Adapting nanobody scaffolds as protein chaperones for effective crystallization and good protein behaviour in cryo-EM



Mingda Ye

Nuffield Department of Clinical Medicine

Centre for Medicines Discovery

Worcester College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Michaelmas 2021/2022

ABSTRACT

In the past three decades, structural biology has developed tremendously with the establishment of the X-ray crystallography and cryo-EM techniques. In order to satisfy the needs of those two techniques, methods have been developed to engineer the proteins of interest. One of the popular methods is binder-assisted structure solution. Among the various kinds of protein binders, nanobodies are one of the most frequently used tool in both X-ray crystallography and cryo-EM. Nanobodies are a small single-domain antibody (VHH) of around 15 kDa, and consist of the complementarity determining regions (CDRs) and the scaffold (non-CDRs). The residues in the nanobody scaffold are relatively constant and the CDRs determine the specificity for the target protein. Until now, there have been hundreds of successful cases of nanobody-assisted structure solution, including both soluble proteins and membrane proteins. However, in crystallography, there has never been reports about the success rate of nanobodies as a crystal chaperone, and the nanobody scaffolds have never been optimized for crystallization purposes. Moreover, in cryo-EM, there was no convenient imaging scaffolds to enable structure solution of small proteins. In this study, I set out to address those two problems.

A few target:nanobody pairs were first selected and evaluated the success rate of crystallization. It turned out that only one pair (MAGEB1:nanobody) successfully crystalized out of the 17 pairs. In order to further demonstrate that the nanobody scaffold was optimizable, I selected one of the failed pairs (RECQL5:nanobody_G0-001) and carried out surface mutagenesis on the nanobody, which successfully led to a 2.8 Å structure. This demonstrated that the nanobody scaffold had large potential to be optimized for crystallization purposes.

Thorough bioinformatics research was then carried out on nanboody-nanobody crystal contacts, in order to find the most suitable surface of the nanobody scaffold for engineering. After the unsupervised density based clustering (DBSCAN), the largest class of nanobody-nanobody crystal contacts was discovered. This class of interfaces was named as ‘Nb-α’. This finding provided guidance on where to start with the scaffold engineering.

Extensive engineering was carried out around the Nb-α following bioinformatics research. 5 iterations of mutagenesis were performed to make the RECQL5:nanobody complex consistently well-diffracting, which

identified 4 key mutations along with the truncation of the C-terminal tail. Those modifications led to the robust dimerization and tetramerization of nanobodies in the crystal lattice, giving the name ‘Gluebody’ to the optimized scaffold. However, the Gluebody scaffold did not show significant improvement for other target proteins, and therefore it did not solve all problems in crystallization.

The tetrameric arrangement of the Gluebodies had two rigid interfaces perpendicular to each other, which if in solution did not dissociate, would make the tetrameric-Gluebody an imaging scaffold with D2 symmetry for cryo-EM. Initial trials of disulfide linked Gluebodies were taken on two target proteins, which led to the 2D classification results as good as the target protein alone when the thesis was written. The work is still on-going to invent a new tool for cryo-EM structure solution.

In conclusion, the discovery process of Gluebodies is the first systematic study on the nanobody scaffold optimization for crystallization purposes. For the first time, the naïve nanobody scaffold is demonstrated to be optimizable for crystallization purposes, and the Gluebody mutations are designed based on bioinformatics research, which are demonstrated to improve the crystal reliability and diversity of packing forms of RECQL5. Meanwhile, the tetrameric-Gluebody assembly is discovered, based on which a symmetric imaging scaffold for cryo-EM could be invented. On the other hand, the Gluebody mutations could not solve all problems in crystallization as it did not show significant assistance for other target proteins. The combination of other approaches would be necessary depending on the target protein.

DECLARATION OF WORK

I hereby declare that this thesis is my own work and effort except in collaborative project, and that it has not been submitted anywhere for any award. The findings from collaborative described here are clearly stated prior to the chapter. Where other sources of information have been used, they have been acknowledged.

Signature



MINGDA YE (printed)

ACKNOWLEDGEMENT

I hereby express great thanks to my supervisors - Prof. Frank von Delft and Dr. Katharina Duerr – for their high-level guidance throughout my DPhil studies. I would also like to thank Michael Fairhead for his experimental guidance. I also want to give thanks to all members of the PX group, Tobias Krojer, Gustavo A. Bezerra, Beth MacLean, Nathan Wright, Andrew Thompson, Lizbe Koekemoer, Elliot Nelson, Janine Gray, Mpho Makola, Yuliya Dubianok, Conor Wild, Mihaela Smilova, Harold Grosjean, Jack Scantlebury, Romain Talon, Ritika Sethi, Nick Pearce, and all members in the Diamond I04-1 group, Alice Douangamath, Daren Fearon, Patrick Collins, Jose Brandao-Neto, Romain Talon, Anthony Aimon, Rachael Skyner, Aisla Powell, Renjie Zhang. They also provided a lot of help throughout my study. Last but not least, I would like to give thanks to my collaborators, Prof. Opher Gileadi, Prof. Robert Gilbert, Prof. Ben Davis, Rod Chalk, Joseph Newman, Ashley Pike, Gamma Chi, Gangshun Yi, Dimitrios Mammalis, Huanyu Li and Victor Lopes Rangel. They helped me with protein samples, cryo-EM and knowledge in chemistry.

LIST OF ABBREVIATIONS

AD	Activation domain
ADP	Adenosine diphosphate
API	Application programming interface
ATP	Adenosine triphosphate
AFM	Atomic force microscopy
ANARCI	Antigen receptor numbering and receptor classification
BRET	Bioluminescence Resonance Energy Transfer
CCD	Charge coupled device
CDR	Complementarity determining region
CK	Crystal Kappa
CNT	Classical nucleation theory
CTAG	Cancer-testis antigen
CTF	Contrast transfer function
CuP	Copper o-phenanthroline
DARPin	Designed ankyrin repeat proteins
DBD	DNA binding domain
DBSCAN	Density-based spatial scattering of applications with noise
DDD	Direct detection device
DLS	Dynamic light scattering
Dab-RT	HIV reverse transcriptase:diabody complex
ELISA	Enzyme-linked immunosorbent assay
EM	Electron microscopy
Fab	The fragment antigen-binding region
FACS	Fluorescence-activated cell sorting
FR	Framework region
FRET	Fluorescence resonance energy transfer
GFP	Green fluorescent protein
GPCR	G protein coupled receptor
GRAVY	Grand average of hydropathy
GST	Glutathione-S-transferase
HTP	High throughput
HRDC	Helicase and RNaseD C-terminal
IG	Immunoglobulin
IgG	Immunoglobulin G
IMGT	ImMunoGeneTics
IR	Infrared light
ITC	Isothermal titration calorimetry
kDa	Kilo Dalton
KIX	Kinase-inducible domain interacting domain
LED	Light-emitting diode
MACS	Magnetic-activated cell sorting
MAGE	Melanoma-Associated Antigen
MBP	Maltose-binding protein
MFS	Major facilitator superfamily
MNP	Myeloproliferative neoplasms
MRL	MAGE-RING ligases
MSI-H	Microsatellite instability-high
MST	Microscale thermophoresis
Nb	Nanobody
Nb-α	Nanobody interface 1
PDB	Protein data bank
PEG	Polyethylene glycol
PREP	Parallel rapid expression and purification
PREPX	Parallel rapid expression, purification, and crystallization
PR(N)P	Prion Protein
RQC	RecQ C-terminal domain
SAXS	Small-angle X-ray scattering
SEC	Size exclusion chromatography
SCE	Side chain entropy
ScFv	Single-chain variable fragment
SGC	Structural Genomics Consortium
SOL	Solubility mutations G40T:Q49E: L52W:I101V
SPR	Surface plasmon resonance
SRI	Set2-Rpb1 interacting
TEM	Transmission electron microscopy
Trx	Thioredoxin
TUT	Terminal uridylyl transferase
V_HH	Single variable domain on a heavy chain
VPP	Volta phase plate
WH	Winged Helix-like
WRN	Werner syndrome helicase
Y2H	Yeast two hybrid
YFP	Yellow fluorescent protein
ZCCHC	Zinc Finger CCHC-Type Containing

Contents

ABSTRACT **2**

DECLARATION OF WORK **4**

ACKNOWLEDGEMENT **5**

LIST OF ABBREVIATIONS **6**

Chapter 1 Introduction and background **11**

1.1 Summary 12
1.1.1 Existing problems in binder-assisted protein crystallography 12
1.1.2 Existing problems in cryo-EM structure solution of small proteins 14

1.2 Approaches assisting crystallization and cryo-EM 15
1.2.1 Crystallography 15
1.2.2 Cryo-EM Imaging scaffolds 27

1.3 Introduction to nanobodies 29
1.3.1 Description, properties, and the structure 29
1.3.2 Residue numbering 31
1.3.3 The selection methods 33
1.3.4 Application in crystallography and cryo-EM 37

1.4 Crystallization mechanisms 42
1.4.1 The depletion mechanism by polyethylene glycol (PEG) 42
1.4.2 Classical nucleation theory (CNT) versus non-classical nucleation theory 43
1.4.3 Key surface factors driving protein crystallization 46

1.5 Methods to detect and quantify protein-protein interactions 47
1.5.1 Biochemical – ELISA and yeast two hybrid 47
1.5.2 Biophysical – ITC, SPR, MST, FRET, BRET and DLS 49

1.6 Cryo-EM 3D reconstruction principles 54
1.6.1 Principle of negative-staining and cryo-EM image formation 54
1.6.2 Principle of 3D reconstruction 55

1.7 Biology of model proteins 57
1.7.1 RECQL5 and WRN 57
1.7.2 CTAG1B 58
1.7.3 MAGEB1 and MAGEC2 59
1.7.4 PRNP 60
1.7.5 TUT4 60

1.8 Overview of the thesis ····· 62

Chapter 2 Nanobody scaffolds have engineering potential for better crystallization ····· 63

2.1 Introduction ····· 64

2.2 Naïve nanobodies are dissatisfactory crystallization chaperones ····· 65
2.2.1 Methods and materials ····· 65
2.2.2 Result ····· 69
2.2.3 Discussion ····· 72

2.3 Ineffective nanobodies improved by surface mutations ····· 72
2.3.1 Strategy of crystal epitope mutagenesis ····· 73
2.3.2 Methods and materials ····· 73
2.3.3 A single nanobody surface mutation yields an effective chaperone for RECQL5 ····· 76
2.3.3 Discussion and conclusion ····· 78

Chapter 3 Bioinformatics research on the crystal contacts between nanobodies ····· 79

3.1 Introduction ····· 80

3.2 Methods ····· 81
3.2.1 Summary ····· 81
3.2.2 Identifying interacting residues in the crystal lattice ····· 81
3.2.3 Grouping interacting residues into interfaces ····· 83
3.2.4 Renumbering the nanobodies from PDB using the IMGT scheme ····· 84
3.2.5 Clustering the nanobody interfaces based on density ····· 85
3.2.6 Visualising clustered nanobody-nanobody interfaces on the structure ····· 86

3.3 Classification of nanobody-nanobody crystal contacts ····· 88
3.3.1 Nb-α (Class 1) interfaces ····· 89
3.3.2 Class 0 interfaces ····· 93
3.3.3 Class 2 interfaces ····· 95
3.3.4 Class 3 interfaces ····· 97
3.3.6 Class 5 interfaces ····· 101
3.3.7 Class 6 interfaces ····· 103
3.3.8 Class 7 interfaces ····· 104
3.3.9 Class 8 interfaces ····· 106

3.4 Discussion and conclusion ····· 108

Chapter 4 Engineering the nanobody scaffold for reliable crystallization ····· 109

4.1 Introduction ····· 110

4.2 Methods ····· 112
4.2.1 High-throughput nanobody cloning method ····· 112
4.2.2 Improving yield of nanobody ····· 114
4.2.3 Application of parallel rapid expression, purification and crystallization (PREPX) method to the nanobody ····· 118

4.2.4 High-throughput crystal harvesting and data collection ····· 118

4.3 Engineering around Nb-α yielded various Gluebodies ····· 118
4.3.1 Roadmap – 5 iterations of mutagenesis ····· 119
4.3.2 Iterations of nanobody scaffold mutagenesis yield consistently diffracting crystals ····· 124
4.3.3 Constitutive nanobody interfaces were present in 6 different crystal forms ····· 125
4.3.4 The name – 'Gluebody' ····· 134
4.3.5 Discussion ····· 134

4.4 Application of Gluebodies to several target proteins ····· 136
4.4.1 Methods and materials ····· 136
4.4.2 Results ····· 138
4.4.3 Discussion and conclusion ····· 141

Chapter 5 Gluebodies as an imaging scaffold in cryo-EM ····· 144

5.1 Introduction ····· 145

5.2 Di-Gluebodies ····· 145
5.2.1 Dimerizing the Gluebody in solution ····· 146
5.2.2 Negative staining EM and cryo-EM analyses of the target:Gluebody dimer complexes ····· 147

5.3 Future perspective: Tetra-Gluebodies ····· 150

Chapter 6 Conclusions and perspectives ····· 152

6.1 Conclusions ····· 153
6.1.1 Naïve nanobody scaffolds have potential to optimize ····· 153
6.1.2 Nb-α is the most predominant nanobody intermolecular crystal interface ····· 153
6.1.3 High-resolution crystal structures of human RECQL5 helicase APO form (nanobody complex) was obtained through engineering Nb-α ····· 153
6.1.4 The Gluebody is a powerful tool but might not be useful for all proteins ····· 154
6.1.5 The Gluebody scaffolds can be potentially used as imaging scaffolds for cryo-EM ····· 154

6.2 Perspectives ····· 154
6.2.1 Nb-Nb v.s. Nb-target interfaces ····· 154
6.2.2 Large v.s. small Nb-Nb interfaces ····· 155
6.2.3 Exploration on nucleation mechanisms ····· 155

Chapter 7 Bibliography ····· 157

Bibliography ····· 157

Chapter 8 Appendix ····· 169

8.1 Source Code for Chapter 3 ····· 170

8.2 Original sequences of the nanobodies in section 4.2.2.1 ····· 185

8.3 Crystal diffraction data (separate file) ····· 188

8.4 Detailed crystallographic and refinement statistics of RECQL5:Gluebody structures (separate files) ····· 188
8.4.1 7ZML Crystal structure of human RECQL5 helicase APO form in complex with Gluebody G1-001 ····· 188
8.4.2 7ZMM Crystal structure of human RECQL5 helicase APO form in complex with Gluebody G2-001 ····· 188
8.4.3 7ZMN Crystal structure of human RECQL5 helicase APO form in complex with Gluebody G3-048 ····· 188
8.4.4 7ZMO Crystal structure of human RECQL5 helicase APO form in complex with Gluebody G3-052 ····· 188
8.4.5 7ZMP Crystal structure of human RECQL5 helicase APO form in complex with Gluebody G3-055 ····· 188
8.4.6 7ZMQ Crystal structure of human RECQL5 helicase APO form in complex with Gluebody G2star-006 ····· 188
8.4.7 7ZMR Crystal structure of human RECQL5 helicase APO form in complex with Gluebody G2star-011 ····· 188
8.4.8 7ZMS Crystal structure of human RECQL5 helicase APO form in complex with Gluebody G4-043 ····· 188
8.4.9 7ZMT Crystal structure of human RECQL5 helicase APO form in complex with Gluebody G5-006 ····· 188
8.4.10 7ZMV Crystal structure of human RECQL5 helicase APO form in complex with Gluebody G5-006 ····· 188

Chapter 1 Introduction and background

1.1 Summary

1.1.1 Existing problems in binder-assisted protein crystallography

Protein X-ray crystallography has long ago become a routinely used method to elucidate the near-atomic structure of proteins of interest. However, getting the protein to crystallize remains highly troublesome, with no substantial methodological breakthroughs for many decades. The mechanistic specifics remain poorly understood, so that there is no systematic route to coax a protein to pack in an ordered lattice (Yau and Vekilov 2000, Chung, Shin et al. 2010, Sleutel, Lutsko et al. 2014, Sauter, Roosen-Runge et al. 2015, Van Driessche, Van Gerven et al. 2018). Nevertheless, numerous techniques have emerged throughout the past decade to aid crystallization and they generally follow one of three general strategies.

The first is to vary the protein environment by exploring crystallization solution, temperature or physical format (Newman, Bolton et al. 2012, Strelov, Kuranova et al. 2014). This approach has been thoroughly commercialized for over two decades, with a large repertoire of crystallization screens purchasable from vendors, and many laboratories equipped with various flavours of automation.

The second strategy is to modify the protein itself to favor crystallization (Price Ii, Chen et al. 2009, Savitsky, Bray et al. 2010, Bukowska and Grütter 2013), by introducing either major changes to the protein through varying the expression construct (Savitsky, Bray et al. 2010), or by subtle changes on the protein surface, such as by surface entropy reduction (Derewenda 2004) or the “crystal epitope” approach(Price Ii, Chen et al. 2009).

The third approach is to introduce “chaperone” proteins that bind to the target of interest and discover additional routes for forming crystal contacts. Those include fusion tags (such as lysozyme (Thorsen, Matt et al. 2014) and BRIL (Chun, Thompson et al. 2012)) and protein binders (such as Fabs (Day, Rasmussen et al. 2007), scFvs (Fairhead, Preger et al. 2019), nanobodies (Rasmussen, Choi et al. 2011) and DARPins (Batyuk, Wu et al. 2016)). Since adding fusion tags might introduce experimental problems, like decreasing purified protein yields (Savitsky, Bray et al. 2010), chaperone-assisted protein crystallization is especially effective on difficult proteins like membrane proteins (Bukowska and Grütter 2013). In addition, with the improvement and diversification of binder selection methods, generating chaperones to assist structural studies is now

routinely attempted in many labs (Winter, Griffiths et al. 1994, Boder, Raeeszadeh-Sarmazdeh et al. 2012, Zimmermann, Egloff et al. 2020).

Other strategies have been reported, but have either not yet borne fruit (e.g. embedding the protein of interest into a highly porous crystal lattice (Maita 2018, Ward and Snow 2020, Zacharchenko and Wright 2021)), or have not achieved widespread usage and thus validation (e.g. imprinted polymers (Saridakis and Chayen 2013), gravity (Strelov, Kuranova et al. 2014)).

The third strategy above, binder-assisted crystallization, holds the promise to address several key crystallization challenges beyond obtaining the first crystal structure (Bukowska and Grütter 2013). Firstly, they can lock the target protein into conformations that cannot be otherwise isolated (Rasmussen, Choi et al. 2011, Steyaert and Kobilka 2011). Secondly, it can help find the considerably more robust crystallization systems required for structure-based lead discovery, including crystal-based fragment screening, since the introduction of new or additional crystal-crystal contacts mediated by the chaperone can systematically increase polymorphism (Collins, Douangamath et al. 2018, Krojer, Fraser et al. 2020). Thirdly, by engineering binder scaffolds, symmetry can be introduced to the binder:target complex to facilitate crystallization processes (Wukovitz and Yeates 1995, Banatao, Cascio et al. 2006).

Nevertheless, the approach does seem to retain significant opportunities for improvement. For instance, a recent study showed that the shortening of the FG loop on human Fabs by only two amino acids can significantly improve crystallization of human Fabs and Fab:antigen complexes (Lieu, Antonysamy et al. 2020). Another more recent study on diabody scaffold engineering also showed that with an additional disulfide between the VH-VH interface on the diabody scaffold, the diabody:HIV-1 reverse transcriptase (RT) complex was able to crystallize in significantly more conditions than the wild type complex and the RT alone (Chesterman and Arnold 2021).

In this study, I select the nanobody as the kind of protein binder and carry out protein engineering on its scaffold regions. The primary aim of the study is to improve its capability to assist the crystallization of its binding partner.

1.1.2 Existing problems in cryo-EM structure solution of small proteins

Since the breakthrough of resolution, single particle cryo-EM has been routely attempted in many structural biology labs (Cheng 2018). Most structures determined by cryo-EM have a molecular weight of above 250 kDa, while only less than 2% of total structures are below 100 kDa, showing that there is a protein size limit for the cryo-EM methodology, mainly due to the low signal-to-noise ratio of small proteins in cryo-EM images (Yeates, Agdanowski et al. 2020). Although there have been successful attempts on some model proteins by improving instruments and algorithms of imaging processing (Khoshouei, Radjainia et al. 2017, Fan, Wang et al. 2019, Herzik, Wu et al. 2019), it is still generally difficult to resolve small proteins by the cryo-EM method. Breaking this size limit is a crucial step to make the cryo-EM a universal methodology.

Instead of pushing the performance of hardware, a systematic method to multiply the protein size is another solution. This will potentially solve two problems – low contrast of small proteins and difficulty in particle alignment (Yeates, Agdanowski et al. 2020). The low contrast is a primary problem, because it makes particles hard to visualize in the image, making both particle picking and alignment difficult. Difficulty in alignment is then an essential factor that prevents high-resolution structure solving. Inaccurate alignment of particles usually results in low signal-to-noise ratio and blurring and thus details are not visible.

Nanobodies were demonstrated helpful in obtaining high-resolution structures of proteins with good contrast in cryo-EM images by enabling easier alignment of the particles. Particularly for small membrane proteins, the nanobodies protruding outside the detergent micelles help alignment against the protein rather than the micelle (Parker, Deme et al. 2021). However, this only added another 15 kDa to the particle. For those particles not visible in the image, it is more difficult. Therefore, binders with additional protein compartments to increase their sizes were developed recently, such as Megabodies and Legobodies (Uchański, Masiulis et al. 2021, Wu and Rapoport 2021). In particular, Legobodies comprise of a maltose binding protein (MBP) and Fab rigidly assembled with the target-binding nanobody, making the complex a ~120 kDa ensemble, and it was demonstrated useful for obtaining high-resolution structures of two small proteins (<50 kDa).

In addition to the binders, multi-valent symmetric imaging scaffolds were also developed and proved useful. A profound example was the 12-fold DARPin cage, which helped the 3.5 Å structure solution of the green fluorescent protein (GFP) (Liu, Huynh et al. 2019). The structure solution was not feasible without the

12-fold symmetry, which was used to bring the signal-to-noise ratio to a higher magnitude.

Since the DARPin cage and Legobodies were only recently reported, whether they really have an impact on real-life target proteins still remains unknown. They both solved the problems of low constrast and alignment difficulty, but they both require additional protein reagents for the assembly, and they both have drawbacks – Legobodies don't possess symmetry and DARPins are not as widely used as nanobodies. To explore the possibility of the nanobody to become a size-multimerizing imaging scaffold with symmetry is therefore the secondary aim of this study.

1.2 Approaches assisting crystallization and cryo-EM

Protein engineering approaches include fusion proteins, binders, and guest-host lattice to assist protein crystallization, and imaging scaffolds to help with cryo-EM. Among them, nanobodies will be described in more detail in section 1.3.

1.2.1 Crystallography

1.2.1.1 Fusion proteins

The idea of creating a fusion protein to assist crystallization is intuitive. There are proteins easy to crystallize, such as the T4 lysozyme (Faber and Matthews 1990, Heinz and Matthews 1994, Zhang, Wozniak et al. 1995), and there are those very hard to crystallize as well. It was later proven that when the T4 lysozyme was fused to a bad crystallizer, such as G protein coupled receptor (GPCR), the whole engineered protein could crystallize (Thorsen, Matt et al. 2014) (Figure 1.1A). Apart from the T4 lysozyme, apocytochrome b562 (BRIL) is also widely used in crystallizing GPCR when it is rigidly fused to the GPCR (Chun, Thompson et al. 2012) (Figure 1.1B). The rationale is that membrane proteins buried in the detergent micelle with minimal exposed surface can hardly form constant crystal contacts between the molecules, but the soluble fusion partner can protrude out of the micelle and mediate more crystal contacts.

Other canonical tags that help crystallization include thioredoxin (Trx) (Stoll, Manohar et al. 1998, Corsini, Hothorn et al. 2008), Glutathione-S-transferase (GST) (Kuge, Fujii et al. 1997, Lally, Newman et al. 1998), Maltose binding protein (MBP) (Moon, Mueller et al. 2010, Waugh 2016) and SUMO fusion (Holcomb, Spellmon et al. 2017). These were initially used as solubility tags to increase the yield of the target protein,

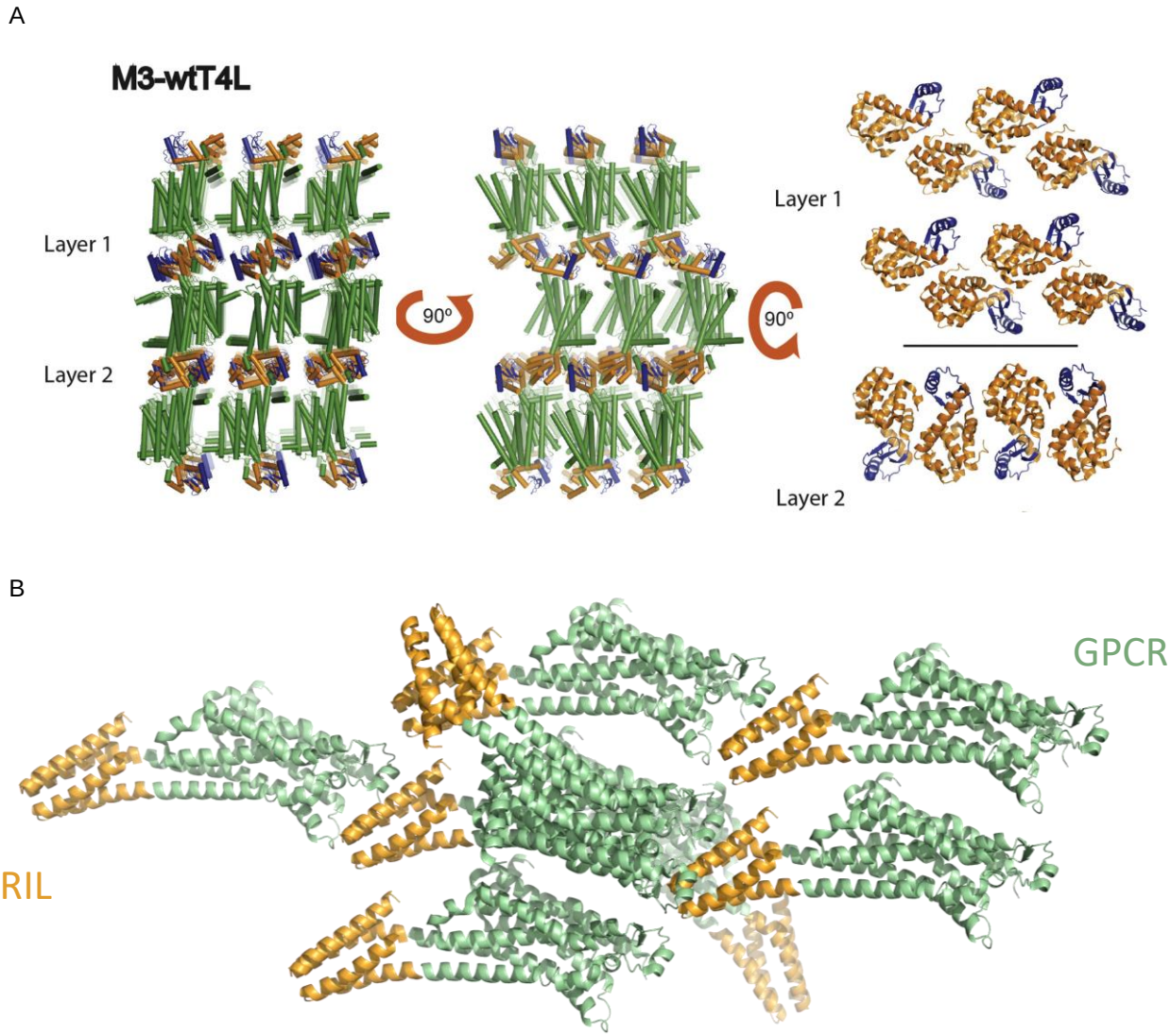


Figure 1.1 Fusion tags facilitate crystallization via creating crystal contacts. A) The modified T4 lysozyme fusion (orange) on the GPCR (green) mediate several crystal contacts enabling crystallogenesis. The figure is taken from (Thorsen, Matt et al. 2014). B) BRIL (orange) fused on the GPCR (green) mediate several GPCR-BRIL and BRIL-BRIL crystal contacts enabling crystallogenensis. The figure is remade from 4EIY in PDB. In both A and B, contacts involving the fusion tags account for most of the crystal contacts.

but with specific linkers, they can also be good crystal chaperones (Lavallie, DiBlasio et al. 1993, Nygren, Stefan et al. 1994, Fox and Waugh 2003, Marblestone, Edavettal et al. 2006, Nallamsetty and Waugh 2006).

1.2.1.2 Protein binders

Binders are proteins that demonstrate binding affinity for certain proteins of interest (antigens). There have been several types of binders used for structural studies (Figure 1.2), such as antigen-binding fragment

A



B

Binder class	Number of all PDB structures	Number of crystal structures	Number of complex crystal structures
Fab	4892	4180	3521
Nanobody	632	468	393
scFv	278	253	200
DARPin	195	185	160
Affimer	18	17	14
Affibody	16	5	5

C

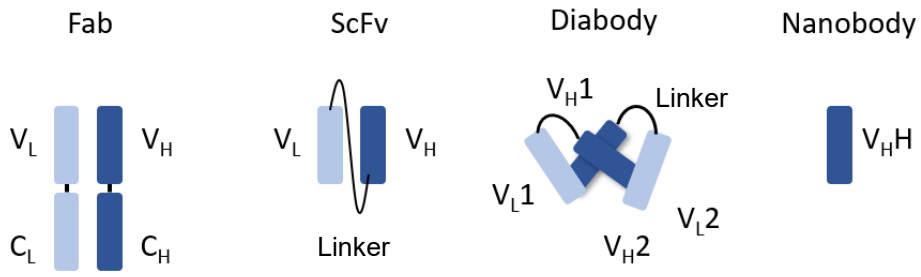
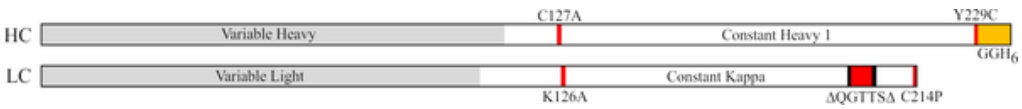


Figure 1.2 Various binder scaffold and their occurrence in PDB. A) Structures of various binder scaffolds with antigen-binding regions shown in green. From Left to right are Fab (PDB code 3TT1), nanobody (PDB code 6R7T), scFv (PDB code 6G8R), DARPin (PDB code 5KNH), Affimer (PDB code 5ML9) and Affibody (PDB code 5DJT) respectively. B) The occurrence of different binder scaffolds in PDB. The numbers are recorded as of Nov 2021. C) Domain arrangements of four types of binders.

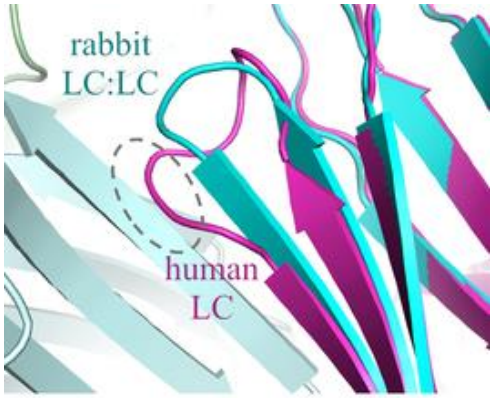
(Fab) (Kovari, Momany et al. 1995), single chain variable fragment (scFv) (Pai, Culver et al. 2011), nanobody, designed ankyrin repeat proteins (DAPRins) (Grubisha, Kaminska et al. 2010, Kummer, Parizek et al. 2012) and diabodies (Chesterman and Arnold 2021).

The Fab, scFv and the nanobody are all derived from antibodies and frequently used as crystal chaperones (Bukowska and Grütter 2013).

A



B



Human Kappa $_{197}$THQGLSSPV$_{205}$
Mouse Kappa THKTSTSPI
Rabbit Kappa T-QGTTS-V

C

Fab	Antigen	Parent Xtal	CK Xtal	Res (Å)	Notes	Reference
G6	Receptor	0	4	3.6	CK1.5	Lilly molecule
Dupilumab	IL4Ra	0	34	3.0	optimized	KEGG D10354
Secukinumab	IL17a	0	54	3.2	optimized	KEGG D09967
h2155	GITR	0	54	2.6	Fab only	US Patent 2013108641
h22G2	TIGIT	0	30	9		US Patent 2016176963
H4	Receptor	0	87	2.6		Lilly molecule

Figure 1.3 Improved Fab scaffold with shortened Kappa loops results in reliable crystal chaperoning. The figures are adapted from (Lieu, Antonysamy et al. 2020). A) Mutational designed of the Fab fragment, including single mutations and the shortening of the Kappa loop. B) Structural comparison of the human (longer) and rabbit (shorter) Kappa loops. C) Significant improvement of the reliability of the Fab as a crystal chaperone with shortened Kappa loops, comparing CK (crystal Kappa) crystal numbers with parental crystal numbers.

The Fab is a region of an antibody that binds the antigens. It is composed of two chains (heavy chain and light chain), and is usually around 48kDa (Gupta and Shukla 2017). The Fab can be expressed in either *E. coli* or insect cells (Furuta, Ogawa et al. 2010, Rouet, Lowe et al. 2012), and it is widely used in structural biology labs. The scFv is a single-gene protein derived from the Fab (Weisser and Hall 2009). On the one hand, the Fab has the largest surface area to introduce additional crystal contacts compared with the nanobody, but on the other hand the Fab is more than three times larger than the nanobody, making it much harder to engineer. However, it has been found there is a significant improvement of its reliability as a crystal chaperone when the Kappa loop is shortened (Lieu, Antonysamy et al. 2020) (Figure 1.3). For 6 different targets, their Fabs with shorter Kappa loops helped the target:Fab complexes crystallized in lots of conditions when the original Fab scaffold could not help with co-crystallization (Figure 1.3C). Still, most crystals diffracted worse than 3 Å, indicating that more is to be optimized on the Fab fragment.

The single-chain variable fragments have the heavy chain (V_H) and light chain (V_L) of the Fab that are connected by a flexible linker (Maynard and Georgiou 2000), so that one open reading frame on the plasmid is enough to produce the binder, as opposed to Fabs where two ORFs are needed and proteins need to assemble after expression leading to mixes of V_H-V_H, V_H-V_L, and V_L-V_L. The scFv also only contains the variable domains, making it much smaller (~28kDa) than the Fab (WoÈrn and PluÈckthun 2001, Holliger and Hudson 2005). Both Fab and scFv both play essential roles in mediating crystal contacts in the lattice (Figure 1.4).

DARPins are a type of non-antibody binder (Figure 1.5) derived from the natural ankyrin repeat protein fold. The affinity of the DARPins is determined by the amino acid sequence of flexible loops present on one face of the ankyrin repeats (Plückthun 2015). There have been a few co-crystal structures with DARPins, including membrane proteins, showcasing its potential as a crystal chaperone (Kohl, Binz et al. 2003, Du, Wang et al. 2014).

Diabodies are derivatives of scFv whose linker between the heavy chain and the light chain is much shorter than scFV, resulting in the dimerization of scFv in vivo. (Chesterman and Arnold 2021). The engineered diabody with additional disulfide in the V_H-V_H interface was recently demonstrated to significantly improve the crystallization propensity compared to the original diabody scaffold (Figure 1.6). The modification on the interface significantly increased the crystallization propensity of the complex as

more crystal hits appeared in different conditions (Figure 1.6B). Although there is only one successful case as proof of principle and the diffraction of the crystals was only 3.0 Å, it strengthens that binder scaffold optimization is a feasible approach to improve crystallization.

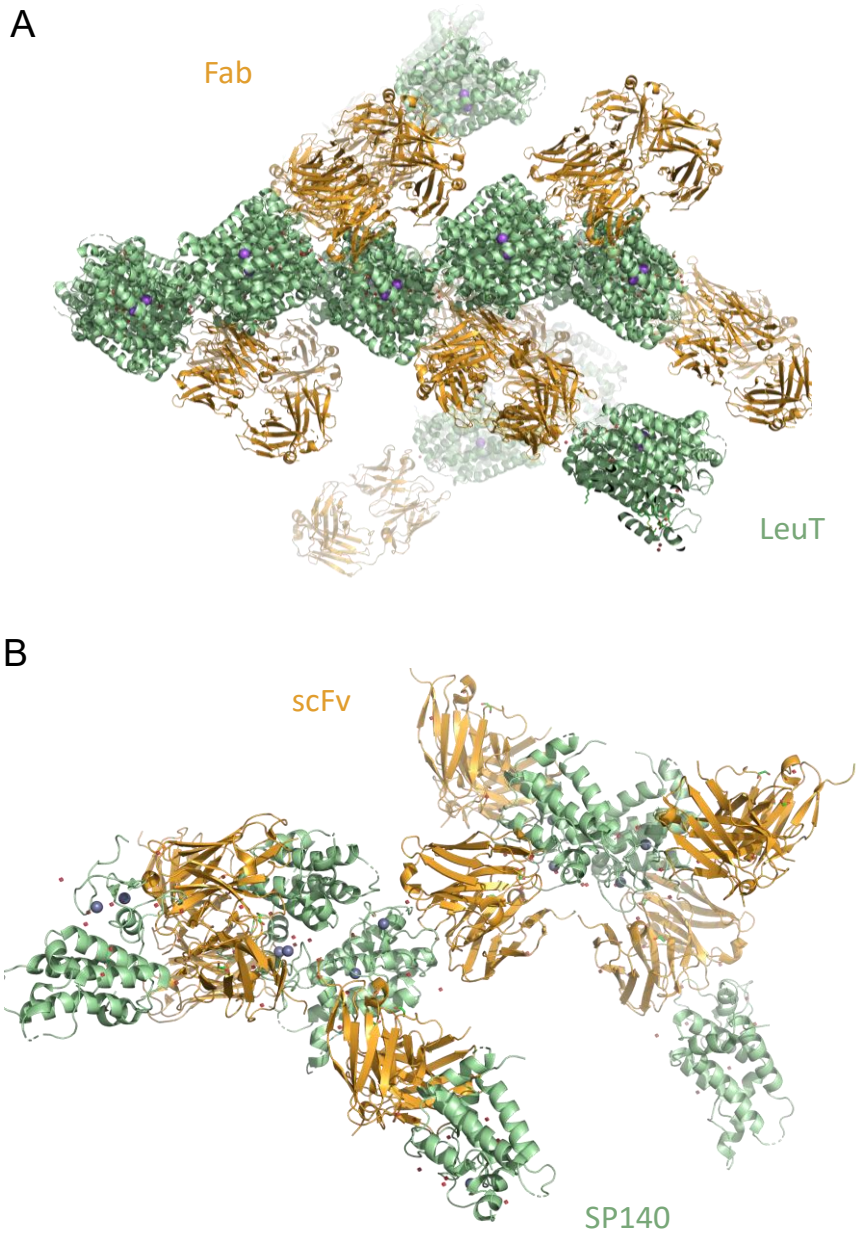


Figure 1.4 Fab and scFv facilitate crystallization via creating crystal contacts. A) The Fab (orange) bound to the LeuT (green) mediate several crystal contacts enabling crystallogenesis. The figure is remade from 3TT1 in PDB. B) ScFv (orange) bound to the SP140 PHD-Bromodomain (green) mediate several crystal contacts enabling crystallogenensis. The figure is remade from 6G8R in PDB.

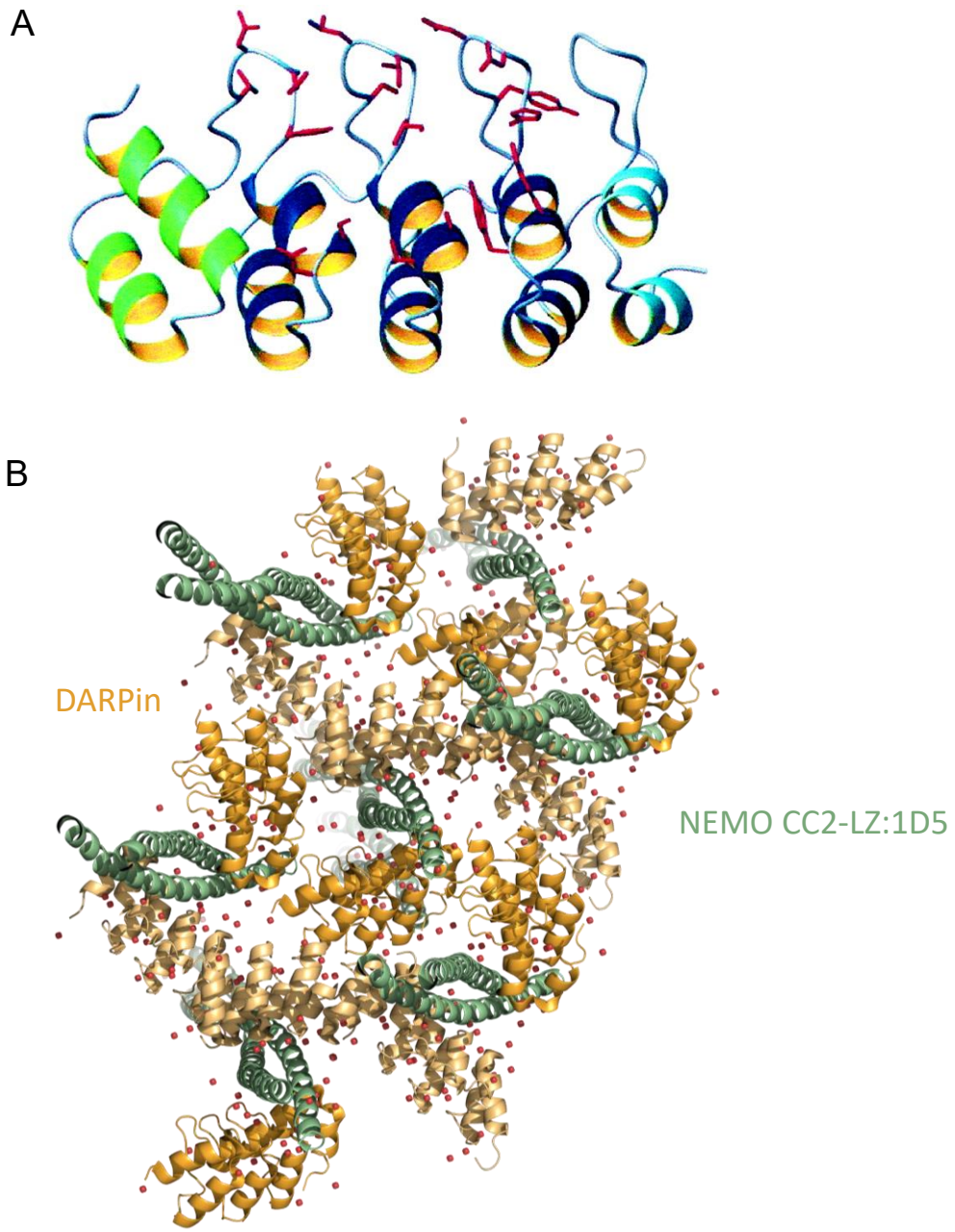


Figure 1.5 DARPin facilitates crystallization via creating crystal contacts. A) The dissection of DAPRin structure. The helices of the N-terminal cap, internal repeats and the C-terminal cap are shown in green, dark blue and light blue respectively. The number of internal helices can vary and helices of randomized sequences can be assembled into one DARPin. Different combinations result in a library of DARPins with different affinity. The figure is taken from (Kohl, Binz et al. 2003). B) DARPin (orange) bound to the NEMO CC2-LZ:1D5 (green) mediate several crystal contacts enabling crystallogenensis. The figure is remade from 2V4H.

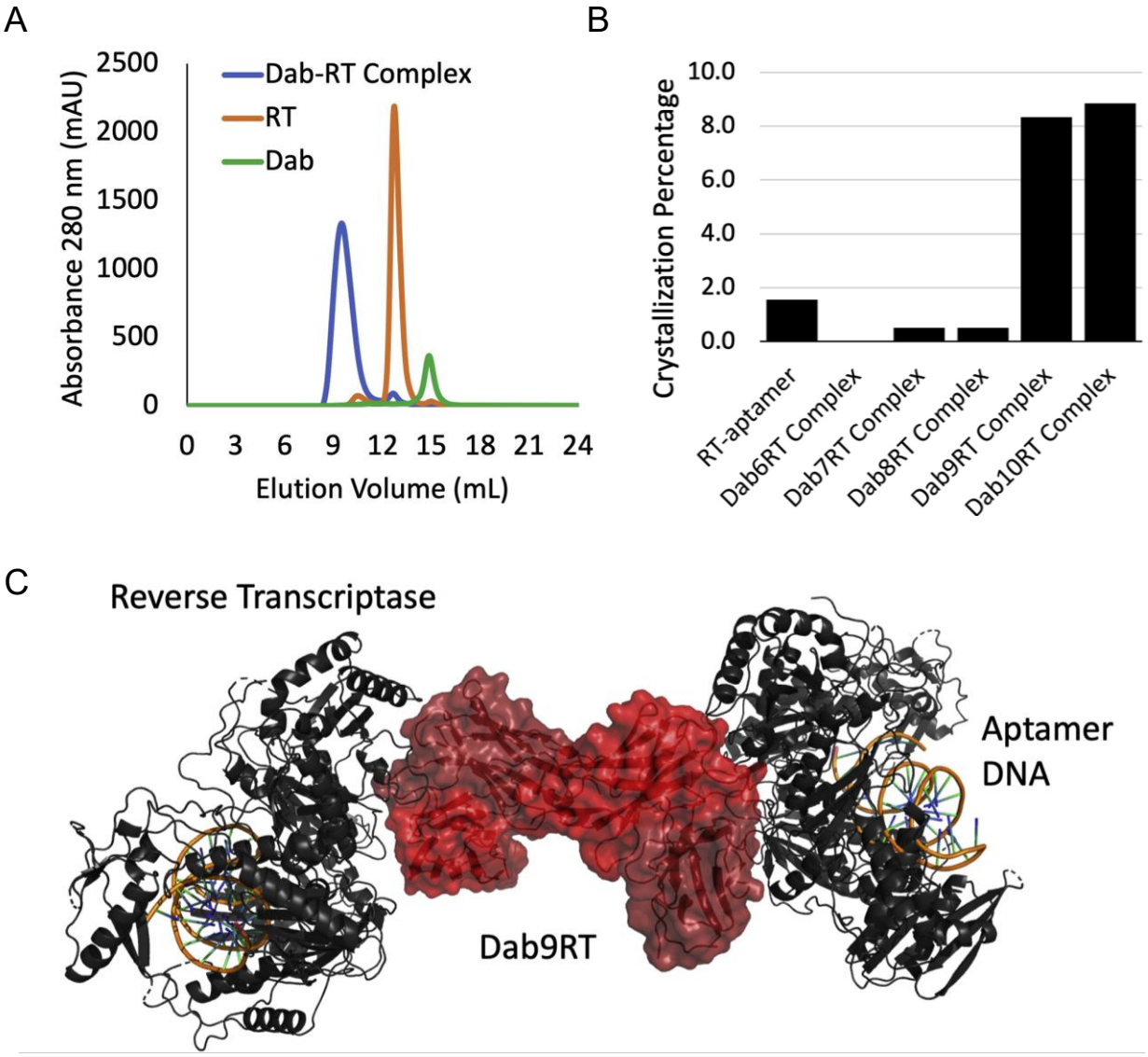


Figure 1.6 Disulfide engineered diabodies significantly increases the crystallization propensity of the HIV-1 reverse transcriptase:diabody (Dab-RT) complex. The figures are adapted from (Chesterman and Arnold 2021). A) The size exclusion chromatography of Dab-RT, reverse transcriptase (RT) and Diabody (Dab). B) HIV-1 reverse transcriptase in complex with disulfide engineered diabodies (Dab8RT, Dab9RT, Dab10RT complex) have higher crystal hit frequencies than the original scaffold (Dab6RT, Dab7RT complex) and the reverse transcriptase alone (RT-aptamer). C) The 3.0 Å structure of the complex of Dab9RT-RT-aptamer.

1.2.1.3 Protein surface engineering

Protein surface engineering includes surface residue mutagenesis (Derewenda 2004) and side chain modifications (such as reductive alkylation) (Rayment 1997). Both methods involve modifying the side chains of certain residues, to create good crystal contacts to trigger crystallization. The change of the surface of a protein that already crystallizes might also lead to a new crystal form (Mateja, Devedjiev et al. 2002).

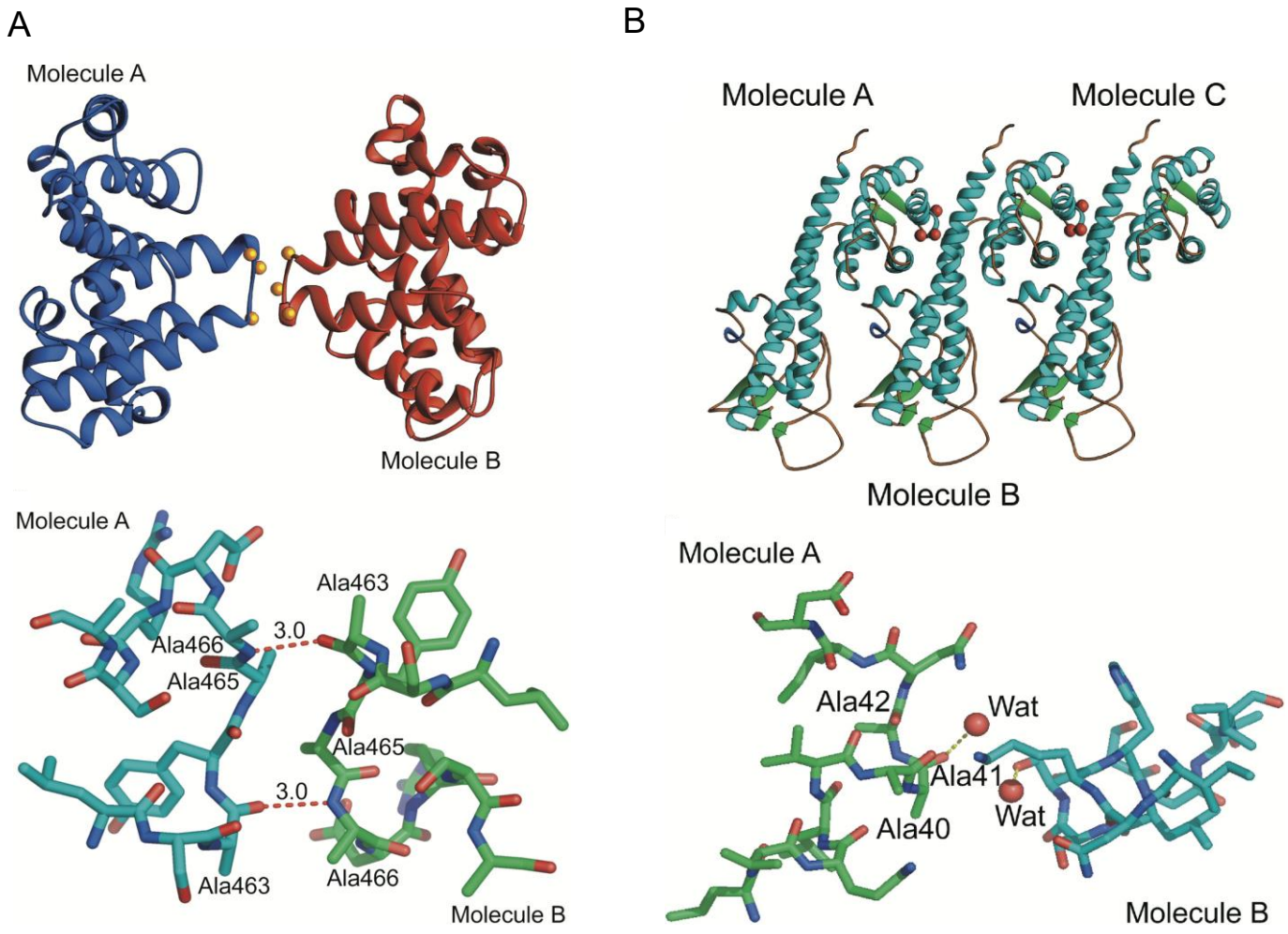


Figure 1.7 Surface Entropy reduction engineering resulted in novel protein structures. The figures are adapted from (Derewenda 2004). A) The RGSL domain of PDZRhoGEF (top) and the close-up view of the mutated crystal contact (bottom). In this case, the triple surface mutant K463A/E465A/E466A yielded high-quality crystal straight from the commercial coarse screens (Hampton Research Crystal I and II). The mutated residues formed key crystal contacts for the crystal. B) The Yersinia pestis LcrV Antigen (top) and the close-up view of the mutated crystal contact (bottom). The triple mutation K41A/D42A/K43A is present in the crystal contact.

For surface residue mutagenesis, there are several strategies, and the representative strategies are surface entropy reduction and the crystal epitope approach.

The surface entropy reduction strategy (Derewenda 2004, Derewenda and Vekilov 2006, Cooper, Boczek et al. 2007) has been frequently used in protein surface engineering for crystallization. The premise of the strategy is when proteins pack in the crystal, the molecules undergo an entropic loss due to stabilization of side chains and the more loss is required the more difficult it is for crystals to form. Therefore, when side chains have less entropy in solution, the more easily they form crystal contacts (Derewenda 2004). Most recommended mutations are lysines or glutamates to alanines (Figure 1.7). Over the past two decades, there have been numerous successful cases with this strategy, producing crystal hits for mutants when wild-type proteins failed in crystallizing (Kim, Dobransky et al. 2005, Anil, Riedinger et al. 2013, Males and Davies 2019). There is also a free server convenient to use to help find the right residues to mutate (Goldschmidt, Cooper et al. 2007).

The concept of crystal epitopes was first introduced by John Hunt (Columbia University) back in 2009 (Price Ii, Chen et al. 2009). A crystal epitope is an array of consecutive 3-6 residues that frequently appear in the crystal contacts generalized from all existing crystal structures in Protein Data Bank. To do crystal epitope engineering is to find out the appropriate successive 3-6 residues exposed on the protein surface and mutate them to one of the crystal epitope patterns. There was a server or plug-in in the ICM-pro programme (Unpublished data) to search for the right residues of any protein to mutate based on this strategy. Michael Fairhead in Centre for Medicines Discovery Department (Previously SGC) of the University of Oxford proved the effectiveness of the strategy (Figure 1.8, unpublished data).

Michael evaluated the strategy for three purposes – obtaining novel structures (11 targets), obtaining constructs yielding reproducible crystals (15 targets) and creating new crystal forms for targets (9 targets). The success rate for the three purposes using this strategy were 27.2%, 20% and 77.8% respectively, demonstrating the effectiveness of the crystal epitope strategy. In addition, nearly half of the structures were obtained with the mutations directly mediating the crystal contacts, indicating the crystal epitopes could influence crystal packing both directly and indirectly. Unfortunately, the server has been down since 2018.

Class of Target	Purpose of the Class	Number of Targets Tested	Mutant Gives Structure	Mutations Involved In Crystal Interfaces	Mutant Yields New Crystal Forms	Mutant Yields Reproducible Crystals	Success Rate	PDB Code/ SGC Model ID
Novel structure	To test whether the crystal epitope strategy helps non-crysatllizing proteins crystallize	11	3	1	n/a	2	27.3%	4UYI 4UEG JMJD2BA-m001
Irreproducible crystals	To test whether the crystal epitope strategy helps proteins crystallize more reliably	15	8	4	4	3	20%	SPIN4A-m004 SPIN1A-m002 SP100A-m020
New crystal form required	To test whether the crystal epitope strategy hels yield more crystal forms from one protein	9	7	2	7	5	77.8%	ATAD2A-m003 BRD1A-m037 BRD1A-m033 JMJD1BA-m003 JMJD1CA-m002 JMJD2AA-023 JMJD2AA-m024 JMJD2AA-m035 JMJD2DA-m025 WSPRPAA-m001

Figure 1.8 Crystallization performance of the crystal epitope strategy guided mutagenesis. (Michael Fairhead, unpublished data). 35 targets were selected in this experiment, and they fell into three classes, i.e. targets without a structure, targets with irreproducible crystals, and targets which require new crystal forms. The crystallization trials were performed using commercially available coarse screens. Numbers in red are used for the calculation of success rates. Details of SGC models can be found on the website https://www.thesgc.org/xtalforms.

1.2.1.4 Guest-host lattice

The idea of a guest-host lattice is to accommodate the protein of interest as a guest protein into a known highly porous scaffold crystal lattice (Ward and Snow 2020). Although there are several types of scaffold crystal lattices available (Ni and Tezcan 2010, Deng, Grunder et al. 2012, Huber, Hartje et al. 2017), only two model protein structures have been solved using this strategy until now (Maita 2018, Zacharchenko and Wright 2021) (Figure 1.9).

One of the successful cases was using R1EN porous crystal to accommodate the Ubiquitin (Ub) (Figure 1.9A) (Maita 2018). It was demonstrated that when ubiquitin was genetically fused to the C-terminal end of R1EN with a certain length of linker, the Ub co-crystallized with R1EN and sat in the hollow pores in the lattice, which enabled the X-ray structure solution of Ub. Additionally, the phasing of the diffraction data only used the R1EN structure as a search model, suggesting that this method might reduce the effort of experimental phasing when determining the structure of novel proteins.

The other successful case was using the highly-porus lattice of the B-cell lymphoma 6 (BCL6) protein to accommodate the short peptides of BCL6-binding domains 1 & 2 of the nuclear receptor corepressor-1 (NcoR1BBD1 and NcoR1BBD2), while NcoR1BBD2 was genetically fused with BCL6 (Figure 1.9B).

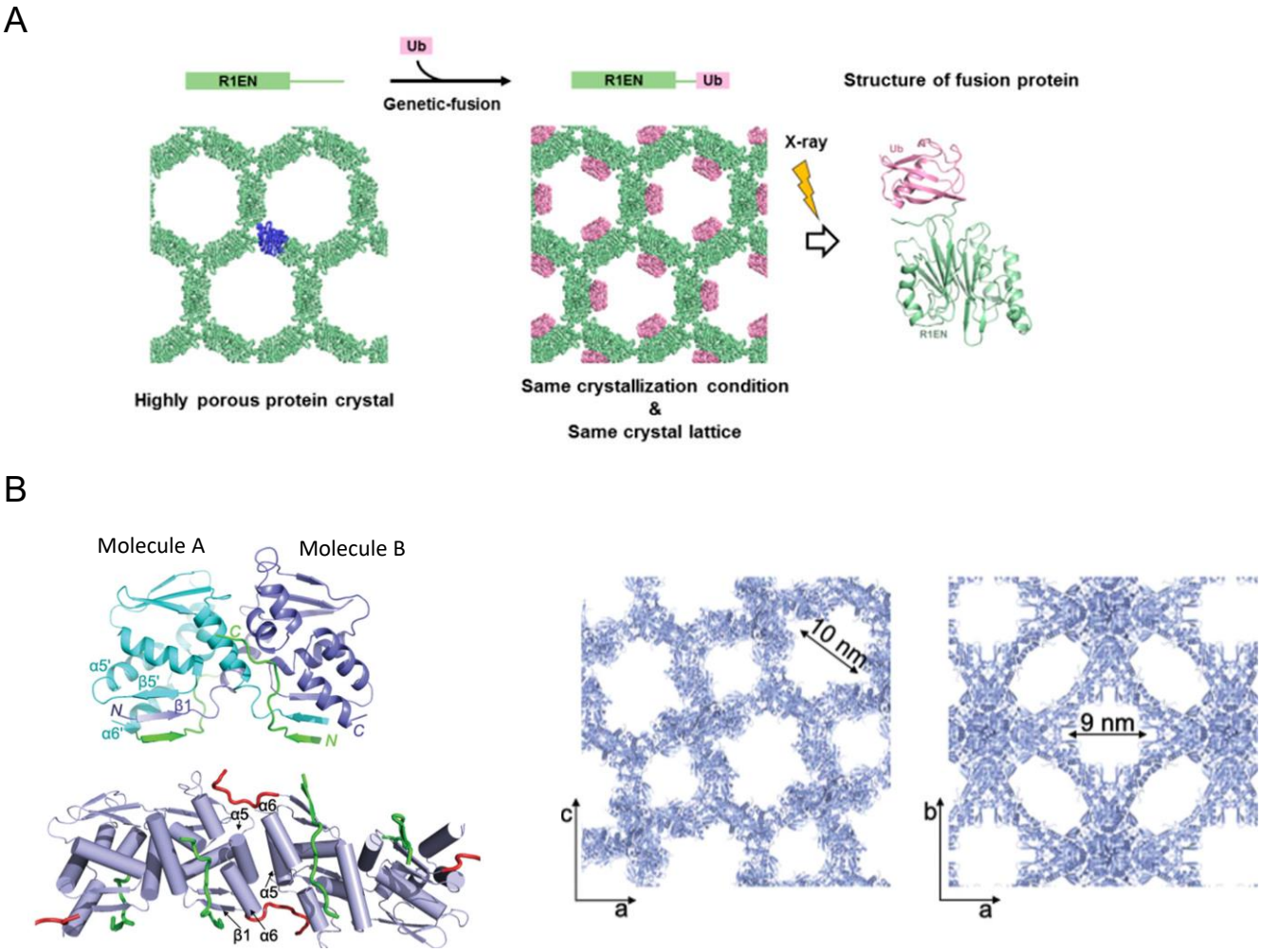


Figure 1.9 Structures solved using the guest-host lattice strategy. A) R1EN (blue) crystallizes into a porous hexagonal crystal lattice (green). The Ubiquitin molecule is shown in pink here. This figure is taken from (Maita 2018). B) The B-cell lymphoma 6 (BCL6) protein fused with the BCL6-binding domain 2 of the nuclear receptor corepressor-1(NcoR1BBD2) forms a complex with the BCL6-binding domain 1 of the nuclear receptor corepressor-1 (NcoR1BBD1). In the cartoon structure, BCL6-NcoR1BBD2 fusion is shown in cyan and blue, and NcoR1BBD1 is shown in green. In the cylinder structure, BCL6 is shown in blue, NcoR1BBD2 is shown in red and NcoR1BBD1 in green. The porous crystal lattice formed by this complex is shown on the right with two vertical views presented. The figure is taken from (Zacharchenko and Wright 2021).

While these are nice proof-of-principle work for the idea, there are drawbacks that may limit the widespread of the method. Due to the size limit of the scaffold lattice, there is a theoretical size limit for the guest protein for each designed lattice. It might also be challenging to apply the method to membrane

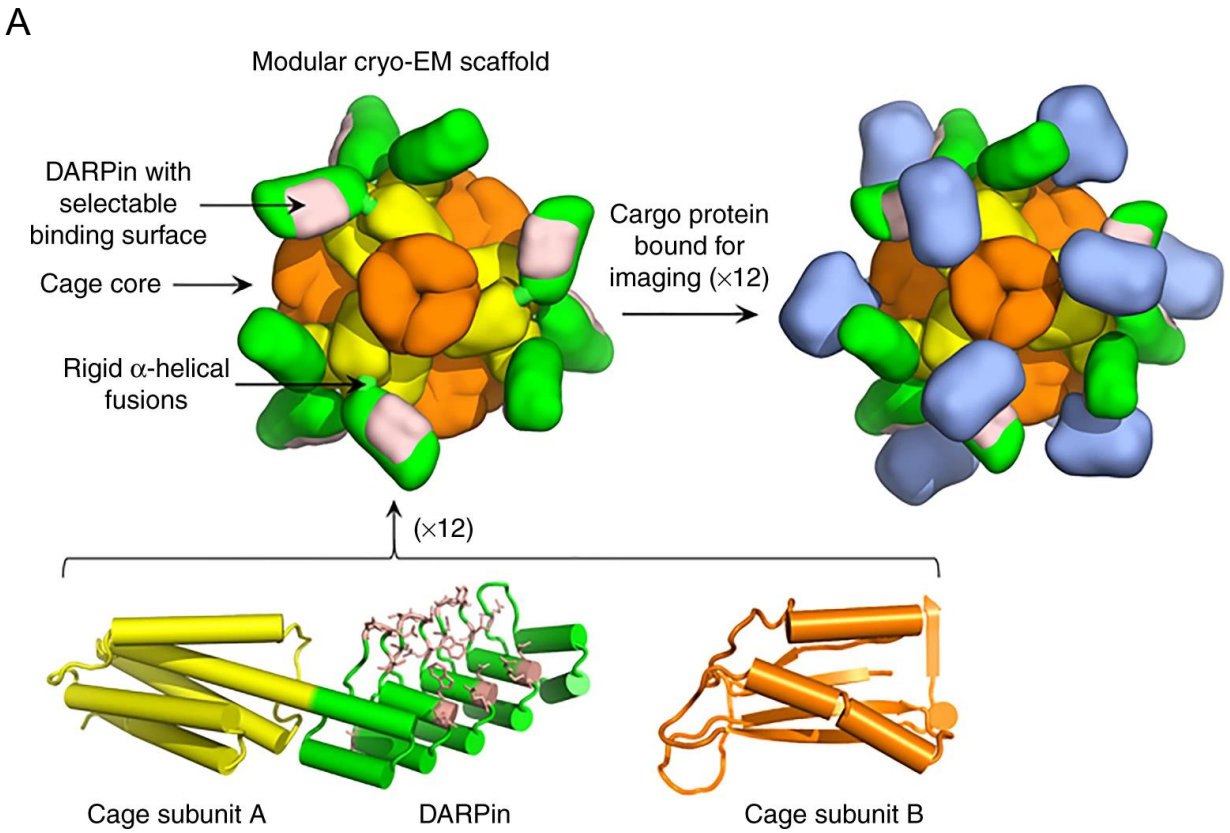
proteins, which need to be surrounded by detergents, lipids, or other molecules to stay dissolved in aqueous solution. Therefore, still considerable amount of time and work is necessary to develop this method.

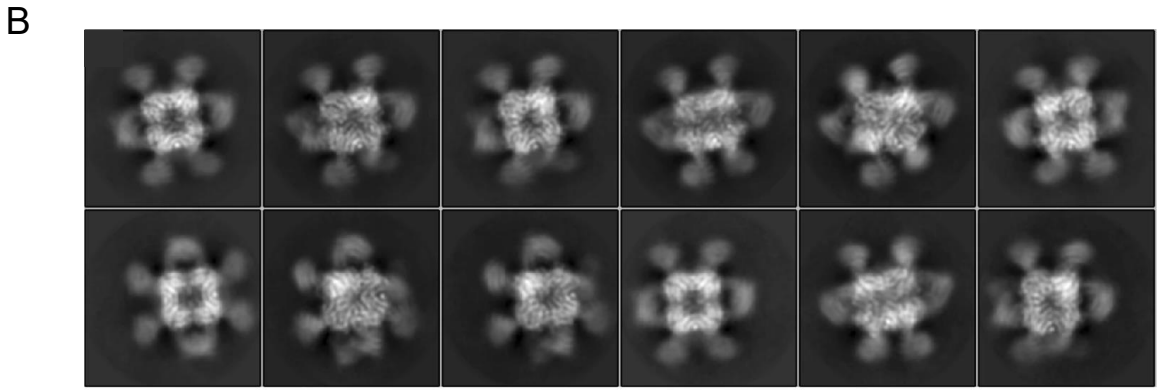
1.2.2 Cryo-EM Imaging scaffolds

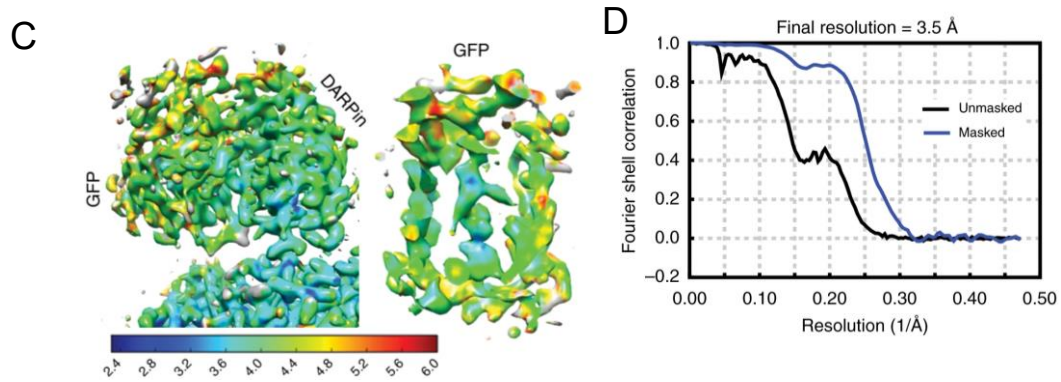
In order to solve the problem of size limit for cryo-EM, as opposed to adding protein chaperones to the small target protein to increase its size, the small proteins can also be assembled onto large symmetric macromolecular cages, so that it is feasible to get the high-resolution structure of the target through solving the structure of the whole assembly (Yeates, Agdanowski et al. 2020). The cages are called the scaffold and the small proteins are also called the cargos. When the cargo is rigidly attached to the scaffold which is known to be easily resolved to high resolution by cryo-EM, even if the cargo is not clearly visible, the scaffold can help with the alignment of the cargo images, and thus lead to high signal-to-noise ratio.

This idea has been proven by the successful structure solution of the 26 kDa green fluorescent proteins (GFPs) symmetrically bound to a high-order cage via rigidly fused DARPins (Liu, Gonen et al. 2018, Liu, Huynh et al. 2019). In this scaffold, 12 pairs of cage subunits (subunit A and B) auto-assembled into an imaging scaffold with 12 fold symmetry (Figure 1.10A). Subunit A was genetically fused with a GFP-binding DARPin molecule through an alpha helix linker. This linker guaranteed the rigid fusion of DARPins. In the 2D and 3D classification results, the scaffold was clearly resolved, but the cargos were still obscure (Figure 1.10B). Therefore, the 12-fold symmetry was taken advantage of, when all the other GFP densities were computationally rotated to contribute to only one copy of the GFP densities on the assembly. This equalled increasing the data size another 11 folds. Finally, the GFP was resolved to 3.5 Å (Figure 1.10C and D).

Figure 1.10 (Next page) Synthetic molecular cage as an imaging scaffold helps solve the high-resolution cryo-EM structure of green fluorescent protein (GFP). The figure is taken from (Liu, Huynh et al. 2019). A) The assembly of the imaging scaffold. 12 sets of cage subunit B (orange) and cage subunit A (yellow) fused with DARPin (green) auto-assemble into an imaging scaffold with 12 vacant positions for the binding of the target protein (blue). B) 2-D classification of the imaging scaffold with GFP (26 kDa) as the target protein. The GFP densities are clearly seen protruding out of the scaffold, but relatively smeared compared to the core. C) High-resolution 3-D reconstruction of the GFP and DARPin colored by local resolution. The reconstruction was done using T symmetry expansion, signal subtraction and 3-D classification without alignment. D) Gold-standard FSC curve for determining the resolution of the reconstruction of GFP.


A
Modular cryo-EM scaffold
DARPin with selectable binding surface
Cage core
Rigid α-helical fusions
Cargo protein bound for imaging (×12)
(×12)
Cage subunit A
DARPin
Cage subunit B


B


C
GFP
DARPin
GFP
2.4
2.8
3.2
3.6
4.0
4.4
4.8
5.2
5.6
6.0
D
Final resolution = 3.5 Å
Fourier shell correlation
Resolution (1/Å)
Unmasked
Masked
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
0.00
0.10
0.20
0.30
0.40
0.50

1.3 Introduction to nanobodies

1.3.1 Description, properties, and the structure

Antibodies, also called immunoglobulins (IG), are a group of glycoproteins that animials produce to act against pathogens, also called antigens (Lefranc and Lefranc 2001). Antibodies act on antigens by specifically binding to their surfaces to inactivate them in the host. There are five different types of IGs which function differently in the body, namely Immnoglobulin G (IgG), Immnoglobulin E (IgE), Immnoglobulin M (IgM), Immnoglobulin A (IgA) and Immnoglobulin (IgD), among which IgG accounts for 75% in blood (Schroeder Jr and Cavacini 2010). The IgGs contain a pair of heavy and light chains, on which there are both constant (C_H on heavy chains and C_L on light chains) and variable domains (V_H on heavy chains and V_L on light chains).

However, antibodies in Alpaca only have the heavy chain and therefore there is only a single variable domain (V_HH) on the heavy chain. Since the V_HH much easier to modify, the nanobody is derived from the Alpaca antibodies (Hamers-Casterman, Atarhouch et al. 1993, Muyldermans 2013). It is comprised of two regions – the Complementarity Determining Regions (CDRs) and non-CDRs (Padlan 1994). The former determines the specificity and affinity of the nanobody to its target protein (Koide, Tereshko et al. 2007), and therefore is highly variable. The latter forms the scaffold of the nanobody, and therefore is relatively constant (Mitchell and Colwell 2018) (Figure 1.11).

The nanobody is a small and structurally conserved protein (Chothia, Lesk et al. 1992). It has around 120 amino acids, which form several β-sheets with three large CDR loops (Figure 1.12). The whole protein is only around 15 kDa, but the variation of CDR loops can provide sufficient diversity (Beghein and Gettemans 2017).

The nanobody is a relatively easy protein to work with due to the following properties. It is well-expressed in the E.coli (BL21) protein expression system (Arbabi-Ghahroudi, Tanha et al. 2005), the yield is up to 45 mg/L (Salema and Fernández 2013, Ta, Steen Redeker et al. 2015, de Marco 2020). The nanobody is also stable at the room and higher temperatures (Arbabi Ghahroudi, Desmyter et al. 1997, Van der Linden, Frenken et al. 1999, Dumoulin, Conrath et al. 2002), which allows purification outside the cold room.

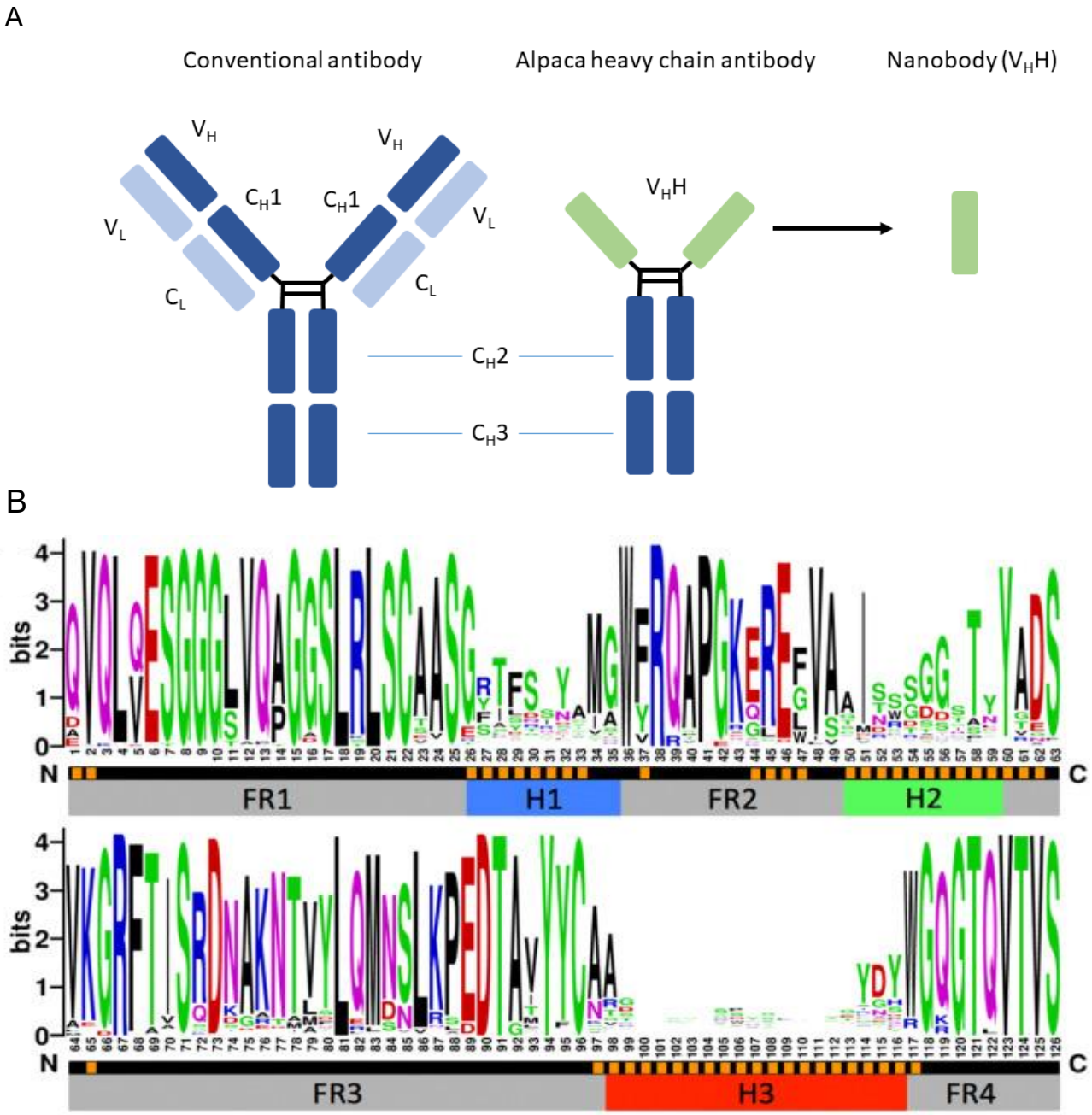


Figure 1.11 Nanobody is a single V_HH comprising of three CDR loops. A) Nanobodies are derived form Alpaca heavy chain antibodies. B) Consensus amino acid analysis of the nanobody. In this analysis, 90 sequences from PDB were analysed. The residue positions that have a >10% chance to contact the antigen is marked as orange. The size of the letter indicates the frequency of the kind of amino acid that occurs in the specific position. Relative constant regions of the nanobody are coloured grey and named as framework regions (FR1- FR4), and as a result four FRs are present in the nanobody. These regions can be defined as the scaffold or non-CDR. On the other hand, variable regions are coloured blue, green or red (H1- H3), which are CDR loops 1 to 3 respectively. This figure is taken from (Mitchell and Colwell 2018).

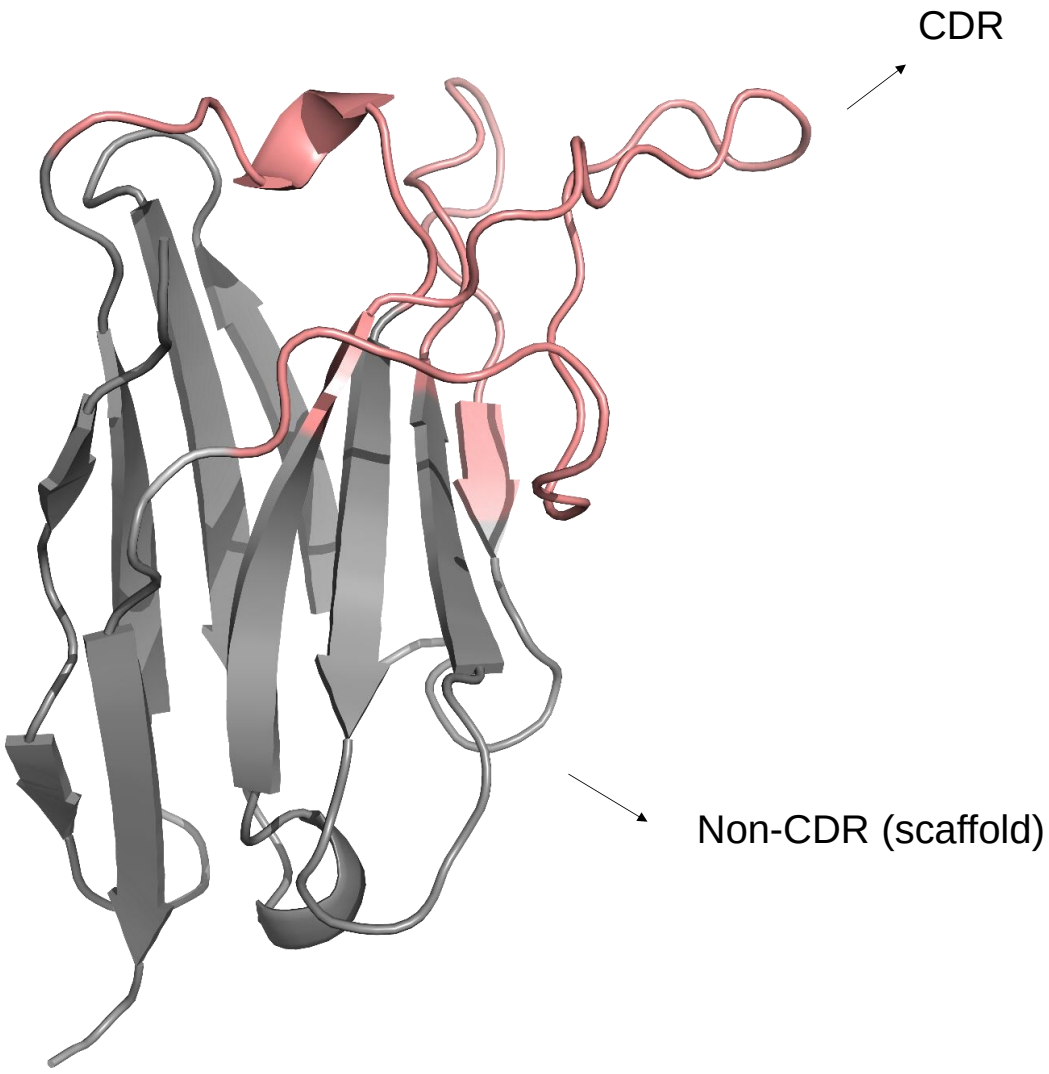


Figure 1.12 3-dimensional structure of a nanobody remade from the nanobody of 5O0W in PDB. The nanobody contains several parallel β-sheet structures, which form most of the framework regions (non-CDR) shown in grey. The CDR regions are shown in light pink. Compared to non-CDRs, the CDRs are more flexible and exposed to the solution.

1.3.2 Residue numbering

For IGs, there have been numerous numbering schemes (Kabat, Te Wu et al. 1992, Al-Lazikani, Lesk et al. 1997, Honegger and PluÈckthun 2001, Lefranc, Pommié et al. 2003, Abhinandan and Martin 2008), so that it is convenient to discuss the residue of the same functional position across different types of IGs. Representative schemes are the Kabat scheme (Kabat, Te Wu et al. 1992) and the international

ImMunoGeneTics (IMGT) scheme (Lefranc, Pommié et al. 2003). Throughout this work, I made use of the Antigen receptor Numbering and Receptor Classification (ANARCI) web server (Dunbar and Deane 2016) and the IMGT scheme (Figure 1.13).

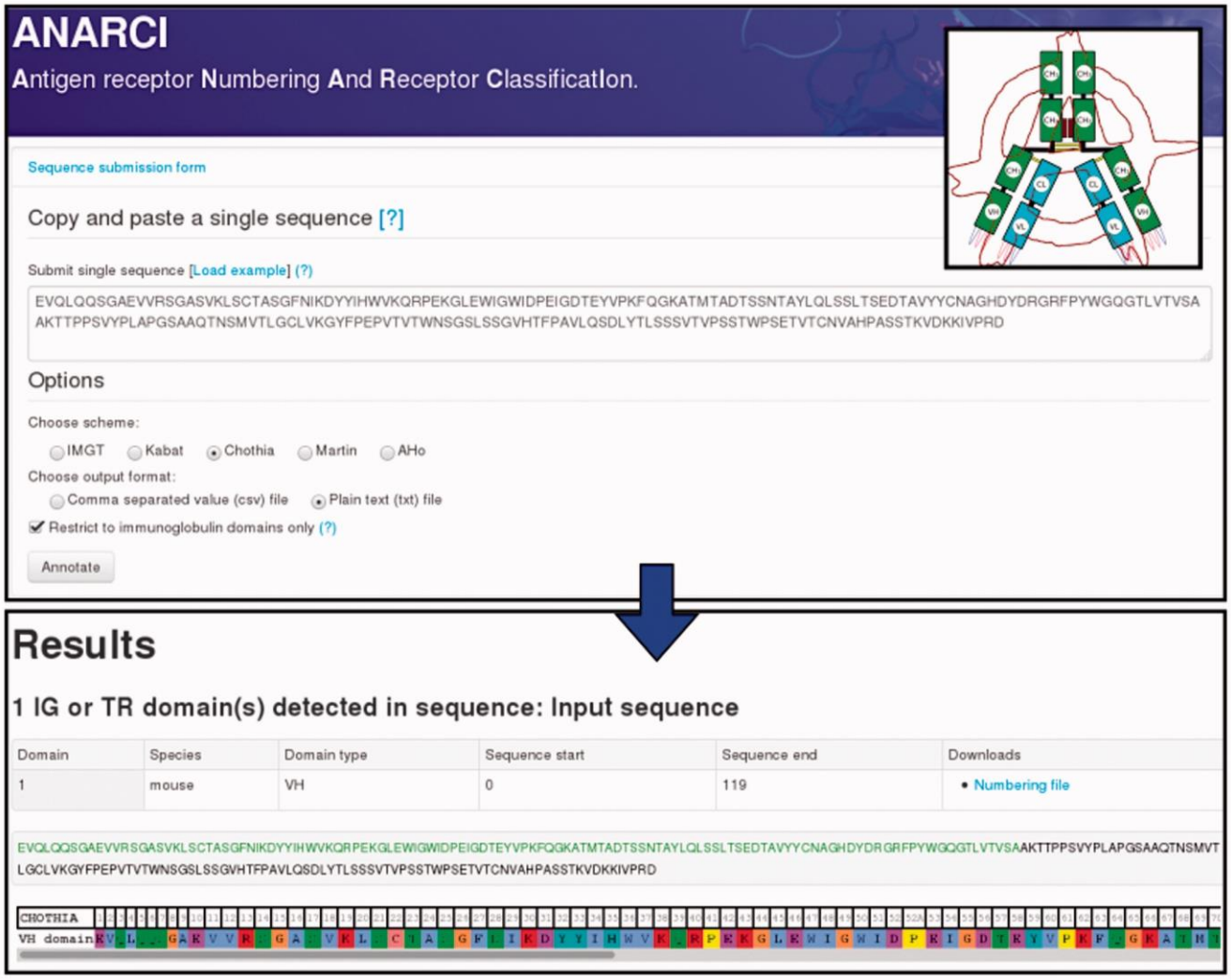


Figure 1.13 ANARCI web server user interface. This is an easy-to-use platform where you only need to upload the sequence of interest and choose from one of the five numbering schemes. The result of the analysis will be the numbering of the sequence and identification of the domain type of this sequence if it is part of the immunoglobulin or T-cell receptor. There is also a choice of multiple sequence input, from which you will also get an alignment of this sequences with the specific numbering you choose. This figure is taken from (Dunbar and Deane 2016).

1.3.3 The selection methods

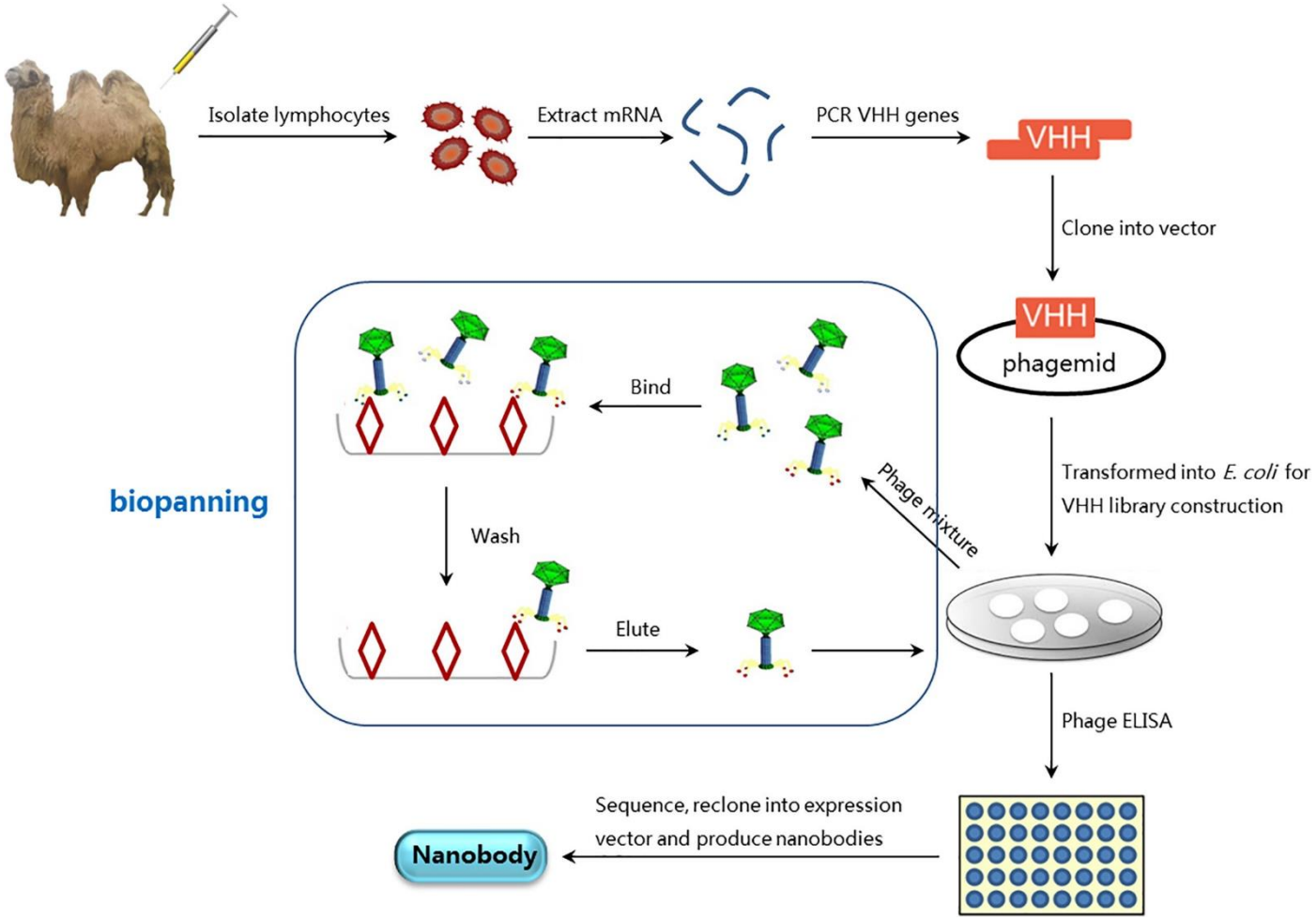


Figure 1.14 The whole nanobody selection process using phage display as the biopanning method. This figure is taken from (Liu, Song et al. 2018).

The whole nanobody selection process basically includes three parts – library construction, biopanning and affinity validation (Figure 1.14)(Liu, Song et al. 2018).

1.3.3.1 Library construction

There are currently three types of libraires for nanobody generation – immune libraries, naïve libraries and synthetic libraries (Liu, Song et al. 2018).

Immune libraries are the classic route to generate nanobodies. Antigens are injected into the vessels of Alpaca/Llama, and then lymphocytes are isolated from the animals to extract the mRNAs, from which V_HH

cDNA libraries were cloned into different vectors, such as phagemids (Figure 1.14). By this method, high affinity nanobodies (pM level) can efficiently be obtained from even small libraries (~10^6 transformants) (Liu, Song et al. 2018). However, due to the immunisation step on the animal, the animal access becomes the speed-limiting step. In addition, libraries against toxic antigens are hard to obtain.

By contrast, naïve libraries and synthetic libraries successfully bypass the need for animals. However, compared to the immune libraries, it is harder to get high affinity binders from them due to lack of the repeated *in-vivo* affinity maturation steps (De Genst, Handelberg et al. 2004). Naïve libraries are constructed from the single domain antibody combinations of 10-20 non-immunized animals (Monegal, Ami et al. 2009). For synthetic nanobody (sybody) libraries, the scaffold is first chosen from one nanobody, and then different strategies are applied to introduce the diversity into the CDRs of nanobodies, one of which is decried in Figure 1.15A. Most widely applied sybody libraries are McMahon and Moutel libraries, which are created for phage display and yeast surface display respectively (McMahon, Baier et al. 2018, Mitchell and Colwell 2018). Although two weeks' animal immunization time is saved using naïve or synthetic libraries, usually larger library size (>10^9 transformants) is required to start the panning with (Muyldermans 2013). Finally, sybodies with 500 pM – 500 nM affinity can be obtained (Zimmermann, Egloff et al. 2020).

1.3.3.2 Biopanning

There are many methods to select nanobody from libraries, such as phage display, yeast/bacteria surface display, mRNA/cDNA display, ribosome display, high-throughput DNA sequencing with mass spectrometry, biotinylated nanobody screening and bacteria two-hybrid system. The most frequently used methods to select nanobodies is phage display (Smith and Petrenko 1997).

Phage display initiated and proceeded rapidly in the late 1980s right after the development of molecular biology and the successful construction of antibody libraries (Smith and Petrenko 1997). Nowadays, binders with nano-molar or even pico-molar affinity can be generated within months (Zanganeh, Nejad et al. 2019). The well-established panning procedure for phage display is described in Figure 1.14.

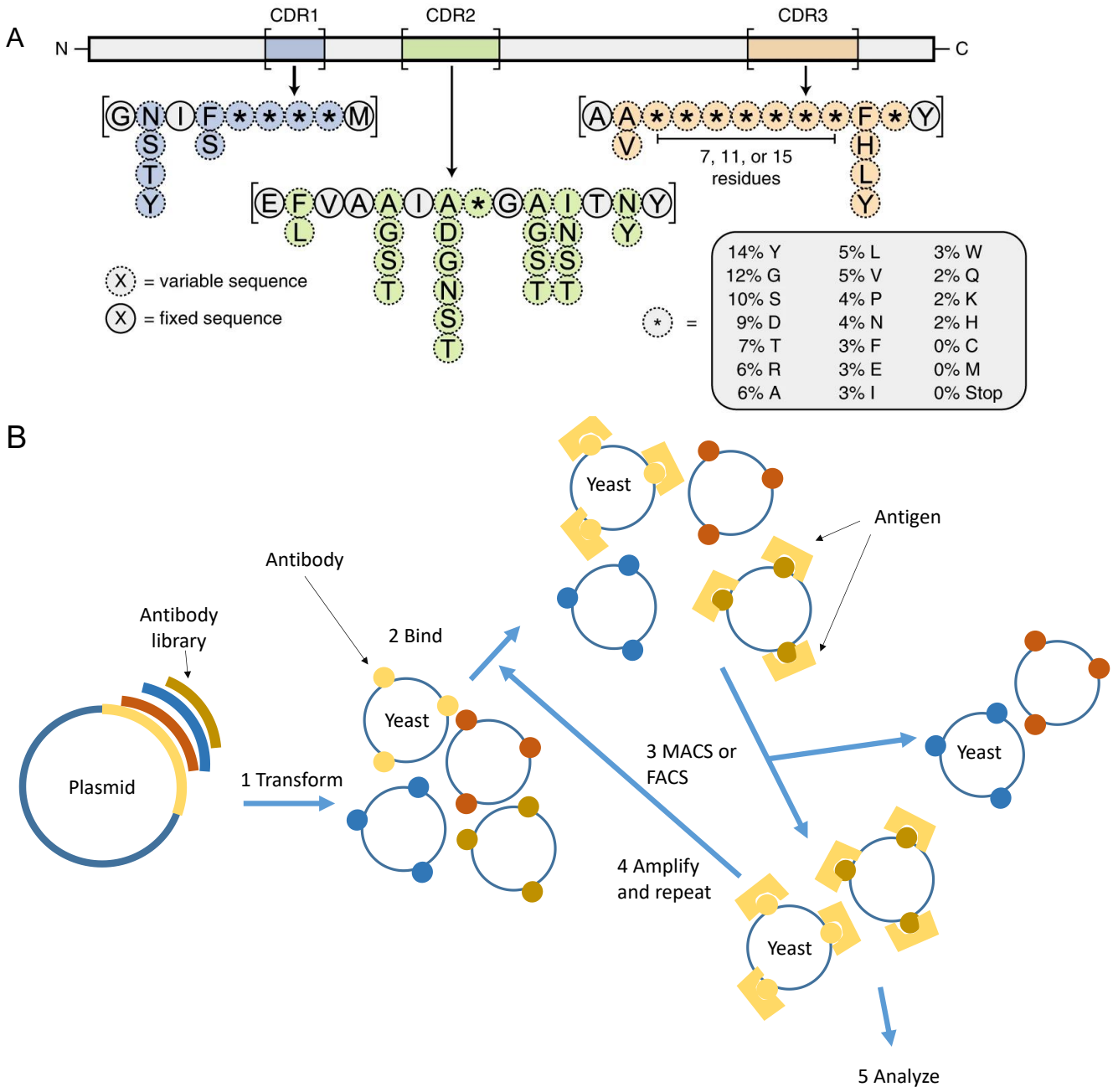


Figure 1.15 Sybody libraries and the pipeline of yeast surface display. A) The construction of sybody libraries (adapted from (Zimmermann, Egloff et al. 2020)). B) There are five steps in this pipeline, some of which may be repeated several times. In the first step, plasmids containing the antibody library are transformed into the yeast, where the yeast expresses the antibodies and displays them on the surface. In the second step, antigens are thoroughly mixed with the yeast cells and bind to specific antibodies. In the third step, Magnetic-activated cell sorting (MACS) or Fluorescence-activated cell sorting (FACS) is used to sort for the cells that bind the antigens, and in the meantime, unbound cells are discarded. Those selected cells will amplify again and repeat the second and third steps for several cycles until the affinity is satisfied. Finally in the step 5, the selected yeast cells are analyzed and sequenced for the antibodies of interest.

Although phage display has been developed quite thoroughly over the decades, the intrinsic property of its production procedure determines that the cycle could not be shortened (Ledsgaard, Kilstrup et al. 2018), which hinders the ambition to use antibody for selection of different conformations of a protein and discovery of its working mechanism (Pardon, Laeremans et al. 2014). In addition, professional facilities and considerable amount of work required for antibody production also restricts the access of small labs to necessary binders. Therefore, the long production cycle, facilities and labour required in the production are the main limitations of phage display as a binder selection method in terms of structural biology studies.

More recently, a yeast display platform has come into use. Together with the use of synthetic nanobody libraries (McMahon, Baier et al. 2018, Zimmermann, Egloff et al. 2020), the procedure bypasses immunization of Llama, which is a standard and long process in traditional nanobody production protocol. With this advantage as well as using yeast rather than phage as a display vector, any well-equipped molecular biology lab can complete nanobody production within three weeks using minimum facility (McMahon, Baier et al. 2018, Uchański, Zögg et al. 2019). The detailed procedure of yeast surface display is described in Figure 1.15B.

1.3.3.3 Affinity validation

The validation of affinity can be performed by various techniques, either biochemically or biophysically. Most commonly used technique is enzyme-linked immunosorbent assay (ELISA) (Figure 1.14) while the microscale thermophoresis (MST) can also be made use of. This will be detailed later in section 1.5 describing the techniques to detect and quantify protein-protein interactions. For structural biology purposes, the nanobodies are uaually proved useful if they co-migrate with the target protein in size exclusion chromatography.

1.3.4 Application in crystallography and cryo-EM

1.3.4.1 Crystallography

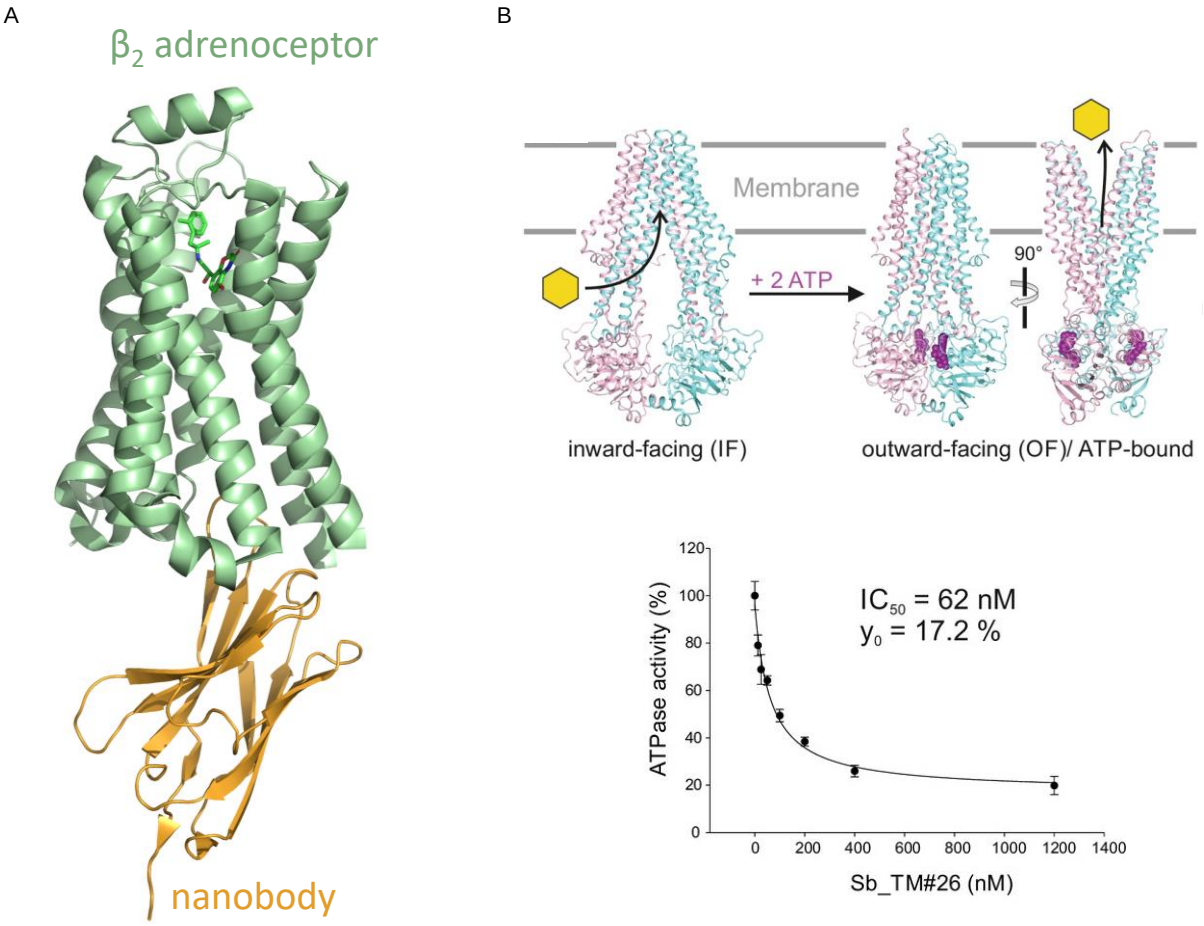


Figure 1.6 The nanobody locks flexible proteins and helps with the crystal structure solution. A) Structure of the β_2 adrenoceptor stabilized by the nanobody in the active state remade from 3P0G in PDB. The β_2 adrenoceptor is shown in green and the nanobody is shown in orange. The nanobody acts as surrogate of the G protein to trap the β_2 adrenoceptor in the active state. B) The synthetic nanobodies trap the ABC transporter in the outward-facing state and prevent it from hydrolyzing ATP. The two states of the ABC transporters are depicted in the top panel as inward-facing (IF) and outward-facing (OF) states. The OF-state and the ATP-bound state share the conformation. The sybody selected against the E517A mutant has high affinity and the binding curve is shown in the bottom panel. The figures are adapted from (Zimmermann, Egloff et al. 2018).

In crystallography, nanobodies are used as crystallization chaperones to lock the flexible parts of the target protein into certain conformations (Rasmussen, Choi et al. 2011, Steyaert and Kobilka 2011, Kruse, Ring et al. 2013, Ring, Manglik et al. 2013, Burg, Ingram et al. 2015, Huang, Manglik et al. 2015, Manglik, Kobilka et al. 2017, Che, Majumdar et al. 2018, Koehl, Hu et al. 2019, Warne, Edwards et al. 2019, Uchański, Pardon et al. 2020). A profound example would be the Nobel Prize winning work by Brian Kobilka, who stabilized the GPCR into the active state with the nanobody as a surrogate and solved the crystal structure of the membrane protein (Figure 1.16A) (Rasmussen, Choi et al. 2011). Different nanobodies/synthetic nanobodies also have a chance to trap the target protein into different conformations, so that the functioning mechanism of the protein can be depicted more clearly with different states captured (Figure 1.16B) (Zimmermann, Egloff et al. 2018, Uchański, Pardon et al. 2020, Zimmermann, Egloff et al. 2020). For example, the outward-facing state is transient and difficult to capture and therefore inward-facing states are present in most structures of ABC transporters. In the case of bacteria ABC transporter TM287/288, the E517A mutant blocks ATP hydrolysis and the outward-facing state can be maintained (Zimmermann, Egloff et al. 2018). Synthetic nanobody selection was performed against the E517A mutant and as a result the selected nanobodies were demonstrated to trap the wild-type transporter in the outward-facing state and prevent the ATP hydrolysis.

1.3.4.2 Cryo-EM

In cryo-EM, the nanobody is also frequently used as an ancillary tool for the image alignment, especially for small membrane proteins. Especially in recent years, it is much clearer that nanobodies can play essential roles in solving high-resolution cryo-EM structures of small proteins, and various nanobody derivatives have been developed as well.

A single 15 kDa nanobody was used to help solve a ~50 kDa membrane protein EM structure at near atomic resolution (Figure 1.17), demonstrating it’s highly helpful (Parker, Deme et al. 2021). This is a milestone work, since there was no work before demonstrating a single nanobody could help solve high-resolution structure of small proteins using cryo-EM. With this pioneer work, as the 2D classification of the PCFT:nanobody complex shows (Figure 1.17B), it is clearer that the role of the EM chaperone is not only

increasing the size of the particle, but also help with the alignment of the protein, the key to high resolution.

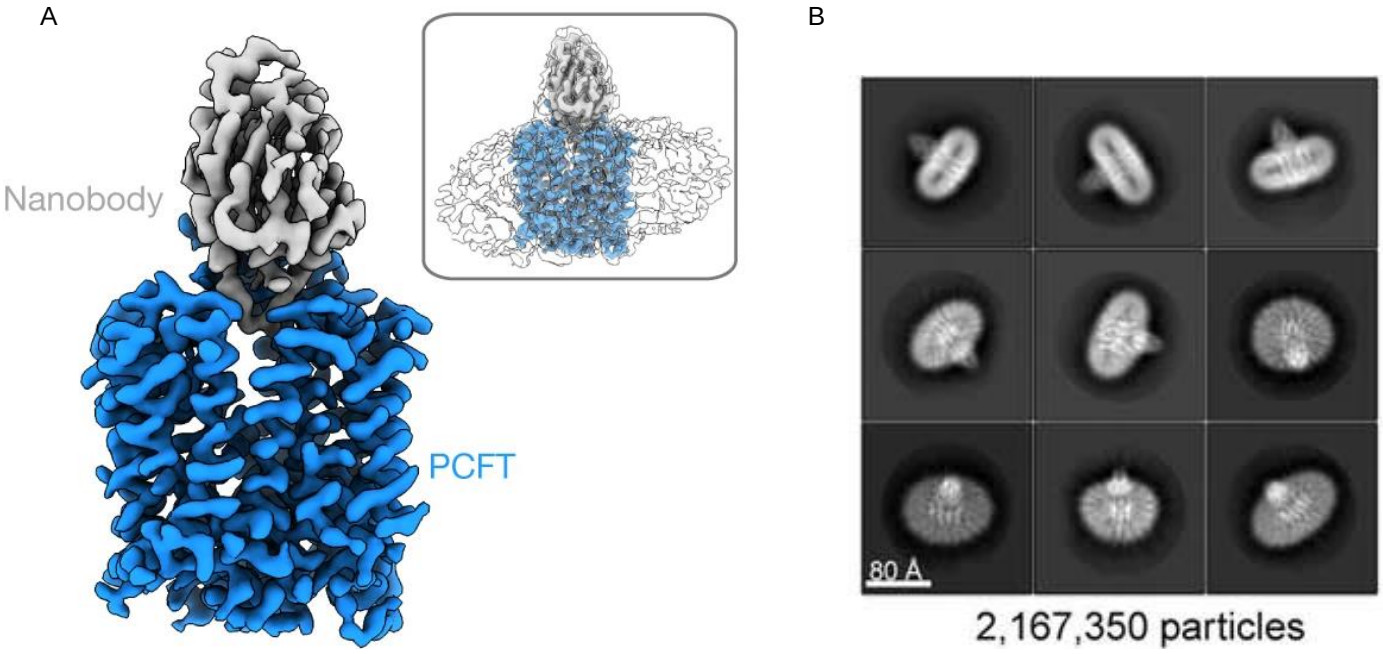


Figure 1.17 Nanobody helps solve the 3.18Å cryo-EM structure of PCFT by minimizing misalignment on detergent micelles. The figures are adapted from (Parker, Deme et al. 2021). A) Cartoon schematic of the PCFT (blue) in comeplex with the nanobody (grey). In the smaller graph on the top right, the detergent micelle around the protein is depicted. B) The 2-D classification result of the PCFT:nanobody complex. The nanobody density is clear protruding out from the micelle, helping with the alignment of the images, which is believed to be one factor for the successful structure solution. The different orientations that the complex particles take are another factor for the high-resolution structure solution.

An advanced nanobody tool 'Legobody' was also invented to further leverage chaperoning ability of the nanobody (Wu and Rapoport 2021). The Legobody is comprised of a nanobody, a secondary antibody (Fab) binding to the nanobody, and a third artificial binding protein (converted from MBP) binding to the nanobody and the Fab (Figure 1.18A). This creates a ~120kDa ensemble more visible in the micrographs. When purifying the Legobody:target complex, there was only a single peak appearing on the SEC profile (Figure 1.18B), indicating the tight binding within the Legobody. The clear shape of the Legobody in the 2D classification results (Figure 1.18C) and the final resolution around 3.5 Å of the refined Legobody structures further demonstrate that the Legobody is a rigid ensemble, which is suitable as an imaging scaffold in cryo-EM.

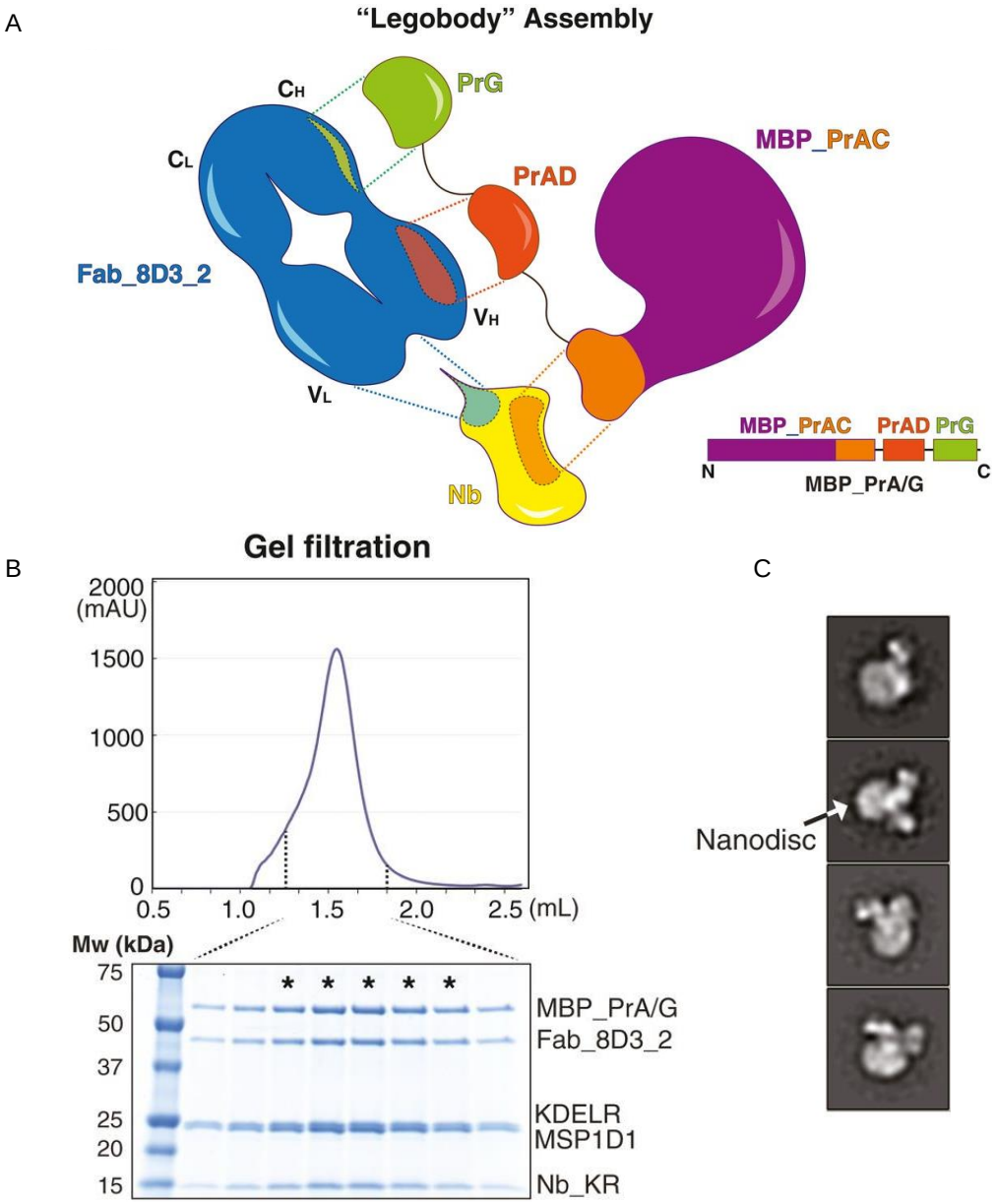


Figure 1.18 Legobodies help with the structure solution of small proteins. The figures are adapted from (Wu and Rapoport 2021). A) The Legobody assembly. There are three components in this assembly, namely the nanobody (yellow), Fab (blue) and a binding protein converted from MBP (magenta). The component directly binding to the antigen is the nanobody, and the rest of the assembly is designed to enlarge the whole assembly. Fab binds to the opposite of CDR loops of the nanobody through its variable domains. The second scaffold is a fusion of MBP and some nanobody/Fab binding domains of protein A. Interfaces are shown in dashed lines on the schematic. B) Purification of Legobody in complex with KDEL. C) 2-D classification of KDEL in complex with Legobody. KDEL is in the nanodisc and the Legobody protruding out of the nanodisc density helps with the alignment of the images.

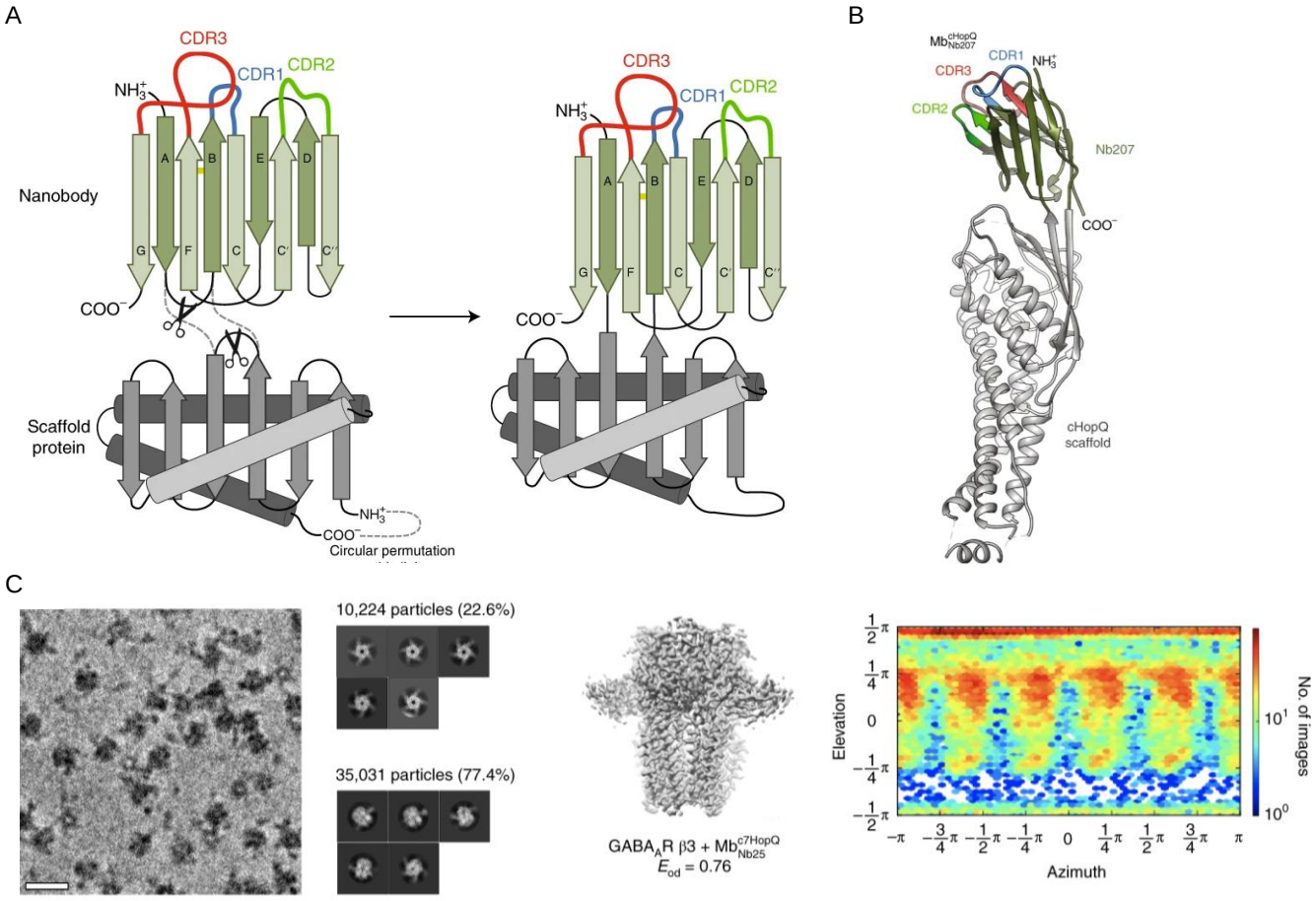


Figure 1.19 The megabody helps solve the high-resolution cryo-EM structure of GABA receptor. Figures are adapted from (Uchański, Masiulis et al. 2021). A) The design of the megabody. The megabody is comprised of a nanobody and HopQ or YgjK. The HopQ or YgjK is fused to the nanobody via circular permutation. B) The cartoon structure of the HopQ megabody. C) The micrograph, 2-D classification, 3-D reconstruction of the GABA receptor in complex with the HopQ megabody and the particle orientation distribution of the constructed map. Phase plate is used to increase the contrast of the particle shown in the micrograph. The 2-D classification result suggests that the HopQ fusion is still flexible due to its smeared density, but this does not affect the alignment of the particles and instead the nanobody helps with the alignment. The 3-D reconstruction shows clear density of the nanobody and no density for the HopQ, indicating the rigid binding of the nanobody to GABA, but a flexible fusion of HopQ. The orientation distribution shows that the megabody will not result in serious orientation preference.

Another nanobody-derived binder called ‘Megabody’ was also reported useful in helping solve high-resolution EM structures of membrane proteins (Figure 1.19) (Laverty, Desai et al. 2019, Uchański, Masiulis et al. 2021). The megabody is comprised of a nanobody and HopQ or YgjK and the HopQ or YgjK is fused to the nanobody via circular permutation. The megabody crystal structures were obtained to showcase the rigidity of the linker (PDB id 6QD6). However, only the nanobody part of the Megabody was visible in the 3D reconstruction of GABA receptor:Megabody complex (Figure 1.19C), showing that its rigidity is not as good as Legobody, which means that the functioning part of the Megabody might just be the nanobody alone. Although the paper claimed that the Megabody functioned to increase the size of the complex, the size of GABA receptor was never a problem since it formed a large pentamer. It was maybe the nanobody that resulted in adequate alignment of the particles, leading to the high resolution structure solution. Therefore, it is still questionable whether it is worth making the Megabody fusion upon the nanobody.

1.4 Crystallization mechanisms

Protein crystallization has been a mysterious process since it was first discovered useful in solving protein structures (Ulrich and Stelzer 2011). Therefore, we still don’t know a systematic approach to make any protein crystallize. Nevertheless, it is still helpful to understand the effect of certain chemicals (such as PEG) on protein crystallization, the effect of surface mutations on the nucleation processes of crystallization, and important protein surface properties that drive crystallization. Here I will discuss the PEG depletion mechanism, non-classical nucleation theory and the factors that drive crystallization based on statistical analysis.

1.4.1 The depletion mechanism by polyethylene glycol (PEG)

PEG is a widely used water-soluble synthetic polymer in crystallization screens at present (Hui and Edwards 2003), and it has been shown that crystals grown in PEG conditions are generally better diffracting than those from high salt conditions (McPherson 1976).Scientists found PEG was able to induce protein precipitation or crystallization around 1980 (McPherson 1976), and then started biochemical and biophysical studies on specific model proteins such as the bovine pancreatic ribonuclease, human albumin and hen egg lysozyme (Atha and Ingham 1981, Kulkarni and Zukoski 2001, Bončina, Reščič et al. 2008). The solubility

studies on different PEG-protein ratios suggested the exclusion of proteins by PEG could not be explained by simple excluded volume theory (Atha and Ingham 1981). Then studies on apoferritin using dynamic light scattering discovered that PEG could significantly induce inter-particle attraction and thus triggered protein crystallization, phase separation or aggregation (Tanaka and Ataka 2002). Therefore, the depletion mechanism by polymers (Asakura and Oosawa 1958, Gast, Hall et al. 1983, Lekkerkerker, Poon et al. 1992, Ilett, Orrock et al. 1995, Asherie, Lomakin et al. 1996, Evans and Poon 1997, Lekkerkerker 1997, Poon, Pirie et al. 1997) can also explain PEG-induced protein crystallization.

1.4.2 Classical nucleation theory (CNT) versus non-classical nucleation theory

In the field of kinetics studies of protein crystallization, there are two main theories on the nucleation of crystals (Erdemir, Lee et al. 2009) – classical and non-classical (two-step) nucleation theories. The CNT claims that protein molecules first reach a metastable phase in the solution directly before they form nucleus, from which crystals grow (Maier, Zocher et al. 2020). Whereas the non-CNT claims there are pre-nucleus clusters forming before nucleus is formed (Gebauer and Cölfen 2011) (Figure 1.20A).

The evidence for non-classical nucleation theory supports that protein crystals can form via different pathways in different conditions. This further provides the reason for creating surface mutants when trying to crystallize the protein of interest, which has already been frequently practiced throughout the past decade (Derewenda 2004, Price Ii, Chen et al. 2009).

In the recent two decades, with the development of advanced detecting methods, such as small-angle X-ray scattering (SAXS) (Sauter, Roosen-Runge et al. 2015) (Figure 1.20B), atomic force microscopy (AFM) (Yau and Vekilov 2000, Chung, Shin et al. 2010, Sleutel, Lutsko et al. 2014) (Figure 1.20C) and transmission electron microscopy (TEM) (Van Driessche, Van Gerven et al. 2018) (Figure 1.21), various types of pre-nucleus clusters were discovered. Pre-nucleus clusters include ordered clusters, such as nanorods, and most other clusters are disordered, like bundles (Van Driessche, Van Gerven et al. 2018). In the crystallization of the same protein, different pre-nucleus clusters can be found when crystals form in different conditions, and the different pathways finally lead to different crystal polymorphs. Additionally, mutations of key residues in the crystal contacts of one crystal polymorph can prevent the crystals forming in relevant conditions with the other crystal polymorph unaffected (Van Driessche, Van Gerven et al. 2018).

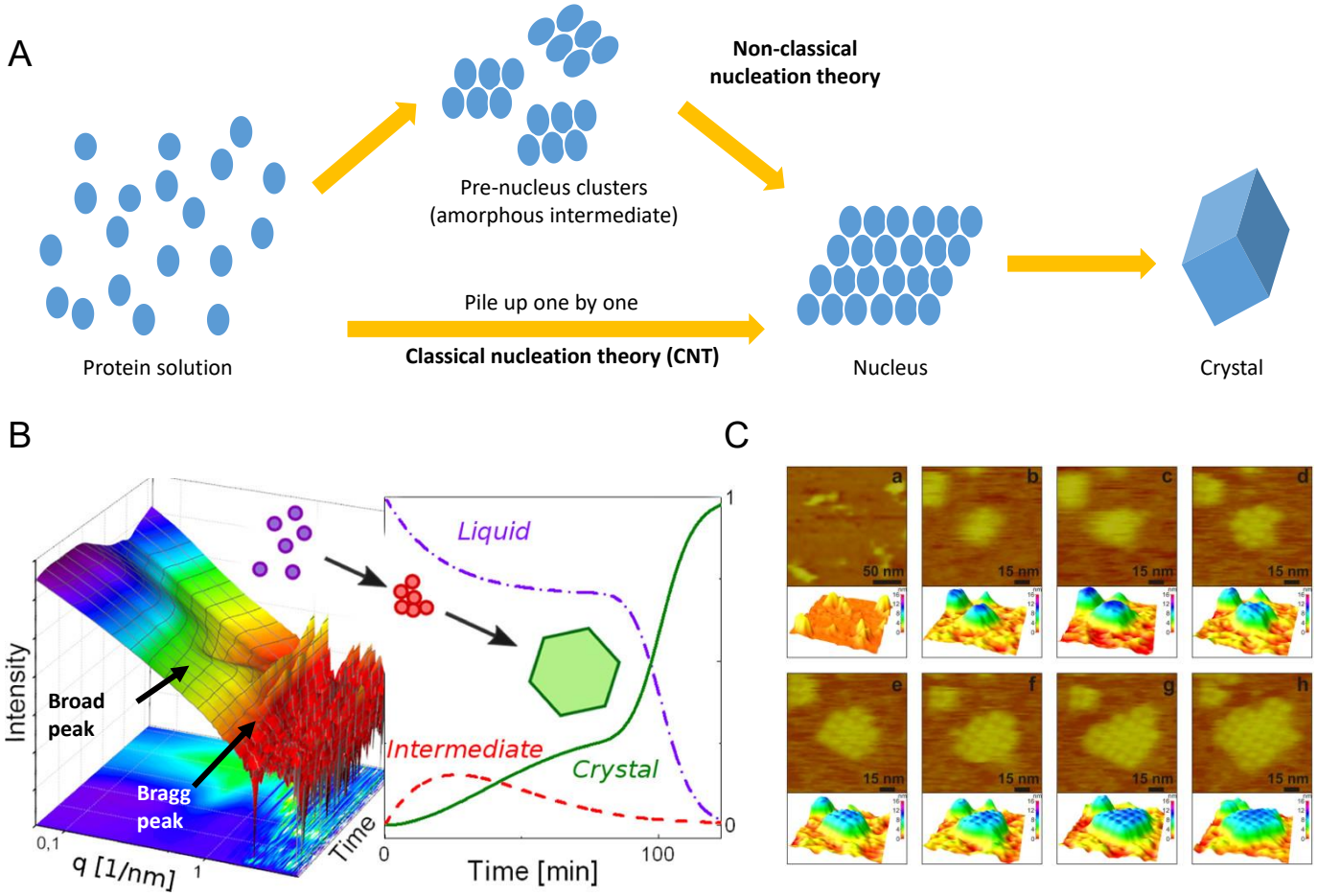


Figure 1.20 Classical and non-classical nucleation theories and evidences for the latter. A) Classical nucleation theory (CNT) claims the nucleus forms directly from a metastable protein solution with the molecules piling up one by one. On the other hand, the non-classical nucleation theory claims there should be an amorphous intermediate stage (formation of pre-nucleus clusters) before nucleus formation. B) Evidence for non-classical nucleation theory by SAXS. Before the nucleation indicated by the appearance of the Bragg peaks, there is a broad peak forming, indicating a new structural feature. After Bragg peaks increase, the broad peaks shrink, indicating the intermediate is gradually converted to the nucleus. The figure is taken from (Sauter, Roosen-Runge et al. 2015). C) Evidence for non-classical nucleation theory by AFM. Small particles first aggregate into a larger cluster with un-ordered arrangement, and then the cluster is gradually relaxed into the lattice. The figure is taken from (Chung, Shin et al. 2010).

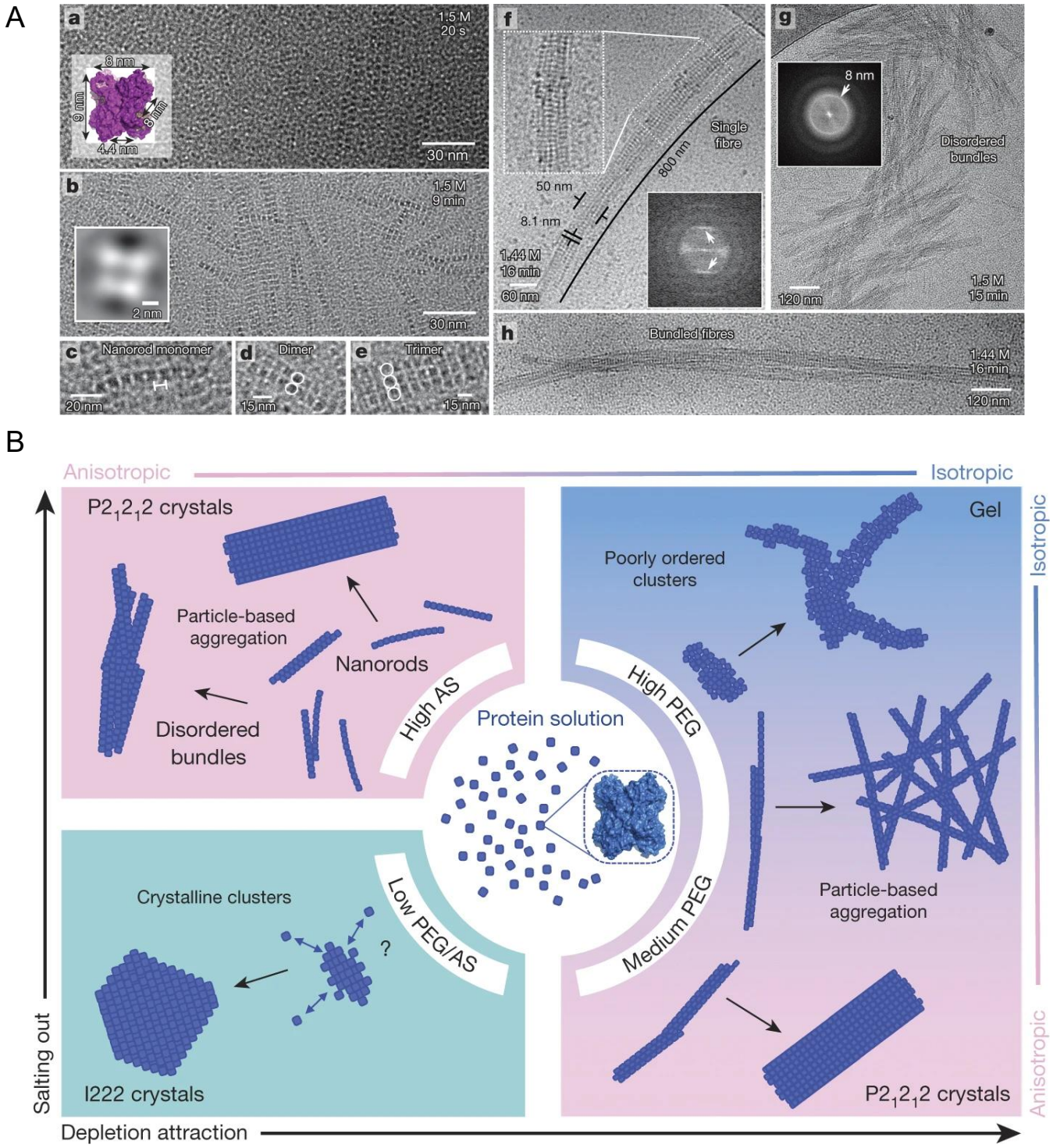


Figure 1.21 Evidence for non-classical nucleation theory by cryo-TEM. The figure is taken from (Van Driessche, Van Gerven et al. 2018). A) cryo-TEM images captured nanorod intermediates during the nucleation of glucose isomerase. The nanorods then formed fibre and bundles before forming $P2_12_12$ crystals. B) Glucose isomerase takes different pathways to form crystals in different conditions.

1.4.3 Key surface factors driving protein crystallization

Although it is hard to understand the protein crystallization mechanism at present, large numbers of crystal structures have accumulated since four decades ago and from statistics some conclusions have emerged. In a study carried out by the Northeast consortium of Structural Genomics (NESG), four properties were determined to significantly affect protein crystallization (Figure 1.22), i.e., the mean side chain entropy (SCE) and grand average of hydropathy (GRAVY), fraction of predicted buried glycines and fraction of

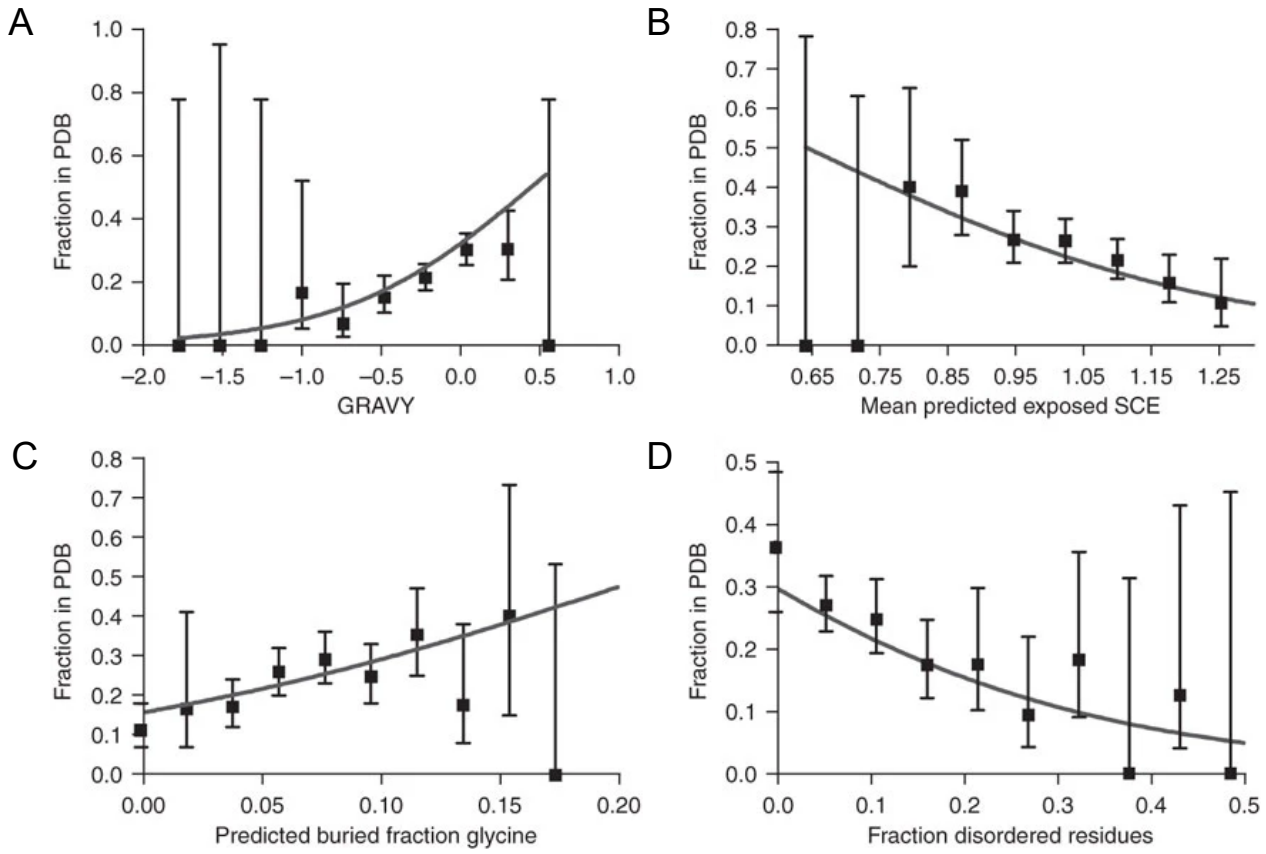


Figure 1.22 Statistical analysis applied into the crystallogenesis prediction. 679 proteins whose structures were successfully determined by NESG are analysed in these graphs. Error bars indicate 95% confidence limits. Fractions of the 679 proteins are plotted against four parameters respectively in these four plots, A) GRAVY, B) SCE by PHD/PROF, C) fraction of predicted buried glycine by PHD/PROF and D) fraction of predicted disordered residues by DISOPRED2. The grey lines are optimized logistic regressions of the data points. The up or down trend of the regression line represents a positive or negative predictor. The figure is taken from (Price Ii, Chen et al. 2009).

predicted disordered residues (Price Ii, Chen et al. 2009, Fusco, Barnum et al. 2014). It provides the rationale of protein surface engineering from the statistics point of view and further demonstrates the important effect of low side chain entropy and high hydropathy on crystallogenesis.

1.5 Methods to detect and quantify protein-protein interactions

The crystal packing involves protein-protein interactions (Janin and Rodier 1995). It would be easier to understand how crystal packing happened if the detailed inter-molecule interactions were detected and measured throughout the whole crystallization process. Unfortunately, since the crystallization mechanisms are not understood thoroughly, there have been no methods to detect and quantify protein-protien interactions in any specific steps of crystallization. There are even no reports on measuring the inter-molecule interactions *in-crystallo*. However, there are still inspirations from the various biochemical and biophysical methods to detect protein-protein interactions. Here I summarize the representative methods to characterise, and then discuss their potential applications in crystallography.

1.5.1 Biochemical – ELISA and yeast two hybrid

The enzyme-linked immunosorbent assay (ELISA) was developed around 50 years ago (Coons, Creech et al. 1941, Yalow and Berson 1960, Miles and Hales 1968, Engvall and Perlmann 1971, Van Weemen and Schuurs 1971). This assay can detect the specific interaction between two proteins. Of all types of ELISA, sandwich ELISA has the highest sensitivity, which can be as high as 39 pg/ml (Dixit, Vashist et al. 2010). This *in vitro* assay is easy to use with only a few steps – fixing the antigen, primary antibody incubation, secondary antibody incubation and the coloration reaction (Kato, Hamaguchi et al. 1977) (Figure 1.23A). The small quantities of materials required also allow high-throughput screening (Mendoza, McQuary et al. 1999).

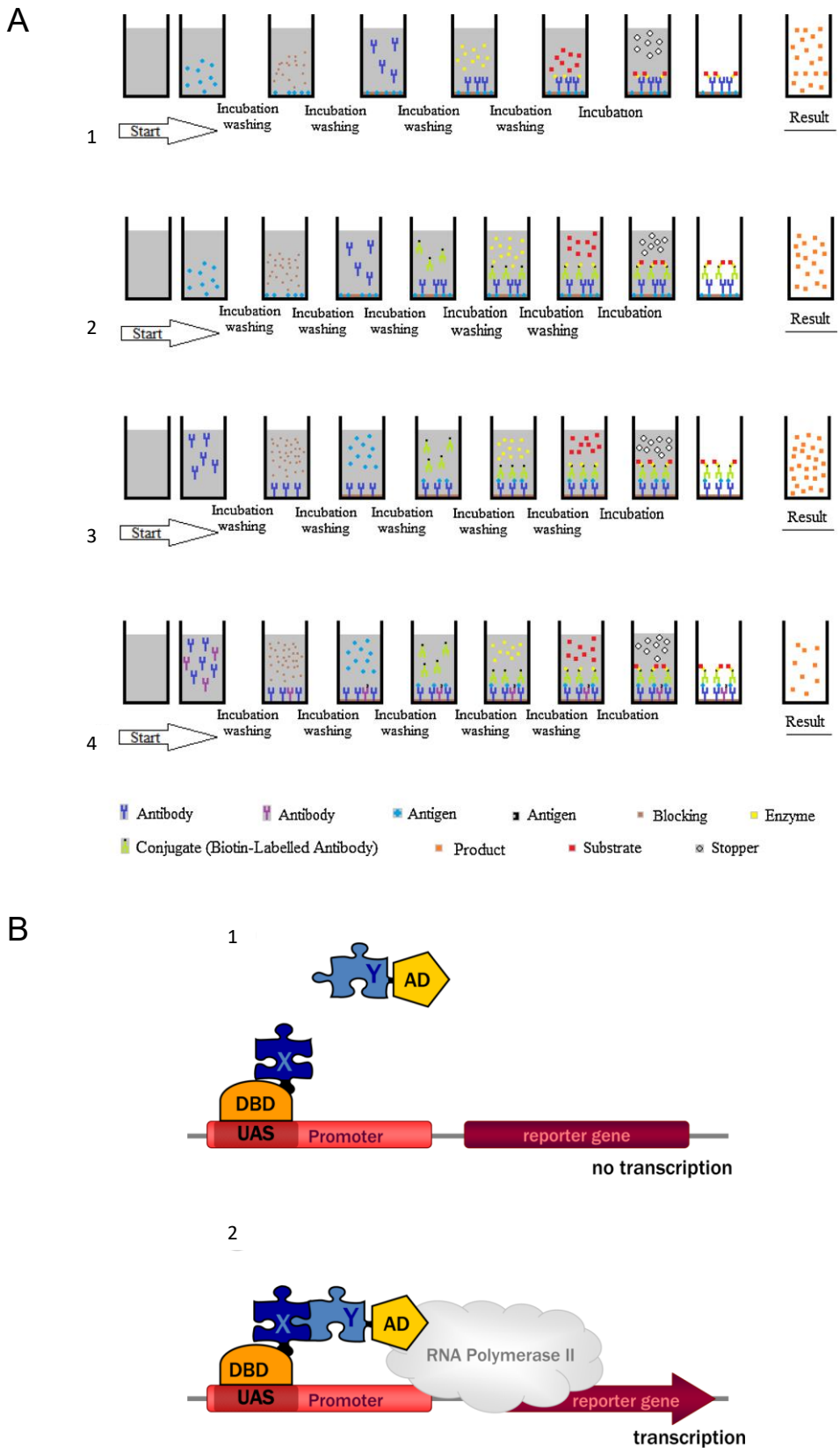


Figure 1.23 Biochemical methods to quantify protein-protein interactions.

Figure 1.23 (Continued) A) Types of ELISA and their workflow. 1-4 are direct, indirect, sandwich and competitive ELISA methods respectively. The figure is taken from (Aydin 2015). B) The mechanism of yeast two hybrid system. Protein X is fused with DNA binding domain (DBD) and protein Y is fused with activation domain (AD). When X and Y does not interact, as shown in 1, the reporter gene does not transcribe. Whereas when X interacts with Y and AD is brought together with DBD, the downstream transcription is triggered. The figure is taken from (Brückner, Polge et al. 2009).

The yeast two hybrid system (Y2H) detects whether two proteins interact by using the yeast transcription system (Fields and Song 1989). If two proteins interact, the DNA binding domain and activation domain of the yeast transcription factor fused respectively to the two proteins will re-assemble and initiate downstream expression of the reporter gene (Chien, Bartel et al. 1991) (Figure 1.23B).

1.5.2 Biophysical – ITC, SPR, MST, FRET, BRET and DLS

Isothermal titration calorimetry (ITC) is a method to measure protein-protein interaction by sensitively detecting the changes in thermal dynamic potentials, such as ΔH, ΔG and ΔS (Freire, Mayorga et al. 1990, Doyle 1997, Jelesarov and Bosshard 1999, Leavitt and Freire 2001). When two different molecules mix in the calorimetric cell, the dissipated or absorbed heat is detected (Figure 1.24A). For self-association proteins, the dissociation constant can also be obtained by measuring the heat in a dilution process happening in the calorimetric cell (Myszka, Sweet et al. 2000).

Surface plasmon resonance (SPR) measures the protein-protein interactions by measuring the change of refractive light on the thin layers of materials fixed on the metal surface (Hou and Cronin 2013). In this method, one protein is fixed onto the metal surface, a beam continuously hits the metal surface, and the refractive light is meanwhile detected by a charged couple device (CCD) chip (Homola, Yee et al. 1999) (Figure 1.24B). When the other protein or ligand flows by the metal surface and interacts with the fixed protein, the refractive light pattern changes and thus the interaction is measured. The sensitivity of SPR can be up to sub nanomolar range (Myszka, Jonsen et al. 1998).

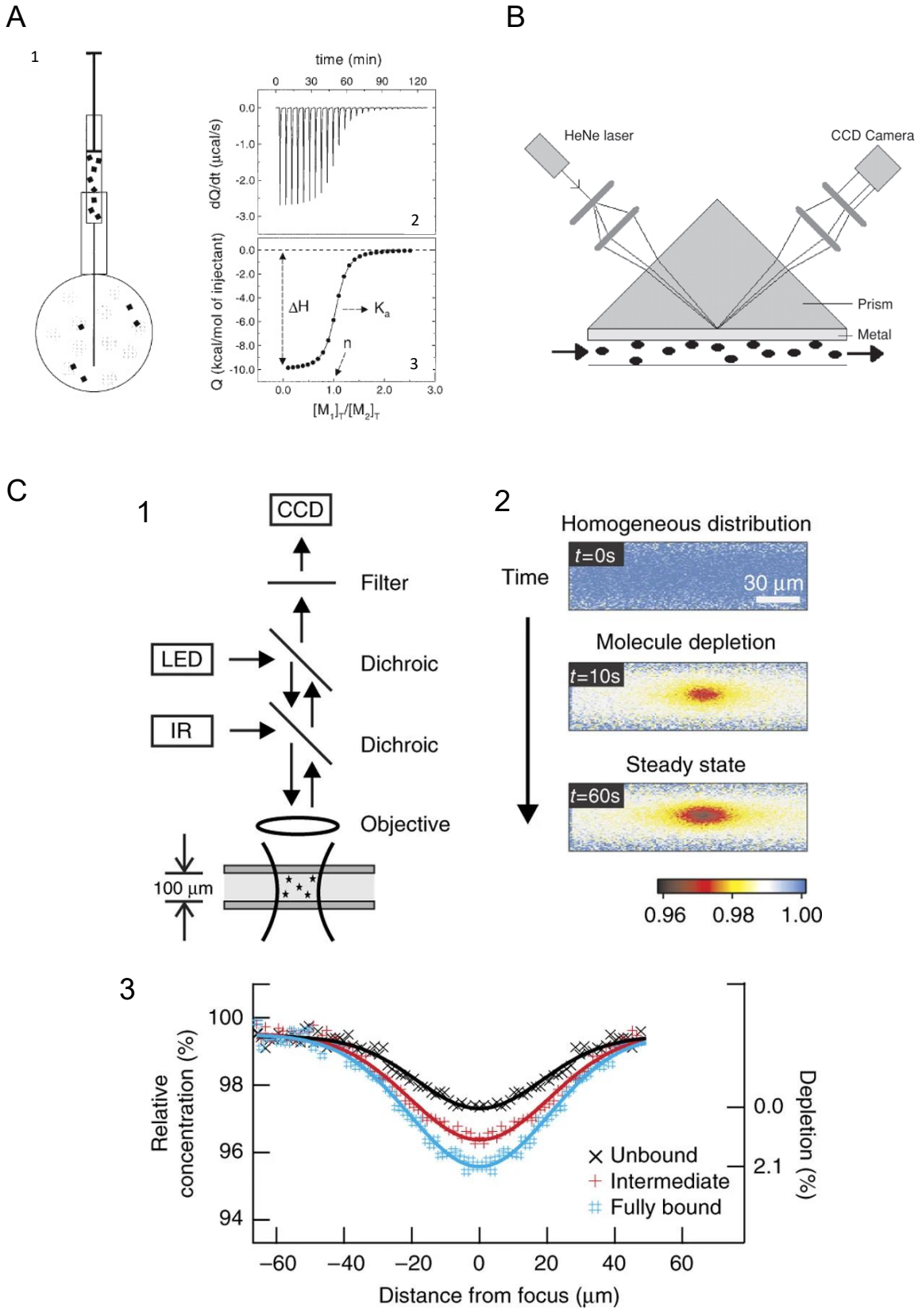


Figure 1.24 ITC, SPR and MST techniques.

Figure 1.24 (Continued) A) The experimental setup of ITC. As shown on the left (1), the reaction happens in the small cell of around 1.4ml and the cell is filled with the protein solution of interest (grey). The ligand or protein solution is in the injection syringe (black). A small volume (10ul) of the syringe solution is repeated injected into the cell at an interval of 400s, and thus a characteristic peak sequence curve is obtained for the reaction (2). Then a curve as an integration of the heat pulse data fitted to the standard binding model is obtained, from where n, K_a, and ΔH is estimated. The figure is taken from (Doyle 1997). B) The experimental setup of SPR. The figure is taken from (Pattnaik 2005). The beam shoots through a prism onto the metal surface where the antigen is fixed and reflects into a CCD camera for recording. The test solution flows by the surface of the metal. The binding on the surface-fix antigen changes the reflected signal. C) The experimental setup of MST. The figure is taken from (Wienken, Baaske et al. 2010). 1) In the epifluorescence microscope, there are two dichroic mirrors. The first above the objective lens is to reflect the infrared-laser so that it can be focused into the silica capillary and introduce temperature gradient. The second dichroic mirror is to reflect the LED light into the capillary so that the fluorescence can be imaged by the CCD camera. 2) Fluorescence images form after the introduction of temperature gradient. After a certain period of time (60s), the molecules come to a steady state. 3) The profile of the steady state is plotted as a function of relative concentration vs location. The profiles of human interferon-γ bound with various concentrations of binding antibodies are plotted here.

Microscale thermophoresis (MST) is based on the observation that molecules tend to migrate along a temperature gradient (Jerabek-Willemsen, Wienken et al. 2011). In this method, one protein is labelled with a dye, and its migration along temperature gradient induced by infrared light is recorded (Wienken, Baaske et al. 2010). When the other interacting protein or ligand is added to the system, the migration pattern changes accordingly, so that the dissociation constant can be determined (Figure 1.24C). The sensitivity of this method can be up to 1 picomolar (Bujalowski and Lohman 1989, Curth, Urbanke et al. 1994).

Fluorescence resonance energy transfer (FRET) and bioluminescence resonance energy transfer (BRET) both detect interactions between two proteins using the resonance energy transfer technique (Förster 1948, Morin and Hastings 1971). Two proteins are fused with either the donor fluorophore or the acceptor fluorophore, and if two proteins interact, the donor and acceptor are brought in proximity and the energy transfer takes place, so that a final excitation signal from the acceptor fluorophore can be detected. The difference of the two methods is that in FRET the donor is green fluorescent protein (GFP), which needs exogenous excitation (Berney and Danuser 2003), whereas in BRET the donor is luciferase, which illuminates

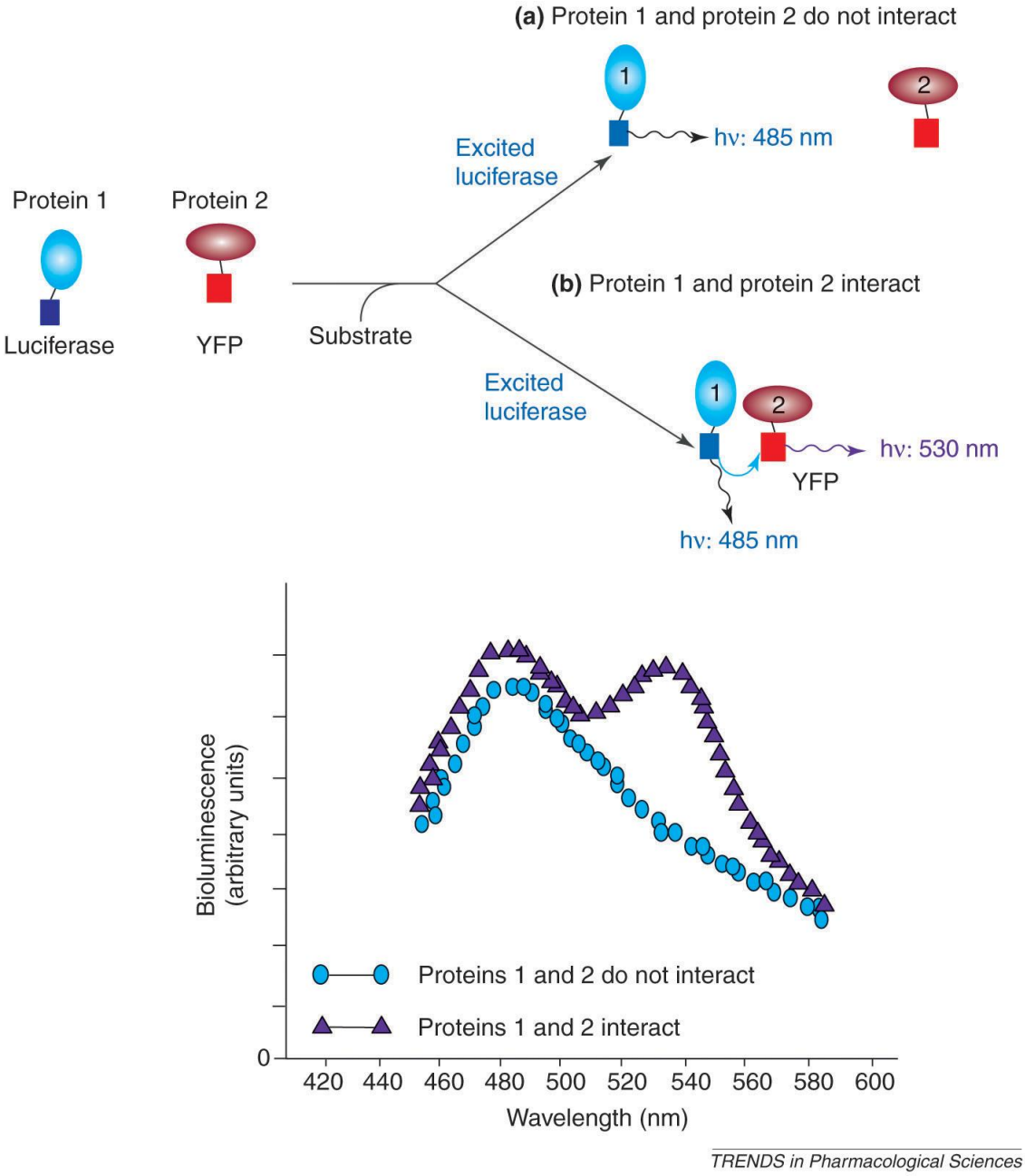


Figure 1.25 Principle of the BRET technology. The figure is taken from (Boute, Jockers et al. 2002). To test the protein-protein interaction between protein 1 and protein 2, protein 1 is fused with luciferase and protein 2 is fused with yellow fluorescent protein (YFP). In any case, luciferase emits light at a wavelength of 485nm. If protein 1 interacts with protein 2, luciferase and YFP are brought into close proximity, and then the emission light from luciferase will excite YFP and finally an emission light at 530 nm will be detected. Otherwise, only an emission at 485nm is detected when YFP is not excited. The bottom graph shows the different results of the two situations.

at certain wavelength itself (Figure 1.25). Therefore, BRET is more commonly used than FRET, since the exogenous excitation can also directly excite the acceptor fluorophore, which often makes the result un-interpretable (Tsien 1998, Boute, Jockers et al. 2002).

The dynamic light scattering (DLS) is often used to determine the diffusion coefficient of a molecule in solution (Cummins, Knable et al. 1964, Pecora 1972, Pike 1972, Bloomfield and Lim 1978). This can also be used in protein-protein interaction studies (Patel, Bernards et al. 2014) (Figure 1.26). If two proteins form a complex, the hydrodynamic radius of the complex should be significantly different from the either of the two proteins alone. Furthermore, the DLS can also measure the autocorrelation function of scattered light intensity, from which the inter-particle interaction potentials can be calculated (Tanaka and Ataka 2002).

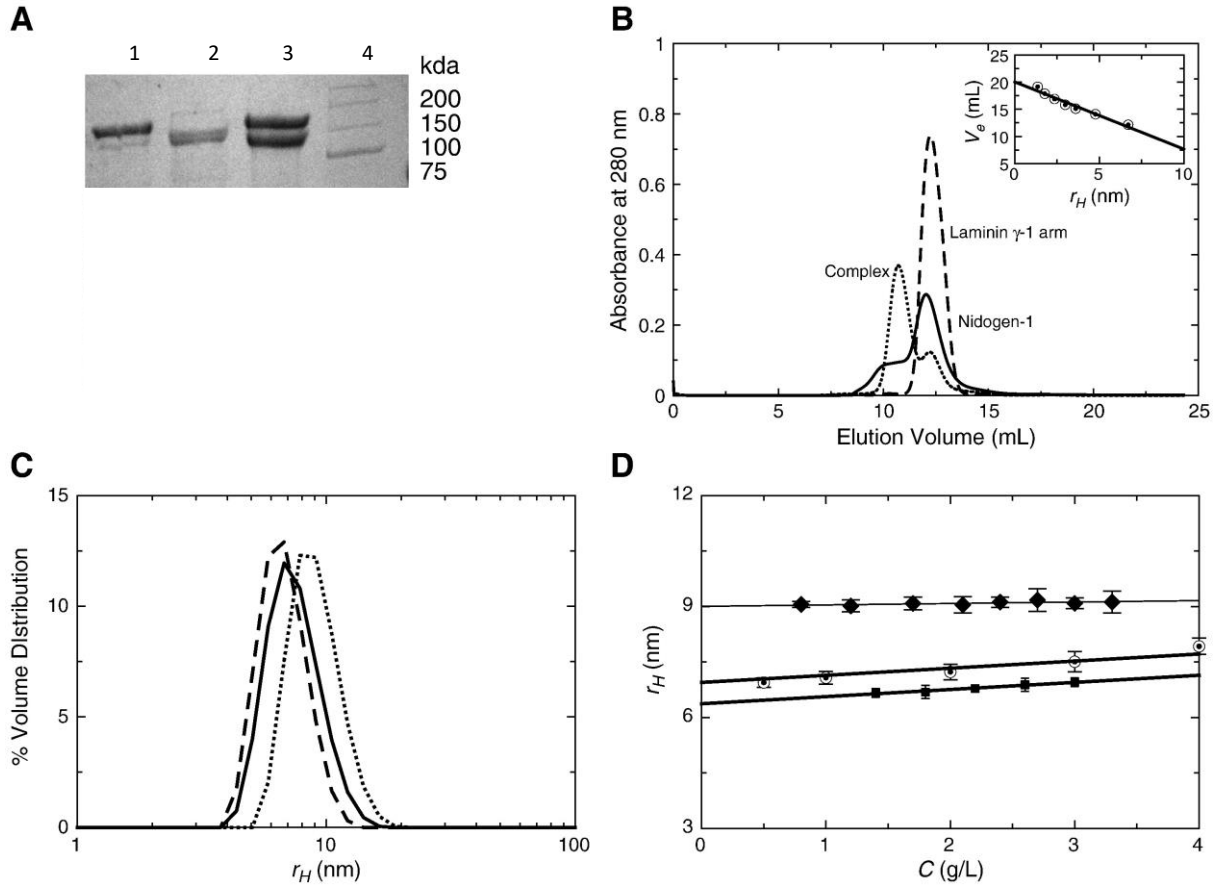


Figure 1.26 The use of dynamic light scattering to identify protein-protein interaction. The figure is adapted from (Patel, Bernards et al. 2014). A) The SDS-PAGE of nidogen-1 (1), laminin γ-1 short arm (2) and the complex (3). B) Size exclusion chromatography profiles of nidogen-1, laminin γ-1 short arm and the complex.

Figure 1.26 (Continued) C) DLS profiles for nidogen-1 (-----), laminin γ-1 short arm (– – –) and their complex (……) at 3.0mg/ml. The radiuses are clearly different between the complex and individual proteins, indicating the interaction between nidogen-1 and laminin γ-1 short arm. D) The mean hydrated radius (r_H) as a function of protein concentration. ○ for nidogen-1, ■ for laminin γ-1 short arm and ♦ for the complex.

1.5.3 The application in protein crystallography

Protein crystallization involves several stages, such as the pre-nucleus stage and the crystal growth stage (Durbin and Feher 1996).

For the crystal growth stage, it seems quite difficult to measure the inter-molecular interactions *in crystallo* by using the above techniques. It is feasible, however, to identify specific side chain interactions after the crystal structure being solved and to calculate the inter-molecular interactions based on the structure (Fusco, Headd et al. 2014).

For the pre-nucleus stage, ITC, SPR, MST, DLS might all be possible to measure the inter-particle interaction potentials of the protein molecules. However, there have been no cases of applying these methods to crystallographic studies except DLS (Tanaka and Ataka 2002).

1.6 Cryo-EM 3D reconstruction principles

The image processing procedure of single particle cryo-EM is a set of complicated computational jobs, including motion correction, CTF find, particle picking, 2D/3D classification, 3D refinement and polishing. Sufficiently many particle images need to be extracted and accurately aligned either 2-dimensionally or 3-dimensionally. When the alignment is effective and as the number of particle images increases, the signal-to-noise ratio should increase, which finally leads to confident high-resolution EM maps. Therefore, the contrast of the particles and the effectiveness of the alignment determines whether the high-resolution maps can be finally achieved. In this section I will briefly discuss the principles of the contrast formation of cryo-EM images and the principle of 3D reconstruction.

1.6.1 Principle of negative-staining and cryo-EM image formation

To see transparent objects one needs contrast, and the source of contrast can normally be amplitude

and phase (Reimer 2013). For negatively stained samples the contrast mainly comes from amplitude contrast, due to the significantly different masses of heavy atoms and the biological sample (Ohi, Li et al. 2004). For ice embedded specimens, however, no heavy metals are around the sample, and thus the contrast is mainly from the phase (Fernandez-Leiro and Scheres 2016). The phase shift is not significant for cryo specimens, and therefore a volta phase plate (VPP) may be necessary in addition to the correct determination of contrast transfer functions (CTF) (Danev and Baumeister 2016, Khoshouei, Radjainia et al. 2016, Khoshouei, Radjainia et al. 2017) (Figure 1.27).

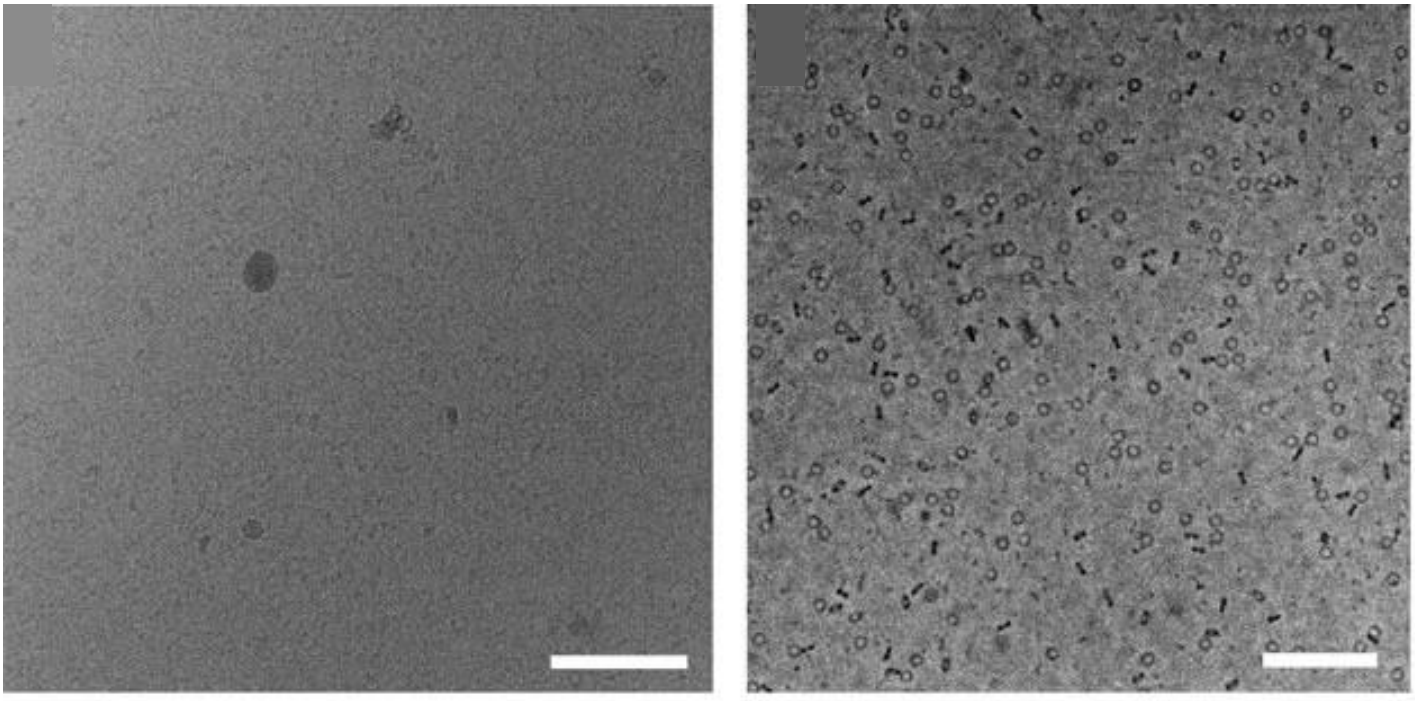


Figure 1.27 Volta phase plate (VPP) helps visualize low-contrast protein particles. The image on the left was taken without VPP and the image on the right was taken with VPP. The scale bar represents 100 nm. The figure is taken from (Khoshouei,

1.6.2 Principle of 3D reconstruction

3D reconstruction is a process of obtaining the 3D structure of an object from 2D photos. A photo of an object is a projection of it from one direction, and it contains compressed signal of the object when some signals are lost (Nogales and Scheres 2015). According to central section theorem, after Fourier transform the projection becomes a central section in the Fourier space (Van Heel 1987). Using this theorem, if photos are taken from different angles, different central sections in the Fourier space can be obtained (Rosenthal

and Henderson 2003). Further, if photos are taken from nearly every angle, then signals in the Fourier space are nearly complete, so the original object can be reconstructed by inverse three-dimensional Fourier transform (Brigham and Morrow 1967, Buijs, Pomerleau et al. 1974), and again if there's a lack of photos from some angles, the signals are lost in the corresponding Fourier space and hence the reconstructed objects become anamorphic (Figure 1.28).

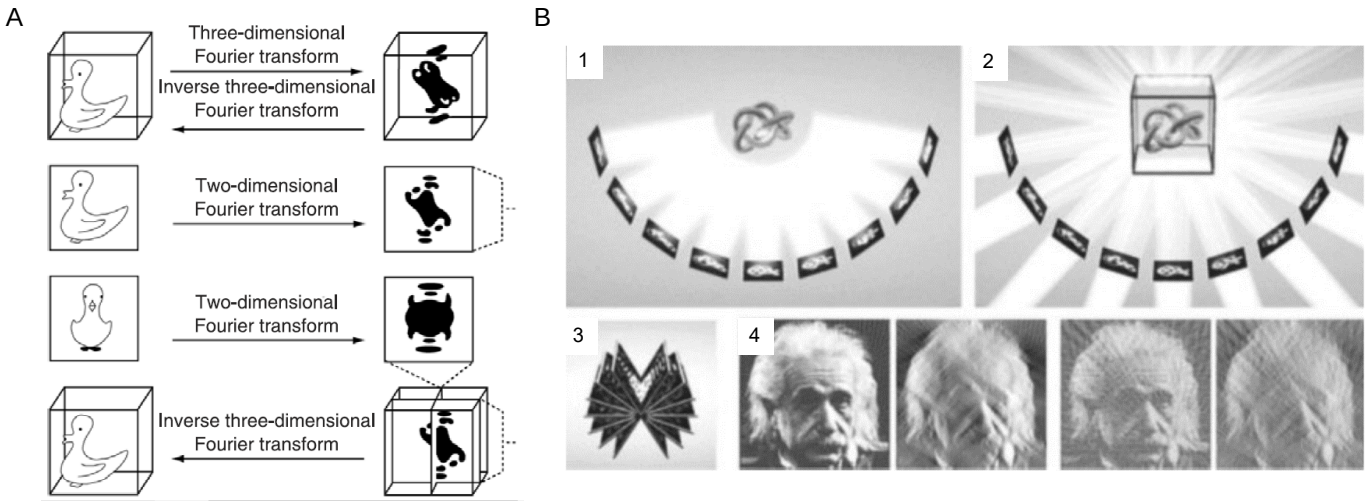


Figure 1.28 Principle of 3-D reconstruction for cryo-EM. The figures are adapted from (Frank 2006). A) A 3-D object (the duck) can be transformed into the Fourier space and vice versa. Projections of the duck can be transformed as a central section into the Fourier space. Then the Fourier space made up sections can be inverse transformed into real space and reconstruct the object. B) Back projections (1 and 2) can be transformed into central sections (3) in the Fourier space. Reconstruction of the object will be more obscure when the central sections in the Fourier space are sparser (4) .

Single particle 3D reconstruction obtains photos of different projections of the object in the way that many different single particles in the specimen take different orientations, so that although the specimen holder does not rotate, different angles of the protein particles can still be obtained passively (Cheng 2015). By the central section theorem, we know that the Fourier space can be almost fulfilled if sufficiently many photos from different angles of projections are obtained, the signals of very few remaining points in the Fourier space that haven't been collected can be calculated by the method of interpolation (Smith 1981). However, some particles have certain preferential orientations, which can lower the resolution of the final reconstruction (Tan, Baldwin et al. 2017, Chen, Noble et al. 2019).

1.7 Biology of model proteins

When I started the project, a few structures were demanded in the SGC. My original goal was solving lots of novel structures. Therefore the model proteins were selected from then, which were often hard-to-crystallize proteins. However, during this process, I found nanobodies did not easily help their targets crystallize, and thus focus was shifted to nanobody scaffold engineering.

1.7.1 RECQL5 and WRN

RECQL5 is one of the five genes of the RecQ family (Hickson 2003). It is involved in processes like DNA repair, recombination and replication (Popuri, Tadokoro et al. 2013). It is widely expressed in the human cells and often acts as a tumour suppressor (Hu, Raynard et al. 2007, He, Qiao et al. 2014). However, in some diseases such as myeloproliferative neoplasms (MNP), RECQL5 is overexpressed, leading to decreased genome integrity (Chen, Ahn et al. 2015). In addition, it was recently discovered that targeting RECQL5 could selectively kill breast cancer (Chakraborty, Dutta et al. 2021). Therefore, it is also worth discovering compounds targeting RECQL5. In this thesis, RECQL5 is the main model protein, and extensive engineering is carried out on its nanobodies to optimize the crystallization of the apo-RECQL5:nanobody complex crystal (without ADP or Mg^{2+}). Currently there is only a 3.4 Å structure for the apo form of RECQL5, which is not sufficient to resolve small compounds. On the other hand, the high resolution 1.8 Å RECQL5 structure has ADP and Mg^{2+} bound, which occupies the active pocket and prevents ligands binding (Newman, Aitkenhead et al. 2017). Therefore, this work aims to improve the resolution of the apo-RECQL5 crystal to 2.5 Å or below with modified nanobodies, so that it is feasible to perform crystallographic fragment screening experiments on apo-RECQL5 (Collins, Douangamath et al. 2018).

Werner syndrome helicase (WRN) is another member of the RecQ family. It is a 1432-amino acid protein, and consists of a helicase core domain, RecQ C-terminal domain, a helicase and RNaseD C-terminal (HRDC) domain and a 3'-5' exonuclease domain (Shen, Gray et al. 1998). The arrangement of the domains are shown in Figure 1.29A. Defects in the WRN helicase can cause the Werner syndrome, which leads to early human aging (Huang, Lee et al. 2006). Moreover, WRN helicase is also a target for microsatellite instability-high (MSI-H) cancers (Chan, Shibue et al. 2019), and thus the structure solution of WRN helicase will be very helpful for drug discovery.

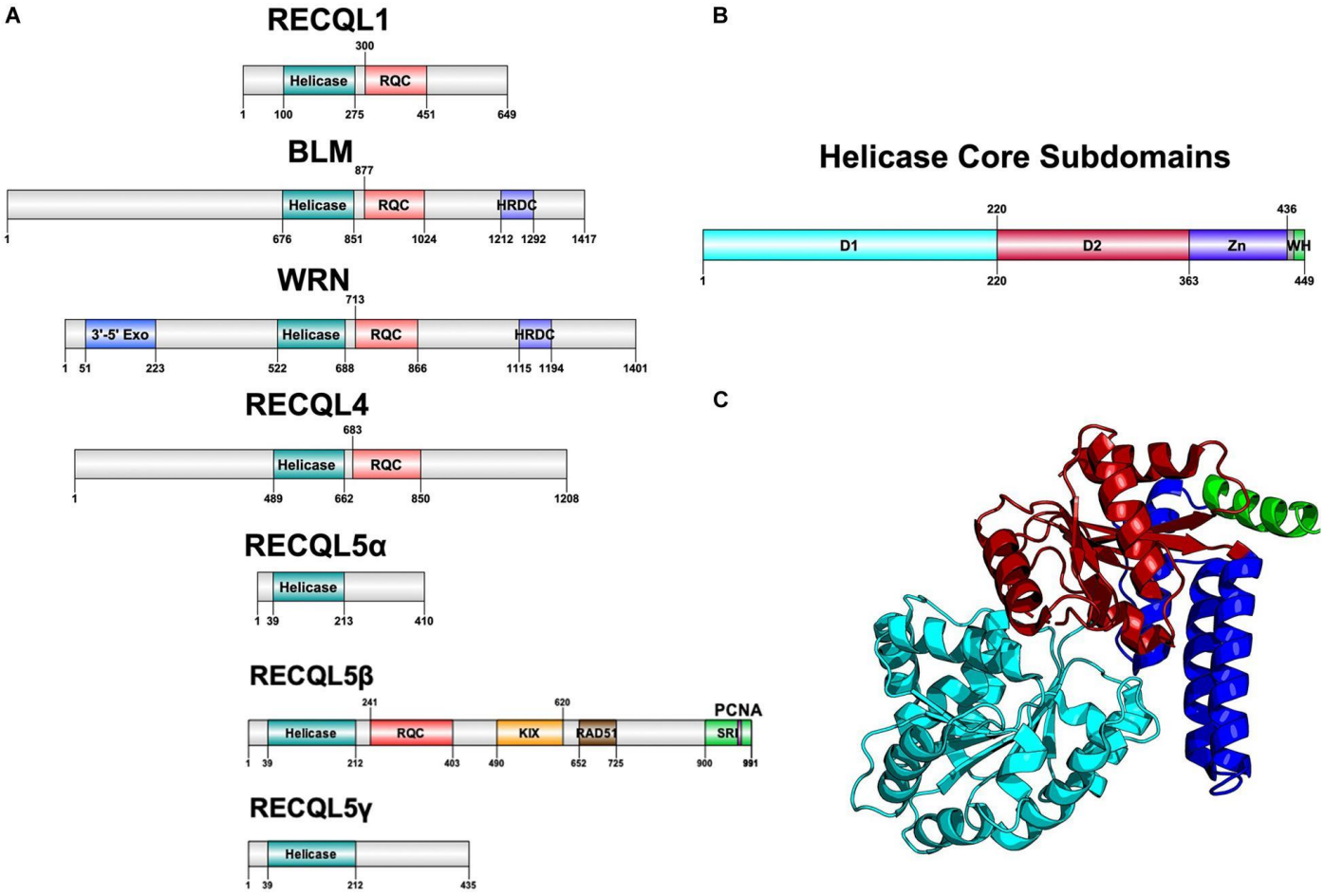


Figure 1.29 The RecQ helicase family and the structure. The figure is taken from (Hamadeh and Lansdorp 2020). A) Domain architecture of the proteins of the RecQ family aligned by the core helicase domain. RQC: RecQ C-terminal domain. HRDC: helicase and RNaseD C-terminal domain. KIX: kinase-inducible domain interacting domain. SRI: Set2-Rpb1 interacting. B) Subdomains of RECQL5 include D1, D2, Zn-binding (Zn), and Winged Helix-like (WH) domains. C) Cartoon representation of the RECQL5 structure. Colors are consistent with the color codes in B).

1.7.2 CTAG1B

The 180 amino-acid protein CTAG1B (Cancer-testis antigen 1B) was first discovered in 1997 (Chen, Scanlan et al. 1997), and then found expressed in different kinds of melanomas, sarcomas and carcinomas (Nicholaou, Ebert et al. 2006). The detailed biological function of this protein still remains unknown, but it is suggested that CTAG1B is involved in cell cycle regulation and apoptosis and is closely related to MAGE proteins. Potentially immunotherapy can be developed against the tumors expressing CTAG1B (Endo, de

Graaff et al. 2015). Until now, there is no full structural information on this protein, the only solved structures are two short truncations (60-72 and 157-165) (Webb, Dunstone et al. 2004, Chan, Gully et al. 2018).

1.7.3 MAGEB1 and MAGEC2

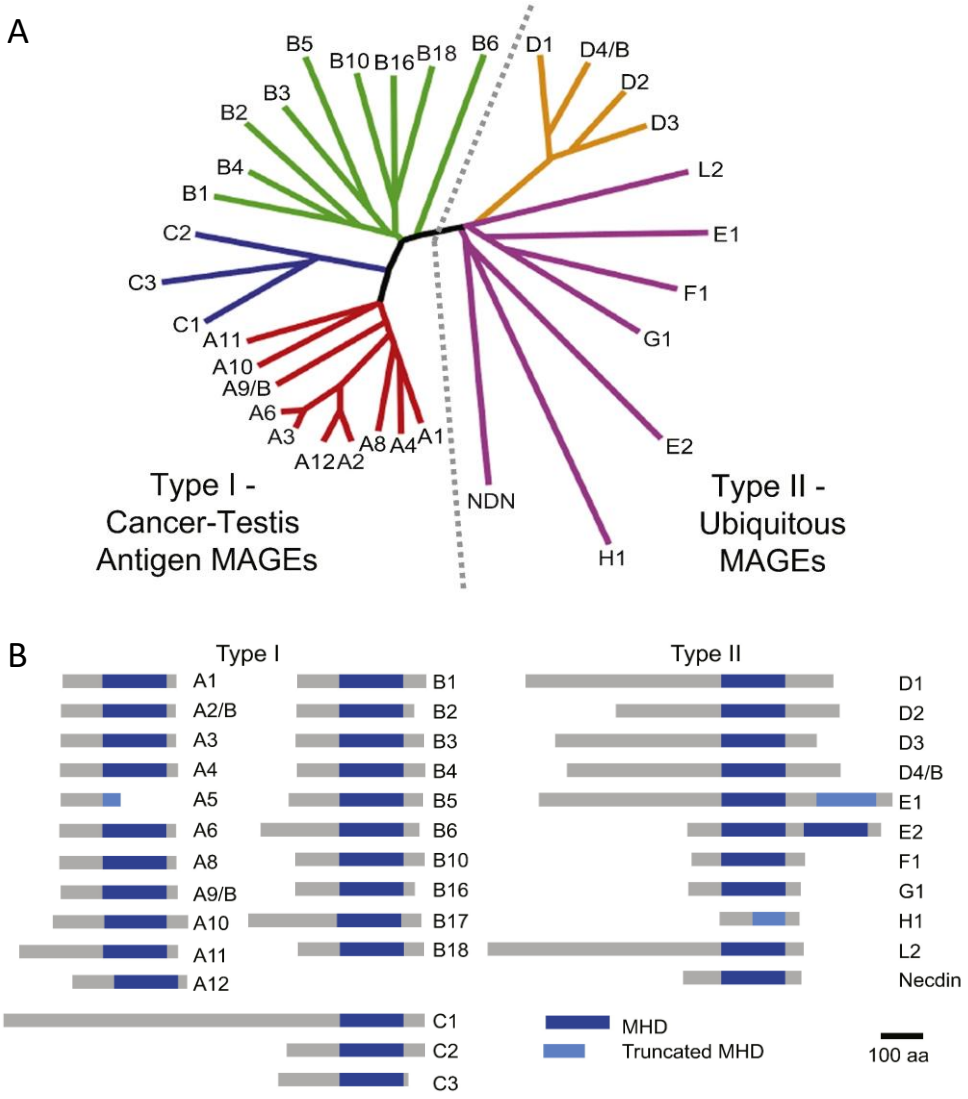


Figure 1.30 The MAGE family. The figures are adapted from (Lee and Potts 2017). A) The evolution of MAGE genes leads to two types. Type I MAGEs, including MAGE-A, MAGE-B and MAGE-C, cluster on the X-chromosome and highly expressed in the testis. Type II MAGEs do not cluster on X-chromosome and therefore are called ubiquitous MAGEs. B) All MAGE proteins more or less share a MAGE-homology domain shown in blue in the architecture.

MAGEB1 and MAGEC2 are two members of melanoma antigen (MAGE) family, which interact with E3 RING ubiquitin ligases and form MAGE-RING ligases (MRL)(Lee and Potts 2017). The MAGE family proteins

have also been detected highly expressed in tumour cells (Mou, Leng et al. 2004). There has been little knowledge about their cellular activities and biological functions, but they are already popular targets for drug discovery. There have been no structural information about MAGEB1 or MAGEC2. Despite existing homolog structures of MAGEA3 (Raman, Rizkallah et al. 2016), MAGEA11 (Yang, Huang et al. 2020) and MAGEG1 (Doyle, Gao et al. 2010), obtaining structures of MAGEB1 and MAGEC2 will broaden the knowledge of the MAGE family, and provides adequate information for drug discovery.

1.7.4 PRNP

The human prion protein (PRNP) plays a key role in forming the human prion disease when the protein is misfolded into a self-templating form (Jucker and Walker 2013). Targeting PRNP using small compounds to prevent its misfolding is a promising anti-prion therapy (Nicoll and Collinge 2009, Barreca, Iraci et al. 2018, Reidenbach, Mesleh et al. 2020). Although there have been several crystal structures obtained so far (Knaus, Morillas et al. 2001, Bortot, Rangel et al. 2021), it is essential to obtain different crystal forms, and the nanobody provides an opportunity for PRNP to crystallize differently (Bortot, Rangel et al. 2021).

1.7.5 TUT4

Terminal uridyltransferase 4 (TUT4/ZCCHC11) is one of the CCHC zinc finger proteins, and is essential in RNA surveillance pathways (Heo, Joo et al. 2009). It functions to add mono or oligo uridines to micro RNAs (miRNA), leading to their maturation or degradation respectively (Heo, Ha et al. 2012, Lim, Ha et al. 2014). There is a crystal structure of the Lin28-interacting module of TUT4 currently available (Yamashita, Nagaike et al. 2019), but there is lack of information on how the domains are spatially arranged, and therefore it is hard to depict its uridyl transfer mechanisms. Recently, a nanobody that can prevent TUT4 from RNA binding was discovered (Yu, Wang et al. 2020). Since the crystallization or cryo-EM study of TUT4:RNA complex is difficult, the structure solution of this TUT4:nanobody will be helpful in providing insight in the interfaces of TUT4 involved in RNA binding. In this thesis, TUT4 is used as a model protein in both crystallography and cryo-EM. In crystallography, the optimized nanobody is used to co-crystallize TUT4. In cryo-EM, the nanobody dimerises to increase the particle size of the TUT4:nanobody complex, aiming to enable particle picking and high-resolution cryo-EM structure solution of TUT4.

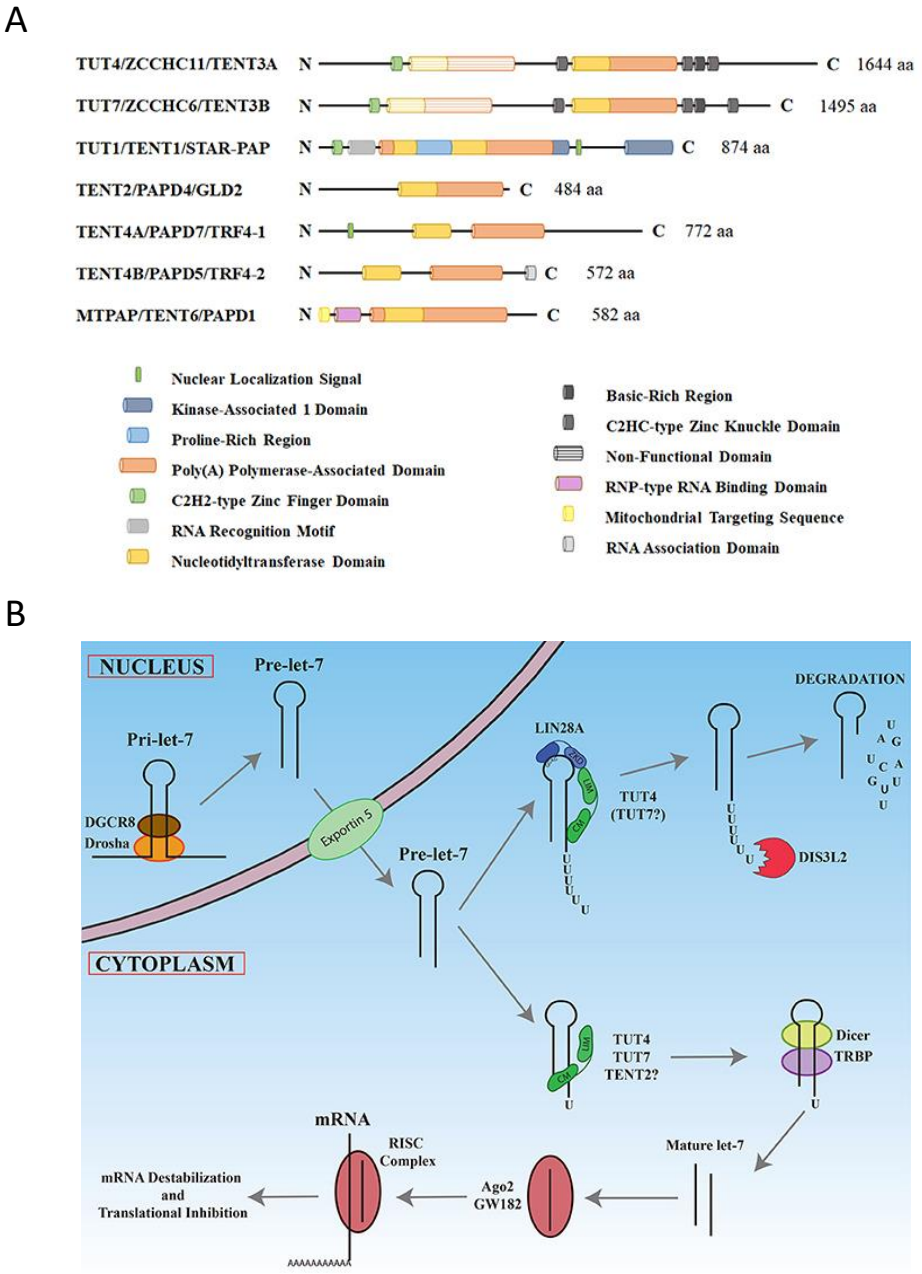


Figure 1.31 The TUT family and the pre-let-7 processing pathway. The figures are taken from (Menezes, Balzeau et al. 2018). A) The domain architecture of all TUT members. All members contain a polyA polymerase-associated domain and a nucleotidyltransferase domain. B) The pre-let-7 processing pathways. After pre-let-7 is exported to the cytoplasm, there are two downstream pathways. In the first pathway, pre-let-7 is oligo-uridylated by Lin28A and TUT4 (or TUT7) and then the oligo-U-tail will be recognized and degraded by DIS3L2. In the second pathway, pre-let-7 is mono-uridylated by TUT4, TUT7 or TENT2, and then the downstream RNA interference is triggered.

1.8 Overview of the thesis

This thesis describes the engineering of the nanobody scaffold for structural biology purposes, specifically crystallography and cryo-EM.

The majority of the thesis is about crystallography, which demonstrates that the scaffold of the nanobody is naïve and has a large potential to be optimized for crystallography.

Chapter 2 argues that naïve nanobodies are dissatisfactory crystallization chaperones based on the data of a selection of crystallization trials. To further support this argument, mutagenesis was carried out on the scaffold of a nanobody which failed to co-crystallize its target. This finally generated an effective nanobody crystal chaperone, further demonstrating that the nanobody scaffold is optimizable.

Chapter 3 focuses on bioinformatics research on nanobody-nanobody crystal contacts based on deposited structures and discovers 'Nb-α' as the most predominant nanobody-nanobody crystal contact. The result of this study supplies an idea where to start the engineering on the nanobody scaffold.

Chapter 4 describes the evolution of Gluebodies from naïve nanobodies. The chapter starts with the optimization of the nanobody engineering pipeline, and proceeds to the five iterations of mutagenesis on the prevalent nanobody-nanobody crystal contact - Nb-α. Finally, it describes the application of the optimized nanobody scaffold version (Gluebody) to a selection of other target proteins.

Chapter 5 describes the efforts on application of Gluebodies to cryo-EM as imaging scaffolds. It describes the current progress of the development of disulfide-linked di-Gluebody as an imaging scaffold for single particle cryo-EM, and also discusses the future perspectives on the tetra-Gluebody as an advanced imaging scaffold.

Finally, Chapter 6 describes further scientific questions behind Gluebodies and future directions of this project.

In conclusion, the thesis demonstrates the necessity for optimizing the nanobody scaffold, and this work provides an optimized version of the nanobody – Gluebody, that can be easily transferable to other target proteins. The discovery of Gluebodies sheds light on a new usage of nanobodies for structural biology in both crystallography and cryo-EM.

Chapter 2 Nanobody scaffolds have engineering potential for better crystallization

2.1 Introduction

The concept of nanobody scaffold engineering to optimize its crystallization potential appears unexplored. Therefore, I started the exploration with a selection of 5 target protein with the related 17 nanobodies to see how many pairs of the protein complexes could yield well-diffracting crystals.

As a result, only one protein:nanobody pair (MAGEB1) yielded a crystal which diffracted to 2.7 Å. This result demonstrates that the success rate of the nanobody is below 10%. Although it is hard to decide whether it is good or not without comparison, subjectively it is not satisfactory, which is the motivation for optimizing the nanobody scaffold for better crystallization.

Subsequently, I set out to explore whether the nanobody scaffold was optimizable by testing 74 nanobody variants with surface mutations based on a RECQL5 nanobody, which failed to co-crystallize RECQL5 in the initial experiment. As a result, several crystals were obtained from 5 variants, and one single key mutation D69Y was sufficiently effective to rescue the nanobody and yielded a crystal diffracting to 2.8 Å.

I further analyzed the structures from the MAGEB1:nanobody complex and RECQL5:nanobody complex. There is only a very small nanobody-nanobody interface in the MAGEB1 structure, whereas a very good large edge-to-edge beta-sheet nanobody interface is present in the RECQL5 structure.

This Chapter concludes that the nanobody scaffold has potential to optimize for better crystallization, and provides a good model protein RECQL5 for subsequent scaffold engineering.

2.2 Naïve nanobodies are dissatisfactory crystallization chaperones

2.2.1 Methods and materials

2.2.1.1 Selected target proteins and nanobodies

To explore how well naïve nanobodies perform as crystallization chaperones, a selection of SGC (Structural Genomics Consortium, Oxford) targets with respective binding nanobodies were selected, expressed and purified to proceed with the crystallization coarse screen.

The IDs and respective sequences of selected target proteins are listed as follows.

Target	SGC construct	Protein sequence (tag cleaved)
CTAG1B	CTAG1BA-c113	MLEFYLAMPFATPMEAELARRSLAQDAPPLPVPGVLLKEFTVSGNILTIRLTAADHRQLQLSISSCLQ QLSLLMWITQCFLPVFLAQAENLYFQ
MAGEB1	MAGEB1A-c001	SMSTESSVKDPVAWEAGMLMHFILRKYKMREPIMKADMLKVVDEKYKEYFTEILNGASRRLELVF GLDLKEDNPSGHTYTLVSKLNLTNDGNLSNDWDFPRNGLLMPLLGVIFLKGNSATEEEIWKFMNVL GAYDGEEHLIYGEPRKFITQDLVQEKYLKYEQVPNSDPPRYQFLWGPRAYAETTKMKVLEFLAKMN GATPRDFPSHYEEALRDEEERAQVRSSVRARRRTTATTFRARSRAPFSRSSHPM
MAGEC2	MAGEC2A-c209	SMESSFTYTLDEKVAELVEFLLLKYEAEEPVTEAEMLMIVIKYKDYFPVILKRAREFMELLFGLALIEVG PDHFCVFANTVGLTDEGSDDEGMPENSLLIIILSVIFIKGNCASEEVIWEVLNAVGVYAGREHFVYGE PRELLTKVWVQGHYLEYREVPHSSPPYYEFLWGPRAHSESIKKKVLEFLAKLNNTVPSSFPSWYKDA LKDVEERVQ
RECQL5	RECQL5A-c104	SMDPERRVRSTLKKVFGFDSFKTPLQESATMAVVKGNKDVFVCMPTGAGKSLCYQLPALLAKGITI VVSPLIALIQDQVDHLLTLKVRVSSLNSKLSAQERKELLADLEREKPQTKILYITPEMAASSSFQPTLNS LVSRHLLSYLVVDEAHCVSQWGHDFRPDYLRLGALRSRLGHAPCVALTATATPQVQEDVFAALHLK KPVAIFKTPCFRANLFYDVQFKELISDPYGNLKDFCLKALGQEADKGLSGCGIVYCRTREACEQLAIEL SCRGVNAKAYHAGLKASERTLVQNDWMEEKVPVIVATISFGMGVDKANVRFVAHWNIAKSMAG YYQESGRAGRDGKPSWCRLYYSRNDRDQVSFLIRKEVAKLQEKRGNKASDKATIMAFDALVTFCEE LGCRHAAIAKYFGDALPACAKGCDHCQNPTAVRRRLEALERSSSW
WRN	WRNA-c060	SMAPNEEQVTCLKMYFGHSSFKPVQWKVIHSVLEERRDNVAVMATGYGKSLCFQYPPVYVGKIGL VISPLISLMEDQVLQLKMSNIPACFLGSAQSENVLTDIKLGKYRIVYVTPEYCSGNMGLLQQLEADIGI TLIAVDEAHCISEWGHDFRDSFRKLGSLKTALPMVPIVALTATASSSIREDIVRCLNLRNPQITCTGFD RPNLYLEVRRKTGNILQDLQPFLVKTSSHWEFEGPTIIYCPSRKMTQQVTGELRKLNLSCGTYHAGM SFSTRKDIHHRFVRDEIQCVIATIAFGMGINKADIRQVIHYGAPKDMESYYQEIGRAGRDGLQSSCH VLWAPADINLNRHLLTEIRNEKFRLYKLKMMAKMEKYLHSSRCRRQIILSHFEDKQVQKASLGIMGT EKCCDNCRSRLDHCYSMD

The IDs and respective sequences of selected nanobodies are listed as follows.

Target construct ID	Nanobody construct ID	Nanobody protein sequence
CTAG1BA-c113	YNCTAG1BA-c001	MKYLLPTAAAGLLLLAAQPAMAQVQLQESGGGLVQAGGSLRLSCAASGRTFSSYSMGWFRQAPG KEREFVAGISWSGGSTDYVDSVKGRFTISRDNAKNTAYLQMNSLKPEDTAVYYCAADRGDGGSNW RLETYYDYWGQGTQVTVS
	YNCTAG1BA-c002	MKYLLPTAAAGLLLLAAQPAMAQVQLQESGGGLVQAGGSLRLSCATSGRTVYTLAMGWFRQAPG KEREFVAAVSTSGGSTFYADSVEDRFTISKDNAKSTVNLHMNSLNPEDTAVYYCAAKMGYTSSRWT NYRLTDVYDYWGQGTQVTVS
	YNCTAG1BA-c003	MKYLLPTAAAGLLLLAAQPAMAQVQLQESGGGLVQPGGSLRLSCAASGFSSDYYEIGWFRQAPGKE REAVSCISSGDNTYYKDSVKGRFTISRDNAKKTVYLQMNSLKSEDTAVYFCAAGLRTVAGSQCPMSR YFYSSWGQGTQVTVS
MAGEB1A-c001	YNMAGEB1A-c001	MKYLLPTAAAGLLLLAAQPAMAQVQLQESGGGLVQAGGSLRLSCAASGFTFDDYAIGWFRQVPGK EREGVSCTSSSVGSTYYADSVKGRFTITSDNAKNTVYLQMNSLKPEDTAVYYCAADSYPYGSTWCRP PFFASWGQGTQVTVS
MAGEC2A-c209	YNMAGEC2A-c001	MKYLLPTAAAGLLLLAAQPAMAQVQLQESGGGLVQAGGSLRLSCKASISLFNRFDVAWYRQAPGK QRDMVARITSGGSVEYSDDVKGRFTIARDNAKNMVLLQMNSLKLEDTAVYYCNLKRRWSWGSSN EFWGQGTQVTVS
RECQL5A-c104	YYRECQL5A-c004	QVQLQESGGGLVQAGGSLRLSCAASGFTFDDYAIGWFRQAPGKEREGVSSIKSSDGSTYYADSVKG RFTISSDNAKNTVYLQMNSLKPEDTAVYYCAATETRPTMDPLDFGYWDQGTQVTVSAENLYFQ
	YYRECQL5A-c005	QVQLQESGGGLVQAGGSLRLSCAASGNIFNINAMGWYRQVPGKQREMVAAITSGGSTNYADSVK GRFTISRDNAKNTVYLQMNSLNPEDTAIYYCEAYGTYTLAPTGEGEYDDYWGQGTQVTVSAENLYF Q
	YYRECQL5A-c006	MQVQLQESGGGLVQAGGSLRLSCAASGSIFSINRMGWYRQAPGKQRELVAAITSGGSTNYADSVK GRFTISRDNAKNTVYLQMNSLKPEDTAIYYCEAYGTYTLAPTGEGEYDDYWGQGTQVTVSAAENLYF Q
WRNA-c060	YNWRNA-c001	QVQLVESGGGLVQAGNSLRLSCLASGRTLASYGMGWFRQAPGKEREFVAAISRSFSNTYYADSVKG RFTISRDNAKNAVYLQMNILQPEDTAAYYCAATSRFAAATDSQYYDYWGQGTQVTVSSHHHHHHE PEA
	YNWRNA-c002	QVQLVESGGGLVQPGGSLRLSCLASGRTLASYGMGWFRQAPGKEREFVAAISRSFSNTYYADSVKG RFTISRDNAKNAVYLQMNILQPEDTAAYYCAATSRFAAATDSQYYDYWGQGTQVTVSSHHHHHHE PEA
	YNWRNA-c005	QVQLVESGGGLVQAGNSLSLSCLASGRTLASYGMGWFRQAPGKEREFVAAISRSFSNTYYADSVKG RFTISRDNAKNAVYLQMNILQPEDTAAYYCAATSRFAAATDSQYYDYWGQGTQVTVSSHHHHHHE PEA
	YNWRNA-c006	QVQLVESGGGLVQTGGSLRLSCAASESITSTMSWYRQAPGKQRELVASITYFGSTDYADFVKGRFSIS RNGARDMLFLQMNKMKPEDTAVYYCKVHTLYGRDYWGQGTQVTVSSHHHHHHEPEA
	YNWRNA-c013	QVQLVESGGGLVQPGGSLRLSCAASGFTFGTYFMNWFRQAPGKGPEWVSGIDPGGDRTWYVESV KGRFTISRDNVKNTLYLQMNSLKPADTAEYYCSPQGRPERGQGTQVTVSSHHHHHHEPEA
	YNWRNA-c014	QVQLVESGGGLVQPGGSLRLSCAASGFTFGTYFMNWFRQAPGKGPEWVSGINPGGDRTSYVESVK GRFTISRDNVKNTLYLQMNSLKPADTAEYYCSPQGRPERGQGTQVTVSSHHHHHHEPEA

Target construct ID	Nanobody construct ID	Nanobody protein sequence
WRNA-c060	YNWRNA-c015	QVQLVESGGGLVQAGGSLRLSCAASGSIFSSNIMGWYRQGPGNTRRFVASIAPGGTTNYERSVKGR FTISRDNAKNTVHLQMNSLKPDDTADYLCNALPADYWGQGTQVTVSSHHHHHHEPEA
	YNWRNA-c017	QVQLVESGGGLVQAGGSLRLSCAASGRTSSIYGMGWFRQAPGKEREFVAAITWILGTAYYGDSVKG RFTISRDNAKNTYLQMNSLKPEDTAVYYCAGQEAPFAPRDPGRVNYWGQGTQVTVSSHHHHHHE PEA
	YNWRNA-c018	QVQLVESGGGLVQPGGSLRLSCAASRSTVSLNVIAWYRQAQGKQREWVANIGSGGSTNYAVSVKG RFTISRDNAKNTAYLQMNNLKPEDTAVYYCQAFAHTTAESPYTDYWGQGTQVTVSSHHHHHHEPE A

2.2.1.2 Purification of the target proteins

CTAG1BA-c001, MAGEC2A-c209, MAGEB1A-c001 and WRNA-c060 protein samples were directly provided by Prof. Opher Gileadi. The RECQL5A-c104 construct plasmid was provided by Prof. Opher Gileadi and its expression and purification were carried out as follows.

The truncated RECQL5 protein (11-453) was subcloned into pNIC-Bsa4 vector with TEV cleavable 6xHIS tag at the N terminus. The protein was expressed using BL21-DE3-pRARE strain in auto-induction TB media (Formedia) containing antibiotics and 0.01% antifoam 204 at 37°C for 5.5 hours followed by 40-44 hours at 18°C. The base buffer we used for purification contained 5% glycerol, 10 mM HEPES (pH 7.5), 500 mM NaCl, 0.5mM TCEP. Bacteria were harvested by centrifugation at 4000 g and resuspended in 3 times volume of base buffer + 30 mM Imidazole, 1% Triton, 0.5 mg/ml lysozyme, 10ug/ml homemade benzonase, followed by storage in -80°C freezer overnight for complete cell lysis. The purification started next day with thawing the frozen bacterial pallets in room temperature water bath, followed by centrifugation at 5000 g for 1h to obtain clear supernatant. The supernatant was then applied to Ni-NTA pre-packed columns (GE healthcare) pre-equilibrated with base buffer + 30 mM imidazole. After thoroughly washing the Ni-NTA columns with base buffer, protein was eluted using 2.5 ml of base buffer + 500 mM imidazole directly into base buffer equilibrated PD-10 columns (GE healthcare). 3.5ml of base buffer was then applied to PD10 columns to elute the RECQL5 protein in 0 imidazole buffer. Then TEV protease was added to protein solution with a 1:10 mass ratio for overnight incubation, 20 mM imidazole was also added in the solution. In the next day, twice quantity of Ni-NTA columns were pre-equilibrated with base buffer + 20 mM imidazole and the RECQL5 solution with TEV was applied to the columns to get rid of RECQL5 with uncleaved 6xHIS tag, TEV protease and contaminants. Flowthrough fractions were collected, and flash frozen for making nanobody complexes.

2.2.1.3 Purification of nanobodies

The nanobody construct plasmids stated in section 2.2.1.1 were all provided by Prof. Opher Gileadi, and their expression and purification were carried out as follows.

Constructs of nanobody variant were transformed into BL21-DE3-pRARE *E.coli* strain. A single colony was inoculated into 1 ml of LB media (Merck) and incubated overnight. Starter culture of each nanobody was then inoculated into 100ml autoinduction media (Formedia) containing 0.01% antifoam 204 and antibiotics, at 37°C for 5.5 hours followed by 40-44 hours at 18°C. The base buffer we used for purification contained 5% glycerol, 10 mM HEPES (pH 7.5), 500 mM NaCl, 0.5 mM TCEP. Bacteria for each nanobody were harvested at 4000g and resuspended in 10ml base buffer + 30 mM imidazole, 1% Triton, 0.5 mg/ml lysozyme, 10 ug/ml homemade benzonase, followed by storage in -80°C freezer overnight for complete cell lysis. The purification started next day with thawing the frozen pallets in room temperature water bath, followed by centrifugation at 5000g for 1 h to obtain clear supernatant. The supernatant for each nanobody was then applied to one 1 ml Ni-NTA pre-packed column (GE healthcare) pre-equilibrated with base buffer + 30 mM imidazole. After thoroughly washing the Ni-NTA columns with base buffer, nanobody was eluted using 2.5 ml of base buffer + 500 mM imidazole directly into base buffer equilibrated PD-10 columns (GE healthcare). 3.5 ml of base buffer was then applied to PD10 columns to elute the nanobody protein in 0 imidazole buffer. Then TEV protease was added to protein solution with a 1:10 mass ratio for overnight incubation, and 10 mM imidazole was also added in the solution. The next day, two Ni-NTA columns were pre-equilibrated for each nanobody with base buffer + 10 mM imidazole and the nanobody solution with TEV was applied to the columns to get rid of nanobody with uncleaved tags, TEV protease and contaminants. Flow-through fractions were collected and 2 ml of base buffer was applied to the Ni-NTA columns to flush all nanobody through. The collected nanobody fractions were ready for complex preparation.

2.2.1.4 Crystallization trial setup of the target:nanobody complexes

The purified binding protein pairs were mixed respectively and put through gel filtration on Superdex 200 or Sepax SRT Sec-300, and peaks corresponding to the complex proteins were pooled and concentrated to 5-10 mg/ml for subsequent crystallization trials.

Two crystallization plates were then set up for each protein complex against two commercially available

coarse screens (Hampton Index and JCSG7) from Molecular dimensions at 20 ^{o}C using Mosquito in the sitting-drop SwissCI-3 drop plates. The plates were then monitored by the Formulatrix crystal plate hotel and imager for 56 days.

2.2.1.5 Diffraction data analysis and structure determination

Diffraction data were processed by multiple auto-processing (including Xia2 and Autoproc) pipelines on ISPYB (Winter 2009, Clemens, Vonrhein et al. 2011, Delagenière, Brenchereau et al. 2011). The results from Xia2-DIALS pipeline were selected for subsequent structure determination. Datasets were truncated using Aimless in CCP4-i suite (Winn, Ballard et al. 2011). Phasing of MAGEB1 was done by Phaser (McCoy, Grosse-Kunstleve et al. 2007), using the MAGEA3 structure (PDB id 4V0P) as a search model. The search model for phasing the nanobody was chain A of 2X1O. Further refinement was done by COOT (Emsley and Cowtan 2004) and Refmac5 (Murshudov, Skubak et al. 2011) in CCP4-i suite.

2.2.2 Result

I altogether selected 17 nanobody-target pairs across 5 different target proteins as stated above and performed standard crystallization trials of the 17 protein complexes and obtained with the protein qualities shown in Figure 2.1. Throughout the whole 56-day observation period, only one protein complex yield crystals in one condition (Figure 2.2) – 0.056M sodium phosphate monobasic plus 1.344M potassium phosphate dibasic with a protein:precipitant ratio of 1:1. The crystal was harvested with 17% ethylene glycol as the cryo-protectant. The 2.7 Å diffraction dataset led to the successful structure solution of the MAGEB1:nanobody protein complex (PDB ID 6R7T). This suggests that the effectiveness of nanobodies as crystallization chaperones is less than 10%, demonstrating the need for dramatic improvement of the approach.

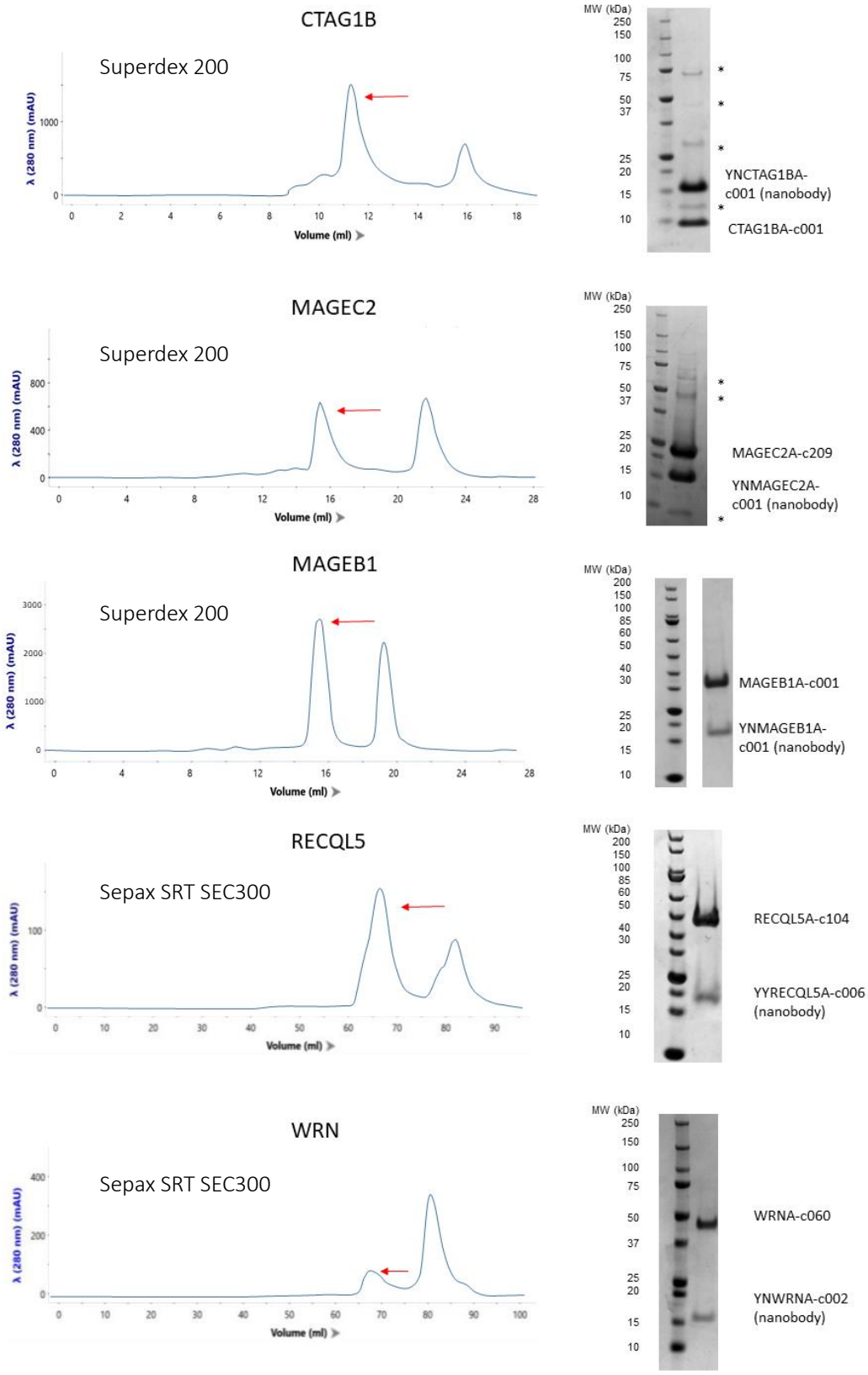


Figure 2.1 The size exlucsion chromatography profiles (left) and SDS-PAGE images (right) for each target protein. The red arrows show the protein peak fractions collected for crystallization trials. '*' indicates contaminants in the protein samples.

A

Target-Nanobody PAIR	STRUCTURE SOLVED?	Target ID
CTAG1BA:YNCTAG1BA-c001	no	CTAG1BA-c001
CTAG1BA:YNCTAG1BA-c002	no	CTAG1BA-c001
CTAG1BA:YNCTAG1BA-c003	no	CTAG1BA-c001
MAGEB1A:YNMAGEB1A-c001	yes PDB id 6R7T	MAGEB1A-c001
MAGEC2A:YNMAGEC2A-c001	no	MAGEC2A-c209
RECQL5A:YYRECQL5A-c004	no	RECQL5A-c104
RECQL5A:YYRECQL5A-c005	no	RECQL5A-c104
RECQL5A:YYRECQL5A-c006	no	RECQL5A-c104
WRNA:YNWRNA-c001	no	WRNA-c060
WRNA:YNWRNA-c002	no	WRNA-c060
WRNA:YNWRNA-c005	no	WRNA-c060
WRNA:YNWRNA-c006	no	WRNA-c060
WRNA:YNWRNA-c013	no	WRNA-c060
WRNA:YNWRNA-c014	no	WRNA-c060
WRNA:YNWRNA-c015	no	WRNA-c060
WRNA:YNWRNA-c017	no	WRNA-c060
WRNA:YNWRNA-c018	no	WRNA-c060

B

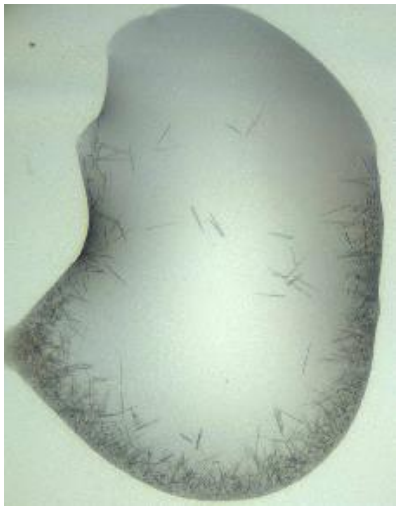


Figure 2.2 The crystallization experiment on a selection of target:nanobody pairs with naïve nanobody scaffolds. A) Summary of the experiment outcome. B) MAGEB1:nanobody complex crystals

In the 2.7 Å MAGEB1 structure, there is only a small nanobody-nanobody crystal contact, and the rest of the contacts are nanobody-MAGEB1 contacts and one MAGE-MAGE contact (Figure 2.3). This suggests that the nanobody indeed helps mediate most crystal contacts which MAGEB1 alone cannot achieve, and thus triggers the crystallization. However, the small nanobody-nanobody interface makes it difficult to engineer and its potential to further improve crystallization is limited.

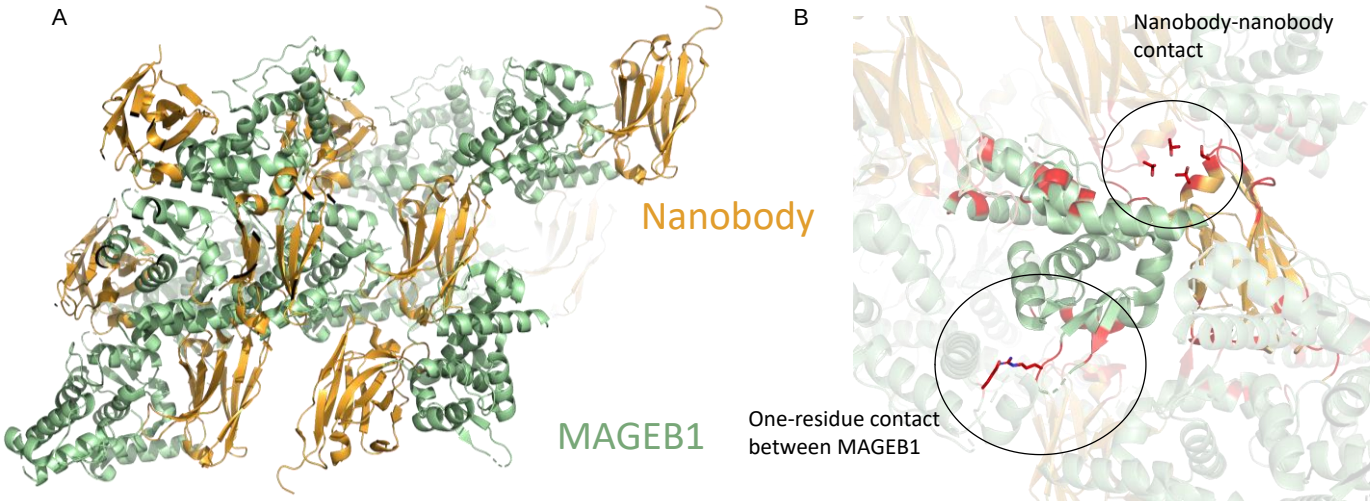


Figure 2.3 The crystal contacts of the MAGEB1:nanobody complex structure. A) MAGEB1 (green) and the nanobody (orange) in the crystal lattice. B) A close-up view of one copy of the MAGEB1:nanobody complex, its contact with the symmetry mates being labelled red. Most of the contacts are MGEB1-nanobody contacts, while only a type of one-residue contact between MAGEB1 molecules takes place. There is also only one type of nanobody-nanobody contact.

2.2.3 Discussion

This experiment was at the very initial stage of the DPhil project, the original goal was to obtain a few crystals (more than one expected) with the help of nanobody. However, the result turned out not satisfactory, and on the other hand demonstrated that the nanobody strategy was not successful.

This experiment could be designed better. For example, more crystal plates could be set up at 4 ^{o}C. However, I obtained a crystal when I finished setting up plates at 20 ^{o}C, and I needed to proceed with the structure solution of the MAGEB1:nanobody complex, so there was no more time for me to spend on this initial experiment. Actually, I did set up some plates at 4 ^{o}C including all RECQL5:nanobody complexes, however I did not obtain any crystals either. Due to the imperfectness of this experiment, I could not draw a strong conclusion, but just expressed that the result was not satisfactory. Nevertheless, this result indeed turned out unexpected, meaning that the nanobodies were expected to be very effective crystal chaperones while they were not in reality. Nanobodie still have potential to improve.

The reason why the success rate of this experiment was so low was that the protein complex was not optimized for crystal packing, which might be a cause of the target protein itself or the nanobody scaffold. The next step was to demonstrate whether the nanobody scaffold had potential to be optimized. The easiest way was not to use the solved MAGEB1:nanobody structure, since it was not obvious where on the nanobody scaffold could be optimized as there was only a very small nanobody-nanobody crystal contact in the structure and the engineering on that would not be universally applicable. Furthermore, to prove whether the modification was helpful was not easy if using diffraction quality as a criteria. On the contrary, it would be much easier to decide whether the modification was successful or not by rescuing an ineffective nanobody scaffold to start yielding crystals. Therefore in the next section, I set out to test several surface mutations on the RECQL5A-c104:YYRECQL5A-c006 complex.

2.3 Ineffective nanobodies improved by surface mutations

The RECQL5 nanobody (YYRECQL5A-c006) is one of the failed nanobodies from the study in the previous section. The reason for choosing this nanobody is that both RECQL5 and the nanobody are highly expressing, speeding up the engineering process. The mutagenesis of the RECQL5 nanobody scaffold is the initial step in

the project.

2.3.1 Strategy of crystal epitope mutagenesis

The concept involved in the first round of engineering is to use crystal epitopes. The reason why I chose this strategy is that previous experiments carried out by Michael Fairhead (SGC, University of Oxford) involving 35 different target proteins have shown that there is a good chance for these crystal epitope mutants to yield reproducible crystals (Figure 1.17).

To find out the right residues to mutate. I used the plug-in named 'Xtal Epitope' in ICM-pro to complete this process, to which I submitted a full sequence of the nanobody, a BLAST result from NCBI website and chain A of the structure 2x1o from PDB as the homology model. The output of the programme is a list of recommended crystal epitope mutations. A screenshot of the software user interface is shown in Figure 2.4.

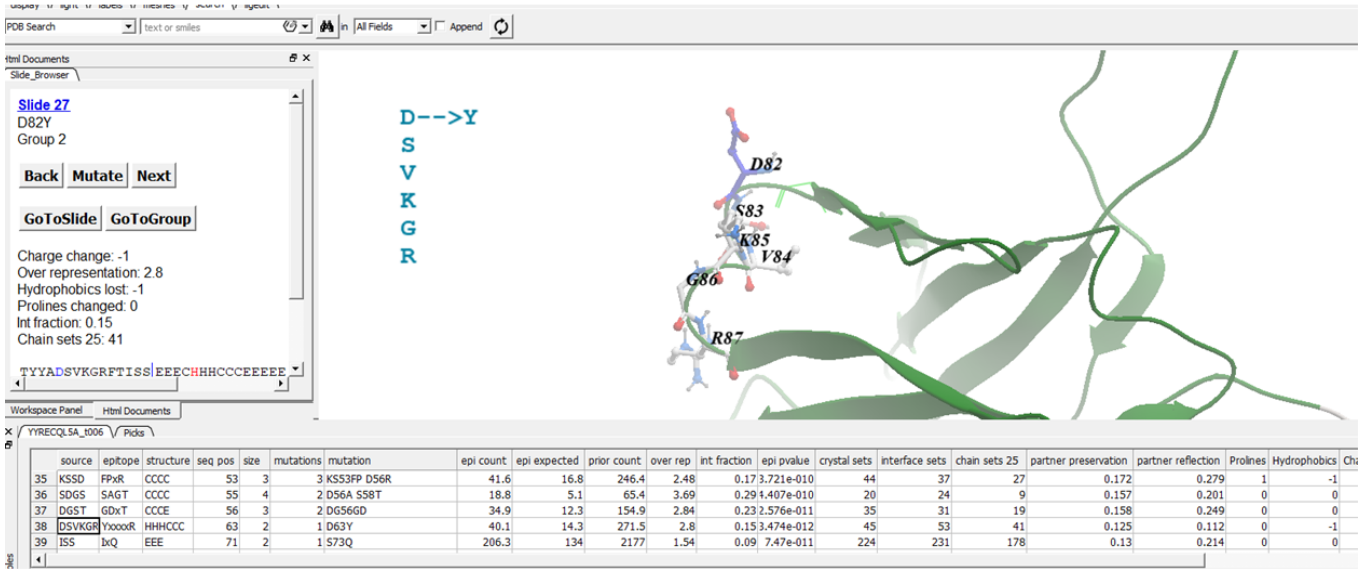


Figure 2.4 ICM-pro crystal epitope helper user interface developed by Claire Strain-Damerell (the Diamond Light Source). The structure on the right-top is adapted from the homology model with potential crystal epitope sequence shown in stick. The table at the bottom presents a list of recommended mutation in the protein.

2.3.2 Methods and materials

In this experiment I designed 74 nanobody variants based on YYRECQL5A-c006, which from this section on is given a new ID – G0-001. The IDs and mutations of the variants are listed in Figure 2.6A. Mutations

IMGT numbering	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
G0-001	Q	V	Q	L	V	E	S	G	G		G	L	V	Q	A	G	G	S	L	R	L	S	C	A	A

IMGT numbering	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50
G0-001	S	G	S	I	F					S	I	N	R	M	G	W	Y	R	Q	A	P	G	K	Q	R

IMGT numbering	51	52	53	54	55	56	57	58	59	60	61	62	63	64	65	66	67	68	69	70	71	72	73	74	75
G0-001	E	L	V	A	A	I	T	S	G				G	S	T	N	Y	A	D	S	V	K		G	R

IMGT numbering	76	77	78	79	80	81	82	83	84	85	86	87	88	89	90	91	92	93	94	95	96	97	98	99	100
G0-001	F	T	I	S	R	D	N	A	K	N	T	V	Y	L	Q	M	N	S	L	K	P	E	D	T	A

IMGT numbering	101	102	103	104	105	106	107	108	109	110	111	111A	111B	111C	111D	112E	112D	112C	112B	112A	112	113	114	115	116
G0-001	I	Y	Y	C	E	A	Y	G	T	Y	T	L	A	P				T	G	E	G	E	Y	D	D

IMGT numbering	117	118	119	120	121	122	123	124	125	126	127	128
G0-001	Y	W	G	Q	G	T	Q	V	T	V	S	

Figure 2.5 The nanobody sequence of G0-001 in the IMGT numbering scheme. The residues in cyan correspond to CDR regions.

annotated with '[]' are crystal epitope mutations, which are separated by ':'. The 'Ctag' at the beginning denotes that the multi-Histidine tag is tethered at the C terminus with a TEV cleavage site separating the nanobody and the tag. L12W is not bracketed because the L12 is an essential residue, which will be detailed in Chapter 4. The template wild type sequence is shown in Figure 2.5 with IMGT numbering. All residue indices in the list below correspond to the numbering in Figure 2.5. There is an additional 'M' before the nanobody starting sequence 'QVQL' and an additional 'AAENLYFQ' after the nanobody ending sequence 'VTVS'.

The nanobody constructs of G0-002 to G0-074 and G1-001 were generated by site-directed mutagenesis method based on G0-001 using pNIC-CTH0 as the vector plasmid. All cloning procedures followed standard SGC ligation independent cloning protocol (Savitsky, Bray et al. 2010).

Constructs of nanobody variants were transformed into Single step (KRX) *E.coli* strain (Promega) at the end of molecular cloning (Hartnett, Gracyalny et al.). A single colony was inoculated into 1 ml of LB media (Merck) and incubated overnight. Starter culture of each nanobody was then inoculated into 100ml autoinduction media (Formedia) containing 0.1% Rhamnose, 0.01% antifoam 204 and antibiotics, at 37°C for

5.5 hours followed by 40-44 hours at 18°C. The base buffer we used for purification contained 5% glycerol, 10 mM HEPES (pH 7.5), 500 mM NaCl, 0.5 mM TCEP. Bacteria for each nanobody were harvested at 4000g and resuspended in 10ml base buffer + 30 mM imidazole, 1% Triton, 0.5 mg/ml lysozyme, 10 ug/ml homemade benzonase, followed by storage in -80°C freezer overnight for complete cell lysis. The purification started next day with thawing the frozen pallets in room temperature water bath, followed by centrifugation at 5000g for 1 h to obtain clear supernatant. The supernatant for each nanobody was then applied to one 1 ml Ni-NTA pre-packed column (GE healthcare) pre-equilibrated with base buffer + 30 mM imidazole. After thoroughly washing the Ni-NTA columns with base buffer, nanobody was eluted using 2.5 ml of base buffer + 500 mM imidazole directly into base buffer equilibrated PD-10 columns (GE healthcare). 3.5 ml of base buffer was then applied to PD10 columns to elute the nanobody protein in 0 imidazole buffer. Then TEV protease was added to protein solution with a 1:10 mass ratio for overnight incubation, and 10 mM imidazole was also added in the solution. The next day, two Ni-NTA columns were pre-equilibrated for each nanobody with base buffer + 10 mM imidazole and the nanobody solution with TEV was applied to the columns to get rid of nanobody with uncleaved tags, TEV protease and contaminants. Flow-through fractions were collected and 2 ml of base buffer was applied to the Ni-NTA columns to flush all nanobody through. The collected fractions were ready for complex preparation.

Each nanobody variant was then mixed with the purified RECQL5 protein (with purification procedures detailed in section 2.2.1.2) at a molar ratio of 1.5:1, and concentrated using 10 kDa concentrator (Amicon and Corning), followed by Sepax SRT SEC-300 HPLC or Superdex-200 SEC. The peak fractions containing both RECQL5 and nanobody were pooled and concentrated to certain extent (Appendix Chater 8.3) for crystallization trials. Crystal plates were set up by Mosquito using Swiss-CI 3 drop plates with precipitants from Hampton Index screen (HIN3 HT-96, Molecular dimensions). Plates were sealed and sent into Formulatrix imager and imaged at 12 hours, 1 day, 4 days, 7 days, 14 days, 28 days and 56 days. Crystals were harvested using Shifter with 17% glycerol for high salt conditions or ethylene glycol for PEG conditions added as cryo-protectant, then snap frozen and shipped to multiple beamlines at Diamond Light Source for X-ray diffraction.

Diffraction data were processed by multiple auto-processing (including Xia2 and Autoproc) pipelines on ISPYB (Winter 2009, Clemens, Vonrhein et al. 2011, Delagenière, Brenchereau et al. 2011). The results from

Xia2-DIALS pipeline were selected for subsequent structure determination. Datasets were truncated using Aimless in CCP4-i suite (Winn, Ballard et al. 2011). Phasing of RECQL5 was done by Phaser (McCoy, Grosse-Kunstleve et al. 2007), using the D1 and D2 domains of RECQCL5 (PDB id 5LB8) as search models. The search model for phasing the nanobody was chain A of 2X1O. Further refinement was done by COOT (Emsley and Cowtan 2004) and Refmac5 (Murshudov, Skubak et al. 2011) in CCP4-i suite.

2.3.3 A single nanobody surface mutation yields an effective chaperone for RECQL5

A

Single Site (22 constructs)

	Nanobody Mutation ID	Protein ID
Site 1	Ctag:[D69Y]	G1-001
	Ctag:[A50F:A51T]	G0-002
	Ctag:[E6Q]	G0-003
	Ctag:[L12W]	G0-004
Site 2	Ctag:[T86A:V87L:Y88S]	G0-005
	Ctag:[T78S]	G0-006
	Ctag:[K84S]	G0-007
Site 3	Ctag:[Q90E]	G0-008
	Ctag:[Y88H]	G0-009
	Ctag:[N82S:N85G]	G0-010
	Ctag:[K95E]	G0-011
	Ctag:[N85G]	G0-012
	Ctag:[N82S]	G0-013
	Ctag:[N85S]	G0-014
	Ctag:[N82K:N85G]	G0-015
	Ctag:[N82G:N85G]	G0-016
	Ctag:[S79Q]	G0-017
	Ctag:[R20K:L21V]	G0-018
	Ctag:[A15P:G17D]	G0-019
	Ctag:[A15P:G17A]	G0-020
	Ctag:[A45V]	G0-021
	Ctag:[Q44R]	G0-022

Double Site (31 constructs)

Site 1 X Site 2 → 3 (G0-023 – 25)

Site 1 X Site 3 → 7 (G0-026 – 32)

Site 2 X Site 3 → 21 (G0-033 – 53)

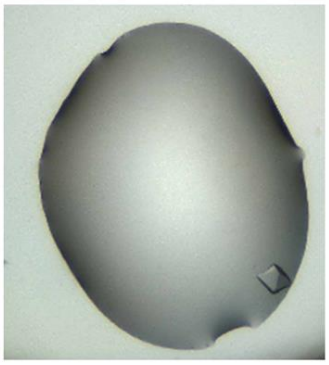
Triple Site (21 constructs)

Site 1 X Site 2 X Site 3 → 21 (G0-054 – 74)

B

Nanobody mutation ID	Protein ID	Concentration (mg/ml)	No. of Drops with Mountable crystals	No. of Drops with Crystalline species
Ctag:[D69Y]	G1-001	11.46	1	1
Ctag:L12W:[D69Y:K84S]	G0-054	10.43	6	6
Ctag:L12W:[D69Y:Y88H]	G0-056	10.15	2	19
Ctag:L12W:[D69Y:N82S:N85G]	G0-057	14.05	4	10
Ctag:L12W:[D69Y:N82G:N85G]	G0-060	16.70	0	4

C



D

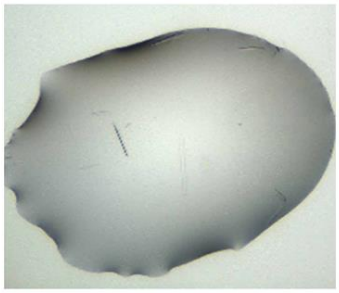


Figure 2.6 The crystal epitope mutation D69Y on the scaffold rescues the RECQL5 nanobody to be effective as a crystal chaperone. A) The combinatorial mutation designs. 22 single site mutations were chosen from the list of mutations provided by the crystal epitope server. 11 of the single site mutations were then selected to generate double-site and triple-site mutations thereafter, which resulted in 74 nanobody variants for crystallization trials. B) The information of the crystal epitope variants that yielded crystals in the Hampton Index 3 Screen. All those effective nanobody variants shared the D69Y mutation, highlighted in red. C) The G1-001 crystal. D) G0-057 crystals.

74 crystal epitope mutations (variant class G0 and variant G1-001) were designed through 3 iterations on the scaffold (non-CDRs) of a RECQL5 nanobody (Figure 2.6A). Purified nanobodies all co-migrated with

RECQL5 on size exclusion chromatography, indicating adequate affinity of the nanobodies. The purified protein complexes were then put into crystallization trials using the Hampton Index coarse screen.

Five out of 74 yielded crystals (Figure 2.6B). These variants shared one common mutation – D69Y, which is also the only mutation of the variant G1-001 (Figure 2.6C). The crystals of the G1-001 complex diffracted to the highest resolution of 2.76 Å. Y69 on the nanobody forms a hydrogen bond with D400 of RECQL5, triggering the crystal formation (Figure 2.7A). These results show that certain mutations on the non-CDRs of the nanobody surface can increase its crystallization propensity while preserving the affinity for its target protein. However, the mutations we found effective were not mediating nanobody-nanobody crystal contacts, and thus could only be useful for their specific target RECQL5.

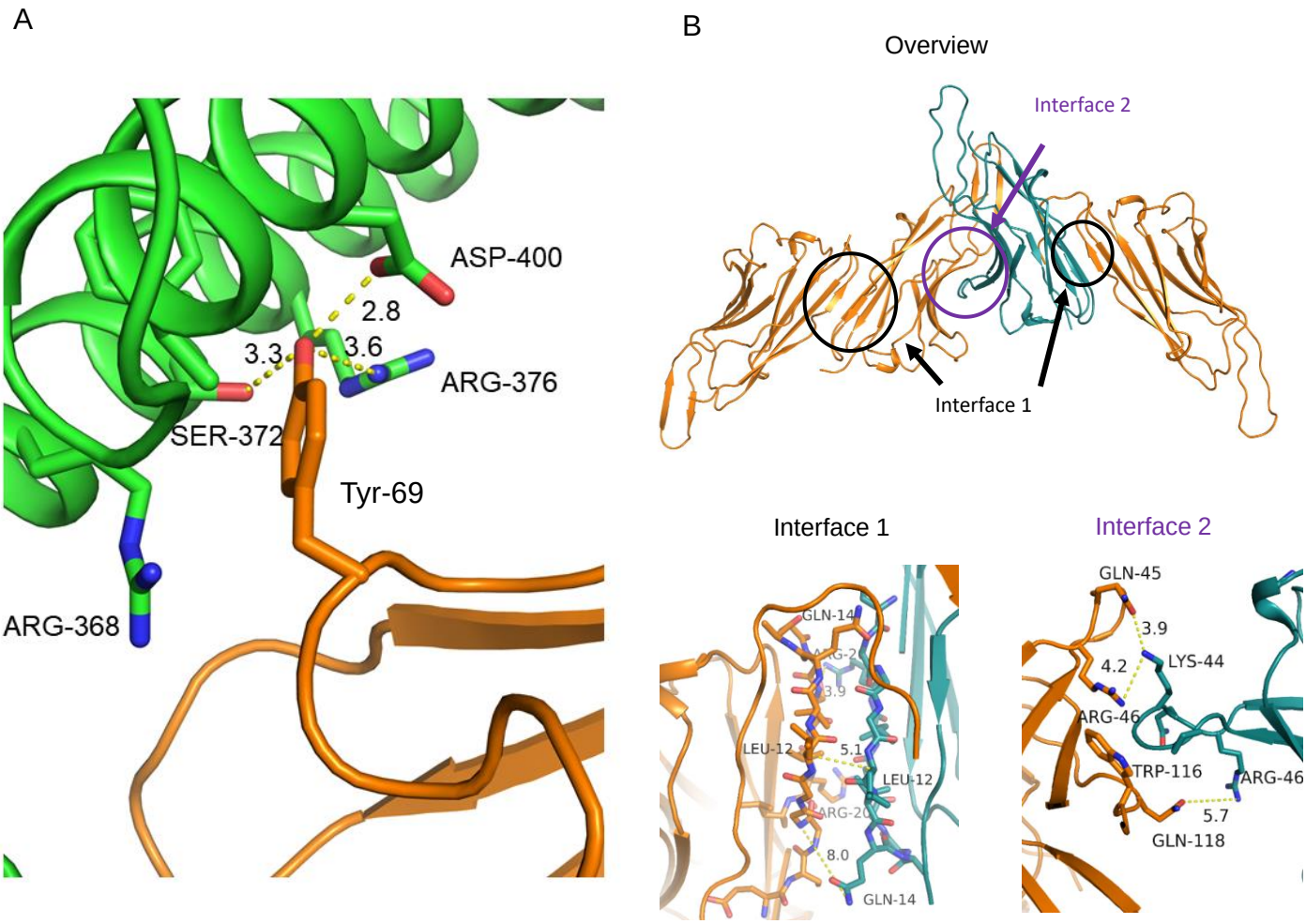


Figure 2.7 RECQL5-nanobody and nanobody-nanobody contacts in G1-001. A) Tyr69 mediates crystal contacts between nanobody (orange) and RECQL5 (green) in G1-001. Tyr69 from the nanobodies contacts an alpha helix of RECQL5.

Figure 2.7 (Continued) B) Two nanobody-nanobody interfaces from G1-001. Interface 1 (later named Nb-α) is indicated in black circle and Interface 2 is indicated in light grey circle. At the bottom are closer views of Interface 1 (left) and Interface 2 (right). Oxygen and Nitrogen atoms in stick representation are colored red and blue respectively. The distances between atoms are displayed in Å.

2.3.3 Discussion and conclusion

This quick experiment which only took three months immediately proves that the nanobody scaffold is optimizable without many efforts. Although the mutation D69Y introduced an additional hydrogen bond between the nanobody and RECQL5, not target independent, still it demonstrates that small changes in the nanobody scaffold can make a big difference. Meanwhile, it is noticed that there are two types of interfaces in the RECQL5:nanobody complex structure, one of which is a large interface of more than 8 interacting residues. This finding makes me interested in how nanobodies often pack on each other *in crystallo*. Therefore, to improve the nanobody scaffold for assisting crystallization of any target of interest, I further devised a strategy to identify mutations mediating crystal contacts that were independent of target proteins.

Chapter 3 Bioinformatics research on the crystal contacts between nanobodies

3.1 Introduction

To be universally applicable, the crystallization of nanobody:protein complexes should rely more on nanobody:nanobody crystal contacts and less on fortuitous interactions between the nanobody scaffold and target proteins. This chapter sets out to discover how nanobodies pack on each other *in crystallo* through data mining on 335 nanobody-containing crystal structures in the protein data bank (PDB).

By using the Pymol package in Python 2.7, I was able to extract all interface information in the 335 crystal structures and focused on the interfaces that only involved nanobodies. The nanobody:nanobody interfaces were represented by index arrays of the residues involved. The index array was then renumbered by the IMGT scheme, enabling comparisons across the nanobodies appearing in different deposited structures. The renumbered arrays were further converted into coordinate matrices, which then immediately shrank into a 4-dimensional array (n^*, μ_x, μ_y, μ_z). The dimensions represent 1/3 of the number of residues in the interface, and the mean values of the X, Y, Z coordinates of the residues in the interface. The 4-dimensional array was then used for unsupervised density-based clustering (DBSCAN), and 9 classes of interfaces were identified.

Among the 9 classes of interfaces, there was one predominant interface, which appeared in 47 structures (39.5% among the structures containing large nanobody:nanobody interfaces). I call this interface 'Nb-α' afterwards. This interface features a beta-sheet interaction near the N-terminus of the nanobody, with two sub-classes – 'aligned' or 'shifted' depending on whether the residue 12 of the nanobody is aligned or not. There is also a detailed analysis of the rest of the interface classes.

This piece of information was important prior to the engineering project, since it provided a direction on where to engineer. Although the number 47 is not a large fraction of 335 (the total number of structures in this analysis), I hypothesized that there is also a decent chance of improving the nanobody scaffold as a crystal chaperone by modifying this interface.

3.2 Methods

3.2.1 Summary

Given a nanobody-containing crystal structure from the pdb, the symmetry mates were first generated to restore the real situation in the crystal lattice. Every interface was then identified and examined. Nanobody-nanobody contacts were subsequently selected. I adopted the criteria of two molecules forming crystal contacts that the nearest atoms are below 4 Å apart, and thus 273 structures were found to contain nanobody-nanobody contacts. The contacts which have more than 5 interacting residues are more suitable for engineering, and therefore only the 356 large contacts across 119 structures were further analysed.

The analysis process started with unifying nanobody residue indexes in different structures with IMGT numbering scheme on the online server ANARCI (Figure 1.3). Selected nanobody interfaces were then mapped onto a single nanobody structure (PDB id 5O0W) using their coordinate information. As numeric representations of the interfaces, 4-dimensional vectors were used and I further performed density-based spatial clustering (DBSCAN) (Ester, Kriegel et al. 1996) on the 356 vectors (Figure 3.4).

All of this analysis was carried out with the Pymol package in Python 2.7. Although the idea was straight forward, it was a bit tedious when realising it through Pymol-API. Below I will describe in detail how to realise the idea step by step, with the source code presented in Appendix section 8.1.

3.2.2 Identifying interacting residues in the crystal lattice

The nanobody-containing crystal structures I obtained from the PDB are listed below as the four-letter codes:

2X1O, 2X1P, 2X1Q, 2X6M, 2X89, 2XT1, 2XV6, 2XXC, 2XXM, 3CFI, 3DWT, 3EAK, 3EBA, 3EZJ, 3G9A, 3K1K, 3K74, 3OGO, 3P0G, 3SN6, 3STB, 3ZKQ, 3ZKS, 4AQ1, 4BEL, 4BFB, 4C57, 4C58, 4C59, 4CDG, 4EIG, 4EIZ, 4EJ1, 4FHB, 4GFT, 4GRW, 4HEM, 4HEP, 4I0C, 4I13, 4I1N, 4IOS, 4JVP, 4KML, 4KRL, 4KRM, 4KRN, 4KRO, 4KRP, 4LDE, 4LDL, 4LDO, 4LGP, 4LGR, 4LGS, 4LHJ, 4LHQ, 4MQS, 4MQT, 4N9O, 4OCL, 4OCM, 4OCN, 4P2C, 4QGY, 4QKX, 4QLR, 4QO1, 4S10, 4S11, 4W6W, 4W6X, 4W6Y, 4WEM, 4WEN, 4WEU, 4WGV, 4WGW, 4X7C, 4X7D, 4X7E, 4X7F, 4XT1, 4Y8D, 4Z9K, 4ZG1, 5BOP, 5BOZ, 5C1M, 5C2U, 5C3L, 5DA0, 5DA4, 5DFZ, 5DXW, 5000, 5E0Q, 5E5M, 5E7B, 5E7F, 5F1K, 5F1O, 5F21, 5F7K, 5F7L, 5F7M, 5F7N, 5F7W, 5F7Y, 5F8Q, 5F8R, 5F93, 5F97, 5F9A, 5F9D, 5FWO, 5G5R, 5G5X, 5GXB, 5H8D, 5H8O, 5HDO, 5HGG, 5HVF, 5HVG, 5HVH, 5IMK, 5IML, 5IMM, 5IMO, 5IP4,

5IVN, 5IVO, 5JA8, 5JA9, 5JDS, 5JQH, 5LHN, 5LHP, 5LHQ, 5LHR, 5LMJ, 5LMW, 5LZ0, 5M13, 5M14, 5M15, 5M2M, 5M2W, 5M30, 5M7Q, 5M94, 5M95, 5MJE, 5MP2, 5MWN, 5MY6, 5MZV, 5NBD, 5NBL, 5NBM, 5NLU, 5NLW, 5NM0, 5NML, 5NQW, 5O02, 5O03, 5O04, 5O05, 5O0W, 5O2U, 5O8F, 5OCL, 5OJM, 5OMM, 5OMN, 5OVW, 5TD8, 5TJW, 5TOJ, 5TOK, 5TP3, 5U64, 5U65, 5UK4, 5UKB, 5USF, 5VAK, 5VAN, 5VL2, 5VLV, 5VM0, 5VM4, 5VM6, 5VNV, 5VNW, 5WB1, 5WB2, 5Y7Z, 5Y80, 6B20, 6B73, 6C5W, 6C9W, 6DBA, 6DBD, 6DBE, 6DBF, 6DBG, 6EHG, 6EQI, 6EY0, 6EY6, 6EZW, 6F0D, 6FE4, 6FUZ, 6FV0, 6GZP, 6H02, 6H15, 6H16, 6H1F, 6H6Y, 6H6Z, 6H70, 6H71, 6H72, 6H7J, 6H7L, 6H7M, 6H7N, 6H7O, 6I6J, 6IBL, 6MXT, 6QTL, 6R7T, 6O3C, 6OS0, 6OS1, 6OS2, 6OYH, 6OYZ, 6OZ6, 6Q6Z, 6QD6, 6QGW, 6QGX, 6QGY, 6QPG, 6QUZ, 6QV0, 6QV1, 6QV2, 6QX4, 6RNK, 6RTW, 6RTY, 6RU3, 6RU5, 6RUL, 6RUM, 6RUV, 6RVC, 6S0Y, 6SGE, 6SSI, 6SSP, 6TEJ, 6TYL, 6SWR, 6U12, 6U14, 6U50, 6U51, 6U52, 6U53, 6U54, 6U55, 6VI4, 6WAQ, 6WAR, 4DK3, 4DK6, 4DKA, 4KDT, 4KSD, 4PIR, 5J1S, 5J1T, 5M2I, 5M2J, 6GJQ, 6GJS, 6GJU, 6GK4, 6GKD, 6GWN, 6GWP, 6GWQ, 6HD8, 6HD9, 6HDA, 6HDB, 6HDC, 6HEQ, 6HER, 6HHD, 6HHU, 6HJX, 6HJY, 6I2G, 6I8G, 6I8H, 6IBB, 6ITP, 6ITQ, 6JB2, 6JB5, 6JB8, 6JB9, 6DO1, 6DYX, 6F2G, 6F2W, 6FPV, 6GCI, 6GS1, 6GS4, 6GS7, 6N4Y, 6N50, 6NFJ.

For each structure, the symmetry mates within 10 Å of the molecule were generated using the ‘symexp’ command in Pymol to restore a local fraction of the crystal lattice. I subsequently call the newly generated molecules ‘symmetry copies’ and the original PDB structure ‘the parent copy’.

In this step, I did not only identify those crystal contacts involving nanobodies, but determined all the interfaces around the parent copy, so that the script could be reused for other research purposes. The interfaces within the parent copy were ignored, because the nanobody binding interface was not of interest. Therefore, it was just necessary to iterate through every residue in the parent copy, select the residues on the symmetry copies that were within 4 Å and record the molecule and chain they belonged to and also their residue index. Finally, a csv file was generated with all the information above and the X, Y, Z coordinates of the Ca atom of the residue, shown in Figure 3.1. The ‘*’ in the first column denotes the residue in the row belongs to the parent copy, and others belong to the symmetry copies. The information fetched in this step will be used in the next step.

6H1F

Annotation	molecule	chain	residue index	residue one-letter code	distance	X	Y	Z
*	6H1F	A	2	V		-18.244	15.181	-10.304
	sym02-10000	B	181	E	7.523105	-19.8545	19.0785	-4.074
*	6H1F	A	3	Q		-20.962	17.674	-11.417
	sym02-10000	B	181	E	7.5577	-19.8545	19.0785	-4.074
*	6H1F	A	11	L		-33.615	25.616	-32.269
	sym03-10100	A	43	K	5.790538	-34.015	30.0105	-36.0185
	sym03-10100	A	42	G	7.966628	-30.608	31.6835	-36.4655
	sym03-10100	A	44	E	7.701558	-37.082	29.1915	-38.1435
*	6H1F	A	12	V		-31.476	26.25	-35.351
	sym03-10100	A	42	G	5.614127	-30.608	31.6835	-36.4655
*	6H1F	A	13	Q		-32.011	26.441	-39.113
	sym03-10100	A	42	G	6.038333	-30.608	31.6835	-36.4655
	sym03000100	A	105	G	7.913763	-27.713	32.4515	-41.9465
	sym03-10100	A	39	Q	9.89837	-37.69	34.4485	-37.8455
	sym03000100	A	104	Y	8.951738	-28.942	31.9395	-45.4755
	sym03000100	A	104	Y	8.875301	-28.988	31.8695	-45.4505
	sym03-10100	A	43	K	5.1316	-34.015	30.0105	-36.0185
	sym03000100	A	106	P	10.66317	-24.448	33.8125	-40.5845
*	6H1F	A	13	Q		-32.056	26.377	-39.083

Figure 3.1 Snapshot of the csv file recording residues in contact. The '*' in the first column denotes the residue in the row belongs to the parent copy, and others belong to the symmetry copies

3.2.3 Grouping interacting residues into interfaces

In this step, using the information obtained in section 3.2.2, interacting residues were to be grouped into interfaces. Each interaction involves two surfaces and each surface has its symmetry copy ID and chain ID, and therefore each pair of interacting residues has a pair of symmetry copy and chain IDs (e.g. chain M of symmetry copy X and chain N of symmetry copy Y), which I call interface ID afterwards. The pairs of interacting residues that have the same interface ID can be grouped into 2 lists, each recording the residue

indices from one surface of the interface (side A or side B). Hence a csv file recording all interfaces around the parent copy was generated at the end of this step, as shown in Figure 3.2.

6QTL			
Interface - side A	Interface - side B	Residues involved in side A	Residues involved in side B
sym01-10001 B	6QTL H	87	64
sym00000001 A	6QTL H	15 17 18	102 109
sym01-10001 E	6QTL B	104	120 15
6QTL C	6QTL F	103 104 29 31 56 76 79	10 11 114 115 116 12 13 14 20 9
sym01000000 C	6QTL D	102 103 104	116 118 120 13
sym01000100 G	6QTL E	43 44 90 91	43 44 91
sym01000001 E	6QTL H	10 11 115 12 7 8 9	56 57 74 76 77
6QTL B	6QTL H	104 105 106 107 108 111 34 39 40 46 48 49 52 54 59 61 63 65	105 106 107 108 111 34 39 40 46 47 48 49 52 54 59 61 63 65
sym0000-100 F	6QTL D	15 16 17 68 86 87	57 58 60 62 67 68 71 72 73 74
sym01000001 E	6QTL B	116 118 12 13	102 104
sym01-10100 H	6QTL C	116 118 119 13 43 90	118 121 13 43 90 91 93
sym01000001 E	6QTL E	3	21 84
sym01-10100 H	6QTL A	43 44 45	44 45
sym000000-1 B	6QTL A	59	87
6QTL H	6QTL E	27 28 29 3 5	56 57 58 59
sym01000100 D	6QTL G	15 18 19	64 67
sym000000-1 H	6QTL G	104	120 121 15
6QTL H	6QTL B	104 105 106 107 108 111 34 39 40 46 47 48 49 52 54 59 61 63 65	105 106 107 108 111 34 39 40 46 48 49 52 54 59 61 63 65
sym01000100 D	6QTL C	116 118 120 13	102 104

Figure 3.2 Snapshot of the csv file recording interfaces. The first two columns together record the interface ID, where the string before the space is the ID of either the parent copy or one of the symmetry copies and the letter after the space is the chain ID. The last two columns record the indices of the residues involved in the interface, using the numbering from the deposited PDB file.

3.2.4 Renumbering the nanobodies from PDB using the IMGT scheme

In this step, information about nanobody-nanobody interfaces were to be extracted from the csv file generated in section 3.2.3, and the residue indices were to be unified.

The csv file from section 3.2.3 contained information about all nanobody:nanobody, nanobody:target and target:taget interfaces. However, I was only interested in the nanobody:nanobody interfaces. Therefore, the nanobody chains need to be identified. To do this, I downloaded the fasta sequences of all PDB structures listed in section 3.2.2, combined them into one multi-fasta file and submitted it to the ANARCI server. The server returned a csv file recording the renumbering result of only the nanobodies, excluding any target proteins. I call this file the reference file afterwards. Therefore, I could easily distinguish whether a certain chain of the structure is a nanobody by deciding whether the specific chain is in the reference file. Then those nanobody chains were renumbered by the IMGT scheme with the reference file, and a renumbered interface file was generated.

In the renumbered interface file, I observed lots of nanobody:nanobody interfaces had communal residues and were symmetric, i.e. both sides of the interface had similar interacting residues. For the convenience of analysis and also a harsher selection criteria, I used the array of communal interacting residues to represent every interface and excluded any interfaces that had fewer than or equal to five interacting residues, since those small interfaces were not useful for subsequent engineering. Finally, 119 structures were found to contain large nanobody:nanobody interfaces and I was able to further analyse those residue index arrays by unsupervised clustering.

3.2.5 Clustering the nanobody interfaces based on density

The index arrays I obtained in 3.2.4 were not easy to analyse since they have different dimensions and the indices are discrete numbers, which cannot represent spatial information. I therefore expanded the index arrays into 3-row matrices containing the 3-dimensional coordinates of each residue index from the original index arrays. The coordinates were taken from the respective residues of the deposited structure 5O0W. I then could immediately shrink the 3-row matrices into a 4-dimensional array (n^*, μ_x, μ_y, μ_z). The dimensions represent 1/3 of the number of residues in the interface, and the mean values of the X, Y, Z coordinates of the residues in the interface. The 4-dimensional array was then used for unsupervised density-based clustering DBSCAN (Density-based spatial clustering of applications with noise). By using this clustering algorithm, I could identify the areas on the nanobody densely populated with nanobody:nanobody interfaces, and thus the easy-forming interfaces could be selected for engineering.

Finally, a result log file of clustering was generated (Figure 3.3). Each row of the file contains the interface name starting with the PDB code of the structure, the IMGT renumbered indices of the residues involved in the interface, the reference interface (the most central point in the cluster), the class the interface belongs to, the distance of the interface to the reference interface, and the coordinates of the pattern represented by the last three dimensions of the 4-dimensional array (n^*, μ_x, μ_y, μ_z). The number after the '_' in the first column represents the index of the interfaces in a certain PDB file if that structure has more than one nanoboby:nanobody interfaces. Meanwhile, the number of structures where each class of interfaces appears was also calculated and reported.

The steps in sections 3.2.2 to 3.2.5 are summarized in Figure 3.4 and the related source codes are written in section 3.3. The clustering results will be analyzed in detail in section 3.4.

Interface name	Interface pattern	Reference interface	Class	Distance	X	Y	Z
2XV6_2	11 12 13 18 19 93	6H7J_1:SYM00010000 D:6H7J C:11 12 127 14 19 5	1	31.4568	46.72133	29.05383	336.3267
3P0G_1	11 12 127 14 19 26 3 5 6 7 8 9	6H7J_1:SYM00010000 D:6H7J C:11 12 127 14 19 5	1	17.16407	41.56283	26.66708	329.2587
3STB_1	8 9 11 122 12 13 14 18 19 20 7	6H7J_1:SYM00010000 D:6H7J C:11 12 127 14 19 5	1	1.726727	43.87073	27.99191	333.6058
4QGY_1	11 12 13 123 14 18 19 20 7 8 9	6H7J_1:SYM00010000 D:6H7J C:11 12 127 14 19 5	1	2.436732	43.98982	28.269	333.773
4QGY_2	11 12 13 123 14 18 19 20 7 8 9	6H7J_1:SYM00010000 D:6H7J C:11 12 127 14 19 5	1	2.436732	43.98982	28.269	333.773
4W6W_1	5 6 7 8 9 11 12 13 14 19 120 122 127	6H7J_1:SYM00010000 D:6H7J C:11 12 127 14 19 5	1	7.959404	42.48369	28.69015	330.8795
4WEU_1	7 8 9 11 12 13 14 18 19 20 90 92 122	6H7J_1:SYM00010000 D:6H7J C:11 12 127 14 19 5	1	7.13852	45.04192	27.13931	333.6886
4WEU_2	7 8 9 11 12 13 14 18 19 20 90 92 122	6H7J_1:SYM00010000 D:6H7J C:11 12 127 14 19 5	1	7.13852	45.04192	27.13931	333.6886
4WGV_1	11 12 13 127 14 19 26 3 5 6 7 8 9	6H7J_1:SYM00010000 D:6H7J C:11 12 127 14 19 5	1	12.98213	41.87646	27.09831	329.9495
4WGV_2	11 12 13 127 14 19 26 3 5 6 7 8 9	6H7J_1:SYM00010000 D:6H7J C:11 12 127 14 19 5	1	12.98213	41.87646	27.09831	329.9495
4WGW_1	11 12 13 127 14 19 26 3 5 6 7 8 9	6H7J_1:SYM00010000 D:6H7J C:11 12 127 14 19 5	1	12.98213	41.87646	27.09831	329.9495
4WGW_2	11 12 13 127 14 19 26 3 5 6 7 8 9	6H7J_1:SYM00010000 D:6H7J C:11 12 127 14 19 5	1	12.98213	41.87646	27.09831	329.9495
4ZG1_3	11 12 13 14 5 6 7 8 9	6H7J_1:SYM00010000 D:6H7J C:11 12 127 14 19 5	1	3.666822	42.001	27.99978	331.5434
4ZG1_4	11 12 14 15 20 28 5 6 7 8 9	6H7J_1:SYM00010000 D:6H7J C:11 12 127 14 19 5	1	8.95115	42.55636	26.20718	330.2288
4ZG1_6	12 13 14 6 7 8 9 11	6H7J_1:SYM00010000 D:6H7J C:11 12 127 14 19 5	1	3.245874	42.543	28.7195	332.7892
4ZG1_8	11 12 13 14 6 7 8 9	6H7J_1:SYM00010000 D:6H7J C:11 12 127 14 19 5	1	3.245874	42.543	28.7195	332.7892
4ZG1_9	11 12 13 14 5 6 7 8 9	6H7J_1:SYM00010000 D:6H7J C:11 12 127 14 19 5	1	3.666822	42.001	27.99978	331.5434
4ZG1_10	11 12 14 19 27 4 5 6 7 8 9	6H7J_1:SYM00010000 D:6H7J C:11 12 127 14 19 5	1	27.75473	41.10109	25.40945	328.5191
5DA4_2	7 8 9 127 11 12 13 14 19 26 5 6	6H7J_1:SYM00010000 D:6H7J C:11 12 127 14 19 5	1	4.868934	42.26092	27.50925	331.2057

Figure 3.3 Snapshot of the csv log file of the clustering result. In the first column, number after the '_' represents the index of the interfaces in a certain PDB file if that structure has more than one nanoboby:nanobody interfaces.

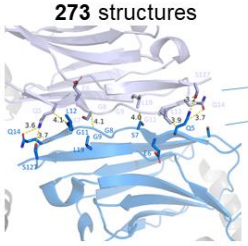
3.2.6 Visualising clustered nanobody-nanobody interfaces on the structure

To visualise the clustering result, each interface is represented by a point in the 3-dimensional space using the last three dimensions of the 4-dimensional vector (n^*, μ_x, μ_y, μ_z) generated for clustering. This point represents the rough position of the interface on the nanobody structure. Since the coordinate information was taken from the nanobody in 5O0W, I manually entered the clustered points into the PDB file of 5O0W as water molecules and displayed them as spheres in different colors in Pymol. Finally I obtained the plots in Figure 3.5B.

Figure 3.4 (Next Page) The computational workflow of nanobody-nanobody crystal contact analysis and classification results. The texts in red represent actions taken in each step. Structure cartoons in marine and light blue represent two nanobody molecules participating in the interface. Residues represented as sticks are interacting residues of the class of nanobody interfaces. Distances by the dashed lines are shown in Å.

335 crystal structures containing nanobody from PDB

273 structures



Nanobody-nanobody interface

Find all interfaces → Filter both sides are nanobodies

Side A Residue: 10 11 12 13 17 18 19 81

Side B Residue: 10 11 113 12 17 18 19 7 8 81 9

IMGT numbering scheme

Side A Residue: 11 12 13 14 18 19 20 90

Side B Residue: 11 12 122 13 19 20 8 9 90 10

Find communal interacting residues

One interface presented in an index array
11 12 13 19 20 90

Filter those having >5 communal residues → Assign coordinates to each residue

X 41.3 42.8 45.6 46.5 45.7 49.8
Y 30.6 33.2 32.3 24.9 21.9 20.5
Z 333.5 335.8 338.2 335.1 332.9 332.2
(Using coordinates on chain E of 5O0W)

Unify array dimension

356 crystal interfaces to analyze (**119** structures)

One interface presented in a coordinate vector

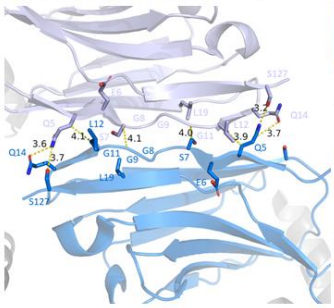
$(n^*, \mu_x, \mu_y, \mu_z)$

(2, 45.3, 27.2, 334.6)

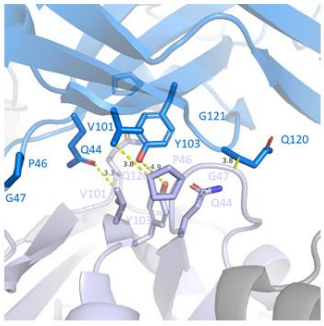
$n^* = n/3$

clustering → Density-based spatial clustering of applications with noise **(DBSCAN)**

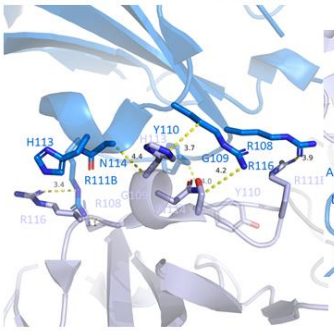
Class 1 (Nb-α) **47 (39.5%)**



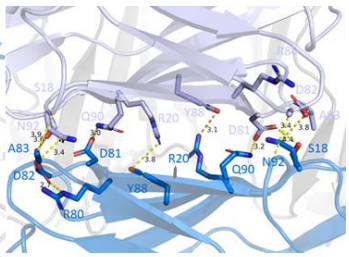
Class 2 **9 (7.6%)**



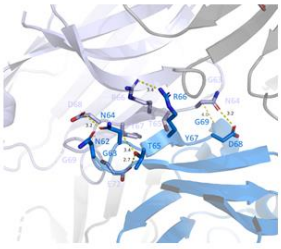
Class 3 **2 (1.7%)**



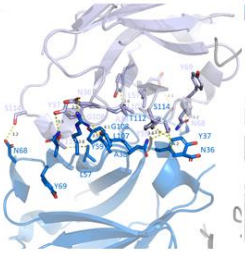
Class 4 **18 (15.1%)**



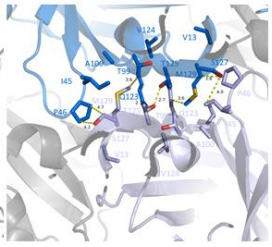
Class 5 **6 (5.0%)**



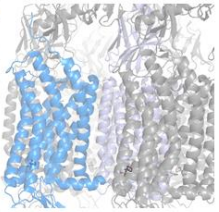
Class 6 **3 (2.5%)**



Class 7 **6 (5.0%)**

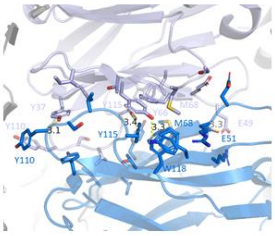


Class 8 **3 (2.5%)**



(Contacts by fusion parts of nanobodies)

Class 0 **6 (5.0%)**



3.3 Classification of nanobody-nanobody crystal contacts

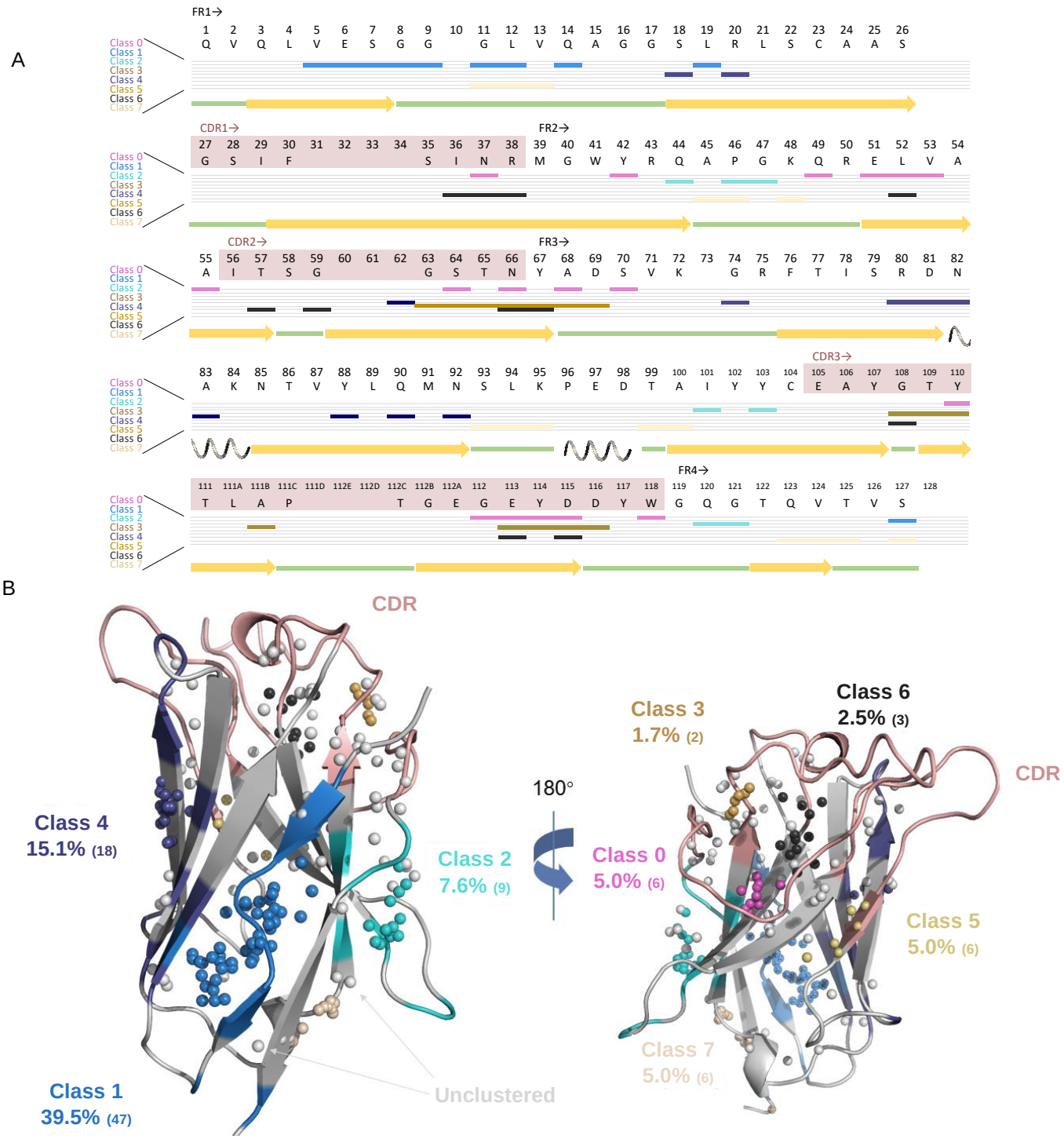


Figure 3.5 Nanobody-nanobody crystal contacts fall into limited classes. A) Standard nanobody residue numbering of RECQL5 nanobody Ctag:WT using IMGT scheme provided by ANARCI. CDRs are in light brown background and non-CDRs are in white background. Participating residues of each category are indicated by a bar below the one-letter code.

Figure 3.5 (Continued) B) DBSCAN clustering result: global view of major nanobody-nanobody crystal interfaces on a model drawn from chain E of 5O0W in PDB. Each colored sphere on the structure represents the average location of participating residues of an interface analyzed. Grey spheres represent ungrouped nanobody interfaces in the analysis. The three largest interface patterns are colored on the cartoon of the nanobody structure with respective class colors. Colours are consistent in A and B.

The clustering result shows 8 classes of nanobody-nanobody crystal contacts, 4 of them away from the CDRs, accounting for 39.5% (Class 1), 7.6% (Class 2), 15.1% (Class 4) and 5.0% (Class 7) of the total (Figure 3.5). The most prevalent 'Nb-α' (39.5%) is an edge-to-edge interaction of the two beta sheets at the N terminus (Figure 3.6). The Nb-α typically has 10 residues interacting in this type of contact, providing good engineering potential to further develop. The detailed analysis of the nine classes of interfaces are below (All residues numbers below are in the IMGT scheme).

3.3.1 Nb-α (Class 1) interfaces

Residues involved in this class of interfaces is typically 11, 12, 127, 14, 19, 5, 6, 7, 8, 9, which is altogether 10 residues. 47 structures out of the total 335 structures (14.03%) and 119 structures (39.50%) containing large nanobody:nanobody interfaces, have this class of interfaces. The PDB four-letter codes of the structures are listed below:

2X1P, 2XV6, 3P0G, 3K1K, 3STB, 4QGY, 4W6W, 4WEU 4WGV, 4WGW, 4ZG1, 5DA4, 5HDO, 5JA9, 5M2M, 5M2W, 5M7Q, 5M94, 5M95, 5O0W, 5U64, 5UK4, 6DBA, 6DBD, 6F0D, 6FUZ, 6FV0, 6H7J, 6H7L, 6H7M, 6H7N, 6H7O, 6IBL, 6OS1, 6OS2, 6Q6Z, 6QGY, 6U52, 6U53, 6U54, 6U55, 6VI4, 6WAR, 4DK3, 4KDT, 6GK4, 6ITQ.

The analysed interfaces belonging to this class are listed below:

Interface name	Interface pattern (Class 1)	Distance to the center of cluster
2X1P_1	11 12 120 121 122 127 5 6 7 8 9	25.2729
2X1P_4	11 12 120 121 122 127 5 6 7 8 9	25.2729
2X1P_7	11 12 120 121 122 127 5 6 7 8 9	25.2729
2X1P_8	11 12 120 121 122 127 5 6 7 8 9	25.2729
2XV6_1	11 12 13 18 19 93	31.4568
2XV6_2	11 12 13 18 19 93	31.4568
3P0G_1	11 12 127 14 19 26 3 5 6 7 8 9	17.1641
3STB_1	8 9 11 122 12 13 14 18 19 20 7	1.7267
4QGY_1	11 12 13 123 14 18 19 20 7 8 9	2.4367

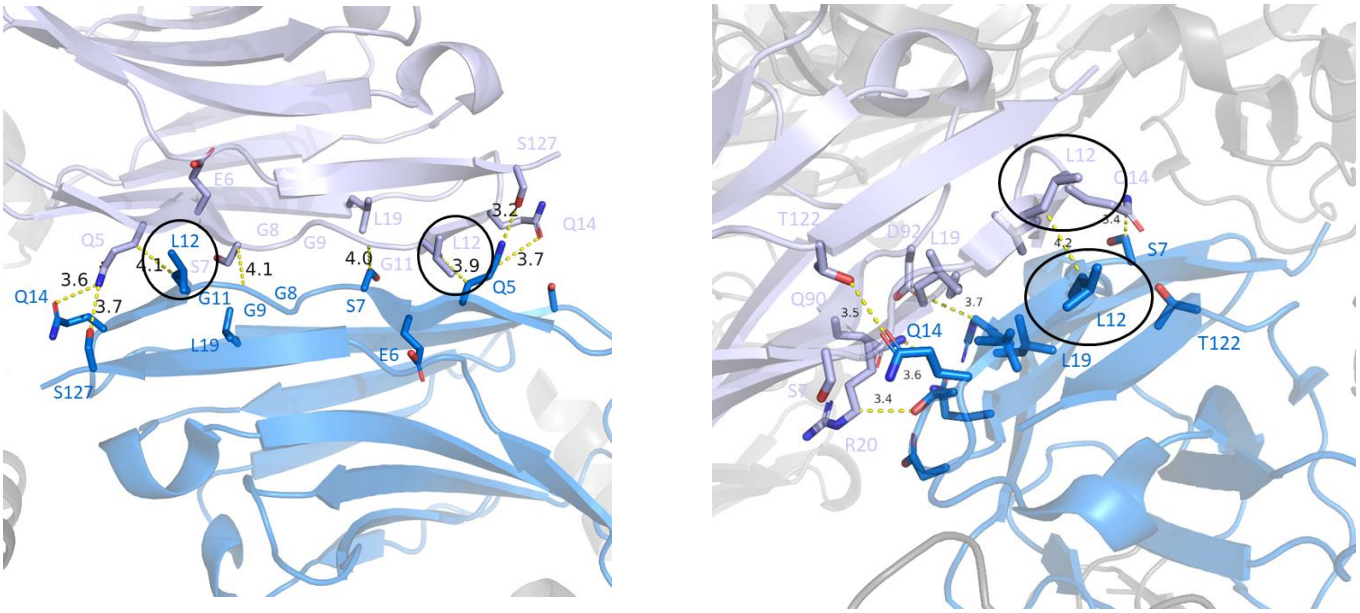
Interface name	Interface pattern (Class 1)	Distance to the center of cluster
4QGY_2	11 12 13 123 14 18 19 20 7 8 9	2.4367
4W6W_1	5 6 7 8 9 11 12 13 14 19 120 122 127	7.9594
4WEU_1	7 8 9 11 12 13 14 18 19 20 90 92 122	7.1385
4WEU_2	7 8 9 11 12 13 14 18 19 20 90 92 122	7.1385
4WGV_1	11 12 13 127 14 19 26 3 5 6 7 8 9	12.9821
4WGV_2	11 12 13 127 14 19 26 3 5 6 7 8 9	12.9821
4WGW_1	11 12 13 127 14 19 26 3 5 6 7 8 9	12.9821
4WGW_2	11 12 13 127 14 19 26 3 5 6 7 8 9	12.9821
4ZG1_3	11 12 13 14 5 6 7 8 9	3.6668
4ZG1_4	11 12 14 15 20 28 5 6 7 8 9	8.9512
4ZG1_6	12 13 14 6 7 8 9 11	3.2459
4ZG1_8	11 12 13 14 6 7 8 9	3.2459
4ZG1_9	11 12 13 14 5 6 7 8 9	3.6668
4ZG1_10	11 12 14 19 27 4 5 6 7 8 9	27.7547
5DA4_2	7 8 9 127 11 12 13 14 19 26 5 6	4.8689
5DA4_3	7 8 9 127 11 12 13 14 19 26 5 6	4.8689
5DA4_4	7 8 9 127 11 12 13 14 19 26 5 6	4.8689
5HDO_1	11 12 18 19 20 90 92	22.5683
5HDO_4	11 12 18 19 20 90 92	22.5683
5JA9_2	9 11 128 12 13 14 19 20 8	14.4289
5JA9_3	11 12 128 13 14 19 20 8 9	14.4289
5M2W_1	11 12 13 14 19 20	22.5182
5M2W_2	12 13 14 19 20 11	22.5182
5M7Q_1	11 12 13 14 18 19 20 7 8 9	3.8045
5M7Q_2	13 14 18 19 20 7 8 9	6.4079
5M94_1	12 13 127 14 19 26 3 5 6 7 8 9	13.1951
5M94_2	12 13 127 14 19 26 3 5 6 7 8 9	13.1951
5M95_1	11 12 13 127 14 19 26 3 5 6 7 8 9	12.9821
5M95_2	11 12 13 127 14 19 26 3 5 6 7 8 9	12.9821
5O0W_1	11 12 13 14 18 19 20 92 9	16.7481
5O0W_2	11 12 13 14 18 19 20 92 9	16.7481
5U64_1	14 16 17 22 23 24 26 11 12 108 13	13.3863
5UK4_1	11 12 13 14 18 19 20 9	14.4650
5UK4_2	11 12 13 14 18 19 20 9	14.4650
5UK4_4	11 12 13 14 18 19 20 9	14.4650
5UK4_5	11 12 13 14 18 19 20 9	14.4650
5UK4_6	11 12 13 14 18 19 20 9	14.4650

Interface name	Interface pattern (Class 1)	Distance to the center of cluster
5UK4_7	11 12 13 14 18 19 20 9	14.4650
5UK4_8	11 12 13 14 18 19 20 9	14.4650
5UK4_9	11 12 13 14 18 19 20 9	14.4650
5UK4_10	11 12 13 14 18 19 20 9	14.4650
5UK4_11	11 12 13 14 18 19 20 9	14.4650
5UK4_12	11 12 13 14 18 19 20 9	14.4650
5UK4_13	11 12 13 14 18 19 20 9	14.4650
5UK4_14	11 12 13 14 18 19 20 9	14.4650
5UK4_15	11 12 13 14 18 19 20 9	14.4650
5UK4_16	11 12 13 14 18 19 20 9	14.4650
5UK4_17	11 12 13 14 18 19 20 9	14.4650
5UK4_18	11 12 13 14 18 19 20 9	14.4650
5UK4_19	11 12 13 14 18 19 20 9	14.4650
5UK4_20	11 12 13 14 18 19 20 9	14.4650
5UK4_22	11 12 13 14 18 19 20 9	14.4650
6DBA_1	11 13 19 20 7 8 93 9	4.2314
6DBA_2	11 13 19 20 7 8 93 9	4.2314
6DBD_3	11 12 13 14 3 5 6 7 8 9	11.1694
6DBD_9	11 12 13 14 3 5 6 7 8 9	11.1694
6F0D_1	9 11 13 14 18 19 20 90 92 8	13.8735
6F0D_3	9 13 14 18 19 20 90 92 8	19.2629
6FUZ_1	11 12 13 18 19 20 90 92 9	13.2421
6FV0_1	11 12 13 14 18 19 20 90 92 9	16.1680
6H7J_1	11 12 127 14 19 5 6 7 8 9	1.4926
6H7J_2	11 12 127 14 19 5 6 7 8 9	1.4926
6H7L_1	11 12 127 13 14 19 5 6 7 8 9	1.5291
6H7L_2	11 12 127 13 14 19 5 6 7 8 9	1.5291
6H7M_1	11 12 127 13 14 19 5 6 7 8 9	1.5291
6H7M_2	11 12 127 13 14 19 5 6 7 8 9	1.5291
6H7N_1	11 12 127 14 19 5 6 7 8 9	1.4926
6H7N_2	11 12 127 14 19 5 6 7 8 9	1.4926
6H7O_1	11 12 127 13 14 19 26 5 6 7 8 9	4.8689
6H7O_2	11 12 127 13 14 19 26 5 6 7 8 9	4.8689
6IBL_1	11 12 127 14 19 5 6 7 8 9	1.4926
6IBL_2	11 12 127 14 19 5 6 7 8 9	1.4926
6OS1_1	11 12 13 14 18 19 20 8 9	8.5963
6OS2_1	11 12 13 14 18 19 20 8 9	8.5963
6Q6Z_1	1 11 12 14 26 3 5 6 7 8 9	61.9271
6QGY_1	9 11 122 12 13 14 17 18 19 20 7 90 8	7.0241

Interface name	Interface pattern (Class 1)	Distance to the center of cluster
6QGY_3	9 11 122 12 13 14 17 18 19 20 7 90 8	7.0241
6U52_1	11 12 13 5 6 7 8 9	11.1736
6U52_2	11 12 13 5 6 7 8 9	11.1736
6U53_1	11 12 13 14 18 19 20 7 8 98 100 9 107	7.9102
6U54_1	11 12 13 14 20 2 5 6 7 8 9	11.4941
6U55_1	11 120 12 127 128 13 14 19 26 27 3 5 6 7 8 9	28.6905
6VI4_1	8 9 11 12 14 17 18 19 20 5 84 6 88 7	14.6106
6VI4_2	8 9 11 12 14 17 18 19 20 5 84 6 88 7	14.6106
6WAR_1	11 120 123 12 26 5 9	49.9616
6WAR_3	11 120 123 12 26 5 9	49.9616
4DK3_1	14 19 26 7 8 9	18.7493
4DK3_3	11 12 120 13 127 14 19 26 5 6 7 8 9	10.6470
4KDT_1	11 12 13 123 15 20 21 27 4 5 6 7 8 9	23.6820
4KDT_2	12 13 14 16 21 22 28 4 5 6 7 8 9 11	19.6922
6GK4_1	11 12 13 14 19 5 6 7 8 9	1.5704
6GK4_2	11 12 13 14 19 5 6 7 8 9	1.5704
6ITQ_2	11 12 24 5 7 8 9	28.4388
6ITQ_3	11 12 24 5 7 8 9	28.4388

There are two subtypes of interfaces within class 1 interfaces, which are shifted or aligned interfaces depending on whether the residues 12 in this interface are shifted or aligned (Figure 3.6). Both subtypes feature an edge-to-edge beta sheet interaction, resulting a large interface area, which makes this class of interfaces ideal for engineering. Since this class of interfaces will be frequently mentioned afterwards and it

is prevalent in existing nanobody structures, it is named as 'Nb-α', meaning the largest class of large nanobody-nanobody crystal interfaces.



Nbinter-1: L12 Shifted

Nbinter-1: L12 Aligned

Figure 3.6 Typical close views of Class 1 (Nb-α) Nanobody:nanobody crystal contacts. L12 Shifted type is shown in the left panel, and L12 Aligned type is shown in the right panel. Structure cartoons in marine and light blue represent two nanobody molecules participating in the interface. Residues represented as sticks are interacting residues of the class of nanobody interfaces. Distances by the dashed lines are shown in Å.

3.3.2 Class 0 interfaces

The Indices of the residues involved in this class of interfaces are typically 110, 112, 113, 114, 115, 118, 37, 42, 49, 51, 52, 55, 64, 66, 68, 70, which is altogether 16 residues. 6 structures out of the total 335 structures (1.79%) and 119 structures (5.04%) containing large nanobody:nanobody interfaces, have this class of interfaces. The PDB four-letter codes of the structures are listed below:

2X1O, 5DA4, 5VL2, 6QTL, 6U50, 6DYX.

The analysed interfaces belonging to this class are listed below:

Interface name	Interface pattern (Class 0)	Distance to the center of cluster
2X1O_1	108 111A 112A 112 114 115 116 117 37 42 49 50 51 52 66 68 69	2.2032

Interface name	Interface pattern (Class 0)	Distance to the center of cluster
2X1O_3	108 111A 112A 112 114 115 116 117 37 42 49 50 51 52 66 68 69	2.2032
5DA4_1	112A 112 113 114 115 116 118 49 50 52 59 64 65 66 67	1.4592
5DA4_5	112A 112 113 114 115 116 118 49 50 52 59 64 65 66 67	1.4592
5DA4_6	111 112A 112 113 114 115 116 118 42 49 50 52 59 64 65 66 67	1.8605
6QTL_1	112 113 114 115 118 37 42 43 49 51 52 55 57 64 66 68 70	0.9259
6QTL_2	112 113 114 115 118 37 42 43 49 51 52 55 57 64 66 68 70	0.9259
6QTL_3	112 113 114 115 118 37 42 43 49 51 52 55 57 64 66 68 70	0.9259
6QTL_4	110 112 113 114 115 118 37 42 43 49 50 51 52 55 64 66 68 70	1.7542
6QTL_5	110 112 113 114 115 118 37 42 49 51 52 55 64 66 68 70	0.2456
6QTL_6	112 113 114 115 118 37 42 43 49 51 52 55 57 64 66 68 70	0.9259
6QTL_7	112 113 114 115 118 37 42 43 49 50 51 52 55 64 66 68 70	2.9195
6QTL_8	112 113 114 115 118 37 42 49 51 52 55 66 68 70	3.7925
6U50_1	115 117 119 38 49 50 52 57 58 66 67 103 108	9.9248
6DYX_1	113 114 115 116 117 38 39 40 54 59 66 68 70 71 72 109	6.1800

Class 0 interfaces only accounts for a small fraction among all kinds of interfaces, and it is not far from the CDR loops (Figure 3.5B). Therefore, even if it typically has 16 interacting residues with large interface area (Figure 3.7), it is not suitable for extensive engineering.

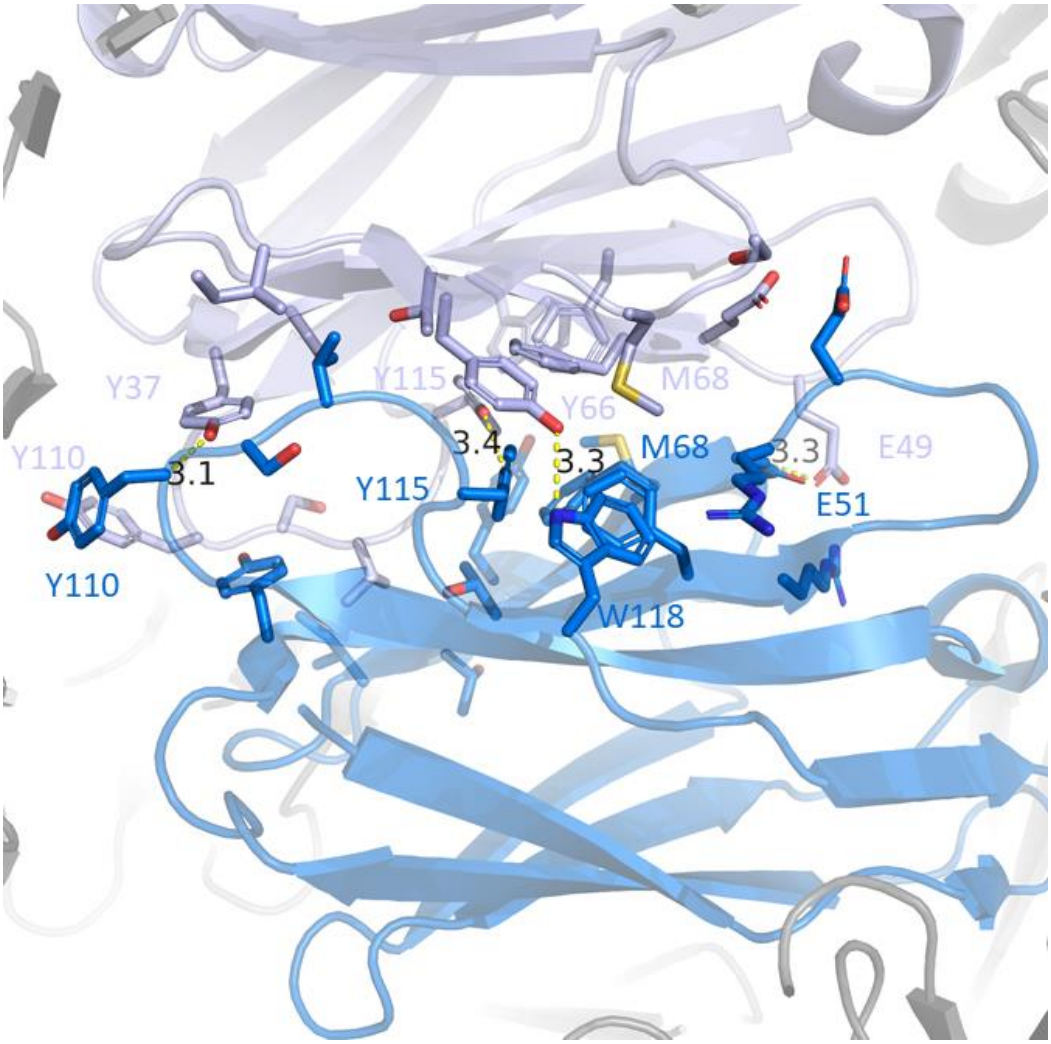


Figure 3.7 Typical close views of Class 0 Nanobody:nanobody crystal contacts. Structure cartoons in marine and light blue represent two nanobody molecules participating in the interface. Residues represented as sticks are interacting residues of the class of nanobody interfaces. Distances by the dashed lines are shown in Å.

3.3.3 Class 2 interfaces

The Indices of the residues involved in this class of interfaces are typically 120, 121, 44, 46, 47, 101, 103, which is altogether 7 residues. 9 structures out of the total 335 structures (2.69%) and 119 structures (7.56%) containing large nanobody:nanobody interfaces, have this class of interfaces. The PDB four-letter codes of the structures are listed below:

2X89, 4GRW, 5BOP, 5HDO, 5O8F, 5OCL, 6DBE, 6GKD, 6N50.

The analysed interfaces belonging to this class are listed below:

Interface name	Interface pattern (Class 2)	Distance to the center of cluster
2X89_1	118 120 123 44 47 50 103	3.3959
2X89_2	118 120 44 47 50 103	8.7538
4GRW_1	120 121 123 125 44 46 47 101 103	6.0060
4GRW_2	120 121 123 125 44 46 47 101 103	6.0060
5BOP_1	120 121 44 46 47 101 103	1.6002
5BOP_2	120 121 44 46 47 101 103	1.6002
5O8F_1	118 120 123 125 44 45 46 47 48 50 99 101 103	10.0068
5O8F_2	118 120 123 125 44 45 46 47 48 50 99 101 103	10.0068
5O8F_3	118 120 125 44 45 46 47 48 50 101 103	4.1238
5O8F_4	118 120 125 44 45 46 47 48 50 101 103	4.1238
5O8F_5	118 120 125 44 46 47 48 50 101 103	1.8238
5OCL_1	120 121 123 44 46 47 48 49 103	2.1430
5OCL_2	120 121 123 44 46 47 48 49 103	2.1430
6DBE_1	120 121 122 125 45 47 48 51 102 104	3.1282
6DBE_2	120 121 122 125 45 47 48 51 102 104	3.1282
6GKD_1	120 121 123 46 47 101 103	5.1517
6GKD_2	120 121 123 46 47 101 103	5.1517
6HD8_1	118 119 120 44 48 49	12.7569
6HD9_1	118 119 120 44 48 49 4	19.2019
6HDB_2	118 119 120 44 48 49	12.7569
6N50_1	115 118 44 45 46 47 101	5.8005

This class is the third largest class from the clustering result and the interface is also away from the CDR loops. However, the interface area is relatively small and therefore engineering potential is limited. The priority of engineering on this class of interfaces is low.

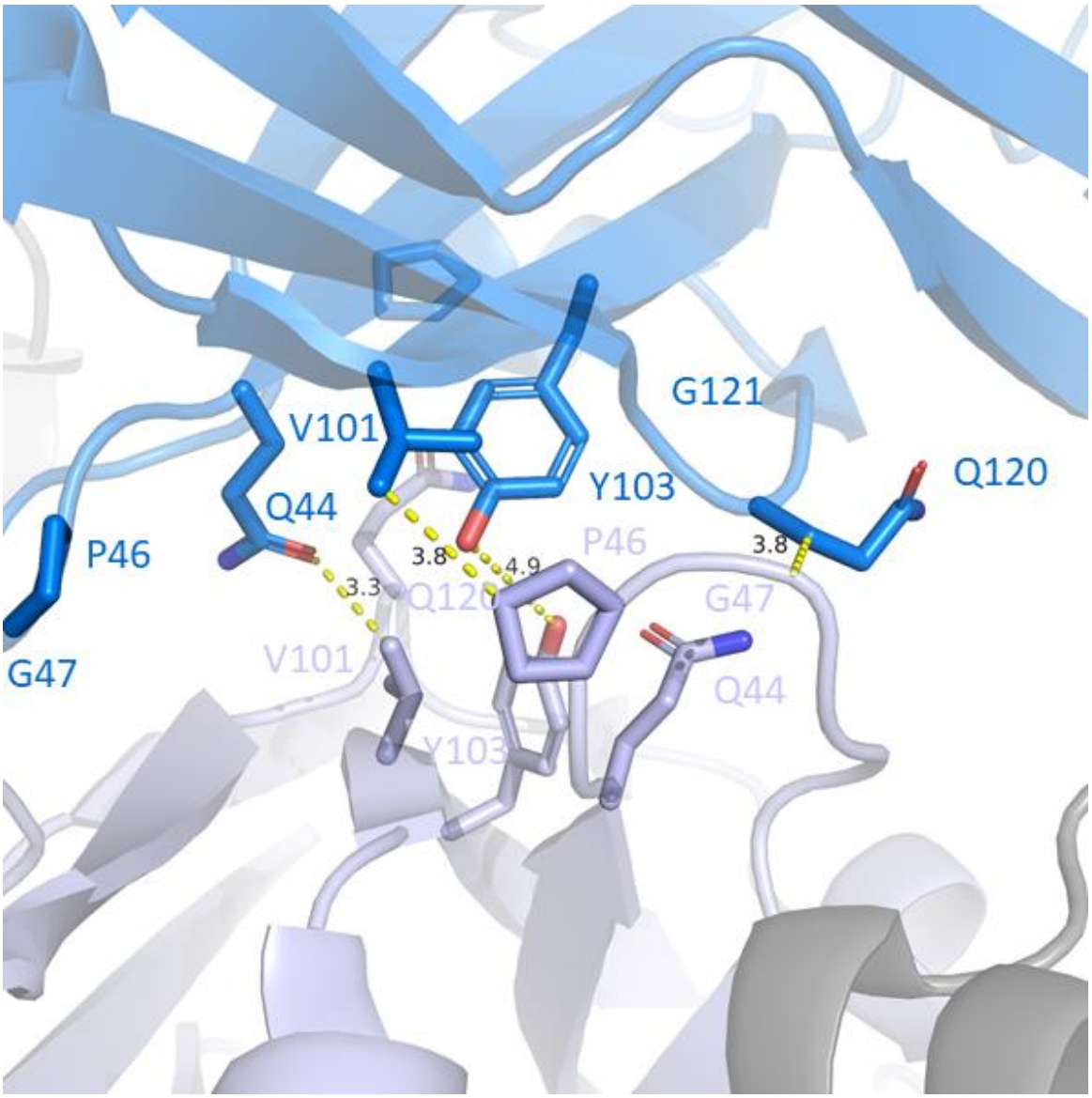


Figure 3.8 Typical close views of Class 2 Nanobody:nanobody crystal contacts. Structure cartoons in marine and light blue represent two nanobody molecules participating in the interface. Residues represented as sticks are interacting residues of the class of nanobody interfaces. Distances by the dashed lines are shown in Å.

3.3.4 Class 3 interfaces

The Indices of the residues involved in this class of interfaces are typically 108, 109, 110, 111B, 113, 114, 115, 116, which is altogether 8 residues. 2 structures out of the total 335 structures (0.60%) and 119 structures (1.68%) containing large nanobody:nanobody interfaces, have this class of interfaces. The PDB four-letter codes of the structures are listed below:

3DWT, 6F0D.

The analysed interfaces belonging to this class are listed below:

Interface name	Interface pattern (Class 3)	Distance to the center of cluster
3DWT_1	107 108 111B 113 114 115 116	1.5163
3DWT_3	108 109 110 111B 113 114 116 36	3.2324
3DWT_4	108 109 110 111B 113 114 116	1.2841
3DWT_5	108 109 110 111B 113 114 115 116 36	2.3922
3DWT_6	107 108 111B 113 114 116	1.0093
3DWT_7	107 108 111B 113 114 116	1.0093
3DWT_8	107 108 111B 113 114 115 116	1.5163
3DWT_9	108 109 110 111B 113 114 115 116	0.4091

This class of interfaces only account for a very small fraction among all kinds of interfaces, and it is also very near the CDR loops (Figure 3.5). Therefore, it is not an ideal site for engineering.

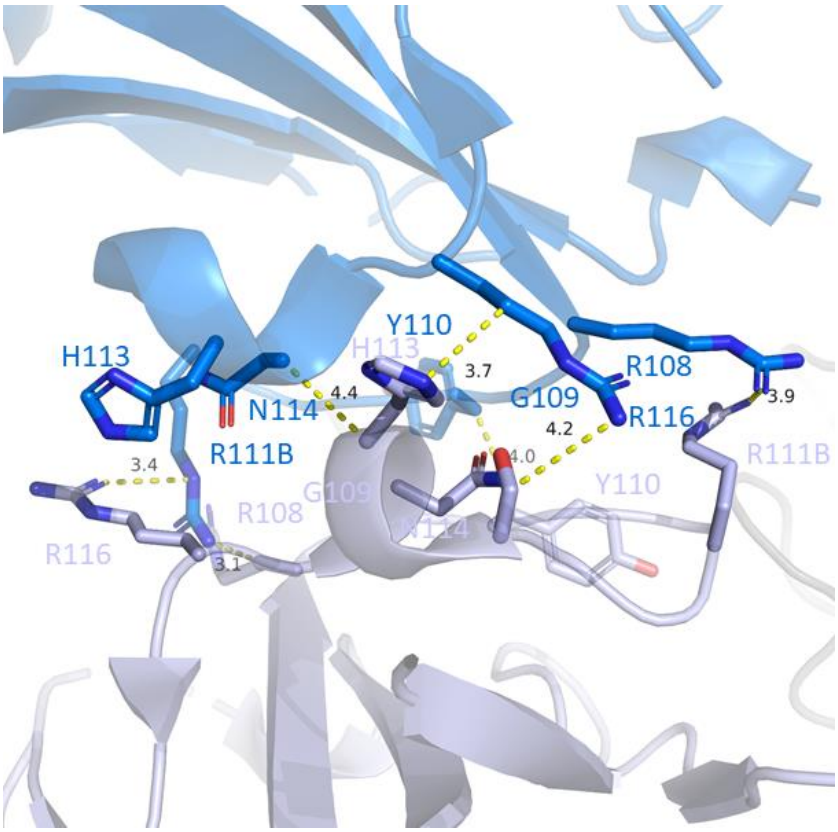


Figure 3.9 Typical close views of Class 3 Nanobody:nanobody crystal contacts.

Figure 3.9 (Continued) Structure cartoons in marine and light blue represent two nanobody molecules participating in the interface. Residues represented as sticks are interacting residues of the class of nanobody interfaces. Distances by the dashed lines are shown in Å.

3.3.5 Class 4 interfaces

The Indices of the residues involved in this class of interfaces are typically 18, 20, 62, 74, 80, 81, 82, 83, 88, 90, 92, which is altogether 11 residues. 18 structures out of the total 335 structures (5.37%) and 119 structures (15.13%) containing large nanobody:nanobody interfaces, have this class of interfaces. The PDB four-letter codes of the structures are listed below:

2X1P, 3EAK, 3K1K, 3OGO, 5E03, 5G5R, 5G5X, 5H8D, 5IMO, 5VL2, 5VNV, 6C9W, 6DBE, 6U51, 4DK3, 6GJQ, 6HHU, 6ITQ.

The analysed interfaces belonging to this class are listed below:

Interface name	Interface pattern (Class 4)	Distance to the center of cluster
3EAK_1	18 20 62 74 77 78 79 80 81 82 83 88 90 92	1.9949
3EAK_2	18 20 62 74 77 78 79 80 81 82 83 88 90 92	1.9949
3K1K_2	18 62 63 74 77 79 80 81 82 83 90 92	4.8158
3K1K_3	18 59 62 63 74 77 79 80 81 82 83 90 92	9.3553
3OGO_1	18 77 79 80 81 82 83 88 90 92	2.3299
3OGO_2	18 20 77 79 80 81 82 83 88 90 92	1.7636
3OGO_3	18 77 79 80 81 82 83 88 90 92	2.3299
3OGO_4	18 20 77 79 80 81 82 83 88 90 92	1.7636
5E03_1	18 20 77 78 79 80 81 82 90 92 93	6.4354
5G5R_1	18 20 63 74 77 78 79 80 81 82 83 90 92	1.0610
5G5X_1	18 20 77 79 80 81 82 83 88 90 92	1.7636
5VL2_1	18 20 62 74 80 81 82 83 90 92	1.4066
5VL2_4	18 20 62 74 80 81 82 90 92	4.3362
5VL2_5	18 20 74 80 81 82 83 90 92	3.9530
5VL2_6	18 80 81 82 83 90 92	12.2326
5VL2_7	18 20 80 81 82 83 88 90 92	7.4631
5VL2_9	18 20 80 81 82 83 90 92	9.5452

Interface name	Interface pattern (Class 4)	Distance to the center of cluster
5VL2_11	18 20 62 74 80 81 82 83 88 90 92	0.3769
5VL2_12	18 20 80 81 82 83 90 92	9.5452
5VNV_1	19 21 64 65 76 77 78 80 82 83 84 85 86 87 91 93 95	12.2789
6C9W_1	17 18 20 62 63 74 77 78 79 80 81 82 83 84 88 90 92	9.2854
6DBE_3	20 64 75 76 78 80 81 82 83 84 85 91 93	2.3372
6U51_1	18 20 59 65 72 74 76 77 79 80 84 90 92	12.8551
4DK3_4	20 59 62 63 74 77 79 80 81 92	14.5372
5E03_1	18 20 77 78 79 80 81 82 90 92 93	6.4354
6GJQ_1	18 20 77 78 79 80 81 82 83 88 90 92	0.8376
6GJQ_2	18 20 77 78 79 80 81 82 83 90 92	0.7444
6GJQ_3	18 20 77 78 79 80 81 82 83 90 92	0.7444
6GJQ_4	18 20 77 78 79 80 81 82 83 88 90 92	0.8376
6HHU_1	14 16 17 18 20 59 63 74 77 78 79 80 81 82 83 88 90 92	19.0106
6ITQ_1	18 20 62 63 74 75 77 78 79 80 81 82 83 90 92	6.1953
6ITQ_4	18 63 74 77 78 79 80 81 82 83 90 92	1.6846
6ITQ_5	18 20 62 63 74 75 77 78 79 80 81 82 83 90 92	6.1953
6ITQ_6	18 63 74 77 78 79 80 81 82 83 90 92	1.6846

This is the second largest class of interfaces from the clustering result. It is also far from the CDR loops (Figure 3.5) and the interaction features a piling of two pairs of beta sheets, resulting in a large interface area (Figure 3.10). This would be ideal for engineering and has the priority second to Nb-α.

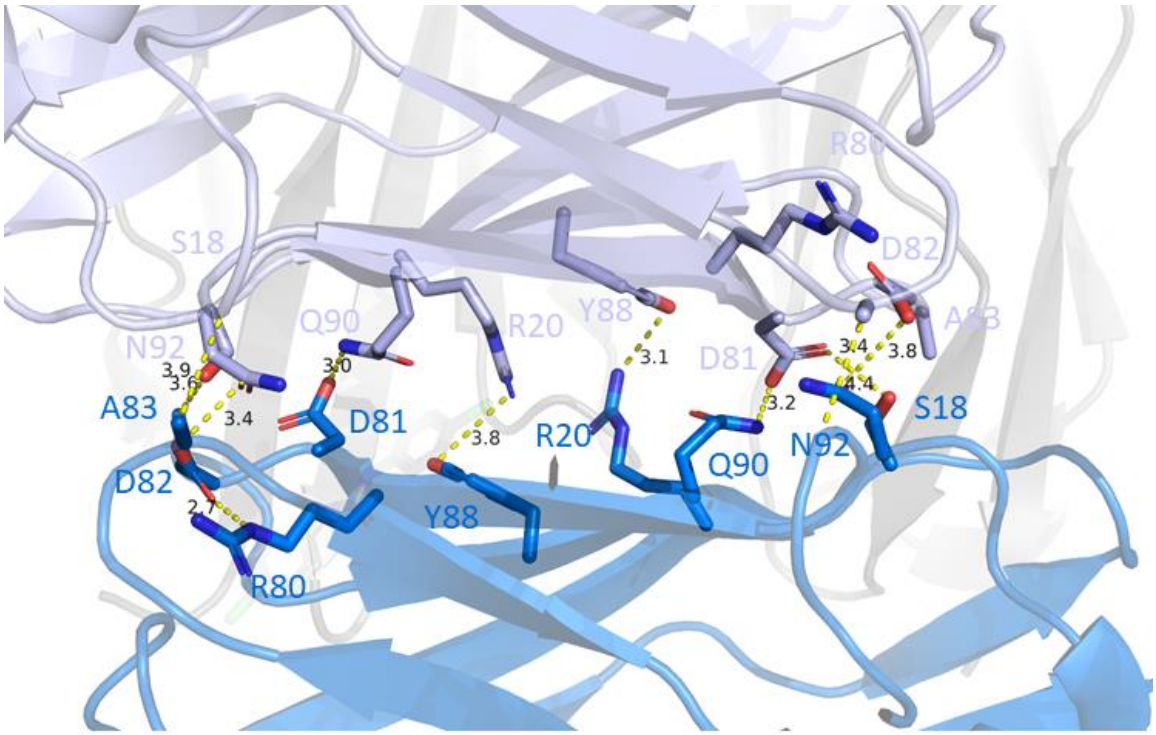


Figure 3.10 Typical close views of Class 4 Nanobody:nanobody crystal contacts. Structure cartoons in marine and light blue represent two nanobody molecules participating in the interface. Residues represented as sticks are interacting residues of the class of nanobody interfaces. Distances by the dashed lines are shown in Å.

3.3.6 Class 5 interfaces

The Indices of the residues involved in this class of interfaces are typically 63, 64, 65, 66, 67, 68, 69, which is altogether 7 residues. 6 structures out of the total 335 structures (1.79%) and 119 (5.04%) structures containing large nanobody:nanobody interfaces, have this class of interfaces. The PDB four-letter codes of the structures are listed below:

3EZJ, 5FWO, 5VNV, 5VM0, 5VM4, 6DBD.

The analysed interfaces belonging to this class are listed below:

Interface name	Interface pattern (Class 5)	Distance to the center of cluster
3EZJ_1	56 59 63 64 65 66 67 69 72	8.4926
3EZJ_2	56 59 63 64 65 66 67 69 72	8.4926
3EZJ_3	56 59 63 64 65 66 67 69 72	8.4926
3EZJ_4	56 59 63 64 65 66 67 69 72	8.4926

Interface name	Interface pattern (Class 5)	Distance to the center of cluster
5FWO_2	112B 63 64 65 66 67 68	3.6087
5VNV_2	40 59 63 66 67 68 71 75	4.4692
6DBD_1	63 64 65 66 67 68 69	1.3184
6DBD_2	63 64 65 66 67 68 69	1.3184
6DBD_6	59 63 64 65 67 68 69	7.2462
6DBD_7	63 64 65 66 67 68 69	1.3184
6HD8_2	52 64 65 66 67 68 69 72	4.6013
6HD9_2	52 64 65 66 67 68 69 72	4.6013
6HDA_2	52 64 65 66 67 68 69 72	4.6013
6HDB_1	52 64 65 66 67 68 69	3.3383
6HDC_1	52 64 65 66 67 68 69 72	4.6013
6ITP_1	111A 111B 112B 49 68 69 72 74 97	19.6507
6ITP_3	111A 111B 112B 49 68 69 72 74 97	19.6507

This class of interfaces only account for a small fraction among all kinds of interfaces and it is situated very close to the CDR loops (Figure 3.5). Although it features a good edge-to-edge beta sheet interaction (Figure 3.11), it is still not suitable for engineering.

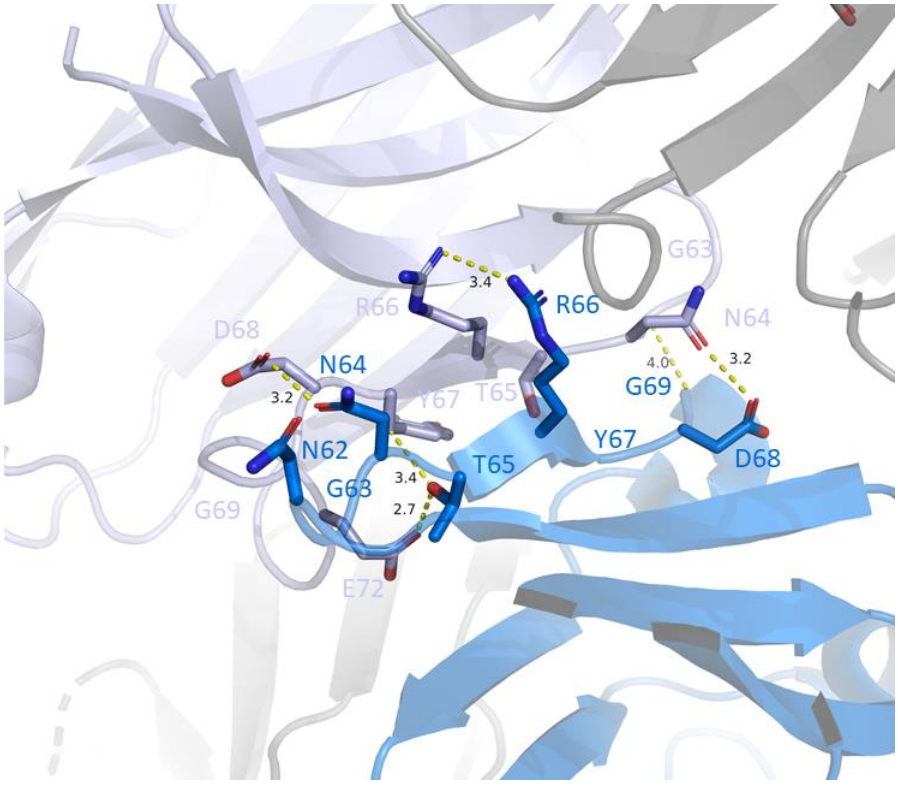


Figure 3.11 Typical close views of Class 5 Nanobody:nanobody crystal contacts.

Figure 3.11 (Continued) Structure cartoons in marine and light blue represent two nanobody molecules participating in the interface. Residues represented as sticks are interacting residues of the class of nanobody interfaces. Distances by the dashed lines are shown in Å.

3.3.7 Class 6 interfaces

The Indices of the residues involved in this class of interfaces are typically 108, 113, 115, 36, 37, 38, 52, 57, 59, 66, 67, which is altogether 11 residues. 3 structures out of the total 335 structures (0.90%) and 119 structures (2.52%) containing large nanobody:nanobody interfaces, have this class of interfaces. The PDB four-letter codes of the structures are listed below:

5NM0, 5NML, 6DBD.

The analysed interfaces belonging to this class are listed below:

Interface name	Interface pattern (Class 6)	Distance to the center of cluster
5NM0_1	110 37 38 39 53 58 59 108 109	5.2285
5NM0_3	110 37 38 39 53 58 59 65 108 109	2.6339
5NML_2	113 115 36 37 38 52 57 58 59 64 66 67 108	3.4722
5NML_3	108 112 114 36 37 38 52 57 59 64 68 107	1.4327
5NML_4	108 113 115 36 37 38 52 57 59 66 67	1.0475
5NML_5	36 37 38 57 64 107	16.6195
5NML_6	108 112 114 36 37 38 52 57 59 64 68 69 107	5.7348
5NML_7	108 112 114 36 37 38 52 57 66 68 107	5.2440
5NML_8	108 112 114 36 37 38 52 57 58 66 68 69 107	8.3581
5NML_9	36 37 38 57 64 107	16.6195
5NML_10	112 36 37 38 52 57 59 64 66 68 69 107	12.0078
5NML_13	112 36 37 38 52 57 66 68 69 107	14.5253
6DBD_4	110 115 116 37 38 57 59 107 109	11.0534
6DBD_5	110 115 116 37 38 57 59 107 109	11.0534

This class of interfaces only account for a small fraction among all kinds of interfaces and it is situated very close to the CDR loops (Figure 3.5). The feature of this interface also seems very random (Figure 3.12), and therefore it is not suitable for engineering.

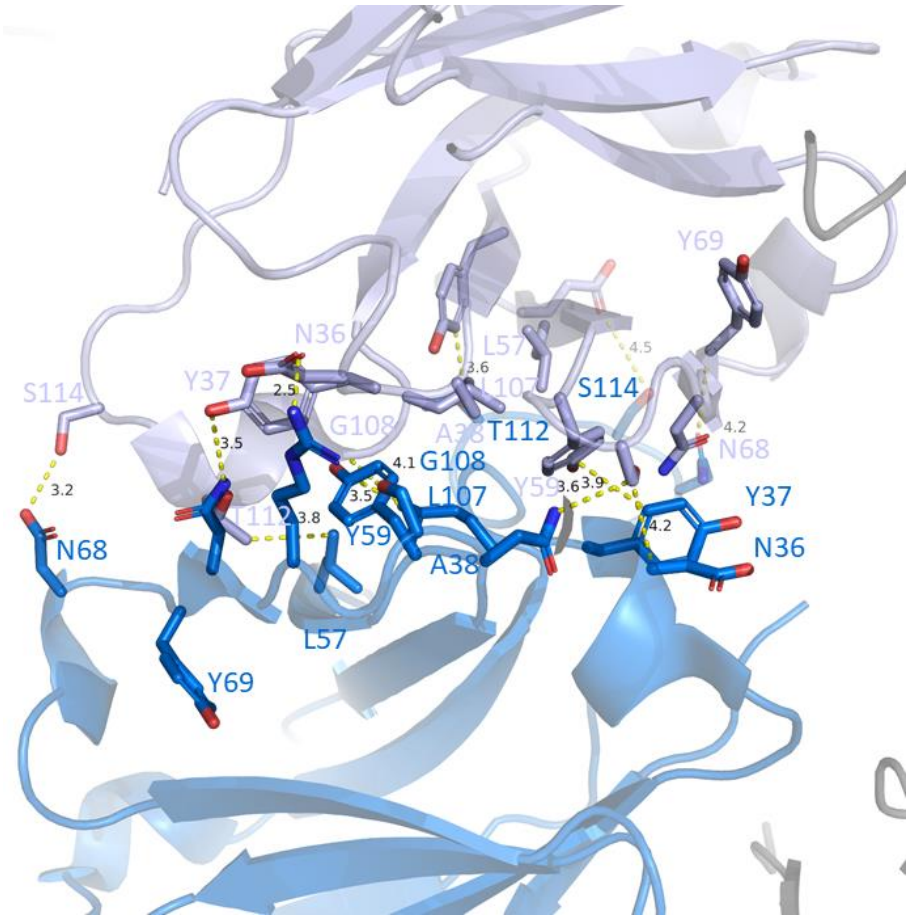


Figure 3.12 Typical close views of Class 6 Nanobody:nanobody crystal contacts. Structure cartoons in marine and light blue represent two nanobody molecules participating in the interface. Residues represented as sticks are interacting residues of the class of nanobody interfaces. Distances by the dashed lines are shown in Å.

3.3.8 Class 7 interfaces

The Indices of the residues involved in this class of interfaces are typically 11, 123, 124, 125, 12, 127, 13, 45, 46, 48, 99, 100, which is altogether 12 residues. 6 structures out of the total 335 structures (1.79%) and 119 structures (5.04%) containing large nanobody:nanobody interfaces, have this class of interfaces. The PDB four-letter codes of the structures are listed below:

5HDO, 5C1M, 5NML, 5OCL, 5VM4, 6QGY.

The analysed interfaces belonging to this class are listed below:

Interface name	Interface pattern (Class 7)	Distance to the center of cluster
5NML_11	123 125 126 127 128 15 46 96 99 101	9.7728
5NML_12	123 125 126 127 128 12 15 46 96 99 101	9.6886
5OCL_3	123 125 126 127 128 12 14 15 46 96 99 100 101	10.6091
5VM4_1	11 123 124 125 12 127 45 46 48 99 100 101	1.6469
5VM4_2	11 123 125 12 127 13 45 46 47 48 99 100	0.9132
5VM4_3	11 123 124 125 12 127 13 45 46 47 48 99 100 101	2.6739
5VM4_5	11 123 125 12 127 45 46 48 99 100 101	2.0428
5VM4_7	11 123 124 125 12 127 45 46 48 99 100 101	1.6469
5VM4_8	11 123 125 12 127 13 45 46 48 99 100 101	0.5429
5VM4_9	11 123 125 12 127 45 46 48 99 100	2.0426
5VM4_10	11 123 124 125 12 127 45 46 48 99 100	0.9810
5VM4_11	11 123 124 125 12 127 13 45 46 47 48 99 100 101	2.6739
5VM4_12	11 123 125 12 127 45 46 48 99 100	2.0426
5VM4_13	11 123 125 12 127 13 45 46 47 48 99 100	0.9132
5VM4_14	11 123 124 125 12 127 13 45 46 48 99 100	0.3967
6QGY_2	123 124 125 126 127 128 15 45 46 96 97 99 100 101	5.5007

This class of interfaces is situated far away from the CDR loops, but the interaction results in the nanobodies packing in an angle (Figure 3.13). Moreover, it only accounts for a small fraction among all kinds of interfaces, and therefore the priority for engineering around this site is low.

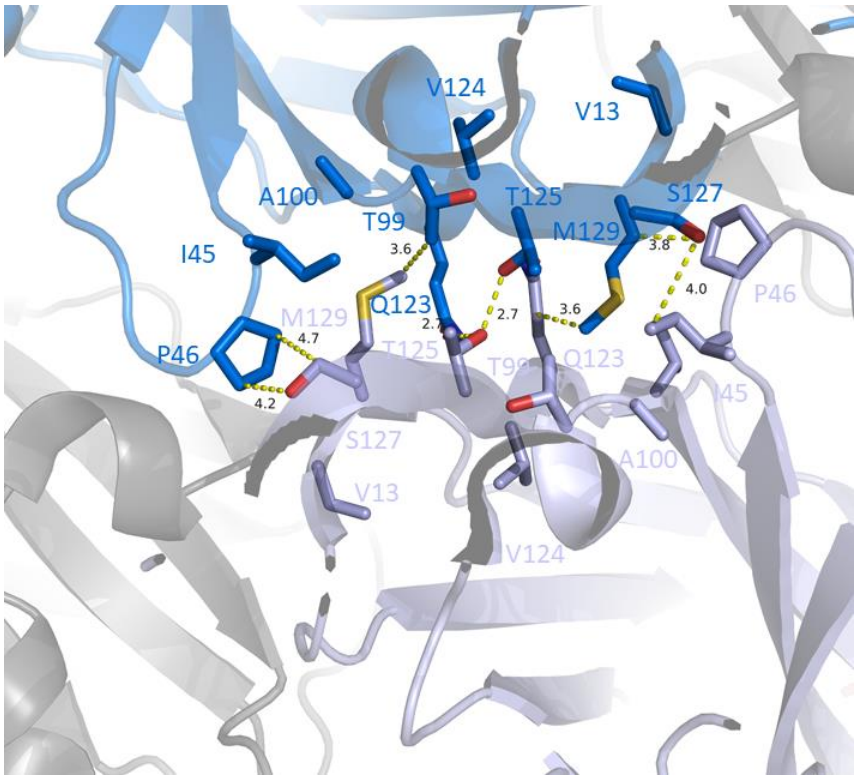


Figure 3.13 Typical close views of Class 7 Nanobody:nanobody crystal contacts. Structure cartoons in marine and light blue represent two nanobody molecules participating in the interface. Residues represented as sticks are interacting residues of the class of nanobody interfaces. Distances by the dashed lines are shown in Å.

3.3.9 Class 8 interfaces

The Indices of the residues involved in this class of interfaces are typically larger than 128, which means the C-terminal fusion parts of the nanobody. 3 structures out of the total 335 structures (0.90%) and 119 structures (2.52%) containing large nanobody:nanobody interfaces, have this class of interfaces. The PDB four-letter codes of the structures are listed below:

5WB1, 5WB2, 6QD6.

The analysed interfaces belonging to this class are listed below:

Interface name	Interface pattern (Class 8)	Distance to the center of cluster
5WB1_1	> 128	22.7599
5WB1_2	> 128	22.7599
5WB2_1	> 128	0.5319
5WB2_2	> 128	9.2336
6QD6_1	> 128	19.9529

Interface name	Interface pattern (Class 8)	Distance to the center of cluster
6QD6_2	> 128	22.7599
6QD6_3	> 128	9.2336
6QD6_4	> 128	2.4968
6QD6_5	> 128	17.5845
6QD6_6	> 128	10.9705
6QD6_7	> 128	9.2336
6QD6_8	> 128	15.1283
6QD6_9	> 128	1.0055
6QD6_10	> 128	13.0757
6QD6_11	> 128	15.1283
6QD6_12	> 128	10.9705
6QD6_15	> 128	4.6547
6QD6_16	> 128	10.9705
6QD6_17	> 128	25.4441

This class of interfaces is between the C-terminal fusions of some special kinds of nanobody derivatives (Figure 3.14), e.g. the megabody. It is not within our range of discussion.

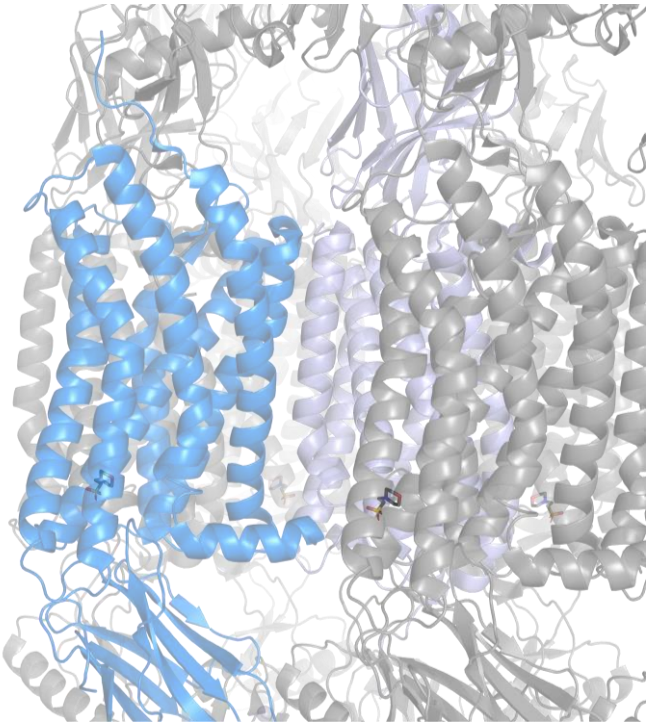


Figure 3.14 Typical close views of Class 8 Nanobody:nanobody crystal contacts. Structure cartoons in marine and light blue represent two nanobody molecules participating in the interface. Residues represented as sticks are interacting residues of the class of nanobody interfaces.

3.4 Discussion and conclusion

Through this bioinformatics analysis, I thoroughly summarised the nanobody-nanobody crystal contacts in the deposited structures. I have concluded that Nb-α is the predominant nanobody crystal contact (Figure 3.15) with the highest priority for engineering, and class 4 interfaces as the second engineering target.

Class	Cluster size	Average distance to the cluster center	Median distance to the cluster center
Nb-α (Class 1)	103	13.5811	13.2421
Class 0	15	2.5137	1.7542
Class 2	21	6.1138	5.1517
Class 3	8	1.5461	1.4002
Class 4	34	5.2980	3.1451
Class 5	17	6.7232	4.6013
Class 6	14	8.2154	7.0465
Class 7	16	3.3605	2.0426
Class 8	19	12.8366	10.9705

Figure 3.15 Nb-α class is the largest among the clustered interfaces but also the most spread out. The cluster size summarized the number of interfaces in each class. The average and median distances to the cluster center indicate how spread out the classes are.

However, if taking account of all 335 analysed structures, the percentage of Nb-α is only around 15%. This means that in the best-case scenario, I can only expect a maximum of 15% improvement of the nanobody as a crystal chaperone, after the scaffold optimization. Nevertheless, it is already the best interface I can start the engineering with. Moreover, it is notable that Nb-α falls into two sub-categories-aligned and shifted sharing similar interfacting residues on the scaffold, which means that this interface has some flexibility. Although Nb-α class is the largest cluster among all 9 interface classes, it is the most spread out (Figure 3.14) and takes a rod shape as is shown in Figure 3.5B. This means that Nb-α tends to slide within a certain region and thus possesses some flexibility. Therefore, mutations that anchor the interface and reduce the flexibility might optimize Nb-α to be a better crystal contact.

Chapter 4 Engineering the nanobody scaffold for reliable crystallization

4.1 Introduction

With the bioinformatics research in Chapter 3, it is confirmed that Nb-α is a prevalent nanobody large crystal contact among deposited structures. Engineering around this interface involves creating hydrogen bonds, salt bridges or covalent bonds connecting the two surfaces of the interface. With this strategy, hundreds of constructs were designed.

The engineering pipeline was then optimized in order to characterize all the designed constructs. High throughput cloning, and the parallel rapid expression purification and crystallization (PREPX) method were adopted for the establishment of the pipeline. The KRX strain (Cloning, expression one step strain) was used when doing cloning, which enabled direct protein expression from the same clone without the need of minipreps and retransformation into protein expression strains. The shifter was used for high-throughput crystal harvesting and the Unattended Data Collection (UDC) was used in Diamond for high throughput unattended overnight data collection. In the meantime, the yield of the nanobodies were further elevated by the introduction of the solubility mutations (G40T:Q49E:L52W:I101V, abbreviated as SOL) and the N-terminal MBP fusion. The whole pipeline was designed to have a capacity of 96-192 samples per run, which lasted a period of 8 days.

The whole engineering process started from the variant G1-001 in section 2.2, which yielded a crystal diffracting to 2.8 Å. The whole process finally took five iterations of mutagenesis of around 500 constructs to achieve the consistently good diffraction of the RECQL5:nanobody crystals. The first two iterations concluded the mutation S7N had an effect of enhancing the Nb-α with a pair of hydrogen bonds and significantly increased the crystallization propensity of the RECQL5:nanobody complex (G2-001). The best crystal could diffract to 2.4 Å. However, those crystals could not consistently diffract better than 2.5 Å, a prerequisite for crystallographic fragment screening. In order to increase the consistency of the diffraction quality, more variants were then explored. The third iteration discovered several different crystal polymorphs, among which the tetrameric nanobody crystal form appeared in many crystals, but they could not achieve high resolution either as the crystals formed in high salt conditions. L12C was confirmed to be crucial to trigger the tetrameric nanobody arrangement. The tetrameric nanobody core of the new crystal form was mediated by two sets of nanobody-nanobody interfaces – the Nb-α and another perpendicular interface featuring two

pairs of cysteines in the center (Cys12), which was then named the C-C interface. Several mutations were made around the C12 to try to enhance the C-C interface in the fourth and fifth iteration. Finally, the G5 variants performed far superior diffraction reliability (consistently <2.5 Å) with two additional key mutations T125M and Q14K in the C-C interface and Nb-α respectively.

Throughout the five iterations of mutagenesis, six different crystal polymorphs of the RECQL5:nanobody complex were discovered mediated by constitutive nanobody-nanobody crystal contacts. The Nb-α appeared in all six crystal forms and the C-C interface appeared in those containing tetrameric nanobodies. This suggested that the modified nanobodies were prone to stick together via those interfaces. Therefore, the modified scaffolds were called 'Gluebodies' and relative mutation sites were called 'Glue sites'.

Since the Glue sites are far from the CDR regions, theoretically they are easily transferable to other nanobodies. I selected another 5 intrinsically difficult target proteins and tested up to 3 G5 mutation sets for each of their nanobodies. As a result, the Gluebodies could help crystallize two of the five targets which could not co-crystallize with the original scaffold. However, those crystals obtained with Gluebodies either could not diffract to high resolution or could not yield a solvable dataset. Therefore, it could be concluded that the Gluebody is not significantly helpful to help intrinsically non-crystallizing crystals crystallize, but it was shown to increase crystal reliability and diversity with the example of RECQL5. Further screening is still on-going with those crystallized targets.

Although the statistics did not show significant improvement of the Gluebody scaffolds compared to the original, the structural comparison between the dimeric nanobody core and the tetrameric Gluebody core in the crystal lattice of the RECQL5:nanobody complex suggested that the C-terminal tail on the nanobody scaffold was likely to hang over the C-C interface to prevent the nanobody from forming further crystal contacts, reducing the crystallization propensity. Therefore, to explore the full potential of the nanobody as a crystallization chaperone, it would be necessary to test several nanobody scaffolds with or without the C-terminal tails and mutations in the Glue sites, namely C-tailed nanobody, di-Gluebody and tetra-Gluebody.

The Gluebody arrangement also has potential to be used as an imaging scaffold in cryo-EM. In the current arrangement, no covalent bond exists between the C12 residues in the C-C interface. Tetra-functional adapters or metal coordination could be explored to make a rigid molecule in solution as a tetravalent

imaging scaffold for cryo-EM.

4.2 Methods

Engineering on the nanobody scaffold involves a large amount of work. This requires a highly efficient engineering pipeline in function. The whole engineering process includes mainly three stages. The first stage is to design and clone the mutant nanobody variant, which requires high-throughput cloning in place. The second stage is to express, purify and crystallize the target:nanobody complex. This stage requires high yield of the nanobody, so that more variants can be expressed and purified in parallel. The final stage is data collection from crystals, which requires high-throughput crystal harvesting and automated data collection. These three stages fulfill a cycle of creating, producing and evaluating the variant nanobody, a prerequisite for good results (Figure 4.1).

4.2.1 High-throughput nanobody cloning method

In Structural Genomics Consortium (SGC), there is a mature ligation independent cloning (LIC) protocol for high-throughput mutagenesis. In this protocol, two PCR reactions are performed to amplify a target DNA fragment with designed point mutations. In the next step, T4 polymerase treatment for the PCR products and annealing of the inserts and digested vectors happen. Finally, annealed plasmids are transformed into DH5α, and selected clones are incubated for minipreps and retransformation into protein expressing strains, such as BL21 (DE3).

In this work, since the nanobody contains only around 120 amino acids, variants with multiple mutations were directly synthesized by TWIST Biosciences followed a one-step LIC cloning in house. These saved weeks of time, when three or more sites of point mutations were to be incorporated. The other improvement I did for this protocol was that I used cloning and expression one step strain (KRX), so that no miniprep and retransformation were required, saving two days in each engineering cycle.

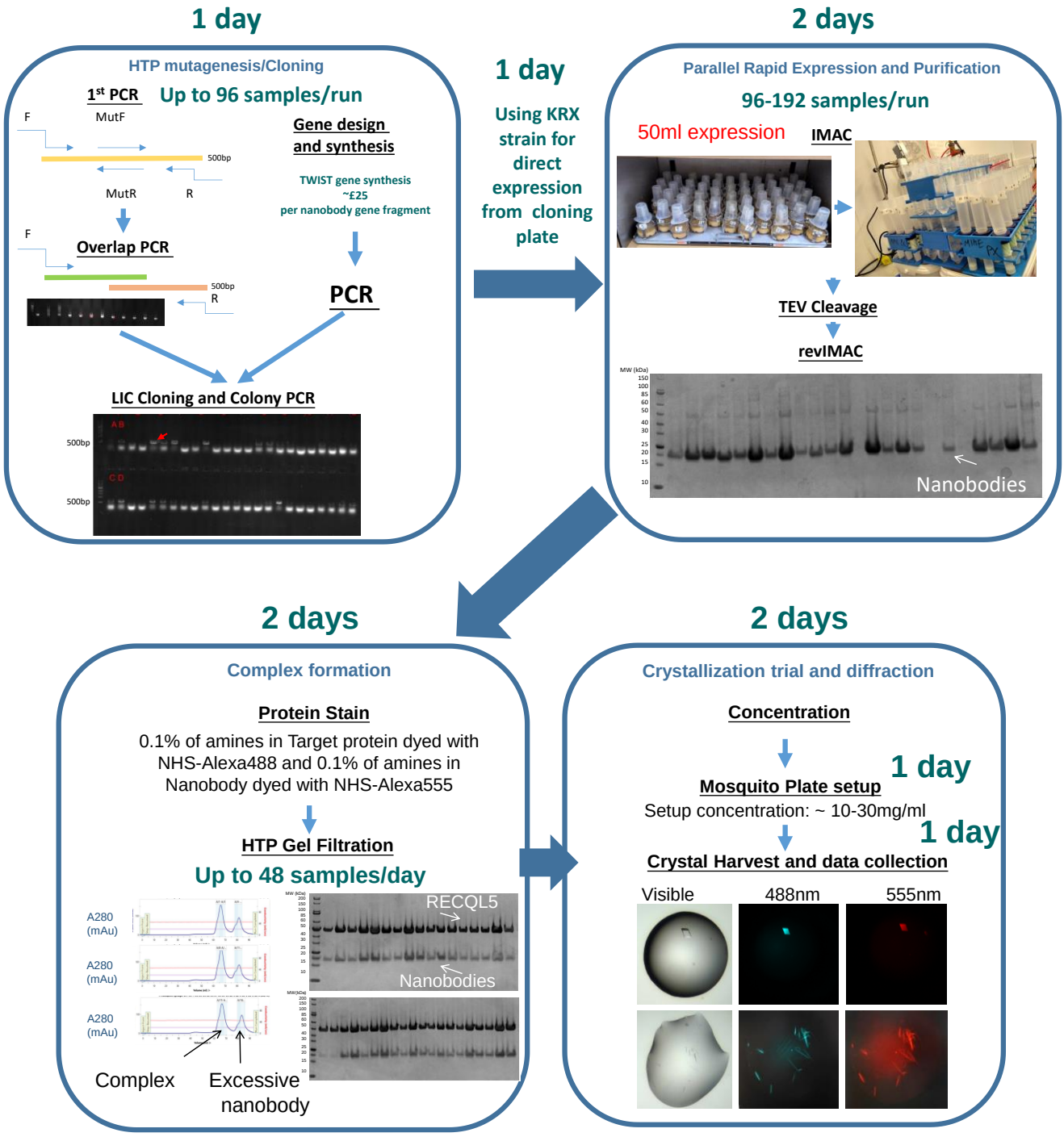


Figure 4.1 Scaffold engineering pipeline. The pipeline has four main processes indicated in each large box. In LIC Cloning and Colony PCR step of HTP mutagenesis, four colonies are normally picked for validation and the red arrow shows the right band position for a successful clone. The lowest band of the DNA ladder indicates 500bp. In PREP process, Ni-NTA columns are used for IMAC and revIMAC. Molecular weight of the marker bands is indicated to the left of the gel images in kDa. The bands around 15 kDa indicate nanobodies.

4.2.2 Improving yield of nanobody

The yield of the nanobody is the key to the high-throughput purification. To realise high-throughput purification, the yield of at least 20mg/L is required. It is necessary to find a generic method to elevate the yield of the nanobody.

In this experiment, I tested several strategies, including single charge mutation on the scaffold, fusing solubility tags on the nanobody and CDR grafting into well-expressing nanobody scaffold.

4.2.2.1 CDR grafting elevates the nanobody yield

The CDR grafting strategy is to graft the CDRs of the poorly expressing nanobodies onto the non-CDRs of the well-expressing nanobodies. Here I selected 34 poorly expressing nanobodies (samples 1-33 in Figure 4.2) and grafted their CDRs into the scaffold of a well-expressing nanobody, whose PDB id is 3K1K. The 66 nanobodies were synthesized by TWIST Biosciences (Sequences listed in Appendix Chapter 8.2) with MBP fusions on the N-terminus, and then I expressed and purified the nanobodies from 50 ml of media each following procedures detailed in section 2.2.1.3. As a result, nearly half of them were rescued to around 20 mg/L when using a highly expressing nanobody scaffold (samples 56-88 in Figure 4.2). The result shows that CDR grafting can effectively increase the yield of nanobody but evaluating the nanobody affinity again after grafting is still necessary.

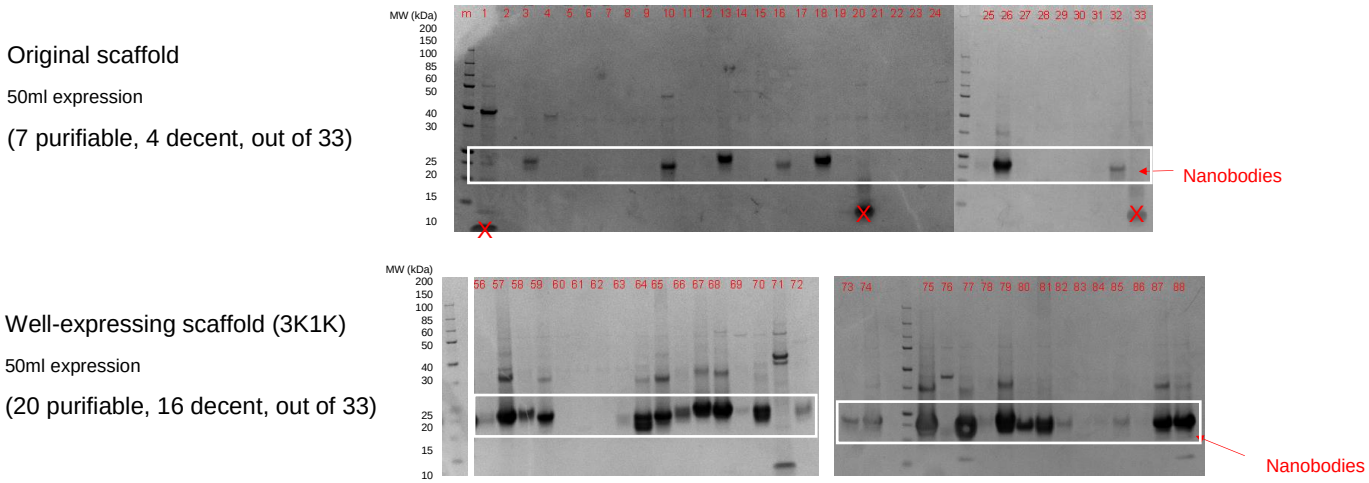


Figure 4.2 Scaffold grafting strategy significantly increase the yield of the nanobody. The bands in the white rectangles correspond to the nanobody bands. 15ul of purified nanobodies of each construct were loaded onto the SDS-PAGE. The details of the nanobody constructs are in Appendix Chapter 8.2.

4.2.2.2 Single charge mutagenesis does not elevate the nanobody yield significantly

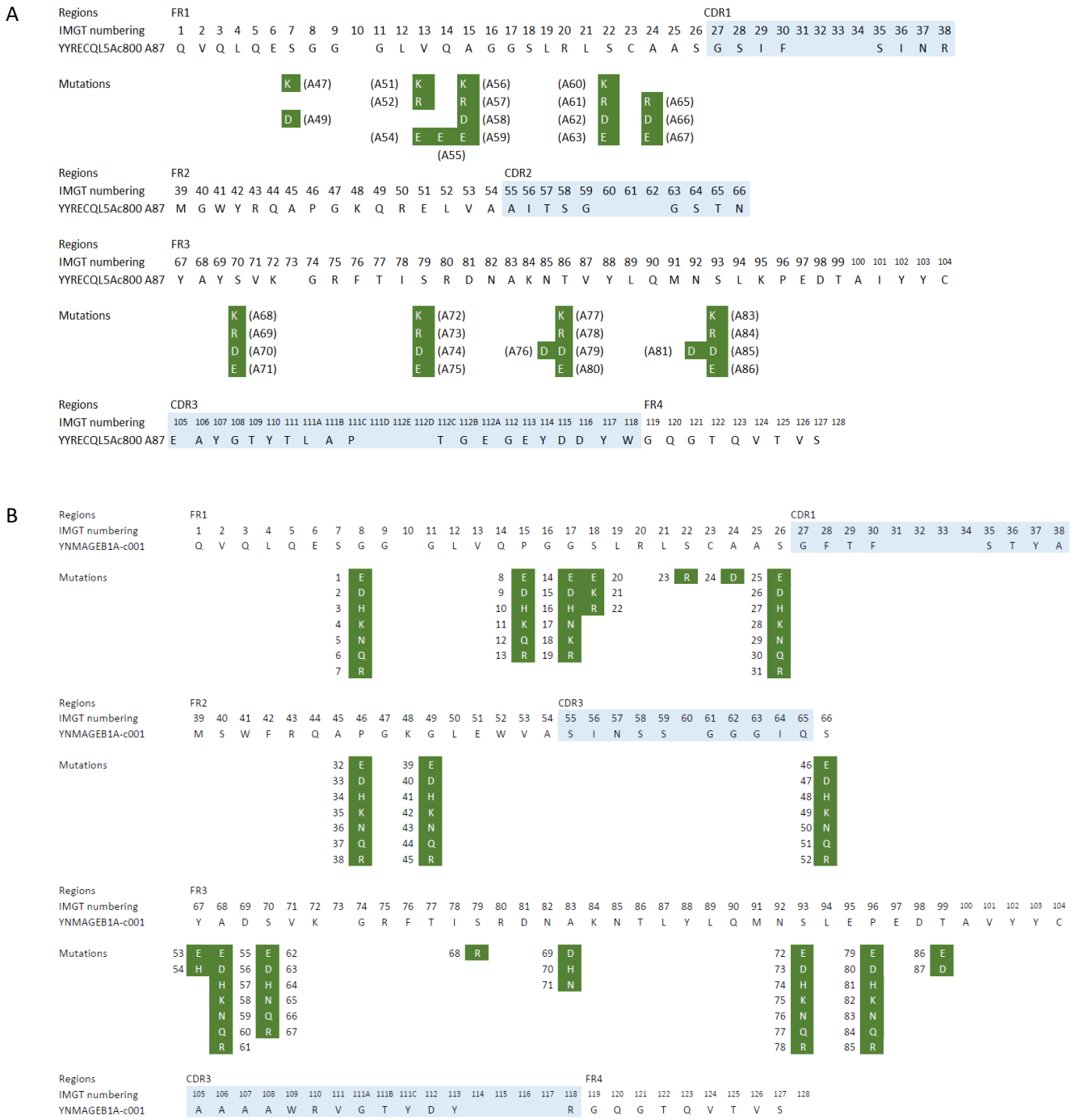


Figure 4.3 The design of charge mutations on nanobody scaffolds. The numbering of the sequences follows the IMGT scheme. Cyan regions indicate CDRs, while the rest are non-CDRs (or FRs). Each green square below the sequence represents a variant which has a mutation at the residue position to the indicated amino acid, with the sample ID labeled. The labels correspond to those in Figure 4.4. A) RECQL5 nanobody charge mutant designs. B) MAGEB1 nanobody charge mutant designs.

I selected two nanobodies as the template for mutagenesis – YYRECQL5Ac800 A87 (Figure 4.3A) and YNMAGEB1A-c001 (Figure 4.3B). Single mutations were designed as shown in Figure 4.3. The aim of the mutations was to introduce charges to the nanobody scaffolds and to evaluate whether it could help increase the yield. A few multi-residue mutants were also densigned for YNMAGEB1A listed below.

YNMAGEB1A-c001 variant	Original sequence	Mutation start	Mutation end	Mutated to
88	PGG	15	17	QGE
89	GS	17	18	EK
90	SCAAS	22	26	RCAAD
91	PGKG	46	49	DGKR
92	SYA	66	68	REE
93	ADS	68	70	EDE
94	SLEP	93	96	ELER
95	SYADS	66	70	REEDD
96	SY	66	67	RE

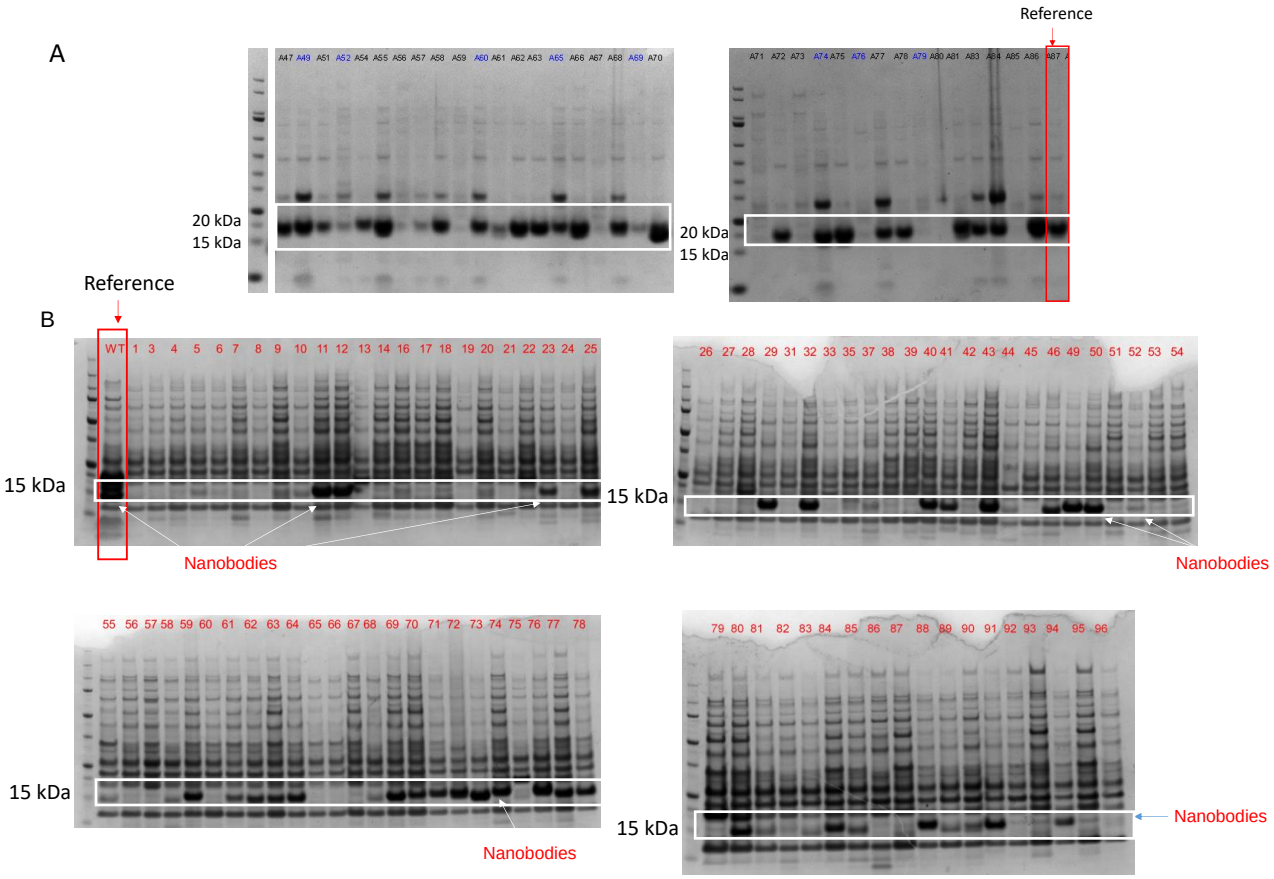


Figure 4.4 Charge mutants only help within the magnitude. The bands within the white rectangle correspond to nanobody bands. The details of the nanobody constructs are shown in Figure 4.3. The lane in the red rectangle is the wild-type construct. A) 15 µl of purified RECQL5 nanobodies with charge mutations were loaded onto the SDS-PAGE. B) 15 µl of MAGEB1 nanobodies after the first step of purification with charge mutations were loaded onto the SDS-PAGE.

I expressed the designed nanobodies in 100 ml culture for each and purified them following the procedures detailed in section 2.2.1.3. The MAGEB1 nanobody variants were only purified with Ni-NTA affinity purification. The result shown in Figure 4.4 indicates that the scaffold charge mutagenesis does not increase the nanobody yield further.

4.2.2.3 MBP fusion raises the nanobody yield

To test whether the MBP fusion raises the yield of nanobody, I selected 12 RECQL5 nanobodies and created their N-terminal MBP fusions (N1 – N12) and C-terminal 6XHIS fusions (C1 – C12). Those constructs were later referred to as G2*-004 – G2*-016 and G2-003 – G2-014. I expressed the designed nanobodies in 100 ml culture for each and purified them following the procedures detailed in section 2.2.1.3. The result shown in Figure 4.5 indicates tht MBP fusion helps raise the yield of RECQL5 nanobodies.

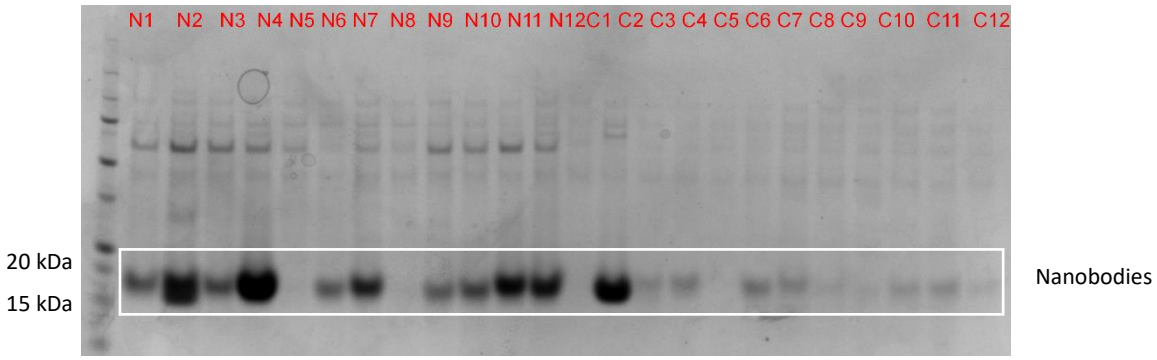


Figure 4.5 MBP fusion raises the nanobody yield. The bands within the white rectangle correspond to nanobody bands. Samples N1-N12 correspond to the 12 RECQL5 nanobodies with N-terminal MBP fusions, whereas C1-C12 correspond to those with C-terminal 6XHis tags. The lane in the red rectangle is the wild-type construct. 15 µl of each purified protein sample was loaded onto the SDS-PAGE.

Combining these result, I applied the CDR grafting strategy on the RECQL5 nanobody scaffold, which introduced 4 additional solubility mutations to the scaffold - G40T:Q49E:L52W:I101V, abbreviated as SOL, together with the N-terminal MBP fusion to further increase the yield of the RECQL5 nanobody.

4.2.3 Application of parallel rapid expression, purification and crystallization (PREPX) method to the nanobody

In the SGC, Michael Fairhead developed a parallel rapid expression, purification and crystallization (PREPX) method, by which up to 48 different constructs can be expressed, purified and put into crystal trials within one week. This is on condition that the selected constructs are highly expressing, and therefore each construct can be expressed in 100 ml of optimized TB autoinduction media. In this method, the most time-consuming step (bacteria cell lysis) in protein purification is replaced by freeze-thaw and chemical lysis, which enables high throughput. The size exclusion chromatography is also equipped with an auto-sampler, so that there is no need for stand-by manual injection. For the nanobody expression, since I further elevated the yield with the solubility mutations and the MBP fusion, only 50 ml of media was required for each construct and therefore the capacity was increased to 96 per week (Figure 4.1).

4.2.4 High-throughput crystal harvesting and data collection

To assess the quality of the crystals, it is necessary to mount them all and send them to the synchrotron for diffraction. The old school crystal harvesting method is very slow. Hence, I used the Shifter to assist with this process and the efficiency was increased by 3-4 fold. The harvested crystals were sent to the Diamond Light Source with unattended data collection (UDC) available on several beamlines. This allows the overnight data collection of hundreds of crystals without manual operation.

4.3 Engineering around Nb-α yielded various Gluebodies

With the crystal contact knowledge from Chapter 3 and the methodology preparation in section 4.1, engineering around Nb-α was able to take place.

4.3.1 Roadmap – 5 iterations of mutagenesis

As shown in Figure 4.6, in this work, all the RECQL5 nanobody variants fall into 6 generations (G0-G5). G0 represents the initial crystal epitope nanobody variants in Chapter 3, of which the best diffracting variant is represented as G1-001 (D69Y). The rest of G1 variants are Nb-α mutants based on G1-001, of which the best construct is represented as G2-001. The rest of G2 variants are Nb-α double mutants based on G2-001, and G2* variants are the N-terminal MBP tagged version of G2 variants. G3 variants were generated by combinatorial mutagenesis of selected G0 and G1 variants based on the G2-001 construct without D69Y. To test the optimal mutation of residue T125, G3* variants were generated on three different G3 constructs, and as a result T125M was selected. Finally, G3 variants plus T125M resulted in G4 and G5 variants. The IDs, mutations and sequences of the nanobody variants shown in Figure 4.6 are listed below (including G0 variants and G1-001). Multiple mutations are separated by ':'. The 'Ntag' or 'Ctag' at the beginning denotes that the tag is tethered at the N terminus or C terminus with a TEV cleavage site separating the nanobody and the tag. The template wild type sequence is shown in Figure 2.4 with IMGT numbering. All residue indices in the list below correspond to the numbering in Figure 2.4. If the varaiant has a 'Ctag', there is an additional

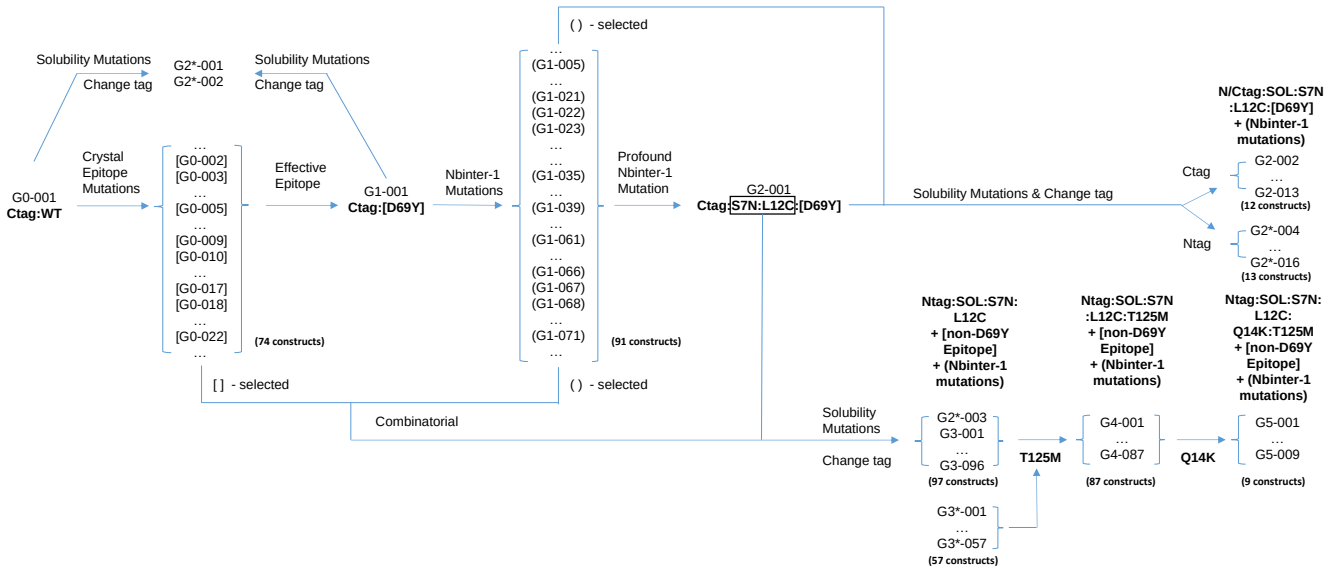


Figure 4.6 The road map of the several iterations of mutagenesis on RECQL5 nanobody scaffold. Variants in '[]' are selected crystal epitope mutants and in '()' are selected Nb-α mutants for subsequent iterations of mutagenesis.

'M' before the nanobody starting sequence 'QVQL' and an additional 'AENLYFQ' after the nanobody ending sequence 'VT(M)VS'. If the varaiant has an 'Ntag', there is an additional 'SMA' before the nanobody starting sequence 'QVQL'.

Nanobody protein ID	Key position mutations (S7/L12/Q14/T125)	Nb-α mutations	Epitope mutations	Solubility mutations? (G40T:Q49E: L52W:I101V)	Tag
G0-001	-	-	-	no	Ctag
G0-002	-	-	A54F:A55T	no	Ctag
G0-003	-	-	E6Q	no	Ctag
G0-004	-	-	L12W	no	Ctag
G0-005	-	-	T86A:V87L:Y88S	no	Ctag
G0-006	-	-	T78S	no	Ctag
G0-007	-	-	K84S	no	Ctag
G0-008	-	-	Q90E	no	Ctag
G0-009	-	-	Y88H	no	Ctag
G0-010	-	-	N82S:N85G	no	Ctag
G0-011	-	-	K95E	no	Ctag
G0-012	-	-	N85G	no	Ctag
G0-013	-	-	N82S	no	Ctag
G0-014	-	-	N85S	no	Ctag
G0-015	-	-	N82K:N85G	no	Ctag
G0-016	-	-	N82G:N85G	no	Ctag
G0-017	-	-	S79Q	no	Ctag
G0-018	-	-	R20K:L21V	no	Ctag
G0-019	-	-	A15P:G17D	no	Ctag
G0-020	-	-	A15P:G17A	no	Ctag
G0-021	-	-	A45V	no	Ctag
G0-022	-	-	Q44R	no	Ctag
G0-023	-	-	L12W:D69Y	no	Ctag
G0-024	-	-	A15P:G17A:D69Y	no	Ctag
G0-025	-	-	A45V:D69Y	no	Ctag
G0-026	-	-	D69Y:K84S	no	Ctag
G0-027	-	-	D69Y:Q90E	no	Ctag
G0-028	-	-	D69Y:Y88H	no	Ctag
G0-029	-	-	D69Y:N82S:N85G	no	Ctag
G0-030	-	-	D69Y:K95E	no	Ctag
G0-031	-	-	D69Y:N82K:N85G	no	Ctag
G0-032	-	-	D69Y:N82G:N85G	no	Ctag
G0-033	L12W	-	K84S	no	Ctag
G0-034	L12W	-	Q90E	no	Ctag
G0-035	L12W	-	Y88H	no	Ctag
G0-036	L12W	-	N82S:N85G	no	Ctag
G0-037	L12W	-	K95E	no	Ctag
G0-038	L12W	-	N82K:N85G	no	Ctag
G0-039	L12W	-	N82G:N85G	no	Ctag
G0-040	-	-	A15P:G17D:K84S	no	Ctag
G0-041	-	-	A15P:G17D:Q90E	no	Ctag
G0-042	-	-	A15P:G17D:Y88H	no	Ctag
G0-043	-	-	A15P:G17D:N82S:N85G	no	Ctag
G0-044	-	-	A15P:G17D:K95E	no	Ctag
G0-045	-	-	A15P:G17D:N82K:N85G	no	Ctag
G0-046	-	-	A15P:G17D:N82G:N85G	no	Ctag
G0-047	-	-	A15P:G17A:K84S	no	Ctag
G0-048	-	-	A15P:G17A:Q90E	no	Ctag
G0-049	-	-	A15P:G17A:Y88H	no	Ctag
G0-050	-	-	A15P:G17A:N82S:N85G	no	Ctag
G0-051	-	-	A15P:G17A:K95E	no	Ctag
G0-052	-	-	A15P:G17A:N82K:N85G	no	Ctag
G0-053	-	-	A15P:G17A:N82G:N85G	no	Ctag
G0-054	L12W	-	D69Y:K84S	no	Ctag
G0-055	L12W	-	D69Y:Q90E	no	Ctag
G0-056	L12W	-	D69Y:Y88H	no	Ctag
G0-057	L12W	-	D69Y:N82S:N85G	no	Ctag
G0-058	L12W	-	D69Y:K95E	no	Ctag
G0-059	L12W	-	D69Y:N82K:N85G	no	Ctag
G0-060	L12W	-	D69Y:N82G:N85G	no	Ctag
G0-061	-	-	D69Y:A15P:G17D:K84S	no	Ctag
G0-062	-	-	D69Y:A15P:G17D:Q90E	no	Ctag
G0-063	-	-	D69Y:A15P:G17D:Y88H	no	Ctag
G0-064	-	-	D69Y:A15P:G17D:N82S:N85G	no	Ctag
G0-065	-	-	D69Y:A15P:G17D:K95E	no	Ctag
G0-066	-	-	D69Y:A15P:G17D:N82K:N85G	no	Ctag
G0-067	-	-	D69Y:A15P:G17D:N82G:N85G	no	Ctag
G0-068	-	-	D69Y:A15P:G17A:K84S	no	Ctag
G0-069	-	-	D69Y:A15P:G17A:Q90E	no	Ctag
G0-070	-	-	D69Y:A15P:G17A:Y88H	no	Ctag
G0-071	-	-	D69Y:A15P:G17A:N82S:N85G	no	Ctag
G0-072	-	-	D69Y:A15P:G17A:K95E	no	Ctag
G0-073	-	-	D69Y:A15P:G17A:N82K:N85G	no	Ctag
G0-074	-	-	D69Y:A15P:G17A:N82G:N85G	no	Ctag
G1-001	-	-	D69Y	no	Ctag
G1-002	-	L12S	D69Y	no	Ctag
G1-003	L12T	-	D69Y	no	Ctag
G1-004	-	L12H	D69Y	no	Ctag
G1-005	L12N	-	D69Y	no	Ctag

Nanobody protein ID	Key position mutations (S7/L12/Q14/T125)	Nb-α mutations	Epitope mutations	Solubility mutations? (G40T:Q49E: L52W:I101V)	Tag
G1-006	-	L12Q	D69Y	no	Ctag
G1-007	-	L12E	D69Y	no	Ctag
G1-008	-	L12D	D69Y	no	Ctag
G1-009	-	L12Y	D69Y	no	Ctag
G1-010	-	S7H	D69Y	no	Ctag
G1-011	-	S7T	D69Y	no	Ctag
G1-012	-	S7N	D69Y	no	Ctag
G1-013	-	S7Q	D69Y	no	Ctag
G1-014	-	S7Y	D69Y	no	Ctag
G1-015	-	G8H	D69Y	no	Ctag
G1-016	-	G8K	D69Y	no	Ctag
G1-017	-	G8N	D69Y	no	Ctag
G1-018	-	G8Q	D69Y	no	Ctag
G1-019	-	G8R	D69Y	no	Ctag
G1-020	-	G8E	D69Y	no	Ctag
G1-021	-	G8D	D69Y	no	Ctag
G1-022	Q14D	-	D69Y	no	Ctag
G1-023	Q14K	-	D69Y	no	Ctag
G1-024	Q14H	-	D69Y	no	Ctag
G1-025	Q14R	-	D69Y	no	Ctag
G1-026	Q14N	-	D69Y	no	Ctag
G1-027	-	G16K	D69Y	no	Ctag
G1-028	-	G16R	D69Y	no	Ctag
G1-029	-	G16D	D69Y	no	Ctag
G1-030	-	G16E	D69Y	no	Ctag
G1-031	-	G16N	D69Y	no	Ctag
G1-032	-	G16Q	D69Y	no	Ctag
G1-033	-	G16Y	D69Y	no	Ctag
G1-034	-	G16H	D69Y	no	Ctag
G1-035	-	G16T	D69Y	no	Ctag
G1-036	-	G17K	D69Y	no	Ctag
G1-037	-	G17S	D69Y	no	Ctag
G1-038	-	G17D	D69Y	no	Ctag
G1-039	-	G17E	D69Y	no	Ctag
G1-040	-	G17N	D69Y	no	Ctag
G1-041	-	G17Q	D69Y	no	Ctag
G1-042	-	G17H	D69Y	no	Ctag
G1-043	-	G17T	D69Y	no	Ctag
G1-044	-	S18K	D69Y	no	Ctag
G1-045	-	S18R	D69Y	no	Ctag
G1-046	-	S18D	D69Y	no	Ctag
G1-047	-	S18E	D69Y	no	Ctag
G1-048	-	S18N	D69Y	no	Ctag
G1-049	-	S18Q	D69Y	no	Ctag
G1-050	-	S18T	D69Y	no	Ctag
G1-051	-	L19K	D69Y	no	Ctag
G1-052	-	L19R	D69Y	no	Ctag
G1-053	-	L19E	D69Y	no	Ctag
G1-054	-	L19Q	D69Y	no	Ctag
G1-055	-	L19N	D69Y	no	Ctag
G1-056	-	L19C	D69Y	no	Ctag
G1-057	-	S18C:R20C	D69Y	no	Ctag
G1-058	-	R20E	D69Y	no	Ctag
G1-059	-	R20D	D69Y	no	Ctag
G1-060	-	R20N	D69Y	no	Ctag
G1-061	-	R20K	D69Y	no	Ctag
G1-061	-	R20K	D69Y	no	Ctag
G1-062	-	R20Y	D69Y	no	Ctag
G1-063	-	R20H	D69Y	no	Ctag
G1-064	-	R20Q	D69Y	no	Ctag
G1-065	-	Q90K	D69Y	no	Ctag
G1-066	-	Q90R	D69Y	no	Ctag
G1-067	-	Q90Y	D69Y	no	Ctag
G1-067	-	Q90Y	D69Y	no	Ctag
G1-068	-	Q90H	D69Y	no	Ctag
G1-068	-	Q90H	D69Y	no	Ctag
G1-069	-	N92K	D69Y	no	Ctag
G1-070	-	N92R	D69Y	no	Ctag
G1-071	-	N92D	D69Y	no	Ctag
G1-071	-	N92D	D69Y	no	Ctag
G1-072	-	N92E	D69Y	no	Ctag
G1-073	-	N92Q	D69Y	no	Ctag
G1-074	-	N92Y	D69Y	no	Ctag
G1-075	-	N92H	D69Y	no	Ctag
G1-076	-	Q90C:N92C	D69Y	no	Ctag
G1-077	-	Q90E:N92E	D69Y	no	Ctag
G1-078	-	R20C	D69Y	no	Ctag
G1-079	-	S18H:R20H	D69Y	no	Ctag
G1-080	-	S18E:R20E	D69Y	no	Ctag

Nanobody protein ID	Key position mutations (S7N:L12Q:L14T:T125)	Nb-α mutations	Epitope mutations	Solubility mutations? (G40T:Q49E:L52W:I101V)	Tag
G1-081	-	S18Y:R20Y	D69Y	no	Ctag
G1-082	-	S18K:R20K	D69Y	no	Ctag
G1-083	-	S18Q:R20Q	D69Y	no	Ctag
G1-084	-	S18C:R20C	D69Y	no	Ctag
G1-085	-	S18H:R20Y	D69Y	no	Ctag
G1-086	-	S18H:R20K	D69Y	no	Ctag
G1-087	-	S18K:R20H	D69Y	no	Ctag
G1-088	-	S18Q:R20H	D69Y	no	Ctag
G1-089	-	S18Y:R20H	D69Y	no	Ctag
G1-090	-	S18H:R20Q	D69Y	no	Ctag
G1-091	-	S18N:R20Q	D69Y	no	Ctag
G2*-001	-	-	-	yes	Ntag
G2*-002	-	-	D69Y	yes	Ntag
G2*-003	S7N:L12C	-	-	yes	Ntag
G2*-004	S7N:L12C	-	D69Y	yes	Ntag
G2*-004	S7N:L12C	-	D69Y	yes	Ntag
G2*-004	S7N:L12C	-	D69Y	yes	Ntag
G2*-004	S7N:L12C	-	D69Y	yes	Ntag
G2*-004	S7N:L12C	-	D69Y	yes	Ntag
G2*-004	S7N:L12C	-	D69Y	yes	Ntag
G2*-005	S7N:L12C	G8D	D69Y	yes	Ntag
G2*-006	S7N:L12N	-	D69Y	yes	Ntag
G2*-007	S7N:L12C:Q14D	-	D69Y	yes	Ntag
G2*-008	S7N:L12C:Q14K	-	D69Y	yes	Ntag
G2*-009	S7N:L12C	G17E	D69Y	yes	Ntag
G2*-010	S7N:L12C	R20K	D69Y	yes	Ntag
G2*-011	S7N:L12C	R20K	D69Y	yes	Ntag
G2*-012	S7N:L12C	Q90Y	D69Y	yes	Ntag
G2*-013	S7N:L12C	Q90H	D69Y	yes	Ntag
G2*-014	S7N:L12C	Q90H	D69Y	yes	Ntag
G2*-015	S7N:L12C	N92D	D69Y	yes	Ntag
G2*-016	S7N:L12C	-	D69Y	yes	Ntag
G2-001	S7N:L12C	-	D69Y	yes	Ctag
G2-002	S7N:L12C	G8D	D69Y	yes	Ctag
G2-003	S7N:L12N	-	D69Y	yes	Ctag
G2-004	S7N:L12C:Q14D	-	D69Y	yes	Ctag
G2-005	S7N:L12C:Q14K		D69Y	yes	Ctag
G2-006	S7N:L12C	G16T	D69Y	yes	Ctag
G2-007	S7N:L12C	G17E	D69Y	yes	Ctag
G2-008	S7N:L12C	R20K	D69Y	yes	Ctag
G2-009	S7N:L12C	Q90R	D69Y	yes	Ctag
G2-010	S7N:L12C	Q90Y	D69Y	yes	Ctag
G2-011	S7N:L12C	Q90H	D69Y	yes	Ctag
G2-012	S7N:L12C	N92D	D69Y	yes	Ctag
G2-013	S7N:L12C	-	D69Y	yes	Ctag
G3*-001	S7N:L12C:T125A	G16T	N82S:N85G	yes	Ntag
G3*-002	S7N:L12C:T125C	G16T	N82S:N85G	yes	Ntag
G3*-003	S7N:L12C:T125D	G16T	N82S:N85G	yes	Ntag
G3*-004	S7N:L12C:T125E	G16T	N82S:N85G	yes	Ntag
G3*-005	S7N:L12C:T125F	G16T	N82S:N85G	yes	Ntag
G3*-006	S7N:L12C:T125G	G16T	N82S:N85G	yes	Ntag
G3*-007	S7N:L12C:T125H	G16T	N82S:N85G	yes	Ntag
G3*-008	S7N:L12C:T125I	G16T	N82S:N85G	yes	Ntag
G3*-009	S7N:L12C:T125K	G16T	N82S:N85G	yes	Ntag
G3*-010	S7N:L12C:T125L	G16T	N82S:N85G	yes	Ntag
G3*-011 =G4-043	S7N:L12C:T125M	G16T	N82S:N85G	yes	Ntag
G3*-012	S7N:L12C:T125N	G16T	N82S:N85G	yes	Ntag
G3*-013	S7N:L12C:T125P	G16T	N82S:N85G	yes	Ntag
G3*-014	S7N:L12C:T125Q	G16T	N82S:N85G	yes	Ntag
G3*-015	S7N:L12C:T125R	G16T	N82S:N85G	yes	Ntag
G3*-016	S7N:L12C:T125S	G16T	N82S:N85G	yes	Ntag
G3*-017	S7N:L12C:T125V	G16T	N82S:N85G	yes	Ntag
G3*-018	S7N:L12C:T125W	G16T	N82S:N85G	yes	Ntag
G3*-019	S7N:L12C:T125Y	G16T	N82S:N85G	yes	Ntag
G3*-020	S7N:L12N:T125A	-	N82S:N85G	yes	Ntag
G3*-021	S7N:L12N:T125C	-	N82S:N85G	yes	Ntag
G3*-022	S7N:L12N:T125D	-	N82S:N85G	yes	Ntag
G3*-023	S7N:L12N:T125E	-	N82S:N85G	yes	Ntag
G3*-024	S7N:L12N:T125F	-	N82S:N85G	yes	Ntag
G3*-025	S7N:L12N:T125G	-	N82S:N85G	yes	Ntag
G3*-026	S7N:L12N:T125H	-	N82S:N85G	yes	Ntag
G3*-027	S7N:L12N:T125I	-	N82S:N85G	yes	Ntag
G3*-028	S7N:L12N:T125K	-	N82S:N85G	yes	Ntag
G3*-029	S7N:L12N:T125L	-	N82S:N85G	yes	Ntag
G3*-030 =G4-050	S7N:L12N:T125M	-	N82S:N85G	yes	Ntag
G3*-031	S7N:L12N:T125N	-	N82S:N85G	yes	Ntag
G3*-032	S7N:L12N:T125P	-	N82S:N85G	yes	Ntag
G3*-033	S7N:L12N:T125Q	-	N82S:N85G	yes	Ntag
G3*-034	S7N:L12N:T125R	-	N82S:N85G	yes	Ntag
G3*-035	S7N:L12N:T125S	-	N82S:N85G	yes	Ntag
G3*-036	S7N:L12N:T125V	-	N82S:N85G	yes	Ntag
G3*-037	S7N:L12N:T125W	-	N82S:N85G	yes	Ntag
G3*-038	S7N:L12N:T125Y	-	N82S:N85G	yes	Ntag
G3*-039	S7N:L12C:T125A	Q90Y	N82S:N85G	yes	Ntag
G3*-040	S7N:L12C:T125C	Q90Y	N82S:N85G	yes	Ntag
G3*-041	S7N:L12C:T125D	Q90Y	N82S:N85G	yes	Ntag

Nanobody protein ID	Key position mutations (S7N:L12Q:L14T:T125)	Nb-α mutations	Epitope mutations	Solubility mutations? (G40T:Q49E:L52W:I101V)	Tag
G3*-042	S7N:L12C:T125E	Q90Y	N82S:N85G	yes	Ntag
G3*-043	S7N:L12C:T125F	Q90Y	N82S:N85G	yes	Ntag
G3*-044	S7N:L12C:T125G	Q90Y	N82S:N85G	yes	Ntag
G3*-045	S7N:L12C:T125H	Q90Y	N82S:N85G	yes	Ntag
G3*-046	S7N:L12C:T125I	Q90Y	N82S:N85G	yes	Ntag
G3*-047	S7N:L12C:T125K	Q90Y	N82S:N85G	yes	Ntag
G3*-048	S7N:L12C:T125L	Q90Y	N82S:N85G	yes	Ntag
G3*-049 =G4-047	S7N:L12C:T125M	Q90Y	N82S:N85G	yes	Ntag
G3*-049 =G4-047	S7N:L12C:T125M	Q90Y	N82S:N85G	yes	Ntag
G3*-050	S7N:L12C:T125N	Q90Y	N82S:N85G	yes	Ntag
G3*-051	S7N:L12C:T125P	Q90Y	N82S:N85G	yes	Ntag
G3*-052	S7N:L12C:T125Q	Q90Y	N82S:N85G	yes	Ntag
G3*-053	S7N:L12C:T125R	Q90Y	N82S:N85G	yes	Ntag
G3*-054	S7N:L12C:T125S	Q90Y	N82S:N85G	yes	Ntag
G3*-055	S7N:L12C:T125V	Q90Y	N82S:N85G	yes	Ntag
G3*-056	S7N:L12C:T125W	Q90Y	N82S:N85G	yes	Ntag
G3*-057	S7N:L12C:T125Y	Q90Y	N82S:N85G	yes	Ntag
G3*-058	S7N:L12A:T125M	G16T	N82S:N85G	yes	Ntag
G3*-059	S7N:L12D:T125M	G16T	N82S:N85G	yes	Ntag
G3*-060	S7N:L12E:T125M	G16T	N82S:N85G	yes	Ntag
G3*-061	S7N:L12F:T125M	G16T	N82S:N85G	yes	Ntag
G3*-062	S7N:L12G:T125M	G16T	N82S:N85G	yes	Ntag
G3*-063	S7N:L12H:T125M	G16T	N82S:N85G	yes	Ntag
G3*-064	S7N:L12I:T125M	G16T	N82S:N85G	yes	Ntag
G3*-065	S7N:L12K:T125M	G16T	N82S:N85G	yes	Ntag
G3*-066	S7N:T125M	G16T	N82S:N85G	yes	Ntag
G3*-067	S7N:L12M:T125M	G16T	N82S:N85G	yes	Ntag
G3*-068	S7N:L12N:T125M	G16T	N82S:N85G	yes	Ntag
G3*-069	S7N:L12P:T125M	G16T	N82S:N85G	yes	Ntag
G3*-070	S7N:L12Q:T125M	G16T	N82S:N85G	yes	Ntag
G3*-071	S7N:L12R:T125M	G16T	N82S:N85G	yes	Ntag
G3*-072	S7N:L12S:T125M	G16T	N82S:N85G	yes	Ntag
G3*-073	S7N:L12V:T125M	G16T	N82S:N85G	yes	Ntag
G3*-074	S7N:L12W:T125M	G16T	N82S:N85G	yes	Ntag
G3*-075	S7N:L12Y:T125M	G16T	N82S:N85G	yes	Ntag
G3*-076	S7N:L12T:T125M	G16T	N82S:N85G	yes	Ntag
G3*-077	S7N:L12A:T125N	Q90Y	N82S:N85G	yes	Ntag
G3*-078	S7N:L12D:T125N	Q90Y	N82S:N85G	yes	Ntag
G3*-079	S7N:L12E:T125N	Q90Y	N82S:N85G	yes	Ntag
G3*-080	S7N:L12F:T125N	Q90Y	N82S:N85G	yes	Ntag
G3*-081	S7N:L12G:T125N	Q90Y	N82S:N85G	yes	Ntag
G3*-082	S7N:L12H:T125N	Q90Y	N82S:N85G	yes	Ntag
G3*-083	S7N:L12I:T125N	Q90Y	N82S:N85G	yes	Ntag
G3*-084	S7N:L12K:T125N	Q90Y	N82S:N85G	yes	Ntag
G3*-085	S7N:T125N	Q90Y	N82S:N85G	yes	Ntag
G3*-086	S7N:L12M:T125N	Q90Y	N82S:N85G	yes	Ntag
G3*-087	S7N:L12N:T125N	Q90Y	N82S:N85G	yes	Ntag
G3*-088	S7N:L12P:T125N	Q90Y	N82S:N85G	yes	Ntag
G3*-089	S7N:L12Q:T125N	Q90Y	N82S:N85G	yes	Ntag
G3*-090	S7N:L12R:T125N	Q90Y	N82S:N85G	yes	Ntag
G3*-091	S7N:L12S:T125N	Q90Y	N82S:N85G	yes	Ntag
G3*-092	S7N:L12V:T125N	Q90Y	N82S:N85G	yes	Ntag
G3*-093	S7N:L12W:T125N	Q90Y	N82S:N85G	yes	Ntag
G3*-094	S7N:L12Y:T125N	Q90Y	N82S:N85G	yes	Ntag
G3*-095	S7N:L12T:T125N	Q90Y	N82S:N85G	yes	Ntag
G3-001	S7N:L12C	G8D	E6Q	yes	Ntag
G3-002	S7N:L12C:Q14D	-	E6Q	yes	Ntag
G3-003	S7N:L12C:Q14K	-	E6Q	yes	Ntag
G3-004	S7N:L12C	G16T	E6Q	yes	Ntag
G3-005	S7N:L12C	G17E	E6Q	yes	Ntag
G3-006	S7N:L12C	R20K	E6Q	yes	Ntag
G3-007	S7N:L12C	Q90R	E6Q	yes	Ntag
G3-008	S7N:L12C	Q90Y	E6Q	yes	Ntag
G3-009	S7N:L12C	Q90H	E6Q	yes	Ntag
G3-010	S7N:L12C	N92D	E6Q	yes	Ntag
G3-011	S7N:L12N	-	E6Q	yes	Ntag
G3-012	S7N:L12C	G8D	T86A:V87L:Y88S	yes	Ntag
G3-013	S7N:L12C:Q14D	-	T86A:V87L:Y88S	yes	Ntag
G3-014	S7N:L12C:Q14K	-	T86A:V87L:Y88S	yes	Ntag
G3-015	S7N:L12C	G16T	T86A:V87L:Y88S	yes	Ntag
G3-016	S7N:L12C	G17E	T86A:V87L:Y88S	yes	Ntag
G3-017	S7N:L12C	R20K	T86A:V87L:Y88S	yes	Ntag
G3-018	S7N:L12C	Q90R	T86A:V87L:Y88S	yes	Ntag
G3-019	S7N:L12C	Q90Y	T86A:V87L:Y88S	yes	Ntag
G3-020	S7N:L12C	Q90H	T86A:V87L:Y88S	yes	Ntag
G3-021	S7N:L12C	N92D	T86A:V87L:Y88S	yes	Ntag
G3-022	S7N:L12N	-	T86A:V87L:Y88S	yes	Ntag
G3-023	S7N:L12C	G8D	Y88H	yes	Ntag
G3-024	S7N:L12C:Q14D	-	Y88H	yes	Ntag
G3-025	S7N:L12C:Q14K	-	Y88H	yes	Ntag
G3-026	S7N:L12C	G16T	Y88H	yes	Ntag
G3-027	S7N:L12C	G17E	Y88H	yes	Ntag
G3-028	S7N:L12C	R20K	Y88H	yes	Ntag
G3-029	S7N:L12C	Q90R	Y88H	yes	Ntag
G3-030	S7N:L12C	Q90Y	Y88H	yes	Ntag
G3-031	S7N:L12C	Q90H	Y88H	yes	Ntag

Nanobody protein ID	Key position mutations (S7N:L12C:Q14K:T125S)	Nb-α mutations	Epitope mutations	Solubility mutations? (G40T:Q48E:L52W:I101V)	Tag
G3-032	S7N:L12C	N92D	Y88H	yes	Ntag
G3-033	S7N:L12N	-	Y88H	yes	Ntag
G3-034	S7N:L12C	G8D	A54F:A55T	yes	Ntag
G3-035	S7N:L12C:Q14D	-	A54F:A55T	yes	Ntag
G3-036	S7N:L12C:Q14K	-	A54F:A55T	yes	Ntag
G3-037	S7N:L12C	G16T	A54F:A55T	yes	Ntag
G3-038	S7N:L12C	G17E	A54F:A55T	yes	Ntag
G3-039	S7N:L12C	R20K	A54F:A55T	yes	Ntag
G3-040	S7N:L12C	Q90R	A54F:A55T	yes	Ntag
G3-041	S7N:L12C	Q90Y	A54F:A55T	yes	Ntag
G3-042	S7N:L12C	Q90H	A54F:A55T	yes	Ntag
G3-043	S7N:L12C	N92D	A54F:A55T	yes	Ntag
G3-044	S7N:L12N	-	A54F:A55T	yes	Ntag
G3-045	S7N:L12C	G8D	N82S:N85G	yes	Ntag
G3-046	S7N:L12C:Q14D	-	N82S:N85G	yes	Ntag
G3-047	S7N:L12C:Q14K	-	N82S:N85G	yes	Ntag
G3-048	S7N:L12C	G16T	N82S:N85G	yes	Ntag
G3-048	S7N:L12C	G16T	N82S:N85G	yes	Ntag
G3-048	S7N:L12C	G16T	N82S:N85G	yes	Ntag
G3-048	S7N:L12C	G16T	N82S:N85G	yes	Ntag
G3-049	S7N:L12C	G17E	N82S:N85G	yes	Ntag
G3-050	S7N:L12C	R20K	N82S:N85G	yes	Ntag
G3-051	S7N:L12C	Q90R	N82S:N85G	yes	Ntag
G3-052	S7N:L12C	Q90Y	N82S:N85G	yes	Ntag
G3-053	S7N:L12C	Q90H	N82S:N85G	yes	Ntag
G3-054	S7N:L12C	N92D	N82S:N85G	yes	Ntag
G3-055	S7N:L12N	-	N82S:N85G	yes	Ntag
G3-055	S7N:L12N	-	N82S:N85G	yes	Ntag
G3-056	S7N:L12C	G8D	K84E	yes	Ntag
G3-057	S7N:L12C:Q14D	-	K84E	yes	Ntag
G3-058	S7N:L12C:Q14K	-	K84E	yes	Ntag
G3-059	S7N:L12C	G16T	K84E	yes	Ntag
G3-060	S7N:L12C	G17E	K84E	yes	Ntag
G3-061	S7N:L12C	R20K	K84E	yes	Ntag
G3-062	S7N:L12C	Q90R	K84E	yes	Ntag
G3-063	S7N:L12C	Q90Y	K84E	yes	Ntag
G3-064	S7N:L12C	Q90H	K84E	yes	Ntag
G3-065	S7N:L12C	N92D	K84E	yes	Ntag
G3-066	S7N:L12N	-	K84E	yes	Ntag
G3-067	S7N:L12C	G8D	S79Q	yes	Ntag
G3-068	S7N:L12C:Q14D	-	S79Q	yes	Ntag
G3-069	S7N:L12C:Q14K	-	S79Q	yes	Ntag
G3-070	S7N:L12C	G16T	S79Q	yes	Ntag
G3-071	S7N:L12C	G17E	S79Q	yes	Ntag
G3-072	S7N:L12C	R20K	S79Q	yes	Ntag
G3-073	S7N:L12C	Q90R	S79Q	yes	Ntag
G3-074	S7N:L12C	Q90Y	S79Q	yes	Ntag
G3-075	S7N:L12C	Q90H	S79Q	yes	Ntag
G3-076	S7N:L12C	N92D	S79Q	yes	Ntag
G3-077	S7N:L12C	G8D	R20K:L21V	yes	Ntag
G3-078	S7N:L12C:Q14D	-	R20K:L21V	yes	Ntag
G3-079	S7N:L12C:Q14K	-	R20K:L21V	yes	Ntag
G3-080	S7N:L12C	G16T	R20K:L21V	yes	Ntag
G3-081	S7N:L12C	G17E	R20K:L21V	yes	Ntag
G3-082	S7N:L12C	R20K	R20K:L21V	yes	Ntag
G3-083	S7N:L12C	Q90R	R20K:L21V	yes	Ntag
G3-084	S7N:L12C	Q90Y	R20K:L21V	yes	Ntag
G3-085	S7N:L12C	Q90H	R20K:L21V	yes	Ntag
G3-086	S7N:L12C	N92D	R20K:L21V	yes	Ntag
G3-087	S7N:L12C	G8D	Q44R	yes	Ntag
G3-088	S7N:L12C:Q14D	-	Q44R	yes	Ntag
G3-089	S7N:L12C:Q14K	-	Q44R	yes	Ntag
G3-090	S7N:L12C	G16T	Q44R	yes	Ntag
G3-091	S7N:L12C	G17E	Q44R	yes	Ntag
G3-092	S7N:L12C	R20K	Q44R	yes	Ntag
G3-093	S7N:L12C	Q90R	Q44R	yes	Ntag
G3-094	S7N:L12C	Q90Y	Q44R	yes	Ntag
G3-095	S7N:L12C	Q90H	Q44R	yes	Ntag
G3-096	S7N:L12C	N92D	Q44R	yes	Ntag
G4-001	S7N:L12C:T125M	G8D	E6Q	yes	Ntag
G4-002	S7N:L12C:Q14D: T125M	-	E6Q	yes	Ntag
G4-003	S7N:L12C:T125M	G16T	E6Q	yes	Ntag
G4-004	S7N:L12C:T125M	G17E	E6Q	yes	Ntag
G4-005	S7N:L12C:T125M	R20K	E6Q	yes	Ntag
G4-006	S7N:L12C:T125M	Q90R	E6Q	yes	Ntag
G4-007	S7N:L12C:T125M	Q90Y	E6Q	yes	Ntag
G4-008	S7N:L12C:T125M	Q90H	E6Q	yes	Ntag
G4-009	S7N:L12C:T125M	N92D	E6Q	yes	Ntag
G4-010	S7N:L12N:T125M	-	E6Q	yes	Ntag
G4-011	S7N:L12C:T125M	G8D	T86A:V87L:Y88S	yes	Ntag
G4-012	S7N:L12C:Q14D: T125M	-	T86A:V87L:Y88S	yes	Ntag
G4-013	S7N:L12C:T125M	G16T	T86A:V87L:Y88S	yes	Ntag
G4-014	S7N:L12C:T125M	G17E	T86A:V87L:Y88S	yes	Ntag

Nanobody protein ID	Key position mutations (S7N:L12C:Q14K:T125S)	Nb-α mutations	Epitope mutations	Solubility mutations? (G40T:Q48E:L52W:I101V)	Tag
G4-015	S7N:L12C:T125M	R20K	T86A:V87L:Y88S	yes	Ntag
G4-016	S7N:L12C:T125M	Q90R	T86A:V87L:Y88S	yes	Ntag
G4-017	S7N:L12C:T125M	Q90Y	T86A:V87L:Y88S	yes	Ntag
G4-018	S7N:L12C:T125M	Q90H	T86A:V87L:Y88S	yes	Ntag
G4-019	S7N:L12C:T125M	N92D	T86A:V87L:Y88S	yes	Ntag
G4-020	S7N:L12N:T125M	-	T86A:V87L:Y88S	yes	Ntag
G4-021	S7N:L12C:T125M	G8D	Y88H	yes	Ntag
G4-022	S7N:L12C:Q14D: T125M	-	Y88H	yes	Ntag
G4-023	S7N:L12C:T125M	G16T	Y88H	yes	Ntag
G4-024	S7N:L12C:T125M	G17E	Y88H	yes	Ntag
G4-025	S7N:L12C:T125M	R20K	Y88H	yes	Ntag
G4-026	S7N:L12C:T125M	Q90R	Y88H	yes	Ntag
G4-027	S7N:L12C:T125M	Q90Y	Y88H	yes	Ntag
G4-028	S7N:L12C:T125M	Q90H	Y88H	yes	Ntag
G4-029	S7N:L12C:T125M	N92D	Y88H	yes	Ntag
G4-030	S7N:L12N:T125M	-	Y88H	yes	Ntag
G4-031	S7N:L12C:T125M	G8D	A54F:A55T	yes	Ntag
G4-032	S7N:L12C:Q14D: T125M	-	A54F:A55T	yes	Ntag
G4-033	S7N:L12C:T125M	G16T	A54F:A55T	yes	Ntag
G4-034	S7N:L12C:T125M	G17E	A54F:A55T	yes	Ntag
G4-035	S7N:L12C:T125M	R20K	A54F:A55T	yes	Ntag
G4-036	S7N:L12C:T125M	Q90R	A54F:A55T	yes	Ntag
G4-037	S7N:L12C:T125M	Q90Y	A54F:A55T	yes	Ntag
G4-038	S7N:L12C:T125M	Q90H	A54F:A55T	yes	Ntag
G4-039	S7N:L12C:T125M	N92D	A54F:A55T	yes	Ntag
G4-040	S7N:L12N:T125M	-	A54F:A55T	yes	Ntag
G4-041	S7N:L12C:T125M	G8D	N82S:N85G	yes	Ntag
G4-042	S7N:L12C:Q14D: T125M	-	N82S:N85G	yes	Ntag
G4-043	S7N:L12C:T125M	G16T	N82S:N85G	yes	Ntag
G4-044	S7N:L12C:T125M	G17E	N82S:N85G	yes	Ntag
G4-045	S7N:L12C:T125M	R20K	N82S:N85G	yes	Ntag
G4-046	S7N:L12C:T125M	Q90R	N82S:N85G	yes	Ntag
G4-047	S7N:L12C:T125M	Q90Y	N82S:N85G	yes	Ntag
G4-048	S7N:L12C:T125M	Q90H	N82S:N85G	yes	Ntag
G4-049	S7N:L12C:T125M	N92D	N82S:N85G	yes	Ntag
G4-050	S7N:L12N:T125M	-	N82S:N85G	yes	Ntag
G4-051	S7N:L12C:T125M	G8D	K84E	yes	Ntag
G4-052	S7N:L12C:Q14D: T125M	-	K84E	yes	Ntag
G4-053	S7N:L12C:T125M	G16T	K84E	yes	Ntag
G4-054	S7N:L12C:T125M	G17E	K84E	yes	Ntag
G4-055	S7N:L12C:T125M	R20K	K84E	yes	Ntag
G4-056	S7N:L12C:T125M	Q90R	K84E	yes	Ntag
G4-057	S7N:L12C:T125M	Q90Y	K84E	yes	Ntag
G4-058	S7N:L12C:T125M	Q90H	K84E	yes	Ntag
G4-059	S7N:L12C:T125M	N92D	K84E	yes	Ntag
G4-060	S7N:L12N:T125M	-	K84E	yes	Ntag
G4-061	S7N:L12C:T125M	G8D	S79Q	yes	Ntag
G4-062	S7N:L12C:Q14D: T125M	-	S79Q	yes	Ntag
G4-063	S7N:L12C:T125M	G16T	S79Q	yes	Ntag
G4-064	S7N:L12C:T125M	G17E	S79Q	yes	Ntag
G4-065	S7N:L12C:T125M	R20K	S79Q	yes	Ntag
G4-066	S7N:L12C:T125M	Q90R	S79Q	yes	Ntag
G4-067	S7N:L12C:T125M	Q90Y	S79Q	yes	Ntag
G4-068	S7N:L12C:T125M	Q90H	S79Q	yes	Ntag
G4-069	S7N:L12C:T125M	N92D	S79Q	yes	Ntag
G4-070	S7N:L12C:T125M	G8D	R20K:L21V	yes	Ntag
G4-071	S7N:L12C:Q14D: T125M	-	R20K:L21V	yes	Ntag
G4-072	S7N:L12C:T125M	G16T	R20K:L21V	yes	Ntag
G4-073	S7N:L12C:T125M	G17E	R20K:L21V	yes	Ntag
G4-074	S7N:L12C:T125M	R20K	R20K:L21V	yes	Ntag
G4-075	S7N:L12C:T125M	Q90R	R20K:L21V	yes	Ntag
G4-076	S7N:L12C:T125M	Q90Y	R20K:L21V	yes	Ntag
G4-077	S7N:L12C:T125M	Q90H	R20K:L21V	yes	Ntag
G4-078	S7N:L12C:T125M	N92D	R20K:L21V	yes	Ntag
G4-079	S7N:L12C:T125M	G8D	Q44R	yes	Ntag
G4-080	S7N:L12C:Q14D: T125M	-	Q44R	yes	Ntag
G4-081	S7N:L12C:T125M	G16T	Q44R	yes	Ntag
G4-082	S7N:L12C:T125M	G17E	Q44R	yes	Ntag
G4-083	S7N:L12C:T125M	R20K	Q44R	yes	Ntag
G4-084	S7N:L12C:T125M	Q90R	Q44R	yes	Ntag
G4-085	S7N:L12C:T125M	Q90Y	Q44R	yes	Ntag
G4-086	S7N:L12C:T125M	Q90H	Q44R	yes	Ntag
G4-087	S7N:L12C:T125M	N92D	Q44R	yes	Ntag
G5-001	S7N:L12C:Q14K:T125M	-	E6Q	yes	Ntag
G5-002	S7N:L12C:Q14K:T125M	-	T86A:V87L:Y88S	yes	Ntag
G5-003	S7N:L12C:Q14K:T125M	-	Y88H	yes	Ntag
G5-004	S7N:L12C:Q14K:T125M	-	A54F:A55T	yes	Ntag
G5-005	S7N:L12C:Q14K:T125M	-	N82S:N85G	yes	Ntag
G5-006	S7N:L12C:Q14K:T125M	-	K84E	yes	Ntag
G5-007	S7N:L12C:Q14K:T125M	-	S79Q:A24V	yes	Ntag
G5-008	S7N:L12C:Q14K:T125M	-	R20K:L21V	yes	Ntag
G5-009	S7N:L12C:Q14K:T125M	-	Q44R	yes	Ntag

The nanobody construct of G2-001 was generated by mutagenesis method based on G1-001 using pNIC-CTH0 as the vector plasmid. The nanobody constructs of G1-002 to G1-091 were generated by mutagenesis method based on G1-001 using pNIC-CTBH as the vector plasmid. The nanobody constructs of G2*-005 to G2*-016 were generated by mutagenesis method based on G2*-004 using pNIC-MBP2-LIC as the vector plasmid and G2-002 to G2-013 using pNIC-CTBH as the vector plasmid with the same method and template DNA. G3*-001 to G3*-019, G3*-020 to G3*-038, G3*-039 to G3*-057, G3*-058 to G3*-076 and G3*-077 to G3*-095 were generated using mutagenesis based on templates G3-048, G3-055, G3-052, G3*-011 and G3*-050 respectively. G3-001 to G3-096 and G2*-001 to G2*-004 were directly synthesized as DNA fragments and subcloned into pNIC-MBP2-LIC vectors. G4-001 to G4-087 and G5-001 to G5-009 were also based on the synthesized DNA fragments using a different set of primers in the PCR step. All cloning procedures followed standard SGC ligation independent cloning protocol (Savitsky, Bray et al. 2010).

Constructs of nanobody variants were transformed into Single step (KRX) *E.coli* strain (Promega) at the end of molecular cloning (Hartnett, Gracyalny et al.). A single colony was inoculated into 1 ml of LB media (Merck) and incubated overnight. Starter culture of each nanobody was then inoculated into 100ml autoinduction media (Formedia) containing 0.1% Rhamnose, 0.01% antifoam 204 and antibiotics, at 37°C for 5.5 hours followed by 40-44 hours at 18°C. The base buffer we used for purification contained 5% glycerol, 10 mM HEPES (pH 7.5), 500 mM NaCl, 0.5 mM TCEP. Bacteria for each nanobody were harvested at 4000g and resuspended in 10ml base buffer + 30 mM imidazole, 1% Triton, 0.5 mg/ml lysozyme, 10 ug/ml homemade benzonase, followed by storage in -80°C freezer overnight for complete cell lysis. The purification started next day with thawing the frozen pallets in room temperature water bath, followed by centrifugation at 5000g for 1 h to obtain clear supernatant. The supernatant for each nanobody was then applied to one 1 ml Ni-NTA pre-packed column (GE healthcare) pre-equilibrated with base buffer + 30 mM imidazole. After thoroughly washing the Ni-NTA columns with base buffer, nanobody was eluted using 2.5 ml of base buffer + 500 mM imidazole directly into base buffer equilibrated PD-10 columns (GE healthcare). 3.5 ml of base buffer was then applied to PD10 columns to elute the nanobody protein in 0 imidazole buffer. Then TEV protease was added to protein solution with a 1:10 mass ratio for overnight incubation, and 10 mM imidazole was also added in the solution. The next day, two Ni-NTA columns were pre-equilibrated for each nanobody with base buffer + 10 mM imidazole and the nanobody solution with TEV was applied to the columns to get

rid of nanobody with uncleaved tags, TEV protease and contaminants. Flow-through fractions were collected and 2 ml of base buffer was applied to the Ni-NTA columns to flush all nanobody through. The collected fractions were ready for complex preparation.

4.3.2 Iterations of nanobody scaffold mutagenesis yield consistently diffracting crystals

Based on the data mining result in Chapter 4 and within our RECQL5:nanobody system, I designed 5 generations of nanobody scaffolds (G1-G5) of more than 200 constructs with different combinations of Nb-α mutations and crystal epitope mutations (Figure 4.6). The rationale of the mutational design around Nb-α was to introduce more interactions to enhance the interface, e.g. salt bridges, hydrogen bonds and potential disulfide bonds.

G1 and G2 were mutants based on the original RECQL5 nanobody scaffold with a C-terminal 6XHis tag, which was cleaved during the purification process and resulted in a 6-residue tail (ENLYFQ) at the C terminus. G2 crystals with the S7N:L12C mutations had higher diffraction rate, however the high-resolution fraction did not significantly increase. In G3, I introduced the solubility mutations (G40T:Q49E:L52W:I101V, abbreviated as SOL), changed the C-terminal 6XHis tag to the N-terminal MBP-6XHis tag which did not yield the 6-residue C-terminal tail, and meanwhile used other crystal epitope mutations rather than D69Y. Although G3 crystals had an even worse diffraction quality, two more mutations (T125M and Q14K) based on G3 scaffolds resulted in significantly improved diffraction quality in G5 (Figure 4.7A).

I then further explored the crystals from the G5 category. There were several crystallization conditions yielding quick crystals within 2 days, which were selected. 10-12 crystals from each variant-condition pair were harvested for diffraction analysis. The pair G5-006-F8c had significantly larger fraction of crystals diffracting better than 2.5 Å (Figure 4.7B), and it went into the further optimization on cryo-protectant and DMSO concentrations (fragment soaking conditions). Out of the 8 conditions we tested, 100% crystals from 4 conditions diffracted better than 2.5 Å (Figure 4.7 C-E), which satisfied crystallographic fragment screening prerequisites. Comparing G5-006 with G2-001 (best variant in G2), diffraction quality of G5-006 crystals was significantly improved.

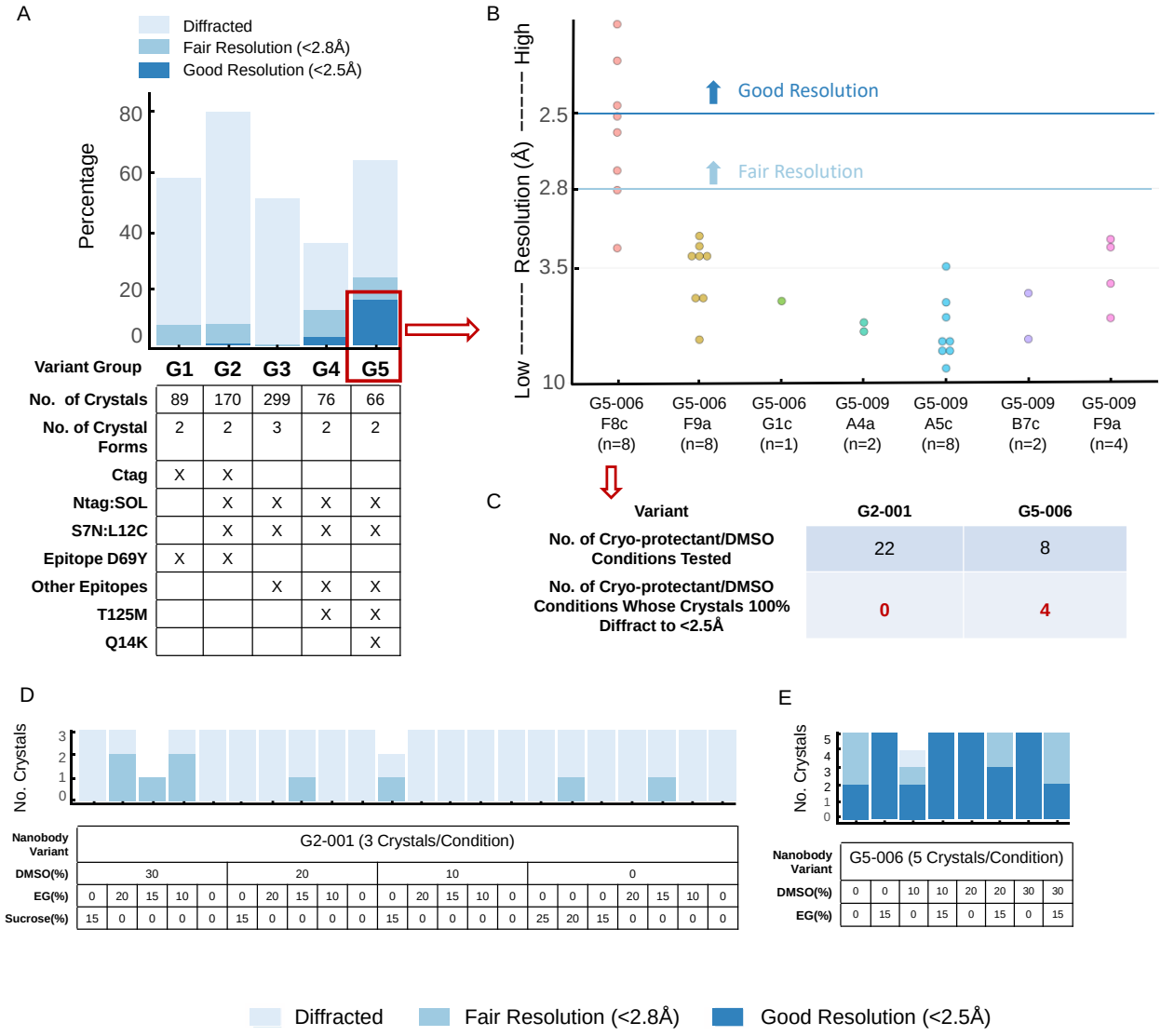
4.3.3 Constitutive nanobody interfaces were present in 6 different crystal forms

In the 5 generations of nanobody scaffolds, I aimed to find constitutive nanobody-nanobody interfaces which were independent of the target protein (RECQL5 in our case).

The interface I started with was Nb-α in G1. Upon combining S7N:L12C with G1, the crystallization propensity of RECQL5:nanobody complex increased dramatically for both N-terminal and C-terminal tagged variants (Figure 4.8A and 4.10B). Comparing the structures of G1-001 and G2-001, there was an additional hydrogen bond in Nb-α introduced by S7N, between the side chain N7 and the carbonyl oxygen of A15 (Figure 4.8C). Therefore, the S7N:L12C mutations were retained in the following iterations of mutagenesis.

Different combinations of crystal epitope mutations and Nb-α mutations were further explored. I replaced D69Y with 9 other crystal epitope mutations in G3, and in the meantime, changed the tag to the N-terminal MBP-6XHis tag. Surprisingly, the crystallization pattern changed dramatically with most G3 crystals in high salt conditions whilst most G2 crystals in PEG conditions (Figure 4.9). It was noticed that Residue Cys12 was essential in the crystallization of G3, since introduction of the C12N mutation decreased the average number of observed hits significantly (Figure 4.9). In addition, when the residue Cys12 was mutated to a selection of other residues, even good crystallizers would immediately stop crystallizing (Figure 4.10A).

Since Cys12 was demonstrated important in improving crystallization, mutagenesis of the neighbouring residues of Cys12 were further explored, especially Thr125. When T125M was introduced in G4 and G5, again the crystallization pattern dramatically changed, expanding hits in both high salt and PEG conditions (Figure 4.9). Furthermore, with the Nb-α mutation Q14K in G5, the crystallization propensity was further improved (Figure 4.10B) and the crystals grew large. However, due to the ambiguity of the electron densities captured from the K14 side chains, the rationale of this improvement remains unknown.



Variant Group	G1	G2	G3	G4	G5
No. of Crystals	89	170	299	76	66
No. of Crystal Forms	2	2	3	2	2
Ctag	X	X			
Ntag:SOL		X	X	X	X
S7N:L12C		X	X	X	X
Epitope D69Y	X	X			
Other Epitopes			X	X	X
T125M				X	X
Q14K					X

Variant	G2-001	G5-006
No. of Cryo-protectant/DMSO Conditions Tested	22	8
No. of Cryo-protectant/DMSO Conditions Whose Crystals 100% Diffract to <2.5Å	0	4

Nanobody Variant	G2-001 (3 Crystals/Condition)																				
DMSO(%)	30					20					10					0					
EG(%)	0	20	15	10	0	0	20	15	10	0	0	20	15	10	0	0	0	0	20	15	10	0
Sucrose(%)	15	0	0	0	0	15	0	0	0	0	15	0	0	0	0	25	20	15	0	0	0	0

Nanobody Variant	G5-006 (5 Crystals/Condition)							
DMSO(%)	0	0	10	10	20	20	30	30
EG(%)	0	15	0	15	0	15	0	15

Figure 4.7 Consistently well-diffracting crystal emerge from several iterations of mutagenesis around Nb-α. A) Diffraction quality of different groups of variants measured by the percentage of crystals which diffract to high resolution. The numbers of crystals and mutations present in each group of variants are indicated in the table below each bar. B) Diffraction quality of different nanobody variant and condition pairs of RECQL5:nanobody variant complexes presented as a dotplot. C) A table summary of results of the fragment soaking condition test of G2-001 and G5-006. The test involves different concentrations of cryo-protectant and DMSO. D)-E) Diffraction quality of G5-006 (D) and G5-009 (E) in different soaking conditions measured by the percentage of crystals which diffract to high resolution. In the barplot, deep blue fractions represent good resolution (<2.5 Å) crystal percentage, blue bar heights represent fair resolution (<2.8 Å) crystal percentage and the total heights of the bars represent percentage of crystals that diffracted. The numbers of crystals, nanobody variants and soaking conditions in each group are indicated in the table below each bar.

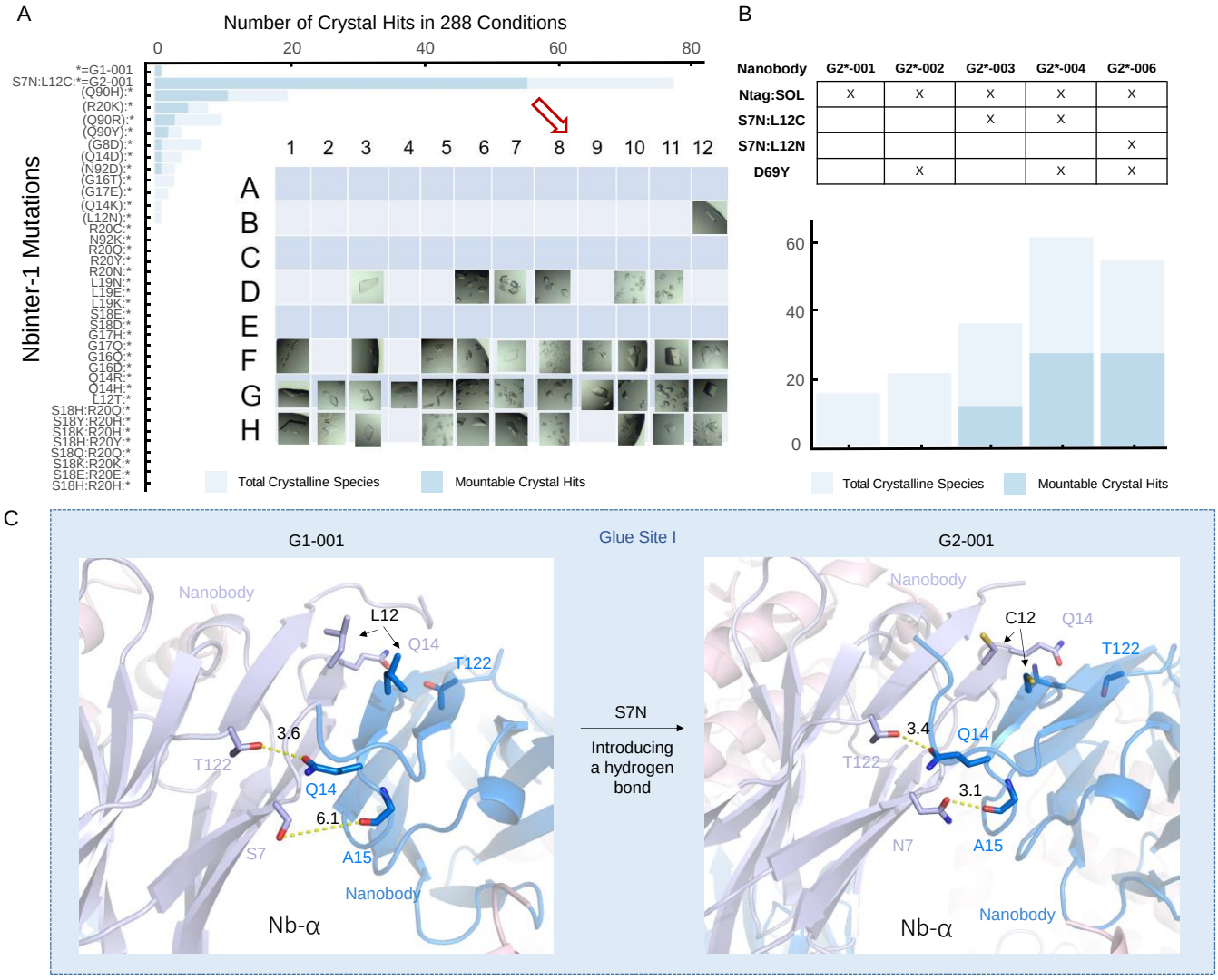
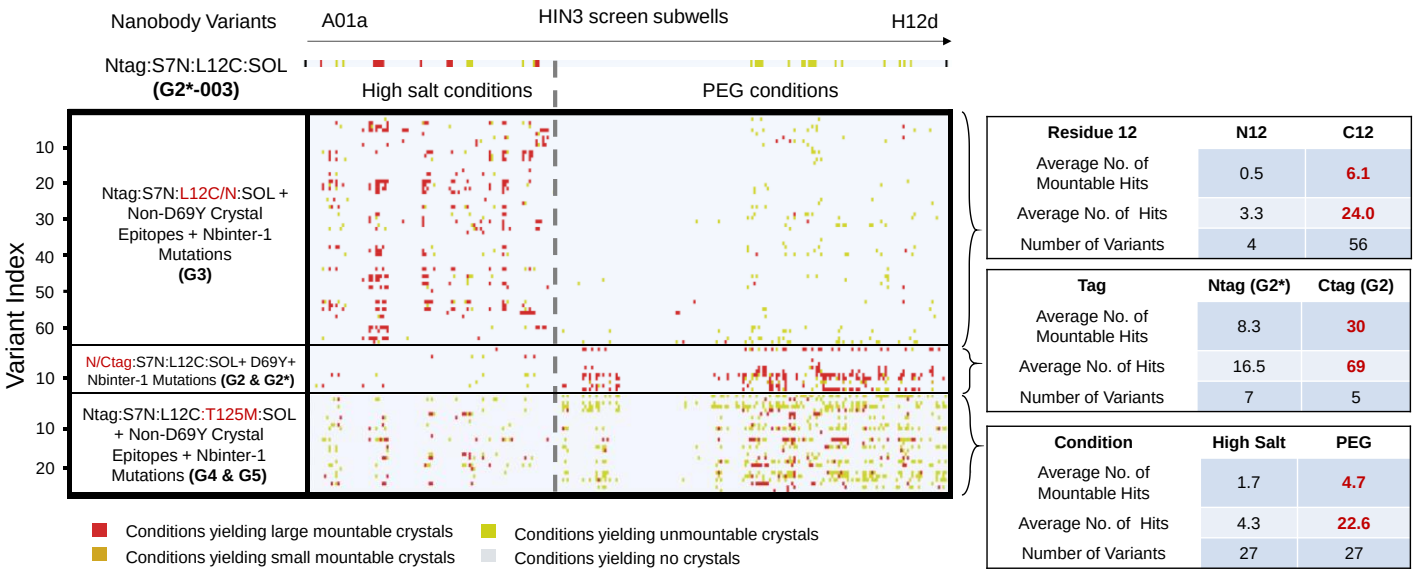


Figure 4.8 S7N enhances Nb-α by introducing an additional pair of hydrogen bonds. A) Hampton Index 3 (HIN3) triplicate drop coarse crystallization screen experiments show that S7N:L12C stands out among Nb-α mutations in significantly increasing the crystallization propensity of the RECQL5:nanobody complex. The asterisk represents the variant complex G1-001. Crystal images are taken from the HIN3 crystal screen of G2-001. B) HIN3 triplicate drop coarse crystallization screen experiments show that S7N:L12C significantly increases the crystallization propensity of the Ntag:SOL variant of RECQL5:nanobody complex with or without D69Y as the crystal epitope mutation. C) Close views of the Glue Site I before and after the mutation. S7N (stick representation) introduces an additional hydrogen bond with the carbonyl oxygen of A15 (stick representation) on the opposite nanobody molecule. Oxygen atoms are colored red and Nitrogen atoms are colored blue. The distances between the atoms connected by dash yellow lines are shown in Å.



Residue 12	N12	C12
Average No. of Mountable Hits	0.5	6.1
Average No. of Hits	3.3	24.0
Number of Variants	4	56

Tag	Ntag (G2*)	Ctag (G2)
Average No. of Mountable Hits	8.3	30
Average No. of Hits	16.5	69
Number of Variants	7	5

Condition	High Salt	PEG
Average No. of Mountable Hits	1.7	4.7
Average No. of Hits	4.3	22.6
Number of Variants	27	27

Figure 4.9 G2&G2*, G3 and G4&5 have different crystallization pattern in the Hampton Index coarse screen. The heat map shows Hampton Index 3 coarse screen triplicate crystal trials of RECQL5:nanobody complex with various mutations on the nanobody scaffold. Each row represents a nanobody construct, and each column in the chart represents a crystallization condition in the Hampton Index 3 coarse screen. The conditions are linearized sequentially from A01a on the left to H12d on the right. To the left of the grey dashed line are high salt conditions and to the right are PEG conditions. The color of each cell in the row represents whether there are crystals growing in that condition and the quality of the crystals. Some statistics of each group of nanobody variants are generalized on the right of the main plot.

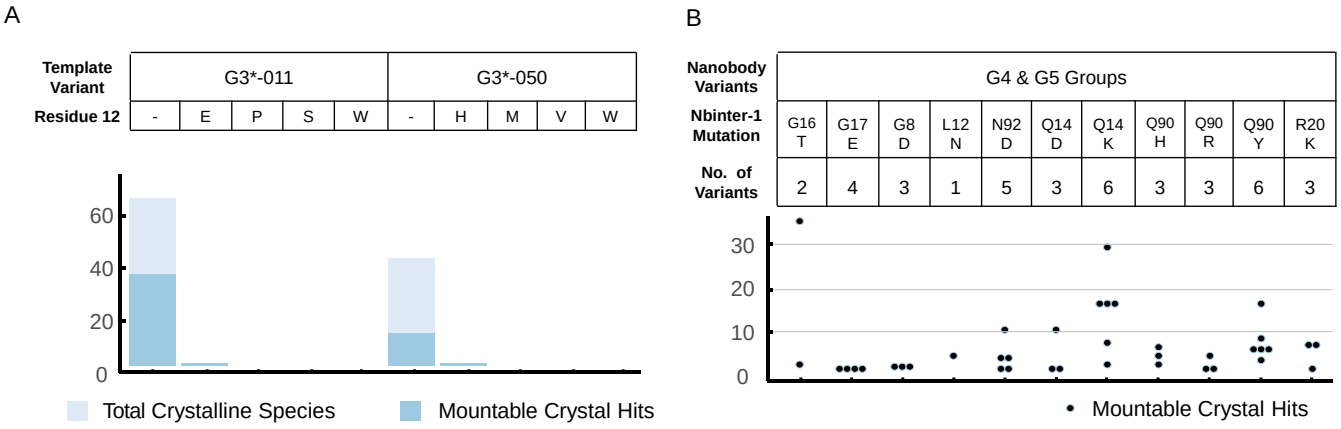


Figure 4.10 G2&G2*, G3 and G4&5 have different crystallization pattern in the Hampton Index coarse screen. A) The RECQL5 nanobody loses crystallization propensity when C12 is mutated to a selection of other residues. B) HIN3 triplicate drop coarse crystallization screen experiments show the effect of different Nb-α mutations to G4 &5 nanobody variants.

The reason why there was a significant crystallization pattern shift from G2 to G3 might be that a totally different nanobody arrangement formed. Instead of a dimer nanobody core (Figure 4.11 left panel) mediated by Nb-α in G1 and G2, I observed a tetrameric nanobody core (Figure 4.11 right panel) in G3. In the tetrameric nanobody, the Cys12 residues were in the core of the interface in contact with each other, and thus I named the new interface perpendicular to Nb-α as 'C-C interface'. Due to the geometric arrangement of the four nanobody molecules, the tetrameric core would not form with the 6-residue C-terminal tail ENLYFQ, which I observed in the G2-001 structure hanging over the Cys12 residues (Figure 4.11 left panel). This could explain why there were no tetrameric nanobodies observed in C-terminal tagged variants (Figure 4.11). The Cys12 itself was also essential in forming the tetrameric nanobody arrangement. When I mutated Cys12 to Asn12, the C-C interface dissociated and only the dimeric nanobody arrangement was observed (Figure 4.12 and Figure 4.13).

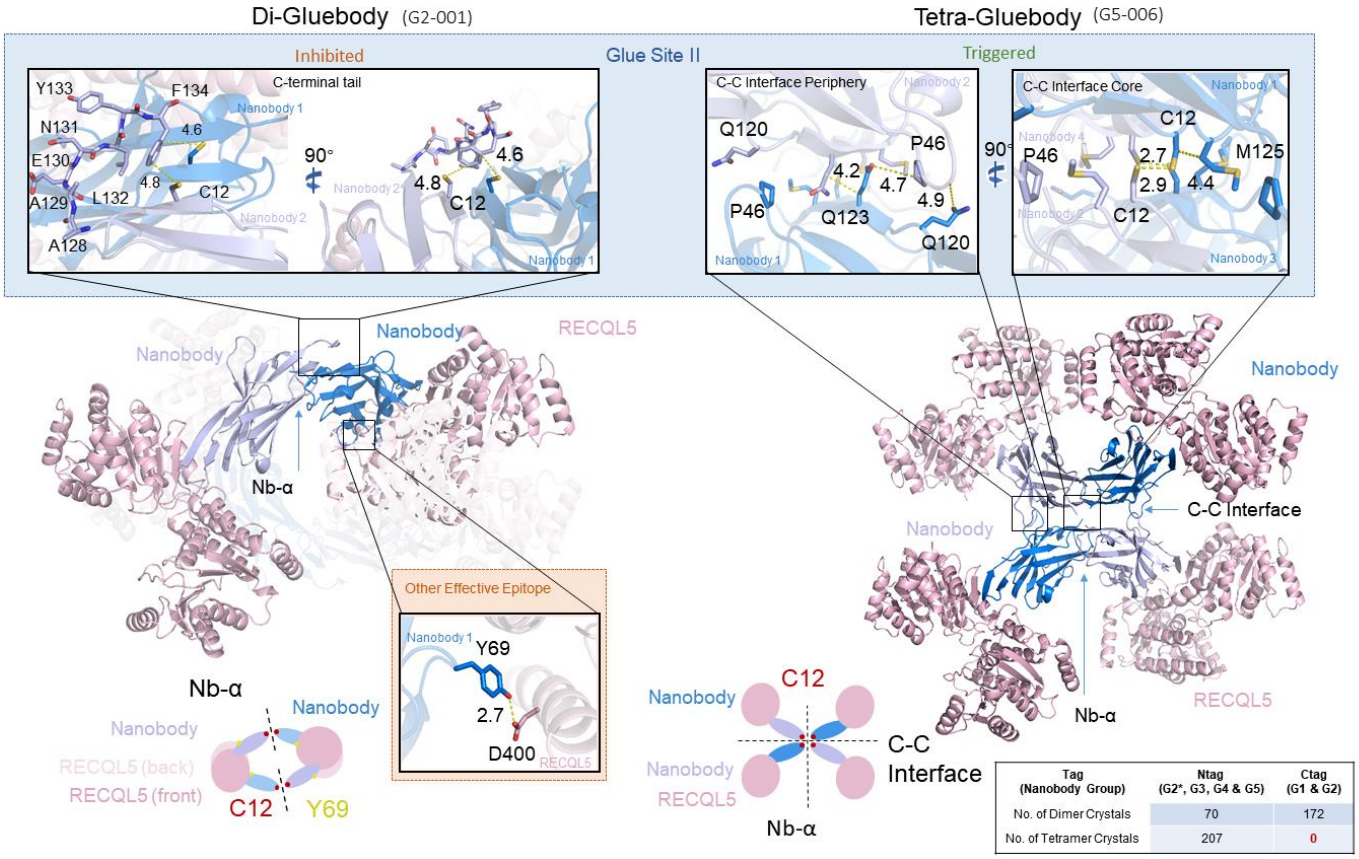


Figure 4.11 Dimeric (Di-Gluebody) and tetrameric (Tetra-Gluebody) crystallization pattern of RECQL5:nanobody with close views of Glue Site I & II and the crystal epitope site of Y69. The left panel shows RECQL5:nanobody complex crystallizes in a dimeric pattern when mutation D69Y and the C-terminal tail exist on the nanobody scaffold. The cartoon structure with high transparency is a symmetry mate of the dimer. The schematic of the dimer pattern is shown on the bottom left. The right panel shows the tetrameric crystallization. The M125 has a longer side chain than the T125 and occupies the cavity around the cysteine core within the C-C interface. The schematic of the tetramer pattern is shown on the bottom right. Oxygen atoms are colored red and Nitrogen atoms are colored blue. The distances between the atoms connected by dash yellow lines are shown in Å.

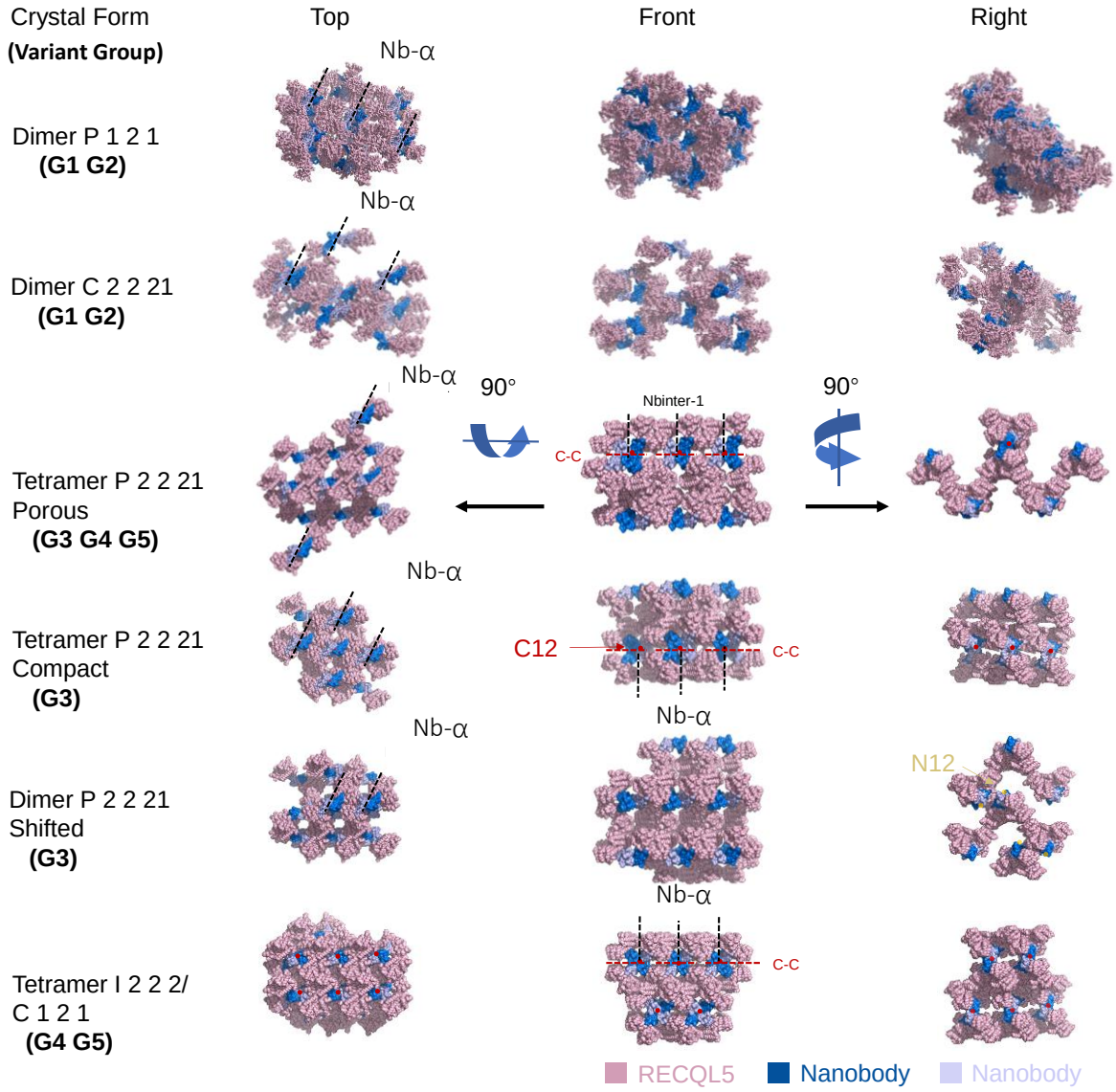


Figure 4.12 Six different crystal forms of the RECQL5:nanobody complex with respective mutations on the nanobody scaffold are presented in three perspectives (top, front and right views). RECQL5 molecules are in light-pink, and nanobody molecules are in marine and light-blue. C12 residues are presented as red dots on the structures and N12 residues are presented as yellow dots. Nb-α and C-C interfaces are indicated by dashed lines in the lattices. The crystal contacts of last four crystal forms are presented in Figure 4.13.

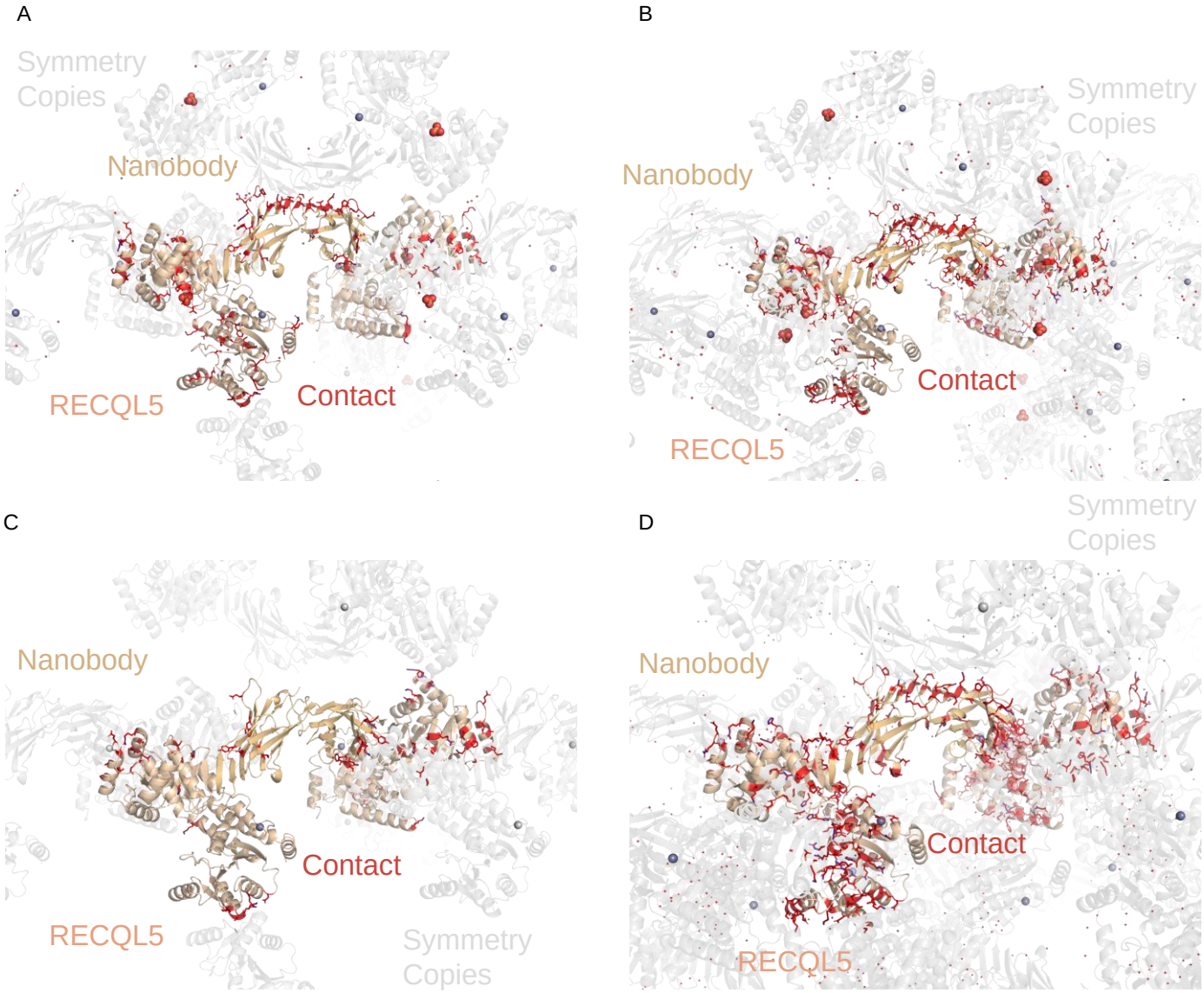


Figure 4.13 Detailed crystal contacts of the RECQL5:nanobody complex from the A) Tetramer P 2 2 21 porous, B) Tetramer P 2 2 21 compact, C) Dimer P 2 2 21 shifted and D) Tetramer C 1 2 1 crystal forms. Crystal contacts are shown in red.

From G1-G3 we already observed 5 different crystal forms (Figure 4.12 and Figure 4.13). Dimer P 1 2 1 and Dimer C 2 2 21 existed mostly in PEG conditions. Tetramer P 2 2 21 Porous/Compact and Dimer P 2 2 21 shifted mostly existed in high salt condition. In G4 and G5 we observed crystals in both high salt and PEG conditions. The high salt crystals of G4&5 were Tetramer P 2 2 21 Porous form, which I already observed in G3. The reason why G4&5 crystallized extensively in PEG conditions might be that it unlocked an additional crystal form Tetramer I 2 2 2/C1 2 1, and this crystal form led to our best-diffracting data (Figure 4.13), indicating that the crystal packing was also improved by the mutation T125M.

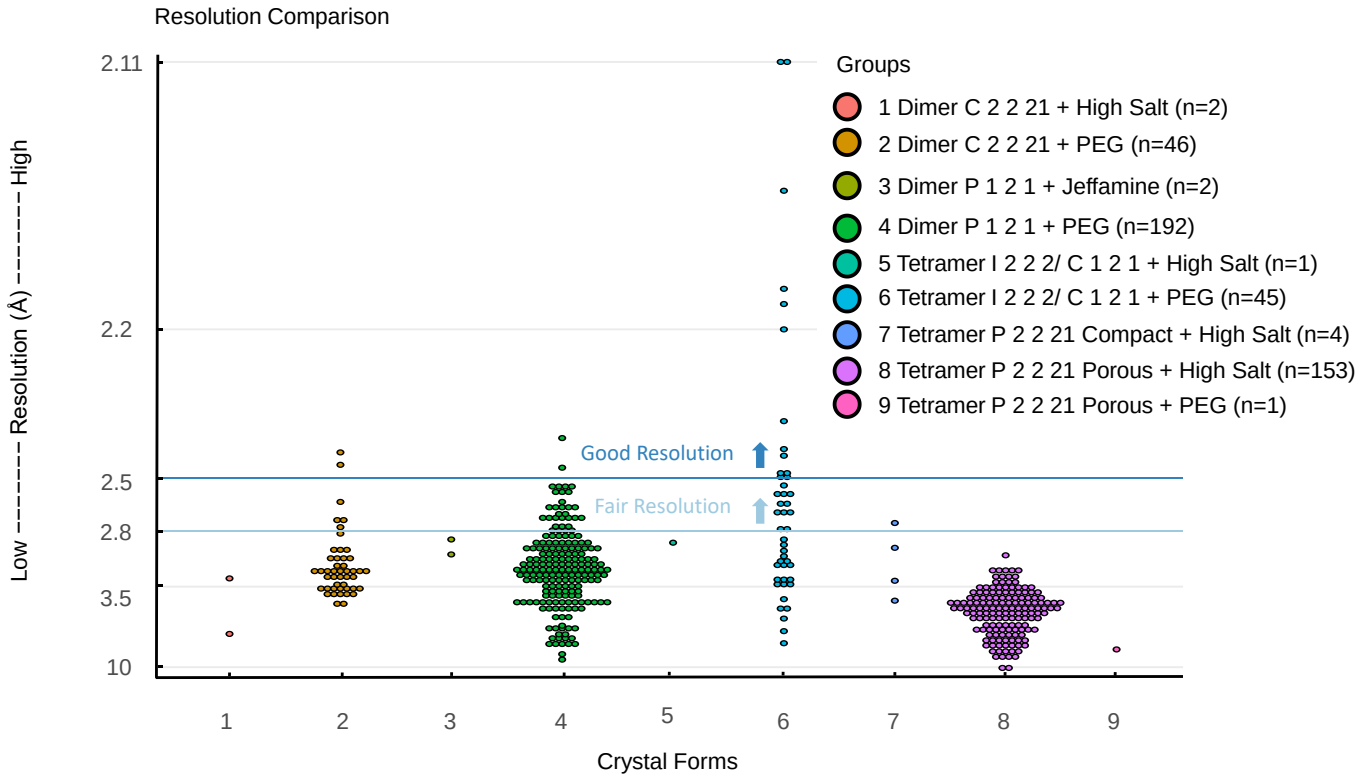


Figure 4.14 Resolution comparison of different crystal forms from RECQL5:nanobody complexes. Each combination of crystal forms and crystallization condition categories are grouped and indicated in different colors.

4.3.4 The name – ‘Gluebody’

Among all 6 crystal forms, the Nb-α was present, and among all tetramer crystal forms, the C-C interface was present (Figure 4.11). This indicated that these two interfaces were favoured. We therefore termed the nanobody scaffolds with our modifications to enhance the two interfaces as ‘Gluebodies’. We also termed interacting residues in Nb-α as ‘Glue site I’ (Figure 4.8C and 4.11) and interacting residues in the C-C interface as ‘Glue site II’ (Figure 4.11).

4.3.5 Discussion

4.3.5.1 Gluebodies are a route to new crystal polymorphs

In this study, I identified multiple crystal polymorphs in the RECQL5:nanobody model system, just by modifying the nanobody scaffold. The key to uncovering more crystal forms might be triggering the tetrameric nanobody pattern, which is the rationale behind the ‘Gluebody’.

The tetrameric nanobody pattern is more difficult to form with the naïve nanobody scaffold. Until recently known, there are only two cases of a tetramer crystal form in PDB (PDB id 6OS1 and 6OS2). The nanobody pattern of 6OS1 and 6OS2 is similar to the tetramer crystal form in our study with Nb-α and the C-C interface perpendicular to each other, but with fewer interacting residues.

Gluebodies can form the tetrameric pattern because they do not have C terminal tails, which is the prerequisite of forming the tetrameric nanobody pattern. However, C-terminal tags are more frequently used in nanobodies than their N-terminal counterparts. In the nanobody interface analysis of 335 crystal structures containing nanobodies, 211 of them (63%) have nanobodies with more than 5 residues as C-terminal tails. In our analysis of the dimer and tetramer crystal forms, C-terminal tails were a strong inhibitor of the formation of the tetramer crystal form, as they hung over the C12 surface, preventing further crystal contacts forming, including the C-C interface.

The four mutations are truly a refinement of intra-nanobody interfaces. In this study, I confirmed that the four key mutations of the Gluebody would be helpful in forming the tetramer crystal form more easily and because those mutated residues only contacted nanobody residues, they would not depend on widely ranging properties of the target proteins.

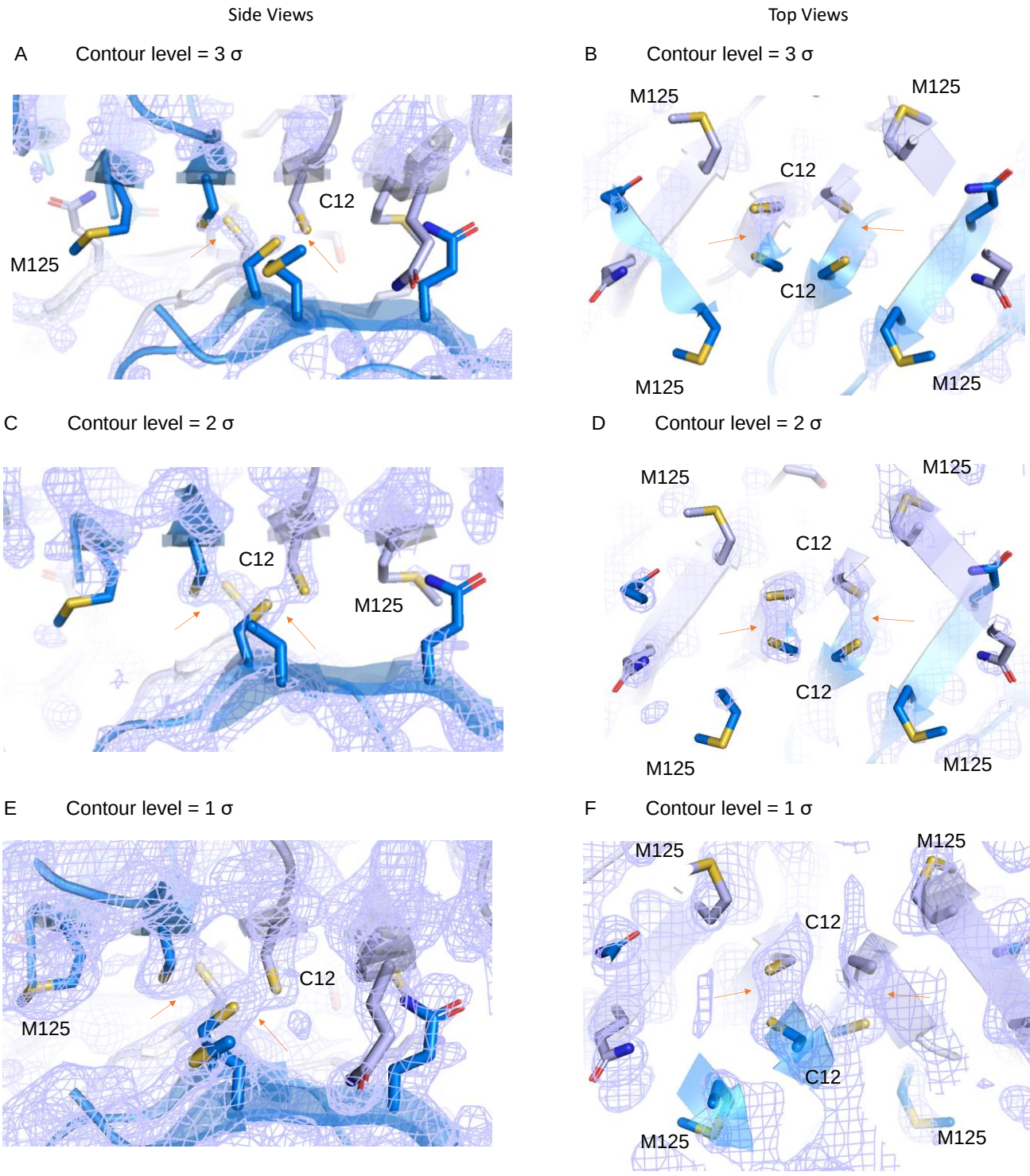


Figure 4.15 Electron density connecting the sulfur atoms in the C-C interface core viewed from the side and the top. C12 and M125 are presented in stick and the sulfur atoms are colored yellow. The electron density around C12 is indicated in light blue mesh with the contour levels at 1σ, 2σ and 3σ.

4.3.5.2 The tetrameric nanobody assembly can be applied in both crystallography and cryo-EM

In this work, it is demonstrated that the Gluebody has facilitated good crystal packing for RECQL5. Furthermore, there is also good chance that some adaptors could make the Gluebody a homotetrameric protein complex in solution. If the complex is rigid enough, it could be used as an imaging scaffold in cryoEM to visualize small soluble proteins more easily with a D2 symmetry, expanding the protein size spectrum of the cryoEM methodology. In the tetrameric crystallization pattern, I observed 4 Cysteine residues in the C-C interface core, possessing further modification potential. There was clearly electron density between the two sulfur atoms in each pair of Cys residues (Figure 4.15), and the distances between two sulfur atoms were 2.9 Å. This was longer than a disulfide bond, which is around 2.05 Å. There is no evidence supporting there exist chemical bonds among the four cysteines, and chances are there is still unoccupied space in the tetrameric nanobody core. Therefore, further metal coordination or organic adaptor molecules could be explored to make the tetrameric nanobody pattern rigid in solution.

4.4 Application of Gluebodies to several target proteins

In this section, I applied the optimized Gluebody scaffold to a couple of other target proteins, including PRNP, WRN, MAGEC2, TUT4 and a major facilitator superfamily (MFS) protein. For the first four targets, they could not co-crystallize their original nanobodies (experiment carried out by collaborators) and for the MFS protein, the Lipid Cubic Phase (LCP) crystal of its nanobody complex could not diffract to high resolution. The aim of this experiment was to test whether the G5 Gluebody could co-crystallize the first four targets and whether it could improve the MFS protein crystal.

4.4.1 Methods and materials

The target protein samples were directly provided by my collaborators. Prof. Opher Gileadi provided MAGEC2 and WRN sampels. Prof. Maria Cristina Nonato provided PRNP. Prof. Robert Gilbert provided TUT4. Dr. Katharina Duerr provided the MFS protein sample. The sequence of MAGEC2 is already stated in section 2.1. The sequences of PRPN, WRN (a longer construct than the WRN in section 2.1) and TUT4 are listed below.

Target	SGC construct ID	Protein sequence
PRPN	PRPNA-c005	MGQGGGTHSQWNKPSKPKTNMKHMAGAAAAGAVVGGLGGYMLGSAMSRPIIHFGSDYEDRYYRE NMHRYPNQVYYRPMDEYSNQNNFVHDCVNITIKQHTVTTTTKGENFTETDVKMMERVVEQMCITQ YERESQAYYQRGSS
WRN	WRNA-c513	SMNEGEEDDDKDFLWPAPNEEQVTCLKMYFGHSSFKPVQWKVIHSVLEERRDNVAVMATGYGKSLC FQYPPVYVGKIGLVISPLISLMEDQVLQLKMSNIPACFLGSAQSENVLTDIKLGKYRIVYVTPEYCSGNM GLLQQLEADIGITLIAVDEAHCISEWGHDFRDSFRKLGSLKTALPMVPIVALTATASSSIREDIVRCLNLRN PQITCTGFDRPNLYLEVRRKTGNILQDLQPFLVKTSSHWEFEGPTIIYCPSRKMTQQVTGELRKLNLSCG TYHAGMSFSTRKDIHHRFVRDEIQCVIATIAFGMGINKADIRQVIHYGAPKDMESYYQEIGRAGRDGL QSSCHVLWAPADINLNRHLLTEIRNEKFRLYKLKMMAKMEKYLHSSRCRRQIILSHFEDKQVQKASLGI MGTEKCCDNCRSRLDHCYSMDDSEDTSWDFGPQAFKLLSAVDILGEKFGIGLPILFLRGSNSQRLADQ YRRHSLFGTGKDQTESWWKAFSRQLITEGFLVEVSRYNKFMKICALTKKGRNWLHKANTESQSLILQA NEELCPKKFLLPSSKTVSSGTKEHCYNQ
TUT4	TUT4B-c001	RNSSKKNKQNDICIEKTEVKSCKVNAANLPGPKDLGLVLRDQSHCKAKKFPNSPVKAEKATISQAKSEK ATSLQAKAEKSPKSPNSVKAEKASSYQMKSEKVPSSPAEAEKGPSLLLKDMRQKTELQQIGKKIPSSFTS VDKVNIEAVGGEKCALQNSPRSQKQQTCTDNTGDSDDSASGIEDVSDDLSKMKNDESNKENSSEMD YLENATVIDESALTPEQRLGLKQAEERLERDHIFRLEKRSPEYTNCRYLCKLCLIHIENIQGAHKHIKEKRH KKNILEKQEESELRSLPPPSPAHLAALSVAVIELAKEHGITDDDLRVRQEIVEEMSKVITTFLPECSLRLYGS SLTRFALKSSDVNIDIKFPPKMNHPDLLIKVLGILKKNVLYVDVESDFHAKVPVVVCRDRKSGLLCRVSAG NDMACLTTDLLTALGKIEPVFIPLVLAFRYWAKLCYIDSQTDGGIPSYCFALMVMFFLQQRKPPLLPCLL GSWIEGFDPKRMDDFQLKGIVE

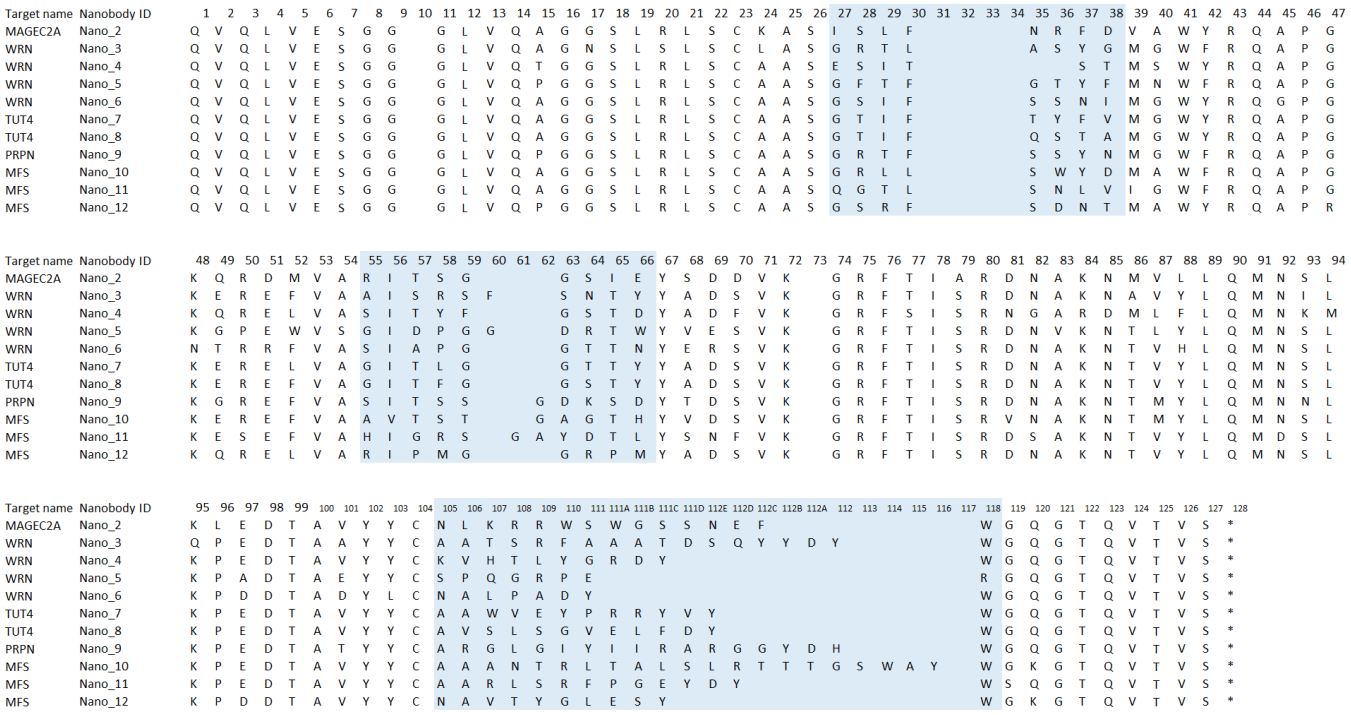
The sequences of the original nanobodies are shown in Figure 4.16 with IMGT numbering. Three sets of G5 Gluebodies were designed based on the original sequences. The mutations are listed below. All residue numbering corresponds to the IMGT numbering scheme in Figure 4.16.

Design 1 Gluebodies S7N:L12C:T125M:Q14K:Q44R

Design 2 Gluebodies S7N:L12C:T125M:Q14K:E6Q

Design 3 Gluebodies S7N:L12C:T125M:Q14K:K84E

The original nanobodies were provided by the collaborators and Gluebody variants were directly synthesized by TWIST biosciences. The nanobodies were expressed and purified following the detailed procedures in section 2.2.1.3.



Target name	Nanobody ID	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47
MAGEC2A	Nano_2	Q	V	Q	L	V	E	S	G	G		G	L	V	Q	A	G	G	S	L	R	L	S	C	K	A	S	I	S	L	F				N	R	F	D	V	A	W	Y	R	Q	A	P	G
WRN	Nano_3	Q	V	Q	L	V	E	S	G	G		G	L	V	Q	A	G	N	S	L	S	L	S	C	L	A	S	G	R	T	L				A	S	Y	G	M	G	W	F	R	Q	A	P	G
WRN	Nano_4	Q	V	Q	L	V	E	S	G	G		G	L	V	Q	T	G	G	S	L	R	L	S	C	A	A	S	E	S	I	T						S	T	M	S	W	Y	R	Q	A	P	G
WRN	Nano_5	Q	V	Q	L	V	E	S	G	G		G	L	V	Q	P	G	G	S	L	R	L	S	C	A	A	S	G	F	T	F				G	T	Y	F	M	N	W	F	R	Q	A	P	G
WRN	Nano_6	Q	V	Q	L	V	E	S	G	G		G	L	V	Q	A	G	G	S	L	R	L	S	C	A	A	S	G	S	I	F				S	S	N	I	M	G	W	Y	R	Q	G	P	G
TUT4	Nano_7	Q	V	Q	L	V	E	S	G	G		G	L	V	Q	A	G	G	S	L	R	L	S	C	A	A	S	G	T	I	F				T	Y	F	V	M	G	W	Y	R	Q	A	P	G
TUT4	Nano_8	Q	V	Q	L	V	E	S	G	G		G	L	V	Q	A	G	G	S	L	R	L	S	C	A	A	S	G	T	I	F				Q	S	T	A	M	G	W	Y	R	Q	A	P	G
PRPN	Nano_9	Q	V	Q	L	V	E	S	G	G		G	L	V	Q	P	G	G	S	L	R	L	S	C	A	A	S	G	R	T	F				S	S	Y	N	M	G	W	F	R	Q	A	P	G
MFS	Nano_10	Q	V	Q	L	V	E	S	G	G		G	L	V	Q	A	G	G	S	L	R	L	S	C	A	A	S	G	R	L	L				S	W	Y	D	M	A	W	F	R	Q	A	P	G
MFS	Nano_11	Q	V	Q	L	V	E	S	G	G		G	L	V	Q	A	G	G	S	L	R	L	S	C	A	A	S	Q	G	T	L				S	N	L	V	I	G	W	F	R	Q	A	P	G
MFS	Nano_12	Q	V	Q	L	V	E	S	G	G		G	L	V	Q	P	G	G	S	L	R	L	S	C	A	A	S	G	S	R	F				S	D	N	T	M	A	W	Y	R	Q	A	P	R

Target name	Nanobody ID	48	49	50	51	52	53	54	55	56	57	58	59	60	61	62	63	64	65	66	67	68	69	70	71	72	73	74	75	76	77	78	79	80	81	82	83	84	85	86	87	88	89	90	91	92	93	94
MAGEC2A	Nano_2	K	Q	R	D	M	V	A	R	I	T	S	G				G	S	I	E	Y	S	D	D	V	K		G	R	F	T	I	A	R	D	N	A	K	N	M	V	L	L	Q	M	N	S	L
WRN	Nano_3	K	E	R	E	F	V	A	A	I	S	R	S	F			S	N	T	Y	Y	A	D	S	V	K		G	R	F	T	I	S	R	D	N	A	K	N	A	V	Y	L	Q	M	N	I	L
WRN	Nano_4	K	Q	R	E	L	V	A	S	I	T	Y	F				G	S	T	D	Y	A	D	F	V	K		G	R	F	S	I	S	R	N	G	A	R	D	M	L	F	L	Q	M	N	K	M
WRN	Nano_5	K	G	P	E	W	V	S	G	I	D	P	G	G			D	R	T	W	Y	V	E	S	V	K		G	R	F	T	I	S	R	D	N	V	K	N	T	L	Y	L	Q	M	N	S	L
WRN	Nano_6	N	T	R	R	F	V	A	S	I	A	P	G				G	T	T	N	Y	E	R	S	V	K		G	R	F	T	I	S	R	D	N	A	K	N	T	V	H	L	Q	M	N	S	L
TUT4	Nano_7	K	E	R	E	L	V	A	G	I	T	L	G				G	T	T	Y	Y	A	D	S	V	K		G	R	F	T	I	S	R	D	N	A	K	N	T	V	Y	L	Q	M	N	S	L
TUT4	Nano_8	K	E	R	E	F	V	A	G	I	T	F	G				G	S	T	Y	Y	A	D	S	V	K		G	R	F	T	I	S	R	D	N	A	K	N	T	V	Y	L	Q	M	N	S	L
PRPN	Nano_9	K	G	R	E	F	V	A	S	I	T	S	S			G	D	K	S	D	Y	T	D	S	V	K		G	R	F	T	I	S	R	D	N	A	K	N	T	M	Y	L	Q	M	N	N	L
MFS	Nano_10	K	E	R	E	F	V	A	A	V	T	S	T			G	A	G	T	H	Y	V	D	S	V	K		G	R	F	T	I	S	R	V	N	A	K	N	T	M	Y	L	Q	M	N	S	L
MFS	Nano_11	K	E	S	E	F	V	A	H	I	G	R	S		G	A	Y	D	T	L	Y	S	N	F	V	K		G	R	F	T	I	S	R	D	S	A	K	N	T	V	Y	L	Q	M	D	S	L
MFS	Nano_12	K	Q	R	E	L	V	A	R	I	P	M	G				G	R	P	M	Y	A	D	S	V	K		G	R	F	T	I	S	R	D	N	A	K	N	T	V	Y	L	Q	M	N	S	L

Target name	Nanobody ID	95	96	97	98	99	100	101	102	103	104	105	106	107	108	109	110	111	111A	111B	111C	111D	112E	112D	112C	112B	112A	112	113	114	115	116	117	118	119	120	121	122	123	124	125	126	127	128
MAGEC2A	Nano_2	K	L	E	D	T	A	V	Y	Y	C	N	L	K	R	R	W	S	W	G	S	S	N	E	F									W	G	Q	G	T	Q	V	T	V	S	*
WRN	Nano_3	Q	P	E	D	T	A	A	Y	Y	C	A	A	T	S	R	F	A	A	A	T	D	S	Q	Y	Y	D	Y						W	G	Q	G	T	Q	V	T	V	S	*
WRN	Nano_4	K	P	E	D	T	A	V	Y	Y	C	K	V	H	T	L	Y	G	R	D	Y													W	G	Q	G	T	Q	V	T	V	S	*
WRN	Nano_5	K	P	A	D	T	A	E	Y	Y	C	S	P	Q	G	R	P	E																R	G	Q	G	T	Q	V	T	V	S	*
WRN	Nano_6	K	P	D	D	T	A	D	Y	L	C	N	A	L	P	A	D	Y																W	G	Q	G	T	Q	V	T	V	S	*
TUT4	Nano_7	K	P	E	D	T	A	V	Y	Y	C	A	A	W	V	E	Y	P	R	R	Y	V	Y											W	G	Q	G	T	Q	V	T	V	S	*
TUT4	Nano_8	K	P	E	D	T	A	V	Y	Y	C	A	V	S	L	S	G	V	E	L	F	D	Y											W	G	Q	G	T	Q	V	T	V	S	*
PRPN	Nano_9	K	P	E	D	T	A	T	Y	Y	C	A	R	G	L	G	I	Y	I	I	R	A	R	G	G	Y	D	H						W	G	Q	G	T	Q	V	T	V	S	*
MFS	Nano_10	K	P	E	D	T	A	V	Y	Y	C	A	A	A	N	T	R	L	T	A	L	S	L	R	T	T	G	S	W	A	Y		W	G	K	G	T	Q	V	T	V	S	*
MFS	Nano_11	K	P	E	D	T	A	V	Y	Y	C	A	A	R	L	S	R	F	P	G	E	Y	D	Y										W	S	Q	G	T	Q	V	T	V	S	*
MFS	Nano_12	K	P	D	D	T	A	V	Y	Y	C	N	A	V	T	Y	G	L	E	S	Y													W	G	K	G	T	Q	V	T	V	S	*

Figure 4.16 Original sequences of nanobodies selected in section 4.4 with IMGT numbering. Residues in cyan indicate CDR regions.

The target:nanobody/Gluebody pairs were mixed and went through the Sepax SRT SEC-300 size exclusion column. Peak fractions corresponding to the protein complexes were pooled, concentrated to 10-20 mg/ml and used to set up sitting-drop crystal trials against Hampton Index 3 or Hampton Crystal screen at both 20°C and 4°C, except the MFS protein complex. Crystals appearing in the sitting-drop crystal plates were then harvested and frozen with the addition of 17% ethylene glycol as the cryo-protectant. Frozen crystals were then sent to the Diamond Light Source (DLS) for diffraction analysis. The MFS protein complex was concentrated to 40 mg/ml and used to set up lipid cubic phase (LCP) plates. Crystals appearing in the plates were harvested, frozen and sent to DLS for diffraction analysis.

4.4.2 Results

For the WRN, PRNP, TUT4 and MAGEC2, the first batch of 16 crystal plates were set up using the design 1 Gluebody variants against the HIN3 (Hampton Index) screen, meaning 2 plates were set up for each target:Gluebody pairs in 20 °C and 4 °C. Fortunately, three out of the four samples yielded crystals. Those of WRN, PRNP and TUT4 were obtained after 56 days, 28 days and 1 day respectively since the day the plates

were set up. The crystals of respective protein samples are shown in Figure 4.17 with the crystallization conditions shown in the figure legend. Afterwards more Gluebody variants were tested, but no more crystals were obtained from them. The total number of tested Gluebody variants and crystal plates are shown in Figure 4.18. This result shows that even with a single mutation difference in the Gluebody scaffold, the crystallization performance can be different. Therefore, the more scaffolds tested, the better chance that a crystal can be obtained.

Subsequently the crystals were sent to the DLS for diffraction experiments. The PRNP and TUT4 crystals could only diffract to 6-8 Å, which could not yield a solvable dataset. When I started doing fine screens for those two crystals to optimize crystallization conditions, PRNP was out of stock and the collaborator stopped the provision of the sample, and therefore the experiment had to stop. On the other hand, TUT4 crystals were repeated successfully and more crystallization conditions were discovered. Now the experiment is still on-going. The WRN crystals diffracted the best to 2 Å. However, that dataset could not be phased using the molecular replacement method by the Phaser programme. The reason could be two – first the structure contained a very different conformation so that it was hard to phase even when using the nanobody; second, it should be noted that in the protein sample there was some contaminant proteins, which might also lead to crystallization. Therefore, it is still not certain now whether this is a positive result.

For the MFS protein, due to the request of the collaborator, no more details will be disclosed in this thesis.

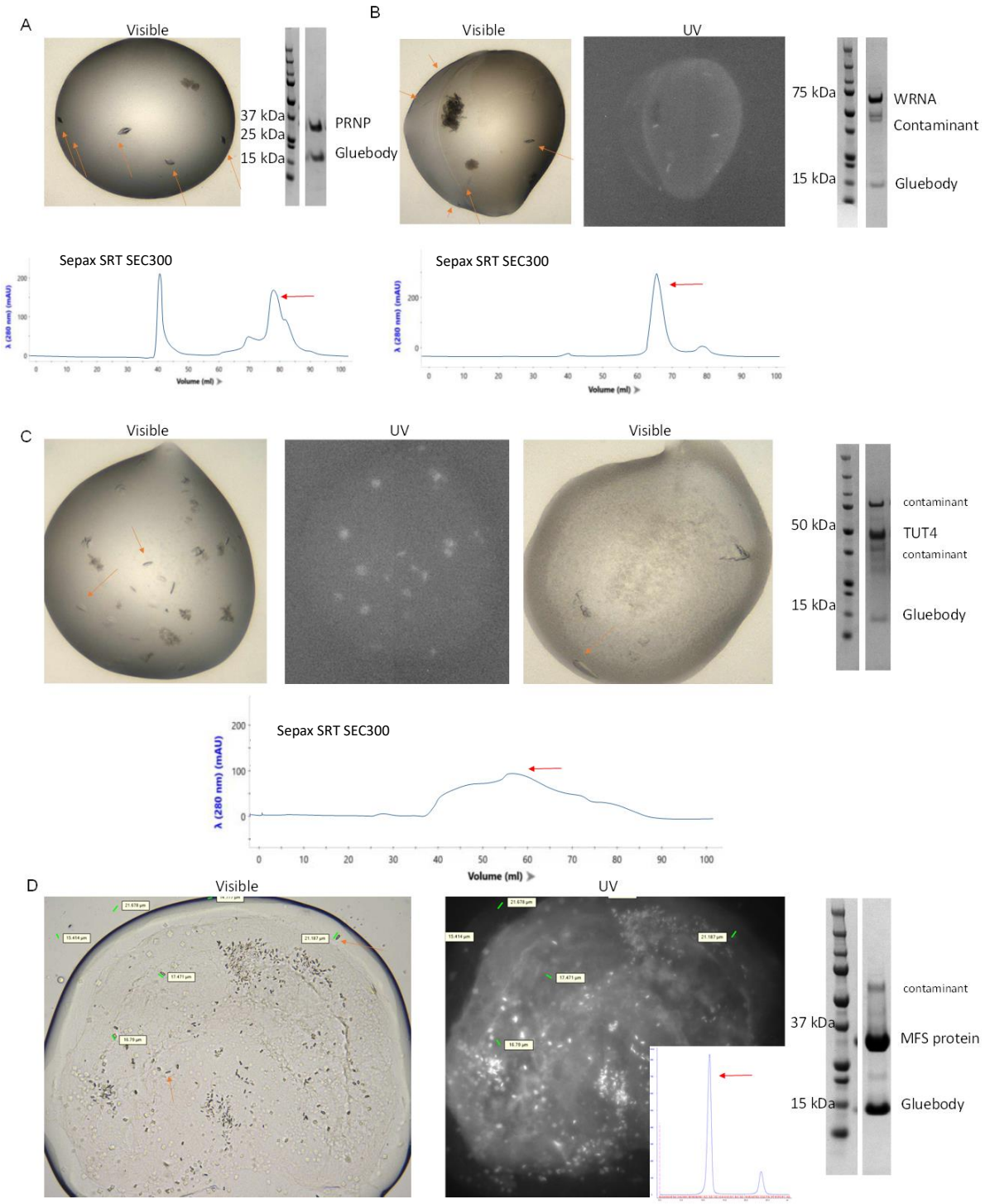


Figure 4.17 Crystals from PRNP, WRNA, TUT4 and MFS protein in complex with respective Tetra-Gluebodies. There are visible and UV images for crystals and SDS-PAGE, SEC profiles showing the quality of the protein. The protein peaks in SEC collected for crystallization trials are indicated by the red arrows. Crystals are indicated by orange arrows. A) PRNP:Gluebody complex crystals. Crystallization happened at 14mg/ml, 20 °C in the condition of 0.2M Magnesium chloride hexahydrate, 0.1M tris pH 8.5 25% w/v PEG3350, with a protein:precipitant ratio of 1:2.

Figure 4.17 (Continued) B) WRNA:Gluebody complex crystals. Crystallization happed at 27.3mg/ml, 20 °C in the condition of 0.15 M Potassium bromide, 30% w/v Polyethylene glycol monomethyl ether 2,000, with a protein:precipitant ratio of 1:2. C) TUT4:Gluebody complex crystals. There are several conditions yielding crystals. In the visible image on the left, the condition is 0.2 M Magnesium chloride hexahydrate, 0.1 M BIS-TRIS pH 6.5, 25% w/v Polyethylene glycol 3,350, with a protein:precipitant ratio of 1:2. In the visible image on the right, the condition is 0.1 M HEPES pH 7.5, 2.0 M Ammonium sulfate, with a protein:precipitant ratio of 1:2. All crystallization happened at 14mg/ml, 20 °C. D) The LCP crystals formed in the condition of 0.1M NaCl 0.1M Li_2SO_4 40%v/v PEG200 0.1M MES pH6 at 40mg/ml, 20 °C

4.4.3 Discussion and conclusion

4.4.3.1 Tetra-Gluebodies do not significantly improve crystallization

Target	Number of Glubodies Tried	Number of Crystal Plates set up with Gluebodies	Lipid Cubic Phase?	Crystals with Gluebody?	Crystals with original nanobody scaffold?	Diffraction	Comment
TUT4	5	20	No	Yes	No	6-8Å	
WRN	8	18	No	Yes	No	2Å	Unsolvable dataset
PRNP	1	2	No	Yes	No	6-8Å	
MFS protein	2	12	Yes	Yes	Yes	Do not diffract	Crystals with original nanobody scaffold diffract to 7-10Å
MAGEC2	3	6	No	No	No	/	

Figure 4.18 Experimental details of the crystallization trials of several targets in complex with respective Gluebodies.

The sets of mutations around the Nb-α on the Gluebody are demonstrated to improve the crystallization propensity of the nanobody with the model protein RECQL5, by enabling various crystal forms via the dimeric or tetrameric Gluebody arrangements. However, the improvement on the nanobody alone is not in our hands sufficient to generate good-quality crystals for non-crystallizing proteins. This was attempted on 5 other proteins (Figure 4.17 and 4.18) yielding crystals for three, two weakly diffracting (TUT4 and PRNP) and the other yielding an unsolvable dataset (WRN). The work is still ongoing. For the MFS protein, the crystals of the Gluebody complex did not diffract as well as the original nanobody scaffold. In such case,

the protein itself will likely need improving, or in the case of LCP, multiple combinations of detergents and LCP lipids need exploring as well.

Although this experiment has not finished yet, the current results do not allow me to claim that the Gluebodies significantly improve crystallization. However, since I obtained crystals for two other targets that the original nanobody could not produce, it was still worth characterising those weakly diffracting crystals, especially the reproducible crystals of the TUT4:Gluebody complex. As I mentioned in Chapter 3, I could only expect a maximum of 15% improvement when the Gluebody is applied to the general target pool. It is obvious that five more targets may not be enough to make fair conclusions for the Gluebody.

4.4.3.2 Multiple Gluebody variants are needed to fully explore the chaperoning ability of nanobodies

In this work, it is demonstrated that different crystal polymorphs could be obtained just by modifying the nanobody scaffold, and that the diffraction characteristics of different crystal polymorphs are different (Figure 4.14). To obtain well-diffracting crystals, it is imperative to discover crystallization conditions that yield a suitable crystal polymorph. With different Gluebody scaffolds, new crystal polymorphs can be obtained, and therefore this allows us to suggest 3 strategies:

1. C-terminal tailed naïve nanobody. There is no reason why the original scaffold should be abandoned completely. When the original nanobody is at hand, using it straight away should also provide some chance for successful co-crystallization of the target protein.
2. Di-Gluebody (Figure 4.19). By introducing Glue Site I mutations (Figure 4.8C, 4.11 and 4.19), (i.e. adding the mutation S7N), in addition to truncating the C-terminal tail, Di-Gluebodies are made. The Nb-α is enhanced in this case, but the C-C interfaces are not necessarily aligned with the Cys residues. There is a case in the PDB (id 6QGY) where two dimer nanobodies pack back-to-back in an angle with an interface perpendicular to Nb-α (Figure 4.20), indicating there is sometimes no need for an anchored C-C interface to assist crystallization.
3. Tetra-Gluebody (Figure 4.19). By introducing Q14K and Glue Site II mutations (Figure 4.11 and 4.19) in addition to Di-Gluebody, Tetra-Gluebodies are made. The 4 key mutations are installed to try to trigger the tetrameric crystallization polymorphs with an anchored C-C interface (Figure 4.11).

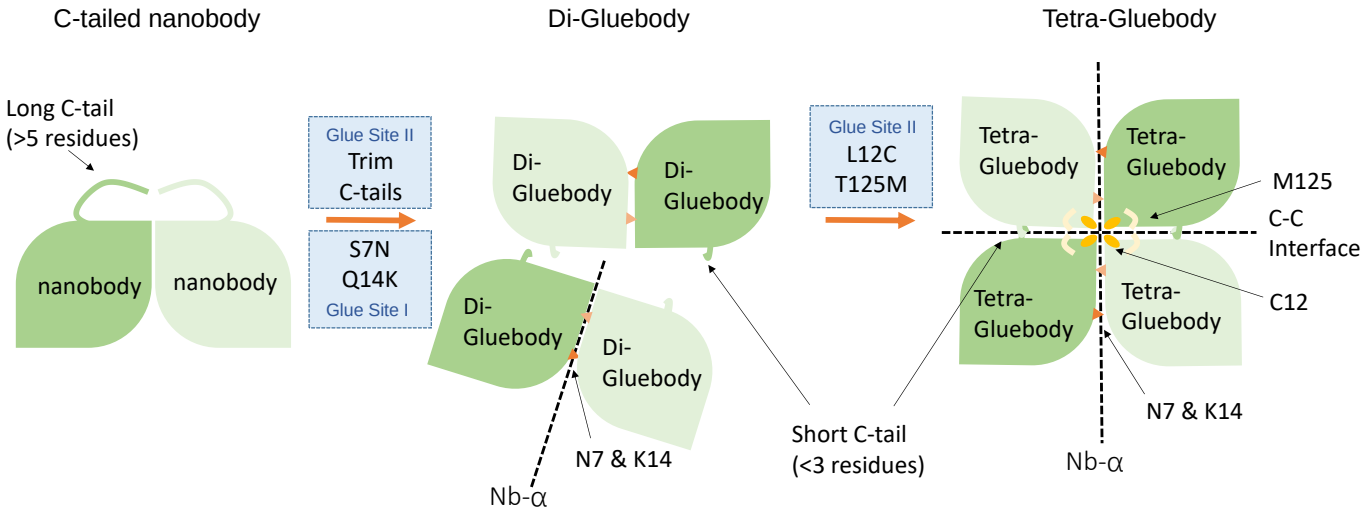


Figure 4.19 Key modifications on the nanobody scaffold in the evolution of the Gluebody. The putative crystallization patterns from the three scaffolds are presented from left to right with the scaffold type at the top of the sketches. The leaf shapes colored in green and light green represent nanobody molecules. The long or short overhangs on the leaf shapes represent the C-terminal tails of the nanobody molecules. The shapes colored in yellow, orange and pale represent the key mutations on the nanobody scaffold. (C12, N7 & K14 and M125, respectively) The dashed lines represent nanobody-nanobody interfaces occurring in the putative crystallization patterns

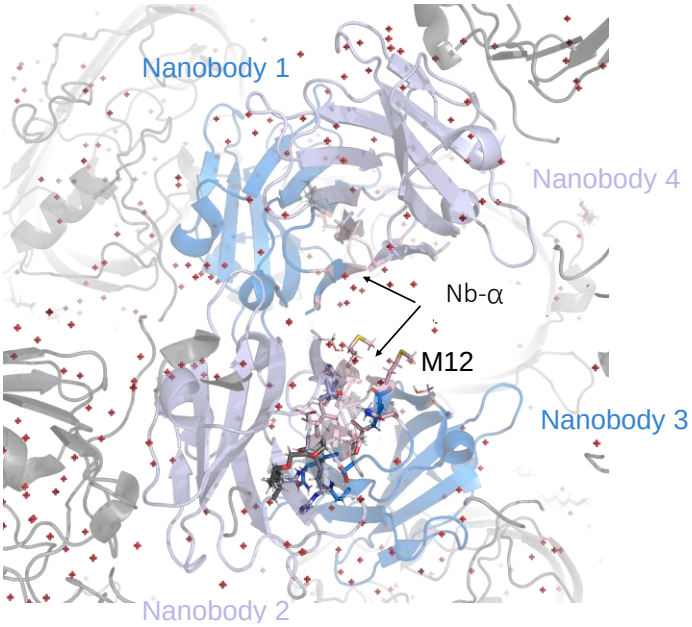


Figure 4.20 Tilted back-to-back packing of nanobody dimers from PDB id 6QGY. Molecules colored in marine and light blue are nanobody molecules. Nb-α is colored pink. The target protein molecules of the nanobodies are colored grey.

Chapter 5 Gluebodies as an imaging scaffold in cryo-EM

5.1 Introduction

As is discussed in Chapter 4, in the RECQL5 system, Gluebodies tend to multimerize *in crystallo*. If this multimerization could happen in solution and used in cryo-EM, Gluebodies could become an imaging scaffold that doubles or quadruples the protein size. For small proteins (<100 kDa), the Gluebody complex could immediately become a large particle expected to be easily visualized in cryo-EM (Figure 5.1). In this chapter, I collaborated with Gangshun Yi from Prof. Robert Gilbert's lab and Dimitrios Mammalis from Prof. Ben Davis' lab, to explore the chemical methods to make Dimer and Tetramer Gluebodies in solution and assess their multimerizing effect in cryo-EM. In this chapter, initial results of the disulfide-linked Gluebodies will be shown. I will refer to covalently linked Gluebodies as Di-Gluebodies or Tetra-Gluebodies depending on whether they are dimeric or tetrameric.

5.2 Di-Gluebodies

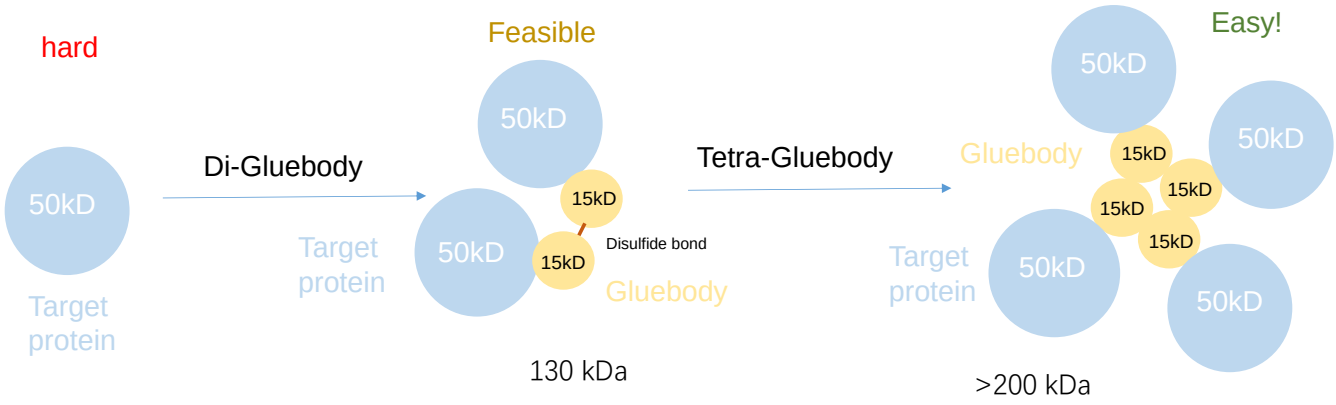


Figure 5.1 The concept of multimerized Gluebodies as an imaging scaffold in cryo-EM.

The Di-Gluebodies described in Chapter 4 are a pair of adjacent nanobodies in the crystal lattice. This dimer arrangement appeared in all six crystal forms obtained from RECQL5:nanobody complexes, indicating Nb-α might be a glue to guarantee the consistent dimer formation in *crystallo*. The idea of a Di-Gluebody in solution is that the introduction of a disulfide bond helps trigger the Nb-α mediated dimerization in solution. Although it has been argued that single chemical bond could also result in flexibility that made Di-Gluebody inadequate as a cryo-EM imaging

scaffold, there is chance that the Nb-α can help rigidify the dimer. In this study, we aim to figure out whether the Di-Gluebody is rigid enough to be an imaging scaffold.

5.2.1 Dimerizing the Gluebody in solution

To obtain a dimer Gluebody in solution, I first used Copper o-phenanthroline (CuP) as an oxidizing reagent and found when the monomeric Gluebody was incubated with 15mM CuP for 15 minutes, the reaction was more than complete (Figure 5.2). Then I dialyzed the solution against at least 10 L of buffer to get rid of the CuP.

However, the dialysis step in the CuP method is hard to deplete CuP completely and any remaining CuP will trigger disulfide formation. Hence, Dimitrios and I optimized a convenient slow disulfide reaction without using CuP. In this reaction, the Gluebody buffer is replaced by 25 mM HEPES, 150 mM NaCl at pH 8.8. The basic condition favours the deprotonation process of the –SH group and thus promotes the disulfide formation. The basic Gluebody solution is left on bench for 7 days at room temperature before most of the dimer molecules form. In order to maintain the dimer, the subsequent purification processes should exclude any reducing agents and have a pH above 8. After mixing the dimer Gluebody and the target protein, a size exclusion chromatography on Superdex 200 is performed to get rid of the remaining monomeric Gluebodies.

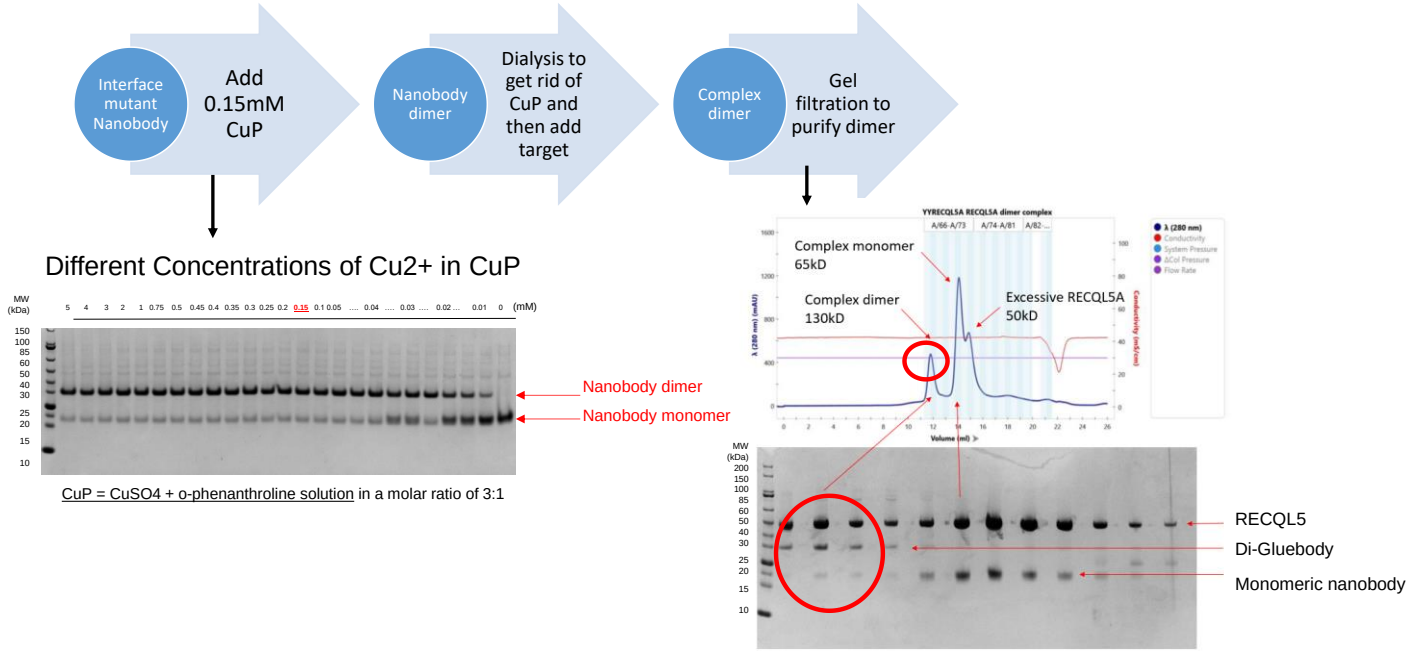


Figure 5.2 The protocol to make Di-Gluebody by the Copper o-phenanthroline method.

5.2.2 Negative staining EM and cryo-EM analyses of the target:Gluebody dimer complexes

In this experiment, RECQL5 (~50kDa) and the full length TUT4 (~120kDa) are used as model proteins to test the dimerizing effect of the Di-Gluebodies. The RECQL5:Gluebody (G2-001 in section 4.3) dimer complex was prepared using the CuP method, and TUT4:Gluebody (Design 1 G5 Gluebody of nano_7 in section 4.4) dimer complex was prepared using the slow oxidation method.

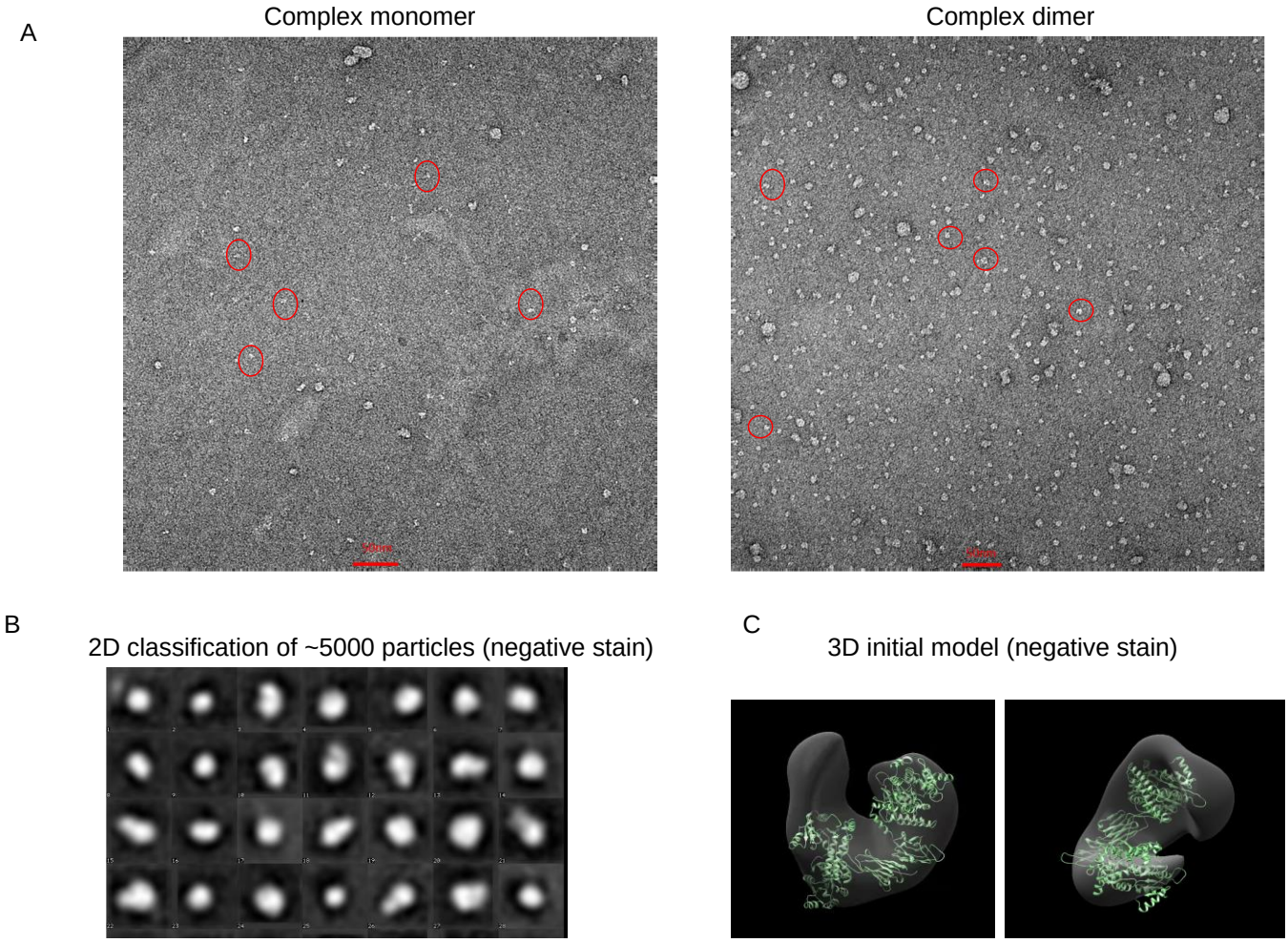


Figure 5.3 Negative staining electron microscopy analysis of the RECQL5:Di-Gluebody complex. The work is in collaboration with Gangshun Yi from Prof. Robert Gilbert's lab. A) Comparison of micrographs taken of RECQL5:nanobody and RECQL5:Di-Gluebody samples. Monomeric and dimeric TUT4:nanobody particles are circled in red in respective images. B) 2-D classification of around 5000 particles from the RECQL5:Di-Gluebody complex. C) 3-D initial model with the RECQL5:Di-Gluebody structure docked into the density.

To prepare the EM grids for negative staining analysis, 4 µl of purified RECQL5:di-Gluebody sample was incubated on glow discharged CF300-CU grid (Electron Microscopy Sciences) for 1 minute, before the grid was washed against uranyl acetate for three times (10 seconds each time). The grid was finally dried with filter paper and stored in the grid box for EM analysis.

The negatively stained EM grids were screened by T12 electron microscopy (Thermo) and collected micrographs were further analysed with RELION 3.1 for 2D classification and 3D reconstruction.

The EM images showed that, compared with monomeric RECQL5, the dimeric complexes enlarge by the eye and distribute evenly and mono-disperse. 2D classification and 3D reconstruction analysis were then carried out. The 3D reconstruction model showed that the complex took a dumbbell shape, suggesting the disulfide linked Gluebody molecules took a V-shape (Figure 5.3).

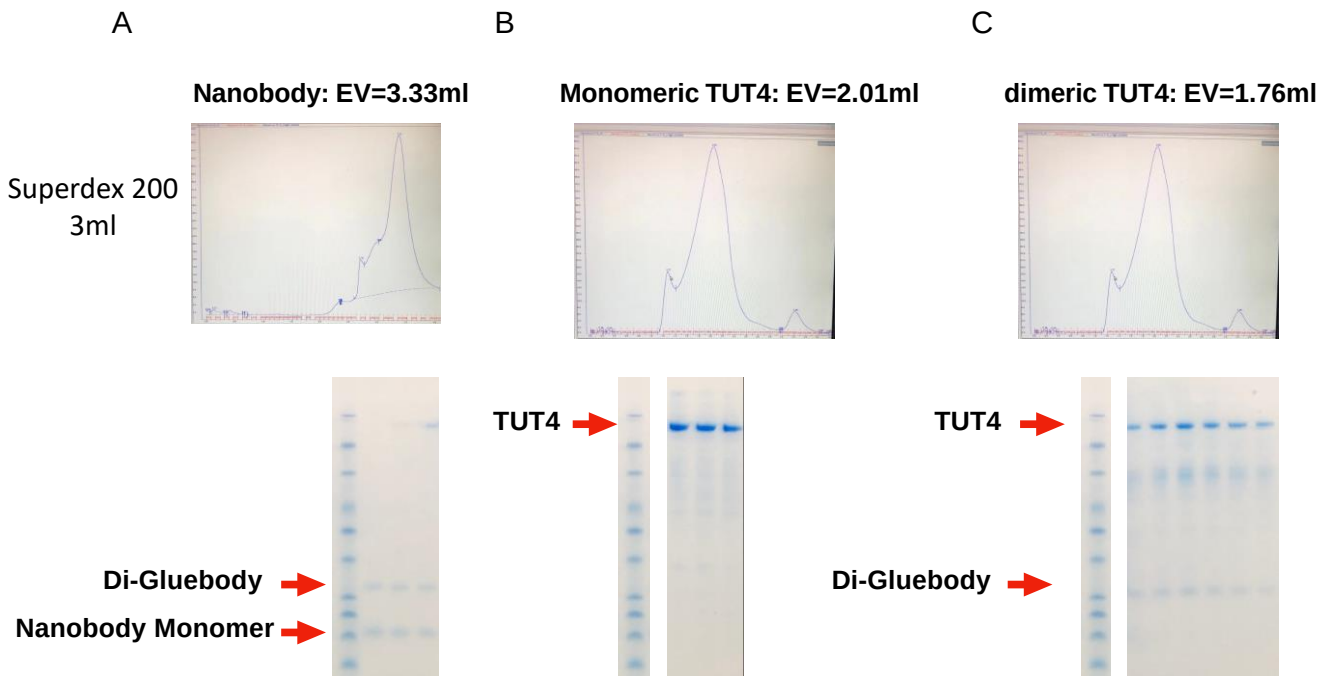


Figure 5.4 Size exclusion chromatography of Di-Gluebody (A), TUT4 (B), and the complex (C) by Superdex 200 5/150. EV is short for elution volume. The data stored in the STRUBI cannot be transferred out due to technical issues and therefore photos of the screen were taken.

The TUT4:Gluebody dimer complex was further purified by size exclusion chromatography (Figure 5.4). The eluted fractions were then collected, concentrated and frozen on the Quantifoil grid and cryo-EM analysis was performed on 200kV Glacios.

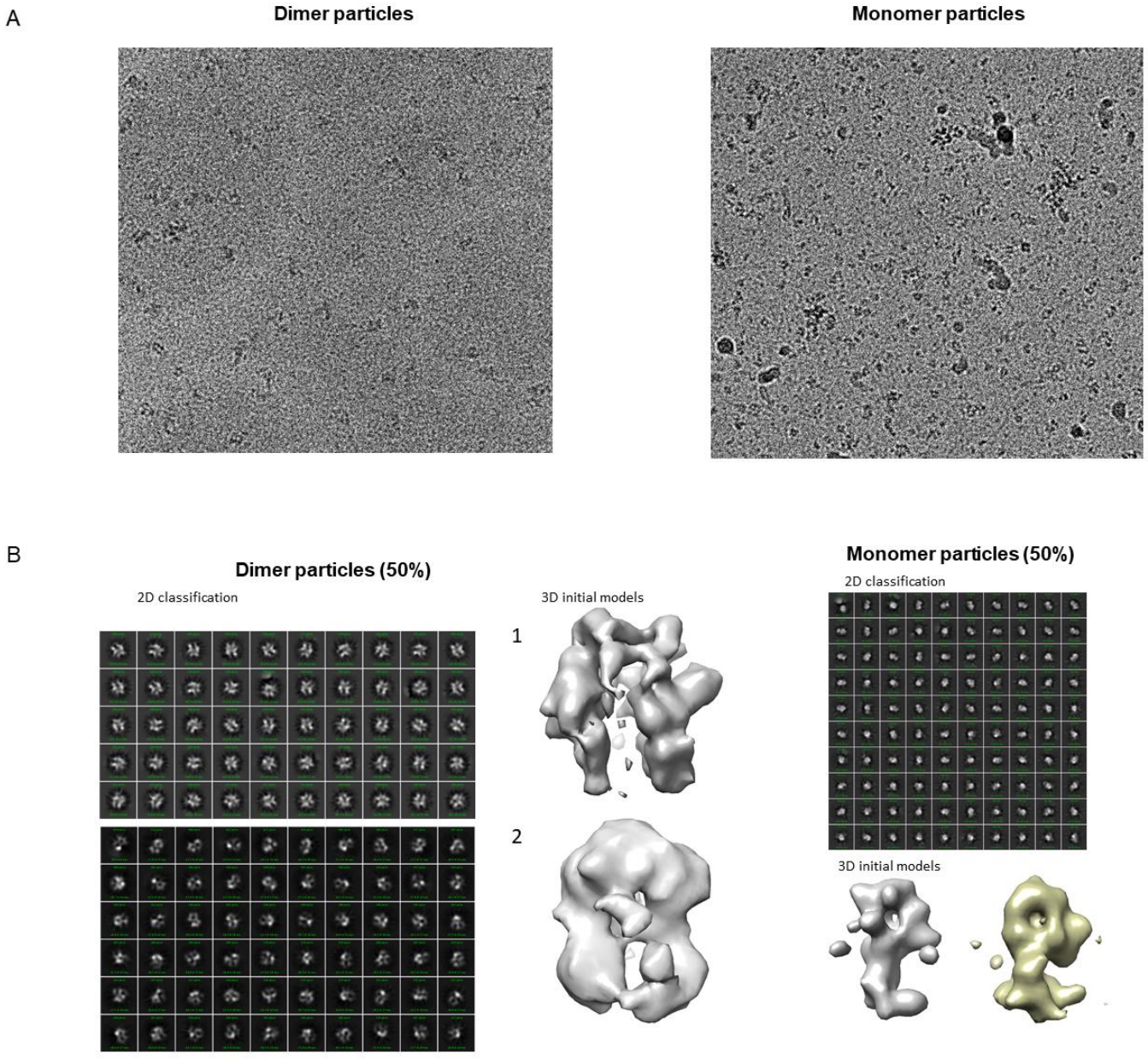


Figure 5.5 Glacios Cryo-EM analysis of the TUT4-DiGluebody complex. A) cryo-EM micrographs showing the dimer TUT4:di-Gluebody complex sample (left) and the monomeric TUT4 sample (right). B) 2-D classification and 3-D initial models of the selected particles. Half of selected particles are dimer particles and the other half are monomeric TUT4-nanobody complex. Dimer particles fall into mainly two conformations, open (1) and close (2), indicating two states of the Di-Gluebody.

To prepare the cryo grids for cryo-EM analysis, 3.5 µl of TUT4:di-Gluebody sample was incubated on glow discharged Quantifoil 2/1 300 mesh copper grid for 5-10 seconds and frozen quickly in liquid ethane with the assistance of Vitrobot. A series of parameters (pressing force, waiting time) were tested for an optimal thickness of ice on the grid.

The cryo grids were screened by 200 kV Glacios electron microscopy (Thermo) and collected micrographs were further analysed with Cryosparc for 2D classification and 3D classification.

The complex particles appeared mono-disperse in the micrograph. Around 10000 particles were picked and 2-D classification was performed. The result (Figure 5.5) showed half of selected particles were dimer particles and the other half were monomeric TUT4-nanobody complex. The 3-D classification of the dimer particles showed that two conformational states were predominant, one open and the other close. The divergence of the conformation might result from different conformational states of the Di-Gluebody or the flexible TUT4. More data collection and further refinement of the structures will be performed.

There have been no solid conclusions on whether the Di-Gluebody a is suitable imaging scaffold yet, and since the TUT4 construct we currently use might be flexible itself, making it hard to decide where the problem lied. The next steps would be trimming the TUT4 construct to only contain a rigid domain and a positive result would be the successful cryo-EM structure solution above 4 Å.

5.3 Future perspective: Tetra-Gluebodies

The tetra-Gluebodies described in Chapter 4 are a tetramer *in crystallo* mediated by two perpendicular interfaces (Nb-α and C-C interface). This arrangement is more likely to be rigid than the disulfide-linked Di-Gluebody. However, the bottleneck is the chemical assembly of the Tetra-Gluebody.

The experiment is still ongoing in collaboration with Dimitrios Mammalis from Prof. Ben Davis' lab. There are several strategies for linking up the tetra-Gluebody in solution – metal coordination, tetra-functional linkers, and metal coordination in combination with bis-functional linkers (Figure 5.6). We hope to obtain a rigidly linked tetra-Gluebody in solution, making use of its D2 symmetry, which would be of great help for cryo-EM structure solution.

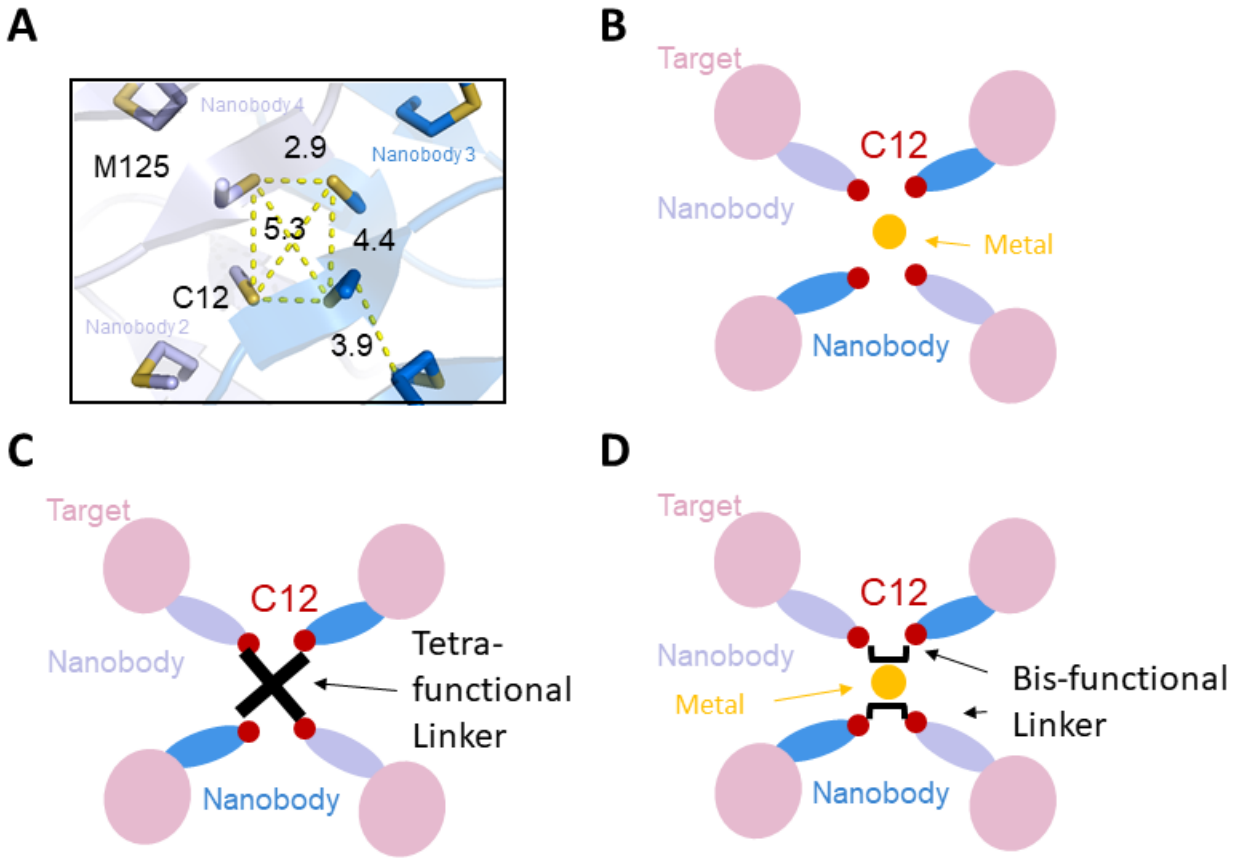


Figure 5.6 Potential Tetra-Gluebody assembly strategies. A) Close view of the C-C interface core of the Tetra-Gluebody. The distances shown are in Å. B) Metal ions might act as glue to trigger Tetra-Gluebody formation in solution. C) Tetra-functional linkers might act as bridge to trigger Tetra-Gluebody by reaction, but the product yield may be very low due to the stereo hindrance. D) Two step reaction to enable Tetra-Gluebody. Bis-functional linkers connect two nanobodies to become dimers and secondly the metal ions coordinates the linker-connected nanobodies to become Tetra-Gluebodies.

Chapter 6 Conclusions and perspectives

6.1 Conclusions

This thesis presents the first systematic engineering study on the optimization of nanobody scaffolds for crystallization purposes, which also sheds light on its potential use as an imaging scaffold for cryo-EM.

6.1.1 Naïve nanobody scaffolds have potential to optimize

I first argued that the naïve nanobodies are dissatisfactory crystal chaperones based on the crystallization result of 17 target:nanobody pairs, all of which are soluble proteins. Only one pair of them resulted in structures. Although nanobody has been demonstrated a very effective crystal chaperone for membrane proteins, it seems that it might not be effective enough to assist crystallization of soluble proteins.

To test this hypothsis, I performed mutagenesis on one of the failed nanobody scaffolds, which immediately rescued its crystallization ability, providing more evidence to support that naïve nanobody scaffolds have potential to optimize.

6.1.2 Nb-α is the most predominant nanobody intermolecular crystal interface

To prepare for the scaffold engineering, bioinformatics research was carried out to guide the engineering. After clustering large nanobody-nanobody interfaces from 335 crystal structures in PDB, 9 classes of Nb-Nb interfaces have emerged, of which Nb-α is the most predominant, accounting for 39.5%. It is a β-sheet-β-sheet interface consisting of around 10 interacting residues. It has two subtyeps depending on whether the L12 residue is aligned or shifted. While Nb-α has the highest priority for engineering, class 2 and class 4 interfaces, which are also away from the CDRs, also have potential for engineering.

6.1.3 High-resolution crystal structures of human RECQL5 helicase APO form (nanobody complex) was obtained through engineering Nb-α

With extensive exploration around Nb-α, I was able to make this interface reliable, which led to multiple polymorphs all mediated by this interface in either dimeric or tetrameric arrangement, and one set of mutations yielding far superior reliability of diffraction.

With the best set of optimized nanobodies, APO form helicase RECQL5 crystals were able to consistently diffract to 2 Å, which was much improved compared to the existing 3.4 Å APO form RECQL5 structure and

enabled crystallographic fragment screening on RECQL5.

Since the modified nanobody scaffolds tend to form dimers or tetramers with constitutive interfaces (Nb-α and C-C interface), I refer to them as “Gluebodies” and conclude that there is a large but under-exploited potential for improving the chaperone behaviour of nanobodies.

6.1.4 The Gluebody is a powerful tool but might not be useful for all proteins

The RECQL5 case shows that the Gluebody is a powerful approach to improve the packing and resolution limit of poorly diffracting crystals. Since the Gluebody mutations are far from the CDRs, they are easily transferrable to other target:nanobody pairs. I then applied the Gluedodies to a set of intrinsically non-crystallizing proteins, but they fell short on helping those proteins crystallize, which shows that the Gluebody is a powerful tool but not uniformly useful for all.

6.1.5 The Gluebody scaffolds can be potentially used as imaging scaffolds for cryo-EM

For the cryo-EM part, I showed initial results on the disulfide linked di-Gluebody used as an imaging scaffold. The di-Gluebody increased the particle size of small proteins, and appeared mono-disperse in the negative staining micrograph and the cryo-EM micrograph. The tetrameric arrangement of the Gluebody also seems useful as an imaging scaffold for cryo-EM once it can be made a rigidly and covalently-linked molecule in solution. Further exploration in chemistry should be carried out.

6.2 Perspectives

6.2.1 Nb-Nb v.s. Nb-target interfaces

In the RECQL5 case, engineering on both Nb-Nb and Nb-target interfaces are demonstrated. In this case, the G5 Gluebodies with Nb-α modifications have outperformed the D69Y mutation in the Nb-target interface. However, this might not be applicable to all proteins.

The advantage of Engineering Nb-Nb interface is obvious. First, it is independent of the target and therefore the optimized scaffolds might become a product, with an estimatable rate of improvement based on a large number of cases. Second, trials can be focused on a reduced number of best-performing variants,

saving the tool users a lot of labour. The disadvantage is that this method has a ceiling in the rate of improvement. In this work, since Nb-α is only present in around 15% of nanobody containing structures, for an unknown protein, mutagenesis on this interface will have a maximum of 15% probability to influence the packing, which is the ceiling of its performance.

Improving Nb-target interfaces is just on the contrary. Its performance is not predictable and therefore no product can be generated, but only an engineering process is transferable, such as the mutagenesis guided by the crystal epitope strategy in this work. This means that for each different target, large numbers of trials have to be conducted each time, and all outcome is random. However, there might not be a performance ceiling or the ceiling is higher than the Nb-Nb engineering, because one can always try so many different engineering strategies and make more variants to achieve the goal.

Therefore both engineering strategies are worth using, and it is not certain that one strategy is better than the other for a specific case.

6.2.2 Large v.s. small Nb-Nb interfaces

In this work, more focus is on large (>5 interactin residues) Nb-Nb interfaces from the start of the bioinformatics study. This is based on the assumption that large interfaces might have a better chance for forming consititutive interfaces. However, this might be untrue and small interfaces may also have profound effects. A single mutation D69Y in this work, although mediates an Nb-target interface, has already been shown powerful in influencing crystal contacts. Therefore, powerful small Nb-Nb interfaces might also exist but remain unexplored. This will provide another direction to study the nanobody scaffolds for crystallization purposes.

6.2.3 Exploration on nucleation mechanisms

In this work, I discovered that different versions of Gluebodies can lead to six different crystal forms in different crystallization conditions. This is a good model to study the nucleation mechanisms behind the various crystal forms. There are two scientific questions regarding this phenomenon – are inter-particle interactions different across the six crystal forms? Do the six crystal forms have pre-nucleus clusters and do the pre-nucleus clusters differ. It is feasible to answer these questions using techniques such as ITC, DLS and

cryo-TEM. This will be helpful for expanding the human knowledge of protein crystallization mechanisms.

Chapter 7 Bibliography

Abhinandan, K. and A. C. Martin (2008). "Analysis and improvements to Kabat and structurally correct numbering of antibody variable domains." Molecular immunology **45**(14): 3832-3839.

Al-Lazikani, B., A. M. Lesk and C. Chothia (1997). "Standard conformations for the canonical structures of immunoglobulins." Journal of molecular biology **273**(4): 927-948.

Anil, B., C. Riedinger, J. A. Endicott and M. E. M. Noble (2013). "The structure of an MDM2-Nutlin-3a complex solved by the use of a validated MDM2 surface-entropy reduction mutant." Acta Crystallographica Section D **69**(8): 1358-1366.

Arbabi-Ghahroudi, M., J. Tanha and R. MacKenzie (2005). "Prokaryotic expression of antibodies." Cancer and Metastasis Reviews **24**(4): 501-519.

Arbabi Ghahroudi, M., A. Desmyter, L. Wyns, R. Hamers and S. Muyldermans (1997). "Selection and identification of single domain antibody fragments from camel heavy-chain antibodies." FEBS letters **414**(3): 521-526.

Asakura, S. and F. Oosawa (1958). "Interaction between particles suspended in solutions of macromolecules." Journal of Polymer Science **33**(126): 183-192.

Asherie, N., A. Lomakin and G. B. Benedek (1996). "Phase Diagram of Colloidal Solutions." Physical Review Letters **77**(23): 4832-4835.

Atha, D. H. and K. C. Ingham (1981). "Mechanism of precipitation of proteins by polyethylene glycols. Analysis in terms of excluded volume." Journal of Biological Chemistry **256**(23): 12108-12117.

Aydin, S. (2015). "A short history, principles, and types of ELISA, and our laboratory experience with peptide/protein analyses using ELISA." Peptides **72**: 4-15.

Banatao, D. R., D. Cascio, C. S. Crowley, M. R. Fleissner, H. L. Tienson and T. O. Yeates (2006). "An approach to crystallizing proteins by synthetic symmetrization." Proceedings of the National Academy of Sciences of the United States of America **103**(44): 16230-16235.

Barreca, M. L., N. Iraci, S. Biggi, V. Cecchetti and E. Biasini (2018). "Pharmacological Agents Targeting the Cellular Prion Protein." Pathogens **7**(1).

Batyuk, A., Y. Wu, A. Honegger, M. M. Heberling and A. Plückthun (2016). "DARPin-based crystallization chaperones exploit molecular geometry as a screening dimension in protein crystallography." Journal of molecular biology **428**(8): 1574-1588.

Beghein, E. and J. Gettemans (2017). "Nanobody technology: a versatile toolkit for microscopic imaging, protein–protein interaction analysis, and protein function exploration." Frontiers in immunology **8**: 771.

Berney, C. and G. Danuser (2003). "FRET or No FRET: A Quantitative Comparison." Biophysical Journal **84**(6): 3992-4010.

Bloomfield, V. A. and T. K. Lim (1978). [19] Quasi-elastic laser light scattering. Methods in Enzymology, Academic Press. **48:** 415-494.

Boder, E. T., M. Raeeszadeh-Sarmazdeh and J. V. Price (2012). "Engineering antibodies by yeast display." Archives of biochemistry and biophysics **526**(2): 99-106.

Bončina, M., J. Reščič and V. Vlachy (2008). "Solubility of Lysozyme in Polyethylene Glycol-Electrolyte Mixtures: The Depletion Interaction and Ion-Specific Effects." Biophysical Journal **95**(3): 1285-1294.

Bortot, L. O., V. L. Rangel, F. A. Pavlovici, K. El Omari, A. Wagner, J. Brandao-Neto, R. Talon, F. von Delft, A. G. Reidenbach, S. M. Vallabh, E. V. Minikel, S. Schreiber and M. C. Nonato (2021). "Novel quaternary structures of the human prion protein globular domain." Biochimie **191**: 118-125.

Boute, N., R. Jockers and T. Issad (2002). "The use of resonance energy transfer in high-throughput screening: BRET versus FRET." Trends in Pharmacological Sciences **23**(8): 351-354.

Brigham, E. O. and R. E. Morrow (1967). "The fast Fourier transform." IEEE Spectrum **4**(12): 63-70.

Brückner, A., C. Polge, N. Lentze, D. Auerbach and U. Schlattner (2009). "Yeast Two-Hybrid, a Powerful Tool for Systems Biology." International Journal of Molecular Sciences **10**(6): 2763-2788.

Buijs, H., A. Pomerleau, M. Fournier and W. Tam (1974). "Implementation of a fast Fourier transform (FFT) for image processing applications." IEEE Transactions on Acoustics, Speech, and Signal Processing **22**(6): 420-424.

Bujalowski, W. and T. M. Lohman (1989). "Negative co-operativity in Escherichia coli single strand binding protein-oligonucleotide interactions: II. Salt, temperature and oligonucleotide length effects." Journal of Molecular Biology **207**(1): 269-288.

Bukowska, M. A. and M. G. Grütter (2013). "New concepts and aids to facilitate crystallization." Current Opinion in Structural Biology **23**(3): 409-416.

Burg, J. S., J. R. Ingram, A. Venkatakrishnan, K. M. Jude, A. Dukkipati, E. N. Feinberg, A. Angelini, D. Waghray, R. O. Dror and H. L. Ploegh (2015). "Structural basis for chemokine recognition and activation of a viral G protein–coupled receptor." Science **347**(6226): 1113-1117.

Chakraborty, S., K. Dutta, P. Gupta, A. Das, A. Das, S. K. Ghosh and B. S. Patro (2021). "Targeting RECQL5 Functions, by a Small Molecule, Selectively Kills Breast Cancer in Vitro and in Vivo." Journal of Medicinal Chemistry **64**(3): 1524-1544.

Chan, E. M., T. Shibue, J. M. McFarland, B. Gaeta, M. Ghandi, N. Dumont, A. Gonzalez, J. S. McPartlan, T. Li and Y. Zhang (2019). "WRN helicase is a synthetic lethal target in microsatellite unstable cancers." Nature **568**(7753): 551-556.

Chan, K. F., B. S. Gully, S. Gras, D. X. Beringer, L. Kjer-Nielsen, J. Cebon, J. McCluskey, W. Chen and J. Rossjohn (2018). "Divergent T-cell receptor recognition modes of a HLA-I restricted extended tumour-associated peptide." Nature communications **9**(1): 1-13.

Che, T., S. Majumdar, S. A. Zaidi, P. Ondachi, J. D. McCorvy, S. Wang, P. D. Mosier, R. Uprety, E. Vardy and B. E. Krumm (2018). "Structure of the nanobody-stabilized active state of the kappa opioid receptor." Cell **172**(1-2): 55-67. e15.

Chen, E., J. S. Ahn, D. B. Sykes, L. J. Breyfogle, A. L. Godfrey, J. Nangalia, A. Ko, D. J. DeAngelo, A. R. Green and A. Mullally (2015). "RECQL5 Suppresses Oncogenic JAK2-Induced Replication Stress and Genomic Instability." Cell reports **13**(11): 2345-2352.

Chen, J., A. J. Noble, J. Y. Kang and S. A. Darst (2019). "Eliminating effects of particle adsorption to the air/water interface in single-particle cryo-electron microscopy: Bacterial RNA polymerase and CHAPSO." Journal of Structural Biology: X **1**: 100005.

Chen, Y.-T., M. J. Scanlan, U. Sahin, Ö. Türeci, A. O. Gure, S. Tsang, B. Williamson, E. Stockert, M. Pfreundschuh and L. J. Old (1997). "A testicular antigen aberrantly expressed in human cancers detected by autologous antibody screening." Proceedings of the National Academy of Sciences **94**(5): 1914-1918.

Cheng, Y. (2015). "Single-Particle Cryo-EM at Crystallographic Resolution." Cell **161**(3): 450-457.

Cheng, Y. (2018). "Single-particle cryo-EM—How did it get here and where will it go." Science **361**(6405): 876-880.

Chesterman, C. and E. Arnold (2021). "Co-crystallization with diabodies: a case study for the introduction of synthetic symmetry." Structure.

Chesterman, C. and E. Arnold (2021). "Co-crystallization with diabodies: a case study for the introduction of synthetic symmetry." Structure **29**(6): 598-605. e593.

Chien, C. T., P. L. Bartel, R. Sternglanz and S. Fields (1991). "The two-hybrid system: a method to identify and clone genes for proteins that interact with a protein of interest." Proceedings of the National Academy of Sciences **88**(21): 9578.

Chothia, C., A. M. Lesk, E. Gherardi, I. M. Tomlinson, G. Walter, J. D. Marks, M. B. Llewelyn and G. Winter (1992). "Structural repertoire of the human VH segments." Journal of molecular biology **227**(3): 799-817.

Chun, E., A. A. Thompson, W. Liu, C. B. Roth, M. T. Griffith, V. Katritch, J. Kunken, F. Xu, V. Cherezov and M. A. Hanson (2012). "Fusion partner toolchest for the stabilization and crystallization of G protein-coupled receptors." Structure **20**(6): 967-976.

Chung, S., S.-H. Shin, C. R. Bertozzi and J. J. De Yoreo (2010). "Self-catalyzed growth of S layers via an amorphous-to-crystalline transition limited by folding kinetics." Proceedings of the National Academy of Sciences **107**(38): 16536-16541.

Clemens, Vonrhein, Claus, Flensburg, Peter, Keller, Andrew, Sharff, Oliver and Smart (2011). "Data processing and analysis with theautoPROCtoolbox." Acta Crystallographica Section D Biological Crystallography **67**(4).

Collins, P. M., A. Douangamath, R. Talon, A. Dias, J. Brandao-Neto, T. Krojer and F. von Delft (2018). "Achieving a good crystal system for crystallographic X-ray fragment screening." Methods in enzymology **610**: 251-264.

Collins, P. M., A. Douangamath, R. Talon, A. Dias, J. Brandao-Neto, T. Krojer and F. von Delft (2018). Chapter Eleven - Achieving a Good Crystal System for Crystallographic X-Ray Fragment Screening. Methods in Enzymology. C. A. Lesburg, Academic Press. **610:** 251-264.

Coons, A. H., H. J. Creech and R. N. Jones (1941). "Immunological Properties of an Antibody Containing a Fluorescent Group." Proceedings of the Society for Experimental Biology and Medicine **47**(2): 200-202.

Cooper, D. R., T. Boczek, K. Grelewska, M. Pinkowska, M. Sikorska, M. Zawadzki and Z. Derewenda (2007). "Protein crystallization by surface entropy reduction: optimization of the SER strategy." Acta Crystallographica Section D **63**(5): 636-645.

Corsini, L., M. Hothorn, K. Scheffzek, M. Sattler and G. Stier (2008). "Thioredoxin as a fusion tag for carrier-driven crystallization."

Protein Science **17**(12): 2070-2079.

Cummins, H. Z., N. Knable and Y. Yeh (1964). "Observation of Diffusion Broadening of Rayleigh Scattered Light." Physical Review Letters **12**(6): 150-153.

Curth, U., C. Urbanke, J. Greipel, H. Gerberding, V. Tiranti and M. Zeviani (1994). "Single-stranded-DNA-binding proteins from human mitochondria and Escherichia coli have analogous physicochemical properties." European Journal of Biochemistry **221**(1): 435-443.

Danev, R. and W. Baumeister (2016). "Cryo-EM single particle analysis with the Volta phase plate." Elife **5**: e13046.

Day, P. W., S. G. Rasmussen, C. Parnot, J. J. Fung, A. Masood, T. S. Kobilka, X.-J. Yao, H.-J. Choi, W. I. Weis and D. K. Rohrer (2007). "A monoclonal antibody for G protein–coupled receptor crystallography." Nature methods **4**(11): 927-929.

De Genst, E., F. Handelberg, A. Van Meirhaeghe, S. Vynck, R. Loris, L. Wyns and S. Muyldermans (2004). "Chemical basis for the affinity maturation of a camel single domain antibody." Journal of Biological Chemistry **279**(51): 53593-53601.

de Marco, A. (2020). "Recombinant expression of nanobodies and nanobody-derived immunoreagents." Protein expression and purification **172**: 105645.

Delagenière, S., P. Brenchereau, L. Launer, A. Ashton, R. Leal, S. Veyrier, J. Gabadinho, E. Gordon, S. Jones, K. Levik and S. McSweeney (2011). "ISPyB: an information management system for synchrotron macromolecular crystallography." Bioinformatics.

Deng, H., S. Grunder, K. E. Cordova, C. Valente, H. Furukawa, M. Hmadeh, F. Gándara, A. C. Whalley, Z. Liu and S. Asahina (2012). "Large-pore apertures in a series of metal-organic frameworks." science **336**(6084): 1018-1023.

Derewenda, Z. S. (2004). "Rational protein crystallization by mutational surface engineering." Structure **12**(4): 529-535.

Derewenda, Z. S. and P. G. Vekilov (2006). "Entropy and surface engineering in protein crystallization." Acta Crystallographica Section D **62**(1): 116-124.

Dixit, C. K., S. K. Vashist, F. T. O'Neill, B. O'Reilly, B. D. MacCraith and R. O'Kennedy (2010). "Development of a High Sensitivity Rapid Sandwich ELISA Procedure and Its Comparison with the Conventional Approach." Analytical Chemistry **82**(16): 7049-7052.

Doyle, J. M., J. Gao, J. Wang, M. Yang and P. R. Potts (2010). "MAGE-RING protein complexes comprise a family of E3 ubiquitin ligases." Molecular cell **39**(6): 963-974.

Doyle, M. L. (1997). "Characterization of binding interactions by isothermal titration calorimetry." Current Opinion in Biotechnology **8**(1): 31-35.

Du, D., Z. Wang, N. R. James, J. E. Voss, E. Klimont, T. Ohene-Agyei, H. Venter, W. Chiu and B. F. Luisi (2014). "Structure of the AcrAB–TolC multidrug efflux pump." Nature **509**(7501): 512-515.

Dumoulin, M., K. Conrath, A. Van Meirhaeghe, F. Meersman, K. Heremans, L. G. Frenken, S. Muyldermans, L. Wyns and A. Matagne (2002). "Single-domain antibody fragments with high conformational stability." Protein Science **11**(3): 500-515.

Dunbar, J. and C. M. Deane (2016). "ANARCI: antigen receptor numbering and receptor classification." Bioinformatics **32**(2): 298-300.

Durbin, S. D. and G. Feher (1996). "PROTEIN CRYSTALLIZATION." Annual Review of Physical Chemistry **47**(1): 171-204.

Emsley, P. and K. Cowtan (2004). "Coot: model-building tools for molecular graphics." Acta Crystallographica Section D **60**(12 Part 1): 2126-2132.

Endo, M., M. A. de Graaff, D. R. Ingram, S. Lim, D. C. Lev, I. H. Briaire-de Bruijn, N. Somaiah, J. V. M. G. Bovée, A. J. Lazar and T. O. Nielsen (2015). "NY-ESO-1 (CTAG1B) expression in mesenchymal tumors." Modern Pathology **28**(4): 587-595.

Engvall, E. and P. Perlmann (1971). "Enzyme-linked immunosorbent assay (ELISA) quantitative assay of immunoglobulin G." Immunochemistry **8**(9): 871-874.

Erdemir, D., A. Y. Lee and A. S. Myerson (2009). "Nucleation of Crystals from Solution: Classical and Two-Step Models." Accounts of Chemical Research **42**(5): 621-629.

Ester, M., H. P. Kriegel, J. Sander and X. Xu (1996). "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise." AAAI Press.

Evans, R. M. L. and W. C. K. Poon (1997). "Diffusive evolution of stable and metastable phases. II. Theory of nonequilibrium behavior in colloid-polymer mixtures." Physical Review E **56**(5): 5748-5758.

Faber, H. and B. Matthews (1990). "A mutant T4 lysozyme displays five different crystal conformations." Nature **348**(6298): 263-

266.

Fairhead, M., C. Preger, E. Wigren, C. Strain-Damerell, E. Ossipova, M. Ye, M. Makola, N. A. Burgess-Brown, H. Persson and F. Von Delft (2019). "Strategies to improve scFvs as crystallization chaperones suggested by analysis of a complex with the human PHD-bromodomain SP140." BioRxiv: 767376.

Fan, X., J. Wang, X. Zhang, Z. Yang, J.-C. Zhang, L. Zhao, H.-L. Peng, J. Lei and H.-W. Wang (2019). "Single particle cryo-EM reconstruction of 52 kDa streptavidin at 3.2 Angstrom resolution." Nature communications **10**(1): 1-11.

Fernandez-Leiro, R. and S. H. W. Scheres (2016). "Unravelling biological macromolecules with cryo-electron microscopy." Nature **537**(7620): 339-346.

Fields, S. and O.-k. Song (1989). "A novel genetic system to detect protein–protein interactions." Nature **340**(6230): 245-246.

Förster, T. (1948). "Zwischenmolekulare Energiewanderung und Fluoreszenz." Annalen der Physik **437**(1-2): 55-75.

Fox, J. D. and D. S. Waugh (2003). Maltose-binding protein as a solubility enhancer. E. coliGene Expression Protocols, Springer**:** 99-117.

Frank, J. (2006). Three-dimensional electron microscopy of macromolecular assemblies: visualization of biological molecules in their native state, Oxford university press.

Freire, E., O. L. Mayorga and M. Straume (1990). "Isothermal titration calorimetry." Analytical Chemistry **62**(18): 950A-959A.

Furuta, T., T. Ogawa, T. Katsuda, I. Fujii and H. Yamaji (2010). "Efficient production of an antibody Fab fragment using the baculovirus–insect cell system." Journal of bioscience and bioengineering **110**(5): 577-581.

Fusco, D., T. J. Barnum, A. E. Bruno, J. R. Luft, E. H. Snell, S. Mukherjee and P. Charbonneau (2014). "Statistical analysis of crystallization database links protein physico-chemical features with crystallization mechanisms." PloS one **9**(7): e101123-e101123.

Fusco, D., J. J. Headd, A. De Simone, J. Wang and P. Charbonneau (2014). "Characterizing protein crystal contacts and their role in crystallization: rubredoxin as a case study." Soft matter **10**(2): 290-302.

Gast, A. P., C. K. Hall and W. B. Russel (1983). "Polymer-induced phase separations in nonaqueous colloidal suspensions." Journal of Colloid and Interface Science **96**(1): 251-267.

Gebauer, D. and H. Cölfen (2011). "Prenucleation clusters and non-classical nucleation." Nano Today **6**(6): 564-584.

Goldschmidt, L., D. R. Cooper, Z. S. Derewenda and D. Eisenberg (2007). "Toward rational protein crystallization: A Web server for the design of crystallizable protein variants." Protein Science **16**(8): 1569-1576.

Grubisha, O., M. Kaminska, S. Duquerroy, E. Fontan, F. Cordier, A. Haouz, B. Raynal, J. Chiaravalli, M. Delepierre and A. Israël (2010). "DARPin-assisted crystallography of the CC2-LZ domain of NEMO reveals a coupling between dimerization and ubiquitin binding." Journal of molecular biology **395**(1): 89-104.

Gupta, S. K. and P. Shukla (2017). "Microbial platform technology for recombinant antibody fragment production: a review." Critical reviews in microbiology **43**(1): 31-42.

Hamadeh, Z. and P. Lansdorp (2020). "RECQL5 at the Intersection of Replication and Transcription." Frontiers in Cell and Developmental Biology **8**(324).

Hamers-Casterman, C., T. Atarhouch, S. a. Muyldermans, G. Robinson, C. Hammers, E. B. Songa, N. Bendahman and R. Hammers (1993). "Naturally occurring antibodies devoid of light chains." Nature **363**(6428): 446-448.

Hartnett, B. J., J. Gracyalny, D. Ph., P. Corporation and M. R. Slater "The Single Step KRX Competent Cells: Efficient Cloning and High Protein Yields." Promega Notes.

He, Y.-J., Z.-Y. Qiao, B. Gao, X.-H. Zhang and Y.-Y. Wen (2014). "Association between RECQL5 genetic polymorphisms and susceptibility to breast cancer." Tumor Biology **35**(12): 12201-12204.

Heinz, D. W. and B. W. Matthews (1994). "Rapid crystallization of T4 lysozyme by intermolecular disulfide cross-linking." Protein Engineering, Design and Selection **7**(3): 301-307.

Heo, I., M. Ha, J. Lim, M.-J. Yoon, J.-E. Park, S. C. Kwon, H. Chang and V. N. Kim (2012). "Mono-Uridylation of Pre-MicroRNA as a Key Step in the Biogenesis of Group II let-7 MicroRNAs." Cell **151**(3): 521-532.

Heo, I., C. Joo, Y.-K. Kim, M. Ha, M.-J. Yoon, J. Cho, K.-H. Yeom, J. Han and V. N. Kim (2009). "TUT4 in Concert with Lin28 Suppresses MicroRNA Biogenesis through Pre-MicroRNA Uridylation." Cell **138**(4): 696-708.

Herzik, M. A., M. Wu and G. C. Lander (2019). "High-resolution structure determination of sub-100 kDa complexes using

conventional cryo-EM." Nature communications **10**(1): 1-9.

Hickson, I. D. (2003). "RecQ helicases: caretakers of the genome." Nature Reviews Cancer **3**(3): 169-178.

Holcomb, J., N. Spellmon, Y. Zhang, M. Doughan, C. Li and Z. Yang (2017). "Protein crystallization: Eluding the bottleneck of X-ray crystallography." AIMS biophysics **4**(4): 557.

Holliger, P. and P. J. Hudson (2005). "Engineered antibody fragments and the rise of single domains." Nature biotechnology **23**(9): 1126-1136.

Homola, J., S. S. Yee and G. Gauglitz (1999). "Surface plasmon resonance sensors: review." Sensors and Actuators B: Chemical **54**(1): 3-15.

Honegger, A. and A. PluÈckthun (2001). "Yet another numbering scheme for immunoglobulin variable domains: an automatic modeling and analysis tool." Journal of molecular biology **309**(3): 657-670.

Hou, W. and S. B. Cronin (2013). "A Review of Surface Plasmon Resonance-Enhanced Photocatalysis." Advanced Functional Materials **23**(13): 1612-1619.

Hu, Y., S. Raynard, M. G. Sehorn, X. Lu, W. Bussen, L. Zheng, J. M. Stark, E. L. Barnes, P. Chi, P. Janscak, M. Jasin, H. Vogel, P. Sung and G. Luo (2007). "RECQL5/Recql5 helicase regulates homologous recombination and suppresses tumor formation via disruption of Rad51 presynaptic filaments." Genes & development **21**(23): 3073-3084.

Huang, S., L. Lee, N. B. Hanson, C. Lenaerts, H. Hoehn, M. Poot, C. D. Rubin, D. F. Chen, C. C. Yang and H. Juch (2006). "The spectrum of WRN mutations in Werner syndrome patients." Human mutation **27**(6): 558-567.

Huang, W., A. Manglik, A. Venkatakrishnan, T. Laeremans, E. N. Feinberg, A. L. Sanborn, H. E. Kato, K. E. Livingston, T. S. Thorsen and R. C. Kling (2015). "Structural insights into μ-opioid receptor activation." Nature **524**(7565): 315-321.

Huber, T. R., L. F. Hartje, E. C. McPherson, A. E. Kowalski and C. D. Snow (2017). "Programmed Assembly of Host–Guest Protein Crystals." Small **13**(7): 1602703.

Hui, R. and A. Edwards (2003). "High-throughput protein crystallization." Journal of Structural Biology **142**(1): 154-161.

Ilett, S. M., A. Orrock, W. C. K. Poon and P. N. Pusey (1995). "Phase behavior of a model colloid-polymer mixture." Physical Review E **51**(2): 1344-1352.

Janin, J. and F. Rodier (1995). "Protein–protein interaction at crystal contacts." Proteins: Structure, Function, and Bioinformatics **23**(4): 580-587.

Jelesarov, I. and H. R. Bosshard (1999). "Isothermal titration calorimetry and differential scanning calorimetry as complementary tools to investigate the energetics of biomolecular recognition." Journal of Molecular Recognition **12**(1): 3-18.

Jerabek-Willemsen, M., C. J. Wienken, D. Braun, P. Baaske and S. Duhr (2011). "Molecular interaction studies using microscale thermophoresis." Assay and drug development technologies **9**(4): 342-353.

Jucker, M. and L. C. Walker (2013). "Self-propagation of pathogenic protein aggregates in neurodegenerative diseases." Nature **501**(7465): 45-51.

Kabat, E. A., T. Te Wu, H. M. Perry, C. Foeller and K. S. Gottesman (1992). Sequences of proteins of immunological interest, DIANE publishing.

Kato, K., Y. Hamaguchi, S. Okawa, E. Ishikawa, K. Kobayashi and N. Katunuma (1977). "Use of Rabbit Antibody IgG Bound onto Plain and Aminoalkylsilyl Glass Surface for the Enzyme-Linked Sandwich Immunoassay." The Journal of Biochemistry **82**(1): 261-266.

Khoshouei, M., M. Radjainia, W. Baumeister and R. Danev (2017). "Cryo-EM structure of haemoglobin at 3.2 Å determined with the Volta phase plate." Nature communications **8**(1): 1-6.

Khoshouei, M., M. Radjainia, A. J. Phillips, J. A. Gerrard, A. K. Mitra, J. M. Plitzko, W. Baumeister and R. Danev (2016). "Volta phase plate cryo-EM of the small protein complex Prx3." Nature communications **7**(1): 1-5.

Kim, A.-R., T. Dobransky, R. J. Rylett and B. H. Shilton (2005). "Surface-entropy reduction used in the crystallization of human choline acetyltransferase." Acta Crystallographica Section D **61**(9): 1306-1310.

Knaus, K. J., M. Morillas, W. Swietnicki, M. Malone, W. K. Surewicz and V. C. Yee (2001). "Crystal structure of the human prion protein reveals a mechanism for oligomerization." Nature Structural Biology **8**(9): 770-774.

Koehl, A., H. Hu, D. Feng, B. Sun, Y. Zhang, M. J. Robertson, M. Chu, T. S. Kobilka, T. Laeremans and J. Steyaert (2019). "Structural insights into the activation of metabotropic glutamate receptors." Nature **566**(7742): 79-84.

Kohl, A., H. K. Binz, P. Forrer, M. T. Stumpp, A. Plückthun and M. G. Grütter (2003). "Designed to be stable: crystal structure of a consensus ankyrin repeat protein." Proceedings of the National Academy of Sciences **100**(4): 1700-1705.

Kohl, A., H. K. Binz, P. Forrer, M. T. Stumpp, A. Plückthun and M. G. Grütter (2003). "Designed to be stable: Crystal structure of a consensus ankyrin repeat protein." Proceedings of the National Academy of Sciences **100**(4): 1700.

Koide, A., V. Tereshko, S. Uysal, K. Margalef, A. A. Kossiakoff and S. Koide (2007). "Exploring the capacity of minimalist protein interfaces: interface energetics and affinity maturation to picomolar KD of a single-domain antibody with a flat paratope." Journal of molecular biology **373**(4): 941-953.

Kovari, L. C., C. Momany and M. G. Rossmann (1995). "The use of antibody fragments for crystallization and structure determinations." Structure **3**(12): 1291-1293.

Krojer, T., J. S. Fraser and F. von Delft (2020). "Discovery of allosteric binding sites by crystallographic fragment screening." Current Opinion in Structural Biology **65**: 209-216.

Kruse, A. C., A. M. Ring, A. Manglik, J. Hu, K. Hu, K. Eitel, H. Hübner, E. Pardon, C. Valant and P. M. Sexton (2013). "Activation and allosteric modulation of a muscarinic acetylcholine receptor." Nature **504**(7478): 101-106.

Kuge, M., Y. Fujii, T. Shimizu, T. Hakoshima, F. Hirose and A. Matsukage (1997). "Use of a fusion protein to obtain crystals suitable for X-ray analysis: Crystallization of a GST-fused protein containing the DNA-binding domain of DNA replication-related element-binding factor, DREF." Protein science **6**(8): 1783-1786.

Kulkarni, A. and C. Zukoski (2001). "Depletion interactions and protein crystallization." Journal of Crystal Growth **232**(1): 156-164.

Kummer, L., P. Parizek, P. Rube, B. Millgramm, A. Prinz, P. R. Mittl, M. Kaufholz, B. Zimmermann, F. W. Herberg and A. Plückthun (2012). "Structural and functional analysis of phosphorylation-specific binders of the kinase ERK from designed ankyrin repeat protein libraries." Proceedings of the National Academy of Sciences **109**(34): E2248-E2257.

Lally, J. M., R. H. Newman, P. P. Knowles, S. Islam, A. I. Coffer, M. Parker and P. S. Freemont (1998). "Crystallization of an intact GST-estrogen receptor hormone binding domain fusion protein." Acta Crystallographica Section D: Biological Crystallography **54**(3): 423-426.

Lavallie, E. R., E. A. DiBlasio, S. Kovacic, K. L. Grant, P. F. Schendel and J. M. McCoy (1993). "A thioredoxin gene fusion expression system that circumvents inclusion body formation in the E. coli cytoplasm." Bio/technology **11**(2): 187-193.

Laverty, D., R. Desai, T. Uchański, S. Masiulis, W. J. Stec, T. Malinauskas, J. Zivanov, E. Pardon, J. Steyaert, K. W. Miller and A. R. Aricescu (2019). "Cryo-EM structure of the human α1β3γ2 GABAA receptor in a lipid bilayer." Nature **565**(7740): 516-520.

Leavitt, S. and E. Freire (2001). "Direct measurement of protein binding energetics by isothermal titration calorimetry." Current Opinion in Structural Biology **11**(5): 560-566.

Ledsgaard, L., M. Kilstrup, A. Karatt-Vellatt, J. McCafferty and A. H. Laustsen (2018). "Basics of antibody phage display technology." Toxins **10**(6): 236.

Lee, A. K. and P. R. Potts (2017). "A Comprehensive Guide to the MAGE Family of Ubiquitin Ligases." Journal of Molecular Biology **429**(8): 1114-1142.

Lefranc, M.-P. and G. Lefranc (2001). The immunoglobulin factsbook, Academic press.

Lefranc, M.-P., C. Pommié, M. Ruiz, V. Giudicelli, E. Foulquier, L. Truong, V. Thouvenin-Contet and G. Lefranc (2003). "IMGT unique numbering for immunoglobulin and T cell receptor variable domains and Ig superfamily V-like domains." Developmental & Comparative Immunology **27**(1): 55-77.

Lekkerkerker, H. N. W. (1997). "Strong, weak and metastable liquids." Physica A: Statistical Mechanics and its Applications **244**(1): 227-237.

Lekkerkerker, H. N. W., W. C. K. Poon, P. N. Pusey, A. Stroobants and P. B. Warren (1992). "Phase Behaviour of Colloid + Polymer Mixtures." Europhysics Letters (EPL) **20**(6): 559-564.

Lieu, R., S. Antonysamy, Z. Druzina, C. Ho, N. R. Kang, A. Pustilnik, J. Wang and S. Atwell (2020). "Rapid and robust antibody Fab fragment crystallization utilizing edge-to-edge beta-sheet packing." PloS one **15**(9): e0232311.

Lim, J., M. Ha, H. Chang, S. C. Kwon, Dhirendra K. Simanshu, Dinshaw J. Patel and V. N. Kim (2014). "Uridylation by TUT4 and TUT7 Marks mRNA for Degradation." Cell **159**(6): 1365-1376.

Liu, W., H. Song, Q. Chen, J. Yu, M. Xian, R. Nian and D. Feng (2018). "Recent advances in the selection and identification of antigen-

specific nanobodies." Molecular immunology **96**: 37-47.

Liu, Y., S. Gonen, T. Gonen and T. O. Yeates (2018). "Near-atomic cryo-EM imaging of a small protein displayed on a designed scaffolding system." Proceedings of the National Academy of Sciences **115**(13): 3362-3367.

Liu, Y., D. T. Huynh and T. O. Yeates (2019). "A 3.8 Å resolution cryo-EM structure of a small protein bound to an imaging scaffold." Nature communications **10**(1): 1-7.

Maier, R., G. Zocher, A. Sauter, S. Da Vela, O. Matsarskaia, R. Schweins, M. Sztucki, F. Zhang, T. Stehle and F. Schreiber (2020). "Protein Crystallization in the Presence of a Metastable Liquid–Liquid Phase Separation." Crystal Growth & Design **20**(12): 7951-7962.

Maita, N. (2018). "Crystal structure determination of ubiquitin by fusion to a protein that forms a highly porous crystal lattice." Journal of the American Chemical Society **140**(42): 13546-13549.

Males, A. and G. J. Davies (2019). "Structural studies of a surface-entropy reduction mutant of O-GlcNAcase." Acta Crystallographica Section D **75**(1): 70-78.

Manglik, A., B. K. Kobilka and J. Steyaert (2017). "Nanobodies to study G protein–coupled receptor structure and function." Annual review of pharmacology and toxicology **57**: 19-37.

Marblestone, J. G., S. C. Edavettal, Y. Lim, P. Lim, X. Zuo and T. R. Butt (2006). "Comparison of SUMO fusion technology with traditional gene fusion systems: enhanced expression and solubility with SUMO." Protein Science **15**(1): 182-189.

Mateja, A., Y. Devedjiev, D. Krowarsch, K. Longenecker, Z. Dauter, J. Otlewski and Z. S. Derewenda (2002). "The impact of Glu-->Ala and Glu-->Asp mutations on the crystallization properties of RhoGDI: the structure of RhoGDI at 1.3 A resolution." Acta Crystallographica Section D **58**(12): 1983-1991.

Maynard, J. and G. Georgiou (2000). "Antibody engineering." Annual review of biomedical engineering **2**(1): 339-376.

McCoy, A. J., R. W. Grosse-Kunstleve, P. D. Adams, M. D. Winn, L. C. Storoni and R. J. Read (2007). "Phaser crystallographic software." Journal of Applied Crystallography **40**(4): 658-674.

McMahon, C., A. S. Baier, R. Pascolutti, M. Wegrecki, S. Zheng, J. X. Ong, S. C. Erlandson, D. Hilger, S. G. Rasmussen and A. M. Ring (2018). "Yeast surface display platform for rapid discovery of conformationally selective nanobodies." Nature structural & molecular biology **25**(3): 289-296.

McPherson, A. (1976). "Crystallization of proteins from polyethylene glycol." Journal of Biological Chemistry **251**(20): 6300-6303.

Mendoza, L. G., P. McQuary, A. Mongan, R. Gangadharan, S. Brignac and M. Eggers (1999). "High-Throughput Microarray-Based Enzyme-Linked Immunosorbent Assay (ELISA)." BioTechniques **27**(4): 778-788.

Menezes, M. R., J. Balzeau and J. P. Hagan (2018). "3′ RNA Uridylation in Epitranscriptomics, Gene Regulation, and Disease." Frontiers in Molecular Biosciences **5**(61).

Miles, L. E. M. and C. N. Hales (1968). "Labelled Antibodies and Immunological Assay Systems." Nature **219**(5150): 186-189.

Mitchell, L. S. and L. J. Colwell (2018). "Comparative analysis of nanobody sequence and structure data." Proteins: Structure, Function, and Bioinformatics **86**(7): 697-706.

Monegal, A., D. Ami, C. Martinelli, H. Huang, M. Aliprandi, P. Capasso, C. Francavilla, G. Ossolengo and A. de Marco (2009). "Immunological applications of single-domain llama recombinant antibodies isolated from a naive library." Protein Engineering, Design & Selection **22**(4): 273-280.

Moon, A. F., G. A. Mueller, X. Zhong and L. C. Pedersen (2010). "A synergistic approach to protein crystallization: combination of a fixed-arm carrier with surface entropy reduction." Protein Science **19**(5): 901-913.

Morin, J. G. and J. W. Hastings (1971). "Energy transfer in a bioluminescent system." Journal of Cellular Physiology **77**(3): 313-318.

Mou, D.-c., X.-s. Leng, J.-r. Peng, L. Zhao, W.-x. Wang, Y. Wang, T. Li, L.-g. Zhang, L. Huang and J.-y. Zhu (2004). "[Expression of MAGE-B genes in hepatocellular carcinoma]." Zhonghua zhong liu za zhi [Chinese journal of oncology] **26**(1): 40-42.

Murshudov, G. N., P. Skubak, A. A. Lebedev, N. S. Pannu, R. A. Steiner, R. A. Nicholls, M. D. Winn, F. Long and A. A. Vagin (2011). "REFMAC5 for the refinement of macromolecular crystal structures." Acta Crystallographica Section D **67**(4): 355-367.

Muyldermans, S. (2013). "Nanobodies: natural single-domain antibodies." Annual review of biochemistry **82**: 775-797.

Myszka, D. G., M. D. Jonsen and B. J. Graves (1998). "Equilibrium Analysis of High Affinity Interactions Using BIACORE." Analytical Biochemistry **265**(2): 326-330.

Myszka, D. G., R. W. Sweet, P. Hensley, M. Brigham-Burke, P. D. Kwong, W. A. Hendrickson, R. Wyatt, J. Sodroski and M. L. Doyle (2000). "Energetics of the HIV gp120-CD4 binding reaction." Proceedings of the National Academy of Sciences **97**(16): 9026.

Nallamsetty, S. and D. S. Waugh (2006). "Solubility-enhancing proteins MBP and NusA play a passive role in the folding of their fusion partners." Protein expression and purification **45**(1): 175-182.

Newman, J., E. E. Bolton, J. Muller-Dieckmann, V. J. Fazio, D. T. Gallagher, D. Lovell, J. R. Luft, T. S. Peat, D. Ratcliffe, R. A. Sayle, E. H. Snell, K. Taylor, P. Vallotton, S. Velanker and F. von Delft (2012). "On the need for an international effort to capture, share and use crystallization screening data." Acta Crystallographica Section F **68**(3): 253-258.

Newman, J. A., H. Aitkenhead, P. Savitsky and O. Gileadi (2017). "Insights into the RecQ helicase mechanism revealed by the structure of the helicase domain of human RECQL5." Nucleic Acids Research **45**(7): 4231-4243.

Ni, T. W. and F. A. Tezcan (2010). "Structural characterization of a microperoxidase inside a metal-directed protein cage." Angewandte Chemie International Edition **49**(39): 7014-7018.

Nicholaou, T., L. Ebert, I. D. Davis, N. Robson, O. Klein, E. Maraskovsky, W. Chen and J. Cebon (2006). "Directions in the immune targeting of cancer: Lessons learned from the cancer-testis Ag NY-ESO-1." Immunology and cell biology **84**(3): 303-317.

Nicoll, J. A. and J. Collinge (2009). "Preventing Prion Pathogenicity by Targeting the Cellular Prion Protein." Infectious Disorders - Drug Targets **9**(1): 48-57.

Nogales, E. and Sjors H. W. Scheres (2015). "Cryo-EM: A Unique Tool for the Visualization of Macromolecular Complexity." Molecular Cell **58**(4): 677-689.

Nygren, P.-Å., S. Stefan and M. Uhlén (1994). "Engineering proteins to facilitate bioprocessing." Trends in biotechnology **12**(5): 184-188.

Ohi, M., Y. Li, Y. Cheng and T. Walz (2004). "Negative staining and image classification—powerful tools in modern electron microscopy." Biological procedures online **6**(1): 23-34.

Padlan, E. A. (1994). "Anatomy of the antibody molecule." Molecular immunology **31**(3): 169-217.

Pai, J. C., J. A. Culver, J. E. Drury, R. S. Motani, R. L. Lieberman and J. A. Maynard (2011). "Conversion of scFv peptide-binding specificity for crystal chaperone development." Protein Engineering, Design & Selection **24**(5): 419-428.

Pardon, E., T. Laeremans, S. Triest, S. G. Rasmussen, A. Wohlkönig, A. Ruf, S. Muyldermans, W. G. Hol, B. K. Kobilka and J. Steyaert (2014). "A general protocol for the generation of Nanobodies for structural biology." Nature protocols **9**(3): 674-693.

Parker, J. L., J. C. Deme, G. Kuteyi, Z. Wu, J. Huo, I. D. Goldman, R. J. Owens, P. C. Biggin, S. M. Lea and S. Newstead (2021). "Structural basis of antifolate recognition and transport by PCFT." Nature **595**(7865): 130-134.

Patel, T. R., C. Bernards, M. Meier, K. McEleney, D. J. Winzor, M. Koch and J. Stetefeld (2014). "Structural elucidation of full-length nidogen and the laminin–nidogen complex in solution." Matrix Biology **33**: 60-67.

Pattnaik, P. (2005). "Surface plasmon resonance." Applied Biochemistry and Biotechnology **126**(2): 79-92.

Pecora, R. (1972). "Quasi-Elastic Light Scattering from Macromolecules." Annual Review of Biophysics and Bioengineering **1**(1): 257-276.

Pike, E. R. (1972). "THE ACCURACY OF DIFFUSION-CONSTANT MEASUREMENTS BY DIGITAL AUTOCORRELATION OF PHOTON-COUNTING FLUCTUATIONS." Le Journal de Physique Colloques **33**: C1-177-C171-180.

Plückthun, A. (2015). "Designed ankyrin repeat proteins (DARPins): binding proteins for research, diagnostics, and therapy." Annual review of pharmacology and toxicology **55**: 489-511.

Poon, W. C. K., A. D. Pirie, M. D. Haw and P. N. Pusey (1997). "Non-equilibrium behaviour of colloid-polymer mixtures." Physica A: Statistical Mechanics and its Applications **235**(1): 110-119.

Popuri, V., T. Tadokoro, D. L. Croteau and V. A. Bohr (2013). "Human RECQL5: guarding the crossroads of DNA replication and transcription and providing backup capability." Critical reviews in biochemistry and molecular biology **48**(3): 289-299.

Price Ii, W. N., Y. Chen, S. K. Handelman, H. Neely, P. Manor, R. Karlin, R. Nair, J. Liu, M. Baran and J. Everett (2009). "Understanding the physical properties that control protein crystallization by analysis of large-scale experimental data." Nature biotechnology **27**(1): 51-57.

Price Ii, W. N., Y. Chen, S. K. Handelman, H. Neely, P. Manor, R. Karlin, R. Nair, J. Liu, M. Baran, J. Everett, S. N. Tong, F. Forouhar, S. S. Swaminathan, T. Acton, R. Xiao, J. R. Luft, A. Lauricella, G. T. DeTitta, B. Rost, G. T. Montelione and J. F. Hunt (2009).

"Understanding the physical properties that control protein crystallization by analysis of large-scale experimental data." Nature Biotechnology **27**(1): 51-57.

Raman, M. C. C., P. J. Rizkallah, R. Simmons, Z. Donnellan, J. Dukes, G. Bossi, G. S. Le Provost, P. Todorov, E. Baston, E. Hickman, T. Mahon, N. Hassan, A. Vuidepot, M. Sami, D. K. Cole and B. K. Jakobsen (2016). "Direct molecular mimicry enables off-target cardiovascular toxicity by an enhanced affinity TCR designed for cancer immunotherapy." Scientific reports **6**: 18851-18851.

Rasmussen, S. G., H.-J. Choi, J. J. Fung, E. Pardon, P. Casarosa, P. S. Chae, B. T. DeVree, D. M. Rosenbaum, F. S. Thian and T. S. Kobilka (2011). "Structure of a nanobody-stabilized active state of the β 2 adrenoceptor." Nature **469**(7329): 175-180.

Rasmussen, S. G. F., H.-J. Choi, J. J. Fung, E. Pardon, P. Casarosa, P. S. Chae, B. T. DeVree, D. M. Rosenbaum, F. S. Thian, T. S. Kobilka, A. Schnapp, I. Konetzki, R. K. Sunahara, S. H. Gellman, A. Pautsch, J. Steyaert, W. I. Weis and B. K. Kobilka (2011). "Structure of a nanobody-stabilized active state of the β2 adrenoceptor." Nature **469**(7329): 175-180.

Rayment, I. (1997). [12] Reductive alkylation of lysine residues to alter crystallization properties of proteins. Methods in Enzymology, Academic Press. **276:** 171-179.

Reidenbach, A. G., M. F. Mesleh, D. Casalena, S. M. Vallabh, J. L. Dahlin, A. J. Leed, A. I. Chan, D. L. Usanov, J. B. Yehl, C. T. Lemke, A. J. Campbell, R. N. Shah, O. K. Shrestha, J. R. Sacher, V. L. Rangel, J. A. Moroco, M. Sathappa, M. C. Nonato, K. T. Nguyen, S. K. Wright, D. R. Liu, F. F. Wagner, V. K. Kaushik, D. S. Auld, S. L. Schreiber and E. V. Minikel (2020). "Multimodal small-molecule screening for human prion protein binders." Journal of Biological Chemistry **295**(39): 13516-13531.

Reimer, L. (2013). Transmission electron microscopy: physics of image formation and microanalysis, Springer.

Ring, A. M., A. Manglik, A. C. Kruse, M. D. Enos, W. I. Weis, K. C. Garcia and B. K. Kobilka (2013). "Adrenaline-activated structure of β 2-adrenoceptor stabilized by an engineered nanobody." Nature **502**(7472): 575-579.

Rosenthal, P. B. and R. Henderson (2003). "Optimal Determination of Particle Orientation, Absolute Hand, and Contrast Loss in Single-particle Electron Cryomicroscopy." Journal of Molecular Biology **333**(4): 721-745.

Rouet, R., D. Lowe, K. Dudgeon, B. Roome, P. Schofield, D. Langley, J. Andrews, P. Whitfeld, L. Jermutus and D. Christ (2012). "Expression of high-affinity human antibody fragments in bacteria." nature protocols **7**(2): 364-373.

Salema, V. and L. Á. Fernández (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.

Saridakis, E. and N. E. Chayen (2013). "Imprinted polymers assisting protein crystallization." Trends in biotechnology **31**(9): 515-520.

Sauter, A., F. Roosen-Runge, F. Zhang, G. Lotze, R. M. Jacobs and F. Schreiber (2015). "Real-time observation of nonclassical protein crystallization kinetics." Journal of the American Chemical Society **137**(4): 1485-1491.

Savitsky, P., J. Bray, C. D. Cooper, B. D. Marsden, P. Mahajan, N. A. Burgess-Brown and O. Gileadi (2010). "High-throughput production of human proteins for crystallization: the SGC experience." Journal of structural biology **172**(1): 3-13.

Savitsky, P., J. Bray, C. D. O. Cooper, B. D. Marsden, P. Mahajan, N. A. Burgess-Brown and O. Gileadi (2010). "High-throughput production of human proteins for crystallization: The SGC experience." Journal of Structural Biology **172**(1): 3-13.

Schroeder Jr, H. W. and L. Cavacini (2010). "Structure and function of immunoglobulins." Journal of Allergy and Clinical Immunology **125**(2): S41-S52.

Sennhauser, G. and M. G. Grütter (2008). "Chaperone-Assisted Crystallography with DARPins." Structure **16**(10): 1443-1453.

Shen, J.-C., M. D. Gray, J. Oshima, A. S. Kamath-Loeb, M. Fry and L. A. Loeb (1998). "Werner syndrome protein: I. DNA helicase and DNA exonuclease reside on the same polypeptide." Journal of Biological Chemistry **273**(51): 34139-34144.

Sleutel, M., J. Lutsko, A. E. Van Driessche, M. A. Durán-Olivencia and D. Maes (2014). "Observing classical nucleation theory at work by monitoring phase transitions with molecular precision." Nature communications **5**(1): 1-8.

Smith, G. P. and V. A. Petrenko (1997). "Phage display." Chemical reviews **97**(2): 391-410.

Smith, P. R. (1981). "The interpolation of fourier transform data in tilted-view three-dimensional reconstruction." Ultramicroscopy **7**(2): 155-159.

Steyaert, J. and B. K. Kobilka (2011). "Nanobody stabilization of G protein-coupled receptor conformational states." Current opinion in structural biology **21**(4): 567-572.

Stoll, V. S., A. V. Manohar, W. Gillon, E. L. Macfarlane, R. C. Hynes and E. F. Pai (1998). "A thioredoxin fusion protein of VanH, a D-

lactate dehydrogenase from Enterococcus faecium: Cloning, expression, purification, kinetic analysis, and crystallization." Protein science **7**(5): 1147-1155.

Strelov, V., I. Kuranova, B. Zakharov and A. Voloshin (2014). "Crystallization in space: Results and prospects." Crystallography Reports **59**(6): 781-806.

Ta, D. T., E. Steen Redeker, B. Billen, G. Reekmans, J. Sikulu, J.-P. Noben, W. Guedens and P. Adriaensens (2015). "An efficient protocol towards site-specifically clickable nanobodies in high yield: cytoplasmic expression in Escherichia coli combined with intein-mediated protein ligation." Protein Engineering, Design and Selection **28**(10): 351-363.

Tan, Y. Z., P. R. Baldwin, J. H. Davis, J. R. Williamson, C. S. Potter, B. Carragher and D. Lyumkis (2017). "Addressing preferred specimen orientation in single-particle cryo-EM through tilting." Nature Methods **14**(8): 793-796.

Tanaka, S. and M. Ataka (2002). "Protein crystallization induced by polyethylene glycol: A model study using apoferritin." The Journal of Chemical Physics **117**(7): 3504-3510.

Thorsen, T. S., R. Matt, W. I. Weis and B. K. Kobilka (2014). "Modified T4 lysozyme fusion proteins facilitate G protein-coupled receptor crystallogenesis." Structure **22**(11): 1657-1664.

Tsien, R. Y. (1998). "THE GREEN FLUORESCENT PROTEIN." Annual Review of Biochemistry **67**(1): 509-544.

Uchański, T., S. Masiulis, B. Fischer, V. Kalichuk, U. López-Sánchez, E. Zarkadas, M. Weckener, A. Sente, P. Ward, A. Wohlkönig, T. Zögg, H. Remaut, J. H. Naismith, H. Nury, W. Vranken, A. R. Aricescu, E. Pardon and J. Steyaert (2021). "Megabodies expand the nanobody toolkit for protein structure determination by single-particle cryo-EM." Nature Methods **18**(1): 60-68.

Uchański, T., E. Pardon and J. Steyaert (2020). "Nanobodies to study protein conformational states." Current opinion in structural biology **60**: 117-123.

Uchański, T., T. Zögg, J. Yin, D. Yuan, A. Wohlkönig, B. Fischer, D. M. Rosenbaum, B. K. Kobilka, E. Pardon and J. Steyaert (2019). "An improved yeast surface display platform for the screening of nanobody immune libraries." Scientific reports **9**(1): 1-12.

Ulrich, J. and T. Stelzer (2011). "Crystallization." Kirk-Othmer Encyclopedia of Chemical Technology: 1-63.

Van der Linden, R., L. Frenken, B. De Geus, M. Harmsen, R. Ruuls, W. Stok, L. De Ron, S. Wilson, P. Davis and C. Verrips (1999). "Comparison of physical chemical properties of llama VHH antibody fragments and mouse monoclonal antibodies." Biochimica et Biophysica Acta (BBA)-Protein Structure and Molecular Enzymology **1431**(1): 37-46.

Van Driessche, A. E., N. Van Gerven, P. H. Bomans, R. R. Joosten, H. Friedrich, D. Gil-Carton, N. A. Sommerdijk and M. Sleutel (2018). "Molecular nucleation mechanisms and control strategies for crystal polymorph selection." Nature **556**(7699): 89-94.

Van Heel, M. (1987). "Angular reconstitution: A posteriori assignment of projection directions for 3D reconstruction." Ultramicroscopy **21**(2): 111-123.

Van Weemen, B. K. and A. H. W. M. Schuurs (1971). "Immunoassay using antigen—enzyme conjugates." FEBS Letters **15**(3): 232-236.

Ward, A. R. and C. D. Snow (2020). "Porous crystals as scaffolds for structural biology." Current opinion in structural biology **60**: 85-92.

Warne, T., P. C. Edwards, A. S. Doré, A. G. Leslie and C. G. Tate (2019). "Molecular basis for high-affinity agonist binding in GPCRs." Science **364**(6442): 775-778.

Waugh, D. S. (2016). "Crystal structures of MBP fusion proteins." Protein Science **25**(3): 559-571.

Webb, A. I., M. A. Dunstone, W. Chen, M.-I. Aguilar, Q. Chen, H. Jackson, L. Chang, L. Kjer-Nielsen, T. Beddoe and J. McCluskey (2004). "Functional and structural characteristics of NY-ESO-1-related HLA A2-restricted epitopes and the design of a novel immunogenic analogue." Journal of Biological Chemistry **279**(22): 23438-23446.

Weisser, N. E. and J. C. Hall (2009). "Applications of single-chain variable fragment antibodies in therapeutics and diagnostics." Biotechnology advances **27**(4): 502-520.

Wienken, C. J., P. Baaske, U. Rothbauer, D. Braun and S. Duhr (2010). "Protein-binding assays in biological liquids using microscale thermophoresis." Nature Communications **1**(1): 100.

Winn, M. D., C. C. Ballard, K. D. Cowtan, E. J. Dodson, P. Emsley, P. R. Evans, R. M. Keegan, E. B. Krissinel, A. G. W. Leslie, A. McCoy, S. J. McNicholas, G. N. Murshudov, N. S. Pannu, E. A. Potterton, H. R. Powell, R. J. Read, A. Vagin and K. S. Wilson (2011). "Overview of the CCP4 suite and current developments." Acta Crystallographica Section D **67**(4): 235-242.

Winter, G. (2009). "xia2: an expert system for macromolecular crystallography data reduction." Journal of Applied Crystallography **43**(1): 186-190.

Winter, G., A. D. Griffiths, R. E. Hawkins and H. R. Hoogenboom (1994). "Making antibodies by phage display technology." Annual review of immunology **12**(1): 433-455.

WoÈrn, A. and A. PluÈckthun (2001). "Stability engineering of antibody single-chain Fv fragments." Journal of molecular biology **305**(5): 989-1010.

Wu, X. and T. A. Rapoport (2021). "Cryo-EM structure determination of small proteins by nanobody-binding scaffolds (Legobodies)." Proceedings of the National Academy of Sciences **118**(41): e2115001118.

Wukovitz, S. W. and T. O. Yeates (1995). "Why protein crystals favour some space-groups over others." Nature Structural Biology **2**(12): 1062-1067.

Yalow, R. S. and S. A. Berson (1960). "Immunoassay of endogenous plasma insulin in man." The Journal of clinical investigation **39**(7): 1157-1175.

Yamashita, S., T. Nagaike and K. Tomita (2019). "Crystal structure of the Lin28-interacting module of human terminal uridylyltransferase that regulates let-7 expression." Nature Communications **10**(1): 1960.

Yang, S. W., X. Huang, W. Lin, J. Min, D. J. Miller, A. Mayasundari, P. Rodrigues, E. C. Griffith, C. T. Gee, L. Li, W. Li, R. E. Lee, Z. Rankovic, T. Chen and P. R. Potts (2020). "Structural basis for substrate recognition and chemical inhibition of oncogenic MAGE ubiquitin ligases." Nature Communications **11**(1): 4931.

Yau, S.-T. and P. G. Vekilov (2000). "Quasi-planar nucleus structure in apoferritin crystallization." Nature **406**(6795): 494-497.

Yeates, T. O., M. P. Agdanowski and Y. Liu (2020). "Development of imaging scaffolds for cryo-electron microscopy." Current opinion in structural biology **60**: 142-149.

Yu, C., L. Wang, R. G. Rowe, A. Han, W. Ji, C. McMahon, A. S. Baier, Y.-C. Huang, W. Marion, D. S. Pearson, A. C. Kruse, G. Q. Daley, H. Wu and P. Sliz (2020). "A nanobody targeting the LIN28:let-7 interaction fragment of TUT4 blocks uridylation of let-7." Proceedings of the National Academy of Sciences **117**(9): 4653.

Zacharchenko, T. and S. Wright (2021). "Functionalization of the BCL6 BTB domain into a noncovalent crystallization chaperone." IUCrJ **8**(2).

Zanganeh, S., H. R. Nejad, J. F. Mehrabadi, R. Hosseini, B. Shahi, Z. Tavassoli and A. Aramvash (2019). "Rapid and sensitive detection of staphylococcal enterotoxin B by recombinant nanobody using phage display technology." Applied biochemistry and biotechnology **187**(2): 493-505.

Zhang, X.-j., J. A. Wozniak and B. W. Matthews (1995). "Protein flexibility and adaptability seen in 25 crystal forms of T4 lysozyme." Journal of molecular biology **250**(4): 527-552.

Zimmermann, I., P. Egloff, C. A. Hutter, F. M. Arnold, P. Stohler, N. Bocquet, M. N. Hug, S. Huber, M. Siegrist and L. Hetemann (2018). "Synthetic single domain antibodies for the conformational trapping of membrane proteins." elife **7**: e34317.

Zimmermann, I., P. Egloff, C. A. Hutter, B. T. Kuhn, P. Bräuer, S. Newstead, R. J. Dawson, E. R. Geertsma and M. A. Seeger (2020). "Generation of synthetic nanobodies against delicate proteins." Nature protocols **15**(5): 1707-1741.

Zimmermann, I., P. Egloff, C. A. J. Hutter, B. T. Kuhn, P. Bräuer, S. Newstead, R. J. P. Dawson, E. R. Geertsma and M. A. Seeger (2020). "Generation of synthetic nanobodies against delicate proteins." Nature Protocols **15**(5): 1707-1741.

Chapter 8 Appendix

8.1 Source Code for Chapter 3

```
import numpy
from sklearn.cluster import DBSCAN
import copy
import pymol
import os
import re
from collections import Counter


#########################################################################################################################
######################################## Finding interacting residues & ##################################################
########################### Grouping interacting residues into interfaces ###############################################
#########################################################################################################################

PDBlist =
['2XT1','5VAN','5VAK','4HEM','2X1O','2X1P','2X1Q','2X6M','2X89','3CFI','3DWT','3EAK','3EBA','3EZJ','3G9A','3K1K',
'3K74','3OGO','3POG','3SN6','3STB','3ZKQ','3ZKS','4AQ1','4BEL','4BFB','4C57','4C58','4C59','4CDG','4DK3','4DK6',
'4DKA','4EIG','4EIZ','4EJ1','4FHB','4GFT','4GRW','4HEP','4IOC','4I13','4I1N','4IOS','4JVP','4KDT','4KML','4KRL',
'4KRM','4KRN','4KRO','4KRP','4KSD','4LDE','4LDL','4LDO','4LGP','4LGR','4LGS','4LHJ','4LHQ','4MQS','4MQT','4N9O',
'4OCL','4OCM','4OCN','4P2C','4PIR','4QGY','4QKX','4QLR','4QO1','4S1O','4S11','4W6W','4W6X','4W6Y','4WEM','4WEN',
'4WEU','4WGV','4WGW','4X7C','4X7D','4X7E','4X7F','4XT1','4Y8D','4Z9K','4ZG1','5BOP','5BOZ','5C1M','5C2U','5C3L',
'5DAO','5DA4','5DFZ','5DXW','5E03','5EOQ','5E5M','5E7B','5E7F','5F1K','5F1O','5F21','5F7K','5F7L','5F7M','5F7N',
'5F7W','5F7Y','5F8Q','5F8R','5F93','5F97','5F9A','5F9D','5FWO','5G5R','5G5X','5GXB','5H8D','5H8O','5HDO','5HGG',
'5HVF','5HVG','5HVH','5IMK','5IML','5IMM','5IMO','5IP4','5IVN','5IVO','5J1S','5J1T','5JA8','5JA9','5JDS','5JQH',
'5LHN','5LHP','5LHQ','5LHR','5LMJ','5LMW','5LZO','5M13','5M14','5M15','5M2I','5M2J','5M2M','5M2W','5M2Y',
'5M30','5M7Q','5M94','5M95','5MJE','5MP2','5MWN','5MY6','5MZV','5NBD','5NBL','5NBM','5NLU','5NLW','5NMO','5NML',
'5NQW','5O02','5O03','5O04','5O05','5OOW','5O2U','5O8F','5OCL','5OJM','5OMM','5OMN','5OVW','5TD8','5TJW','5TOJ',
'5TOK','5TP3','5U64','5U65','5UK4','5UKB','5USF','5VL2','5VLV','5VMO','5VM4','5VM6','5VNV','5VNW','5WB1','5WB2',
'5Y7Z','5Y80','6B20','6B73','6C5W','6C9W','6DBA','6DBD','6DBE','6DBF','6DBG','6DO1','6DYX','6EHG','6EQI','6EYO',
'6EY6','6EZW','6FOD','6F2G','6F2W','6FE4','6FPV','6FUZ','6FVO','6GCI','6GJQ','6GJS','6GJU','6GK4','6GKD','6GS1',
'6GS4','6GS7','6GWN','6GWP','6GWQ','6GZP','6H02','6H15','6H16','6H1F','6H6Y','6H6Z','6H70','6H71','6H72','6H7J',
'6H7L','6H7M','6H7N','6H7O','6HD8','6HD9','6HDA','6HDB','6HDC','6HEQ','6HER','6HHD','6HHU','6HJX','6HJY','6I2G',
'6I6J','6I8G','6I8H','6IBB','6IBL','6ITP','6ITQ','6JB2','6JB5','6JB8','6JB9','6MXT','6N4Y','6N5O','6NFJ',
'6O3C','6OSO','6OS1','6OS2','6OYH','6OYZ','6OZ6','6Q6Z','6QD6','6QGW','6QGX','6QGY','6QNW','6QPG','6QTL','6QUZ',
'6QV0','6QV1','6QV2','6QX4','6R7T','6RNK','6RTW','6RTY','6RU3','6RU5','6RUL','6RUM','6RUV','6RVC','6SOY','6SGE',
'6SSI','6SSP','6TEJ','6TYL','6U12','6U14','6U50','6U51','6U52','6U53','6U54','6U55','6VI4','6WAQ','6WAR','2XV6',
'2XXC','2XXM','6SWR']

pymol.finish_launching(['pymol', '-qc'])



# out files
```

```
output = open('P:\\Documents\\Protein Interface\\contact finding script\\contactResidue.csv','w')
output2 = open('P:\\Documents\\Protein Interface\\contact finding script\\contactSurface.csv','w')

for structure in PDBlist:
    print 'Processing ' + structure + '...'
    output.write(structure + '\n')
    pymol.importing.fetch(structure)
    pymol.stored.profile = []
    pymol.cmd.iterate_state(1, structure + ' and name CA', "stored.profile.append((model, chain, resi, oneletter, '
', str(x), str(y), str(z)))")
    pymol.creating.symexp('sym', structure, '(' + structure + ')', 10)
    contactprofileA = {}
    contactprofileB = {}
    contactRes = []
    for i in range(len(pymol.stored.profile)):
        info = pymol.stored.profile[i]
        pymol.stored.coordx = float(info[5])
        pymol.stored.coordy = float(info[6])
        pymol.stored.coordz = float(info[7])
        pymol.stored.contact_ini = []
        pymol.stored.contact = []
        pymol.selecting.select(info[0] + ' and Chain ' + info[1] + ' and resi ' + info[2])
        pymol.selecting.select('sele around 4')
        pymol.selecting.select('contact', 'sele and (not (' + info[0] + ' and Chain ' + info[1] + '))')
        pymol.cmd.iterate('contact', 'stored.contact_ini.append((model, chain, resi, oneletter))')
        pymol.stored.contact_ini = list(set(pymol.stored.contact_ini))
        if pymol.stored.contact_ini:
            for j in range(len(pymol.stored.contact_ini)):
                residue = pymol.stored.contact_ini[j]
                selected = pymol.selecting.select(residue[0] + ' and Chain ' + residue[1] + ' and resi ' +
residue[2] + ' and name CA')
                if selected:
                    pymol.cmd.iterate_state(1, 'sele', 'stored.contact.append((model, chain, resi, oneletter,'
                                            ' str(((pymol.stored.coordx - x)**2 + (pymol.stored.coordy - y)**2 + '
                                            '(pymol.stored.coordz - z)**2)**0.5), str(x), str(y), str(z)))')
            if pymol.stored.contact:

                output.write('*,' + ','.join(info) + '\n')
                contactRes.append(info)
                otherside = []
                for j in range(len(pymol.stored.contact)):
                    otherside.append((pymol.stored.contact[j][0],pymol.stored.contact[j][1]))
                    output.write(' ,' + ','.join(pymol.stored.contact[j]) + '\n')
                counts = Counter(otherside).values()
                interface = ' '.join(Counter(otherside).keys()[counts.index(max(counts))]) + ',' + '
'.join(info[:2])
```

```
                if contactprofileB.has_key(interface):
                    contactprofileB[interface].append(info[2])
                else:
                    contactprofileB[interface] = []
                    contactprofileB[interface].append(info[2])
                for j in range(len(pymol.stored.contact)):
                    if ' '.join((pymol.stored.contact[j][0],pymol.stored.contact[j][1])) + ',' + ' '.join(info[:2])
== interface:
                        if contactprofileA.has_key(interface):
                            if not pymol.stored.contact[j][2] in contactprofileA[interface]:
                                contactprofileA[interface].append(pymol.stored.contact[j][2])
                            else:
                                pass
                        else:
                            contactprofileA[interface] = []
                            contactprofileA[interface].append(pymol.stored.contact[j][2])

    output2.write(structure + '\n')
    output2.write('Interface - side A,Interface - side B,Residues involved in side A,Residues involved in side B\n')
    for i in range(len(contactprofileB)):
        contactprofileB.values()[i].sort()
        contactprofileA.values()[contactprofileA.keys().index(contactprofileB.keys()[i])].sort()
        output2.write(contactprofileB.keys()[i] + ',' + '
'.join(contactprofileA.values()[contactprofileA.keys().index(contactprofileB.keys()[i])]) + ',' + '
'.join(contactprofileB.values()[i]) + '\n')

    pymol.cmd.delete('all')

output.close()
output2.close()
pymol.cmd.quit()

##########################################################################################################################
##########################################################################################################################




##########################################################################################################################
############################## Renumbering the nanobodies from PDB using the IMGT scheme ################################
##########################################################################################################################

def fetchfasta(PDBlist, outputadd="F:\\research\\supercharge\\sequences\\nanoseqs.txt"):
    outfasta = open(outputadd, 'w')
    pymol.finish_launching(['pymol', '-qc'])
    for structure in PDBlist:
```

```
        print 'Processing ' + structure + '...'
        try:
            pymol.importing.fetch(structure)
        except Exception,e:
            print e
            pymol.cmd.load('F:/research/supercharge/sequences/RCSBPDB/' + structure.lower() + '.cif')
        pymol.stored.chains = []
        pymol.cmd.iterate('name CA','stored.chains.append(chain)')
        chainids = list(set(pymol.stored.chains))
        for i in chainids:
            fasta = pymol.cmd.get_fastastr('chain ' + i)
            prefix = len(structure) + 1
            mendedfasta = '>' + structure + '_' + i + fasta[prefix:]
            outfasta.write(mendedfasta)
        pymol.cmd.delete('all')
    outfasta.close()
    return

def correctResi(PDBlist,chaindict):
    pymol.finish_launching(['pymol', '-qc'])
    residict = {}
    for structure in PDBlist:
        print 'Processing ' + structure + '...'
        try:
            pymol.importing.fetch(structure)
        except Exception,e:
            print e
            pymol.cmd.load('F:/research/supercharge/sequences/RCSBPDB/' + structure.lower() + '.cif')
        pymol.stored.chains = []
        pymol.cmd.iterate('name CA', 'stored.chains.append(chain)')
        chainids = list(set(pymol.stored.chains))
        for i in chainids:
            if chaindict.has_key(structure + '_' + i):
                start = int(chaindict[structure + '_' + i][4])
                seqdict = {}
                pymol.stored.seq = []
                pymol.cmd.iterate('name CA and chain ' + i, 'stored.seq.append(resi)')
                for count in range(len(pymol.stored.seq)):
                    seqdict[pymol.stored.seq[count]] = count + 1 - start
                residict[structure + '_' + i] = copy.copy(seqdict)
        pymol.cmd.delete('all')
    return residict

def findref(seq,ref,goal):
    original_goal = goal
    count = 0
```

```
    while goal > 0:
        if seq[count] != '-':
            goal -= 1
            count += 1
        else:
            count += 1
    return ref[count - 1]

def
contactRegion(PDBlist,input="F:\\research\\supercharge\\sequences\\contactSurface.csv",ref="F:\\research\\supercharg
e\\sequences\\ref.csv",output="F:\\research\\supercharge\\sequences\\CCS.csv"):
    interfacefile = open(input,'rb')
    referencefile = open(ref,'rb')
    result = open(output,'w')

    #parse reference
    chaindict = {}
    for line in referencefile:
        info = line.split('\r\n')[0].split(',')
        if info[0] == 'Name':
            header = copy.copy(info)
        else:
            chaindict[info[0]] = copy.copy(info)

    #prepare residue index dictionary
    residict = correctResi(PDBlist,chaindict)

    #parse interface

    reference = header[7:]
    for line in interfacefile:
        if line.split(',')[1] == '':
            print 'Processing ' + line[:4] + ' ... '
            target = line[:4]
            result.write(line[:-4] + '\n')
            continue
        elif line.split(',')[0] == 'Interface - side A' or line.split(',')[0] == 'INTERFACE - SIDE A':
            result.write(line[:-4] + '\n')
            continue
        else:
            if not target in PDBlist:
                continue
            info = line.split(',')
            sideA = '_'.join(info[0].split(' '))
            sideAresi = info[2].split(' ')
            sideB = '_'.join(info[1].split(' '))
```

```
            sideBresi = info[3].split('\r\n')[0].split(' ')
            if sideA[0:3] == 'sym' or sideA[0:3] == 'SYM':
                sideA = target + '_' + info[0].split(' ')[1]
            if sideB[0:3] == 'sym' or sideB[0:3] == 'SYM':
                sideB = target + '_' + info[1].split(' ')[1]
            if chaindict.has_key(sideA):
                seq = chaindict[sideA][7:]
                resiref = residict[sideA]
                correctAresi = []
                for goal in sideAresi:
                    try:
                        pos = findref(seq, reference, resiref[goal])
                        correctAresi.append(pos)
                    except IndexError,e:
                        print e
                        print 'resi ' + str(resiref[goal]) + ' is out of range.'
            else:
                correctAresi = copy.copy(sideAresi)
            if chaindict.has_key(sideB):
                seq = chaindict[sideB][7:]
                resiref = residict[sideB]
                correctBresi = []
                for goal in sideBresi:
                    try:
                        pos = findref(seq, reference, resiref[goal])
                        correctBresi.append(pos)
                    except IndexError,e:
                        print e
                        print 'resi ' + str(resiref[goal]) + ' is out of range.'
            else:
                correctBresi = copy.copy(sideBresi)

            outseq = info[0] + ',' + info[1] + ',' + ' '.join(correctAresi) + ',' + ' '.join(correctBresi) + '\n'
            result.write(outseq)

    result.close()
    interfacefile.close()
    referencefile.close()
    return


contactRegion(PDBlist)

##########################################################################################################################
##########################################################################################################################
```

```
#########################################################################################################################
################################## Clustering the nanobody interfaces based on density ##################################
#########################################################################################################################


infileadd = "F:\\research\\supercharge\\sequences\\CCS.csv"
refadd="F:\\research\\supercharge\\sequences\\ref.csv"
NanoCoordadd = 'F:\\research\\supercharge\\sequences\\nanobodyCoords.csv'
try:
    os.mkdir('F:\\research\\supercharge\\sequences\\NewCommonPatterns')
    os.mkdir('F:\\research\\supercharge\\sequences\\NewCommonPatterns\\pictures')
except WindowsError:
    print 'Directory::: F:\\research\\supercharge\\sequences\\NewCommonPatterns\\pictures::: already exists!'
outpngadd = "F:/research/supercharge/sequences/NewCommonPatterns/pictures/"

usingResiCoor = 1
symthres = -1
epsthres = 3
sizethres = 8
listlenthres = 5
vector_dim = 3
if usingResiCoor == 0:
    outfileadd = "F:\\research\\supercharge\\sequences\\NewCommonPatterns\\ClusterResultIndex.csv"
    logadd = "F:\\research\\supercharge\\sequences\\NewCommonPatterns\\ClusterResultIndexLog.csv"
else:
    outfileadd = "F:\\research\\supercharge\\sequences\\NewCommonPatterns\\ClusterResultCoor.csv"
    logadd = "F:\\research\\supercharge\\sequences\\NewCommonPatterns\\ClusterResultCoorLog.csv"

def sortList(list):
    newlist = []
    for i in list:
        try:
            newlist.append(int(i))
        except ValueError:
            newlist.append(int(re.findall(r'\d+',i)[0]))
    newlist.sort()
    newstrlist = []
    for i in newlist:
        newstrlist.append(str(i))
    return newstrlist

def findoripos(header, chainseq, pos):
    start = int(chainseq[4])
    for i in range(len(header[7:])):
        if header[7 + i] == pos:
            index = i
```

```
            break
        else:
            continue
    blank = 0
    for j in range(index + 1):
        if chainseq[7 + j] == '-':
            blank += 1
            continue
    ori = index - blank + 1
    return ori

def restoreResi(PDBlist,chaindict):
    pymol.finish_launching(['pymol', '-qc'])
    residict = {}
    for structure in PDBlist:
        print 'Processing ' + structure + '...'
        try:
            pymol.importing.fetch(structure)
        except Exception,e:
            print e
            pymol.cmd.load('F:/research/supercharge/sequences/RCSBPDB/' + structure.lower() + '.cif')
        pymol.stored.chains = []
        pymol.cmd.iterate('name CA', 'stored.chains.append(chain)')
        chainids = list(set(pymol.stored.chains))
        for i in chainids:
            if chaindict.has_key(structure + '_' + i):
                start = int(chaindict[structure + '_' + i][4])
                seqdict = {}
                pymol.stored.seq = []
                pymol.cmd.iterate('name CA and chain ' + i, 'stored.seq.append(resi)')
                for count in range(len(pymol.stored.seq)):
                    seqdict[count + 1 - start] = pymol.stored.seq[count]
                residict[structure + '_' + i] = copy.copy(seqdict)
        pymol.cmd.delete('all')
    return residict

def parseCoord(nanoCoordadd):
    infile = open(nanoCoordadd,'rb')
    coords ={}
    for line in infile:
        info = line.split('\r\n')[0].split(',')
        coords[info[2]] = [float(info[4]),float(info[5]),float(info[6])]
    return coords

def scoreInter(listA, listB, noRef):
    noA = listA.split(' ')
```

```
    noB = listB.split(' ')
    samelist = []
    wholelist = []
    same = 0
    sameR = 0
    for i in noA:
        wholelist.append(i)
        if i in noB:
            same += 1
            samelist.append(i)
    for j in noB:
        if not j in wholelist:
            wholelist.append(j)
    for k in wholelist:
        if k in noRef:
            sameR += 1
    scoreA = float(same)/float(len(noA))
    scoreB = float(same)/float(len(noB))
    scoreC = float(sameR)/float(len(noRef))
    return scoreA, scoreB, scoreC, samelist, wholelist

def listtoArray(newlist):
    A = numpy.zeros(len(newlist))
    for i in range(len(newlist)):
        try:
            A[i] = int(newlist[i])
        except ValueError,e:
            if newlist[i] == '33A':
                A[i] = 33.5
            elif newlist[i] == '60A':
                A[i] = 60.33
            elif newlist[i] == '61A':
                A[i] = 60.66
            elif newlist[i] == '82A':
                A[i] = 82.2
            elif newlist[i] == '82B':
                A[i] = 82.4
            elif newlist[i] == '82C':
                A[i] = 82.6
            elif newlist[i] == '82D':
                A[i] = 82.8
            elif newlist[i] == '111A':
                A[i] = 111.0625
            elif newlist[i] == '111B':
                A[i] = 111.125
            elif newlist[i] == '111C':
```

```
                A[i] = 111.1872
            elif newlist[i] == '111D':
                A[i] = 111.25
            elif newlist[i] == '111E':
                A[i] = 111.3125
            elif newlist[i] == '111F':
                A[i] = 111.375
            elif newlist[i] == '111G':
                A[i] = 111.4375
            elif newlist[i] == '112H':
                A[i] = 111.5
            elif newlist[i] == '112G':
                A[i] = 111.5625
            elif newlist[i] == '112F':
                A[i] = 111.625
            elif newlist[i] == '112E':
                A[i] = 111.6875
            elif newlist[i] == '112D':
                A[i] = 111.75
            elif newlist[i] == '112C':
                A[i] = 111.8125
            elif newlist[i] == '112B':
                A[i] = 111.875
            else:
                A[i] = 111.9375
    return A

def listtoArrayCoor(samelist, nanoCoordsadd):
    nanoCoords = parseCoord(nanoCoordsadd)
    A = numpy.zeros(len(samelist))
    B = numpy.zeros(len(samelist))
    C = numpy.zeros(len(samelist))
    for i in range(len(samelist)):
        A[i] = nanoCoords[samelist[i]][0]
        B[i] = nanoCoords[samelist[i]][1]
        C[i] = nanoCoords[samelist[i]][2]
    return A,B,C

def getVector(pattern, dim):
    vector = numpy.zeros(dim)
    miu = float(sum(pattern))/float(len(pattern))
    vector[0] = float(len(pattern))/3.0
    vector[1] = miu
    vector[2] = numpy.median(pattern) - numpy.median(pattern)
    #vector[2] = numpy.median(pattern)
    for i in range(dim-3):
```

```
        vector[i+3] = sum((pattern - miu)**(i+4))/(((float(len(pattern)))**(2*i+4)))
    return vector

def getCluster(vectorlist, threshold, sizethres):
    distances = []
    patternNos = []
    X = numpy.vstack(tuple(vectorlist))
    db = DBSCAN(eps=threshold, min_samples=sizethres).fit(X)
    labels = db.labels_
    n_clusters_ = len(set(labels)) - (1 if -1 in labels else 0)
    for i in range(n_clusters_):
        ave = sum(X[labels == i])/len(X[labels == i])
        Y = list(sum(numpy.transpose((X - ave)**2)))
        patternNos.append(Y.index(min(Y)))
        distances.append(Y)
    return patternNos, list(labels), n_clusters_, distances

interfile = open(infileadd,'rb')
referencefile = open(refadd,'rb')
resultfile = open(outfileadd,'w')
logfile = open(logadd,'w')
namelist = []
vectorlist = []
reflist = []
seq1 = []
seq2 = []
face1 = []
face2 = []
#parse reference
chaindict = {}
for line in referencefile:
    info = line.split('\r\n')[0].split(',')
    if info[0] == 'Name':
        header = copy.copy(info)
    else:
        chaindict[info[0]] = copy.copy(info)

#prepare residue index dictionary and define restore index function
backresidict = restoreResi(PDBlist,chaindict)
def restoreIndex(structure_chainid, seq, chaindict,header,Residict=backresidict):
    orilist = []
    seqlist = seq.split(' ')
    for ind in seqlist:
        ori = findoripos(header, chaindict[structure_chainid],ind)
        orilist.append(Residict[structure_chainid][ori])
    return orilist
```

```
#parse interface file
totaltarget = []
hasnanobody = []
hasnanonanointer = []
for line in interfile:
    info = line.split('\r\n')[0].split(',')
    if info[1] == '':
        target = info[0]
        if target == '5.00E+03':
            target = '5E03'
        if not target in totaltarget:
            totaltarget.append(target)
        print 'Processing ' + info[0] + '...'
        count = 0
        continue
    elif info[0] == 'Interface - side A' or info[0] == 'INTERFACE - SIDE A':
        continue
    else:
        if not target in PDBlist:
            continue
        sideA = '_'.join(info[0].split(' '))
        sideAresi = info[2]
        sideB = '_'.join(info[1].split(' '))
        sideBresi = info[3]
        if sideA[0:3] == 'sym' or sideA[0:3] == 'SYM':
            sideA = target + '_' + info[0].split(' ')[1]
        if sideB[0:3] == 'sym' or sideB[0:3] == 'SYM':
            sideB = target + '_' + info[1].split(' ')[1]
        if chaindict.has_key(sideA) or chaindict.has_key(sideB):
            if not target in hasnanobody:
                hasnanobody.append(target)
        if chaindict.has_key(sideA) and chaindict.has_key(sideB):
            if not target in hasnanonanointer:
                hasnanonanointer.append(target)
            scoreA, scoreB, scoreC, samelist, wholelist = scoreInter(sideAresi, sideBresi, ['1'])
            if float((scoreA + scoreB))/2.0 > symthres and len(samelist) > listlenthres:
                count += 1
                namelist.append(target + '_' + str(count))
                if usingResiCoor == 0:
                    vectorlist.append(getVector(listtoArray(samelist),vector_dim))
                if usingResiCoor == 1:
                    A, B, C = listtoArrayCoor(samelist,NanoCoordadd)

vectorlist.append(numpy.vstack((getVector(A,vector_dim),getVector(B,vector_dim),getVector(C,vector_dim))).flatten())
                reflist.append(samelist)
```

```
                seq1.append(sideAresi)
                seq2.append(sideBresi)
                face1.append(info[0])
                face2.append(info[1])
print 'Total Target: ' + str(len(totaltarget))
print 'Has nanobody: ' + str(len(hasnanobody))
print 'Has nanobody-nanobody interface: ' + str(len(hasnanonanointer))
print 'Item not in PDBlist:'
for item in totaltarget:
    if not item in PDBlist:
        print item
print 'Item not in interface file but in PDBlist:'
for item in PDBlist:
    if not item in totaltarget:
        print item



patternNos, labels, n_clusters_,distances = getCluster(vectorlist, epsthres, sizethres)
commonInter = {}
patternlist = []
totallist = []
for i in range(n_clusters_):
    pattern = ' '.join(reflist[patternNos[i]])
    commonInter[':'.join([namelist[patternNos[i]],face1[patternNos[i]],face2[patternNos[i]],pattern,str(i)])] = []
    patternlist.append(':'.join([namelist[patternNos[i]],face1[patternNos[i]],face2[patternNos[i]],pattern,str(i)]))

logfile.write('interface name, interface pattern, reference interface, class, distance, X, Y, Z \n')
StructureRec = []
DisRec = []
InterFaceRec = []
InterResiRec = []
RefRec = []

for i in range(len(labels)):
    if labels[i] == -1:
        logfile.write(namelist[i] + ',' + ' '.join(reflist[i]) + ',none,' + str(labels[i]) + ',none,' +
str(vectorlist[i][1]) + ',' + str(vectorlist[i][4]) + ',' + str(vectorlist[i][7]) + '\n')
    else:
        logfile.write(namelist[i] + ',' + ' '.join(reflist[i]) + ',' + patternlist[labels[i]] + ',' + str(labels[i])
+ ',' + str(
            distances[labels[i]][i]) + ',' + str(vectorlist[i][1]) + ',' + str(vectorlist[i][4]) + ',' +
str(vectorlist[i][7]) + '\n')
    if not namelist[i].split('_')[0] in totallist:
        totallist.append(namelist[i].split('_')[0])
    if not labels[i] == -1:
```

```
        if not namelist[i].split('_')[0] in commonInter[patternlist[labels[i]]]:
            commonInter[patternlist[labels[i]]].append(namelist[i].split('_')[0])
        if not namelist[i].split('_')[0] + '_' + str(labels[i]) in StructureRec:
            StructureRec.append(namelist[i].split('_')[0] + '_' + str(labels[i]))
            InterFaceRec.append((face1[i],face2[i]))
            InterResiRec.append((restoreIndex(namelist[i].split('_')[0] + '_' + face1[i].split('
')[1],seq1[i],chaindict,header),
                                 restoreIndex(namelist[i].split('_')[0] + '_' + face2[i].split('
')[1],seq2[i],chaindict,header)))
            RefRec.append(patternlist[labels[i]])
            DisRec.append(distances[labels[i]][i])
        if (namelist[i].split('_')[0] + '_' + str(labels[i]) in StructureRec) and distances[labels[i]][i] <
DisRec[StructureRec.index(namelist[i].split('_')[0] + '_' + str(labels[i]))]:
            newind = StructureRec.index(namelist[i].split('_')[0] + '_' + str(labels[i]))
            InterFaceRec[newind] = (face1[i], face2[i])
            InterResiRec[newind] = (restoreIndex(namelist[i].split('_')[0] + '_' + face1[i].split('
')[1],seq1[i],chaindict,header),
                                 restoreIndex(namelist[i].split('_')[0] + '_' + face2[i].split('
')[1],seq2[i],chaindict,header))
            DisRec[newind] = distances[labels[i]][i]

resultfile.write('category, pattern name, pattern, pattern length, cases,case total,' + ','.join(header[7:]) + '\n')
for common in commonInter.keys():
    structures = ' '.join(commonInter[common])
    resultfile.write(common.split(':')[4] + ',' + common.split(':')[0] + ',' + common.split(':')[3] + ',' +
str(len(common.split(':')[3].split(' '))) + ',' + structures + ',' + str(len(commonInter[common])))
    commonList = common.split(':')[3].split(' ')
    for index in header[7:]:
        if index in commonList:
            resultfile.write(',1')
        else:
            resultfile.write(',0')
    resultfile.write('\n')
resultfile.write(',,,cases having symmetric nanobody-nanobody interfaces that share >5 interactive residues:,' + '
'.join(totallist) + ',' + str(len(totallist)) + '\n')
resultfile.write(',,,cases having nanobody-nanobody interfaces:,' + ' '.join(hasnanonanointer) + ',' +
str(len(hasnanonanointer)) + '\n')
resultfile.write(',,,cases having nanobody:,' + ' '.join(hasnanobody) + ',' + str(len(hasnanobody)) + '\n')
resultfile.write(',,,total cases in interface file:,' + ' '.join(totaltarget) + ',' + str(len(totaltarget)) + '\n')
logfile.close()
interfile.close()
referencefile.close()
resultfile.close()

# generate PSE
pymol.finish_launching(['pymol', '-qc'])
```

```
for i in range(len(StructureRec)):
    print 'Drawing ' + StructureRec[i] + '...'
    structure = StructureRec[i].split('_')[0]
    patternresi1 = '(resi ' + ' or resi '.join(InterResiRec[i][0]) + ')'
    patternresi2 = '(resi ' + ' or resi '.join(InterResiRec[i][1]) + ')'
    refseq = restoreIndex(structure + '_' + InterFaceRec[i][0].split(' ')[1],
RefRec[i].split(':')[3],chaindict,header)
    newrefseq = sortList(refseq)
    zoomregion = newrefseq[0] + '-' + newrefseq[-1]
    highlight1 = newrefseq[int(len(newrefseq) / 2)]
    highlight2 = '(resi ' + ' or resi '.join(refseq) + ')'
    pymol.importing.fetch(structure)
    pymol.cmd.set('cartoon_side_chain_helper','1')
    pymol.stored.seq = []
    pymol.creating.symexp('sym',structure,'(' + structure + ')', '10')
    pymol.cmd.color('grey', 'all')
    pymol.cmd.color('cyan', InterFaceRec[i][0].split(' ')[0] + ' and Chain ' + InterFaceRec[i][0].split(' ')[1])
    pymol.cmd.color('yellow', InterFaceRec[i][1].split(' ')[0] + ' and Chain ' + InterFaceRec[i][1].split(' ')[1])
    #pymol.cmd.hide('cartoon', InterFaceRec[i][0].split(' ')[0] + ' and Chain ' + InterFaceRec[i][0].split(' ')[1] +
' and resi ' + zoomregion)
    #pymol.cmd.hide('cartoon', InterFaceRec[i][1].split(' ')[0] + ' and Chain ' + InterFaceRec[i][1].split(' ')[1] +
' and resi ' + zoomregion)
    pymol.cmd.show('lines', InterFaceRec[i][0].split(' ')[0] + ' and Chain ' + InterFaceRec[i][0].split(' ')[1] + '
and ' + highlight2)
    pymol.cmd.show('lines', InterFaceRec[i][1].split(' ')[0] + ' and Chain ' + InterFaceRec[i][1].split(' ')[1] + '
and ' + highlight2)
    pymol.cmd.show('sticks', InterFaceRec[i][0].split(' ')[0] + ' and Chain ' + InterFaceRec[i][0].split(' ')[1] + '
and ' +  patternresi1)
    pymol.cmd.show('sticks', InterFaceRec[i][1].split(' ')[0] + ' and Chain ' + InterFaceRec[i][1].split(' ')[1] + '
and ' +  patternresi2)
    pymol.cmd.color('pink', InterFaceRec[i][1].split(' ')[0] + ' and Chain ' + InterFaceRec[i][1].split(' ')[1] + '
and ' + highlight2)
    pymol.cmd.color('pink', InterFaceRec[i][0].split(' ')[0] + ' and Chain ' + InterFaceRec[i][0].split(' ')[1] + '
and ' + highlight2)
    pymol.cmd.color('red', InterFaceRec[i][1].split(' ')[0] + ' and Chain ' + InterFaceRec[i][1].split(' ')[1] + '
and resi ' + highlight1)
    pymol.cmd.color('red', InterFaceRec[i][0].split(' ')[0] + ' and Chain ' +InterFaceRec[i][0].split(' ')[1] + '
and resi ' + highlight1)
    pymol.cmd.color('atomic', '(not elem C)')
    pymol.cmd.zoom('(' + InterFaceRec[i][0].split(' ')[0] + ' and Chain ' + InterFaceRec[i][0].split(' ')[1] + ' and
resi ' + zoomregion +') or (' + InterFaceRec[i][1].split(' ')[0] + ' and Chain ' + InterFaceRec[i][1].split(' ')[1]
+ ' and resi ' + zoomregion + ')')
    pymol.cmd.save(outpngadd + StructureRec[i] + '.pse')
    pymol.cmd.delete('all')
```

```
pymol.cmd.quit()


################################################################################################################################
################################################################################################################################
```

8.2 Original sequences of the nanobodies in section 4.2.2.1

Sample ID	Scaffold Type	Nanobody name	Nanobody sequence
1	original scaffold	DYG02	QVQLQESGGGLVQAGGSLRLSCAASGNIFNINAMGWYRQVPGKQREMVAAITSGGSTNYAYSVKGRFTISRDNAKNTVY LQMNSLNPEDTAIYYCEAYGTYTLAPTGEGEYDDYWGQGTQVTVS
2	original scaffold	DYG01	QVQLQESGGGLVQAGGSLRLSCAASGFTFDDYAIGWFRQAPGKEREGVSSIKSSDGSTYYAYSVKGRFTISSDNAKNTVYL QMNSLKPEDTAVYYCAATETRPTMDPLDFGYWDQGTQVTVS
3	original scaffold	DYF12	QVQLQESGGGLVQAGGSLRLSCAASGSIFSINRMGWYRQAPGKQRELVAAITSGGSTNYAYSVKGRFTISRDNAKNTVYL QMNSLKPEDTAIYYCEAYGTYTLAPTGEGEYDDYWGQGTQVTVS
4	original scaffold	DYF11	QVQLQESGGGLVQAGGSLRLSCKASISLFNRFDVAWYRQAPGKQRDMVARITSGGSVEYSYDVKGRFTIARDNAKNMVL LQMNSLKLEDTAVYYCNLKRRWSWGSSNEFWGQGTQVTVS
5	original scaffold	DYF07	QVQLQESGGGLVQPGGSLRLSCAASGFTFSTYAMSWFRQAPGKGLEWVASINSSGGGIQSYAYSVKGRFTISRDNAKNTL YLQMNSLEPEDTAVYYCAAAAWRVGTYDYRGQGTQVTVS
6	original scaffold	DYF06	QVQLQESGGGLVQAGGSLRLSCAASGFTFDDYAIGWFRQVPGKEREGVSCTSSSVGSTYYAYSVKGRFTITSDNAKNTVYL QMNSLKPEDTAVYYCAADSYPYGSTWCRPPFFASWGQGTQVTVS
7	original scaffold	DYE03	QVQLQESGGGLVQAGGSLRLSCAASGRTFSRYVMAWFRQAPGKEREFVARITSTGGTIDHAYSVKGRFTISRDNAKNTVY LQMNSLKPEDTAVYYCAAARWVSVVRVNTVVDYEYWGQGTQVTVS
8	original scaffold	DYE02	QVQLQESGGGLVQPGGSLRLSCAASGFTFGIYDMSWVRQAPGKGPEWVSTINDDGGSKSYAYSVKGRFTISRDNANNTL YLQMNSLKPEDTAVYYCAKSYYAGSFYYSAYDYWGQGTQVTVS
9	original scaffold	DYE01	QVQLQESGGGLVQAGGSLRLSCAASGIWFSINNMAWFRQAPGKQRERIATITSGGTANLVYSVKGRFTISRDNAENTMYL QMNSLIPEDTAVYYCNLIADYDMGFQSFWGRGTQVTVS
10	original scaffold	DYD12	QVQLQESGGGLVQAGGSLRLSCAASGIWFSINNMAWYRQTPGKQRERIAIITSAGTTNYVYSVKGRFTISRDDAKNTMYL QMNSLIPEDTAVYYCNLVADYDMGFQSFWGRGTQVTVS
11	original scaffold	DYF02	QVQLQESGGGLVQAGGSLRLSCTASGSIRRPNAMGWYRQAPGPQRELVAVITSGGSTTYAYSEKGRFTISRDSAENTVYL QMNSLKSEDTAVYYCRLLQYGQIYWGQGTQVTVS
12	original scaffold	IntG02	QVQLQENGGGCVQAGGSLRLSCAASGNIFNINAMGWYRQVPGKQREMVAAITSGGSTNYADSVKGRFTISRDNAKNTV YLQMNSLNPEDTAIYYCEAYGTYTLAPTGEGEYDDYWGQGTQVTVS
13	original scaffold	IntG01	QVQLQENGGGCVQAGGSLRLSCAASGFTFDDYAIGWFRQAPGKEREGVSSIKSSDGSTYYADSVKGRFTISSDNAKNTVYL QMNSLKPEDTAVYYCAATETRPTMDPLDFGYWDQGTQVTVS
14	original scaffold	IntF12	QVQLQENGGGCVQAGGSLRLSCAASGSIFSINRMGWYRQAPGKQRELVAAITSGGSTNYADSVKGRFTISRDNAKNTVYL QMNSLKPEDTAIYYCEAYGTYTLAPTGEGEYDDYWGQGTQVTVS
15	original scaffold	IntF11	QVQLQENGGGCVQAGGSLRLSCKASISLFNRFDVAWYRQAPGKQRDMVARITSGGSVEYSDDVKGRFTIARDNAKNMV LLQMNSLKLEDTAVYYCNLKRRWSWGSSNEFWGQGTQVTVS
16	original scaffold	IntF07	QVQLQENGGGCVQPGGSLRLSCAASGFTFSTYAMSWFRQAPGKGLEWVASINSSGGGIQSYADSVKGRFTISRDNAKNT LYLQMNSLEPEDTAVYYCAAAAWRVGTYDYRGQGTQVTVS

Sample ID	Scaffold Type	Nanobody name	Nanobody sequence
17	original scaffold	IntF06	QVQLQENGGGCVQAGGSLRLSCAASGFTFDDYAIGWFRQVPGKEREGVSCTSSSVGSTYYADSVKGRFTITSDNAKNTVY LQMNSLKPEDTAVYYCAADSYPYGSTWCRPPFFASWGQGTQVTVTS
18	original scaffold	IntE03	QVQLQENGGGCVQAGGSLRLSCAASGRTFSRYVMAWFRQAPGKEREFVARITSTGGTIDHADSVKGRFTISRDNAKNTV YLQMNSLKPEDTAVYYCAAARWVSVVRVNTVVDYEYWGQGTQVTVTS
19	original scaffold	IntE02	QVQLQENGGGCVQPGGSLRLSCAASGFTFGIYDMSWVRQAPGKGPEWVSTINDDGGSKSYADSVKGRFTISRDNANNT LYLQMNSLKPEDTAVYYCAKSYYAGSFYYSAYDYWGQGTQVTVTS
20	original scaffold	IntE01	QVQLQENGGGCVQAGGSLRLSCAASGIWFSINNMAWFRQAPGKQRERIATITSGGTANLVDSVKGRFTISRDNAENTMY LQMNSLIPEDTAVYYCNLIADYDMGFQSFWGRGTQVTVTS
21	original scaffold	IntD12	QVQLQENGGGCVQAGGSLRLSCAASGIWFSINNMAWYRQTPGKQRERIAIITSAGTTNYVDSVKGRFTISRDDAKNTMY LQMNSLIPEDTAVYYCNLVADYDMGFQSFWGRGTQVTVTS
22	original scaffold	IntF02	QVQLQENGGGCVQAGGSLRLSCTASGSIRRPNAMGWYRQAPGPQRELVAVITSGGSTTYADSEKGRFTISRDSAENTVYL QMNSLKSEDTAVYYCRLLQYGQIYWGQGTQVTVTS
23	original scaffold	bothG02	QVQLQENGGGCVQAGGSLRLSCAASGNIFNINAMGWYRQVPGKQREMVAAITSGGSTNYAYSVKGRFTISRDNAKNTV YLQMNSLNPEDTAIYYCEAYGTYTLAPTGEGEYDDYWGQGTQVTVTS
24	original scaffold	bothG01	QVQLQENGGGCVQAGGSLRLSCAASGFTFDDYAIGWFRQAPGKEREGVSSIKSSDGSTYYAYSVKGRFTISSDNAKNTVYL QMNSLKPEDTAVYYCAATETRPTMDPLDFGYWDQGTQVTVTS
25	original scaffold	bothF12	QVQLQENGGGCVQAGGSLRLSCAASGSIFSINRMGWYRQAPGKQRELVAAITSGGSTNYAYSVKGRFTISRDNAKNTVYL QMNSLKPEDTAIYYCEAYGTYTLAPTGEGEYDDYWGQGTQVTVTS
26	original scaffold	bothF11	QVQLQENGGGCVQAGGSLRLSCKASISLFNRFDVAWYRQAPGKQRDMVARITSGGSVEYSYDVKGRFTIARDNAKNMV LLQMNSLKLEDTAVYYCNLKRRWSWGSSNEFWGQGTQVTVTS
27	original scaffold	bothF07	QVQLQENGGGCVQPGGSLRLSCAASGFTFSTYAMSWFRQAPGKGLEWVASINSSGGGIQSYAYSVKGRFTISRDNAKNT LYLQMNSLEPEDTAVYYCAAAAWRVGTYDYRGQGTQVTVTS
28	original scaffold	bothF06	QVQLQENGGGCVQAGGSLRLSCAASGFTFDDYAIGWFRQVPGKEREGVSCTSSSVGSTYYAYSVKGRFTITSDNAKNTVY LQMNSLKPEDTAVYYCAADSYPYGSTWCRPPFFASWGQGTQVTVTS
29	original scaffold	bothE03	QVQLQENGGGCVQAGGSLRLSCAASGRTFSRYVMAWFRQAPGKEREFVARITSTGGTIDHAYSVKGRFTISRDNAKNTVY LQMNSLKPEDTAVYYCAAARWVSVVRVNTVVDYEYWGQGTQVTVTS
30	original scaffold	bothE02	QVQLQENGGGCVQPGGSLRLSCAASGFTFGIYDMSWVRQAPGKGPEWVSTINDDGGSKSYAYSVKGRFTISRDNANNT LYLQMNSLKPEDTAVYYCAKSYYAGSFYYSAYDYWGQGTQVTVTS
31	original scaffold	bothE01	QVQLQENGGGCVQAGGSLRLSCAASGIWFSINNMAWFRQAPGKQRERIATITSGGTANLVYSVKGRFTISRDNAENTMY LQMNSLIPEDTAVYYCNLIADYDMGFQSFWGRGTQVTVTS
32	original scaffold	bothD12	QVQLQENGGGCVQAGGSLRLSCAASGIWFSINNMAWYRQTPGKQRERIAIITSAGTTNYVYSVKGRFTISRDDAKNTMYL QMNSLIPEDTAVYYCNLVADYDMGFQSFWGRGTQVTVTS
33	original scaffold	bothF02	QVQLQENGGGCVQAGGSLRLSCTASGSIRRPNAMGWYRQAPGPQRELVAVITSGGSTTYAYSEKGRFTISRDSAENTVYL QMNSLKSEDTAVYYCRLLQYGQIYWGQGTQVTVTS
56	3K1K scaffold	DYG02	QVQLVESGGGLVQAGGSLRLSCAASGNIFNINAMTWYRQAPGKEREWVAAITSGGSTNYAYSVKGRFTISRDNAKNTVYL QMNSLKPEDTAVYYCEAYGTYTLAPTGEGEYDDYWGQGTQVTVTS
57	3K1K scaffold	DYG01	QVQLVESGGGLVQAGGSLRLSCAASGFTFDDYAMTWYRQAPGKEREWVASIKSSDGSTYYAYSVKGRFTISRDNAKNTVY LQMNSLKPEDTAVYYCAATETRPTMDPLDFGYWGQGTQVTVTS
58	3K1K scaffold	DYF12	QVQLVESGGGLVQAGGSLRLSCAASGSIFSINRMTWYRQAPGKEREWVAAITSGGSTNYAYSVKGRFTISRDNAKNTVYL QMNSLKPEDTAVYYCEAYGTYTLAPTGEGEYDDYWGQGTQVTVTS
59	3K1K scaffold	DYF11	QVQLVESGGGLVQAGGSLRLSCAASISLFNRFDMTWYRQAPGKEREWVARITSGGSVEYAYSVKGRFTISRDNAKNTVYL QMNSLKPEDTAVYYCNLKRRWSWGSSNEFWGQGTQVTVTS
60	3K1K scaffold	DYF07	QVQLVESGGGLVQAGGSLRLSCAASGFTFSTYAMTWYRQAPGKEREWVASINSSGGGIQSYAYSVKGRFTISRDNAKNTV YLQMNSLKPEDTAVYYCAAAAWRVGTYDYRGQGTQVTVTS

Sample ID	Scaffold Type	Nanobody name	Nanobody sequence
61	3K1K scaffold	DYF06	QVQLVESGGGLVQAGGSLRLSCAASGFTFDDYAMTWYRQAPGKEREWVACTSSSVGSTYYAYSVKGRFTISRDNAKNTV YLQMNSLKPEDTAVYYCAADSYPYGSTWCRPPFFASWGQGTQVTVS
62	3K1K scaffold	DYE03	QVQLVESGGGLVQAGGSLRLSCAASGRTFSRYVMTWYRQAPGKEREWVARITSTGGTIDYAYSVKGRFTISRDNAKNTVY LQMNSLKPEDTAVYYCAAARWVSVVRVNTVVDYEYWGQGTQVTVS
63	3K1K scaffold	DYE02	QVQLVESGGGLVQAGGSLRLSCAASGFTFGIYDMTWYRQAPGKEREWVATINDDGGSKSYAYSVKGRFTISRDNAKNTV YLQMNSLKPEDTAVYYCAKSYYAGSFYYSAYDYWGQGTQVTVS
64	3K1K scaffold	DYE01	QVQLVESGGGLVQAGGSLRLSCAASGIWFSINNMTWYRQAPGKEREWVATITSGGTANYAYSVKGRFTISRDNAKNTVY LQMNSLKPEDTAVYYCNLIADYDMGFQSFWGQGTQVTVS
65	3K1K scaffold	DYD12	QVQLVESGGGLVQAGGSLRLSCAASGIWFSINNMTWYRQAPGKEREWVAIITSAGTTNYAYSVKGRFTISRDNAKNTVYL QMNSLKPEDTAVYYCNLVADYDMGFQSFWGQGTQVTVS
66	3K1K scaffold	DYF02	QVQLVESGGGLVQAGGSLRLSCAASGSIRRPNAMTWYRQAPGKEREWVAVITSGGSTTYAYSVKGRFTISRDNAKNTVYL QMNSLKPEDTAVYYCRLLQYGQIYWGQGTQVTVS
67	3K1K scaffold	IntG02	QVQLVENGGGCVQAGGSLRLSCAASGNIFNINAMTWYRQAPGKEREWVAAITSGGSTNYADSVKGRFTISRDNAKNTVY LQMNSLKPEDTAVYYCEAYGTYTLAPTGEGEYDDYWGQGTQVTVS
68	3K1K scaffold	IntG01	QVQLVENGGGCVQAGGSLRLSCAASGFTFDDYAMTWYRQAPGKEREWVASIKSSDGSTYYADSVKGRFTISRDNAKNTV YLQMNSLKPEDTAVYYCAATETRPTMDPLDFGYWGQGTQVTVS
69	3K1K scaffold	IntF12	QVQLVENGGGCVQAGGSLRLSCAASGSIFSINRMTWYRQAPGKEREWVAAITSGGSTNYADSVKGRFTISRDNAKNTVYL QMNSLKPEDTAVYYCEAYGTYTLAPTGEGEYDDYWGQGTQVTVS
70	3K1K scaffold	IntF11	QVQLVENGGGCVQAGGSLRLSCAASISLFNRFDMTWYRQAPGKEREWVARITSGGSVEYADSVKGRFTISRDNAKNTVY LQMNSLKPEDTAVYYCNLKRRWSWGSSNEFWGQGTQVTVS
71	3K1K scaffold	IntF07	QVQLVENGGGCVQAGGSLRLSCAASGFTFSTYAMTWYRQAPGKEREWVASINSSGGGIQSYADSVKGRFTISRDNAKNT VYLQMNSLKPEDTAVYYCAAAAWRVGTYDYRGQGTQVTVS
72	3K1K scaffold	IntF06	QVQLVENGGGCVQAGGSLRLSCAASGFTFDDYAMTWYRQAPGKEREWVACTSSSVGSTYYADSVKGRFTISRDNAKNT VYLQMNSLKPEDTAVYYCAADSYPYGSTWCRPPFFASWGQGTQVTVS
73	3K1K scaffold	IntE03	QVQLVENGGGCVQAGGSLRLSCAASGRTFSRYVMTWYRQAPGKEREWVARITSTGGTIDYADSVKGRFTISRDNAKNTV YLQMNSLKPEDTAVYYCAAARWVSVVRVNTVVDYEYWGQGTQVTVS
74	3K1K scaffold	IntE02	QVQLVENGGGCVQAGGSLRLSCAASGFTFGIYDMTWYRQAPGKEREWVATINDDGGSKSYADSVKGRFTISRDNAKNT VYLQMNSLKPEDTAVYYCAKSYYAGSFYYSAYDYWGQGTQVTVS
75	3K1K scaffold	IntE01	QVQLVENGGGCVQAGGSLRLSCAASGIWFSINNMTWYRQAPGKEREWVATITSGGTANYADSVKGRFTISRDNAKNTV YLQMNSLKPEDTAVYYCNLIADYDMGFQSFWGQGTQVTVS
76	3K1K scaffold	IntD12	QVQLVENGGGCVQAGGSLRLSCAASGIWFSINNMTWYRQAPGKEREWVAIITSAGTTNYADSVKGRFTISRDNAKNTVY LQMNSLKPEDTAVYYCNLVADYDMGFQSFWGQGTQVTVS
77	3K1K scaffold	IntF02	QVQLVENGGGCVQAGGSLRLSCAASGSIRRPNAMTWYRQAPGKEREWVAVITSGGSTTYADSVKGRFTISRDNAKNTVY LQMNSLKPEDTAVYYCRLLQYGQIYWGQGTQVTVS
78	3K1K scaffold	bothG02	QVQLVENGGGCVQAGGSLRLSCAASGNIFNINAMTWYRQAPGKEREWVAAITSGGSTNYAYSVKGRFTISRDNAKNTVY LQMNSLKPEDTAVYYCEAYGTYTLAPTGEGEYDDYWGQGTQVTVS
79	3K1K scaffold	bothG01	QVQLVENGGGCVQAGGSLRLSCAASGFTFDDYAMTWYRQAPGKEREWVASIKSSDGSTYYAYSVKGRFTISRDNAKNTV YLQMNSLKPEDTAVYYCAATETRPTMDPLDFGYWGQGTQVTVS
80	3K1K scaffold	bothF12	QVQLVENGGGCVQAGGSLRLSCAASGSIFSINRMTWYRQAPGKEREWVAAITSGGSTNYAYSVKGRFTISRDNAKNTVYL QMNSLKPEDTAVYYCEAYGTYTLAPTGEGEYDDYWGQGTQVTVS
81	3K1K scaffold	bothF11	QVQLVENGGGCVQAGGSLRLSCAASISLFNRFDMTWYRQAPGKEREWVARITSGGSVEYAYSVKGRFTISRDNAKNTVYL QMNSLKPEDTAVYYCNLKRRWSWGSSNEFWGQGTQVTVS
82	3K1K scaffold	bothF07	QVQLVENGGGCVQAGGSLRLSCAASGFTFSTYAMTWYRQAPGKEREWVASINSSGGGIQSYAYSVKGRFTISRDNAKNT VYLQMNSLKPEDTAVYYCAAAAWRVGTYDYRGQGTQVTVS

Sample ID	Scaffold Type	Nanobody name	Nanobody sequence
83	3K1K scaffold	bothF06	QVQLVENGGGCVQAGGSLRLSCAASGFTFDDYAMTWYRQAPGKEREWVACTSSSVGSTYYAYSVKGRFTISRDNAKNTV YLQMNSLKPEDTAVYYCAADSYPYGSTWCRPPFFASWGQGTQVTVS
84	3K1K scaffold	bothE03	QVQLVENGGGCVQAGGSLRLSCAASGRTFSRYVMTWYRQAPGKEREWVARITSTGGTIDYAYSVKGRFTISRDNAKNTV YLQMNSLKPEDTAVYYCAAARWVSVVRVNTVVDYEYWGQGTQVTVS
85	3K1K scaffold	bothE02	QVQLVENGGGCVQAGGSLRLSCAASGFTFGIYDMTWYRQAPGKEREWVATINDDGGSKSYAYSVKGRFTISRDNAKNTV YLQMNSLKPEDTAVYYCAKSYYAGSFYYSAYDYWGQGTQVTVS
86	3K1K scaffold	bothE01	QVQLVENGGGCVQAGGSLRLSCAASGIWFSINNMTWYRQAPGKEREWVATITSGGTANYAYSVKGRFTISRDNAKNTVY LQMNSLKPEDTAVYYCNLIADYDMGFQSFWGQGTQVTVS
87	3K1K scaffold	bothD12	QVQLVENGGGCVQAGGSLRLSCAASGIWFSINNMTWYRQAPGKEREWVAIITSAGTTNYAYSVKGRFTISRDNAKNTVY LQMNSLKPEDTAVYYCNLVADYDMGFQSFWGQGTQVTVS
88	3K1K scaffold	bothF02	QVQLVENGGGCVQAGGSLRLSCAASGSIRRPNAMTWYRQAPGKEREWVAVITSGGSTTYAYSVKGRFTISRDNAKNTVY LQMNSLKPEDTAVYYCRLLQYGQIYWGQGTQVTVS

8.3 Crystal diffraction data (separate file)

8.4 Detailed crystallographic and refinement statistics of RECQL5:Gluebody structures (separate files)

8.4.1 7ZML Crystal structure of human RECQL5 helicase APO form in complex with Gluebody G1-001

8.4.2 7ZMM Crystal structure of human RECQL5 helicase APO form in complex with Gluebody G2-001

8.4.3 7ZMN Crystal structure of human RECQL5 helicase APO form in complex with Gluebody G3-048

8.4.4 7ZMO Crystal structure of human RECQL5 helicase APO form in complex with Gluebody G3-052

8.4.5 7ZMP Crystal structure of human RECQL5 helicase APO form in complex with Gluebody G3-055

8.4.6 7ZMQ Crystal structure of human RECQL5 helicase APO form in complex with Gluebody G2star-006

8.4.7 7ZMR Crystal structure of human RECQL5 helicase APO form in complex with Gluebody G2star-011

8.4.8 7ZMS Crystal structure of human RECQL5 helicase APO form in complex with Gluebody G4-043

8.4.9 7ZMT Crystal structure of human RECQL5 helicase APO form in complex with Gluebody G5-006

8.4.10 7ZMV Crystal structure of human RECQL5 helicase APO form in complex with Gluebody G5-006