as measured by SLS at 266 nm (Figure 4.17C, Figure 4.19).

AMS134<sup>M123K VH</sup> has a similar Tonset of aggregation compared with AMS134<sup>WT</sup> (Table 4.2, Figure 4.17C, Figure 4.19). This is consistent with the fact that M123 was only just at the  $2\sigma$  cutoff of normalised mutation frequency, and was below the  $2\sigma$  cutoff of log<sub>2</sub>(fold change), so was not expected to confer vast improvements in stability or aggregation resistance. AMS134<sup>R93G VL</sup> has a similar Tm to AMS134<sup>WT</sup>, but has significantly increased aggregation (Table 4.2, Figure 4.17C, Figure 4.19). This is consistent with the

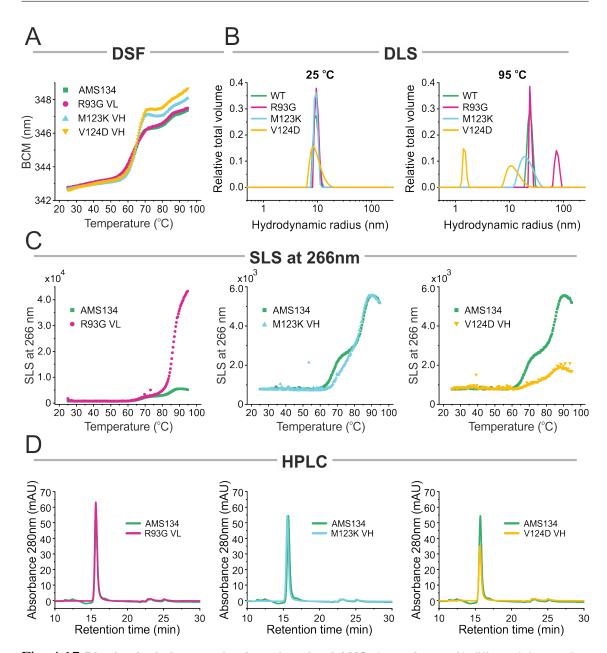


Fig. 4.17 **Biophysical characterisation of evolved AMS134 variants.** A) differential scanning fluormietry (DSF) using intrinsic protein fluorescence by exciting with a 266 nm laser and measuring emission from 315-430 nm. The first transition mid-point temperature (Tm1) was calculated based on the barycentric mean (BCM) of the fluorescence intensity curves from 315-430 nm. B) Dynamic light scattering (DLS) was measured at 25 °C and 95 °C to see the change in hydrodynamic radius following thermal melt. C) Static light scattering (SLS) was measured at each temperature to calculate the temperature onset of aggregation (Tonset) and to delineate unfolding and aggregation. Data collected using UNcle system (Unchained Labs). D) HPLC chromatograms of the 3 full-length AMS134 point mutants compared with AMS<sup>WT</sup> (1 mg/mL, PBS) on an XBridge Protein BEH SEC Column, 200Å, 3.5  $\mu$ m (Waters). The retention time is shown by an elution peak for each IgG and the corresponding absorbance at 280 nm.

170

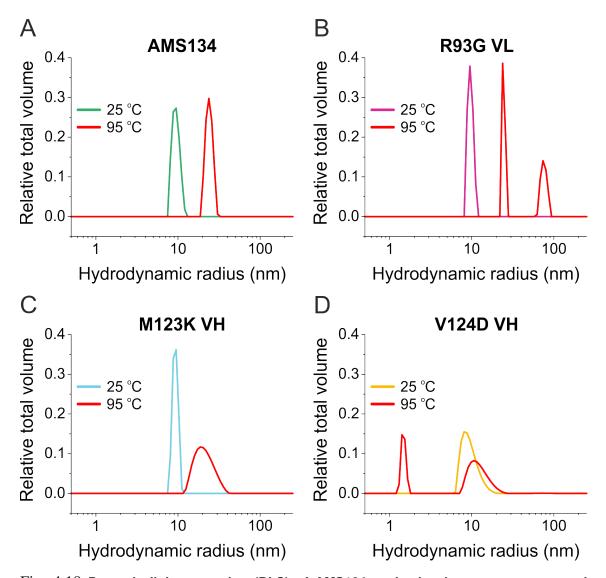


Fig. 4.18 Dynamic light scattering (DLS) of AMS134 evolved point mutants measured before and after thermal melt. Dynamic light scattering (DLS) measured before (25 °C) and after (95 °C) thermal melt to see the change in hydrodynamic radius. Variants shown are (A) AMS134<sup>WT</sup>, (B) AMS134<sup>R93G VL</sup>, (C) AMS134<sup>M123K VH</sup>, and (D) AMS134<sup>V124D VH</sup>.

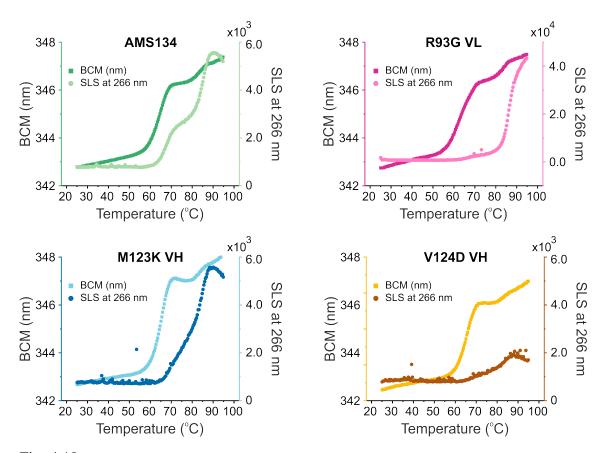


Fig. 4.19 **Thermal stability and aggregation behaviour of evolved AMS134 variants.** differential scanning fluormietry (DSF) using intrinsic protein fluorescence by exciting with a 266 nm laser and measuring emission from 315-430 nm. The first transition mid-point temperature (Tm1) was calculated based on the barycentric mean (BCM) of the fluorescence intensity curves from 315-430 nm. Static light scattering (SLS) at 266 nm was measured at each temperature to calculate the temperature onset of aggregation (Tonset) and to delineate unfolding and aggregation.

fact that R93 was below the  $2\sigma$  cutoff of both normalised mutation frequency and  $\log_2(\text{fold change})$ , so was not expected to improve the stability or aggregation resistance.

The variants were analysed by HPLC to measure their retention time as well as their proportions of high- and low- molecular weight species. All variants showed the same retention time for the monomeric species (Figure 4.17D). They all showed similar levels of high-molecular weight species, but only AMS134<sup>WT</sup> showed significant levels of low-molecular weight species as well as a small shoulder in the monomeric peak. This suggests the wild-type is more susceptible to degradation than the evolved point mutants, which may give some insight into the fact these variants showed improved *in vivo* growth scores in TPBLA.

The variants were assessed by AC-SINS to measure their propensity to self-associate. AMS134<sup>R93G VL</sup> showed an increased wavelength shift (increased self-association) compared with AMS134<sup>WT</sup> (one way ANOVA: p < 0.001, Tukey's HSD: p < 0.001), consistent with the SLS and DLS data showing the increase in aggregation (Figure 4.20). AMS134<sup>M123K VH</sup> showed an almost 1 nm reduction in wavelength shift (reduced self-association) compared with AMS134<sup>WT</sup> (one way ANOVA: p < 0.001, Tukey's HSD: p < 0.001) (Figure 4.20). AMS134<sup>V124D VH</sup> showed a wavelength shift of 0.5 nm less than compared with AMS134<sup>WT</sup>, indicating reduced self-association (one way ANOVA: p < 0.001, Tukey's HSD: p < 0.001, Tukey's HSD: p < 0.05), consistent with the decrease in aggregation seen in the SLS and DLS data (Figure 4.20).

All three point mutants show improvement in in vivo growth scores in TPBLA compared with AMS134<sup>WT</sup>, although it is not clear exactly what is driving this change. AMS134<sup>V124D VH</sup> no longer aggregates at high temperatures as measured by SLS, but it self-associates in AC-SINS at almost the same level as AMS134<sup>WT</sup>. AMS134<sup>R93G VL</sup> has a higher growth score in TPBLA, but has a massive increase in aggregation compared with AMS134<sup>WT</sup> as measured by SLS. AMS134<sup>M123K VH</sup> has no change in aggregation or thermal stability compared with AMS134<sup>WT</sup>, but shows a slight reduction in self-association as measured by AC-SINS. Nevertheless, the improvements that are seen with these point mutants are very slight. However, the library was designed to contain on average one amino acid substitution per scFv gene. It is likely that to get significant improvements in stability and aggregation behaviour this mutation rate needs to be increased to allow multiple mutations. However, this would not be possible with our currently employed shotgun Illumina sequencing methodology, where shorter fragments are aligned on a longer reference sequence, as it makes it impossible to know if identified point mutations are found alone or in combination with others. Furthermore, this makes it complicated to try and understand exactly how TPBLA is improving these test proteins, as well as the

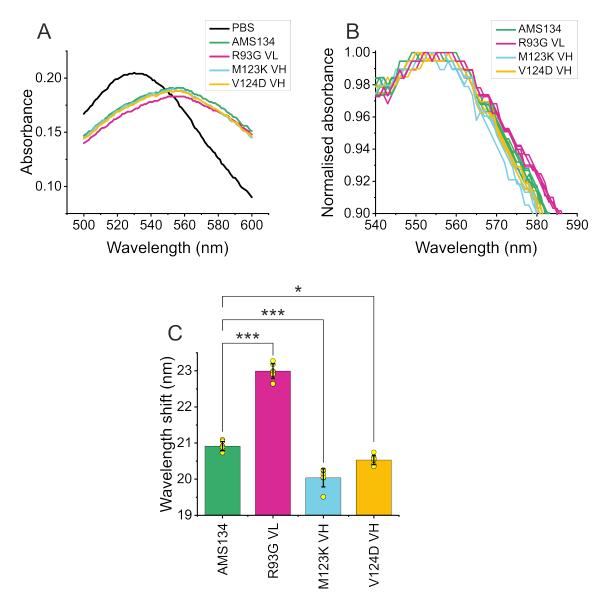


Fig. 4.20 AC-SINS absorbtion spectra and wavelength shifts of AMS134 evolved point mutants. A) Affinity-Capture Self-Interaction Nanoparticle Spectroscopy (AC-SINS) of AMS134<sup>WT</sup> (green) and 3 evolved point mutants formatted as full-length IgGs (1 mg/mL, PBS), and AuNP alone (black). Larger plasmon wavelength shifts correlate with higher self-association. Representative spectra from 1 of 5 technical repeats. B) Normalised absorbance spectra of all 5 repeats between 540 - 590 nm. C) Plasmon wavelength shifts calculated compared with AuNP alone. Data points are from five technical repeats, error bars show standard deviation. (\*' = p < 0.05, (\*\*\*') = p < 0.001.

individual impact of the single point mutations. This could be addressed by analysing the VH and VL domains separately as individually they are small enough to be read in a single Illumina read. Alternatively, barcoding or a long-read sequencing technique such as Pacbio could be used to read the entire mutated gene in one run and therefore enabling analysis of multiple mutations.

# 4.2.6 Analysis of 35 clinical late stage therapeutics demonstrates no correlation between MIC and any single developability assay

The AMSCI dataset was useful initially to demonstrate the potential of TPBLA to (1) screen and rank test proteins, and (2) evolve test proteins to improve their developability. The models presented in Section 4.2.3 demonstrate links between TPBLA and mAb stability, solubility, and aggregation. However, the small size of the dataset limits the applicability of these models. In order to probe this further, we utilised a well-characterised dataset of late stage clinical therapeutics (Phase 2, Phase 3, or Approved) from a study by Jain et al. (2017) characterised using a suite of developability assays summarised in Table 4.4. The scFv encoding 35 of these mAbs were screened using TPBLA to correlate with other developability assays and identify the main factors influencing TPBLA growth score and to identify correlations.

The *in vivo* growth score of *E. coli* SCS1 cells expressing the scFv of each of the 35 chosen mAbs within the TPBLA construct was measured in a 48 well plate format over an ampicillin concentration range of 0-140  $\mu$ g/mL. The dataset showed a wide range of responses (Figure 4.21A). Looking at the *in vivo* growth score of the scFvs based on their approval rating, there is a statistically significant difference between Phase 2, Phase 3, or Approved scFvs when analysed using ANOVA (p = 0.04). However, further assessing this using a post-hoc Tukey test demonstrates the individual differences are not statistically significant between Phase 2 and Phase 3 (p = 0.96). The differences between Approved and Phase 2 or Phase 3 also do not reach the threshold of statistical significance (p = 0.08), however the median and mean *in vivo* growth scores for Approved scFvs are improved compared with Phase 2 and Phase 3 (Figure 4.21B). Furthermore, the interquartile range of Approved scFv TPBLA growth scores is tighter and shifted up compared with Phase 2 and Phase 3, which agrees with the idea that to be approved the antibodies need a certain level of stability and aggregation resistance. The approval ratings used were true as of 2017, when the paper by Jain et al. (2017) was published. Since then, many of the Phase 2 and Phase 3 mAbs have been discontinued. The difference between the TPBLA growth scores of the approved mAbs compared with the discontinued mAbs is statistically significant (Welch two sample t-test: t = 3.28, df 81.5, p = 0.0015), where the discontinued mAbs have on average a lower TPBLA score than the approved mAbs (Figure 4.21C). There are some outliers that were discontinued but have high TPBLA growth scores. However, these could have been discontinued for reasons other than aggregation or stability problems, such as potency issues or activity. Overall, this demonstrates the ability of TPBLA to identify and separate poorly behaved antibody fragments.

Assay	Description	Biophysical property assessed
HEK titer	The expression titer (mg/L) of mAb produced in HEK cells	Aggregation propensity and stability
Thermal midpoint (Tm) determination using differential scanning fluorimetry (DSF) (He et al., 2011)	mAb mixed with fluorescent dye (SYPRO orange) which is sensitive to protein unfolding. The fluorescent signal is measured over 40 °C to 95 °C in 0.25 °C per minute steps.	Thermal stability
Hydrophobic Interaction Chromatography (HIC) (Estep et al., 2015)	Retention time of mAbs on a butyl-NP5 HIC column which contains resin with hydrophobic groups. Longer retention times correlate with increased hydrophobicity.	Hydrophobicity
Standup Monolayer Adsorption Chromatography (SMAC) (Kohli et al., 2015)	Retention time of mAbs on a Zenix SEC-300 column which contains a monolayer of silica. Longer retention times correlate with poor colloidal stability.	Colloidal stability
Cross Interaction Chromatography (CIC) (Jacobs et al., 2010)	Retention time of mAbs on a column where polyclonal antibodies have been conjugated to the column matrix. Longer retention times correlate with increased polyspecificty.	Polyspecificity
Polyspecificity Reagent (PSR) binding (Xu et al., 2013)	Biotinylated membrane proteins used as the polyspecificity reagent (PSR). Binding was measured by presenting IgGs on the surface of yeast and incubating with PSR before quantifying binding using a fluorescent signal.	Polyspecificity

Table 4.4 Developability assays used to characterise the Jain dataset (Jain et al., 2017;
Willis et al., 2020)

Accelerated stability (AS)	mAbs (1 mg/mL) were incubated at 40 °C for 30 days. Timepoints were taken over the 30 days and the amount of aggregate calculated using gel filtration chromatography.	Aggregation propensity, thermal stability, shelf life
Affinity-Capture Self-Interaction Nanoparticle Spectroscopy (AC-SINS) (Liu et al., 2014)	Gold nanoparticles conjugated to anti-human Fc IgGs and incubated with mAb of interest. The absorbance spectra of the nanoparticles is measured. Aggregation of the nanoparticles by IgG self-interaction results in a wavelength shift.	Aggregation propensity, self-association
Salt-Gradient AC-SINS (SGAC-SINS) (Estep et al., 2015)	Gold nanoparticles (as above) incubated with mAb of interest then diluted in 0.3-1M ammonium sulphate. The wavelength shift in absorbance is plotted against salt concentration.	Aggregation propensity, solubility
Clone Self-Interaction by Biolayer Interferometry (CSI-BLI) (Sun et al., 2013)	mAb of interest in solution was incubated with a biosensor with surface immobilised mAb of interest. Binding was measured using bio-layer interferometry system Octet (Section 1.4)	Self-association, aggregation
Enzyme-Linked Immunosorbent Assay (ELISA) (Mouquet et al., 2010)	mAb of interest is screened for binding against multiple antigens (Cardiolipin, keyhole limpet haemocyanin, lipopolysacchaaride, ss- and ds-DNA, and insulin). Antigens immobilised on ELISA plates, incubated with mAb of interest. Binding detected using anti-human IgG-HRP.	Promiscuous binding, Polyspecificity
Baculovirus particle (BVP) assay (Hötzel et al., 2012)	As in ELISA, except uses baculovirus particles in place of antigens. Binding of mAb to BVP correlates with rapid clearance <i>in vivo</i> .	Unfavourable pharmokinetics (e.g. rapid clearance)
Extensional and shear flow device (EFD) (Willis et al., 2020)	mAb of interest is subjected to defined hydrodynamic forces which mimic those experienced in bioprocessing. The affect of these flow forces on the amount of monomeric species is quantified and compared to before flow.	Flow-induced aggregation

To understand which biophysical properties have the biggest impact on *in vivo* growth in TPBLA of the 35 late stage clinical therapeutics, this was correlated with their performance in 13 commonly employed developability assays (Table 4.4) (Jain et al., 2017; Willis et al., 2020). These assays individually probe the proteins' thermal stability, aggregation propensity, self-association, hydrophobicity, colloidal stability, hydrophobicity, polyspecificity, solubility, shelf-life, promiscuous binding, and unfavourable pharmokinetics (Jain

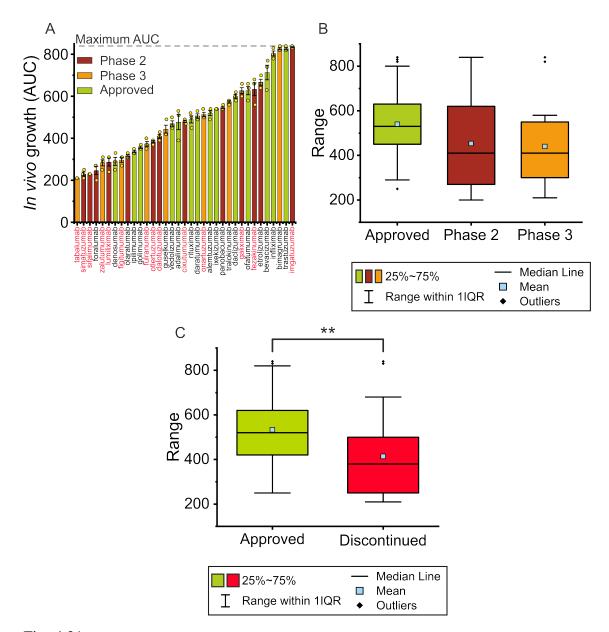


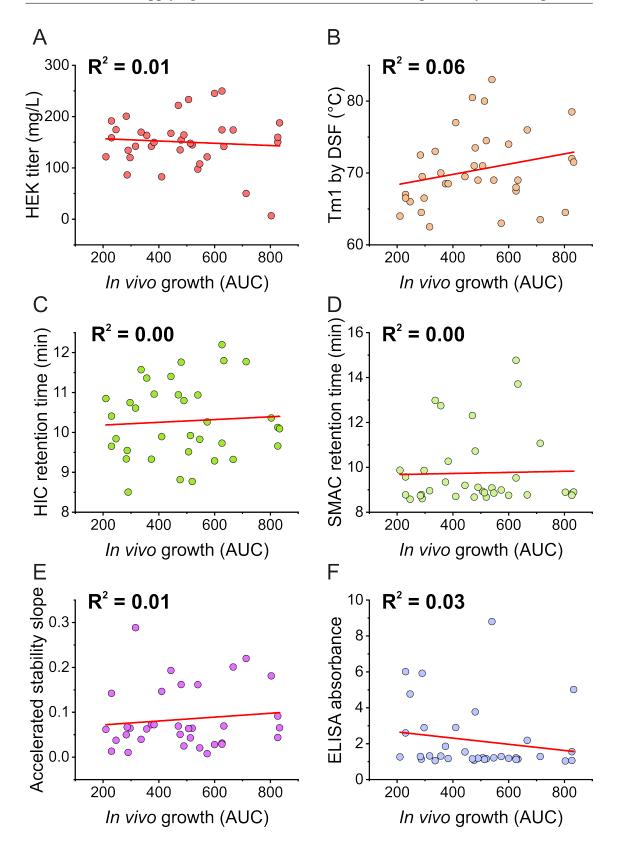
Fig. 4.21 **TPBLA analysis of 35 chosen Jain mAbs as scFvs.** A) Area under the antibiotic survival curve calculated for the scFv regions of the 35 Jain mAbs, screened at 0-140  $\mu$ g/mL ampicillin. Data points are from three independent experiments. Error bars show standard error of the mean. Bars coloured by approval rating as of 2017 (Jain et al., 2017). mAbs that have been discontinued since 2017 are coloured in red. B) Box plot showing TPBLA scores based on approval rating published by Jain et al. (2017). Outliers are values that fall outside 1 interquartile range (IQR). C) Box plot showing TPBLA scores of approved mAbs vs those that have been discontinued since Jain et al. (2017). Outliers are values that fall outside 1 interquartile range (IQR). (\*\*\*' = p < 0.01.

et al., 2017). (Figure 4.22, Figure 4.23). None of the 13 developability assays showed a correlation with TPBLA score ( $\mathbb{R}^2 < 0.15$ ). When analysed using a Spearmans rank correlation, Fab Tm by DSF had the highest correlation with TPBLA score (Spearman's rank = 0.28) (Figure 4.24A, Figure 4.23B). Although this is a weak correlation, it is similar to that seen within the AMSCI mAbs dataset. Using a Spearmans rank correlation, TPBLA shows an unexpected negative correlation with ELISA, although this again very weak (Spearman's rank = -0.27, Figure 4.23C). The Spearmans rank correlations between TP-BLA and the other 13 developability assays were used in a hierarchical clustering analysis to identify related assays. TPBLA clusters with Fab Tm by DSF, which is then most closely related to HEK titer and Salt-Gradient AC-SINS (SGAC) (Figure 4.24B). All other assays cluster together on a separate branch, indicating TPBLA is reporting on something different and novel that is not encompassed by these commonly employed developability assays (Figure 4.24B).

## 4.2.7 Multiple regression models can be used to rationalise TPBLA score

In Section 4.2.3 we attempted to explain TPBLA score by performing a multiple regression model and identifying parameters that influence the *in vivo* growth of the AMSCI mAbs. However, the dataset was too small for statistically significant analysis. The larger size of the Jain dataset enables more statistically robust analysis of correlations between TPBLA and these 13 developability assays, all of which probe particular and individual biophysical characteristics.

A multiple linear regression model was used to test which parameters can significantly predict TPBLA score. An initial model included 7 assays: Fab Tm by DSF, Standup Monolayer Adsorption Chromatography (SMAC), Accelerated Stability (AS), Polyspecificity Reagent (PSR) binding, Cross Interaction Chromatography (CIC), theoretical pI, and scFv molecular weight. The overall regression was statistically significant (f(7, 27) = 3.968, R<sup>2</sup> = 0.51, r = 0.71, p = 0.004) (Figure 4.25A). Detailed statistics for the model can be found in Appendix D, Table D.1. Individual parameters were removed systematically to find the most parsimonious model. Removing AS gives a statistically significant model (f(6, 28) = 3.913, R<sup>2</sup> = 0.46, r = 0.68, p = 0.005). Detailed statistics for the model can be found in Appendix D, Table D.2. However, using this model PSR and molecular weight do not significantly predict TPBLA score (p > 0.1) (Figure 4.25B).



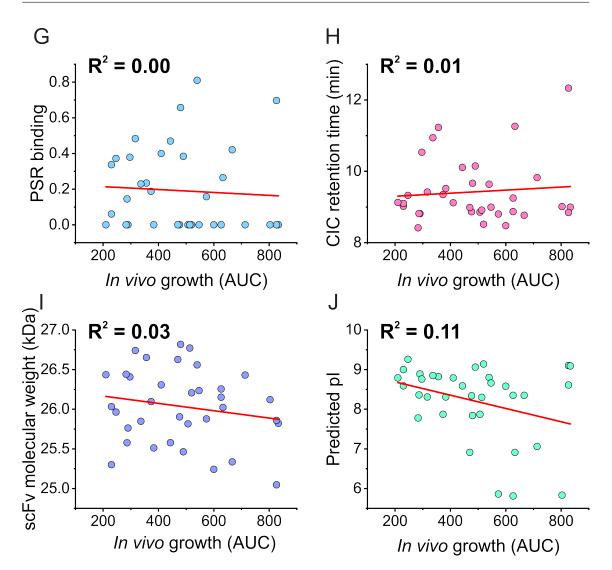


Fig. 4.22 **TPBLA** plotted against biophysical characterisation of 35 Jain mAbs.*in vivo* growth score from TPBLA plotted against (A) HEK titer, (B) Fab transition mid-point temperatures (Tm) by DSF, (C) Hydrophobic Interaction Chromatography (HIC), (D) Standup Monolayer Adsorption Chromatography (SMAC) retention time, (E) Accelerated stability slope, (F) Enzyme-Linked Immunosorbent Assay (ELISA), (G) Polyspecificity Reagent (PSR) binding, (H) Cross Interaction Chromatography (CIC) retention time, (I) scFv molecular weight, and (J) theoretical pl. Data for developability assays plotted against TPBLA from A-J was taken from Jain et al. (2017). Linear regression was performed using Orign pro.

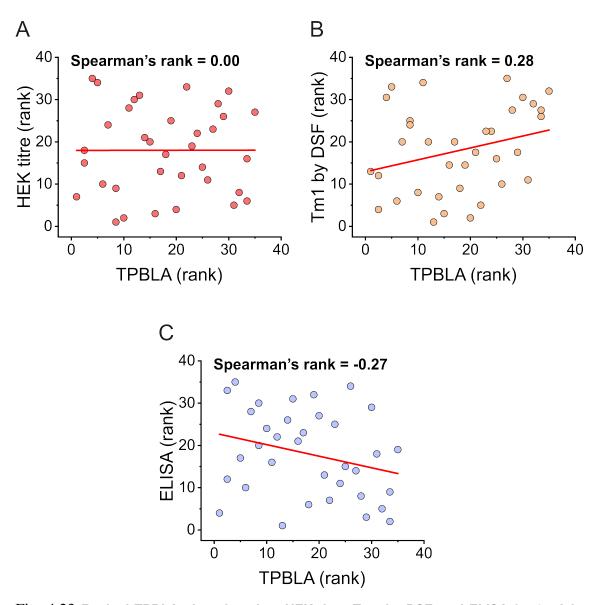


Fig. 4.23 Ranked TPBLA plotted against HEK titre, Tm1 by DSF, and ELISA for 35 Jain mAbs. Ranked *in vivo* growth score from TPBLA plotted against ranked (A) HEK titer, (B) Fab transition mid-point temperatures (Tm) by DSF, (C) Enzyme-Linked Immunosorbent Assay (ELISA). HEK titre, Tm, and ELISA data taken from Jain et al. (2017). Spearmans rank and linear regression was performed using Orign pro.

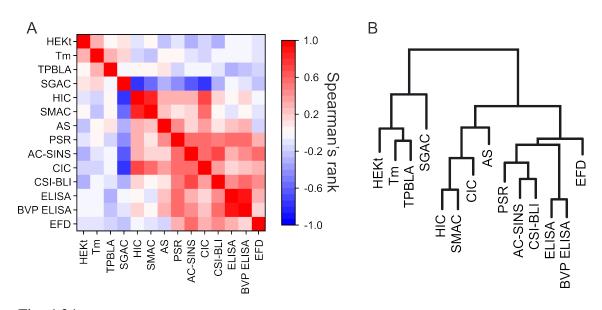


Fig. 4.24 Spearmans rank correlation and hierachical clustering of developability assays compared with TPBLA for 35 Jain mAbs. A) Matrix showing Spearmans rank correlation of developability assays. B) Hierachical clustering of developability assays.

Removing molecular weight gives a statistically significant regression model (f(5, 29) = 4.116, R<sup>2</sup> = 0.42, r = 0.64, p = 0.006) where all parameters significantly predict TPBLA score (p < 0.1) (Figure 4.25C). Detailed statistics for the model can be found in Appendix D, Table D.3.

In agreement with the models generated with the AMSCI mAbs dataset, this model demonstrates that Tm and pI have the most significant impact on TPBLA score (p < 0.01). The link between TPBLA and pI is likely due to protein solubility, in agreement with the data presented in Chapter 3. CIC and SMAC have a significant impact on TPBLA score (p < 0.05), likely due to them measuring unwanted protein-protein interactions which can lead to protein aggregation. Similarly, PSR has a moderately significant impact on TPBLA score (p < 0.1).

Using a Spearmans rank to assess correlations between the ranked predicted and experimental TPBLA scores for the models gave similar R<sup>2</sup> values as the unranked data (Figure 4.26).

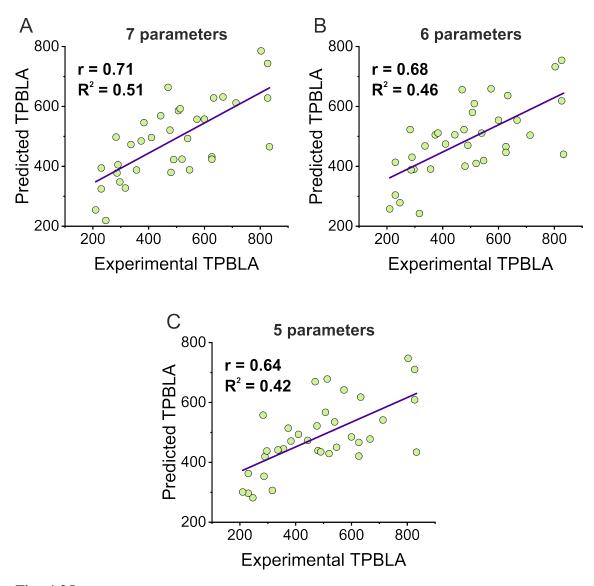


Fig. 4.25 Multiple regression models to predict TPBLA score from biophysical parameters. Experimental TPBLA score plotted against predicted TPBLA score using multiple regression models utilising different developability assays and biophysical parameters. A) Uses Fab Tm by DSF, Standup Monolayer Adsorption Chromatography (SMAC) retention time, Accelerated Stability (AS), Polyspecificity Reagent (PSR) binding, Cross Interaction Chromatography (CIC) retention time, theoretical pl, and scFv molecular weight. B) Removes AS and uses only Tm, SMAC, PSR, CIC, pl, and molecular weight. C) The most parsimonious model. Removes molecular weight and uses only Tm, SMAC, PSR, CIC, and pl.

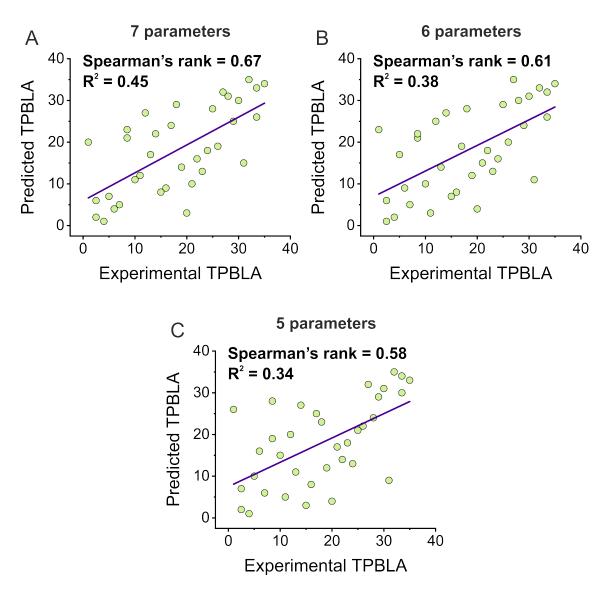


Fig. 4.26 **Multiple regression models to predict TPBLA score from biophysical parameters.** Ranked Experimental TPBLA score plotted against ranked predicted TPBLA score using multiple regression models utilising different developability assays and biophysical parameters. A) Uses Fab Tm by DSF, Standup Monolayer Adsorption Chromatography (SMAC) retention time, Accelerated Stability (AS), Polyspecificity Reagent (PSR) binding, Cross Interaction Chromatography (CIC) retention time, theoretical pl, and scFv molecular weight. B) Removes AS and uses only Tm, SMAC, PSR, CIC, pl, and molecular weight. C) The most parsimonious model. Removes molecular weight and uses only Tm, SMAC, PSR, CIC, and pl.

# 4.2.8 Multiple regression models can predict TPBLA score based on performance in other developability assays

To assess the potential for this multiple regression model to accurately predict TPBLA score, 6 mAbs were removed at random from the dataset. A multiple regression was performed to generate a model using Tm, SMAC, PSR, CIC, and pI to predict TPBLA score. The model was statistically significant (f(5, 23) = 3.375,  $R^2 = 0.42$ , r = 0.64, p = 0.02) (Figure 4.27A). Detailed statistics for the model can be found in Appendix D, Table D.4. This was then used to predict TPBLA for the 6 random mAbs that had been removed from the dataset (Figure 4.27B). The predicted TPBLA scores for the 6 mAbs correlated well with the experimental TPBLA scores ( $R^2 = 0.65$ , r = 0.78, Spearmans rank = 0.83).

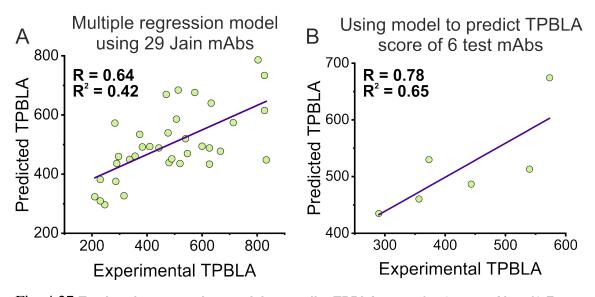


Fig. 4.27 Testing the regression model to predict TPBLA score for 6 test mAbs. A) Experimental TPBLA score plotted against predicted TPBLA score using multiple regression models utilising Tm, SMAC, PSR, CIC, and pl. 6 test mAbs were removed from the dataset of 35 Jain mAbs to create the model. B) The model using the 29 mAbs was used to predict the TPBLA score for the 6 test mAbs.

These multiple linear regression models are not able to perfectly predict TPBLA score. Clearly TPBLA reports on a complex set of metrics, including; thermal stability, protein aggregation, solubility, colloidal stability, and self-association. The fact that the models are unable to perfectly predict TPBLA score is likely that TPBLA is probing novel characteristics these developability assays do not monitor, highlighting its potential for use as a developability assay within the bioprocessing pipeleine. While TPBLA is not always able to identify the best antibodies, it is able to distinguish problematic sequences and would be a useful tool for identifying problematic molecules from a panel of variants prior to protein purification.

## 4.3 Discussion

Biopharmaceutical aggregation can occur at any stage of its lifetime, from expression in cell culture through to purification, formulation, transportation, and storage (Willis et al., 2020; Fukuhara et al., 2021). The resulting particulates can compromise the safety and efficacy of the final product (Starr et al., 2021). Therefore, protein aggregation can be a major hinderance to biopharmaceutical development as the identification and removal of these aggregates can be both costly and time-consuming (Jiskoot et al., 2012; Wolf Pérez et al., 2019). It is therefore of interest to the biopharmaceutical industry to develop methods to identify aggregation-prone sequences early in the developmental pipeline in order to minimise unnecessary time and expense. However, identification of aggregation-prone regions (APRs) can be difficult as these may be buried and only exposed upon partial or full unfolding of the protein, which may only occur under certain stresses or during manufacture (Wang and Roberts, 2018; Eyes et al., 2019).

There are many techniques available for identifying protein aggregates and assessing the developability of a biopharmaceutical candidate, as discussed in Section 1.4. However, this still remains a significant feat as it requires extensive purification of many different candidates and characterisation using a plethora of developability assays. In contrast, TPBLA could be employed as a developability screen following affinity maturation to filter out poorly developable candidates prior to protein purification as it can screen a large number of variants relatively quickly. Unlike some other assays that measure protein solubility, TPBLA enables the proper formation of disulfide bonds within antibody fragments as it is carried out in the E. coli periplasm (Cabantous and Waldo, 2006; Morell et al., 2011; Espargaró et al., 2012; McLure et al., 2022). TPBLA uses no perturbant to accelerate aggregation or destabilise the test protein, such as increased temperature, pH, or chemical denaturant, so it measures the innate stability of the molecule within a cellular environment. The multiple regression models presented in this chapter demonstrate TPBLA is influenced by a multitude of factors, including; thermal stability, solubiltiy, aggregation propensity, self-association, and colloidal stability. Therefore, TPBLA may give a more rounded readout of a molecule compared with developability assays that measure only a single characteristic.

This chapter focusses mainly on IgGs, as the biopharmaceutical sector is currently dominated by mAbs (Walsh, 2018; Khetan et al., 2022). We demonstrate how TPBLA is able to rank a panel of antibodies and identify problematic sequences. All of the AMSCI mAbs, apart from AMS197 and AMS214, score worse than the AMS106 aggregation-prone control (infliximab), consistent with the fact they were designed to have developability

problems. All of the variants that have low Tm and Tonset, high HIC retention times, or large AC-SINS wavelength shifts score below AMS106 in TPBLA. This suggests TPBLA could be used to identify these problematic variants. Characterising multiple approved mAbs would be useful to ascertain a relevant threshold TPBLA score, below which would be considered 'poorly developable' and discarded. AMS197 and AMS214 both have similar Tm, Tonset, and HIC retention times to the aggregation-resistant control AMS122 (trastuzumab), with AMS197 having a higher Tonset (65.09 °C) than AMS122 (63.51 °C). Unlike all the other AMSCI mAbs, they both score better than AMS106 in TPBLA. This is likely due to a combination of improved stability, aggregation-resistance, and solubility.

A downside of TPBLA is that it only measures the scFv. If the aggregation of a mAb is due to its framework regions, this will not be represented in the assay. This is demonstrated by the difference in AC-SINS wavelength shift between therapeutic infliximab and AMS106 (infliximab scFv on a different scaffold), where therapeutic infliximab shows a wavelength shift of 35.85 nm compared with AMS106 shifting 0.64 nm. The aggregation mechanism of therapeutic infliximab has been shown to be as a result of both Fab-Fab interactions in a head-to-tail conformation as well as Fab-Fc interactions (Lerch et al., 2017; Domnowski et al., 2021). As AMS106 only contains the variable domains of therapeutic infliximab with different constant regions, it is possible this disrupted the interfaces involved in these aberrant interactions.

We demonstrate the evolution methodology developed in Section 3 can be applied to therapeutic scaffolds, however evolution of an already stable variant (AMS197) likely requires a high mutation rate library and to be evolved at a higher selection pressure (Drummond et al., 2005). For AMS134, hotspot residues on the surface were generally substituted for less hydrophobic residues, consistent with the idea that aggregation could be driven through self-association via hydrophobic surface patches. These hotspots were mainly clustered in and around the HCDR3, highlighting a potential affinity-stability trade-off. This could potentially be problematic, as evolution using TPBLA does not include a selection for binding affinity, which may result in evolved antibodies that no longer bind to their therapeutic targets. This issue is investigated further in Chapter 5. Characterising three evolved point mutants of AMS134 (R93G, M123K, and V124D) indicates that the absolute intensity of the log<sub>2</sub>(fold change) is indicative of the mutants beneficial capacity. AMS134<sup>R93G VL</sup> was below both  $2\sigma$  thresholds from normalised mutation frequency and log<sub>2</sub>(fold change) and displayed increased aggregation propensity in both SLS and AC-SINS. AMS134<sup>M123K VH</sup> was below the  $2\sigma$  threshold for  $\log_2(\text{fold change})$ , but above the threshold for normalised mutation frequency, but displayed similar aggregation behaviour to AMS134<sup>WT</sup> in both SLS and AC-SINS. AMS134<sup>V124D VH</sup> was above both  $2\sigma$  thresholds from normalised mutation frequency and  $log_2$ (fold change) and displayed reduced self-association in AC-SINS as well as abolished aggregation at high temperatures in SLS.

The error-prone PCR library size was estimated to be  $\sim 2.4 \times 10^{11}$  for AMS134 and  $\sim 1.9 \times x \ 10^{11}$  for AMS197, based on the number of colony forming units growing on an agar plate. This is likely an overestimation, as every individual colony is unlikely to contain a unique clone. The sequencing methodology using Illumina shotgun libraries is limited as it is unable to identify multiple mutants, but it can be used to identify hotspot residues and APRs within the molecule. The libraries generated in the chapter were designed to have an average of one mutation per gene, therefore limiting the potential for improvement as multiple mutations are often required for significat improvements in protein stability or aggregation-resistance (Yu and Dalby, 2018). To identify significantly improved variants, a library with a high mutation frequency should be used. This library could then be plated at a high selection pressure, and individual colonies picked and assessed using TPBLA before sequencing to identify significantly improved variants. In this instance, NGS using the shotgun library approach developed in Section 3 would still be a useful step in measuring the mutational frequency accross the whole molecule, identifying APRs, and seeing the overall frequency of individual mutations to give an idea of which ones are found more often and therefore are influencing TPBLA the most.

The Jain dataset assessing 35 late stage clinical mAbs using TPBLA demonstrated there was no significant difference in TPBLA score between the approval ratings (Phase 2, Phase 3, or Approved). However, this was using the approval rating published in the 2017 paper by Jain et al. (2017). Since then, many of the mAbs that were in Phase 2 or Phase 3 clinical trials have been discontinued. Comparing the TPBLA score between these Discontinued mAbs and the approved mAbs showed a statistically significant difference, where the Discontinued mAbs showed a lower TPBLA score compared with the Approved mAbs. There were some mAbs that had high scores in TPBLA but have been discontinued, however this could be due to issues unrelated to stability or aggregation, such as potency issues or activity. This demonstrates how TPBLA could be used to identify poorly developable candidates from a pool of variants. It is important to note that all the Jain Abs are late-stage therapeutics, and are unlikely to have any significant developability issues. If used early in the developmental pipeline, prior to protein purification, TPBLA could be used to identify sequences that do have significant developability issues (such as low solubility, high aggregation, high self-association, or low stability) before they reach pre-clinical trials.

The multiple regression models using a combination of classic developability assays and biophysical parameters are able to reasonably predict TPBLA score for the Jain mAbs. The most parsimonious model included Tm, SMAC, PSR, CIC, and pI, all of which have a significant impact on predicting TPBLA score. This is the first example of evidence that TPBLA is influenced by a combination of thermal stability, self-association, colloidal stability, and solubility and could therefore be useful at ranking test proteins based on these characteristics. The model is not perfect, indicating TPBLA is measuring some other paramater that is not measured by any of the other 13 developability assays tested in this study. Therefore, alongside a combination of these parameters that affect a proteins' developability, TPBLA is measuring something novel which highlights the relevance of using TPBLA to rank candidate mAbs. Furthermore, whatever TPBLA is reporting on is clearly important as it is able to distinguish between the late-stage clinical failures and approved mAbs in the Jain dataset (Figure 4.21). The success of these multiple regression models indicate experimental and predicted biophysical parameters could be used to design a machine learning model to predict TPBLA. The data generated in this chapter could be used to inform such a model, and could be used to gain a better understanding of what properties TPBLA is measuring that the other developability assays are not.

In summary, the data presented in this chapter demonstrates the potential of TPBLA for screening and ranking candidate mAbs based on their developability. The high-throughput directed evolution methodology developed in Section 3 can be used to selectively evolve antibody fragments to improve their developability, however to gain significant improvements it likely requires a higher mutation rate library than those used in this study. Additionally, the directed evolution methodology enables hotspot residues involved in mediating protein aggregation to be identified, which could be used to guide rational design of better biologics. The multiple regression models highlight the influence of thermal stability, self-association, colloidal stability, and solubility on TPBLA, demonstrating the relevance of the *in vivo* growth score to mAb developability. The fact that TPBLA can be reasonably explained using these experimental and predictive parameters suggests this data could be used to inform a novel machine learning model to predict TPBLA score.

# **Chapter 5**

# Towards simultaneous improvement of both aggregation and binding with TPBLA

### 5.1 Introduction

It is well known in the protein engineering field that there is a trade-off between different biophysical properties in proteins, such as between stability and function (McLure et al., 2022). Engineering of biopharmaceuticals is generally to enhance a number of drug-like properties, such as; low clearance rates, low self-association (homotypic interactions) or aggregation, low off-target binding (heterotypic interactions), high stability, high solubility, and low viscosity at high concentrations (Starr and Tessier, 2019). Arguably the most important property is that the biopharmaceutical must bind to its target and produce the desired effect. Often affinity-matured antibodies have decreased stability or increased aggregation, due to the trade-off between different properties - when you evolve one you lose another (McLure et al., 2022; Rabia et al., 2018). There have been a number of studies detailing this trade-off during directed evolution, particularly between affinity and stability or specificity (Julian et al., 2017; Tiller et al., 2017b; Stimple et al., 2020). In Chapters 3 and 4, TPBLA is used to assess and evolve both increased solubility and reduced aggregation in a variety of different proteins, including biopharmaceuticals. However, selecting for these properties while neglecting function could result in evolved antibodies that no longer bind to their target. A way to address this would be to modify the assay to include a selection for binding. This chapter investigates the possibility of introducing a

selection for function into TPBLA to create the Solubility 'n' Affinity Coselection (SnAC) assay. We assess the potential for this new assay to both screen and select for binding affinity alongside beneficial biophysical characteristics.

#### 5.1.1 Aims of the study

This chapter adapts the previously developed TPBLA (Chapter 3) to introduce a coselection for cognate binding alongside the selection for aggregation resistance to create the SnAC assay by correlating binding to expression of a fluorescent reporter protein.

### 5.2 Results

#### 5.2.1 Split fluorescent proteins as sensors

Various fluorescent proteins have been utilised as biosensors and correlated with protein stability, solubility, or its ability to interact with a target (Ghosh et al., 2000; Lindman et al., 2010; Magliery et al., 2005; Golinski et al., 2021). Green fluorescent protein (GFP) can be split into two halves which are able to fluoresce upon interaction; if the two halves are fused to two different proteins, the ensuing fluorescent signal can be correlated with proteinprotein interactions (Baird et al., 1999; Ghosh et al., 2000). The  $\beta$ -lactamase construct in TPBLA is directed to the periplasm, therefore a fluorescent protein that is active in this oxidising environment is required. GFP is inactive if translocated to the periplasm prior to folding, however the red fluorescent mCherry and green fluorescent mNeonGreen are active in the periplasm (Dammeyer and Tinnefeld, 2012). A red fluorescent mCherry was engineered for improved stability and fluorescence, dubbed superfolder mCherry or sfCherry, however its split version has low levels of fluorescence (Feng et al., 2017). This split variant was engineered to create sfCherry2 which showed a 10-fold increased fluorescence compared to its split superfolded mCherry counterpart (Feng et al., 2017). This study also reported an engineered split mNeonGreen variant (split-mNeonGreen2 or mNG2) that has improved background to noise ratio compared with superfolder GFP (sfGFP). These engineered split proteins were selected in this chapter to assess the potential for adding a selection for function to the TPBLA. This same group that designed split sfCherry2 and mNG2 published a study in 2019 with further work on their split sfCherry2 construct (Feng et al., 2019). Here they showed the limiting factor to fluorescence in these split fluorescent proteins is the association of the two split fragments. To improve

their split FP they used a SpyTag/SpyCatcher interaction where they fused SpyCatcher to  $sfCherry2_{1-10}$  and SpyTag to  $sfCherry2_{11}$  to enhance the association of the fluorescent protein fragments. This demonstrates how fusing two fragments of a split fluorescent protein (sfCherry2) to two proteins that bind together (SpyTag/SpyCatcher) can be used to link protein-protein interactions to a fluroescent output.

To assess the potential of introducing a selection for binding affinity into TPBLA, the HA4 monobody was utilised as a model system as a single point mutation Y87A can inhibit binding while maintaining stability (Figure 5.1) (Wojcik et al., 2010). HA4 binds to the SH2 domain of AB1 kinase with high affinity ( $K_d \sim 7$  nM), the oncogenic counterpart of which is implicated in chronic leukaemia (Wojcik et al., 2010). HA4 and SH2 are both around 10 kDa single domain proteins, representing a simple model system to use when designing our dual selection assay. Furthermore, the HA4:SH2 system has been used previously to develop techniques for the directed evolution of protein binding affinity (Wang et al., 2018; Morrison et al., 2021). The assay was designed so that the fluorescence intensity of the split fluorescent protein would be correlated to the binding affinity between the test proteins (HA4:SH2). Adding one fragment of the split fluorescent protein, and the other corresponding fragment as a fusion protein with SH2 domain, cells expressing both constructs could be sorted based on the fluorescence signal of the split fluorescent protein, which would increase as a result of increased HA4/SH2 interaction (Figure 5.2).

One issue would be if there was differential expression of  $\beta$ -lactamase fusion between different cells. To correct for this, another fluorescent protein (mScarlet-I) was included on the same transcript as the tripartite  $\beta$ -lactamase construct. This protein would remain in the cytoplasm, and its fluorescence intensity would be correlated to the expression of the tripartite  $\beta$ -lactamase construct. While this would not account for post-translational changes in expression levels between cells, such as due to protease degradation, this is not an issue as it is exactly the kind of instability we aim to be able to test for using this assay. With these two different fluorescent signals, two colour FACS could be used to correct for changes in the split fluorescent protein emission as a result of differential expression.

# 5.2.2 Split mNeonGreen2 combined with TPBLA is not able to detect binding affinity *in vivo*

In order to introduce a selection for binding into the TPBLA, binding was linked to a fluorescent output. As described in Section 5.2.1, a split fluorescent biosensor was utilised

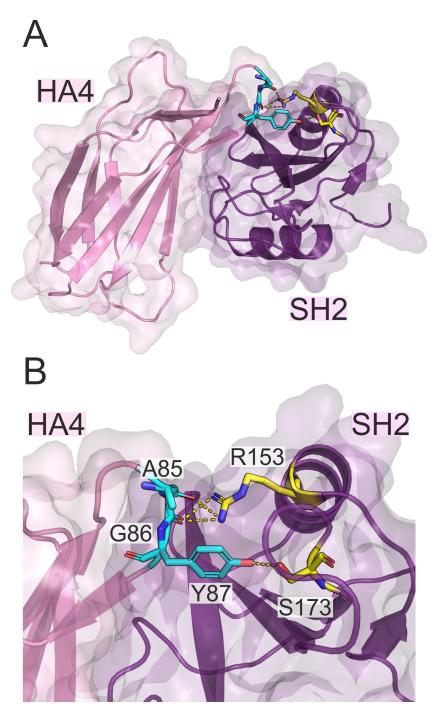


Fig. 5.1 Crystal structure of the HA4 monobody bound to the SH2 domain of human Ab1 kinase A) HA4 (pink) bound to Ab1 SH2 domain (purple). The residues involved in the binding interaction are highlighted. B) Zoomed view of the binding interaction. A85, G86, and Y87 of HA4 interact with R153 and S173 of Ab1 SH2. Polar interactions are shown by dashed yellow lines. Mutating Y87 in HA4 to alanine removes key interaction with S173 of Ab1 SH2, thereby reducing the binding affinity (Wojcik et al., 2010). PDB ID 3K2M.

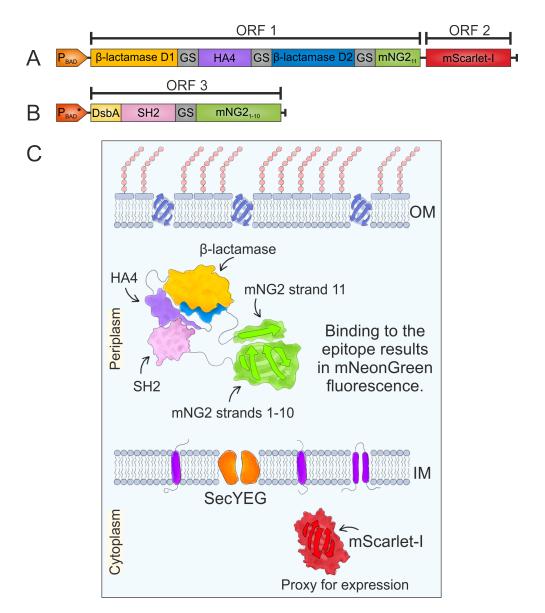


Fig. 5.2 Overview of Solubility 'n' Affinity Coselection (SnAC) 1.0 to simultaneously assess affinity and aggregation propensity. A) The DNA sequence encoding the 11th strand of a split fluorescent protein (mNeonGreen2, mNG2) is added onto the C-terminal end of the TPBLA construct (with the HA4 monobody as the test protein) via a flexible linker under the control of the pBAD promoter. mScarlet-I is expressed as the second cistron in a bicistronic construct to provide a proxy for expression and allow two colour FACS to correct for different levels of mNG2 fluorescence as a result of differential expression rather than improved binding affinity and complementation. The two open reading frames (ORF) are shown as ORF 1 and ORF 2. B) A second construct expressing SH2 (the binding partner of HA4) fused to the corresponding strands 1-10 of the split fluorescence protein via a flexible linker is controlled via the pBAD weak promoter (pBAD<sup>\*</sup>), a variant of pBAD with a single point mutation incorporated to reduce its sensitivity to arabinose and therefore reduce expression levels. This is to ensure the construct does not aggregate in the cytoplasm. This construct is directed to the periplasm via the DsbA signal sequence. C) Both fusion constructs are directed to the periplasm. Binding of HA4 to SH2 brings the two split fluorescent protein fragments in close proximity, forming the active fluorescent protein and giving a readout that can be assessed using FACS.

to achieve this to develop SnAC 1.0 (Figure 5.2). The 11th strand of mNeonGreen2 (mNG2<sub>11</sub>) was fused to the C-terminus of the TPBLA construct (TPBLA-mNG2<sub>11</sub>), which was expressing the HA4 monobody as the test protein (blaHA4-mNG2<sub>11</sub>). A second construct, expressed under the control of the pBAD weak promoter, contains strands 1-10 of mNG2 (mNG2<sub>1-10</sub>) fused to the SH2 domain of human AB1 kinase, the epitope for HA4. mNG2<sub>1-10</sub> can fold into a barrel but cannot form the fluorescent chromophore without fusing with mNG2<sub>11</sub>. Binding of the HA4 monobody in TPBLA-mNG2<sub>11</sub> to the SH2 domain fused to mNG2<sub>1-10</sub> would bring these two fragments into contact, enabling them to come together and form the complete fluorescent protein and give a fluorescent signal (Figure 5.2). This would enable binding of two proteins in the periplasm to be correlated with a fluorescent signal. To measure this, the wild type HA4 monobody (HA4<sup>WT</sup>) and the reduced binding point mutant Y87A (HA4<sup>Y87A</sup>) were utilised (Wojcik et al., 2010). HA4<sup>Y87A</sup> was designed by removing the key residue in binding SH2 and showed no binding when using a fluorescence polarization competition assay (Wojcik et al., 2010). A second fluorescent protein (mScarlet-I) was expressed bicistronically with TPBLA. This would enable us to correct for expression levels, where there may be increases in the fluorescent signal but only due to increases in expression.

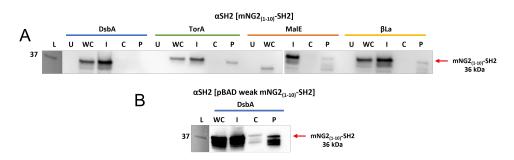


Fig. 5.3 Western blot against SH2 showing SH2-mNG2<sub>1-10</sub> localisation in the cell. L, ladder; U, uninduced; WC, whole cell; I, insoluble; C, cytoplasm; P, periplasm. A) Levels of SH2-mNG2<sub>1-10</sub> in different cellular fractions when directed to the periplasm using different signal sequences and a pBAD WT promoter. B) Reducing the overall expression level using a pBAD weak promoter increases soluble protein expression in the periplasm.

The SH2-mNG2<sub>1-10</sub> construct was expressed under the pBAD weak promoter and directed to the periplasm using a DsbA signal sequence, which enables co-translational export of a passenger protein via the Sec pathway (Schierle et al., 2003). The reduction in expression levels compared with pBAD wild-type prevented the fusion protein from just aggregating in the cytoplasm and not being exported to the periplasm (Figure 5.3). A TorA signal peptide with the wild-type pBAD promoter gave low levels of export to the periplasm, however this utilises the more complex and less well-studied Tat pathway exporting folded proteins, compared with the more commonly used Sec pathway exporting unfolded proteins either co- or post- translationally (Palmer and Berks, 2012). However, cells expressing the

blaHA4<sup>WT</sup>-mNG2<sub>11</sub> and blaHA4<sup>Y87A</sup>-mNG2<sub>11</sub> constructs alongside mNG2<sub>1-10</sub> showed little difference in their fluorescent signal after 3hr expression (Figure 5.4A). Furthermore, the positive control expressing split mNG2 (mNG2<sub>1-10</sub> and mNG2<sub>11</sub> fused via a flexible linker) targeted to the periplasm using a DsbA signal sequence had a slightly red-shifted emission spectra compared with the HA4<sup>WT</sup> and HA4<sup>Y87A</sup> (Figure 5.4A). Two negative controls were designed where either mNG2<sub>1-10</sub> or mNG2<sub>11</sub> was deleted from the HA4<sup>WT</sup> plasmid, creating  $\Delta$ mNG2<sub>1-10</sub> and  $\Delta$ mNG2<sub>11</sub>, respectively. Unfortunately, cells expressing both of these negative controls gave similar fluorescence emission spectra and intensities as the those expressing the HA4<sup>WT</sup> and HA4<sup>Y87A</sup> constructs (Figure 5.4B). The fluorescent signal over time of cells expressing a positive control (expressing split mNG2), negative control (blaGG<sub>STOP</sub>-mNG2<sub>11</sub>, only expresses the first domain of β-lactamase due to a premature stop codon and frame shift mutation between the G/S linkers, so does not express mNG2<sub>11</sub>), alongside HA4<sup>WT</sup> and HA4<sup>Y87A</sup>, growing at 37 °C over 400 minutes was measured, which showed no difference in fluorescence signal between induced and uninduced cells from either HA4<sup>WT</sup> and HA4<sup>Y87A</sup> (Figure 5.4C).

The low fluorescence signal of the periplasmic split mNG2 positive control, as well as the lack of fluorescence output from the HA4<sup>WT</sup>, highlighted the need for a stronger reporter protein. As mNG2 is the brightest split fluorescent protein developed with the ability to be periplasmically expressed, the assay needed to find a way to link binding in the periplasm to a cytoplasmic reporter protein.

#### 5.2.3 CadC periplasmic sensor for screening binding affinity

For the split fluorescent protein system to work, the pool of potential reporters is limited by its need to be active in the oxidising environment of the periplasm. Furthermore, the fluorescence intensity of split fluorescent proteins is never as high as their intact counterparts, which could limit the potential dynamic range of the assay (Feng et al., 2017). An alternative option would be to use an intact fluorescent reporter in the *E. coli* cytoplasm. To do this, protein binding in the periplasm needed to be linked to to mRNA transcription in the cytoplasm. The *E. coli* transmembrane transcriptional activator CadC has been successfully utilised to screen and evolve binding affinity in therapeutic scaffolds (Morrison et al., 2021). As described in Section 1.6.2, CadC is made up of an N-terminal cytosolic DNA binding domain (DBD) and a C-terminal periplasmic pH sensor domain (Lindner and White, 2014; Chang et al., 2018). It senses acidic pH and high lysine levels in the periplasm, causing the periplasmic sensor domain to dimerise and bind two motifs on the CadBA promoter and initiate gene transcription (Kuper and Jung,

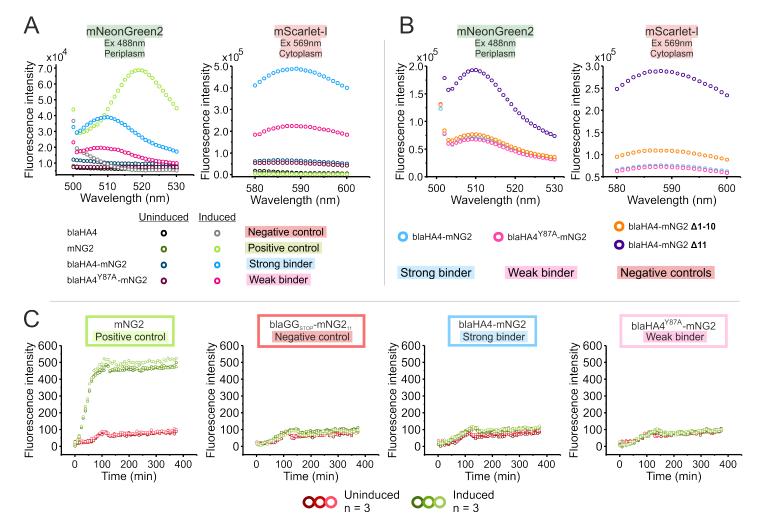


Fig. 5.4 SnAC 1.0 using a split fluorescent protein is not able to measure binding affinity *in vivo*. A) Fluorescence spectra of periplasm and cytoplasm fractions from cells grown at 37 °C, 200 rpm expressing SH2-mNG2<sub>1-10</sub> and either blaHA4<sup>WT</sup>-mNG2<sub>11</sub> (strong binder) or blaHA4<sup>Y87A</sup>-mNG2<sub>11</sub> (weak binder) before and after a 3 hr induction. blaHA4 and mNG2 were measured as negative and positive controls, respectively. mNG2 and mScarlet-I were excited at 488 and 569 nm, respectively. B) Control fluorescence spectra of periplasm and cytoplasm fractions from cells grown at 37 °C, 200 rpm expressing SH2-mNG2<sub>11</sub> or blaHA4<sup>WT</sup>-mNG2<sub>11</sub> or blaHA4<sup>WT</sup>-mNG2<sub>11</sub> before and after a 3 hr induction. blaHA4<sup>WT</sup>-mNG2<sub>11</sub> or blaHA4<sup>WT</sup>-mNG2<sub>11</sub> before and after a 3 hr induction. blaHA4<sup>WT</sup>-mNG2<sub>11</sub> or blaHA4<sup>WT</sup>  $\Delta$ mNG2<sub>11</sub> were used as negative controls. C) mNG2 fluorescence signal over time from uninduced and induced cells expressing blaHA4<sup>WT</sup>-mNG2<sub>11</sub> or blaHA4<sup>Y87A</sup>-mNG2<sub>11</sub> or blaHA4<sup>Y87A</sup>-mNG2<sub>11</sub>. Cells were grown overnight at 37 °C, 200 rpm.

2005). By replacing the periplasmic sensor domain with an caffeine-inducible dimerising nanobody (VHH), and placing sfGFP under the control of the CadBA promoter, gene transciption can be switched on via addition of caffeine (causing the dimerisation of CadC and enabling it to bind to CadBA) (Chang et al., 2018) (Figure 5.5). We hypothesised the system could be adapted and combined with TPBLA to enable dual screeing and evolution of developable characteristics alongside binding affinity within therapeutic proteins. This would work by initially fusing a C-terminal caffeine-inducible dimerising nanobody to the TPBLA construct, which would contain the HA4 monobody between the two domains of  $\beta$ -lactamase (Figure 5.5A). The periplasmic sensor domain of CadC would be replaced by the SH2 domain, and sfGFP would be placed under the control of the CadBA promoter (Figure 5.5B). Upon addition of caffeine, the VHH domain on the C-terminus of TPBLA would dimerise, enabling the now dimeric construct to bind to the SH2 domain on CadC and cause CadC to dimerise (Figure 5.5D). This allows CadC to bind to the CadBA promoter and induce transciption of sfGFP, therefore enabling the fluorescence intensity of sfGFP to be correlated to the binding affinity of HA4 for SH2 (Figure 5.5D). Cells could be sorted by FACS pre- or post-evolution to separate out variants that no longer bind to their target, with the intention in the future to carry out the evolution and FACS within a single experiment.

### 5.2.4 Adapting the *E. coli* CadC transmembrane transcriptional activator to include a selection for binding affinity into TPBLA

As described in Section 5.2.3, the wild-type *E. coli* transcriptional activator CadC has the potential to be exploited to link protein-protein interactions in the periplasm to gene expression in the cytoplasm (Figure 5.5). For SnAC 2.0, a construct from a previous study which replaced the transmembrane domain with an artificial transmembrane domain made of 16 leucine repeat residues, Leu(16), was utilised (Chang et al., 2018). This has previously been shown to enable the formation of correctly oriented chimeric CadC into the inner membrane (Chang et al., 2018; Lindner and White, 2014). Previous studies using CadC with the periplasmic sensor domain replaced with the dimerising leucine zipper domain (forcing dimerisation and therefore gene expression) demonstrated higher activity when using a Leu(16) transmembrane domain compared with the wild type transmembrane domain of CadC was replaced with a caffeine inducible nanobody (VHH) to create CadC-VHH (Chang et al., 2018). sfGFP was placed under the control of the CadBA promoter, thereby enabling gene transcription to be switched on via the addition of

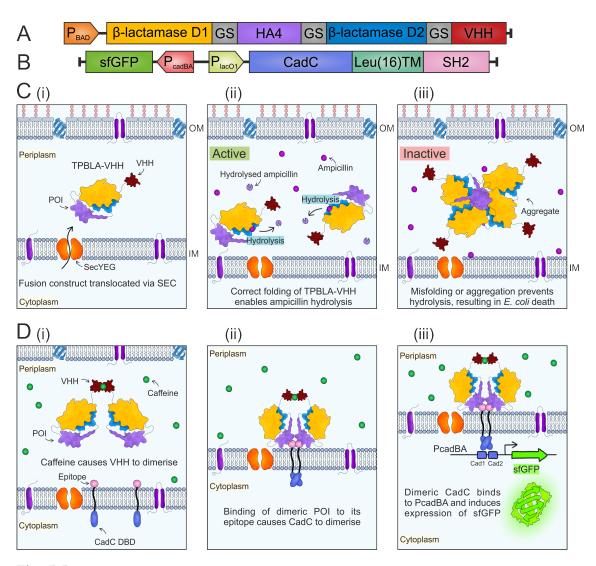


Fig. 5.5 Overview of TPBLA-SnAC 2.0 to simultaneously assess affinity and aggregation propensity. A) A caffeine-inducible dimerising nanobody (VHH) is fused to the C-terminus of the TPBLA construct, which contains the HA4 monobody within the two domains of  $\beta$ -lactamase. B) A second plasmid where SH2 (the binding partner of HA4) replaces the periplasmic sensor domain of E. coli transmembrane transcriptional activator CadC and is connected to the DNA binding domain (DBD) via an artificial transmembrane domain composed of 16 Leucine repeat residues, which has previously been demonstrated to support the expression of correctly oriented chimeric CadC proteins into the E. coli inner membrane (Lindner and White, 2014; Chang et al., 2018; Morrison et al., 2021). On this same plasmid, sfGFP is placed under the control of the CadBA promoter. C) (i) The aggregation screen will be carried out in the same way as TPBLA, by screening for ampicillin resistance. (ii) Correct folding of the test protein enables the two domains of  $\beta$ -lactamase to come together and form the active enzyme, with the ability to hydrolyse  $\beta$ lactam antibiotics. (iii) Misfolding, aggregation or instability of the test protein blocks association of  $\beta$ -lactamase, inhibiting formation of the catalytic site. D) (i) Upon addition of caffeine, the TPBLA construct will dimerise via its C-terminal VHH domain and can bind to the SH2 domain, which has replaced the periplasmic sensor domain of CadC. (ii) This forces dimerisation of the chimeric CadC construct, (iii) thereby enabling it to bind to the two operator motifs of the CadBA promoter and induce expression of sfGFP. Carrying out this screening assay alongside or following ampicillin selection could result in identification of proteins with improved biophysical properties that maintain binding to their targets.

caffeine (causing the dimerisation of VHH and therefore of CadC, enabling it to bind to CadBA) (Chang et al., 2018) (Figure 5.5).

To assess whether this system could be used in conjunction with TPBLA, the dimerising nanobody VHH was fused to the C-terminus of the TPBLA construct, which was itself expressing the HA4 monobody as the test protein, enabling caffeine-inducible dimerisation of TPBLA (Figure 5.5A). The periplasmic sensor domain of CadC was replaced with the SH2 domain (Figure 5.5B). In the presence of caffeine, binding of the HA4 within dimeric TPBLA to the SH2 domain fused to CadC should induce dimerisation of CadC, and enabling binding to the CadBA promoter and inducing expression of sfGFP (Figure 5.5D). To assess the potential of this system, HA4<sup>WT</sup> and HA4<sup>Y87A</sup> were used to see if SnAC 2.0 could distinguish between a strong and weak binder (blaHA4<sup>WT</sup>-VHH or blaHA4<sup>Y87A</sup>-VHH, respectively). Cells would be co-transformed with CadC-SH2 and TPBLA-VHH containing either HA4<sup>WT</sup> or HA4<sup>Y87A</sup>, and the fluorescence intensity of sfGFP measured over time. As controls, cells would be transformed with CadC with the periplasmic sensor domain replaced with either SH2 (CadC-SH2, negative control) or VHH (CadC-VHH, positive control) only.

## 5.2.5 Fusion of caffeine-inducible dimerising nanobody (VHH) to TPBLA does not inhibit β-lactamase activity

In order to co-evolve aggregation resistance and binding affinity using our new assay, fusion of the caffeine-inducible dimerising nanobody VHH to TPBLA must not inhibit β-lactamase activity. A traditional TPBLA screen was performed on cells co-transformed with CadC-SH2 and TPBLA-VHH containing a HA4 variant. Therefore, the TPBLA plates contained both tetracycline and kanamycin, as they are the corresponding resistance markers on the TPBLA and CadC-SH2 plasmids, respectively. As well as HA4<sup>WT</sup> and HA4<sup>Y87A</sup>, two destabilised mutants were designed, HA4<sup>2A</sup> and HA4<sup>Y87A 2A</sup>, where two isoleucine to alanine mutations (I39A and I75A) were introduced within HA4 to destabilise the core by creating a cavity (Eriksson et al., 1992), similar to MBP<sup>4A</sup> in Chapter 3. The traditional TPBLA screen was able to distinguish between the stable and destabilised variants of HA4 using their β-lactamase activity (Figure 5.6A,B). The two destabilised variants have a higher AUC than the negative control (blaGG<sub>STOP</sub>-VHH + CadC-SH2), indicating they are not completely unfolded and therefore are still somewhat selective. The fact that all variants are better than the negative control, and that the destabilised variants are better than the stable variants, shows the fusion of VHH to the C-terminus of β-lactamase does not completely inhibit enzymatic activity. Furthermore, the dimerisation

of TPBLA with HA4<sup>WT</sup> via its VHH domain has no impact on  $\beta$ -lactamase activity (Figure 5.6C). Therefore, this would enable screening of a library by its ampicillin resistance to identify the most stable variants.

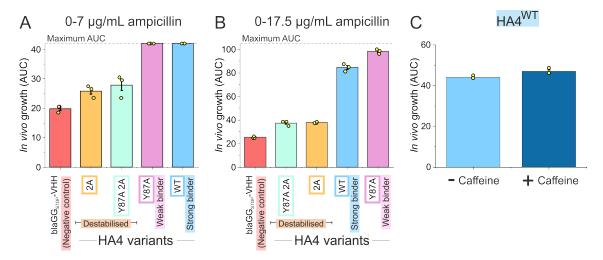


Fig. 5.6 TPBLA screen of cells co-transformed with CadC-SH2 and either blaHA4<sup>WT</sup>-VHH (strong binder), blaHA4<sup>Y87A</sup>-VHH (weak binder), blaHA4<sup>2A</sup>-VHH (destabilised), blaHA4<sup>Y87A</sup> <sup>2A</sup>-VHH (destabilised), or blaGG<sub>STOP</sub>-VHH (negative control). Area under the antibiotic survival curve (AUC) calculated for blaHA4<sup>WT</sup>-VHH, blaHA4<sup>Y87A</sup>-VHH, blaHA4<sup>2A</sup>-VHH, blaHA4<sup>Y87A</sup> <sup>2A</sup>-VHH, or blaGG<sub>STOP</sub>-VHH, screened at (A) 0-7  $\mu$ g/mL or (B) 0-17.5  $\mu$ g/mL ampicillin. (C) Separate experiments of AUC calculated for blaHA4<sup>WT</sup>-VHH at 0-17.5  $\mu$ g/mL ampicillin with and without the addition of 100 mM caffeine. Error bars show standard error of the mean (S.E.M) from three independent experiments.

#### 5.2.6 CadC can be used to measure cognate binding in the periplasm

The expression conditions were optimised by varying the amounts of IPTG (induces CadC expression), caffeine (induces VHH dimerisation), and arabinose (induces TPBLA expression) to find a condition whereby HA4<sup>WT</sup> (strong binder) and HA4<sup>Y87A</sup> (weak binder) can be clearly distinguished. Furthermore, optimal screening condition should display low non-specific activation of CadC-SH2 alone (negative control), yet be able to activate CadC-VHH (positive control). Initially, the IPTG concentration (25, 50, or 100  $\mu$ M) and caffeine concentration (50 or 100  $\mu$ M) was varied, while keeping the arabinose concentration the same as was used in traditional TPBLA experiments (0.075 % (*w/v*)). The fluorescence intensity of sfGFP in cells transformed with CadC-SH2 or CadC-VHH alone, or co-transformed with CadC-SH2 and TPBLA-VHH containing either HA4<sup>WT</sup> (blaHA4<sup>WT</sup>-VHH) or HA4<sup>Y87A</sup> (blaHA4<sup>Y87A</sup>-VHH) was measured over time (Figure 5.7). Looking at the endpoint fluorescence intensity after 1500 minutes and assessing the relative increase in fluorescence intensity compared with uninduced samples, high IPTG concentrations (50 and 100  $\mu$ M) resulted in increased levels of non-specific activation of

CadC-SH2, whereas at 25  $\mu$ M this was less prominent (Figure 5.9). Caffeine concentration has less of an impact, but looking at the negative control CadC-SH2 at 100  $\mu$ M there is less non-specific activation of CadC compared with 50  $\mu$ M. Future experiments consequently used 25  $\mu$ M IPTG and 100  $\mu$ M caffeine, which was consistent with conditions used in previous studies using a CadC-VHH chimera (Chang et al., 2018).

It is known that the Y87A point mutation in HA4 has a significant impact on binding affinity, where it has previously been shown to completely abolish binding using a fluorescence polarization competition assay (Wojcik et al., 2010). However, as it was a competition assay it does not completely preclude the possibility of very weak binding. We hypothesised that the positive fluorescence signal from HA4<sup>Y87A</sup> was because the localised concentration in the periplasm was so high. Therefore, the fluorescence intensity of cells co-transformed with CadC-SH2 and TPBLA-VHH containing either HA4<sup>WT</sup> (blaHA4<sup>WT</sup>-VHH) or HA4<sup>Y87A</sup> (blaHA4<sup>Y87A</sup>-VHH) induced with varying concentrations of arabinose (0 mM, 0.001 - 10 mM in 10-fold steps) was measured over time. At 0.01 mM arabinose, blaHA4<sup>WT</sup>-VHH gives a strong positive signal whereas blaHA4<sup>Y87A</sup>-VHH has the same signal as the 0 mM arabinose (uninduced) sample (Figure 5.8). Concentrations of arabinose above 0.01 mM resulted in a fluorescent signal from blaHA4<sup>Y87A</sup>-VHH (Figure 5.8). The increase in green fluorescence signal from the 0 mM arabinose (uninduced) sample over time is likely due to autofluorescence increasing as the  $OD_{600}$  of the cells increases. Nevertheless, this data shows the CadC strategy can be used to distinguish between strong and weak binding in the E. coli periplasm.

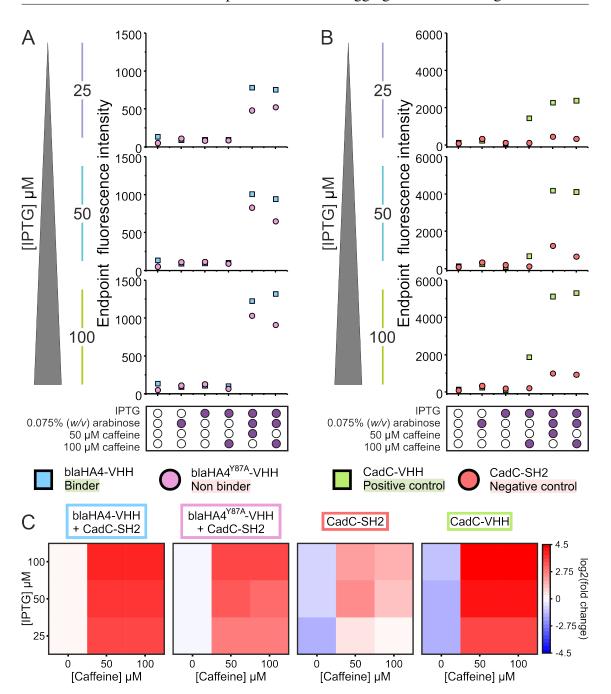


Fig. 5.7 Fluorescence intensity endpoints of cells co-transformed with CadC-SH2 and either blaHA4<sup>WT</sup>-VHH (strong binder) or blaHA4<sup>Y87A</sup>-VHH (weak binder), or transformed with only CadC-SH2 (negative control) or CadC-VHH (positive control) alone. Endpoint sfGFP fluorescence intensity of whole cells grown overnight at 37 °C, 200 rpm, induced at a range of IPTG (25, 50, 100  $\mu$ M) and caffeine (50, 100 $\mu$ M) concentrations, with 0.075 % (*w/v*) arabinose. Cells transformed with (A) CadC-SH2 and either blaHA4<sup>WT</sup>-VHH or blaHA4<sup>Y87A</sup>-VHH, or (B) CadC-VHH or CadC-SH2. C) Heatmaps showing the log<sub>2</sub>(fold change) of fluorescence intensity endpoints compared with the uninduced sample of cells co-transformed with CadC-SH2 and either blaHA4<sup>WT</sup>-VHH (strong binder) or blaHA4<sup>Y87A</sup>-VHH (weak binder), or transformed with only CadC-SH2 (negative control) or CadC-VHH (positive control) alone.

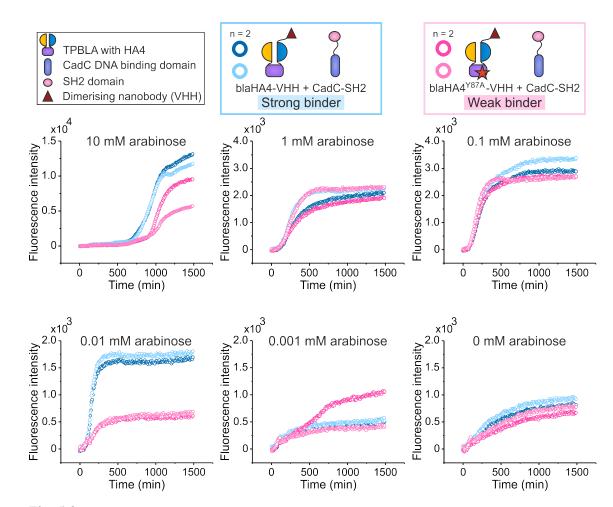


Fig. 5.8 Fluorescence intensity over time of cells co-transformed with CadC-SH2 and either blaHA4<sup>WT</sup>-VHH (strong binder) or blaHA4<sup>Y87A</sup>-VHH (weak binder) at varying arabinose concentrations. sfGFP fluorescence spectra over time of whole cells co-transformed with CadC-SH2 and either blaHA4<sup>WT</sup>-VHH or blaHA4<sup>Y87A</sup>-VHH, grown overnight at 37 °C, 200 rpm and induced at a range of arabinose concentrations (10, 1, 0.1, 0.01, 0.001, 0 mM) with 25  $\mu$ M IPTG and 100 $\mu$ M caffeine. The two shades of blue (HA4<sup>WT</sup>) or pink (HA4<sup>Y87A</sup>) represent two technical repeats (n = 2).

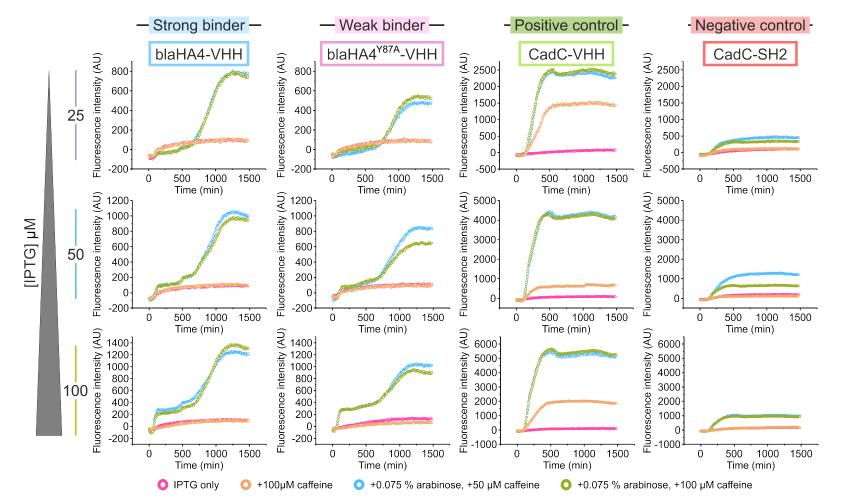


Fig. 5.9 Fluorescence intensity over time of cells co-transformed with CadC-SH2 and either blaHA4<sup>WT</sup>-VHH (strong binder) or blaHA4<sup>Y87A</sup>-VHH (weak binder), or transformed with only CadC-SH2 (negative control) or CadC-VHH (positive control) alone. sfGFP fluorescence intensity over time of whole cells grown overnight at 37 °C, 200 rpm, induced at a range of IPTG (25, 50, 100  $\mu$ M) and caffeine (50, 100 $\mu$ M) concentrations, with 0.075 % (*w/v*) arabinose. At each IPTG concentration, blaHA4<sup>WT</sup>-VHH (strong binder) and blaHA4<sup>Y87A</sup>-VHH (weak binder) are plotted on the same Y-axis. Similarly, at each IPTG concentration CadC-SH2 (negative control) and CadC-VHH (positive control) are plotted on the same Y-axis.

#### 5.2.7 Flow cytometry can be used to identify binders

To screen a library of variants based on their binding affinity and identify the best binders, individual positive cells need to be distinguished from individual negative cells. This can be acheived using flow cytometry, or fluorescence activated cell sorting (FACS). Cells co-transformed with blaMBP and either CadC-SH2 (negative control) or CadC-VHH (positive control) were grown and the sfGFP fluorescence was measured over time (Figure 5.10A). blaMBP was included as a control so that the *E. coli* could be grown in both tetracycline and carbenicillin, to maintain both plasmids expressing blaMBP and CadC, respectively. Furthermore, the addition of blaMBP means that the cellular machinery is split between expressing both fusion proteins, enabling a more fair comparison to those expressing blaHA4 and CadC-SH2. Following an overnight grow, these cells were visualised using flow cytometry. In the CadC-VHH cells, there is a clear emergence of a strong sfGFP positive population (Figure 5.10B, C), showing FACS could be used to sort the positive variants from the negative variants to identify binders.

To identify the optimal arabinose concentration for both distinguishing strong and weak binding, as well as reducing autofluorescence background versus true signal, the fluorescence intensity over time of cells co-transformed with CadC-SH2 and TPBLA-VHH containing either HA4<sup>WT</sup> (blaHA4<sup>WT</sup>-VHH), HA4<sup>Y87A</sup> (blaHA4<sup>Y87A</sup>-VHH), HA4<sup>2A</sup> (blaHA4<sup>2A</sup>-VHH), or HA4<sup>Y87A 2A</sup> (blaHA4<sup>Y87A 2A</sup>-VHH) induced with 0.01 mM - 0.05 mM arabinose in 0.01 mM steps was measured (Figure 5.11A). After an overnight grow, these cells were visualised using flow cytometry to see the proportions of sfGFP positive cells (Figure 5.11B). At 0.01 mM arabinose, there is a clear positive population emerging with around a 50:50 split between positive:negative in the blaHA4<sup>WT</sup>-VHH cells compared with the uninduced cells (Figure 5.11B). At this concentration, the other three variants show no or little positive cells. As the arabinose concentration increases, so does the amount of sfGFP positive cells in the blaHA4<sup>Y87A</sup>-VHH, blaHA4<sup>2A</sup>-VHH, or blaHA4<sup>Y87A 2A</sup>-VHH samples. This could be due to the expression levels increasing the localised concentration of monobody in the periplasm enabling the monobody to bind SH2, whereas at the lower concentrations only the strongest binder (blaHA4<sup>WT</sup>-VHH) is able to bind. Therefore, for further studies the 0.01 mM arabinose concentration was chosen, as at this condition there is a clear difference between blaHA4<sup>WT</sup>-VHH and the other variants.

Originally, the two destabilised variants (HA4<sup>2A</sup> and HA4<sup>Y87A 2A</sup>) were designed to represent a destabilised strong binder (HA4<sup>2A</sup>) and a destabilised weak binder (HA4<sup>Y87A 2A</sup>), as the destabilised strong binder would retain the wild-type binding residue Y87. However, both destabilised variants give the same fluorescent signal as the weak binder HA4<sup>Y87A</sup>. It

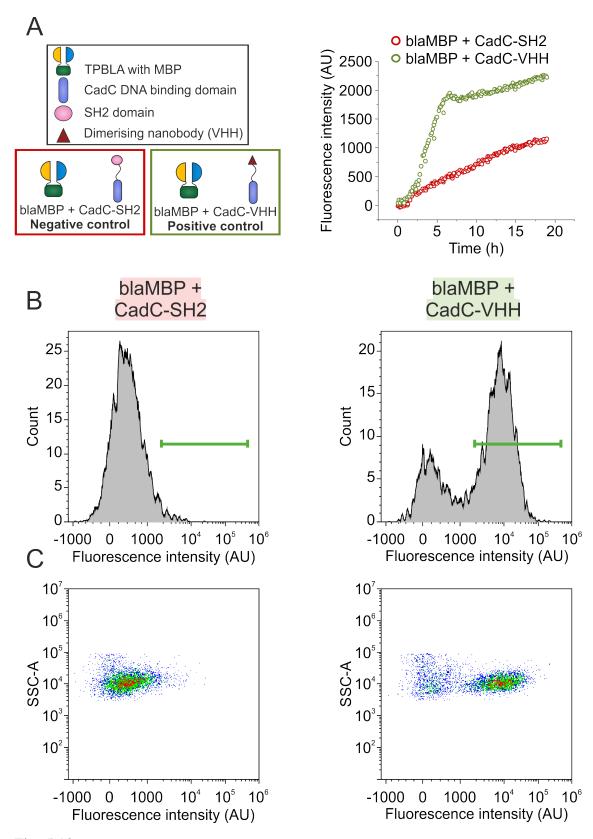


Fig. 5.10 CadC-based sensor can measure binding in the periplasm and be used to sort positive cells using FACS. A) sfGFP fluorescence intensity over time from cells co-transformed with blaMBP and either CadC-SH2 (negative control) or CadC-VHH (positive control). Cells were induced with 0.01 mM arabinose, 25  $\mu$ M IPTG, and 100 $\mu$ M caffeine. B) FACS histogram showing individual cell fluorescence intensities from these cells expressing blaMBP and CadC-SH2 or CadC-VHH following an overnight grow. Green line shows GFP positive cells. C) Side scattering vs fluorescence intensity from these cells following an overnight grow.

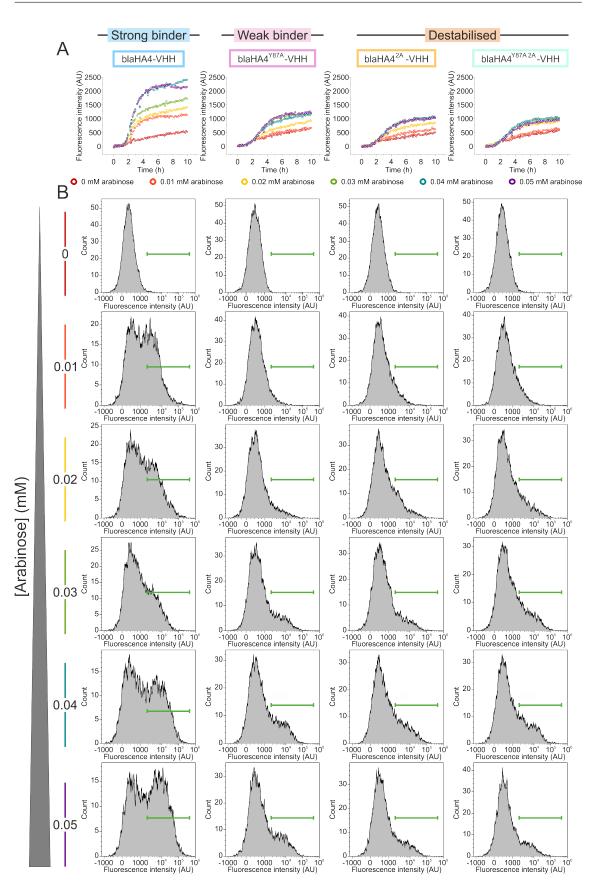
is possible that HA4<sup>2A</sup> does have a higher affinity than HA4<sup>Y87A 2A</sup>, but that the destabilisation reduces the active concentration of protein in the periplasm and therefore results in reduced binding. However, without purifying these proteins and performing binding assays this cannot be known for certain. Nevertheless, both destabilised variants represent useful tools for assessing the potential of SnAC to screen out variants with poor stability and affinity.

# 5.2.8 SnAC can be used to screen a library of variants to identify the most stable and highest affinity variant

The power of the SnAC assay would be to use it to screen a library to identify both high solubility and high affinity variants from a pool. To assess the potential of this assay, we developed a proof of principle experiment. A mock library was created mixing equal amounts of blaHA4<sup>WT</sup>-VHH (strong binder), blaHA4<sup>Y87A</sup>-VHH (weak binder), blaHA4<sup>2A</sup>-VHH (destabilised), and blaHA4<sup>Y87A 2A</sup>-VHH (destabilised). This was co-transformed with CadC-SH2 into SCS1 cells and screened for their resistance to ampicillin (for methods, see Section 2.10.5.1). This removed the two destabilised variants, as confirmed by next-generation sequencing (Figure 5.12). The resulting 'TPBLA screened' library was retransformed into *E. coli* SCS1 cells and sorted using a FACS melody to identify positive fluorescent clones (Section 2.10.4).

From the naive to the TPBLA screened library, the normalised frequency of the 2A point mutations (I39A and I75A) dropped to almost zero, with a log<sub>2</sub>(fold change) reduction of 4.9 and 5.0, respectively (Figure 5.12, Figure 5.13A). The frequency of HA4<sup>WT</sup> dropped as a result of the ampicillin screen (log<sub>2</sub>(fold change) = -2.2), whereas HA4<sup>Y87A</sup> frequency increased (log<sub>2</sub>(fold change) = 0.84). This demonstrates the need for this new assay, as by only using the original ampicillin resistance screen the only variant enriched would be the weak binder HA4<sup>Y87A</sup>.

From the TPBLA screened to the FACS screened library, the frequency of HA4<sup>WT</sup> increased from 0.11 to 0.55 (log<sub>2</sub>(fold change) = 2.4), whereas the frequency of HA4<sup>Y87A</sup> decreased from 0.89 to 0.44 (log<sub>2</sub>(fold change) = -0.99) (Figure 5.12). Following a second round of FACS, the frequency of HA4<sup>WT</sup> increased to 0.83, whereas the frequency of HA4<sup>Y87A</sup> decreased to 0.17. Prior to FACS, the frequencies of HA4<sup>WT</sup> and HA4<sup>Y87A</sup> were 0.11 and 0.89, respectively. Therefore, from this after the second round of FACS the frequency of HA4<sup>WT</sup> had increased by a log<sub>2</sub>(fold change) of 2.9, whereas HA4<sup>Y87A</sup> had decreased by -2.4. This demonstrates how the SnAC assay can successfully screen a



210 Towards simultaneous improvement of both aggregation and binding with TPBLA

Fig. 5.11 SnAC assay can identify positive binders in the periplasm and be used to sort positive cells using FACS. A)sfGFP fluorescence spectra over time of whole cells co-transformed with CadC-SH2 and either blaHA4<sup>WT</sup>-VHH, blaHA4<sup>Y87A</sup>-VHH, blaHA4<sup>2A</sup>-VHH, or blaHA4<sup>Y87A</sup> <sup>2A</sup>-VHH, grown overnight at 37 °C, 200 rpm and induced at a range of arabinose concentrations (0 -0.5 mM) with 25  $\mu$ M IPTG and 100 $\mu$ M caffeine. B) FACS histogram showing fluorescence intensities from these cells following an overnight grow. Green line shows GFP positive cells.

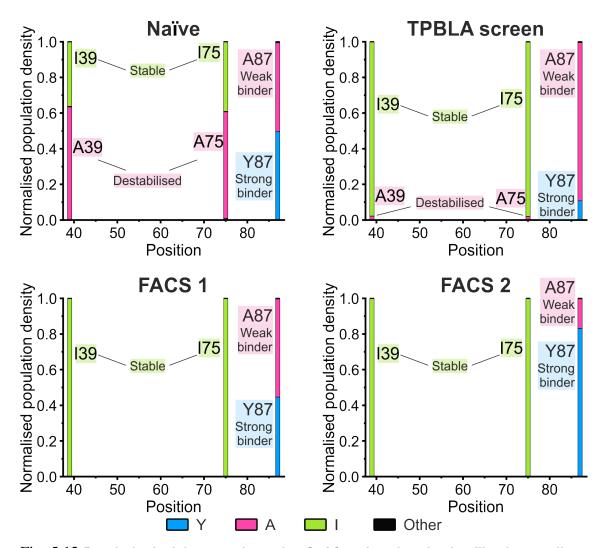
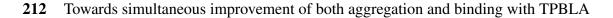


Fig. 5.12 **Proof of principle screening using SnAC and analysed using Illumina amplicon sequencing.** Normalised population density of each residue at each position within a fragment of HA4 for the respective libraries indicated above each panel. The wild type residue at positions other than 39, 75, and 87 were omitted for clarity. This shows the proportion of HA4<sup>WT</sup> (strong binder), HA4<sup>Y87A</sup> (weak binder), and the two destabilising 2A mutations (I39A and I75A) from HA4<sup>2A</sup> and HA4<sup>Y87A 2A</sup> in the unselected library (Naive), post-ampicillin screen (TPBLA screen), first FACS screen (FACS 1), and second FACS screen (FACS 2). The proportion of alanine (pink), tyrosine (blue), and isoleucine (green) residues at each position are highlighted. n = 1.



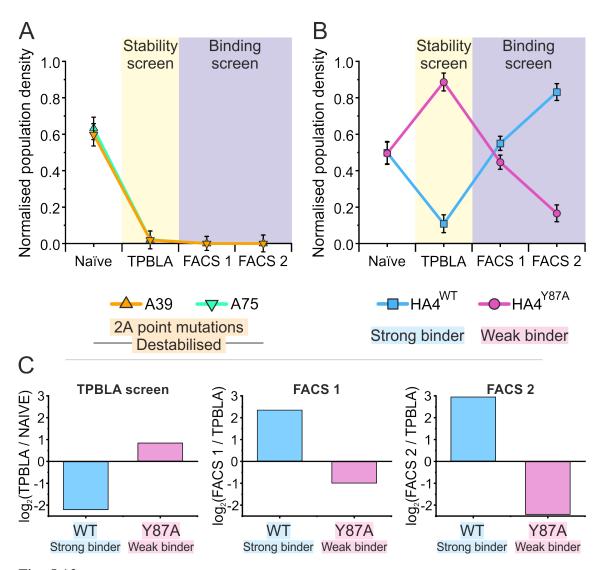


Fig. 5.13 CadC-based sensor can measure binding in the periplasm and be used to sort positive cells using FACS. Normalised frequencies of (A) the 2A point mutations (I39A and I75A), and (B) HA4<sup>WT</sup> and HA4<sup>Y87A</sup>, at each step over the selection experiment. Error bars show two standard deviations from the average normalised frequency at every position except 39, 75, and 87. This assumes mutations detected at any position other than 39, 75, and 87 are sequencing errors. (C) log<sub>2</sub>(fold change) of the frequency calculated using the naive and TPBLA selected libraries, and the TPBLA selected libraries vs FACS 1 and FACS 2, for HA4<sup>WT</sup> and HA4<sup>Y87A</sup>. n = 1.

library of variants based on their stability and binding affinity in order to identify the most stable and highest affinity variant, HA4<sup>WT</sup>.

#### 5.3 Discussion

Having established that TPBLA can be used to screen and evolve protein solubility in Chapter 3, used it as a developability screen on 35 clinically relevant scFv's and to evolve aggregation resistance in Chapter 4, this chapter sought to expand the assay further to introduce a selection for binding affinity. Often affinity matured antibodies have a reduction in stability; one study showed nanobodies affinity-matured using a single round of yeast surface display resulted in 18 °C reduction in melting temperature (Julian et al., 2015; Rabia et al., 2018). This highlights the need for a method that enables co-selection of both stability and aggregation resistance. Phage (Wojcik et al., 2010), yeast surface (Julian et al., 2019) and ribosome (Buchanan et al., 2012) display techinques have all been utilised to evolve affinity in antibody based drugs. These methods can be utilised to evolve thermodynamic stability and aggregation resistance by modifying the protein folding conditions, such as carrying out display experiments at increased temperatures (phage and yeast surface display) (Jespers et al., 2004; Park et al., 2006; Jones et al., 2011; Pavoor et al., 2012) or in the presence of reductants such as DTT (ribosome display) (Buchanan et al., 2012). Phage display has been modified to add in a selection for protein stability by displaying nanobodies on the surface of phage, heating to induce unfolding followed by cooling then screening displayed nanobodies against a coformational ligand specific for folded V<sub>H</sub> domains (protein A), resulting in identification of soluble, aggregation-resistant nanobodies (Jespers et al., 2004). Yeast surface display has been modified in a similar way, by utilising protein A to include a selection for protein stability (Julian et al., 2015).

Periplasmic phage-assisted continuous evolution (pPACE, Section 1.6.2) has been recently developed to co-evolve binding affinity and soluble expression of scFvs in the *E. coli* periplasm (Morrison et al., 2021). While a powerful tool, pPACE has a number of challenges: experiments are complex and require specialist equipment, they require the ability to genetically modify phage, they have a high failure rate whereby if the selection pressure is too high, phage expressing the evolving protein frequently "wash out", where the phage are diluted in host *E. coli* faster than they can propogate, and experiments are difficult to multiplex (DeBenedictis et al., 2022). In contrast, the SnAC assay uses commonly available lab reagents and equipment, can be carried out in small volumes, and the fluorescence-based binding affinity screen can be carried out in a 96-well plate with

the potential for multiplexing. This can potentially be used to carry out multiple biological repeats of evolution experiments simultaneously to enable exploration of a wider sequence space.

SnAC 1.0 utilising split fluorescent proteins was unsuccessful; while all the fusion proteins were successfully secreted to the periplasm (Figure 5.3), the fluorescent reporter signal was not bright enough to be seen - if the method was working at all. The signal from positive control (split mNG2) was already not very bright highlighting the need for a brighter fluorescent reporter protein. mNG2 is currently the brightest split FP available that would work in the periplasm, therefore we needed to turn to cytoplasmic fluorescent reporters and find a way to link periplasmic protein-protein interactions to gene expression in the cytoplasm. In contrast, SnAC 2.0 utilising CadC was successfully able to link proteinprotein interactions in the periplasm to sfGFP expression in the cytoplasm by fusing SH2 to CadC and having blaHA4-VHH construct binding to SH2 as a dimer. The caffeineinducible dimerising nanobody is important as it enables switching the dimerisation on and off, in case this inhibits β-lactamase activity and hinders the TPBLA part of the SnAC method. However, a traditional TPBLA of blaHA4<sup>WT</sup>-VHH with and without 100 mM caffeine showed no significant differences in growth, indicating eventually the system could move towards selection of binding affinity and aggregation resistance simultaneously (Figure 5.6C).

As it stands, SnAC requires optimisation of IPTG, arabinose, and caffeine concentrations for each system. As shown by comparing signals from *E. coli* co-transformed with CadC-SH2 and either blaHA4<sup>WT</sup>-VHH or blaHA4<sup>Y87A</sup>-VHH, if the expression levels are too high, you can get non-specific activation of the CadC. This is likely either due to over-expressing CadC so that single CadC molecules are more likely to find another by chance and dimerise, or by over-expressing blaHA4<sup>Y87A</sup>-VHH thereby increasing the active concentration in the periplasm enabling binding to SH2. However, by reducing the expression levels and reducing the active concentration of TPBLA-VHH, only those with the strongest affinity for the target (HA4<sup>WT</sup>) are able to bind. Future experiments could look at changing the pBAD promoter for a more dialable promoter, or changing the plasmid origin to a higher copy number to widen the dynamic range of expression levels.

HA4 binds to SH2 with high affinity ( $K_d \sim 7 \text{ nM}$ ), demonstrating SnAC can be used to assess and maintain binding affinities at this high level (Wojcik et al., 2010). This is important, as affinities of approved antibodies are generally within the range of picomolar to nanomolar (Brown et al., 2020). It is possible the fluorescence part of SnAC could be used alone to screen for both stability and binding affinity. If a test protein has reduced stability or increased aggregation propensity, this limits the active concentration of protein available to bind to its target. By limiting the expression levels, the only way a test protein would be able to give a detectable fluorescence signal is to either (1) evolve stronger binding affinity, or (2) increase the active concentration of protein by improving its stability and/or aggregation resistance. This is demonstrated by HA4<sup>2A</sup>, which still retains the binding site residue Y87 but does not give a fluorescence signal at low expression levels (0.01 mM arabinose), likely due to reduced stability or increased aggregation resulting in a reduction of the concentration of active protein free to bind to SH2 (Figure 5.11).

FACS can successfully be used to screen and sort negative and positive binders. In cells transformed with CadC-VHH (positive control) and induced with caffeine and IPTG, there is a clear emergence of a positive fluorescent population which is not visible in the CadC-SH2 (negative control) cells (Figure 5.10). When looking at the fluorescent signal over time, the CadC-SH2 cells show a low uptick in green fluorescence. It is known that *E. coli* autofluoresce green due in part to increased expression of flavins, which are involved in energy production and reactive oxygen species detoxification, indicating a response to cellular stresses (Mihalcescu et al., 2015; Surre et al., 2018). This could result in individual cells being falsely sorted as positive, if their autofluorescence is bright enough. Changing the reporter protein from sfGFP for a different fluorescent protein, such as the red fluorescent mScarlet-I, could alleviate this.

To demonstrate the potential of the SnAC method for evolving both stability and binding affinity, four variants of HA4 with a mix of stabilities and binding affinities were identified. These were initially screened based on their antibiotic resistance (correlated to stability and aggregation-resistance), which removed the two destabilised variants with a  $\log_2(\text{fold change})$  frequency reduction of ~5. This screen reduced the frequency of the high-affinity binder HA4<sup>WT</sup> ( $\log_2(\text{fold change}) = -2.2$ ), whereas the frequency of HA4<sup>Y87A</sup> increased ( $\log_2(\text{fold change}) = 0.84$ ). Tyrosine residues are often involved in antibody CDRs where they are generally thought to be "sticky" and promiscuous binders (Collis et al., 2003; Clark et al., 2006), in contrast to arginine and other charged residues which are thought to mediate more precise electrostatic interactions (Sheinerman et al., 2000). However, it has been shown that tyrosine residues are capable of mediating high affinity and specificity interactions (Birtalan et al., 2008). While these residues are important for binding and mediating protein-protein interactions with an antigen, they can also result in increased non-specific self-association due to this "stickyness" resulting in formation of amorphous aggregates. This is demonstrated by the fact that HA4<sup>Y87A</sup> performs slightly better than HA4<sup>WT</sup> in the TPBLA (Figure 5.6B). Therefore, it makes sense that HA4<sup>Y87A</sup> would outcompete HA4<sup>WT</sup> in the ampicillin resistance screen. Traditionally, this is where the evolution screen would end. HA4<sup>Y87A</sup> would be identified as a stable and aggregationresistant variant, whereas HA4<sup>WT</sup>, HA4<sup>2A</sup>, and HA4<sup>Y87A 2A</sup> would be discarded. This again highlights the importance of developing an assay that enables co-selection of these two desirable properties.

The remaining variants can then be sorted using FACS based on sfGFP expression, and therefore binding affinity. The first FACS screen resulted in an frequency increase of HA4<sup>WT</sup> from 0.11 to 0.55, while the frequency of HA4<sup>Y87A</sup> decreased from 0.89 to 0.44 (Figure 5.12). Often screening using FACS requires multiple rounds of selection, with some methods having up to 8 (Desai et al., 2021; Lou et al., 2021). The second FACS screen increased the frequency of HA4<sup>WT</sup> to 0.83, and reduced the frequency of HA4<sup>Y87A</sup> to 0.17. From the original TPBLA screened library to the final FACS screened library, the log<sub>2</sub>(fold change) of frequencies for HA4<sup>WT</sup> and HA4<sup>Y87A</sup> were 2.9 and -2.4, respectively. This demonstrates the ability of the SnAC assay to selectively enrich highly stable and high affinity variants.

Future work with SnAC could look at changing the pipeline. For example, take a library and first screen for binding affinity to remove all non binders. Then take this library and screen for the most stable. This would reduce the chances of a highly stable non-binder from outcompeting the binders in the antibiotic resistance screen, and therefore reduces the library size to only functionally active variants.

In summary, the work in this chapter presents a novel and powerful assay for screening protein aggregation and binding affinity. SnAC can be used to screen and sort a library of variants to identify the most stable, without compromising on binding affinity, and has the potential to be used as a directed evolution tool to co-evolve both aggregation resistance and binding affinity in biotherapeutics.

## **Chapter 6**

# **Final conclusions**

### 6.1 Overall conclusion of results

Understanding the underlying cause of protein instability and aggregation is of great importance to both the biopharmaceutical industry and to human health and disease. The cost of bringing a single biopharmaceutical drug to market is reported to be upwards of \$2 billion, with some estimates putting it as high as \$4 billion (Farid et al., 2020). As a biopharmaceutical can aggregate at any point throughout its lifetime, from expression and purification through to storage and administration to patients, it is important to have methods to identify aggregation-prone or unstable molecules early in development as well as to engineer them to improve their biophysical properties in order to minimise unnecessary time and expense (Willis et al., 2020; Fukuhara et al., 2021). Additionally, the formation of amyloid deposits has been associated with more than 50 human diseases, including Alzheimer's, Parkinson's and type II diabetes (Guthertz et al., 2022; Chiti and Dobson, 2017). Understanding the mechanisms underlying protein stability and aggregation is therefore of ever-increasing interest in order to develop treatments for such diseases.

Understanding the molecular mechanisms governing protein aggregation in a single protein is no small feat. It is often a result of a combination of competing interactions governing protein solubility, stability, hydrophobicity, and inherent aggregation-propensity (Ebo et al., 2020a). This can be inherent in the molecule due to the primary and tertiary sequence, and/or due to various environmental factors that can cause aggregation, such as hydrodynamic flow which can unfold proteins and expose otherwise buried aggregation-prone regions (Willis et al., 2020). Various methods exist to characterise a proteins biophysical properties, however many of these interrogate only a single property; such as HIC measuring a protein's hydrophobicity, or DSF measuring thermal stability. As aggregation results from this combination of interactions, these methods cannot always easily identify aggregation-prone candidates. Furthermore, they are often laborious as they require variants to be expressed and purified prior to characterisation.

The TriPartite β-Lactamase Assay (TPBLA) was previously developed to assess protein thermodynamic stability (Foit et al., 2009) and aggregation propensity (Ebo et al., 2020a), as well as to identify small molecule inhibitors of amyloid formation (Saunders et al., 2016). It has also been used as a directed evolution screen to increase the thermodynamic stability of Im7 (Foit et al., 2009), to reduce the aggregation propensity of the Fv region of the aggregation-prone IgG WFL (Ebo et al., 2020a), and to understand the aggregation mechanisms of β-2-microglobulin (Guthertz et al., 2022). As this all occurs in the E. coli periplasm, the oxidising environment supports proper formation of disulfide bonds, which are key components of antibody-based therapeutics. Previous work using TPBLA had only compared point mutants of a single scFv to themselves, rather than comparing the scores of different scFvs. Furthermore, the directed evolution methodology was limited by first-generation sequencing making it laborious, low-throughput, and high-cost. The work in Chapter 3 aimed to develop TPBLA into a high-throughput screening technology capable of screening hundreds to thousands of variants in a single experiment based on their stability and aggregation-propensity. This chapter also developed a more efficient methodology to create error-prone PCR libraries, using Golden Gate assembly to clone the error-prone PCR fragment seamlessly into the  $\beta$ -lactamase vector at 100 % efficiency. This enabled the generation of libraries containing  $\sim \times 10^9$  to  $\times 10^{11}$  variants, estimated based on the number of colony forming units. This is a 3- to 5-fold improvement compared with the original megaprimer method which generated libraries containing  $\sim \times 10^6$  variants (Ebo et al., 2020a). The increase in library size increases the probability the library contains optimal variants by enabling exploration of a wider sequence space (Saito et al., 2021). TPBLA was developed into a high-throughput screening assay using these large error-prone libraries to evolve variants of MBP and combining this with the power of next-generation sequencing enabled rapid and robust identification of hotspot regions and beneficial point mutations. Initially, short-read Illumina sequencing was used to identify these hotspot regions and improved variants, whereas long-read Pacbio was used to see whether these point mutants were found alone or in combination with other mutations. Pacbio sequencing of the evolved MBP<sup>4A</sup> library showed it was made up of mostly single point mutations (84% of the top 100 mutants were single), demonstrating Illumina short read sequencing was sufficient for accurately assessing mutational profiles within these libraries. The methodology has the potential for use as a deep mutational scanning screen, where libraries

are synthesised to have every single point mutant at every position. As this library would be synthesised to contain only single point mutants, the Illumina shotgun approach would be sufficient to assess changes in mutation frequency following evolution. Ideally a higher read depth would be used when sequencing both naive and selected libraries. This would enable more robust log<sub>2</sub>(fold change) calculations as the frequency of erroneous reads at each position would be lower. Moving TPBLA from solid to liquid culture could represent a powerful assay for deep mutational scanning experiments. This would enable timepoints to be taken over the course of the evolution experiment to look at mutational frequency changes over time, which could give more insight into the positions and residues involved in governing stability and aggregation. However, previous work attempting to carry out TPBLA in liquid culture instead of solid agar where score was measured as E. coli growth rates over time found a counterintuitive correlation between growth score and developability (Golinski et al., 2021). This could have been due to increased protein production of highly expressed and stable variants resulting in a decrease in E. coli growth rate (Golinski et al., 2021). Furthermore, it has been demonstrated that  $\beta$ -lactamases can be secreted into the extracellular medium, a process shown to be partially dependent on Type I Secretion System component TolC (Rangama et al., 2021). Therefore, carrying out the TPBLA evolution screen in liquid culture has the potential to result in 'protection' of cells expressing inactive  $\beta$ -lactamase (test protein aggregated/unstable) by  $\beta$ -lactamase being secreted into the media. It has been shown that outer membrane vesicles (OMVs) from a  $\beta$ -lactam resistant *E. coli* strain contain  $\beta$ -lactamase and when purified they exhibit  $\beta$ -lactamase activity (Kim et al., 2018). Additionally, supplementing  $\beta$ -lactam susceptible E. coli with OMVs from the  $\beta$ -lactam resistant strain allows them to grow in the presence of ampicillin (Kim et al., 2018). Therefore, converting TPBLA from solid to liquid medium likely requires significant optimisation to overcome these issues.

Chapter 4 applied TPBLA to biopharmaceutically relevant proteins and assessed its potential for use as a devleopability screen during initial drug development. Multiple regression models demonstrated TPBLA is influenced by a multitude of factors, including; thermal stability, solubility, aggregation propensity, self-association, and colloidal stability. Therefore, TPBLA may give a more rounded readout of a molecule compared with developability assays that measure only a single characteristic which could provide invaluable information for identifying poorly developable candidates. The high-throughput directed evolution methodology developed in Chapter 3 was applied to two antibody fragments, however evolution of an already stable variant (AMS197) likely requires a high mutation rate library and to be evolved at a higher selection pressure (Drummond et al., 2005). Evolution of an aggregation-prone candidate (AMS134) identified multiple improved variants, with the highest performing variant in TPBLA completely abolishing

aggregation at high temperatures when measured by SLS. Furthermore, assessing evolved mutations of AMS134 demonstrated the validity of using the log<sub>2</sub>(fold change) value to identify improved variants. Three point mutants were assessed; one below the  $2\sigma$  threshold (AMS134<sup>R93G VL</sup>), one just at the threshold (AMS134<sup>M123K VH</sup>), and one above the threshold (AMS134<sup>V124D VH</sup>). The variant below the  $2\sigma$  threshold (AMS134<sup>M123K VH</sup>) showed an increase in aggregation propensity, the variant at the threshold (AMS134<sup>M123K VH</sup>) showed no significant difference, and the variant above the threshold (AMS134<sup>V124D VH</sup>) showed aggregation was abolished at high temperatures. The libraries used for this evolution were designed to have one mutation per gene on average, which therefore limits the potential for improving as multiple mutations are often required for significant improvements in protein stability or aggregation-resistance (Yu and Dalby, 2018). However, increasing the mutation rate could result in accumulating mutations that could then have a knock-on effect on target affinity, as many hotspots identified in the evolution of AMS134 were within HCDR3. This highlights the importance of including a selection for binding affinity into TPBLA when evolving antibody-based therapeutics.

Chapter 4 assessed 35 late stage clinical mAbs using TPBLA and compared this to their performance in other common developability assays. The TPBLA score showed no significant difference between the approval ratings (Phase 2, Phase 3, or Approved) using those published in 2017 (Jain et al., 2017). Since then, many of those in Phase 2 and 3 clinical trials have been discontinued. Comparing TPBLA score between the discontinued mAbs and those already approved showed a significant difference in TPBLA score, where the discontinued mAbs generally showed lower growth scores, demonstrating how TPBLA could be used to identify candidates likely to fail at clinical trials. Multiple regression models were used to explain TPBLA score using the other common developability assays, highlighting links between TPBLA and assays that measure thermal stability, polyspecificty, colloidal stability, self-interactions, and solubility, demonstrating TPBLA could be useful at ranking candidate mAbs based on these properties. The model is not perfect, indicating TPBLA is probing an unknown mechanism or parameter not measured by any of the other 13 developability assays. Since the discontinued mAbs showed significantly lower TPBLA scores compared with the approved mAbs, whatever TPBLA is measuring is clearly relevant when identifying positive candidates. The TPBLA scores for these 35 mAbs compared with their performance in the other 13 developability assays could be used to inform a machine learning model to predict TPBLA score, and potentially even antibody developability. This could be used to gain a better understanding of to what extent these different properties are influencing TPBLA, and identify the property or properties TPBLA is measuring that the other developability assays are not.

The basic rule of directed evolution is 'you get what you screen for' (You and Arnold, 1996), and evolving to improve a proteins' thermodynamic stability or reduce their aggregation propensity commonly comes at a cost to target affinity (McLure et al., 2022; Rabia et al., 2018). Chapter 5 investigated the potential of introducing a selection for binding into TPBLA to create the Solubility 'n' Affinity Coselection (SnAC) assay. By using E. coli transmembrane transcriptional activator CadC, SnAC was successfully able to link proteinprotein interactions in the periplasm to sfGFP expression in the cytoplasm. SnAC enables the fluorescence intensity of sfGFP to be correlated to the binding of HA4 to SH2 and positive binders to be isolated by FACS. SnAC was successfully used to sort HA4 variants with a mix of stabilities and binding affinities to enrich for the most stable and high-affinity variant, HA4<sup>WT</sup>. E. coli co-transformed with CadC-SH2 and TPBLA-VHH containing a HA4 variant (HA4<sup>WT</sup> (strong binder), HA4<sup>Y87A</sup> (weak binder), HA4<sup>2A</sup> (destabilised), and HA4<sup>Y87A 2A</sup> (destabilised)) were initially screened by their antibiotic resistance, which removed the two destabilised variants. The stability screen also reduced the frequency of strong binder HA4<sup>WT</sup>, while increasing the frequency of weak binder HA4<sup>Y87A</sup>. Tyrosine residues are commonly involved in antibody CDR regions due to their "stickiness" and involvement in promiscuous binding, and it is due to this propensity for mediating protein-protein interactions that tyrosines can result in increased self-interactions or aggregation (Collis et al., 2003; Clark et al., 2006; Ausserwöger et al., 2022). This explains how HA4<sup>Y87A</sup> outcompetes HA4<sup>WT</sup> in the antibiotic resistance screen, and highlights the importance of including a selection for binding affinity into TPBLA. Traditionally, the antibiotic resistance screen is the only step in the directed evolution experiment, which in this case would identify weak binder HA4<sup>Y87A</sup> as a stable and aggregation-resistant variant, whereas HA4<sup>WT</sup>, HA4<sup>2A</sup>, and HA4<sup>Y87A 2A</sup> would be discarded. The FACS screen sorting HA4<sup>WT</sup> and HA4<sup>Y87A</sup> based on their sfGFP fluorescence increased the frequency of HA4<sup>WT</sup> to 0.83, and reduced the frequency of HA4<sup>Y87A</sup> to 0.17, after two rounds of FACS sorting. Looking at the log<sub>2</sub>(fold change) in frequency of these two variants from the original TPBLA screened library to the second FACS screened library, HA4<sup>WT</sup> and HA4<sup>Y87A</sup> were 2.9 and -2.4, respectively. This demonstrates the power of the SnAC assay as a directed evolution screen to selectively sort a library of variants to enrich highly-stable and high affinity variants.

#### 6.2 Future work

The directed evolution methodology developed in this thesis has great potential for use as a deep mutational scanning screen to better understand the aggregation mechanisms of disease-relevant aggregation-prone proteins, such as  $\alpha$ -synuclein (Newberry et al., 2020; Doherty et al., 2020) or amyloid- $\beta$ 42 (A $\beta$ 42) (Seuma et al., 2021), both of which are currently being investigated within the laboratory. The data collected from such experiments could be used to inform machine learning algorithms to better understand the aggregation mechanisms of these disease-relevant proteins.

Chapter 4 demonstrated TPBLA could be used as a developability screen to rank candidates based on their beneficial biophysical properties. The data collected in this chapter on the 35 clinically relevant mAbs alongside biophysical metrics calculated from the primary and tertiary structure could be used to inform machine learning algorithms to predict TPBLA score. This could also be used to better understand what properties are influencing TPBLA and to what extent, to understand what property or properties TPBLA is measuring that the other 13 developability assays are missing, and how this could all be used to predict mAb developability.

The SnAC assay developed in Chapter 5 has wide and exciting potential. Further work could be done to characterise the range of stabilities and affinities attainable with the assay; this could be done by using a well-characterised model system, such as Im7/Im9/Im2, all of which share high sequence identity and bind colicin E9 with different affinities (Meenan et al., 2010; Friel et al., 2009; Foit et al., 2009). Furthermore, there have been a range point mutations particularly of Im7 which have been well-characterised for their effect on thermodynamic stability and binding affinity (Friel et al., 2009; Foit et al., 2009). This or a similar system would be useful to see how accurately SnAC is able to specifically correlate binding affinity with fluorescence intensity, as the HA4/SH2 system used in Chapter 5 only assessed positive or negative binding. For the directed evolution methodology, changing the order of the screens could be beneficial. For example, if the FACS screen was carried out first to screen the library based on their binding affinity. This would ensure only binders would be taken forward and screened based on their stability. This FACS screened library would be taken and screened based on the E. coli antibiotic resistance, reducing the chances of a highly stable non-binder (such as HA4<sup>Y87A</sup>) from outcompeting the slightly less stable binders in the this screen. Additionally, this reduces the library size for the stability screen to only functionally active variants.

223

SnAC could also be used to engineer biopharmaceuticals to bind to otherwise undruggable targets. For example, G protein-coupled receptors (GPCRs) are involved in over 100 human diseases and approximately 34 % of all drugs approved by the Food and Drug Administration (FDA) target members of this family (Hauser et al., 2018; He et al., 2022; Meltzer et al., 2022). However, they are difficult to purify for use in drug discovery, both structure-based or for use in affinity maturation techniques such as phage or yeast display (Magnani et al., 2016). As SnAC occurs in vivo, it could be used as a directed evolution screen to engineer stable and high-affinity binders to these otherwise difficult to target epitopes. Furthermore, extensive research currently goes in to engineering GPCRs to improve their thermostability for use in drug discovery (Magnani et al., 2016). SnAC could be adapted to engineer the epitope rather than the biopharmaceutical by replacing the periplasmic sensor domain of CadC with the dimerising nanobody (VHH), and in the place of the transmembrane domain insert a GPCR to be targetted. Therefore, the only way to switch on sfGFP expression using CadC is for the GPCR to fold stably into the membrane enabling dimerisation by VHH. The GPCR could be evolved to improve its stability by limiting the expression level of CadC-GPCR-VHH, introducing variance into the GPCR, and sorting cells based on their sfGFP expression. This could drive a selection for stability and proper folding to raise the active concentration of GPCR.

### 6.3 Final remarks

The adapted TPBLA for high-throughput directed evolution developed in this thesis has been shown to rapidly evolve protein solubility, improve thermodynamic stability, and to identify mutational hotspots involved in limiting protein developability. The combination with next-generation sequencing enables a more comprehensive analysis of a protein's mutational landscape following evolution, allowing a more in-depth understanding of the individual proteins stability and aggregation mechanisms. Furthermore, the highthroughput TPBLA has the potential for use as a deep mutational scanning screen to better understand the molecular mechanisms underlying protein aggregation. This could be of particular use for understanding disease relevant proteins, such as  $\alpha$ -synuclein or A $\beta$ 42, which has the potential to lead to the development of novel therapeutic strategies and earlier identification of disease. TPBLA has the potential for use as a biopharmaceutical developability screen, as it can identify problematic sequences based on a combination of their stability, solubility, and aggregation, without the need for purified protein. Furthermore, TPBLA has been successfully adapted into SnAC, enabling co-screening of binding affinity alongside stability to ensure resulting molecules retain their function following evolution. With the biopharmaceutical industry beginning to develop novel therapeutic scaffolds (Luo et al., 2022), TPBLA and SnAC can be exploited to assess these for their aggregation and binding behaviours, in order to both predict and evolve the developability and manufacturability of these novel molecules.

Directed evolution is a powerful tool for improving the biophysical properties of proteins for biopharmaceutical and industrial processes, for engineering new functions such as enzymes with new activity or tRNAs that incorporate noncanonical amino acids (DeBenedictis et al., 2022; Chin et al., 2002), as well as DMS experiments to uncover protein fitness landscapes and understand proteins in disease models (Bolognesi et al., 2019; Newberry et al., 2020; Seuma et al., 2021). While there are many potential avenues that remain to be explored, the palette of selection techniques now available to researchers, combined with the advent of low-cost NGS to allow high-throughput identification and analysis of the variants unmasked by these screens, is democratizing access to this incredible tool. The screens developed in this thesis could enable high-throughput directed evolution of stability and affinity, and have wide potential for application to biopharmaceuticals and beyond. It is now up to researchers to unleash this power onto new and exciting targets, but all without forgetting the golden rule of directed evolution: 'you get what you screen for'.

### References

- Abbas, S. A., Sharma, V. K., Patapoff, T. W., and Kalonia, D. S. (2013). Characterization of Antibody– Polyol Interactions by Static Light Scattering: Implications for Physical Stability of Protein Formulations. *International Journal of Pharmaceutics*, 448(2):382– 389.
- Abou-Nader, M. and Benedik, M. J. (2010). Rapid Generation of Random Mutant Libraries. *Bioengineered bugs*, 1(5):337–40.
- Akamatsu, Y., Pakabunto, K., Xu, Z., Zhang, Y., and Tsurushita, N. (2007). Whole IgG surface display on mammalian cells: Application to isolation of neutralizing chicken monoclonal anti-IL-12 antibodies. *Journal of Immunological Methods*, 327(1-2):40–52.
- Al-Hilaly, Y. K., Biasetti, L., Blakeman, B. J. F., Pollack, S. J., Zibaee, S., Abdul-Sada, A., Thorpe, J. R., Xue, W.-F., and Serpell, L. C. (2016). The Involvement of Dityrosine Crosslinking in α-Synuclein Assembly and Deposition in Lewy Bodies in Parkinson's Disease. *Scientific Reports*, 6:39171.
- Alejaldre, L., Pelletier, J. N., and Quaglia, D. (2021). Methods for Enzyme Library Creation: Which One Will You Choose? *BioEssays*, 43(8):2100052.
- Alkan, S. S. (2004). Monoclonal Antibodies: The Story of a Discovery That Revolutionized Science and Medicine. *Nature Reviews Immunology*, 4(2):153.
- Allen, J. M., Simcha, D. M., Ericson, N. G., Alexander, D. L., Marquette, J. T., Van Biber, B. P., Troll, C. J., Karchin, R., Bielas, J. H., Loeb, L. A., and Camps, M. (2011). Roles of DNA Polymerase I in Leading and Lagging-Strand Replication Defined by a High-Resolution Mutation Footprint of ColE1 Plasmid Replication. *Nucleic Acids Research*, 39(16):7020–7033.
- Álvarez, B., Mencía, M., de Lorenzo, V., and Fernández, L. Á. (2020). In Vivo Diversification of Target Genomic Sites Using Processive Base Deaminase Fusions Blocked by dCas9. *Nature Communications*, 11(1):6436.
- Amin, S., Barnett, G. V., Pathak, J. A., Roberts, C. J., and Sarangapani, P. S. (2014). Protein Aggregation, Particle Formation, Characterization & Rheology. *Current Opinion in Colloid and Interface Science*, 19(5):438–449.
- Anfinsen, C. B. (1973). Principles That Govern the Folding of Protein Chains. *Science*, 181(4096):223–30.

- Anfinsen, C. B., Haber, E., Sela, M., and White, F. H. (1961). The Kinetics of Formation of Native Ribonuclease during Oxidation of the Reduced Polypeptide Chain. *Proceedings* of the National Academy of Sciences of the United States of America, 47(9):1309–1314.
- Arzumanyan, G. A., Gabriel, K. N., Ravikumar, A., Javanpour, A. A., and Liu, C. C. (2018). Mutually Orthogonal DNA Replication Systems in Vivo. ACS Synthetic Biology, 7(7):1722–1729.
- Ausserwöger, H., Schneider, M. M., Herling, T. W., Arosio, P., Invernizzi, G., Knowles, T. P. J., and Lorenzen, N. (2022). Non-specificity as the sticky problem in therapeutic antibody development. *Nature Reviews Chemistry*, 6(12):844–861.
- Azzi, A. (1974). The Use of Fluorescent Probes for the Study of Membranes. *Methods in Enzymology*, 32(1970):234–46.
- Badran, A. H. and Liu, D. R. (2015). Development of Potent in Vivo Mutagenesis Plasmids with Broad Mutational Spectra. *Nature Communications*, 6:8425.
- Bailly, M., Mieczkowski, C., Juan, V., Metwally, E., Tomazela, D., Baker, J., Uchida, M., Kofman, E., Raoufi, F., Motlagh, S., Yu, Y., Park, J., Raghava, S., Welsh, J., Rauscher, M., Raghunathan, G., Hsieh, M., Chen, Y.-L., Nguyen, H. T., Nguyen, N., Cipriano, D., and Fayadat-Dilman, L. (2020). Predicting Antibody Developability Profiles through Early Stage Discovery Screening. *mAbs*, 12(1):1743053.
- Baird, G. S., Zacharias, D. A., and Tsien, R. Y. (1999). Circular Permutation and Receptor Insertion within Green Fluorescent Proteins. *Proceedings of the National Academy of Sciences*, 96(20):11241–11246.
- Balleza, E., Kim, J. M., and Cluzel, P. (2018). Systematic Characterization of Maturation Time of Fluorescent Proteins in Living Cells. *Nature Methods*, 15(1):47–51.
- Bates, A. and Power, C. A. (2019). David vs. Goliath: The Structure, Function, and Clinical Prospects of Antibody Fragments. *Antibodies*, 8(2):28.
- Bee, C., Abdiche, Y. N., Pons, J., and Rajpal, A. (2013). Determining the Binding Affinity of Therapeutic Monoclonal Antibodies towards Their Native Unpurified Antigens in Human Serum. *PLOS ONE*, 8(11):e80501.
- Benaissa, H., Ounoughi, K., Aujard, I., Fischer, E., Goïame, R., Nguyen, J., Tebo, A. G., Li, C., Le Saux, T., Bertolin, G., Tramier, M., Danglot, L., Pietrancosta, N., Morin, X., Jullien, L., and Gautier, A. (2021). Engineering of a Fluorescent Chemogenetic Reporter with Tunable Color for Advanced Live-Cell Imaging. *Nature Communications*, 12(1):6989.
- Bencurova, E., Pulzova, L., Flachbartova, Z., and Bhide, M. (2015). A rapid and simple pipeline for synthesis of mRNA-ribosome-V(H)H complexes used in single-domain antibody ribosome display. *Molecular bioSystems*, 11(6):1515–1524.
- Berkowitz, S. A. (2006). Role of Analytical Ultracentrifugation in Assessing the Aggregation of Protein Biopharmaceuticals. *The AAPS journal*, 8(3):E590–E605.
- Berkowitz, S. A., Engen, J. R., Mazzeo, J. R., and Jones, G. B. (2013). Analytical Tools for Characterizing Biopharmaceuticals and the Implications for Biosimilars. *Nature Reviews Drug Discovery*, 11(7):527–540.

- Bhirde, A. A., Chiang, M.-J. J., Venna, R., Beaucage, S., and Brorson, K. (2018). High-Throughput in-Use and Stress Size Stability Screening of Protein Therapeutics Using Algorithm-Driven Dynamic Light Scattering. *Journal of Pharmaceutical Sciences*, 107(8):2055–2062.
- Birtalan, S., Zhang, Y., Fellouse, F. A., Shao, L., Schaefer, G., and Sidhu, S. S. (2008). The intrinsic contributions of tyrosine, serine, glycine and arginine to the affinity and specificity of antibodies. *Journal of Molecular Biology*, 377(5):1518–1528.
- Blum, T. R., Liu, H., Packer, M. S., Xiong, X., Lee, P.-G., Zhang, S., Richter, M., Minasov, G., Satchell, K. J. F., Dong, M., and Liu, D. R. (2021). Phage-Assisted Evolution of Botulinum Neurotoxin Proteases with Reprogrammed Specificity. *Science*, 371(6531):803–810.
- Bodelón, G., Palomino, C., and Fernández, L. Á. (2013). Immunoglobulin Domains in Escherichia Coli and Other Enterobacteria: From Pathogenesis to Applications in Antibody Technologies. *FEMS Microbiology Reviews*, 37(2):204–250.
- Boder, E. T. and Wittrup, K. D. (1997). Yeast Surface Display for Screening Combinatorial Polypeptide Libraries. *Nature Biotechnology*, 15(6):553–557.
- Bolognesi, B., Faure, A. J., Seuma, M., Schmiedel, J. M., Tartaglia, G. G., and Lehner, B. (2019). The Mutational Landscape of a Prion-like Domain. *Nature Communications*, 10(1):4162.
- Bozcal, E. and Dagdeviren, M. (2017). Toxicity of  $\beta$ -Lactam antibiotics: Pathophysiology, molecular biology and possible recovery strategies. In Malangu, N., editor, *Poisoning*, chapter 5. IntechOpen, Rijeka.
- Brown, M. E., Bedinger, D., Lilov, A., Rathanaswami, P., Vásquez, M., Durand, S., Wallace-Moyer, I., Zhong, L., Nett, J. H., Burnina, I., Caffry, I., Lynaugh, H., Sinclair, M., Sun, T., Bukowski, J., Xu, Y., and Abdiche, Y. N. (2020). Assessing the binding properties of the anti-PD-1 antibody landscape using label-free biosensors. *PloS One*, 15(3):e0229206.
- Buchanan, A., Ferraro, F., Rust, S., Sridharan, S., Franks, R., Dean, G., McCourt, M., Jermutus, L., and Minter, R. (2012). Improved Drug-like Properties of Therapeutic Proteins by Directed Evolution. *Protein Engineering, Design and Selection*, 25(10):631– 638.
- Buck, P. M., Kumar, S., Wang, X., Agrawal, N. J., Trout, B. L., Singh, S. K., L Trout, B., and Singh, S. K. (2012). *Computational Methods to Predict Therapeutic Protein Aggregation*, volume 899. Humana Press, Totowa, NJ.
- Bushnell, B., Rood, J., and Singer, E. (2017). BBMerge Accurate paired shotgun read merging via overlap. *PloS One*, 12(10):e0185056.
- Cabantous, S. and Waldo, G. S. (2006). In vivo and in vitro protein solubility assays using split GFP. *Nature Methods*, 3(10):845–854.
- Cai, H. H. (2018). Therapeutic Monoclonal Antibodies Approved by FDA in 2017. *MOJ Immunology*, 6(3):82–84.

- Camps, M., Naukkarinen, J., Johnson, B. P., and Loeb, L. A. (2003). Targeted Gene Evolution in Escherichia Coli Using a Highly Error-Prone DNA Polymerase I. *Proceedings of the National Academy of Sciences*, 100(17):9727–9732.
- Carpenter, J. F., Randolph, T. W., Jiskoot, W., Crommelin, D. J. A., Middaugh, C. R., and Winter, G. (2010). Potential Inaccurate Quantitation and Sizing of Protein Aggregates by Size Exclusion Chromatography: Essential Need to Use Orthogonal Methods to Assure the Quality of Therapeutic Protein Products. *Journal of Pharmaceutical Sciences*, 99(5):2200–2208.
- Cavrois, M., de Noronha, C., and Greene, W. C. (2002). A sensitive and specific enzymebased assay detecting HIV-1 virion fusion in primary T lymphocytes. *Nature Biotechnology*, 20(11):1151–1154.
- Centre for Process Innovation (2022). BioStreamline | CPI. https://www.uk-cpi.com/case-studies/biostreamline.
- Chang, H.-J., Mayonove, P., Zavala, A., De Visch, A., Minard, P., Cohen-Gonsaud, M., and Bonnet, J. (2018). A Modular Receptor Platform to Expand the Sensing Repertoire of Bacteria. ACS Synthetic Biology, 7(1):166–175.
- Chapman, B. and Chang, J. (2000). Biopython: Python Tools for Computational Biology. *ACM SIGBIO Newsletter*, 20(2):15–19.
- Chaturvedi, S. K., Siddiqi, M. K., Alam, P., and Khan, R. H. (2016). Protein Misfolding and Aggregation: Mechanism, Factors and Detection. *Process Biochemistry*, 51(9):1183– 1192.
- Chen, K. and Arnold, F. H. (1991). Enzyme Engineering for Nonaqueous Solvents: Random Mutagenesis to Enhance Activity of Subtilisin E in Polar Organic Media. *Bio/Technology*, 9(11):1073–1077.
- Chen, K. and Arnold, F. H. (1993). Tuning the Activity of an Enzyme for Unusual Environments: Sequential Random Mutagenesis of Subtilisin E for Catalysis in Dimethylformamide. *Proceedings of the National Academy of Sciences of the United States of America*, 90(12):5618–5622.
- Chen, V. B., Arendall, W. B., Headd, J. J., Keedy, D. A., Immormino, R. M., Kapral, G. J., Murray, L. W., Richardson, J. S., and Richardson, D. C. (2010). MolProbity: All-Atom Structure Validation for Macromolecular Crystallography. *Acta Crystallographica*. *Section D, Biological Crystallography*, 66(Pt 1):12–21.
- Chen, Y., Wei, G., Zhao, J., Nussinov, R., and Ma, B. (2020). Computational Investigation of Gantenerumab and Crenezumab Recognition of A $\beta$  Fibrils in Alzheimer's Disease Brain Tissue. *ACS Chemical Neuroscience*, 11(20):3233–3244.
- Cherf, G. M. and Cochran, J. R. (2015). Applications of Yeast Surface Display for Protein Engineering. *Methods in Molecular Biology*, 1319:155–175.
- Chin, J. W., Martin, A. B., King, D. S., Wang, L., and Schultz, P. G. (2002). Addition of a Photocrosslinking Amino Acid to the Genetic Code of Escherichia Coli. *Proceedings of* the National Academy of Sciences of the United States of America, 99(17):11020–11024.

- Chiti, F. and Dobson, C. M. (2017). Protein Misfolding, Amyloid Formation, and Human Disease: A Summary of Progress Over the Last Decade. *Annual Review of Biochemistry*, 86:27–68.
- Cho, T. J. and Hackley, V. A. (2010). Fractionation and Characterization of Gold Nanoparticles in Aqueous Solution: Asymmetric-Flow Field Flow Fractionation with MALS, DLS, and UV-vis Detection. *Analytical and bioanalytical chemistry*, 398(5):2003–18.
- Christiansen, A., Kringelum, J. V., Hansen, C. S., Bøgh, K. L., Sullivan, E., Patel, J., Rigby, N. M., Eiwegger, T., Szépfalusi, Z., Masi, F. D., Nielsen, M., Lund, O., and Dufva, M. (2015). High-Throughput Sequencing Enhanced Phage Display Enables the Identification of Patient-Specific Epitope Motifs in Serum. *Scientific Reports*, 5(1):12913.
- Chudasama, V., Maruani, A., and Caddick, S. (2016). Recent Advances in the Construction of Antibody-Drug Conjugates. *Nature Chemistry*, 8(2):114–119.
- Chun, S., Strobel, S., Bassford, P., and Randall, L. (1993). Folding of maltose-binding protein. Evidence for the identity of the rate-determining step in vivo and in vitro. *The Journal of Biological Chemistry*, 268(28):20855–20862.
- Ciaccio, N. A. and Laurence, J. S. (2009). Effects of Disulfide Bond Formation and Protein Helicity on the Aggregation of Activating Transcription Factor 5 (ATF5). *Molecular pharmaceutics*, 6(4):1205–1215.
- Ciesielski, G. L., Hytönen, V. P., and Kaguni, L. S. (2016). Biolayer Interferometry: A Novel Method to Elucidate Protein–Protein and Protein–DNA Interactions in the Mitochondrial DNA Replisome. In *Methods in Molecular Biology (Clifton, N.J.)*, volume 1351, pages 223–231.
- Clackson, T., Hoogenboom, H. R., Griffiths, A. D., and Winter, G. (1991). Making Antibody Libraries Using Phage Display Libraries. *Nature*, 352(6336):624–628.
- Clark, L. A., Ganesan, S., Papp, S., and van Vlijmen, H. W. T. (2006). Trends in antibody sequence changes during the somatic hypermutation process. *Journal of Immunology* (*Baltimore, Md.: 1950*), 177(1):333–340.
- Cock, P. J. A., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., and de Hoon, M. J. L. (2009). Biopython: Freely Available Python Tools for Computational Molecular Biology and Bioinformatics. *Bioinformatics*, 25(11):1422–1423.
- Cole, J. L., Lary, J. W., P Moody, T., Laue, T. M., P. Moody, T., Laue, T. M., P Moody, T., and Laue, T. M. (2008). Analytical Ultracentrifugation: Sedimentation Velocity and Sedimentation Equilibrium. *Methods in Cell Biology*, 84:143–179.
- Collis, A. V. J., Brouwer, A. P., and Martin, A. C. R. (2003). Analysis of the antigen combining site: Correlations between length and sequence composition of the hypervariable loops and the nature of the antigen. *Journal of Molecular Biology*, 325(2):337–354.
- Conchillo-Solé, O., de Groot, N. S., Avilés, F. X., Vendrell, J., Daura, X., and Ventura, S. (2007). AGGRESCAN: A Server for the Prediction and Evaluation of "Hot Spots" of Aggregation in Polypeptides. *BMC Bioinformatics*, 8:65.

- Cornwell, O., Radford, S. E., Ashcroft, A. E., and Ault, J. R. (2018). Comparing Hydrogen Deuterium Exchange and Fast Photochemical Oxidation of Proteins: A Structural Characterisation of Wild-Type and  $\Delta N6 \beta(2)$ -Microglobulin. *Journal of the American Society for Mass Spectrometry*, 29(12):2413–2426.
- Corpet, F. (1988). Multiple Sequence Alignment with Hierarchical Clustering. *Nucleic Acids Research*, 16(22):10881–10890.
- Costa, S., Almeida, A., Castro, A., and Domingues, L. (2014). Fusion tags for protein solubility, purification and immunogenicity in Escherichia coli: The novel Fh8 system. *Frontiers in Microbiology*, 5:63.
- Courtois, F., Agrawal, N. J., Lauer, T. M., and Trout, B. L. (2016). Rational Design of Therapeutic mAbs against Aggregation through Protein Engineering and Incorporation of Glycosylation Motifs Applied to Bevacizumab. *mAbs*, 8(1):99–112.
- Cromwell, M. E. M., Hilario, E., and Jacobson, F. (2006). Protein Aggregation and Bioprocessing. *The AAPS Journal*, 8(3):E572–E579.
- Crook, N., Abatemarco, J., Sun, J., Wagner, J. M., Schmitz, A., and Alper, H. S. (2016). In Vivo Continuous Evolution of Genes and Pathways in Yeast. *Nature Communications*, 7(1):13051.
- Currin, A., Parker, S., Robinson, C. J., Takano, E., Scrutton, N. S., and Breitling, R. (2021). The Evolving Art of Creating Genetic Diversity: From Directed Evolution to Synthetic Biology. *Biotechnology Advances*, 50:107762.
- Dammeyer, T. and Tinnefeld, P. (2012). Engineered Fluorescent Proteins Illuminate the Bacterial Periplasm. *Computational and Structural Biotechnology Journal*, 3(4):e201210013.
- D'Angelo, S., Velappan, N., Mignone, F., Santoro, C., Sblattero, D., Kiss, C., and Bradbury, A. R. (2011). Filtering "genic" open reading frames from genomic DNA samples for advanced annotation. *BMC Genomics*, 12(1):S5.
- Datta, N. and Kontomichalou, P. (1965). Penicillinase synthesis controlled by infectious R factors in Enterobacteriaceae. *Nature*, 208(5007):239–241.
- Daugherty, P. S., Chen, G., Iverson, B. L., and Georgiou, G. (2000). Quantitative Analysis of the Effect of the Mutation Frequency on the Affinity Maturation of Single Chain Fv Antibodies. *Proceedings of the National Academy of Sciences of the United States of America*, 97(5):2029–34.
- De Bernardez Clark, E., Schwarz, E., and Rudolph, R. B. T. M. i. E. (1999). Inhibition of Aggregation Side Reactions during in Vitro Protein Folding. In *Amyloid, Prions, and Other Protein Aggregates*, volume 309, pages 217–236. Academic Press.
- De Groot, A. S. and Scott, D. W. (2007). Immunogenicity of Protein Therapeutics. *Trends in Immunology*, 28(11):482–490.
- Deatherage, D. E. and Barrick, J. E. (2014). Identification of Mutations in Laboratory-Evolved Microbes from next-Generation Sequencing Data Using Breseq. *Methods in molecular biology (Clifton, N.J.)*, 1151:165–88.

- DeBenedictis, E. A., Carver, G. D., Chung, C. Z., Söll, D., and Badran, A. H. (2021). Multiplex Suppression of Four Quadruplet Codons via tRNA Directed Evolution. *Nature Communications*, 12(1):5706.
- DeBenedictis, E. A., Chory, E. J., Gretton, D. W., Wang, B., Golas, S., and Esvelt, K. M. (2022). Systematic Molecular Evolution Enables Robust Biomolecule Discovery. *Nature Methods*, 19(1):55–64.
- Den Engelsman, J., Garidel, P., Smulders, R., Koll, H., Smith, B., Bassarab, S., Seidl, A., Hainzl, O., and Jiskoot, W. (2011). Strategies for the Assessment of Protein Aggregates in Pharmaceutical Biotech Product Development. *Pharmaceutical Research*, 28(4):920– 933.
- Desai, A. A., Smith, M. D., Zhang, Y., Makowski, E. K., Gerson, J. E., Ionescu, E., Starr, C. G., Zupancic, J. M., Moore, S. J., Sutter, A. B., Ivanova, M. I., Murphy, G. G., Paulson, H. L., and Tessier, P. M. (2021). Rational Affinity Maturation of Anti-Amyloid Antibodies with High Conformational and Sequence Specificity. *The Journal* of Biological Chemistry, 296:100508.
- Dill, K. A. and Chan, H. S. (1997). From Levinthal to Pathways to Funnels. *Nature Structural and Molecular Biology*, 4(1):10–19.
- Doherty, C. P., Ulamec, S. M., Maya-Martinez, R., Good, S. C., Makepeace, J., Khan, G. N., van Oosten-Hawle, P., Radford, S. E., and Brockwell, D. J. (2020). A Short Motif in the N-terminal Region of  $\alpha$ -Synuclein Is Critical for Both Aggregation and Function. *Nature Structural and Molecular Biology*, 27(3):249–259.
- Domnowski, M., Maruno, T., Enomoto, K., Kummer, F., Kulakova, A., Harris, P., Uchiyama, S., Jaehrling, J., and Friess, W. (2021). A Multi-Method Approach to Assess the Self-Interaction Behavior of Infliximab. *Journal of Pharmaceutical Sciences*, 110(5):1979–1988.
- Dou, J., Vorobieva, A. A., Sheffler, W., Doyle, L. A., Park, H., Bick, M. J., Mao, B., Foight, G. W., Lee, M. Y., Gagnon, L. A., Carter, L., Sankaran, B., Ovchinnikov, S., Marcos, E., Huang, P.-S., Vaughan, J. C., Stoddard, B. L., and Baker, D. (2018). De Novo Design of a Fluorescence-Activating β-Barrel. *Nature*, 561(7724):485–491.
- Drummond, D. A., Iverson, B. L., Georgiou, G., and Arnold, F. H. (2005). Why high-errorrate random mutagenesis libraries are enriched in functional and improved proteins. *Journal of Molecular Biology*, 350(4):806–816.
- Duan, X., Hall, J. A., Nikaido, H., and Quiocho, F. A. (2001). Crystal structures of the maltodextrin/maltose-binding protein complexed with reduced oligosaccharides: Flexibility of tertiary structure and ligand binding. *Journal of Molecular Biology*, 306(5):1115–1126.
- Duff, M. R., Grubbs, J., and Howell, E. E. (2011). Isothermal Titration Calorimetry for Measuring Macromolecule-Ligand Affinity. *Journal of Visualized Experiments : JoVE*, (55):2796.
- Dyson, M. R., Masters, E., Pazeraitis, D., Perera, R. L., Syrjanen, J. L., Surade, S., Thorsteinson, N., Parthiban, K., Jones, P. C., Sattar, M., Wozniak-Knopp, G., Rueker, F., Leah, R., and McCafferty, J. (2020). Beyond Affinity: Selection of Antibody Variants

with Optimal Biophysical Properties and Reduced Immunogenicity from Mammalian Display Libraries. *mAbs*, 12(1):1829335.

- Ebo, J. S., Guthertz, N., Radford, S. E., and Brockwell, D. J. (2020a). Using Protein Engineering to Understand and Modulate Aggregation. *Current Opinion in Structural Biology*, 60:157–166.
- Ebo, J. S., Saunders, J. C., Devine, P. W. A., Gordon, A. M., Warwick, A. S., Schiffrin, B., Chin, S. E., England, E., Button, J. D., Lloyd, C., Bond, N. J., Ashcroft, A. E., Radford, S. E., Lowe, D. C., and Brockwell, D. J. (2020b). An in Vivo Platform to Select and Evolve Aggregation-Resistant Proteins. *Nature Communications*, 11(1):1816.
- Ecker, D. M., Jones, S. D., and Levine, H. L. (2015). The Therapeutic Monoclonal Antibody Market. *mAbs*, 7(1):9–14.
- Edwards, W. R., Busse, K., Allemann, R. K., and Jones, D. D. (2008). Linking the Functions of Unrelated Proteins Using a Novel Directed Evolution Domain Insertion Method. *Nucleic Acids Research*, 36(13):e78.
- Edwards, W. R., Williams, A. J., Morris, J. L., Baldwin, A. J., Allemann, R. K., and Jones, D. D. (2010). Regulation of Beta-Lactamase Activity by Remote Binding of Heme: Functional Coupling of Unrelated Proteins through Domain Insertion. *Biochemistry*, 49(31):6541–6549.
- Egan, A. J. F., Errington, J., and Vollmer, W. (2020). Regulation of peptidoglycan synthesis and remodelling. *Nature Reviews Microbiology*, 18(8):446–460.
- Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., Peluso, P., Rank, D., Baybayan, P., Bettman, B., Bibillo, A., Bjornson, K., Chaudhuri, B., Christians, F., Cicero, R., Clark, S., Dalal, R., deWinter, A., Dixon, J., Foquet, M., Gaertner, A., Hardenbol, P., Heiner, C., Hester, K., Holden, D., Kearns, G., Kong, X., Kuse, R., Lacroix, Y., Lin, S., Lundquist, P., Ma, C., Marks, P., Maxham, M., Murphy, D., Park, I., Pham, T., Phillips, M., Roy, J., Sebra, R., Shen, G., Sorenson, J., Tomaney, A., Travers, K., Trulson, M., Vieceli, J., Wegener, J., Wu, D., Yang, A., Zaccarin, D., Zhao, P., Zhong, F., Korlach, J., and Turner, S. (2009). Real-Time DNA Sequencing from Single Polymerase Molecules. *Science*, 323(5910):133–138.
- Elgundi, Z., Reslan, M., Cruz, E., Sifniotis, V., and Kayser, V. (2017). The State-of-Play and Future of Antibody Therapeutics. *Advanced Drug Delivery Reviews*, 122:2–19.
- Elledge, S. K., Zhou, X. X., Byrnes, J. R., Martinko, A. J., Lui, I., Pance, K., Lim, S. A., Glasgow, J. E., Glasgow, A. A., Turcios, K., Iyer, N. S., Torres, L., Peluso, M. J., Henrich, T. J., Wang, T. T., Tato, C. M., Leung, K. K., Greenhouse, B., and Wells, J. A. (2021). Engineering Luminescent Biosensors for Point-of-Care SARS-CoV-2 Antibody Detection. *Nature Biotechnology*, 39(8):928–935.
- Emsley, P., Lohkamp, B., Scott, W. G., and Cowtan, K. (2010). Features and Development of Coot. Acta Crystallographica. Section D, Biological Crystallography, 66(Pt 4):486– 501.
- Englander, S. W. and Mayne, L. (2014). The Nature of Protein Folding Pathways. *Proceedings of the National Academy of Sciences of the United States of America*, 111(45):15873–15880.

- Eriksson, A. E., Baase, W. A., Zhang, X. J., Heinz, D. W., Blaber, M., Baldwin, E. P., and Matthews, B. W. (1992). Response of a Protein Structure to Cavity-Creating Mutations and Its Relation to the Hydrophobic Effect. *Science*, 255(5041):178–183.
- Espargaró, A., Sabate, R., and Ventura, S. (2012). Thioflavin-S staining coupled to flow cytometry. A screening tool to detect in vivo protein aggregation. *Molecular bioSystems*, 8(11):2839–2844.
- Estep, P., Caffry, I., Sun, T., Cao, Y., Lynaugh, H., Jain, T., Vásquez, M., Tessier, P. M., and Xu, Y. (2015). An Alternative Assay to Hydrophobic Interaction Chromatography for High-Throughput Characterization of Monoclonal Antibodies. *mAbs*, 7(3):553–561.
- Esvelt, K. M., Carlson, J. C., and Liu, D. R. (2011). A System for the Continuous Directed Evolution of Biomolecules. *Nature*, 472(7344):499–503.
- Evans, P. (2006). Scaling and Assessment of Data Quality. *Acta Crystallographica. Section D*, *Biological Crystallography*, 62(Pt 1):72–82.
- Evans, P. R. and Murshudov, G. N. (2013). How Good Are My Data and What Is the Resolution? Acta Crystallographica Section D: Biological Crystallography, 69(Pt 7):1204–1214.
- Eyes, T. J., Austerberry, J. I., Dearman, R. J., Johannissen, L. O., Kimber, I., Smith, N., Thistlethwaite, A., and Derrick, J. P. (2019). Identification of B Cell Epitopes Enhanced by Protein Unfolding and Aggregation. *Molecular Immunology*, 105:181–189.
- Farid, S. S., Baron, M., Stamatis, C., Nie, W., and Coffman, J. (2020). Benchmarking biopharmaceutical process development and manufacturing cost contributions to R&D. *mAbs*, 12(1):1754999.
- Feng, S., Sekine, S., Pessino, V., Li, H., Leonetti, M. D., and Huang, B. (2017). Improved Split Fluorescent Proteins for Endogenous Protein Labeling. *Nature Communications*, 8(1):370.
- Feng, S., Varshney, A., Coto Villa, D., Modavi, C., Kohler, J., Farah, F., Zhou, S., Ali, N., Müller, J. D., Van Hoven, M. K., and Huang, B. (2019). Bright Split Red Fluorescent Proteins for the Visualization of Endogenous Proteins and Synapses. *Communications biology*, 2(1):344.
- Fernandez-Escamilla, A. M., Rousseau, F., Schymkowitz, J., and Serrano, L. (2004). Prediction of Sequence-Dependent and Mutational Effects on the Aggregation of Peptides and Proteins. *Nature Biotechnology*, 22(10):1302–1306.
- Foit, L., Morgan, G. J., Kern, M. J., Steimer, L. R., von Hacht, A. A., Titchmarsh, J., Warriner, S. L., Radford, S. E., and Bardwell, J. C. A. (2009). Optimizing Protein Stability in Vivo. *Molecular Cell*, 36(5):861–871.
- Fonzé, E., Charlier, P., To'th, Y., Vermeire, M., Raquet, X., Dubus, A., and Frère, J. M. (1995). TEM1 beta-lactamase structure solved by molecular replacement and refined structure of the S235A mutant. *Acta Crystallographica. Section D, Biological Crystallography*, 51(Pt 5):682–694.
- Fowler, D. M. and Fields, S. (2014). Deep Mutational Scanning: A New Style of Protein Science. *Nature Methods*, 11(8):801–807.

- Fox, J. D., Kapust, R. B., and Waugh, D. S. (2001). Single Amino Acid Substitutions on the Surface of Escherichia Coli Maltose-Binding Protein Can Have a Profound Impact on the Solubility of Fusion Proteins. *Protein Science*, 10(3):622–630.
- Fraunhofer, W. and Winter, G. (2004). The Use of Asymmetrical Flow Field-Flow Fractionation in Pharmaceutics and Biopharmaceutics. *European Journal of Pharmaceutics and Biopharmaceutics*, 58(2):369–383.
- Frei, J. and Lai, J. (2016). Protein and Antibody Engineering by Phage Display. *Methods in Enzymology*, 580:45–87.
- Friel, C. T., Smith, D. A., Vendruscolo, M., Gsponer, J., and Radford, S. E. (2009). The mechanism of folding of Im7 reveals competition between functional and kinetic evolutionary constraints. *Nature Structural & Molecular Biology*, 16(3):318–324.
- Frokjaer, S. and Otzen, D. E. (2005). Protein Drug Stability: A Formulation Challenge. *Nature Reviews Drug Discovery*, 4(4):298–306.
- Fukuhara, A., Anzai, Y., Osawa, K., Umeda, M., Minemura, H., Shiramizu, N., Yokoyama, M., and Uchiyama, S. (2021). Plate Reader-Based Analytical Method for the Size Distribution of Submicron-Sized Protein Aggregates Using Three-Dimensional Homodyne Light Detection. *Journal of Pharmaceutical Sciences*, pages S0022–3549(21)00424–X.
- Furniss, R. C. D., Kaderabkova, N., Barker, D., Bernal, P., Maslova, E., Antwi, A. A., McNeil, H. E., Pugh, H. L., Dortet, L., Blair, J. M. A., Larrouy-Maumus, G., McCarthy, R. R., Gonzalez, D., and Mavridou, D. A. I. (2022). Breaking antimicrobial resistance by disrupting extracytoplasmic protein folding. *eLife*, 11:e57974.
- Galán, A., Comor, L., Horvatić, A., Kuleš, J., Guillemin, N., Mrljak, V., and Bhide, M. (2016). Library-Based Display Technologies: Where Do We Stand? *Molecular BioSystems*, 12(8):2342–2358.
- Galarneau, A., Primeau, M., Trudeau, L.-E., and Michnick, S. W. (2002). Beta-Lactamase Protein Fragment Complementation Assays as in Vivo and in Vitro Sensors of Protein Protein Interactions. *Nature Biotechnology*, 20(6):619–622.
- Garcia-Pardo, J., Graña-Montes, R., Fernandez-Mendez, M., Ruyra, A., Roher, N., Aviles, F. X., Lorenzo, J., and Ventura, S. (2014). Amyloid Formation by Human Carboxypeptidase D Transthyretin-like Domain under Physiological Conditions. *Journal of Biological Chemistry*, 289(49):33783–33796.
- Ghosh, I., Hamilton, A. D., and Regan, L. (2000). Antiparallel Leucine Zipper-Directed Protein Reassembly: Application to the Green Fluorescent Protein. *Journal of the American Chemical Society*, 122(23):5658–5659.
- Gibney, E., Van Noorden, R., Ledford, H., Castelvecchi, D., and Warren, M. (2018). 'Test-tube' evolution wins Chemistry Nobel Prize. *Nature*, 562(7726):176.
- Gibson, T. J., Mccarty, K., McFadyen, I. J., Cash, E., Dalmonte, P., Hinds, K. D., Dinerman, A. A., Alvarez, J. C., and Volkin, D. B. (2011). Application of a High-Throughput Screening Procedure with PEG-induced Precipitation to Compare Relative Protein Solubility during Formulation Development with IgG1 Monoclonal Antibodies. *Journal* of Pharmaceutical Sciences, 100(3):1009–1021.

- Gil-Garcia, M., Bañó-Polo, M., Varejão, N., Jamroz, M., Kuriata, A., Díaz-Caballero, M., Lascorz, J., Morel, B., Navarro, S., Reverter, D., Kmiecik, S., Ventura, S., Banó-Polo, M., Varejao, N., Jamroz, M., Kuriata, A., Díaz-Caballero, M., Lascorz, J., Morel, B., Navarro, S., Reverter, D., Kmiecik, S., Ventura, S., Bañó-Polo, M., Varejão, N., Jamroz, M., Kuriata, A., Díaz-Caballero, M., Lascorz, J., Morel, B., Navarro, S., Reverter, D., Kmiecik, S., Ventura, S., Banó-Polo, M., Varejao, N., Jamroz, M., Kuriata, A., Díaz-Caballero, M., Lascorz, J., Morel, B., Navarro, S., Reverter, D., Kmiecik, S., Ventura, S., Banó-Polo, M., Varejao, N., Jamroz, M., Kuriata, A., Díaz-Caballero, M., Lascorz, J., Morel, B., Navarro, S., Reverter, D., Kmiecik, S., and Ventura, S. (2018). Combining Structural Aggregation Propensity and Stability Predictions to Redesign Protein Solubility. *Molecular Pharmaceutics*, 15(9):3846–3859.
- Goldenzweig, A., Goldsmith, M., Hill, S. E., Gertman, O., Laurino, P., Ashani, Y., Dym, O., Unger, T., Albeck, S., Prilusky, J., Lieberman, R. L., Aharoni, A., Silman, I., Sussman, J. L., Tawfik, D. S., and Fleishman, S. J. (2016). Automated Structure- and Sequence-Based Design of Proteins for High Bacterial Expression and Stability. *Molecular Cell*, 63(2):337–346.
- Golinski, A. W., Mischler, K. M., Laxminarayan, S., Neurock, N. L., Fossing, M., Pichman, H., Martiniani, S., and Hackel, B. J. (2021). High-Throughput Developability Assays Enable Library-Scale Identification of Producible Protein Scaffold Variants. *Proceedings of the National Academy of Sciences of the United States of America*, 118(23):e2026658118.
- Greener, A., Callahan, M., and Jerpseth, B. (1997). An Efficient Random Mutagenesis Technique Using an E.Coli Mutator Strain. *Molecular Biotechnology*, 7(2):189–195.
- Guntas, G., Mansell, T. J., Kim, J. R., and Ostermeier, M. (2005). Directed Evolution of Protein Switches and Their Application to the Creation of Ligand-Binding Proteins. *Proceedings of the National Academy of Sciences of the United States of America*, 102(32):11224–11229.
- Guntas, G. and Ostermeier, M. (2004). Creation of an Allosteric Enzyme by Domain Insertion. *Journal of Molecular Biology*, 336(1):263–273.
- Guthertz, N., van der Kant, R., Martinez, R. M., Xu, Y., Trinh, C., Iorga, B. I., Rousseau, F., Schymkowitz, J., Brockwell, D. J., and Radford, S. E. (2022). The effect of mutation on an aggregation-prone protein: An in vivo, in vitro, and in silico analysis. *Proceedings of the National Academy of Sciences of the United States of America*, 119(22):e2200468119.
- Gutzeit, C., Chen, K., and Cerutti, A. (2018). The Enigmatic Function of IgD: Some Answers at Last. *European Journal of Immunology*, 48(7):1101–1113.
- Hailu, T. T., Foit, L., and Bardwell, J. C. A. (2013). In Vivo Detection and Quantification of Chemicals That Enhance Protein Stability. *Analytical Biochemistry*, 434(1):181–186.
- Halperin, S. O., Tou, C. J., Wong, E. B., Modavi, C., Schaffer, D. V., and Dueber, J. E. (2018). CRISPR-guided DNA Polymerases Enable Diversification of All Nucleotides in a Tunable Window. *Nature*, 560(7717):248–252.
- Hamers-Casterman, C., Atarhouch, T., Muyldermans, S., Robinson, G., Hammers, C., Songa, E. B., Bendahman, N., and Hammers, R. (1993). Naturally Occurring Antibodies Devoid of Light Chains. *Nature*, 363:446–448.

- Hanes, J., Pluckthun, A., and Plückthun, A. (1997). In Vitro Selection and Evolution of Functional Proteins by Using Ribosome Display. *Proceedings of the National Academy* of Sciences of the United States of America, 94(10):4937–4942.
- Hao, J. and Berry, A. (2004). A Thermostable Variant of Fructose Bisphosphate Aldolase Constructed by Directed Evolution Also Shows Increased Stability in Organic Solvents. *Protein Engineering, Design and Selection*, 17(9):689–697.
- Hartl, F. U., Bracher, A., and Hayer-Hartl, M. (2011). Molecular Chaperones in Protein Folding and Proteostasis. *Nature*, 475:324–334.
- Hauser, A. S., Chavali, S., Masuho, I., Jahn, L. J., Martemyanov, K. A., Gloriam, D. E., and Babu, M. M. (2018). Pharmacogenomics of GPCR Drug Targets. *Cell*, 172(1-2):41– 54.e19.
- Haverick, M., Mengisen, S., Shameem, M., and Ambrogelly, A. (2014). Separation of mAbs Molecular Variants by Analytical Hydrophobic Interaction Chromatography HPLC: Overview and Applications. *mAbs*, 6(4):852–8.
- He, F., Woods, C. E., Becker, G. W., Narhi, L. O., and Razinkov, V. I. (2011). High-throughput assessment of thermal and colloidal stability parameters for monoclonal antibody formulations. *Journal of Pharmaceutical Sciences*, 100(12):5126–5141.
- He, X.-h., You, C.-z., Jiang, H.-l., Jiang, Y., Xu, H. E., and Cheng, X. (2022). AlphaFold2 versus experimental structures: Evaluation on G protein-coupled receptors. *Acta Pharmacologica Sinica*, pages 1–7.
- He, Y., Lei, J., Pan, X., Huang, X., and Zhao, Y. (2020). The hydrolytic water molecule of Class A  $\beta$ -lactamase relies on the acyl-enzyme intermediate ES\* for proper coordination and catalysis. *Scientific Reports*, 10(1):10205.
- Heater, B. S., Chan, W. S., Lee, M. M., and Chan, M. K. (2019). Directed Evolution of a Genetically Encoded Immobilized Lipase for the Efficient Production of Biodiesel from Waste Cooking Oil. *Biotechnology for Biofuels*, 12(1):1–14.
- Hermeling, S., Crommelin, D. J. A., Schellekens, H., and Jiskoot, W. (2004). Structure-Immunogenicity Relationships of Therapeutic Proteins. *Pharmaceutical Research*, 21(6):897–903.
- Heyduk, E. and Heyduk, T. (2014). Ribosome Display Enhanced by next Generation Sequencing: A Tool to Identify Antibody-Specific Peptide Ligands. *Analytical Biochemistry*, 464:73–82.
- Hingorani, A. D., Kuan, V., Finan, C., Kruger, F. A., Gaulton, A., Chopade, S., Sofat, R., MacAllister, R. J., Overington, J. P., Hemingway, H., Denaxas, S., Prieto, D., and Casas, J. P. (2019). Improving the odds of drug development success through human genomics: Modelling study. *Scientific Reports*, 9(1):18911.
- Ho, J. G. S., Middelberg, A. P. J., Ramage, P., and Kocher, H. P. (2003). The Likelihood of Aggregation during Protein Renaturation Can Be Assessed Using the Second Virial Coefficient. *Protein Science*, 12(4):708–716.

- Ho, M., Nagata, S., and Pastan, I. (2006). Isolation of anti-CD22 Fv with high affinity by Fv display on human cells. *Proceedings of the National Academy of Sciences of the United States of America*, 103(25):9637–9642.
- Ho, M. and Pastan, I. (2009). Mammalian Cell Display for Antibody Engineering. *Methods* in molecular biology (Clifton, N.J.), 525:337–xiv.
- Hong, J. and Gierasch, L. M. (2010). Macromolecular Crowding Remodels the Energy Landscape of a Protein by Favoring a More Compact Unfolded State. *Journal of the American Chemical Society*, 132(30):10445–10452.
- Hötzel, I., Theil, F.-P., Bernstein, L. J., Prabhu, S., Deng, R., Quintana, L., Lutman, J., Sibia, R., Chan, P., Bumbaca, D., Fielder, P., Carter, P. J., and Kelley, R. F. (2012). A strategy for risk mitigation of antibodies with fast clearance. *mAbs*, 4(6):753–760.
- Hu, Y., Liu, C., and Muyldermans, S. (2017). Nanobody-Based Delivery Systems for Diagnosis and Targeted Tumor Therapy. *Frontiers in Immunology*, 8:1442.
- Huang, S., Li, R., Zhang, Z., Li, L., Gu, X., Fan, W., Lucas, W. J., Wang, X., Xie, B., Ni, P., Ren, Y., Zhu, H., Li, J., Lin, K., Jin, W., Fei, Z., Li, G., Staub, J., Kilian, A., van der Vossen, E. A. G., Wu, Y., Guo, J., He, J., Jia, Z., Ren, Y., Tian, G., Lu, Y., Ruan, J., Qian, W., Wang, M., Huang, Q., Li, B., Xuan, Z., Cao, J., Asan, n., Wu, Z., Zhang, J., Cai, Q., Bai, Y., Zhao, B., Han, Y., Li, Y., Li, X., Wang, S., Shi, Q., Liu, S., Cho, W. K., Kim, J.-Y., Xu, Y., Heller-Uszynska, K., Miao, H., Cheng, Z., Zhang, S., Wu, J., Yang, Y., Kang, H., Li, M., Liang, H., Ren, X., Shi, Z., Wen, M., Jian, M., Yang, H., Zhang, G., Yang, Z., Chen, R., Liu, S., Li, J., Ma, L., Liu, H., Zhou, Y., Zhao, J., Fang, X., Li, G., Fang, L., Li, Y., Liu, D., Zheng, H., Zhang, Y., Qin, N., Li, Z., Yang, G., Yang, S., Bolund, L., Kristiansen, K., Zheng, H., Li, S., Zhang, X., Yang, H., Wang, J., Sun, R., Zhang, B., Jiang, S., Wang, J., Du, Y., and Li, S. (2009). The Genome of the Cucumber, Cucumis Sativus L. *Nature Genetics*, 41(12):1275–1281.
- Jacobs, S. A., Wu, S.-J., Feng, Y., Bethea, D., and O'Neil, K. T. (2010). Cross-Interaction Chromatography: A Rapid Method to Identify Highly Soluble Monoclonal Antibody Candidates. *Pharmaceutical Research*, 27(1):65–71.
- Jahn, T. R. and Radford, S. E. (2005). The Yin and Yang of Protein Folding. *FEBS Journal*, 272(23):5962–5970.
- Jain, T., Sun, T., Durand, S., Hall, A., Houston, N. R., Nett, J. H., Sharkey, B., Bobrowicz, B., Caffry, I., Yu, Y., Cao, Y., Lynaugh, H., Brown, M., Baruah, H., Gray, L. T., Krauland, E. M., Xu, Y., Vásquez, M., and Wittrup, K. D. (2017). Biophysical Properties of the Clinical-Stage Antibody Landscape. *Proceedings of the National Academy of Sciences* of the United States of America, 114(5):944–949.
- Javanpour, A. A. and Liu, C. C. (2019). Genetic Compatibility and Extensibility of Orthogonal Replication. *ACS Synthetic Biology*, 8(6):1249–1256.
- Javanpour, A. A. and Liu, C. C. (2021). Evolving Small-Molecule Biosensors with Improved Performance and Reprogrammed Ligand Preference Using OrthoRep. ACS Synthetic Biology, 10(10):2705–2714.
- Jelsch, C., Lenfant, F., Masson, J. M., and Samama, J. P. (1992). Crystallization and preliminary crystallographic data on Escherichia coli TEM1 beta-lactamase. *Journal of Molecular Biology*, 223(1):377–380.

- Jennewein, M. F. and Alter, G. (2017). The Immunoregulatory Roles of Antibody Glycosylation. *Trends in Immunology*, 38(5):358–372.
- Jespers, L., Schon, O., Famm, K., and Winter, G. (2004). Aggregation-Resistant Domain Antibodies Selected on Phage by Heat Denaturation. *Nature Biotechnology*, 22(9):1161– 1165.
- Jiskoot, W., Randolph, T. W., Volkin, D. B., Middaugh, C. R., Schöneich, C., Winter, G., Friess, W., Crommelin, D. J. A., and Carpenter, J. F. (2012). Protein Instability and Immunogenicity: Roadblocks to Clinical Application of Injectable Protein Delivery Systems for Sustained Release. *Journal of Pharmaceutical Sciences*, 101(3):946–954.
- Jo, B. H. (2022). An Intrinsically Disordered Peptide Tag that Confers an Unusual Solubility to Aggregation-Prone Proteins. *Applied and Environmental Microbiology*, 88(7):e0009722.
- Joh, N. H., Grigoryan, G., Wu, Y., and DeGrado, W. F. (2017). Design of self-assembling transmembrane helical bundles to elucidate principles required for membrane protein folding and ion transport. *Philosophical Transactions of the Royal Society of London*. *Series B, Biological Sciences*, 372(1726):20160214.
- Johnson, I. S. (1983). Human Insulin from Recombinant DNA Technology. *Science*, 219(4585):632 LP 637.
- Johnston, C. W., Badran, A. H., and Collins, J. J. (2020). Continuous Bioactivity-Dependent Evolution of an Antibiotic Biosynthetic Pathway. *Nature Communications*, 11(1):4202.
- Jones, D. S., Tsai, P.-C., and Cochran, J. R. (2011). Engineering Hepatocyte Growth Factor Fragments with High Stability and Activity as Met Receptor Agonists and Antagonists. *Proceedings of the National Academy of Sciences*, 108(32):13035–13040.
- Jones, P. T., Dear, P. H., Foote, J., Neuberger, M. S., and Winter, G. (1986). Replacing the Complementarity-Determining Regions in a Human Antibody with Those from a Mouse. *Nature*, 321(6069):522–525.
- Julian, M. C., Lee, C. C., Tiller, K. E., Rabia, L. A., Day, E. K., Schick, A. J., and Tessier, P. M. (2015). Co-Evolution of Affinity and Stability of Grafted Amyloid-Motif Domain Antibodies. *Protein Engineering, Design and Selection*, 28(10):339–350.
- Julian, M. C., Li, L., Garde, S., Wilen, R., and Tessier, P. M. (2017). Efficient Affinity Maturation of Antibody Variable Domains Requires Co-Selection of Compensatory Mutations to Maintain Thermodynamic Stability. *Scientific Reports*, 7:1–13.
- Julian, M. C., Rabia, L. A., Desai, A. A., Arsiwala, A., Gerson, J. E., Paulson, H. L., Kane, R. S., and Tessier, P. M. (2019). Nature-Inspired Design and Evolution of Anti-Amyloid Antibodies. *Journal of Biological Chemistry*, 294(21):8438–8451.
- Kaderabkova, N., Bharathwaj, M., Furniss, R. C. D., Gonzalez, D., Palmer, T., and Mavridou, D. A. I. (2022). The biogenesis of  $\beta$ -lactamase enzymes. *Microbiology* (*Reading, England*), 168(8).
- Kaplon, H., Chenoweth, A., Crescioli, S., and Reichert, J. M. (2022). Antibodies to Watch in 2022. *mAbs*, 14(1):2014296.

- Kaplon, H., Muralidharan, M., Schneider, Z., and Reichert, J. M. (2020). Antibodies to Watch in 2020. *mAbs*, 12(1):1703531.
- Karplus, P. A. and Diederichs, K. (2012). Linking Crystallographic Model and Data Quality. *Science*, 336(6084):1030–1033.
- Kastritis, P. L. and Bonvin, A. M. J. J. (2013). On the binding affinity of macromolecular interactions: Daring to ask why proteins interact. *Journal of the Royal Society, Interface*, 10(79):20120835.
- Kay, J., Matteson, E. L., Dasgupta, B., Nash, P., Durez, P., Hall, S., Hsia, E. C., Han, J., Wagner, C., Xu, Z., Visvanathan, S., and Rahman, M. U. (2008). Golimumab in Patients with Active Rheumatoid Arthritis despite Treatment with Methotrexate: A Randomized, Double-Blind, Placebo-Controlled, Dose-Ranging Study. *Arthritis & Rheumatism*, 58(4):964–975.
- Kennedy, P. J., Oliveira, C., Granja, P. L., and Sarmento, B. (2017). Antibodies and Associates: Partners in Targeted Drug Delivery. *Pharmacology and Therapeutics*, 177:129–145.
- Kessler, M., Goldsmith, D., and Schellekens, H. (2006). Immunogenicity of Biopharmaceuticals. Nephrology, dialysis, transplantation : official publication of the European Dialysis and Transplant Association - European Renal Association, 21:v9–12.
- Khatib, S. E. and Salla, M. (2022). The mosaic puzzle of the therapeutic monoclonal antibodies and antibody fragments - A modular transition from full-length immunoglobulins to antibody mimetics. *Leukemia Research Reports*, 18:100335.
- Kheddo, P., Tracka, M., Armer, J., Dearman, R. J., Uddin, S., van der Walle, C. F., and Golovanov, A. P. (2014). The Effect of Arginine Glutamate on the Stability of Monoclonal Antibodies in Solution. *International Journal of Pharmaceutics*, 473(1-2):126–133.
- Khetan, R., Curtis, R., Deane, C. M., Hadsund, J. T., Kar, U., Krawczyk, K., Kuroda, D., Robinson, S. A., Sormanni, P., Tsumoto, K., Warwicker, J., and Martin, A. C. R. (2022). Current Advances in Biopharmaceutical Informatics: Guidelines, Impact and Challenges in the Computational Developability Assessment of Antibody Therapeutics. *mAbs*, 14(1):2020082.
- Kim, S. W., Park, S. B., Im, S. P., Lee, J. S., Jung, J. W., Gong, T. W., Lazarte, J. M. S., Kim, J., Seo, J.-S., Kim, J.-H., Song, J.-W., Jung, H. S., Kim, G. J., Lee, Y. J., Lim, S.-K., and Jung, T. S. (2018). Outer Membrane Vesicles from  $\beta$ -Lactam-Resistant Escherichia Coli Enable the Survival of  $\beta$ -Lactam-Susceptible E. Coli in the Presence of  $\beta$ -Lactam Antibiotics. *Scientific Reports*, 8(1):5402.
- Kim, W. and Hecht, M. H. (2006). Generic Hydrophobic Residues Are Sufficient to Promote Aggregation of the Alzheimer's Abeta42 Peptide. *Proceedings of the National Academy of Sciences of the United States of America*, 103(43):15824–15829.
- Kohler, G. and Milstein, C. (1975). Continuous Cultures of Fused Cells Secreting Antibody of Predefined Specificity. *Nature*, 256(1):495–497.

- Kohli, N., Jain, N., Geddie, M. L., Razlog, M., Xu, L., and Lugovskoy, A. A. (2015). A Novel Screening Method to Assess Developability of Antibody-like Molecules. *mAbs*, 7(4):752–758.
- Kothawala, A., Kilpatrick, K., Novoa, J. A., and Segatori, L. (2012). Quantitative Analysis of  $\alpha$ -Synuclein Solubility in Living Cells Using Split GFP Complementation. *PloS One*, 7(8):e43505.
- Krause, M. E., Martin, T. T., and Laurence, J. S. (2012). Mapping Site-Specific Changes That Affect Stability of the N-terminal Domain of Calmodulin. *Molecular Pharmaceutics*, 9(4):734–743.
- Kumar, A. (2022). Interpreting f-statistics in linear regression: Formula, Examples. https://vitalflux.com/interpreting-f-statistics-in-linear-regression-formula-examples/.
- Kumari, A., Rajput, R., Shrivastava, N., Somvanshi, P., and Grover, A. (2018). Synergistic Approaches Unraveling Regulation and Aggregation of Intrinsically Disordered  $\beta$ -Amyloids Implicated in Alzheimer's Disease. *The International Journal of Biochemistry* & *Cell Biology*, 99:19–27.
- Kunamneni, A., Ogaugwu, C., Bradfute, S., and Durvasula, R. (2020). Ribosome Display Technology: Applications in Disease Diagnosis and Control. *Antibodies*, 9(3):28.
- Kunamneni, A., Ye, C., Bradfute, S. B., and Durvasula, R. (2018). Ribosome display for the rapid generation of high-affinity Zika-neutralizing single-chain antibodies. *PloS One*, 13(11):e0205743.
- Kuper, C. and Jung, K. (2005). CadC-mediated Activation of the cadBA Promoter in Escherichia Coli. *Journal of Molecular Microbiology and Biotechnology*, 10(1):26–39.
- Kuriakose, A., Chirmule, N., and Nair, P. (2016). Immunogenicity of Biotherapeutics: Causes and Association with Posttranslational Modifications. *Journal of Immunology Research*, 2016:1298473.
- Kuriata, A., Iglesias, V., Pujols, J., Kurcinski, M., Kmiecik, S., and Ventura, S. (2019). Aggrescan3D (A3D) 2.0: Prediction and Engineering of Protein Solubility. *Nucleic Acids Research*, 47(W1):W300–W307.
- Kurki, P., Barry, S., Bourges, I., Tsantili, P., and Wolff-Holz, E. (2021). Safety, Immunogenicity and Interchangeability of Biosimilar Monoclonal Antibodies and Fusion Proteins: A Regulatory Perspective. *Drugs*, 81(16):1881–1896.
- Ladbury, J. E. and Chowdhry, B. Z. (1996). Sensing the heat: The application of isothermal titration calorimetry to thermodynamic studies of biomolecular interactions. *Chemistry* & *Biology*, 3(10):791–801.
- Langan, R. A., Boyken, S. E., Ng, A. H., Samson, J. A., Dods, G., Westbrook, A. M., Nguyen, T. H., Lajoie, M. J., Chen, Z., Berger, S., Mulligan, V. K., Dueber, J. E., Novak, W. R. P., El-Samad, H., and Baker, D. (2019). De novo design of bioactive protein switches. *Nature*, 572(7768):205–210.
- Lee, C. C., Julian, M. C., Tiller, K. E., Meng, F., DuConge, S. E., Akter, R., Raleigh, D. P., and Tessier, P. M. (2016). Design and Optimization of Anti-Amyloid Domain Antibodies Specific for  $\beta$ -Amyloid and Islet Amyloid Polypeptide. *Journal of Biological Chemistry*, 291(6):2858–2873.

- Lee, E.-C., Liang, Q., Ali, H., Bayliss, L., Beasley, A., Bloomfield-Gerdes, T., Bonoli, L., Brown, R., Campbell, J., Carpenter, A., Chalk, S., Davis, A., England, N., Fane-Dremucheva, A., Franz, B., Germaschewski, V., Holmes, H., Holmes, S., Kirby, I., and Bradley, A. (2014). Complete Humanization of the Mouse Immunoglobulin Loci Enables Efficient Therapeutic Antibody Discovery. *Nature Biotechnology*, 32:356–363.
- Leem, J., Dunbar, J., Georges, G., Shi, J., and Deane, C. M. (2016). ABodyBuilder: Automated Antibody Structure Prediction with Data– Driven Accuracy Estimation. *mAbs*, 8(7):1259–1268.
- Lénon, M., Ke, N., Ren, G., Meuser, M. E., Loll, P. J., Riggs, P., and Berkmen, M. (2021). A useful epitope tag derived from maltose binding protein. *Protein Science : A Publication of the Protein Society*, 30(6):1235–1246.
- Lerch, T. F., Sharpe, P., Mayclin, S. J., Edwards, T. E., Lee, E., Conlon, H. D., Polleck, S., Rouse, J. C., Luo, Y., and Zou, Q. (2017). Infliximab crystal structures reveal insights into self-association. *mAbs*, 9(5):874–883.
- Leung, D., Chen, E., and Goeddel, D. (1989). A Method for Random Mutagenesis of a Defined DNA Segment Using a Modified Polymerase Chain Reaction. *Technique*, 1:11–15.
- Levinthal, C. (1968). Are There Pathways for Protein Folding? *Journal de Chimie Physique*, 65(1):44–45.
- Li, H. (2011). A Statistical Framework for SNP Calling, Mutation Discovery, Association Mapping and Population Genetical Parameter Estimation from Sequencing Data. *Bioinformatics*, 27(21):2987–93.
- Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R. (2009). The Sequence Alignment/Map Format and SAMtools. *Bioinformatics*, 25(16):2078–2079.
- Lindman, S., Hernandez-Garcia, A., Szczepankiewicz, O., Frohm, B., and Linse, S. (2010). In Vivo Protein Stabilization Based on Fragment Complementation and a Split GFP System. *Proceedings of the National Academy of Sciences*, 107(46):19826–19831.
- Lindner, E. and White, S. H. (2014). Topology, Dimerization, and Stability of the Single-Span Membrane Protein CadC. *Journal of Molecular Biology*, 426(16):2942–2957.
- Lippow, S. M., Wittrup, K. D., and Tidor, B. (2007). Computational Design of Antibody-Affinity Improvement beyond in Vivo Maturation. *Nature Biotechnology*, 25(10):1171– 6.
- Littlechild, J. A. (2015). Enzymes from Extreme Environments and Their Industrial Applications. *Frontiers in Bioengineering and Biotechnology*, 3:1–9.
- Liu, Y., Caffry, I., Wu, J., Geng, S. B., Jain, T., Sun, T., Reid, F., Cao, Y., Estep, P., Yu, Y., Vásquez, M., Tessier, P. M., and Xu, Y. (2014). High-Throughput Screening for Developability during Early-Stage Antibody Discovery Using Self-Interaction Nanoparticle Spectroscopy. *mAbs*, 6(2):483–492.

- Lonberg, N., Taylor, L. D., Harding, F. A., Trounstine, M., Higgins, K. M., Schramm, S. R., Kuo, C. C., Mashayekh, R., Wymore, K., McCabe, J. G., Munoz-O'Regan, D., O'Donnell, S. L., Lapachet, E. S., Bengoechea, T., Fishwild, D. M., Carmack, C. E., Kay, R. M., and Huszar, D. (1994). Antigen-Specific Human Antibodies from Mice Comprising Four Distinct Genetic Modifications. *Nature*, 368(6474):856–859.
- Lou, W., Stimple, S. D., Desai, A. A., Makowski, E. K., Kalyoncu, S., Mogensen, J. E., Spang, L. T., Asgreen, D. J., Staby, A., Duus, K., Amstrup, J., Zhang, Y., and Tessier, P. M. (2021). Directed Evolution of Conformation-specific Antibodies for Sensitive Detection of Polypeptide Aggregates in Therapeutic Drug Formulations. *Biotechnology* and Bioengineering, 118(2):797–808.
- Louros, N., Konstantoulea, K., De Vleeschouwer, M., Ramakers, M., Schymkowitz, J., and Rousseau, F. (2020). WALTZ-DB 2.0: An updated database containing structural information of experimentally determined amyloid-forming peptides. *Nucleic Acids Research*, 48(D1):D389–D393.
- Lu, P., Min, D., DiMaio, F., Wei, K. Y., Vahey, M. D., Boyken, S. E., Chen, Z., Fallas, J. A., Ueda, G., Sheffler, W., Mulligan, V. K., Xu, W., Bowie, J. U., and Baker, D. (2018). Accurate computational design of multipass transmembrane proteins. *Science (New York, N.Y.)*, 359(6379):1042–1046.
- Luo, R., Liu, H., and Cheng, Z. (2022). Protein scaffolds: Antibody alternatives for cancer diagnosis and therapy. *RSC Chem. Biol.*, 3(7):830–847.
- Maas, C., Hermeling, S., Bouma, B., Jiskoot, W., and Gebbink, M. F. B. G. (2007). A Role for Protein Misfolding in Immunogenicity of Biopharmaceuticals. *Journal of Biological Chemistry*, 282(4):2229–2236.
- Magliery, T. J., Wilson, C. G. M., Pan, W., Mishler, D., Ghosh, I., Hamilton, A. D., and Regan, L. (2005). Detecting Protein-Protein Interactions with a Green Fluorescent Protein Fragment Reassembly Trap: Scope and Mechanism. *Journal of the American Chemical Society*, 127(1):146–157.
- Magnani, F., Serrano-Vega, M. J., Shibata, Y., Abdul-Hussein, S., Lebon, G., Miller-Gallacher, J., Singhal, A., Strege, A., Thomas, J. A., and Tate, C. G. (2016). A mutagenesis and screening strategy to generate optimally thermostabilized membrane proteins for structural studies. *Nature Protocols*, 11(8):1554–1571.
- Mahler, H.-C., Friess, W., Grauschopf, U., and Kiese, S. (2009). Protein Aggregation: Pathways, Induction Factors and Analysis. *Journal of Pharmaceutical Sciences*, 98(9):2909–2934.
- Mahler, S. M., Marquis, C. P., Brown, G., Roberts, A., and Hoogenboom, H. R. (1997). Cloning and Expression of Human V-genes Derived from Phage Display Libraries as Fully Assembled Human Anti-TNF $\alpha$  Monoclonal Antibodies. *Immunotechnology*, 3(1):31–43.
- Marcos, E., Chidyausiku, T. M., McShan, A. C., Evangelidis, T., Nerli, S., Carter, L., Nivón, L. G., Davis, A., Oberdorfer, G., Tripsianes, K., Sgourakis, N. G., and Baker, D. (2018). De Novo Design of a Non-Local β-Sheet Protein with High Stability and Accuracy. *Nature Structural and Molecular Biology*, 25(11):1028–1034.

- Martin, M. (2011). Cutadapt Removes Adapter Sequences from High-Throughput Sequencing Reads. *EMBnet.journal*, 17(1):10.
- Marvin, J. S. and Lowman, H. B. (2003). Redesigning an Antibody Fragment for Faster Association with Its Antigen. *Biochemistry*, 42(23):7077–83.
- Mateu, M. G., Andreu, D., Carreño, C., Roig, X., Cairó, J.-J. J., Camarero, J. A., Giralt, E., and Domingo, E. (1992). Non-Additive Effects of Multiple Amino Acid Substitutions on Antigen-Antibody Recognition. *European Journal of Immunology*, 22(6):1385–9.
- Maurer-Stroh, S., Debulpaep, M., Kuemmerer, N., Lopez de la Paz, M., Martins, I. C., Reumers, J., Morris, K. L., Copland, A., Serpell, L., Serrano, L., Schymkowitz, J. W. H., and Rousseau, F. (2010). Exploring the sequence determinants of amyloid structure using position-specific scoring matrices. *Nature Methods*, 7(3):237–242.
- McCoy, A. J., Grosse-Kunstleve, R. W., Adams, P. D., Winn, M. D., Storoni, L. C., and Read, R. J. (2007). Phaser Crystallographic Software. *Journal of Applied Crystallography*, 40(Pt 4):658–674.
- McLure, R. J., Radford, S. E., and Brockwell, D. J. (2022). High-Throughput Directed Evolution: A Golden Era for Protein Science. *Trends in Chemistry*, 4(5).
- Meenan, N. A. G., Sharma, A., Fleishman, S. J., MacDonald, C. J., Morel, B., Boetzel, R., Moore, G. R., Baker, D., and Kleanthous, C. (2010). The structural and energetic basis for high selectivity in a high-affinity protein-protein interaction. *Proceedings of* the National Academy of Sciences, 107(22):10080–10085.
- Meltzer, M., Zvagelsky, T., Hadad, U., Papo, N., and Engel, S. (2022). Yeast-based directed-evolution for high-throughput structural stabilization of G protein-coupled receptors (GPCRs). *Scientific Reports*, 12(1):8657.
- Mihalcescu, I., Van-Melle Gateau, M., Chelli, B., Pinel, C., and Ravanat, J.-L. (2015). Green autofluorescence, a double edged monitoring tool for bacterial growth and activity in micro-plates. *Physical Biology*, 12(6):066016.
- Miller, S. M., Wang, T., Randolph, P. B., Arbab, M., Shen, M. W., Huang, T. P., Matuszek, Z., Newby, G. A., Rees, H. A., and Liu, D. R. (2020). Continuous Evolution of SpCas9 Variants Compatible with Non-G PAMs. *Nature Biotechnology*, 38(4):471–481.
- Moore, C. L., Papa, L. J., and Shoulders, M. D. (2018). A Processive Protein Chimera Introduces Mutations across Defined DNA Regions in Vivo. *Journal of the American Chemical Society*, 140(37):11560–11564.
- Morell, M., de Groot, N. S., Vendrell, J., Avilés, F. X., and Ventura, S. (2011). Linking Amyloid Protein Aggregation and Yeast Survival. *Molecular BioSystems*, 7(4):1121–1128.
- Morrison, M. S., Wang, T., Raguram, A., Hemez, C., and Liu, D. R. (2021). Disulfide-Compatible Phage-Assisted Continuous Evolution in the Periplasmic Space. *Nature Communications*, 12(1):5959.
- Morrison, S. L., Johnson, M. J., Herzenberg, L. A., and Oi, V. T. (1984). Chimeric Human Antibody Molecules: Mouse Antigen-Binding Domains with Human Constant Region Domains. *Proceedings of the National Academy of Sciences of the United States of America*, 81(21):6851–6855.

- Mosch, R. and Guchelaar, H.-J. (2022). Immunogenicity of Monoclonal Antibodies and the Potential Use of HLA Haplotypes to Predict Vulnerable Patients. *Frontiers in Immunology*, 13:885672.
- Mouquet, H., Scheid, J. F., Zoller, M. J., Krogsgaard, M., Ott, R. G., Shukair, S., Artyomov, M. N., Pietzsch, J., Connors, M., Pereyra, F., Walker, B. D., Ho, D. D., Wilson, P. C., Seaman, M. S., Eisen, H. N., Chakraborty, A. K., Hope, T. J., Ravetch, J. V., Wardemann, H., and Nussenzweig, M. C. (2010). Polyreactivity increases the apparent affinity of anti-HIV antibodies by heteroligation. *Nature*, 467(7315):591–595.
- Mudedla, S. K., Murugan, N. A., and Agren, H. (2018). Free Energy Landscape for Alpha-Helix to Beta-Sheet Interconversion in Small Amyloid Forming Peptide under Nanoconfinement. *Journal of Physical Chemistry B*, 122(42):9654–9664.
- Mullard, A. (2022). 2021 FDA approvals. Nature Reviews. Drug Discovery, 21(2):83-88.
- Murshudov, G. N., Skubák, P., Lebedev, A. A., Pannu, N. S., Steiner, R. A., Nicholls, R. A., Winn, M. D., Long, F., and Vagin, A. A. (2011). REFMAC5 for the Refinement of Macromolecular Crystal Structures. *Acta crystallographica. Section D, Biological crystallography*, 67(Pt 4):355–367.
- Myszka, D. G. and Rich, R. L. (2000). Implementing surface plasmon resonance biosensors in drug discovery. *Pharmaceutical Science & Technology Today*, 3(9):310–317.
- Navarro, S. and Ventura, S. (2022). Computational methods to predict protein aggregation. *Current Opinion in Structural Biology*, 73:102343.
- Nearmnala, P., Thanaburakorn, M., Panbangred, W., Chaiyen, P., and Hongdilokkul, N. (2021). An in Vivo Selection System with Tightly Regulated Gene Expression Enables Directed Evolution of Highly Efficient Enzymes. *Scientific Reports*, 11(1):11669.
- Nelson, A. L. (2010). Antibody Fragments: Hope and Hype. *mAbs*, 2(1):77–83.
- Newberry, R. W., Leong, J. T., Chow, E. D., Kampmann, M., and DeGrado, W. F. (2020). Deep Mutational Scanning Reveals the Structural Basis for  $\alpha$ -Synuclein Activity. *Nature Chemical Biology*, 16(6):653–659.
- Ng, A. H., Nguyen, T. H., Gómez-Schiavon, M., Dods, G., Langan, R. A., Boyken, S. E., Samson, J. A., Waldburger, L. M., Dueber, J. E., Baker, D., and El-Samad, H. (2019). Modular and tunable biological feedback control using a de novo protein switch. *Nature*, 572(7768):265–269.
- Nguyen, M. T., Krupa, M., Koo, B.-K., Song, J.-A., Vu, T. T. T., Do, B. H., Nguyen, A. N., Seo, T., Yoo, J., Jeong, B., Jin, J., Lee, K. J., Oh, H.-B., and Choe, H. (2016). Prokaryotic Soluble Overexpression and Purification of Human VEGF165 by Fusion to a Maltose Binding Protein Tag. *PLoS ONE*, 11(5):e0156296.
- Nilsson, M. R., Driscoll, M., and Raleigh, D. P. (2002). Low Levels of Asparagine Deamidation Can Have a Dramatic Effect on Aggregation of Amyloidogenic Peptides: Implications for the Study of Amyloid Formation. *Protein Science*, 11(2):342–349.
- Ohta, Y. and Flajnik, M. (2006). IgD, like IgM, Is a Primordial Immunoglobulin Class Perpetuated in Most Jawed Vertebrates. *Proceedings of the National Academy of Sciences of the United States of America*, 103(28):10723–8.

- Packer, M. S., Rees, H. A., and Liu, D. R. (2017). Phage-Assisted Continuous Evolution of Proteases with Altered Substrate Specificity. *Nature Communications*, 8(1):956.
- Pai, J. C., Entzminger, K. C., and Maynard, J. A. (2012). Restriction enzyme-free construction of random gene mutagenesis libraries in E. coli. *Analytical Biochemistry*, 421(2):640–648.
- Palmer, T. and Berks, B. C. (2012). The twin-arginine translocation (Tat) protein export pathway. *Nature Reviews. Microbiology*, 10(7):483–496.
- Palzkill, T. (2018). Structural and Mechanistic Basis for Extended-Spectrum Drug-Resistance Mutations in Altering the Specificity of TEM, CTX-M, and KPC  $\beta$ lactamases. *Frontiers in Molecular Biosciences*, 5.
- Park, H. and Kim, S. (2021). Gene-Specific Mutagenesis Enables Rapid Continuous Evolution of Enzymes in Vivo. *Nucleic Acids Research*, 49(6):e32.
- Park, S., Xu, Y., Stowell, X. F., Gai, F., Saven, J. G., and Boder, E. T. (2006). Limitations of Yeast Surface Display in Engineering Proteins of High Thermostability. *Protein Engineering, Design and Selection*, 19(5):211–217.
- Patro, S. Y. and Przybycien, T. M. (2000). Self-Interaction Chromatography: A Tool for the Study of Protein-Protein Interactions in Bioprocessing Environments. *Biotechnology* and Bioengineering, 52(2):193–203.
- Pavoor, T. V., Wheasler, J. A., Kamat, V., and Shusta, E. V. (2012). An Enhanced Approach for Engineering Thermally Stable Proteins Using Yeast Display. *Protein Engineering*, *Design and Selection*, 25(10):625–630.
- Peeler, J. C. and Mehl, R. A. (2012). Site-Specific Incorporation of Unnatural Amino Acids as Probes for Protein Conformational Changes. *Methods in Molecular Biology* (*Clifton, N.J.*), 794:125–134.
- Perez-Riba, A. and Itzhaki, L. S. (2017). A Method for Rapid High-Throughput Biophysical Analysis of Proteins. *Scientific Reports*, 7(1):1–6.
- Peyvandi, F., Scully, M., Kremer Hovinga, J. A., Cataland, S., Knöbl, P., Wu, H., Artoni, A., Westwood, J.-P., Mansouri Taleghani, M., Jilma, B., Callewaert, F., Ulrichts, H., Duby, C., and Tersago, D. (2016). Caplacizumab for Acquired Thrombotic Thrombocytopenic Purpura. *New England Journal of Medicine*, 374(25):2497.
- Pollo, M., Mehta, A., Torres, K., Thorne, D., Zimmermann, D., and Kolhe, P. (2019). Contribution of Intravenous Administration Components to Subvisible and Submicron Particles Present in Administered Drug Product. *Journal of Pharmaceutical Sciences*, 108(7):2406–2414.
- Qin, J., Li, R., Raes, J., Arumugam, M., Burgdorf, K. S., Manichanh, C., Nielsen, T., Pons, N., Levenez, F., Yamada, T., Mende, D. R., Li, J., Xu, J., Li, S., Li, D., Cao, J., Wang, B., Liang, H., Zheng, H., Xie, Y., Tap, J., Lepage, P., Bertalan, M., Batto, J.-M., Hansen, T., Le Paslier, D., Linneberg, A., Nielsen, H. B., Pelletier, E., Renault, P., Sicheritz-Ponten, T., Turner, K., Zhu, H., Yu, C., Li, S., Jian, M., Zhou, Y., Li, Y., Zhang, X., Li, S., Qin, N., Yang, H., Wang, J., Brunak, S., Doré, J., Guarner, F., Kristiansen, K., Pedersen, O., Parkhill, J., Weissenbach, J., Bork, P., Ehrlich, S. D., and Wang, J. (2010).

A Human Gut Microbial Gene Catalogue Established by Metagenomic Sequencing. *Nature*, 464(7285):59–65.

- Quiocho, F. A., Spurlino, J. C., and Rodseth, L. E. (1997). Extensive Features of Tight Oligosaccharide Binding Revealed in High-Resolution Structures of the Maltodextrin Transport/Chemosensory Receptor. *Structure*, 5(8):997–1015.
- Rabia, L. A., Desai, A. A., Jhajj, H. S., and Tessier, P. M. (2018). Understanding and overcoming trade-offs between antibody affinity, specificity, stability and solubility. *Biochemical engineering journal*, 137:365–374.
- Rader, R. A. (2008). (Re)Defining Biopharmaceutical. *Nature Biotechnology*, 26(7):743–751.
- Rangama, S., Lidbury, I. D. E. A., Holden, J. M., Borsetto, C., Murphy, A. R. J., Hawkey, P. M., and Wellington, E. M. H. (2021). Mechanisms Involved in the Active Secretion of CTX-M-15  $\beta$ -Lactamase by Pathogenic Escherichia coli ST131. *Antimicrobial Agents* and Chemotherapy, 65(10):e0066321.
- Raran-Kurussi, S., Sharwanlal, S. B., Balasubramanian, D., and Mote, K. R. (2022). A comparison between MBP- and NT\* as N-terminal fusion partner for recombinant protein production in E. coli. *Protein Expression and Purification*, 189:105991.
- Ratanji, K. D., Derrick, J. P., Dearman, R. J., and Kimber, I. (2014). Immunogenicity of Therapeutic Proteins: Influence of Aggregation. *Journal of Immunotoxicology*, 11(2):99–109.
- Ravikumar, A., Arrieta, A., and Liu, C. C. (2014). An Orthogonal DNA Replication System in Yeast. *Nature Chemical Biology*, 10(3):175–177.
- Raybould, M. I., Marks, C., Krawczyk, K., Taddese, B., Nowak, J., Lewis, A. P., Bujotzek, A., Shi, J., and Deane, C. M. (2019). Five Computational Developability Guidelines for Therapeutic Antibody Profiling. *Proceedings of the National Academy of Sciences of the United States of America*, 116(10):4025–4030.
- Rayfield, W. J., Kandula, S., Khan, H., and Tugcu, N. (2017). Impact of Freeze/Thaw Process on Drug Substance Storage of Therapeutics. *Journal of Pharmaceutical Sciences*, 106(8):1944–1951.
- Ren, C., Wen, X., Mencius, J., and Quan, S. (2021). An Enzyme-Based Biosensor for Monitoring and Engineering Protein Stability in Vivo. *Proceedings of the National Academy of Sciences of the United States of America*, 118(13):e2101618118.
- Rhoads, A. and Au, K. F. (2015). PacBio Sequencing and Its Applications. *Genomics, Proteomics and Bioinformatics*, 13(5):278–289.
- Ribatti, D. (2014). From the discovery of monoclonal antibodies to their therapeutic application: An historical reappraisal. *Immunology Letters*, 161(1):96–99.
- Richardson, J. S. and Richardson, D. C. (2002). Natural beta-sheet proteins use negative design to avoid edge-to-edge aggregation. *Proceedings of the National Academy of Sciences of the United States of America*, 99(5):2754–2759.

- Richter, M. F., Zhao, K. T., Eton, E., Lapinaite, A., Newby, G. A., Thuronyi, B. W., Wilson, C., Koblan, L. W., Zeng, J., Bauer, D. E., Doudna, J. A., and Liu, D. R. (2020). Phage-Assisted Evolution of an Adenine Base Editor with Improved Cas Domain Compatibility and Activity. *Nature Biotechnology*, 38(7):883–891.
- Rix, G., Watkins-Dulaney, E. J., Almhjell, P. J., Boville, C. E., Arnold, F. H., and Liu, C. C. (2020). Scalable Continuous Evolution for the Generation of Diverse Enzyme Variants Encompassing Promiscuous Activities. *Nature Communications*, 11(1):5644.
- Robert, X. and Gouet, P. (2014). Deciphering Key Features in Protein Structures with the New ENDscript Server. *Nucleic Acids Research*, 42(W1).
- Roberts, C. J. (2014a). Protein Aggregation and Its Impact on Product Quality. *Current Opinion in Biotechnology*, 30:211–217.
- Roberts, C. J. (2014b). Therapeutic Protein Aggregation: Mechanisms, Design, and Control. *Trends in Biotechnology*, 32(7):372–380.
- Rodgers, K. R. and Chou, R. C. (2016). Therapeutic Monoclonal Antibodies and Derivatives: Historical Perspectives and Future Directions. *Biotechnology Advances*, 34(6):1149–1158.
- Roopenian, D. C. and Akilesh, S. (2007). FcRn: The Neonatal Fc Receptor Comes of Age. *Nature Reviews Immunology*, 7(9):715–725.
- Rosowski, S., Becker, S., Toleikis, L., Valldorf, B., Grzeschik, J., Demir, D., Willenbücher, I., Gaa, R., Kolmar, H., Zielonka, S., and Krah, S. (2018). A Novel One-Step Approach for the Construction of Yeast Surface Display Fab Antibody Libraries. *Microbial Cell Factories*, 17(1):1–11.
- Roth, T. B., Woolston, B. M., Stephanopoulos, G., and Liu, D. R. (2019). Phage-Assisted Evolution of Bacillus Methanolicus Methanol Dehydrogenase 2. ACS Synthetic Biology, 8(4):796–806.
- Rousseau, F., Schymkowitz, J., van der Kant, R., De Baets, G., and Van Durme, J. (2015). Solubis: Optimize Your Protein. *Bioinformatics*, 31(15):2580–2582.
- Rousseau, F., Serrano, L., and Schymkowitz, J. W. (2006). How Evolutionary Pressure against Protein Aggregation Shaped Chaperone Specificity. *Journal of Molecular Biology*, 355(5):1037–1047.
- Rozbeh, R. and Forchhammer, K. (2021). Split NanoLuc Technology Allows Quantitation of Interactions between PII Protein and Its Receptors with Unprecedented Sensitivity and Reveals Transient Interactions. *Scientific Reports*, 11(1):12535.
- Rubin, A. F., Gelman, H., Lucas, N., Bajjalieh, S. M., Papenfuss, A. T., Speed, T. P., and Fowler, D. M. (2017). A Statistical Framework for Analyzing Deep Mutational Scanning Data. *Genome Biology*, 18(1):1–15.
- Safarnejad, M. R., Jouzani, G. S., Tabatabaie, M., Twyman, R. M., and Schillberg, S. (2011). Antibody-Mediated Resistance against Plant Pathogens. *Biotechnology Advances*, 29(6):961–971.

- Saito, Y., Oikawa, M., Sato, T., Nakazawa, H., Ito, T., Kameda, T., Tsuda, K., and Umetsu, M. (2021). Machine-Learning-Guided Library Design Cycle for Directed Evolution of Enzymes: The Effects of Training Data Composition on Sequence Space Exploration. ACS Catalysis, 11(23):14615–14624.
- Salema, V. and Fernández, L. Á. (2013). High Yield Purification of Nanobodies from the Periplasm of E. Coli as Fusions with the Maltose Binding Protein. *Protein Expression* and Purification, 91(1):42–48.
- Salverda, M. L. M., De Visser, J. A. G. M., and Barlow, M. (2010). Natural evolution of TEM-1  $\beta$ -lactamase: Experimental reconstruction and clinical relevance. *FEMS microbiology reviews*, 34(6):1015–1036.
- Sant'Anna, R., Braga, C., Varejão, N., Pimenta, K. M., Graña-Montes, R., Alves, A., Cortines, J., Cordeiro, Y., Ventura, S., and Foguel, D. (2014). The Importance of a Gatekeeper Residue on the Aggregation of Transthyretin. *Journal of Biological Chemistry*, 289(41):28324–28337.
- Santos, J., Pujols, J., Pallarès, I., Iglesias, V., and Ventura, S. (2020). Computational prediction of protein aggregation: Advances in proteomics, conformation-specific algorithms and biotechnological applications. *Computational and Structural Biotechnology Journal*, 18:1403–1413.
- Saunders, J. C. (2014). An in Vivo Platform for Identifying Protein Aggregation Inhibitors. PhD thesis, University of Leeds.
- Saunders, J. C., Young, L. M., Mahood, R. A., Jackson, M. P., Revill, C. H., Foster, R. J., Smith, D. A., Ashcroft, A. E., Brockwell, D. J., and Radford, S. E. (2016). An in Vivo Platform for Identifying Inhibitors of Protein Aggregation. *Nature Chemical Biology*, 12(2):94–101.
- Schadt, E. E., Turner, S., and Kasarskis, A. (2010). A Window into Third-Generation Sequencing. *Human Molecular Genetics*, 19(R2):R227–240.
- Schierle, C. F., Berkmen, M., Huber, D., Kumamoto, C., Boyd, D., and Beckwith, J. (2003). The DsbA Signal Sequence Directs Efficient, Cotranslational Export of Passenger Proteins to the Escherichia Coli Periplasm via the Signal Recognition Particle Pathway. *Journal of Bacteriology*, 185(19):5706–5713.
- Schroeder, H. W. and Cavacini, L. (2010). Structure and Function of Immunoglobulins. *Journal of Allergy and Clinical Immunology*, 125(2):S41–S52.
- Secco, P., D'Agostini, E., Marzari, R., Licciulli, M., Di Niro, R., D'Angelo, S., Bradbury, A. R. M., Dianzani, U., Santoro, C., and Sblattero, D. (2009). Antibody library selection by the  $\beta$ -lactamase protein fragment complementation assay. *Protein engineering, design & selection: PEDS*, 22(3):149–158.
- Seuma, M., Faure, A. J., Badia, M., Lehner, B., and Bolognesi, B. (2021). The Genetic Landscape for Amyloid Beta Fibril Nucleation Accurately Discriminates Familial Alzheimer's Disease Mutations. *eLife*, 10:e63364.
- Shah, M. (2018). Commentary: New Perspectives on Protein Aggregation during Biopharmaceutical Development. *International Journal of Pharmaceutics*, 552(1-2):1–6.

- Sharff, A. J., Rodseth, L. E., Spurlino, J. C., and Quiocho, F. A. (1992). Crystallographic Evidence of a Large Ligand-Induced Hinge-Twist Motion between the Two Domains of the Maltodextrin Binding Protein Involved in Active Transport and Chemotaxis. *Biochemistry*, 31(44):10657–10663.
- Sheinerman, F. B., Norel, R., and Honig, B. (2000). Electrostatic aspects of protein-protein interactions. *Current Opinion in Structural Biology*, 10(2):153–159.
- Shifman, J. M. (2008). Intricacies of  $\beta$  Sheet Protein Design. *Structure*, 16(12):1751–1752.
- Siepen, J. A., Radford, S. E., and Westhead, D. R. (2003). Beta edge strands in protein structure prediction and aggregation. *Protein Science: A Publication of the Protein Society*, 12(10):2348–2359.
- Silva, D.-A., Yu, S., Ulge, U. Y., Spangler, J. B., Jude, K. M., Labão-Almeida, C., Ali, L. R., Quijano-Rubio, A., Ruterbusch, M., Leung, I., Biary, T., Crowley, S. J., Marcos, E., Walkey, C. D., Weitzner, B. D., Pardo-Avila, F., Castellanos, J., Carter, L., Stewart, L., Riddell, S. R., Pepper, M., Bernardes, G. J. L., Dougan, M., Garcia, K. C., and Baker, D. (2019). De novo design of potent and selective mimics of IL-2 and IL-15. *Nature*, 565(7738):186–191.
- Smith, G. P. (1985). Filamentous Fusion Phage: Novel Expression Vectors That Display Cloned Antigens on the Virion Surface. *Science*, 228(4705):1315–1317.
- Sormanni, P., Aprile, F. A., and Vendruscolo, M. (2015a). The CamSol Method of Rational Design of Protein Mutants with Enhanced Solubility. *Journal of Molecular Biology*, 427(2):478–490.
- Sormanni, P., Aprile, F. A., and Vendruscolo, M. (2015b). Rational Design of Antibodies Targeting Specific Epitopes within Intrinsically Disordered Proteins. *Proceedings of the National Academy of Sciences of the United States of America*, 112(32):9902–9907.
- Sparrer, H., Rutkat, K., and Buchner, J. (1997). Catalysis of Protein Folding by Symmetric Chaperone Complexes. *Proceedings of the National Academy of Sciences of the United States of America*, 94(4):1096–1100.
- Spotts, J. M., Dolmetsch, R. E., and Greenberg, M. E. (2002). Time-lapse imaging of a dynamic phosphorylation-dependent protein–protein interaction in mammalian cells. *Proceedings of the National Academy of Sciences*, 99(23):15142–15147.
- Starr, C. G., Makowski, E. K., Wu, L., Berg, B., Kingsbury, J. S., Gokarn, Y. R., and Tessier, P. M. (2021). Ultradilute Measurements of Self-Association for the Identification of Antibodies with Favorable High-Concentration Solution Properties. *Molecular Pharmaceutics*.
- Starr, C. G. and Tessier, P. M. (2019). Selecting and Engineering Monoclonal Antibodies with Drug-like Specificity. *Current Opinion in Biotechnology*, 60:119–127.
- Stefanis, L. (2012). α-Synuclein in Parkinson's Disease. *Cold Spring Harbor Perspectives in Medicine*, 2(2):a009399–a009399.
- Stemmer, W. P. (1994). DNA Shuffling by Random Fragmentation and Reassembly: In Vitro Recombination for Molecular Evolution. *Proceedings of the National Academy of Sciences*, 91(22):10747–10751.

- Stimple, S. D., Smith, M. D., and Tessier, P. M. (2020). Directed Evolution Methods for Overcoming Trade-Offs between Protein Activity and Stability. *AIChE Journal*, 66(3).
- Strausberg, R. L., Levy, S., and Rogers, Y.-H. (2008). Emerging DNA Sequencing Technologies for Human Genomic Medicine. *Drug Discovery Today*, 13(13-14):569– 577.
- Stricher, F., Rousseau, F., Borg, J., Schymkowitz, J., Serrano, L., and Nys, R. (2005). The FoldX Web Server: An Online Force Field. *Nucleic Acids Research*, 33:W382–W388.
- Sun, T., Reid, F., Liu, Y., Cao, Y., Estep, P., Nauman, C., and Xu, Y. (2013). High Throughput Detection of Antibody Self-Interaction by Bio-Layer Interferometry. *mAbs*, 5(6):838–41.
- Surre, J., Saint-Ruf, C., Collin, V., Orenga, S., Ramjeet, M., and Matic, I. (2018). Strong increase in the autofluorescence of cells signals struggle for survival. *Scientific Reports*, 8(1):12088.
- Swindells, M. B., Porter, C. T., Couch, M., Hurst, J., Abhinandan, K., Nielsen, J. H., Macindoe, G., Hetherington, J., and Martin, A. C. (2017). abYsis: Integrated Antibody Sequence and Structure— Management, Analysis, and Prediction. *Journal of Molecular Biology*, 429(3):356–364.
- Syedbasha, M., Linnik, J., Santer, D., O'Shea, D., Barakat, K., Joyce, M., Khanna, N., Tyrrell, D. L., Houghton, M., and Egli, A. (2016). An ELISA Based Binding and Competition Method to Rapidly Determine Ligand-receptor Interactions. *Journal of Visualized Experiments : JoVE*, (109):53575.
- Tamura, R., Jiang, F., Xie, J., and Kamiyama, D. (2021). Multiplexed Labeling of Cellular Proteins with Split Fluorescent Protein Tags. *Communications Biology*, 4(1):1–8.
- Taverna, D. M. and Goldstein, R. A. (2002). Why Are Proteins Marginally Stable? *Proteins: Structure, Function and Genetics*, 46(1):105–109.
- Thuronyi, B. W., Koblan, L. W., Levy, J. M., Yeh, W.-H., Zheng, C., Newby, G. A., Wilson, C., Bhaumik, M., Shubina-Oleinik, O., Holt, J. R., and Liu, D. R. (2019). Continuous Evolution of Base Editors with Expanded Target Compatibility and Improved Activity. *Nature Biotechnology*, 37(9):1070–1079.
- Tian, Y. and Ruotolo, B. T. (2018). The Growing Role of Structural Mass Spectrometry in the Discovery and Development of Therapeutic Antibodies. *Analyst*, 143(11):2459–2468.
- Tiller, K. E., Chowdhury, R., Li, T., Ludwig, S. D., Sen, S., Maranas, C. D., and Tessier, P. M. (2017a). Facile Affinity Maturation of Antibody Variable Domains Using Natural Diversity Mutagenesis. *Frontiers in Immunology*, 8:986.
- Tiller, K. E., Li, L., Kumar, S., Julian, M. C., Garde, S., and Tessier, P. M. (2017b). Arginine Mutations in Antibody Complementarity-Determining Regions Display Context-Dependent Affinity/Specificity Trade-Offs. *Journal of Biological Chemistry*, 292(40):16638–16652.

- Tipper, D. J. and Strominger, J. L. (1965). Mechanism of action of penicillins: A proposal based on their structural similarity to acyl-D-alanyl-D-alanine. *Proceedings of the National Academy of Sciences of the United States of America*, 54(4):1133.
- Tomar, D. S., Kumar, S., Singh, S. K., Goswami, S., and Li, L. (2016). Molecular Basis of High Viscosity in Concentrated Antibody Solutions: Strategies for High Concentration Drug Product Development. *mAbs*, 8(2):216–228.
- Tooke, C. L., Hinchliffe, P., Bragginton, E. C., Colenso, C. K., Hirvonen, V. H., Takebayashi, Y., and Spencer, J. (2019).  $\beta$ -Lactamases and  $\beta$ -Lactamase Inhibitors in the 21st Century. *Journal of Molecular Biology*, 431(18):3472–3500.
- Tou, C. J., Schaffer, D. V., and Dueber, J. E. (2020). Targeted Diversification in the S. Cerevisiae Genome with CRISPR-guided DNA Polymerase I. ACS Synthetic Biology, 9(7):1911–1916.
- Trinh, C. H., Smith, D. P., Kalverda, A. P., Phillips, S. E. V., and Radford, S. E. (2002). Crystal structure of monomeric human beta-2-microglobulin reveals clues to its amyloidogenic properties. *Proceedings of the National Academy of Sciences of the United States of America*, 99(15):9771–9776.
- Trivedi, M., Laurence, J., and Siahaan, T. (2009). The Role of Thiols and Disulfides on Protein Stability. *Current Protein & Peptide Science*, 10(6):614–625.
- Trovato, A., Seno, F., and Tosatto, S. C. E. (2007). The PASTA Server for Protein Aggregation Prediction. *Protein Engineering, Design and Selection*, 20(10):521–523.
- Turner, P. J. (2005). Extended-Spectrum  $\beta$ -Lactamases. *Clinical Infectious Diseases*, 41(Supplement\_4):S273–S275.
- van der Kant, R., Karow-Zwick, A. R., Van Durme, J., Blech, M., Gallardo, R., Seeliger, D., Aßfalg, K., Baatsen, P., Compernolle, G., Gils, A., Studts, J. M., Schulz, P., Garidel, P., Schymkowitz, J., and Rousseau, F. (2017). Prediction and Reduction of the Aggregation of Monoclonal Antibodies. *Journal of Molecular Biology*, 429(8):1244–1261.
- van Reis, R. and Zydney, A. (2007). Bioprocess Membrane Technology. *Journal of Membrane Science*, 297(1-2):16–50.
- Vandenameele, J., Lejeune, A., Di Paolo, A., Brans, A., Frère, J.-M., Schmid, F. X., and Matagne, A. (2010). Folding of Class A β-Lactamases Is Rate-Limited by Peptide Bond Isomerization and Occurs via Parallel Pathways. *Biochemistry*, 49(19):4264–4275.
- Ventura, S. (2005). Sequence Determinants of Protein Aggregation: Tools to Increase Protein Solubility. *Microbial Cell Factories*, 4(1):11.
- Vodnik, M., Zager, U., Strukelj, B., and Lunder, M. (2011). Phage Display: Selecting Straws Instead of a Needle from a Haystack. *Molecules*, 16(1):790–817.
- Vorobieva, A. A., White, P., Liang, B., Horne, J. E., Bera, A. K., Chow, C. M., Gerben, S., Marx, S., Kang, A., Stiving, A. Q., Harvey, S. R., Marx, D. C., Khan, G. N., Fleming, K. G., Wysocki, V. H., Brockwell, D. J., Tamm, L. K., Radford, S. E., and Baker, D. (2021). De Novo Design of Transmembrane  $\beta$  Barrels. *Science*, 371(6531).

- Vuignier, K., Schappler, J., Veuthey, J.-L., Carrupt, P.-A., and Martel, S. (2010). Drugprotein binding: A critical review of analytical tools. *Analytical and Bioanalytical Chemistry*, 398(1):53–66.
- Walia, A., Guleria, S., Mehta, P., Chauhan, A., and Parkash, J. (2017). Microbial Xylanases and Their Industrial Application in Pulp and Paper Biobleaching: A Review. *3 Biotech*, 7(1):1–12.
- Walsh, G. (2018). Biopharmaceutical Benchmarks 2018. Nature Biotechnology, 36:1136.
- Wang, H., Wu, L., Wang, C., Xu, J., Yin, H., Guo, H., Zheng, L., Shao, H., and Chen, G. (2021). Biosimilar or Not: Physicochemical and Biological Characterization of MabThera and Its Two Biosimilar Candidates. ACS Pharmacology & Translational Science, 4(2):790–801.
- Wang, J., Zhang, T., Liu, R., Song, M., Wang, J., Hong, J., Chen, Q., and Liu, H. (2017). Recurring sequence-structure motifs in  $(B\alpha)$ 8-barrel proteins and experimental optimization of a chimeric protein designed based on such motifs. *Biochimica et Biophysica Acta (BBA) Proteins and Proteomics*, 1865(2):165–175.
- Wang, L., Brock, A., and Schultz, P. G. (2002). Adding L-3-(2-Naphthyl)Alanine to the Genetic Code of E. Coli. *Journal of the American Chemical Society*, 124(9):1836–1837.
- Wang, T., Badran, A. H., Huang, T. P., and Liu, D. R. (2018). Continuous Directed Evolution of Proteins with Improved Soluble Expression. *Nature Chemical Biology*, 14(10):972–980.
- Wang, W. (2005). Protein Aggregation and Its Inhibition in Biopharmaceutics. International Journal of Pharmaceutics, 289(1-2):1–30.
- Wang, W., Nema, S., and Teagarden, D. (2010). Protein Aggregation Pathways and Influencing Factors. *International Journal of Pharmaceutics*, 390(2):88–89.
- Wang, W. and Roberts, C. J. (2018). Protein Aggregation Mechanisms, Detection, and Control. *International Journal of Pharmaceutics*, 550(1):251–268.
- Wang, W., Singh, S., Zeng, D. L., King, K., and Nema, S. (2007). Antibody Structure, Instability, and Formulation. *Journal of Pharmaceutical Sciences*, 96(1):1–26.
- Wang, X., Das, T. K., Singh, S. K., and Kumar, S. (2009). Potential Aggregation Prone Regions in Biotherapeutics: A Survey of Commercial Monoclonal Antibodies. *mAbs*, 1(3):254–267.
- Warszawski, S., Katz, A. B., Lipsh, R., Khmelnitsky, L., Nissan, G. B., Javitt, G., Dym, O., Unger, T., Knop, O., Albeck, S., Diskin, R., Fass, D., Sharon, M., and Fleishman, S. J. (2019). Optimizing Antibody Affinity and Stability by the Automated Design of the Variable Light-Heavy Chain Interfaces. *PLoS Computational Biology*, 15(8):1–24.
- Wehrman, T., Kleaveland, B., Her, J.-H., Balint, R. F., and Blau, H. M. (2002). Protein-protein interactions monitored in mammalian cells via complementation of  $\beta$ lactamase enzyme fragments. *Proceedings of the National Academy of Sciences*, 99(6):3469–3474.

- Wellner, A., McMahon, C., Gilman, M. S. A., Clements, J. R., Clark, S., Nguyen, K. M., Ho, M. H., Hu, V. J., Shin, J.-E., Feldman, J., Hauser, B. M., Caradonna, T. M., Wingler, L. M., Schmidt, A. G., Marks, D. S., Abraham, J., Kruse, A. C., and Liu, C. C. (2021). Rapid Generation of Potent Antibodies by Autonomous Hypermutation in Yeast. *Nature Chemical Biology*, 17(10):1057–1064.
- Wheeler, D. A., Srinivasan, M., Egholm, M., Shen, Y., Chen, L., McGuire, A., He, W., Chen, Y.-J., Makhijani, V., Roth, G. T., Gomes, X., Tartaro, K., Niazi, F., Turcotte, C. L., Irzyk, G. P., Lupski, J. R., Chinault, C., Song, X.-z., Liu, Y., Yuan, Y., Nazareth, L., Qin, X., Muzny, D. M., Margulies, M., Weinstock, G. M., Gibbs, R. A., and Rothberg, J. M. (2008). The Complete Genome of an Individual by Massively Parallel DNA Sequencing. *Nature*, 452(7189):872–876.
- White, D. A., Buell, A. K., Knowles, T. P., Welland, M. E., and Dobson, C. M. (2010). Protein Aggregation in Crowded Environments. *Journal of the American Chemical Society*, 132(14):5170–5175.
- Willander, M. and Al-Hilli, S. (2009). Analysis of biomolecules using surface plasmons. Methods in Molecular Biology (Clifton, N.J.), 544:201–229.
- Willis, J. C. W. and Chin, J. W. (2018). Mutually Orthogonal Pyrrolysyl-tRNA Synthetase/tRNA Pairs. *Nature Chemistry*, 10(8):831–837.
- Willis, L. F., Kumar, A., Jain, T., Caffry, I., Xu, Y., Radford, S. E., Kapur, N., Vásquez, M., and Brockwell, D. J. (2020). The Uniqueness of Flow in Probing the Aggregation Behavior of Clinically Relevant Antibodies. *Engineering Reports*, 2:e12147.
- Winter, G. (2010). Xia2: An Expert System for Macromolecular Crystallography Data Reduction. *Journal of Applied Crystallography*, 43(1):186–190.
- Winter, G., Griffiths, A. D., Hawkins, R. E., and Hoogenboom, H. R. (1994). Making Antibodies by Phage Display Technology. Annual Review of Immunology, 12(1):433– 455.
- Winter, G., Waterman, D. G., Parkhurst, J. M., Brewster, A. S., Gildea, R. J., Gerstel, M., Fuentes-Montero, L., Vollmar, M., Michels-Clark, T., Young, I. D., Sauter, N. K., and Evans, G. (2018). DIALS: Implementation and Evaluation of a New Integration Package. Acta Crystallographica. Section D, Structural Biology, 74(Pt 2):85–97.
- Wojcik, J., Hantschel, O., Grebien, F., Kaupe, I., Bennett, K. L., Barkinge, J., Jones, R. B., Koide, A., Superti-Furga, G., and Koide, S. (2010). A Potent and Highly Specific FN3 Monobody Inhibitor of the Abl SH2 Domain. *Nature Structural and Molecular Biology*, 17(4):519–27.
- Wolf Pérez, A.-M., Sormanni, P., Andersen, J. S., Sakhnini, L. I., Rodriguez-Leon, I., Bjelke, J. R., Gajhede, A. J., De Maria, L., Otzen, D. E., Vendruscolo, M., and Lorenzen, N. (2019). In Vitro and in Silico Assessment of the Developability of a Designed Monoclonal Antibody Library. *mAbs*, 11(2):388–400.
- Wu, J., Schultz, J. S., Weldon, C. L., Sule, S. V., Chai, Q., Geng, S. B., Dickinson, C. D., and Tessier, P. M. (2015). Discovery of Highly Soluble Antibodies Prior to Purification Using Affinity-Capture Self-Interaction Nanoparticle Spectroscopy. *Protein Engineering, Design and Selection*, 28(10):403–414.

- Xiong, P., Wang, M., Zhou, X., Zhang, T., Zhang, J., Chen, Q., and Liu, H. (2014). Protein Design with a Comprehensive Statistical Energy Function and Boosted by Experimental Selection for Foldability. *Nature Communications*, 5:5330.
- Xu, Y., Roach, W., Sun, T., Jain, T., Prinz, B., Yu, T.-Y., Torrey, J., Thomas, J., Bobrowicz, P., Vasquez, M., Wittrup, K. D., and Krauland, E. (2013). Addressing Polyspecificity of Antibodies Selected from an in Vitro Yeast Presentation System: A FACS-based, High-Throughput Selection and Analytical Tool. *Protein Engineering, Design and Selection*, 26(10):663–670.
- Yan, Q., Huang, M., Lewis, M. J., and Hu, P. (2018). Structure Based Prediction of Asparagine Deamidation Propensity in Monoclonal Antibodies. *mAbs*, 10(6):901–912.
- Yao, Z., Drecun, L., Aboualizadeh, F., Kim, S. J., Li, Z., Wood, H., Valcourt, E. J., Manguiat, K., Plenderleith, S., Yip, L., Li, X., Zhong, Z., Yue, F. Y., Closas, T., Snider, J., Tomic, J., Drews, S. J., Drebot, M. A., McGeer, A., Ostrowski, M., Mubareka, S., Rini, J. M., Owen, S., and Stagljar, I. (2021). A Homogeneous Split-Luciferase Assay for Rapid and Sensitive Detection of Anti-SARS CoV-2 Antibodies. *Nature Communications*, 12(1):1806.
- You, L. and Arnold, F. H. (1996). Directed Evolution of Subtilisin E in Bacillus Subtilis to Enhance Total Activity in Aqueous Dimethylformamide. *Protein Engineering*, 9(1):77– 83.
- Yu, H. and Dalby, P. A. (2018). Coupled Molecular Dynamics Mediate Long- and Short-Range Epistasis between Mutations That Affect Stability and Aggregation Kinetics. *Proceedings of the National Academy of Sciences of the United States of America*, 115(47):E11043–E11052.
- Zambrano, R., Jamroz, M., Szczasiuk, A., Pujols, J., Kmiecik, S., and Ventura, S. (2015). AGGRESCAN3D (A3D): Server for Prediction of Aggregation Properties of Protein Structures. *Nucleic Acids Research*, 43(W1):W306–W313.
- Zapadka, K. L., Becher, F. J., Gomes Dos Santos, A. L., and Jackson, S. E. (2017). Factors Affecting the Physical Stability (Aggregation) of Peptide Therapeutics. *Interface Focus*, 7(6):20170030.
- Zhang, Z., Pan, H., and Chen, X. (2009). Mass Spectrometry for Structural Characterization of Therapeutic Antibodies. *Mass Spectrometry Reviews*, 28(1):147–176.
- Zhao, X.-L., Chen, W.-Q., Yang, Z.-H., Li, J.-M., Zhang, S.-J., and Tian, L.-F. (2009). Selection and affinity maturation of human antibodies against rabies virus from a scFv gene library using ribosome display. *Journal of Biotechnology*, 144(4):253–258.
- Zheng, K., Bantog, C., and Bayer, R. (2011). The Impact of Glycosylation on Monoclonal Antibody Conformation and Stability. *mAbs*, 3(6):568–576.

# Appendix A

# Primers used in this study

Code	Sequence	Use
RJM001 RJM002	GAAGAGTTGGcGAAAGATCCGC CTCGTAAGACTTCAGCGC	V312A into lab MBP
RJM003 RJM004	GAAAGATACCgatccgAAAGTCACCG TCGAATTTCTTACCGACTTC	G32D+I33P into MBP
RJM005 RJM006	CCTCGAAAACgacCTGCTGACTG AACTCTTTTGCCAGCTCTTTG	Y283D into MBP
RJM007 RJM008	CACCTGGCCGgcgATTGCTGCTG AAGTACGGTTCTTGCAGG	L160A into MBP
RJM009 RJM010	accaagAACAACAACAATAACAATAACAAC gatacgAGTCTGCGCGTCTTTCAG	Mutate TSSS -> RITK C-term MBP
RJM011 RJM012	taaAACAACAACAATAACAATAAC CTTGGTGATACGAGTCTG	Add stop codon to MBP
RJM013 RJM014	CTTTCAGGGAAAAATCGAAGAAGGTAAACTG TACAGGTTTTCGTGATGGTGATGGTGATG	Insert TEV cleavage site after his tag in pMal-c5x MBP
RJM015 RJM016	CTGGCCGGCGgcgGCTGCTGACG GTGAAGTACGGTTCTTGCAGG	I161A into L160A MBP
RJM017 RJM018	GAAAGCGGGTgcgACCTTCCTGGTTG GCGCCAGCGTTATCCACG	L192A into L160A + I161A MBP
RJM019 RJM020	TGCGACCTTCgcgGTTGACCTGATTAAAAACAAACACATGAATGCAGAC CCCGCTTTCGCGCCAGCG	L195A into L160A + I161A + L192A MBP
RJM021 RJM022	ttcgcgGTTGACCTGATTAAAAACAAACACATGAATGCAGACAC ggtcgcACCCGCTTTCGCGCCAGC	L192A + L195A into L160A + I161A MBP
RJM023 RJM024	ATAACGGTCTgGCTGAAGTCG AGCCTTTATCGCCGTTAATC	Remove Bsal site from MBP
RJM025 RJM026	GTGAGCGTGGLTCTCGCGGTA CGGCTCCAGATTTATCAGCAATAAAC	Remove Bsal site in bla domain 2
RJM027 RJM028	GGTAGTGTGGaGTCTCCCCATGC ATCGGCGCTACGGCGTTT	Remove Bsal site in bla vector non coding region
RJM029 RJM030	CACCTGGCCGgcggcgGCTGCTGACG AAGTACGGTTCTTGCAGG	L160A + I161A into TEV MBP STOP
RJM035 RJM036	gagaccTCGAGCTCAGGATCCGGG GCCACCACCAGCAACC	Add first Bsal site to blaGS
RJM037	ggtctcGGAGCGGTTCCGGAAGCG	Add second Bsal site to blaGS

#### Table A.1 Primers used in Q5 mutagenesis and restriction digest cloning

	Use
CGGATCCTGAGCTCGAGG	
taaggTCAGGATCCGGGTCTCGG ggttaGCTCGAGGTCTCGCCACC	Add Bsu36I site and stop codon betweer BsaI sites in blaGG
ggggatccTTCCGGGAGCGGGAGCTC accctcgagTGAGCCACCACCAGATC	Swap XhoI and BamHI restriction sites in b-lactamase 64 GS linker
TGACATTATCattTGGGCACACGACCGC GGGCCATCGCCAGTTGCC	F61I substitution in MBP 4A
TATCTTCTGGaccCACGACCGCTTTGG ATGTCAGGGCCATCGCCA	A63T substitution in MBP 4A
AGAACTGAAAgtgAAAGGTAAGAGCG TTATCCAGCGCCGGGATC	A141V substitution in MBP 4A
CACCTGGCCGctGGCGGCTGCT AAGTACGGTTCTTGCAGGTTG	A160L substitution in MBP 4A
TAACGCTGGCaccAAAGCGGGTG TCCACGCCCACGTCTTTA	A188T substitution in MBP 4A
GAAAGCGGGTctgACCTTCGCGG GCGCCAGCGTTATCCACG	A192L substitution in MBP 4A
TGCGACCTTCctgGTTGACCTGATTAAAAACAAACACATGAATGCAGAC CCCGCTTTCGCGCCAGCG	A195L substitution in MBP 4A
CAACATCGACattAGCAAAGTGAATTATGGTGTAACGGTACTGCCG GACCATGCCCACGGGCCG	T237I substitution in MBP 4A
AAACAAACACgTGAATGCAGAC TTAATCAGGTCAACCGCG	M204V substitution in MBP 4A
CATGAATGCAŁACCCGATTAC TGTTTGTTTTTAATCAGGTC	D207Y substitution in MBP 4A
ATGAATGCAGECACCGATTAC GTGTTTGTTTTTAATCAGGTC	D207V substitution in MBP 4A
GCTGCCTTTALTAAAGGCGAAAC TTCTGCGATGGAGTAATC	N218I substitution in MBP 4A
TACTGCGGTG&TCAACGCCGC CGCACGGCATACCAGAAAG	I348L substitution in MBP 4A
TAGCGCGGGTgcgATGTTCATGTATTCTCC TCTTCGCCCCATGCATAA	Y87A substitution in HA4 monobody in MBP 4A
tagTCCATAAGATTAGCGGATC tcttGCTATGGCATAGCAAAGTG	5 substitution mutations to create pBADSTRONG
GGTGGTGGCTCGAGCggcagctc ACCGCTCCCGGATCCgcaggtgc	Clone HA4 into bla vector
CACCTGGCCGctgGCGGCTGCTG AAGTACGGTTCTTGCAGGTTGAAC	A160L substitution in MBP 4A
GGGTGCGACCatcGCGGTTGACC GCTTTCGCGCCAGCGTTATC	F194I substitution in MBP 4A
TGACATTATCatcTGGGCACACGACCG GGGCCATCGCCAGTTGCC	F61I mutation in MBP 4A
TGGCGCGAAAtcgGGTGCGACCT GCGTTATCCACGCCCACGTC	A190S in MEP 4A
CCGCTTTGGTgacTACGCTCAAT TCGTGTGCCCAGAAGATAATG	G69D in MBP 4A
TAAAAACAAAtacATGAATGCAGTCAC ATCAGGTCAACCGCGAAG	H203Y in MBP 4A
CGCGAAAGCGagt GCGACCTTCG CCAGCGTTATCCACGCCC	G191S in MBP 4A
TCAGACTGTCaatGAAGCCCTGAAAGACGC CGACCGCTGGCGGCGTTG	D358N in MBP 4A
	taaggTCAGGATCCGGGGTCTGG         ggggataGCTTGCGGGACGGGACGGCACGACCGACGACCGACGGACG

	Sequence	Use
RJM187	CTTCCAGGTCTCCGAAAAATCGAAGAAGGTAAACTGG	Amplify MBP 4A evolved variant from bla
RJM188	CTTCCAGGTCTCCTACTTGGTGATACGAGTCTGC	vector to clone into pMAL vector
RJM193	GTGAGCAAGGGTGAGGAGGATAACATGG	
RJM194	ACCAGGGCCCCCGCTACC	remove ATG from mNG2 1-10
RJM223	TAAAAACAAAtACATGAATGCAGACAC	H203Y in MBP 4A
RJM224	ATCAGGTCAACCGCGAAG	NZUSI III MBF 4A
RJM273	tcgctgttcatgttgccagctttttcgAGTCTGGAAAAACACAGC	Replace DsbA signal peptide with TorT
RJM274	caaaagtaaaaacagcaggacccgcatGAATTCCTCCTGGTACCG	
RJM275	teteattetetgggegagegtaetgeaegegAGTCTGGAAAAACACAGC	
RJM276	aacccaaaagcgacacgtaatgcctgtttcatGAATTCCTCCTGGTACCG	Replace DsbA signal peptide with TolB
RJM286	ATGCCATAGCtTTTTTATCCATAAGATTAGC	Current a PAP work another
RJM287	AGCAAAGTGTGACGCCGT	Create pBAD weak promoter
RJM290	ATGAAAAAGATTTGGCTGGCGC	Amplify pSNAC to remove split bla and
RJM291	gaatteeteetggtaeegage	keep only pBAD DsbA-SH2-mNG2
RJM292	GGTACAGGTCTCCTaagaaaccaattgtccatattgc	Amplify pSNAC DsbA to add split
RJM293	GGTACAGGTCTCCttctctgaatggcgggagta	bla-mNG2 and mScarlet-I
RJM294	GGTACAGGTCTCCagaagaaaccaattgtccatattgc	
RJM294	GGTACAGGTCTCCcttAAATAAACAAAAGAGTTTGTAGAAACG	
RJM298	TGACGAGGTCTCCAACTGCAGGTAATTAAATAAGCTTCA	
		Annelify solid bla sNCO and sCooplat T
RJM299	TGACGAGGTCTCCcttAAATAAACAAAAGAGTTTGTAGAAACG	Amplify split bla-mNG2 and mScarlet-I
RJM300	TGACGAGGTCTCGTaagaaaccaattgtccatattgc	to clone into pSNAC DsbA
RJM301	TGACGAGGTCTCGAGTTAcatggagtgcttgagc	
RJM303	GTGGTGGCTCGAGCGAACTGAAAAATAGTATTAGTGAT	
RJM304	GCTCCCGGATCCGCCCTGTTTAAATCCT	Amplify Im7 to clone into pSNAC
400000	Gereceggareegeerarranteer	
RJM309	GTTCGAGGTCTCCtcgggaagggaaggttct	
RJM310	GTTCGAGGTCTCCACCTAGTTCGCCAGTTAATAGTTTG	pSNAC_mNG2_weakDsbA
RJM311	GTTCGAGGTCTCGAGGTGGTGGTGGTTCTGGT	- PP 200 bl-CC CTOP
RJM312	GTTCGAGGTCTCGccgaCCAATGCTTAATCAGTGAGG	pBR322_blaGG_STOP
RJM315	GGGAATGGTCTCGGTGGCTCGAGCGAACTGAAAAATAGTATTAGTGAT	Amplify Im7 to clone into pSNAC_GG_STOP
RJM323	ACATGCGGTCTCCGCTCCCGGATCCGCCCTGTTTAAATCCT	mapility may co crone inco pomic_00_bior
RJM367		Cut off mNG2(11) from pSNAC weak mNG2
KUM300	CCAATGCTTAATCAGTGAG	
RJM369	tcaagcactccatgTAACTGCAGGTAAT	Cut off mNG2(1-10) from pSNAC weak mNG2
RJM369		Cut off mNG2(1-10) from pSNAC weak mNG2
RJM369 RJM370	tcaagcactccatgTAACTGCAGGTAAT	Cut off mNG2(1-10) from pSNAC weak mNG2
RJM369 RJM370 RJM371	tcaagcactccatgTAACTGCAGGTAAT TTAgcgtttcggcgccggataa	Cut off mNG2(1-10) from pSNAC weak mNG2 LllR mutant in bevacizumab scFab
RJM369 RJM370 RJM371	tcaagcactccatgTAACTGCAGGTAAT TTAgcgtttcggcgccggataa CCCCAGTTCGcgtTCAGCCTCGG	
RJM369 RJM370 RJM371 RJM372	tcaagcactccatgTAACTGCAGGTAAT TTAgcgtttcggcgccggataa CCCCAGTTCGcgtTCAGCCTCGG	L11R mutant in bevacizumab scFab
RJM369 RJM370 RJM371 RJM372 RJM373	tcaagcactccatgTAACTGCAGGTAAT TTAgcgtttcggcgccggataa CCCCAGTTCGcgtTCAGCCTCGG GATTGCGTCATTTGGATGTCGCTC	
RJM369 RJM370 RJM371 RJM372 RJM373 RJM374	tcaagcactccatgTAACTGCAGGTAAT TTAgcgtttcggcgccggataa CCCCAGTTCGcgtTCAGCCTCGG GATTGCGTCATTTGGATGTCGCTC AGCCTCGGTTcgtGACCGTGTTA GAAAGCGAACTGGGGGATTG	L11R mutant in bevacizumab scFab
RJM369 RJM370 RJM371 RJM372 RJM373 RJM374 RJM375	tcaagcactccatgTAACTGCAGGTAAT TTAgcgtttcggcgccggataa CCCCAGTTCGcgtTCAGCCTCGG GATTGCGTCATTTGGATGTCGCTC AGCCTCGGTTcgtGACCGTGTTA GAAAGCGAACTGGGGGATTG TGACCGTGTTgcaATCACTTGTTC	L11R mutant in bevacizumab scFab
RJM369 RJM370 RJM371 RJM372 RJM373 RJM374 RJM375	tcaagcactccatgTAACTGCAGGTAAT TTAgcgtttcggcgccggataa CCCCAGTTCGcgtTCAGCCTCGG GATTGCGTCATTTGGATGTCGCTC AGCCTCGGTTcgtGACCGTGTTA GAAAGCGAACTGGGGGATTG	L11R mutant in bevacizumab scFab G16R mutant in bevacizumab scFab
RJM369 RJM370 RJM371 RJM372 RJM373 RJM374 RJM375 RJM376	tcaagcactccatgTAACTGCAGGTAAT TTAgcgtttcggcgccggataa CCCCAGTTCGcgtTCAGCCTCGG GATTGCGTCATTTGGATGTCGCTC AGCCTCGGTTcgtGACCGTGTTA GAAAGCGAACTGGGGGATTG TGACCGTGTTgcaATCACTTGTTC CCAACCGAGGCTGAAAGC	L11R mutant in bevacizumab scFab G16R mutant in bevacizumab scFab
RJM369 RJM370 RJM371 RJM372 RJM373 RJM374 RJM375 RJM376 RJM377	tcaagcactccatgTAACTGCAGGTAAT TTAgcgtttcggcgccggataa CCCCAGTTCGcgtTCAGCCTCGG GATTGCGTCATTTGGATGTCGCTC AGCCTCGGTTcgtGACCGTGTTA GAAAGCGAACTGGGGGATTG TGACCGTGTTgcaATCACTTGTTC CCAACCGAGGCTGAAAGC CCGTGTTACAttcACTTGTTCAGC	L11R mutant in bevacizumab scFab G16R mutant in bevacizumab scFab
RJM369 RJM370 RJM371 RJM372 RJM373 RJM374 RJM375 RJM376 RJM377	tcaagcactccatgTAACTGCAGGTAAT TTAgcgtttcggcgccggataa CCCCAGTTCGcgtTCAGCCTCGG GATTGCGTCATTTGGATGTCGCTC AGCCTCGGTTcgtGACCGTGTTA GAAAGCGAACTGGGGGATTG TGACCGTGTTgcaATCACTTGTTC CCAACCGAGGCTGAAAGC	L11R mutant in bevacizumab scFab G16R mutant in bevacizumab scFab T20A mutant in bevacizumab scFab
RJM369 RJM370 RJM371 RJM372 RJM373 RJM374 RJM376 RJM377 RJM378	tcaagcactccatgTAACTGCAGGTAAT TTAgcgtttcggcgccggataa CCCCAGTTCGcgtTCAGCCTCGG GATTGCGTCATTTGGATGTCGCTC AGCCTCGGTTcgtGACCGTGTTA GAAAGCGAACTGGGGGATTG TGACCGTGTTgcaATCACTTGTTC CCAACCGAGGCTGAAAGC CCGTGTTACAttcACTTGTTCAGC TCACCAACCGAGGCTGAA	L11R mutant in bevacizumab scFab G16R mutant in bevacizumab scFab T20A mutant in bevacizumab scFab I21F mutant in bevacizumab scFab
RJM369 RJM370 RJM371 RJM372 RJM373 RJM374 RJM375 RJM376 RJM377 RJM379	tcaagcactccatgTAACTGCAGGTAAT TTAgcgtttcggcgccggataa CCCCAGTTCGcgtTCAGCCTCGG GATTGCGTCATTTGGATGTCGCTC AGCCTCGGTTcgtGACCGTGTTA GAAAGCGAACTGGGGGATTG TGACCGTGTTgcaATCACTTGTTC CCAACCGAGGCTGAAAGC CCGTGTTACAttcACTGTTCAGC TCACCAACCGAGGCTGAA CCCGGAAGGTAccgATCTACTTCAC	L11R mutant in bevacizumab scFab G16R mutant in bevacizumab scFab T20A mutant in bevacizumab scFab
RJM369 RJM370 RJM371 RJM372 RJM373 RJM374 RJM375 RJM376 RJM377 RJM379	tcaagcactccatgTAACTGCAGGTAAT TTAgcgtttcggcgccggataa CCCCAGTTCGcgtTCAGCCTCGG GATTGCGTCATTTGGATGTCGCTC AGCCTCGGTTcgtGACCGTGTTA GAAAGCGAACTGGGGGATTG TGACCGTGTTgcaATCACTTGTTC CCAACCGAGGCTGAAAGC CCGTGTTACAttcACTTGTTCAGC TCACCAACCGAGGCTGAA	L11R mutant in bevacizumab scFab G16R mutant in bevacizumab scFab T20A mutant in bevacizumab scFab I21F mutant in bevacizumab scFab
RJM369 RJM370 RJM371 RJM372 RJM373 RJM374 RJM376 RJM377 RJM378 RJM379 RJM379 RJM380	tcaagcactccatgTAACTGCAGGTAAT TTAgcgtttcggcgccggataa CCCCAGTTCGcgtTCAGCCTCGG GATTGCGTCATTTGGATGTCGCTC AGCCTCGGTTcgtGACCGTGTTA GAAAGCGAACTGGGGGATTG TGACCGTGTTgcaATCACTTGTTC CCAACCGAGGCTGAAAGC CCGTGTTACAttcACTGTTCAGC TCACCAACCGAGGCTGAA CCCGGAAGGTAccgATCTACTTCAC	L11R mutant in bevacizumab scFab G16R mutant in bevacizumab scFab T20A mutant in bevacizumab scFab I21F mutant in bevacizumab scFab L47P mutant in bevacizumab scFab
RJM369 RJM370 RJM371 RJM372 RJM373 RJM374 RJM376 RJM377 RJM378 RJM379 RJM380 RJM381	tcaagcactccatgTAACTGCAGGTAAT TTAgcgtttcggcgccggataa CCCCAGTTCGcgtTCAGCCTCGG GATTGCGTCATTTGGATGTCGCTC AGCCTCGGTTcgtGACCGTGTTA GAAAGCGAACTGGGGGATTG TGACCGTGTTgcaATCACTTGTTC CCAACCGAGGCTGAAAGC CCGTGTTACAttcACTTGTTCAGC TCACCAACCGAGGCTGAA CCCGGAAGGTAccgATCTACTTCAC GCCTTACCTGGCTTTTGT	L11R mutant in bevacizumab scFab G16R mutant in bevacizumab scFab T20A mutant in bevacizumab scFab I21F mutant in bevacizumab scFab
RJM369 RJM370 RJM371 RJM372 RJM374 RJM374 RJM376 RJM377 RJM378 RJM379 RJM380 RJM381 RJM381	tcaagcactccatgTAACTGCAGGTAAT TTAgcgtttcggcgccggataa CCCCAGTTCGcgtTCAGCCTCGG GATTGCGTCATTTGGATGTCGCTC AGCCTCGGTTgtGACCGTGTTA GAAGCGAACTGGGGGATTG TGACCGTGTTgcaATCACTTGTTC CCAACCGAGGCTGAAAGC CCGTGTTACAttcACTTGTTCAGC TCACCAACCGAGGCTGAA CCCCGAAGGTAccgATCTACTTCAC GCCTTACCTGGCTTTGT GTTACACTCTGgtGTCCCAAGTC GAGGATGTGAAGTAGATG	L11R mutant in bevacizumab scFab G16R mutant in bevacizumab scFab T20A mutant in bevacizumab scFab I21F mutant in bevacizumab scFab L47P mutant in bevacizumab scFab
RJM369 RJM370 RJM371 RJM372 RJM374 RJM374 RJM376 RJM376 RJM377 RJM378 RJM379 RJM380 RJM381 RJM382	tcaagcactccatgTAACTGCAGGTAAT TTAgcgtttcggcgccggataa CCCCAGTTCGcgtTCAGCCTCGG GATTGCGTCATTTGGATGTCGCTC AGCCTCGGTTcgtGACCGTGTTA GAAAGCGAACTGGGGGATTG TGACCGTGTTgcaATCACTTGTTC CCAACCGAGGCTGAAAGC CCGTGTTACAttcACTTGTTCAGC TCACCAACGGAGGCTGAA CCCCGAGGTAccgATCTACTTCAC GCCTTACCTGGCTTTGT GTTACACTCTagtGTCCCAAGTC	L11R mutant in bevacizumab scFab G16R mutant in bevacizumab scFab T20A mutant in bevacizumab scFab I21F mutant in bevacizumab scFab L47P mutant in bevacizumab scFab G57S mutant in bevacizumab scFab
RJM369 RJM370 RJM371 RJM372 RJM373 RJM374 RJM376 RJM377 RJM378 RJM379 RJM380 RJM381 RJM382 RJM383	tcaagcactccatgTAACTGCAGGTAAT TTAgcgtttcggcgccggataa CCCCAGTTCGcgtTCAGCCTCGG GATTGCGTCATTTGGATGTCGCTC AGCCTCGGTTgtGACCGTGTTA GAAGCGAACTGGGGGATTG TGACCGTGTTgcaATCACTTGTTC CCAACCGAGGCTGAAAGC CCGTGTTACAttcACTTGTTCAGC TCACCAACCGAGGCTGAA CCCCGAAGGTAccgATCTACTTCAC GCCTTACCTGGCTTTGT GTTACACTCTGgtGTCCCAAGTC GAGGATGTGAAGTAGATG	L11R mutant in bevacizumab scFab G16R mutant in bevacizumab scFab T20A mutant in bevacizumab scFab I21F mutant in bevacizumab scFab L47P mutant in bevacizumab scFab
RJM369 RJM370 RJM371 RJM372 RJM374 RJM375 RJM376 RJM376 RJM377 RJM378 RJM379 RJM380 RJM381 RJM383 RJM384	tcaagcactccatgTAACTGCAGGTAAT TTAgcgtttcggcgccggataa CCCCAGTTCGcgtTCAGCCTCGG GATTGCGTCATTTGGATGTCGCTC AGCCTCGGTTcgtGACCGTGTTA GAAAGCGAACTGGGGGATTG TGACCGTGTTgcaATCACTTGTTC CCAACCGAGGCTGAAAGC CCGTGTTACAttcACTTGTTCAGC TCACCAACCGAGGCTGAA CCCGAAGGTAccgATCTACTTCAC GCCTTACCTGGCTTTGT GTTACACTCTagtGTCCCAAGTC GAGGATGTGAAGTAGATG TCAACAATACccgACTGTTCCCTG CAGTAGTAAGTAGGAGG	L11R mutant in bevacizumab scFab G16R mutant in bevacizumab scFab T20A mutant in bevacizumab scFab I21F mutant in bevacizumab scFab L47P mutant in bevacizumab scFab G57S mutant in bevacizumab scFab
RJM368 RJM370 RJM371 RJM371 RJM373 RJM373 RJM374 RJM375 RJM376 RJM377 RJM377 RJM378 RJM379 RJM380 RJM381 RJM381 RJM383 RJM384 RJM385	tcaagcactccatgTAACTGCAGGTAAT         TTAgcgtttcggcgccggataa         CCCCAGTTCGcgtTCAGCCTCGG         GATTGCGTCATTTGGATGTCGCTC         AGCCTCGGTTcgtGACCGTGTTA         GAAAGCGAACTGGGGGATTG         TGACCGTGTTgcaATCACTTGTC         CCCAACCGAGGCTGAAAGC         CCGTGTTACAttcACTTGTTCAGC         TCACCAACCGAGGCTGAA         CCCGAAGGTAccgATCTACTTCAC         GCCTACCTAGTGTCCCAAGTC         GATACATACCGAACTGTTCCCTG         CAACAATACCcgACTGTTCCCTG         CAATACTCTGctGTTCCCTGGA	L11R mutant in bevacizumab scFab G16R mutant in bevacizumab scFab T20A mutant in bevacizumab scFab I21F mutant in bevacizumab scFab L47P mutant in bevacizumab scFab G57S mutant in bevacizumab scFab
RJM369 RJM370 RJM371 RJM372 RJM374 RJM375 RJM376 RJM376 RJM377 RJM378 RJM379 RJM380 RJM381 RJM383 RJM384	tcaagcactccatgTAACTGCAGGTAAT TTAgcgtttcggcgccggataa CCCCAGTTCGcgtTCAGCCTCGG GATTGCGTCATTTGGATGTCGCTC AGCCTCGGTTcgtGACCGTGTTA GAAAGCGAACTGGGGGATTG TGACCGTGTTgcaATCACTTGTTC CCAACCGAGGCTGAAAGC CCGTGTTACAttcACTTGTTCAGC TCACCAACCGAGGCTGAA CCCGAAGGTAccgATCTACTTCAC GCCTTACCTGGCTTTGT GTTACACTCTagtGTCCCAAGTC GAGGATGTGAAGTAGATG TCAACAATACccgACTGTTCCCTG CAGTAGTAAGTAGGAGG	L11R mutant in bevacizumab scFab G16R mutant in bevacizumab scFab T20A mutant in bevacizumab scFab I21F mutant in bevacizumab scFab L47P mutant in bevacizumab scFab G57S mutant in bevacizumab scFab S92P mutant in bevacizumab scFab
RJM369 RJM370 RJM371 RJM372 RJM373 RJM374 RJM375 RJM376 RJM376 RJM377 RJM378 RJM380 RJM381 RJM381 RJM383 RJM384 RJM385 RJM386	tcaagcactccatgTAACTGCAGGTAAT TTAgcgtttcggcgccggataa CCCCAGTTCGcgtTCAGCCTCGG GATTGCGTCATTTGGATGTCGCTC AGCCTCGGTCgtGACCGTGTTA GAAAGCGAACTGGGGGATTG TGACCGTGTTgcaATCACTTGTTC CCAACCGAGGCTGAAAGC CCGTGTTACAttcACTTGTTCAGC TCACCAACGAGGGCTGAA CCCGAAGGTAccgATCTACTTCAC GCCTTACCTGGCTTTACT GTTACACTCTgtGTCCCAGTC GAGGATGTGAAGTAGGAG TCAACAATACTCTgctGTTCCCTG CAGTAGTAAGTAGCGAAG ACAATACTCTgctGTTCCCTGGA TGACAGTAGTAAGTAGGG	L11R mutant in bevacizumab scFab G16R mutant in bevacizumab scFab T20A mutant in bevacizumab scFab I21F mutant in bevacizumab scFab L47P mutant in bevacizumab scFab G57S mutant in bevacizumab scFab S92P mutant in bevacizumab scFab
RJM369 RJM370 RJM371 RJM372 RJM374 RJM374 RJM376 RJM376 RJM377 RJM378 RJM378 RJM380 RJM381 RJM381 RJM384 RJM384 RJM385 RJM386 RJM387	tcaagcactccatgTAACTGCAGGTAAT TTAgcgtttcggcgccggataa CCCCAGTTCGcgtTCAGCCTCGG GATTGCGTCATTTGGATGTCGCTC AGCCTCGGTTcgtGACCGTGTTA GAAAGCGAACTGGGGGATTG TGACCGTGTTgcaATCACTTGTTC CCAACCGAGGCTGAAAGC CCGTGTTACAttcACTTGTTCAGC TCACCAACCGAGGCTGAA CCCGAAGGTAccgATCTACTTCAC GCCTTACCAtGAGTACTACTTCAC GCCTTACCTGGCTTTGT GTTACACTCTagtGTCCCAAGTC GAGGATGTGAAGTAGATG TCAACAATACCcgACTGTTCCCTG CAGTAGTAAGTAGCGAAG ACAATACTCTgctGTTCCCTGGA TGACAGTAGTAAGTAGCG CCTGGACATTCgatCAAGGTACTAAG	L11R mutant in bevacizumab scFab G16R mutant in bevacizumab scFab T20A mutant in bevacizumab scFab I21F mutant in bevacizumab scFab L47P mutant in bevacizumab scFab G57S mutant in bevacizumab scFab S92P mutant in bevacizumab scFab
RJM369 RJM370 RJM371 RJM372 RJM373 RJM374 RJM375 RJM376 RJM376 RJM377 RJM378 RJM380 RJM381 RJM381 RJM384 RJM385 RJM386	tcaagcactccatgTAACTGCAGGTAAT TTAgcgtttcggcgccggataa CCCCAGTTCGcgtTCAGCCTCGG GATTGCGTCATTTGGATGTCGCTC AGCCTCGGTCgtGACCGTGTTA GAAAGCGAACTGGGGGATTG TGACCGTGTTgcaATCACTTGTTC CCAACCGAGGCTGAAAGC CCGTGTTACAttcACTTGTTCAGC TCACCAACGAGGGCTGAA CCCGAAGGTAccgATCTACTTCAC GCCTTACCTGGCTTTACT GTTACACTCTgtGTCCCAGTC GAGGATGTGAAGTAGGAG TCAACAATACTCTgctGTTCCCTG CAGTAGTAAGTAGCGAAG ACAATACTCTgctGTTCCCTGGA TGACAGTAGTAAGTAGGG	L11R mutant in bevacizumab scFab G16R mutant in bevacizumab scFab T20A mutant in bevacizumab scFab I21F mutant in bevacizumab scFab L47P mutant in bevacizumab scFab G57S mutant in bevacizumab scFab S92P mutant in bevacizumab scFab
RJM369 RJM370 RJM371 RJM372 RJM374 RJM375 RJM376 RJM376 RJM377 RJM378 RJM378 RJM380 RJM381 RJM383 RJM384 RJM385 RJM386 RJM387 RJM387 RJM387	tcaagcactccatgTAACTGCAGGTAAT         TTAgcgtttcggcgccggataa         CCCCAGTTCGcgtTCAGCCTCGG         GATTGCGTCATTTGGATGTCGCTC         AGCCTCGGTCgtGACCGTGTTA         GAAAGCGAACTGGGGGATTG         TGACCGTGTTgcaATCACTTGTC         CCAACCGAGGCTGAAAGC         CCCGAGGTACcgATCACTTGTCAGC         TCAACCAACCGAGGCTGAA         CCCGAAGGTAccgATCTACTTCAC         GCCTTACCTGGCTTTGT         GTTACACTCTGGTCCCAAGTC         GAGCATGTGAAGTAGCAG         TCAACAATACccgACTGTTCCCTG         CAGTAGTAAGTAGCGAAG         ACAATACTCTgctGTTCCCTGGA         TGACAGTAGTAAGTAGCG         CTGGACATTCgatCAAGGTACTAAG         GTTGCAGTAGTAAGTAGCG	L11R mutant in bevacizumab scFab G16R mutant in bevacizumab scFab T20A mutant in bevacizumab scFab I21F mutant in bevacizumab scFab L47P mutant in bevacizumab scFab G57S mutant in bevacizumab scFab S92P mutant in bevacizumab scFab
RJM369           RJM370           RJM371           RJM371           RJM373           RJM374           RJM375           RJM376           RJM377           RJM376           RJM377           RJM378           RJM378           RJM378           RJM378           RJM388           RJM388           RJM384           RJM385           RJM387           RJM388           RJM389	tcaagcactccatgTAACTGCAGGTAAT         TTAgcgtttcggcgccggataa         CCCCAGTTCGcgtTCAGCCTCGG         GATTGCGTCATTTGGATGTCGCTC         AGCCTCGGTTcgtGACCGTGTTA         GAAAGCGAACTGGGGGATTG         TGACCGTGTgcaATCACTTGTTC         CCAACCGAGGCTGAAAGC         CCGTGTTACAttcACTTGTTCAGC         TCACCAACCGAGGGCTGAA         CCCGAAGGTAccgATCTACTTCAC         GCTTACCAttcACTTGTTCAC         GCCTACCTAGGTCTACTTCAC         GCTTACCTGGCTTTTGT         GTTACACTCTGGTCCCAAGTC         GAGGATGTGAAGTAGCGAAG         ACAATACCTGgtGTTCCCTGGA         TGACAGTAGTAAGTAGCGA         CTGGACATTCgatCAAGGTACTAAG         GACAGTAGAAGTATTGTTGAC         GACATTCGGTCAGGTACTAAGG	L11R mutant in bevacizumab scFab G16R mutant in bevacizumab scFab T20A mutant in bevacizumab scFab I21F mutant in bevacizumab scFab L47P mutant in bevacizumab scFab G57S mutant in bevacizumab scFab S92P mutant in bevacizumab scFab
RJM369           RJM370           RJM371           RJM371           RJM373           RJM374           RJM375           RJM376           RJM377           RJM376           RJM377           RJM378           RJM378           RJM378           RJM388           RJM387           RJM387           RJM388	tcaagcactccatgTAACTGCAGGTAAT         TTAgcgtttcggcgccggataa         CCCCAGTTCGcgtTCAGCCTCGG         GATTGCGTCATTTGGATGTCGCTC         AGCCTCGGTCgtGACCGTGTTA         GAAAGCGAACTGGGGGATTG         TGACCGTGTTgcaATCACTTGTC         CCAACCGAGGCTGAAAGC         CCCGAGGTACcgATCACTTGTCAGC         TCAACCAACCGAGGCTGAA         CCCGAAGGTAccgATCTACTTCAC         GCCTTACCTGGCTTTGT         GTTACACTCTGGTCCCAAGTC         GAGCATGTGAAGTAGCAG         TCAACAATACccgACTGTTCCCTG         CAGTAGTAAGTAGCGAAG         ACAATACTCTgctGTTCCCTGGA         TGACAGTAGTAAGTAGCG         CTGGACATTCgatCAAGGTACTAAG         GTTGCAGTAGTAAGTAGCG	L11R mutant in bevacizumab scFab G16R mutant in bevacizumab scFab T20A mutant in bevacizumab scFab I21F mutant in bevacizumab scFab L47P mutant in bevacizumab scFab G57S mutant in bevacizumab scFab S92P mutant in bevacizumab scFab G99D mutant in bevacizumab scFab
RJM369           RJM370           RJM371           RJM371           RJM373           RJM374           RJM375           RJM376           RJM377           RJM376           RJM377           RJM378           RJM378           RJM378           RJM378           RJM388	tcaagcactccatgTAACTGCAGGTAAT         TTAgcgtttcggcgccggataa         CCCCAGTTCGcgtTCAGCCTCGG         GATTGCGTCATTTGGATGTCGCTC         AGCCTCGGTTcgtGACCGTGTTA         GAAAGCGAACTGGGGGATTG         TGACCGTGTgcaATCACTTGTTC         CCAACCGAGGCTGAAAGC         CCGTGTTACAttcACTTGTTCAGC         TCACCAACCGAGGGCTGAA         CCCGAAGGTAccgATCTACTTCAC         GCTTACCAttcACTTGTTCAC         GCCTACCTAGGTCTACTTCAC         GCTTACCTGGCTTTTGT         GTTACACTCTGGTCCCAAGTC         GAGGATGTGAAGTAGCGAAG         ACAATACCTGgtGTTCCCTGGA         TGACAGTAGTAAGTAGCGA         CTGGACATTCgatCAAGGTACTAAG         GACAGTAGAAGTATTGTTGAC         GACATTCGGTCAGGTACTAAGG	L11R mutant in bevacizumab scFab G16R mutant in bevacizumab scFab T20A mutant in bevacizumab scFab I21F mutant in bevacizumab scFab L47P mutant in bevacizumab scFab G57S mutant in bevacizumab scFab S92P mutant in bevacizumab scFab G99D mutant in bevacizumab scFab

Code	Sequence	Use
RJM393 RJM394	GTTCAATCGTagtGAGTGTGGTG GACTTCGTGACAGGAGATG	G212S mutant in bevacizumab scFab
RJM395 RJM396	AGCCCCAGGAatgGGACTTGAGT TGACGAACCCAATTCATACCG	K317M mutant in bevacizumab scFab
RJM397 RJM398	AGGTACGTTAgatACTGTTAGTAGTG TGACCCCAAACGTCGAAG	V393D mutant in bevacizumab scFab
RJM399 RJM400	CCCTTCATCAaacTCTACAAGTGGC GCAAGAGGGAAGACACTTG	K413N mutant in bevacizumab scFab
RJM401 RJM402	GAAGGTTGAGaccAAGAGTTGTG TTGTCTACCTTCGTATTGC	P497T mutant in bevacizumab scFab
RJM485 RJM486	aggtcaggatccgggtctcGGAGCGGTTCCGGAAGCG taggttagctcgaggtctcGCCACCACCAGCAACC	Create pSNAC blaGGSTOP-VHH
акм3 акм4	CCGTCAAGCCctcGGCCAAGGAC ACCCAGTGCATATAGTGTCCGGTGAAC	Create bla AMS134 P41L
акм5 акм6	AGGATACTTTcgcTATTTCGACTTATGGGGTCGCGG GGGCGACCAGGGTCACGA	Create bla AMS134 W107R
акм7 акм8	ATACTTTTGGcatTTCGACTTATGGGGTCGC CCTGGGCGACCAGGGTCA	Create bla AMS134 Y108H
акм9 акм10	CTTTTGGTATtccGACTTATGGGGTCG TATCCTGGGCGACCAGGG	Create bla AMS134 F109S
AKM11 AKM12	CGACTTATGGgatCGCGGCACCA AAATACCAAAAGTATCCTGGGCGACC	Create bla AMS134 G113D
AKM13 AKM14	ATGGGGTCGCgatACCATGGTCA AAGTCGAAATACCAAAAGTATCCTG	Create bla AMS134 G115D
AKM15 AKM16	TCGCGGCACCaaaGTCATCGTCT CCCCATAAGTCGAAATACC	Create bla AMS134 M117K
AKM17 AKM18	CGGCACCATGgatATCGTCTCTT CGACCCCATAAGTCGAAATAC	Create bla AMS134 V118D
AKM19 AKM20	CACCATGGTCaccGTCTCTTCTG CCGCGACCCCATAAGTCG	Create bla AMS134 I119T
AKM21 AKM22	GAAAATCAGCggtGTAGAGGCGG AGCGTGAAATCGGTGCCT	Create bla AMS134 R220G
AKM23 AKM24	ATCAGGATTCgcaTTCAGTAATTACGACATGGCGTG GCCGCGCACGACAGACGC	Create bla AMS197 T28A
AKM25 AKM26	CAGTAATTACaacATGGCGTGGG AATGTGAATCCTGATGCC	Create bla AMS197 D33N
акм27 акм28	CCGTGGTACCcatGTGACAGTGT CCCCAATACACGAACGGAATG	Create bla AMS197 L114H
AKM29 AKM30	CTATAAGGCTaatCGTCTTCAATCAGGG ATGAGCAACTCGGGCGCC	Create bla AMS197 S187N

#### Table A.2 Primers used in golden gate cloning

	Sequence	Use
		Add Fwd BsaI site to blaMBP epPCR
RJM033	GGGAATGGTCTCGGTGGCTCGAGC	product to clone into blaGGSTOP
		Add Rev BsaI site to blaMBP epPCR
RJM034	ACATGCGGTCTCCGGCTCCCGGATCC	product to clone into blaGGSTOP
RJM086	CGAAGTGGTCTCGaaTCTAGAGTCGACCTGCAG	
RJM087	CGAAGTGGTCTCGATgaattcctcctggtaccg	
RJM088	CGAAGTGGTCTCCtcATGAAAAAGATTTGGCTGGC	
RJM089	CGAAGTGGTCTCCtCGCCGATGCGCTAAACGC	
RJM090	CGAAGTGGTCTCCGCGatggtgagcaagggtgag	
RJM090	CGAAGTGGTCTCCCGAttacatcatatcggtaaaggcc	
RJM092	GTGGAAGGTCTCCaccTAATCTAGAGTCGACCTGCA	
RJM093	GTGGAAGGTCTCCTgaattcctcctggtaccg	
RJM094	GTGGAAGGTCTCGttcATGAAAAAGATTTGGCTGGCG	
RJM095	GTGGAAGGTCTCGcatCGCCGATGCGCTAAACGC	
RJM096	GTGGAAGGTCTCGGatgqagqaqqacaacatq	Create control vector with split
RJM097		-
	GTGGAAGGTCTCGtcgtcctcgttgtggctggt	sfCherry2
RJM098	GTGGAAGGTCTCGacgacgttggtggtggcgga	
RJM099	GTGGAAGGTCTCGatagaccccccgccagcgct	
RJM100	GTGGAAGGTCTCGctatgtacaccatcgtggag	
RJM101	GTGGAAGGTCTCGAggtgctgtgtctggcctc	
RJM102	TCCGATGGTCTCGATGTTTATTTTTCTAAATACATGCGGC	
RJM103	TCCGATGGTCTCGaCCAATGCTTAATCAGTGAGG	
RJM104	TCCGATGGTCTCGTGGtcgggaagggaaggttct	
RJM105	TCCGATGGTCTCGgattacatcatatcggtaaaggcc	
RJM106	TCCGATGGTCTCGaatctacaaataattttgtttaacttttctag	
RJM107	TCCGATGGTCTCGTCCttatttgtatagttcatccatgcc	
RJM108	TCCGATGGTCTCGaGGAGTCTCCCCATGCGAG	
RJM109	TCCGATGGTCTCGcttAAATAAACAAAAGAGTTTGTAGAAACGC	
RJM110	TCCGATGGTCTCCTaagaaaccaattgtccatattgc	
RJM111	TCCGATGGTCTCCTgaattcctcctggtaccg	
RJM112	TCCGATGGTCTCCttcATGAAAAAGATTTGGCTGGCG	
RJM113	TCCGATGGTCTCCataccagggcccccgctacc	
RJM114	TCCGATGGTCTCCgtatggtgagcaagggtgag	Create pSNAC with mNeonGreen2
RJM115	TCCGATGGTCTCCAGTTAcatggagtgcttgagctc	create power with anotherenz
RJM116	TCCGATGGTCTCGAACTGCAGGTAATTAAATAAGCTTCA	
RJM117	TCCGATGGTCTCGACATTTGTCCTACTCAGGAGAGC	
RJM118	TTCGTCGGTCTCCgacTAACTGCAGGTAATTAAATAAGCTTCAAATAAAACGAAAGGCT	
RJM119	TTCGTCGGTCTCCtagaccccccgccagcgct	
RJM120	TTCGTCGGTCTCCtctatgtacaccatcgtggag	
RJM121	TTCGTCGGTCTCCgTTAggtgctgtgtctggcctc	
RJM122	TTCGTCGGTCTCCTAAcctctagaaataattttgtttaactttaag	
RJM123	TTCGTCGGTCTCCCCtttatacagttcatccataccgt	Create pSNAC with sfCherry2
RJM124	TTCGTCGGTCTCGaaGGAGTCTCCCCATGCGAG	
RJM125		
	TTCGTCGGTCTCGataccagggcccccgctacc	
RJM126	TTCGTCGGTCTCGgtatggaggaggacaacatg	
RJM127	TTCGTCGGTCTCGAgtectcgttgtggetggt	
RJM130	GATGACGGTCTCGaaTCTAGAGTCGACCTGCAG	
RJM131	GATGACGGTCTCGatAGCAAAAACAGGAAGGCA	Create control vector with mNG2 - usin
RJM132	GATGACGGTCTCGCTatggtgagcaagggtgag	bla signal sequence
RJM133	GATGACGGTCTCGGAttacatcatatcggtaaaggcc	
RJM225	AGCAGAGGTCTCgagtctggaaaaacacagct	Amplify pSNAC delta mScarlet-I to
RJM226	AGCAGAGGTCTCGggaattcctcctggtaccg	replace DsbA signal sequence
		Amplify pold signal seguence to poplar
RJM227 RJM228	GATGCAGGTCTCCttccatgaaatacctgctgccgaccgc GATGCAGGTCTCCCgactggccatcgccggctgggc	Amplify pelB signal sequence to replace DsbA in pSNAC delta mScarlet-I
RJM228	GATGCAGGTCTCCgactggccatcgccggctgggc	DsbA in pSNAC delta mScarlet-I
RJM228 RJM229	GATGCAGGTCTCCgactggccatcgccggctgggc	DsbA in pSNAC delta mScarlet-I Amplify Tora signal sequence to replace
RJM228 RJM229	GATGCAGGTCTCCgactggccatcgccggctgggc	DsbA in pSNAC delta mScarlet-I
RJM228 RJM229 RJM230	GATGCAGGTCTCCgactggccatcgccggctgggc TGTCACGGTCTCCttccatgaacaacaacgatctgtttcaggcgagccgcc TGTCACGGTCTCCgactcgcggtcgcgcgggg	DsbA in pSNAC delta mScarlet-I Amplify Tora signal sequence to replace DsbA in pSNAC delta mScarlet-I
RJM228 RJM229 RJM230 RJM231	GATGCAGGTCTCCgactggccatcgccggctgggc TGTCACGGTCTCCttccatgaacaacaacgatctgtttcaggcgagccgcc TGTCACGGTCTCCgactcgcggtcgcgcggggg CGACAAGGTCTCCGttccatgaaaattaaaaccggcgggg	DsbA in pSNAC delta mScarlet-I Amplify Tora signal sequence to replace DsbA in pSNAC delta mScarlet-I Amplify malE signal sequence to replace
	GATGCAGGTCTCCgactggccatcgccggctgggc TGTCACGGTCTCCttccatgaacaacaacgatctgtttcaggcgagccgcc TGTCACGGTCTCCgactcgcggtcgcgcgggg	DsbA in pSNAC delta mScarlet-I Amplify Tora signal sequence to replace DsbA in pSNAC delta mScarlet-I
RJM228 RJM229 RJM230 RJM231 RJM232	GATGCAGGTCTCCgactggccatcgccggctgggc TGTCACGGTCTCCttccatgaacaacaacgatctgtttcaggcgagccgcc TGTCACGGTCTCCgactcgcggtcgcggcgggg CGACAAGGTCTCGttccatgaaaattaaaaccggcgcggg CGACAAGGTCTCGgactggccagcgcgctagcgct	DsbA in pSNAC delta mScarlet-I Amplify Tora signal sequence to replace DsbA in pSNAC delta mScarlet-I Amplify malE signal sequence to replace DsbA in pSNAC delta mScarlet-I
RJM228 RJM229 RJM230 RJM231 RJM232 RJM233	GATGCAGGTCTCCgactggccatcgccggctgggc TGTCACGGTCTCCttccatgaacaacaacgatctgtttcaggcgagccgcc TGTCACGGTCTCCgactcgcggtcgcgcgggg CGACAAGGTCTCCgtccatgaaaattaaaaccggcggcgg CGACAAGGTCTCCgactggccagcggcgtagcgct CAGTGAGGTCTCCttccatgaaaaaaccgcgattgcgattgcgg	DsbA in pSNAC delta mScarlet-I Amplify Tora signal sequence to replace DsbA in pSNAC delta mScarlet-I Amplify malE signal sequence to replace DsbA in pSNAC delta mScarlet-I Amplify ompA signal sequence to replace
RJM228 RJM229 RJM230 RJM231 RJM232 RJM233	GATGCAGGTCTCCgactggccatcgccggctgggc TGTCACGGTCTCCttccatgaacaacaacgatctgtttcaggcgagccgcc TGTCACGGTCTCCgactcgcggtcgcggcgggg CGACAAGGTCTCGttccatgaaaattaaaaccggcgcggg CGACAAGGTCTCGgactggccagcgcgctagcgct	DsbA in pSNAC delta mScarlet-I Amplify Tora signal sequence to replace DsbA in pSNAC delta mScarlet-I Amplify malE signal sequence to replace DsbA in pSNAC delta mScarlet-I
RJM228 RJM229 RJM230 RJM231 RJM232 RJM233 RJM233	GATGCAGGTCTCCgactggccatcgccggctgggc TGTCACGGTCTCCttccatgaacaacaacgatctgtttcaggcgagccgcc TGTCACGGTCTCCgactgcggtcgcggcggggg CGACAAGGTCTCCGttccatgaaaattaaaaccggcgcggg CGACAAGGTCTCCGgactggccagggctagcgct CAGTGAGGTCTCCttccatgaaaaaaaccgcgattgcgattgcgg CAGTGAGGTCTCCgactggcctgcgccacggtcgc	DsbA in pSNAC delta mScarlet-I Amplify Tora signal sequence to replac DsbA in pSNAC delta mScarlet-I Amplify malE signal sequence to replac DsbA in pSNAC delta mScarlet-I Amplify ompA signal sequence to replac DsbA in pSNAC delta mScarlet-I
RJM228 RJM229 RJM230 RJM231 RJM232 RJM233 RJM233 RJM235	GATGCAGGTCTCCgactggccatcgccggctgggc         TGTCACGGTCTCCttccatgaacaacaacgatctgtttcaggcgagccgcc         TGTCACGGTCTCCgactgcggtcgcgcggcggg         CGACAAGGTCTCCGttccatgaacattaacacggcgcgcg         CGACAAGGTCTCCGgccggccggcgcggg         CGACGAGGTCTCCGtccatgaacattaacacgggcgggg         CGACAAGGTCTCCGgccggccggcgggg         CGACGAGGTCTCCGtccatgaacaacaacgggcgggg         CAGTGAGGTCTCCttccatgaacaacacgggtggggggg         CAGTGAGGTCTCCttccatgaacaacacgggtggg         CAGTGAGGTCTCCgactggcctgggccacggtcgc         TCGAGTGGTCTCGttccatGaGTATTCAACATTTCCGTG	DsbA in pSNAC delta mScarlet-I         Amplify Tora signal sequence to replac         DsbA in pSNAC delta mScarlet-I         Amplify malE signal sequence to replac         DsbA in pSNAC delta mScarlet-I         Amplify ompA signal sequence to replac         DsbA in pSNAC delta mScarlet-I         Amplify blactamase signal sequence to replac
RJM228 RJM229 RJM230 RJM231 RJM232	GATGCAGGTCTCCgactggccatcgccggctgggc TGTCACGGTCTCCttccatgaacaacaacgatctgtttcaggcgagccgcc TGTCACGGTCTCCgactgcggtcgcggcggggg CGACAAGGTCTCCGttccatgaaaattaaaaccggcgcggg CGACAAGGTCTCCGgactggccagggctagcgct CAGTGAGGTCTCCttccatgaaaaaaaccgcgattgcgattgcgg CAGTGAGGTCTCCgactggcctgcgccacggtcgc	DsbA in pSNAC delta mScarlet-I Amplify Tora signal sequence to replace DsbA in pSNAC delta mScarlet-I Amplify malE signal sequence to replace DsbA in pSNAC delta mScarlet-I Amplify ompA signal sequence to replace DsbA in pSNAC delta mScarlet-I
RJM228 RJM229 RJM230 RJM231 RJM232 RJM233 RJM233 RJM235	GATGCAGGTCTCCgactggccatcgccggctgggc         TGTCACGGTCTCCttccatgaacaacaacgatctgtttcaggcgagccgcc         TGTCACGGTCTCCgactgcggtcgcgcggcggg         CGACAAGGTCTCCGttccatgaacattaacacggcgcgcg         CGACAAGGTCTCCGgccggccggcgcggg         CGACGAGGTCTCCGtccatgaacattaacacgggcgggg         CGACAAGGTCTCCGgccggccggcgggg         CGACGAGGTCTCCGtccatgaacaacaacgggcgggg         CAGTGAGGTCTCCttccatgaacaacacgggtggggggg         CAGTGAGGTCTCCttccatgaacaacacgggtggg         CAGTGAGGTCTCCgactggcctgggccacggtcgc         TCGAGTGGTCTCGttccatGaGTATTCAACATTTCCGTG	DsbA in pSNAC delta mScarlet-I         Amplify Tora signal sequence to replace         DsbA in pSNAC delta mScarlet-I         Amplify malE signal sequence to replace         DsbA in pSNAC delta mScarlet-I         Amplify ompA signal sequence to replace         DsbA in pSNAC delta mScarlet-I         Amplify bplactamase signal sequence to replace

Code	Sequence	Use
RJM325	ACGTCAGGTCTCGatagaccccccgccagcgct	
RJM326	ACGTCAGGTCTCCctatgtacaccatcgtggagc	
RJM327	ACGTCAGGTCTCCcTTAggtgctgtgtctggcctc	
RJM328	ACGTCAGGTCTCCTAAgaaataattttgtttaactttaagaagga	
RJM329	ACGTCAGGTCTCCgaTTAtttatacagttcatccataccgt	
RJM330	ACGTCAGGTCTCGAAtcccgccattcagagaag	
RJM331	ACGTCAGGTCTCGGCATGTATTTAGAAAAATAAACATTTGTCC	
RJM332	GGCTTAGGTCTCCACAAATGTTTATTTTTCTAAATACATGCGGCCGCTCATG	
RJM333	GGCTTAGGTCTCCagacccccccccccccccccccccccccccccccc	
RJM334	GGCTTAGGTCTCGgtctatgtacaccatcgtggagc	
RJM335	GGCTTAGGTCTCGTAggtgctgtgtctggcctc	
RJM336	GGCTTAGGTCTCCccTAAcctctagaaataattttgtttaactttaag	
RJM337	GGCTTAGGTCTCCagtTTAtttatacagttcatccataccgt	
RJM338	GGCTTAGGTCTCCAacttttcatactcccgccattcaga	
RJM339	GGCTTAGGTCTCCtcaccagggcccccgctacc	
RJM340	GGCTTAGGTCTCGgtgaggaggacaacatggcc	
RJM341	GGCTTAGGTCTCGGCAGGTTAgtcctcgttgtggctggt	
RJM342	GGCTTAGGTCTCGCTGCAGGTAATTAAATAAGCTTCA	
RJM343	GGCTTAGGTCTCGTTGTCCTACTCAGGAGAGC	
RJM439	GAACACGGTCTCGTAAGTCTCCCCATGCGAGAGT	
RJM440	GAACACGGTCTCGGAACCGCCACTTCCGCCTGATCCACCCCAATGCTTAATCAGTGAGG	
RJM441	GAACACGGTCTCGGTTCAGGTGGGAGTGGTGGCAGCgaagttcaactgcaagctt	Create pSNAC HA4 VHH-caffeine
RJM442	GAACACGGTCTCGCTTATTAacgaggttccagaggatc	
RJM443	TGTGGAGGTCTCGgctaaactttatctgagaatagtcaatcttcg	
RJM444	TGTGGAGGTCTCGctgccacctggacccaacag	
RJM445	TGTGGAGGTCTCGqcaqtctqqaaaaacacaqctqqt	Create CadC-SH2
RJM446	TGTGGAGGTCTCGtagcgtttcggcgccggata	
RJM481	GGTACTGGTCTCCCTAAtaaactttatctgagaatagtcaatcttcg	
RJM482	GGTACTGGTCTCCTTgccacctggacccaacag	Amplify pHJ12 CadC-SH2 to make CadC-MH
RJM483	GGTACTGGTCTCGgcAAAATCGAAGAAGGTAAACTGG	Amplify wild type MBP from RM11 blaMB
RJM484	GGTACTGGTCTCGTTAGGATCCCTTGGTGATACG	no BsaI to make CadC-MBP

#### Table A.3 Primers used in sequencing

Code	Sequence	Sequencing use
Fullb-lac-F Fullb-lac-R	TAGCGGATCATACCTGACG CGCTTCTGCGTTCTGAT	Full b-lactamase
b-lac-linker-F b-lac-linker-R	CGGAGCTGAATGAAGCCATACC TCACCGGCTCCAGATTTATCAGC	Between 28-GS linker
RJM083	aagaaaccaattgtccatat	From pBAD promoter
RJM150 RJM151 RJM152 RJM153	GACCACTTCTGCGCTCGGCC agacgatgggttgggaggcg ACAAACTCTTTTGTTTATTT aacatggcctctctcccagc	Walking primers for pSNAC splitFP
RJM154	cgatctcgatcccgcgaaat	pET sfCh2
RJM159	GCCAGCCAAATCTTTTCAT	pSNAC from DsbA
RJM160	aggtcacccttggtggactt	pSNAC sequencing primer reverse
RJM163	ACGCTACTCACAGAGCTGCC	Reverse sequencing of domain 1 bla in pSNAC
RJM164	tccttgaagttgagctcggt	Reverse sequencing of domain 2 bla in pSNAC
RJM237	caggtattcggccgcgttac	pSNAC sequencing primer reverse SH2 for signal sequence
RJM285	tctgatccgccaccaccaac	pSNAC reverse sequencing primer SH2-mNG2 16aa link
RJM302	atatcccggactattata	mScarlet-I
RJM352 RJM353 RJM354	ataatttcaacagccataat ATGAAAAGATTTGGCTGGC acggcggcgtggtgaccgtg	Walking primers for pSNAC-VHH
RJM453	ccccagttaaaagcaaacga	Sequencing primer binding at C-term of CadC
RJM454	aacaagatactgagcacagc	Sequencing primer binding upstream of lac operator

Table A.4 Primers used to amplify libraries for NGS	
---	--

Code	Sequence	Use
b-lac-linker-F b-lac-linker-R	CGGAGCTGAATGAAGCCATACC TCACCGGCTCCAGATTTATCAGC	Amplify between 28-GS linker for Illumina and Pacbio sequencing
RJM493 RJM494	ACACTCTTTCCCTACACGACGCTCTTCCGATCTTCTAGTAGCTCTGTGTATTAC GACTGGAGTTCAGACGTGTGCTCTTCCGATCTACTAATCGGAGAATACATGAA	Amplify HA4 WT and variants for EZ-amplicon

**Appendix B** 

### First derivatives of DSF and SLS data for AMSCI mAbs

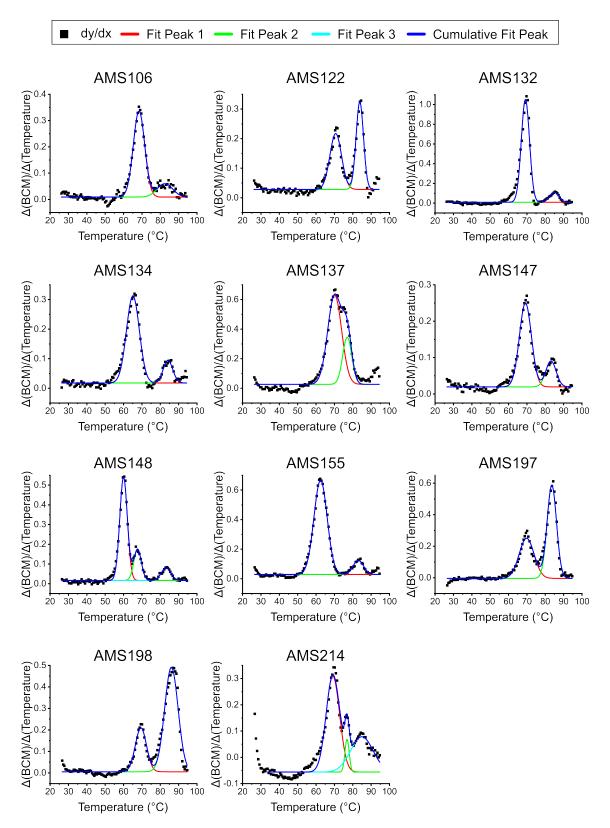


Fig. B.1 Thermal stability of 11 AMSCI mAbs. Differential scanning fluorimetry (DSF) measurements of the full-length AMSCI mAbs. Protein unfolding was measured using intrinsic protein fluorescence. Data is presented as first derivatives of barycentric mean (BCM) versus temperature (°C). The transition mid-point temperatures (Tm) were calculated using the first derivative of the fluorescence raw data. This was fitted to one or more gaussians using Origin Pro 2020 version 9.7.0.118. calculated by differential scanning fluorimetry (DSF). Protein unfolding was measured using intrinsic protein fluorescence. The barycentric mean (BCM) of the fluorescence intensity curves from 315-430 nm was used to plot the Tm curves (Section 2.4.8). Data collected using UNcle system (Unchained Labs).

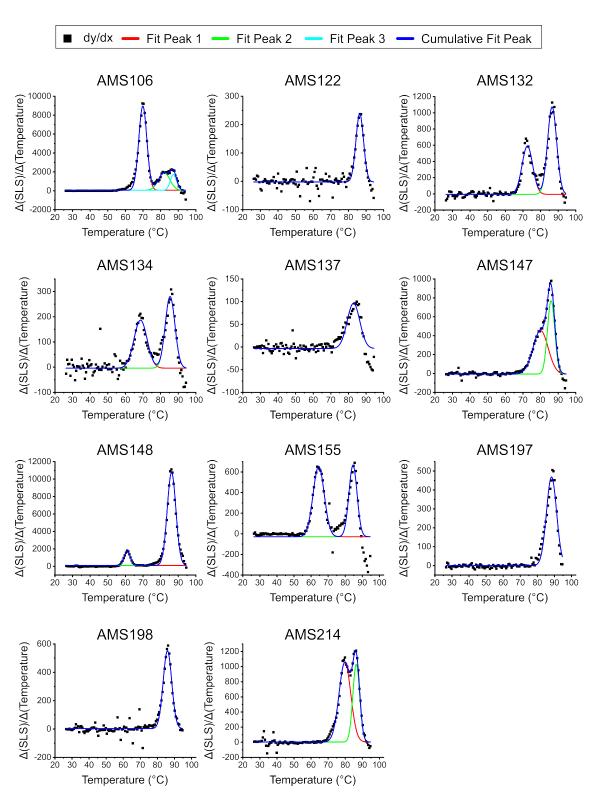


Fig. B.2 Biophysical characterisation of 11 AMSCI mAbs. Static light scattering (SLS) was measured at each temperature throughout the differential scanning fluorimetry (DSF) experiment. Data is presented as first derivatives of SLS versus temperature (°C). The temperature onset of aggregation (Tonset) was calculated from the first derivative of the SLS raw data. This was fitted to one or more gaussians using Origin Pro 2020 version 9.7.0.118. The fit from the first gaussian was normalised between 0 and 1, and the Tonset was defined as the point at which the slope (first derivative) exceeded 0.1 % of the peak value of the first derivative. Data collected using UNcle system (Unchained Labs).

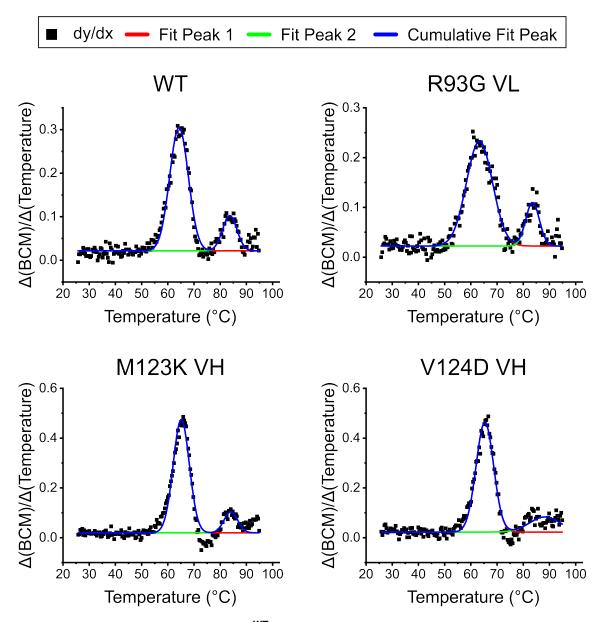


Fig. B.3 **Thermal stability of AMS134<sup>WT</sup> and evolved variants.** Differential scanning fluorimetry (DSF) measurements of the full-length AMSCI mAbs. Protein unfolding was measured using intrinsic protein fluorescence. Data is presented as first derivatives of barycentric mean (BCM) versus temperature (°C). The transition mid-point temperatures (Tm) were calculated using the first derivative of the fluorescence raw data. This was fitted to one or more gaussians using Origin Pro 2020 version 9.7.0.118. calculated by differential scanning fluorimetry (DSF). Protein unfolding was measured using intrinsic protein fluorescence. The barycentric mean (BCM) of the fluorescence intensity curves from 315-430 nm was used to plot the Tm curves (Section 2.4.8). Data collected using UNcle system (Unchained Labs).

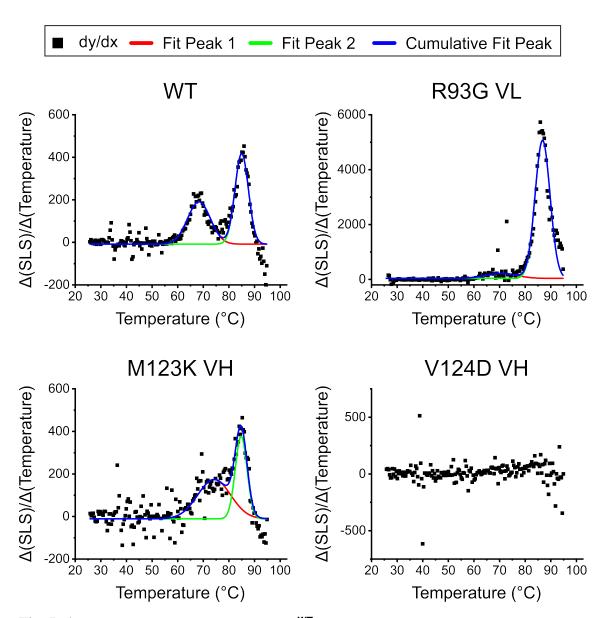


Fig. B.4 **Aggregation behaviour of AMS134<sup>WT</sup> and evolved variants.** Static light scattering (SLS) was measured at each temperature throughout the differential scanning fluorimetry (DSF) experiment. Data is presented as first derivatives of SLS versus temperature (°C). The temperature onset of aggregation (Tonset) was calculated from the first derivative of the SLS raw data. This was fitted to one or more gaussians using Origin Pro 2020 version 9.7.0.118. The fit from the first gaussian was normalised between 0 and 1, and the Tonset was defined as the point at which the slope (first derivative) exceeded 0.1 % of the peak value of the first derivative. Data collected using UNcle system (Unchained Labs).

**Appendix C** 

# Multiple regression model statistics for AMSCI mAbs

Table C.1 Linear regression statistics for AMSCI mAbs regression model with 5 parameters. Includes: theoretical pI, Tonset by DLS, Tonset by SLS, Tagg by SLS, and Camsol score. f(5, 5) = 5.142,  $R^2 = 0.787$ , r = 0.887, p = 0.04829.

	<b>Estimate</b> (β)	Standard error	t value	p value
(Intercept)	1490.76	835.41	1.784	0.1344
Theoretical pI	-398.92	143.17	-2.786	0.0386
TOnset by DLS	50.33	21.07	2.388	0.0625
TOnset by SLS	-54.75	34.54	-1.585	0.1738
Tagg by SLS	36.2	29.11	1.244	0.2688
Camsol score	476.7	235.46	2.025	0.0988

Table C.2 Linear regression statistics for AMSCI mAbs regression model with 4 parameters. Includes: theoretical pI, Tonset by DLS, Tonset by SLS, and Camsol score. f(4, 6) = 5.537,  $R^2 = 0.837$ , r = 0.915, p = 0.03255.

	<b>Estimate</b> (β)	Standard error	t value	p value
(Intercept)	1093.66	806.38	1.356	0.2238
Theoretical pI	-282.79	113.37	-2.494	0.0469
TOnset by DLS	45.8	21.68	2.112	0.0791
TOnset by SLS	-18	18.69	-0.963	0.3727
Camsol score	512.47	244.11	2.099	0.0805

Table C.3 Linear regression statistics for AMSCI mAbs regression model with 3 parameters. Includes: Tonset DLS, Camsol score, and theoretical pI. f(3, 7) = 7.147,  $R^2 = 0.754$ , r = 0.868, p = 0.01548.

	<b>Estimate</b> (β)	Standard error	t value	p value
(Intercept)	756.443	722.622	1.047	0.32998
Theoretical pI	-226.442	96.598	-2.344	0.05153
TOnset by DLS	25.882	6.475	3.997	0.00521
Camsol score	493.901	242.082	2.04	0.08069

Table C.4 Linear regression statistics for AMSCI mAbs regression model with 2 parameters. Includes: Tonset DLS and theoretical pI. f(2, 8) = 6.192,  $R^2 = 0.608$ , r = 0.779, p = 0.0237.

	<b>Estimate</b> (β)	Standard error	t value	p value
(Intercept)	1495.59	738.56	2.025	0.0775
Theoretical pI	-311.62	102.9	-3.028	0.0163
TOnset by DLS	24.24	7.59	3.194	0.0127

**Appendix D** 

## Multiple regression model statistics for Jain mAbs

Table D.1 Linear regression statistics for Jain mAbs regression model with 7 parameters. The model includes: Fab Tm by DSF (TM), Standup Monolayer Adsorption Chromatography (SMAC), Accelerated Stability (AS), Polyspecificity Reagent (PSR) binding, Cross Interaction Chromatography (CIC), theoretical pI (PI), and scFv molecular weight (MW). f(7, 27) = 3.968,  $R^2 = 0.51$ , r = 0.71, p = 0.004.

	<b>Estimate</b> (β)	Standard error	t value	p value
(Intercept)	2402.421	1.66E+03	1.446463	0.159557
TM	19.57136	5.29E+00	3.702668	0.000967
SMAC	-44.9978	2.35E+01	-1.91229	0.066503
AS	783.8614	4.69E+02	1.671366	0.106202
PSR	-362.74	1.66E+02	-2.17947	0.038193
CIC	122.7285	4.57E+01	2.687265	0.012181
PI	-122.172	3.44E+01	-3.54679	0.001448
MW	-0.11543	6.13E-02	-1.88315	0.070495

Table D.2 Linear regression statistics for Jain mAbs regression model with 6 parameters. The model includes: Fab Tm by DSF (TM), Standup Monolayer Adsorption Chromatography (SMAC), Polyspecificity Reagent (PSR) binding, Cross Interaction Chromatography (CIC), theoretical pI (PI), and scFv molecular weight (MW). f(6, 28) = 3.913,  $R^2 = 0.46$ , r = 0.68, p = 0.005.

	<b>Estimate</b> (β)	Standard error	t value	p value
(Intercept)	2075.572	1701.342238	1.219961	0.232656
TM	18.09389	5.3756093	3.365923	0.00223
SMAC	-50.2372	24.0563486	-2.08831	0.045984
PSR	-218.105	146.6511444	-1.48724	0.148128
CIC	115.6893	46.9095995	2.466219	0.020042
PI	-135.806	34.5208732	-3.93403	0.000501
MW	-0.08861	0.0610233	-1.45205	0.157602

Table D.3 Linear regression statistics for Jain mAbs regression model with 5 parameters. The model includes: Fab Tm by DSF (TM), Standup Monolayer Adsorption Chromatography (SMAC), Polyspecificity Reagent (PSR) binding, Cross Interaction Chromatography (CIC) and theoretical pI (PI). f(5, 29) = 4.116,  $R^2 = 0.42$ , r = 0.64, p = 0.006.

	<b>Estimate</b> (β)	Standard error	t value	p value
(Intercept)	-268.19	547.979393	-0.48942	0.628228
ТМ	17.47368	5.460062	3.200271	0.003316
SMAC	-61.6977	23.154955	-2.66456	0.01246
PSR	-263.781	145.949603	-1.80734	0.081094
CIC	134.404	45.958224	2.924482	0.006633
PI	-132.775	35.110052	-3.78168	0.000721

Table D.4 Linear regression statistics for 29 Jain mAbs regression model with 5 parameters. The model includes: Fab Tm by DSF (TM), Standup Monolayer Adsorption Chromatography (SMAC), Polyspecificity Reagent (PSR) binding, Cross Interaction Chromatography (CIC) and theoretical pI (PI). f(5, 23) = 3.375,  $R^2 = 0.42$ , r = 0.64, p = 0.01971.

	<b>Estimate</b> (β)	Standard error	t value	p value
(Intercept)	-158.507	725.375	-0.462	0.64872
ТМ	16.164	6.855	2.805	0.01032
SMAC	-65.309	27.399	-2.573	0.01733
PSR	-288.447	186.172	-1.739	0.096
CIC	142.548	51.558	2.901	0.00828
PI	-137.56	45.525	-3.131	0.00486