CORA Cork Open Research Archive
Cartlann Taighde Oscailte Chorcaí

Title	In silico-aided design, build and test of synthetic proteins
Authors	Yallapragada, V. V. B.
Publication date	2019-12-20
Original Citation	Yallapragada, V. V. B. 2019. In silico-aided design, build and test of synthetic proteins. PhD Thesis, University College Cork.
Type of publication	Doctoral thesis
Rights	© 2019, Venkata Vamsi Bharadwaj Yallapragada. - https://creativecommons.org/licenses/by-nc-nd/4.0/
Download date	2026-11-16 10:07:23
Item downloaded from	https://hdl.handle.net/10468/10076



UCC
University College Cork, Ireland
Coláiste na hOllscoile Corcaigh

Ollscoil na Eireann, Corcaigh

THE NATIONAL UNIVERSITY OF IRELAND, CORK

School of Surgery



In silico-aided design, build and test of synthetic proteins

Thesis presented by,

Venkata Vamsi Bharadwaj Yallapragada B.Tech, M.Sc

under the supervision of

Dr Mark Tangney

For the degree of
Doctor of Philosophy

University College Cork

December 2019

ABSTRACT

Since the discovery of proteins in 1838, the field of protein engineering and our understanding of proteins have improved exponentially. Synthetic proteins have found applications in various biomedical, food and material-based settings. This rise in synthetic proteins was complemented with the parallel expansion in the availability of *in silico* tools for protein modelling. The complexity in the composition and design of synthetic proteins requires careful *in silico* validation to screen for potential pitfalls in the design. *In silico* tools for protein modelling and design have been used extensively to computationally validate the structure and functioning of the synthetic proteins prior to wet-lab testing.

In this thesis, the workflow of design-model-build-test of synthetic proteins with novel applications in imaging is described. The *in silico*-aided design, screening and the *in vitro* testing of synthetic proteins targeting *S. aureus* surface antigen Clumping factor A are discussed in Chapter 2. In this chapter, a suitable candidate worthy of examining in a future *in vivo* setting was identified. During the *in silico*-aided screening, the complexity of data obtained from various *in silico* tools posed new challenges. This was termed as 'the *in silico* myriad problem'. In Chapter 3, a mathematical strategy (Function2Form bridge) was tested to address the *in silico* myriad problem, by combining the scores of different design parameters pertaining to the synthetic protein being analysed into a single easily interpreted output describing overall performance. The strategy comprises 1. A mathematical strategy combining data from a myriad of *in silico* tools into an Overall Performance-score (a singular score informing on a user-defined overall performance); 2. The F2F-Plot, a graphical means of informing the wet-lab biologist holistically on designed construct suitability in the context of multiple parameters, highlighting scope for improvement. F2F bridge was implemented during the design process of all the synthetic proteins in Chapter 4 and Chapter 5.

The synthetic protein design strategy used in Chapter 2 was implemented to design synthetic proteins targeting cancer cells, and to assess their potential as *in vivo* imaging agents in Chapter 4. For both MUC1 and ClfA targeted proteins, *in vivo* luminescence imaging studies involving systemic intravenous administration of proteins, validated synthetic protein specific accumulation at target cell locations within mice as evidenced by localised luminescence. Dose response

studies indicated that luminescence output was both target cell and administered protein quantity related.

In Chapter 5, a self-assembling protein 'cage' was designed, built and tested *in vitro*. An accompanying novel fluorescence-based protein-protein interaction reporting strategy was introduced, involving incorporation of cysteine residues at the interaction interface of monomeric proteins of the self-assembling protein cage. *In silico* tools were used to ensure the conformational and functional stability. FlAsH EDT2 (**fl**uorescin **ar**senical **h**airpin binder-**e**thane**di**thiol) mediated fluorescence was used to confirm the self-assembly. This demonstrates the level of accuracy and detail that can be incorporated into synthetic protein design using *in silico* tools.

In Chapter 6, the scope of introducing miniaturised optical devices to aid biological experimentation was explored. A novel handheld device for monitoring continuous bacterial growth, with prospects of measuring biofluorescence was developed. The device was tested using different bacterial strains and showed accuracy levels similar to a standard benchtop spectrophotometer.

This thesis demonstrates the use of computational methods and various *in silico* tools for protein design. Modern day biomedical science demands novel concepts with deployable technology to assist their translation into user-based settings. In this thesis, various interdisciplinary concepts have been applied to deliver on a holistic end-goal.

Declaration

This is to certify that the work I am submitting is my own and has not been submitted for another degree, either at University College Cork or elsewhere. All external references and sources are clearly acknowledged and identified within the contents. I have read and understood the regulations of University College Cork concerning plagiarism.

Venkata Vamsi Bharadwaj Yallapragada
Author

To,

Ramana
Sailaja
Ooha

Hakuna Matata …

Acknowledgements

First and foremost, I would like to express my sincere gratitude to my supervisor Dr Mark Tangney, for his support and guidance throughout the course of my PhD. His encouragement, patience and foresight for me has improved my scientific thinking over the last 3.142857 years. I would also like to extend my gratitude to everyone at CancerResearch@UCC, who have created a wonderful working environment. A big thanks to Sidney and Ciaran, who have been my go-to people in the lab.

There are many friends who have rubbed a ton of positivity whenever I needed. Chinna and Uday for all the memories down the lane, I will always be thankful. Joe and Liam for all those late-night hangouts and countless 'craic', I owe them a lifetime of pints. Life during PhD wouldn't have been same without them. A special thanks to my close friends Dora and David, who have managed to steal me into non-study-based settings. That was very necessary. I would like to thank Nana, Glenn, Sini, Madhu, Adarsh, Fang, Sandeep, Krishna, Alan, and Eamonn for their continuous support in the last few years.

Finally, my thanks go to my family. My dad, who is the best teacher I have known, my mom, who's patience and cooking skills, I am yet to master. My sister, whom I always look up to. Their support through this journey is invaluable. Thank you all.

LIST OF ABBREVIATIONS

Å	Angstroms
Amp	Ampicillin
ATCC	American Type Culture Collection
ATP	Adenosine triphosphate
BLI	Bioluminescence imaging
CBR	Click Beetle Red
cfu	Colony forming units
CLIA	Clinical Laboratory Improvement Amendments
COOH	Carboxylic Acid
DIG	Digoxigenin
DMEM	Dulbecco's Modified Eagle's Medium
DMSO	Dimethyl sulfoxide
De novo	From the beginning
G	Grams
GLuc	Gaucia Luciferase
GOI	Gene of Interest
H	hours
HPRA	Health Products Regulatory Agency
HPLC	High-performance liquid chromatography
IFP	Inverse Folding Problem
I-Int	Integrated Intensity
In Silico	Use of computational methods
IP	Intraperitoneal
IPTG	Isopropyl β-D-1-thiogalactopyranoside

IT	Intratumoural
IV	Intravenous
IVIS	In Vivo Imaging System
kDa	Kilodalton
kg	Kilogram
LB	Luria Bertani broth
LUCIDs	Luciferase Based indicators of Drugs
Luc	Luciferase
mM	Millimolar
ml	Millilitre
MRI	Magnetic resonance imaging
NADH	Nicotinamide adenine dinucleotide
NMR	Nuclear Magnetic Resonance
nm	Nanometre
NH2	Amino Group
ng	Nanograms
NTR	Nitroreductase
O/N	Overnight
OD_{600}	Optical density at 600 nanometres
OI	Optical imaging
PBS	Phosphate buffered saline
PCR	Polymerase Chain Reaction
PDB	Protein Data Bank
PET	Positron emission tomography
POCT	Point of Care Test

PPI	Protein-protein Interactions
POI	Protein of interest
PVDF	Polyvinylidene difluoride
p/sec/cm^2/sr	Photons emitted per second per cm^2 per steradian
ROI	Region of Interest
RMSD	Root mean square difference
SD	Standard deviation
SEM	Standard error of the mean
SDS-PAGE	Sodium dodecyl sulfate polyacrylamide gel electrophoresis
SP	Scaffold Proteins
TBM	Template Based Modelling
Tetra-Cyst	Tera-Cystine residues
UV	Ultra violet
Wt	Wild type
µg	Micrograms
µl	Microliter
µM	Micromolar
^{o}C	Degrees Celsius
2ME	β-mercaptoethanol
RC plot	Ramachandran plot
TM	Melting temperature
VNTR	Variable number tandem repeat

Table of Contents

Chapter 1: The current state of protein design technology 1

Chapter 2: Design, build and *in vitro* test of a targeted synthetic protein strategy ..39

Chapter 3: F2F Bridge: A synthetic protein holistic performance prediction strategy 97

Chapter 4: A novel *in vivo* imaging strategy using synthetic protein engineering 146

Chapter 5: *In silico* aided-engineering of self-assembling protein cages 200

Chapter 6: Development and validation of a novel miniaturised optical imaging device 248

DISCUSSION AND FUTURE PROSPECTS.....**273**

Chapter 1

The current state of protein design technology

Table of Contents

1.1 ABSTRACT 3
1.2.1 Man's desire to design 4
1.2.2 Proteins for natural and human directed benefits 4
1.2.3 Milestones and advances in our understanding of proteins 5
1.2.4 Designing synthetic proteins 7
1.2.5 Current *in silico* tools aiding protein design 7
1.2.5.1 Understanding protein folding 10
1.2.5.1.1 Protein modelling algorithms for finding the energy minima(s) of a protein 10
1.2.5.2 Protein structure prediction 14
1.2.5.3 *De novo* design and structural remodelling 18
1.2.5.4 Predicting protein-protein interactions 23
1.2.5.5 Visualisation tools aiding protein design 23
1.2.6.1 Current tools for visualising proteins 26
1.2.7 Commercial landscape of protein design technology 26
1.2.8 How far are we from designer proteins 32
1.2.8.1 Costs of protein production and testing 32
1.2.8.3 Deploying synthetic proteins 33
1.2.8.4 Regulations and ethics 33
1.3 CONCLUSION 34
1.4 REFERENCES 35

1.1 ABSTRACT

Understanding the complexity of nature and introducing controlled modifications to biological systems paved the way for game-changing technology in biomedical sciences of the 21st century. Since the discovery of proteins in 1838, the field of protein engineering and our understanding of proteins have improved significantly. Synthetic proteins have found applications in various biomedical, food and material-based settings. This rise in synthetic proteins was complemented with the parallel expansion in the availability of *in silico* tools for protein modelling. The complexity in the composition and design of synthetic proteins requires careful *in silico* validation to screen for potential pitfalls in the design. Protein modelling, protein design and visualisation are the key concepts that need to be understood in this context. In this literature review, underlying concepts behind current protein modelling and design approaches are discussed.

1.2.1 Man's desire to design

"What I cannot create I do not understand", a quote by a theoretical physicist Richard Feynman, accurately represents the value of 'design' in man's quest for answers in science and also reflects a philosophical comprehension of our shortcomings in understanding the complexity of life. Having control over biological design and the ability to recreate/redesign life has been man's dream since the dawn of genetic engineering [1]. The commercial and ethical barriers provided the necessary tension to contain the scientific creativity within its most useful uses, and this tension between scientific creativity and ethical scepticism has resulted in astounding results. Understanding the complexity of nature and introducing controlled modifications to biological systems paved the way for game-changing technology in biomedical sciences of the 21st century [2]. DNA- and protein-based developments in recent years have benefited heavily from the arrival of synthetic biology. The availability of large databases and computational tools, low-cost DNA synthesis, high-throughput testing systems and the parallel rise in deployment-enabling technology have contributed immensely towards improving the ease and pace of scientific research.

1.2.2 Proteins for natural and human directed benefits

The diversity of the functional capabilities of proteins is unmatched with any other class of molecules. Proteins have established themselves as the primary building blocks of life. Natural proteins perform and mediate many critical functions such as providing structure and stability, mobility, pathogen clearing and other molecular sensory and regulatory functions [3]. This versatility of protein function is an attribute of the (i) amino acid composition (ii) 3D structure and (iii) interaction with other proteins and various different molecules. Engineering proteins with added functionalities has found numerous biomedical applications. As of 2010, over 200 protein-based therapeutics have been marketed [4, 5]. Protein-based therapies are being developed to target various infectious diseases and cancer , while engineered proteins are also extensively used in food, biotech and material technology-based industries [4, 6, 7].

1.2.3 Milestones and advances in our understanding of proteins

Since the discovery of proteins in 1838 [8, 9], the field of protein engineering and our understanding of proteins have improved significantly. The growth in knowledge and products based on proteins has been complimented with the parallel rise in technology aiding the study of their structure and functioning. In over a century of exploration in protein science, the road towards complete control over protein design has been marked with several important technological and regulatory milestones (Figure 1.1).

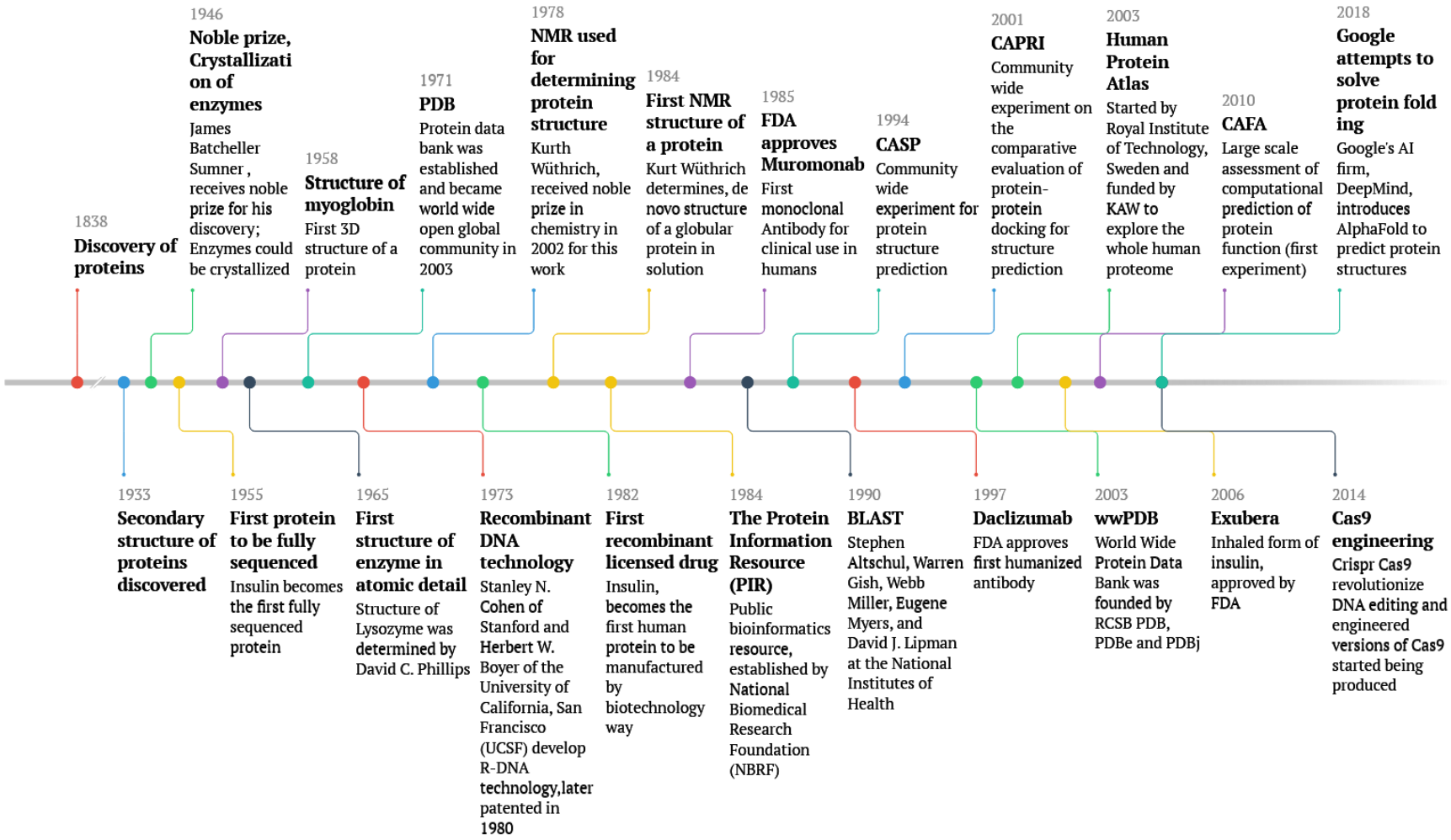


Figure 1.1: Milestones and advances in our understanding of proteins. The above timeline tracks major events upto 2018.

1.2.4 Designing synthetic proteins

In literature, the term 'synthetic proteins' has been used to define various types of engineered proteins in various contexts. In simple terms, proteins that are produced by human intervention using recombinant DNA are defined as synthetic proteins. Synthetic proteins broadly encompass (i) recombinant proteins with added functionalities by fusing one or two naturally existing protein/protein fragments, (ii) *de novo* proteins that, as whole or part-wise, never existed in nature, and (iii) proteins with unnatural amino acids and protein-like molecules (peptidomimetics). In all these cases, proteins are engineered using one or more of the above-mentioned ways to perform a user-defined function. A user-defined function may contain multiple subfunctions that are associated with different structural parts of the synthetic protein.

For any designed protein to function as intended, it is important that the structure is stable, energetically feasible, and supports all the subparts in intended locations and conformations. Appropriate exposure of the subparts is crucial for the protein to perform the user-defined function at its best capability. Techniques such as adding small peptide tags, creating point mutations and engineering backbones are commonly observed in protein engineering. In certain occasions, completely novel peptides are generated by *de novo* design. Post-modification, the structure is re-modelled, and the process is repeated until satisfactory mathematical confirmations are obtained. *In silico* aided designing stands to benefit immensely from the computational tools that predict protein interactions, perform backbone engineering, predict function and toxicity etc. [10]. The complexity in the composition and design of synthetic proteins requires careful *in silico* validation to screen for potential pitfalls in the design.

1.2.5 Current *in silico* tools aiding protein design

The rise in engineered proteins was complemented with the parallel expansion in the availability of *in silico* tools for protein modelling [10]. *In silico* tools for proteins could be broadly classified either as structure-based tools or sequence-based tools based on the input to the program. A majority of these computational tools help visualise the 3D structure of the query protein and provide graphical/mathematical

readout about the quality of a particular *in silico* parameter, such as hydrophobicity, active site, etc. Such an analysis and visualisation of the proteins provides deep insights into protein structure and forms a pivotal component in informed protein engineering and *de novo* protein design. 3D structure prediction, studying protein interactions and *de novo* design are the key aspects that need to be understood in this context. Figure 1.2 shows the commonly-used computational tools for protein modelling and design.

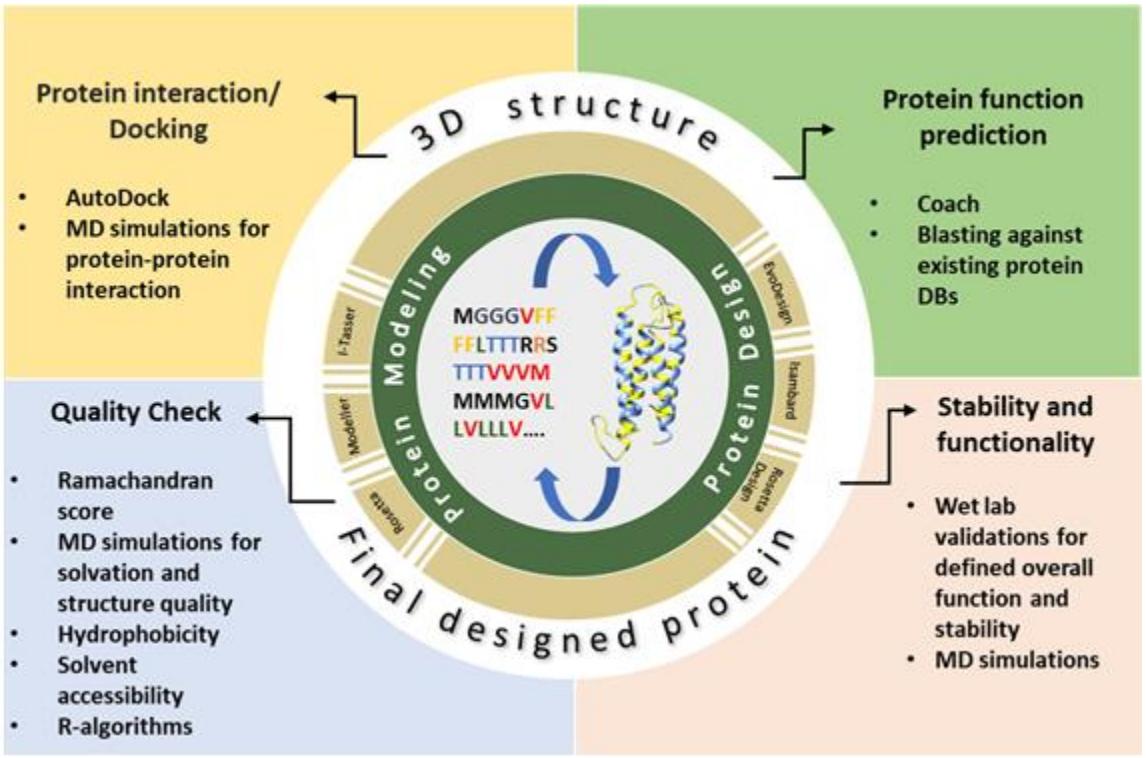


Figure 1.2 Commonly used in silico tools for protein modeling and design

1.2.5.1 Understanding protein folding

Elucidating the structure of proteins revolutionized the field of protein science and paved the way for establishment of massive databases. Traditionally, physical methods such as NMR spectroscopy and X-Ray crystallography are deployed to elucidate and study the 3D structure of proteins. The recent advances in computational sciences have resulted in sophisticated algorithms for predicting and modelling the 3D structure of a protein from its corresponding amino acid sequence. The concepts of protein fold prediction revolve around free energy models, stereochemistry and nature directed homology [11]. The degree of freedom in a conformation (steric and torsional effects), chemical interactions with neighbouring residues and physical forces surrounding the residues are the main elements that dictate protein folding [12, 13]. Properties such as charge, polarity and hydrophobicity of the residues also have significant influence on the final structure. This enormous number of variables and the combinatorial explosion of the feasible conformations make the protein-folding problem highly complicated.

1.2.5.1.1 Protein modelling algorithms for finding the energy minima(s) of a protein

American biochemist Christian Anfinsen, in his thermodynamic hypothesis, known as Anfinsen's dogma, proposed that "in the right physiological conditions, a protein will always fold into its native state". This native state of a protein is defined as the lowest Gibbs-free energy state that can be achieved by its amino acid sequence. This argument was backed-up by physical structure determination techniques such as X-ray crystallography, which soon was adopted as a standard method for protein structure determination. Crystallography, however, forces proteins into a single diffractible crystal and thus may exhibit only one native structure [14]. This 'one protein – one structure' supposition was later challenged through the discovery of 'chameleonic' sequences, metamorphic proteins and intrinsically disordered proteins [14]. The free energy landscape of a protein has multiple local minima in which a protein can settle. This presence of multiple local minima poses the greatest challenge in computational protein structure prediction. Over the years, various mathematical

strategies were proposed to provide a solution to the free energy landscape problem. Figure 1.3 illustrates the free energy landscape of a protein. Table 1.1 details the modelling algorithms commonly used in various protein structure prediction tools. Understanding the merits and demerits of the type of modelling algorithm used is important when choosing an appropriate structure prediction tool.

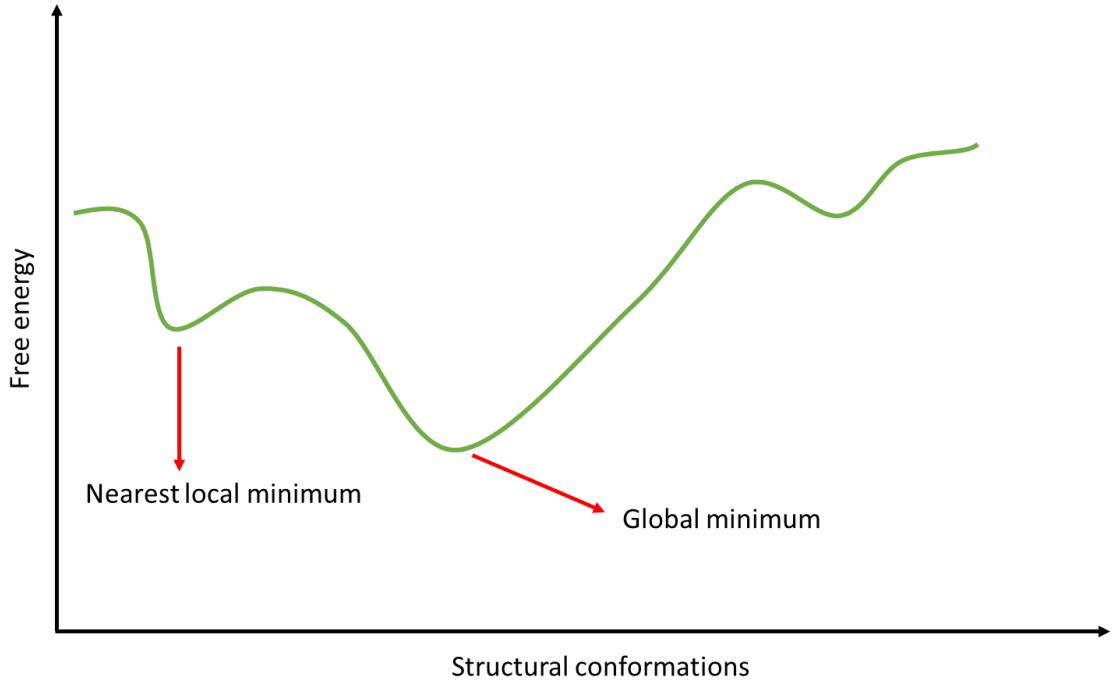


Figure 1.3: Free energy landscape of a protein. *The energy landscape of a protein has multiple mimima. The first minimum from the starting set of atomic positions is called the nearest local minimum. The lowest energy state is the global minimum.*

Algorithm	Description
Gradient based minimization	Proceeds in the direction in which free energy decreases rapidly. **Finds the nearest local minimum effectively** **Fails to generate global minimum**
Monte Carlo based sampling	Proceeds by random sampling by accepting/rejecting moves based on free energy **Finds global minimum effectively**
Molecular dynamic simulations	Uses Newton's laws of motion and calculates the force acting on each atom, due to surrounding atoms in the protein and environment. Alternative to Monte Carlo sampling. **Highly time consuming and requires high computational power**

Table 1.1: Commonly used modelling algorithms for protein structure prediction.

Pros and cons are listed in colour.

1.2.5.2 Protein structure prediction

The protein folding problem was historically solved with appreciable accuracy using two strategies: i. *Ab initio* methods, and ii. Homology-based methods [15].

1.2.5.2.1 *Ab initio* methods (Template-free method)

Ab initio refers to first principle methods. Tools that rely on *ab initio* methods for predicting use laws of physics and do not rely on protein databases [4]. *Ab initio* methods simulate the conformations using free energy function to describe the internal free energy of a protein structure and the interactions of the 3D structure with the surrounding environment. The ultimate goal of the strategy is to seek the conformation with the lowest free energy, that corresponds to the functional state of the protein. [11].

1.2.5.2.2 Homology-based methods (Template-based method)

In this method, the secondary structure of a protein is obtained by comparing the fragments with sequence homology in existing protein databases [11]. The accuracy of prediction depends on the availability of homologous sequences in the databases. Therefore, most naturally existing proteins could be modelled using homology modelling with high accuracy. However, while predicting the structure of synthetic proteins, the confidence in predicted structure depends on how close the synthetic protein resembles the naturally existing proteins.

Today's modelling uses a combination of both homology-based and *ab initio* methods to elucidate the 3D structure. Once the query sequence is given as an input, sequence alignment tools perform a thorough analysis to find similarities in the databases. In most cases, the similarities occur in small fragments all through the sequence. Computational tools are used to model the homologous parts using the proteins in the Protein Database (PDB) as a template. The sections of the sequence with minimal or no similarity are subjected to *ab initio* methods. Such an approach provides an optimal result in most cases. This process gives a linear chain of secondary structure fragments. Once the whole sequence has been modelled, the fragments are assembled by

threading, and refined to rank the predicted conformations based on free energy principles. Commonly used tools for protein modeling are listed in Table 1.2

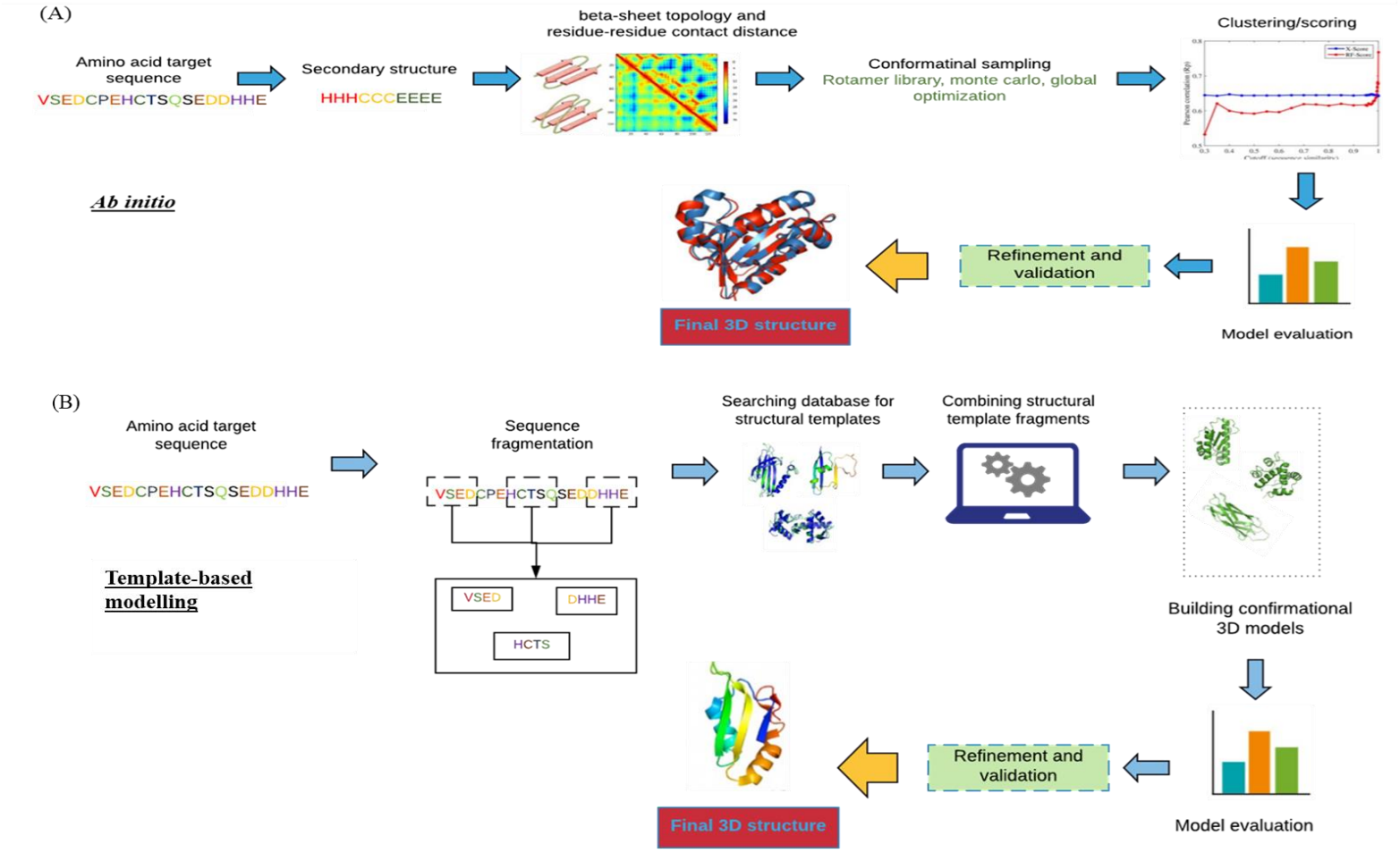


Figure 1.4: Protein structure prediction methods. *Sequence of events in ab initio and template-based modelling methods.*

Computational tools for protein modelling and protein interaction studies		
3D structure prediction tools (based on popularity and CASP)		
I-Tasser [16]	Iterative Threading ASSEmbly Refinement developed by Zhang Lab: Uses fold recognition (threading) and *Ab initio* methods to detect structure templates from existing databases (PDB). 3D models of the query sequence are constructed by reassembling structural fragments by replica exchange Monte Carlo simulations.	http://zhanglab.ccmb.med.umich.edu/I-TASSER/
Modeller [17]	Template based homology modelling program developed by Andrej Sali (University of California, San Francisco, Accelrys). Uses *Ab initio* structure prediction for regions with high variability.	https://salilab.org/modeller/
Rosetta	Automated protein structure prediction server developed by Baker's lab. Offers non-commercial comparative and ab initio modeling.	https://www.rosettacommons.org/
Raptor X [18]	Developed by Xu group, Raptor X uses remote homology recognition/protein threading for structure prediction. Provided best alignments for difficult targets in CASP 9.	http://raptorx.uchicago.edu/
Tools for studying protein interactions (based on popularity and CAPRI)		
ClusPro [19]	Widely used tool for protein-protein docking, developed by S Vajda *et al* (ABC Group and Structural Bioinformatics Lab Boston University and Stony Brook University). ClusPro also has an antibody mode. Uses PIPER (FFT based) for rigid body docking and refinement by energy minimization.	https://cluspro.org/
Z Dock [20]	Developed by Weng Z *et al* (University of Massachusetts Medical School and Boston University), predicts structures of protein-protein complexes and symmetrical multimers. Performs rigid body search for docking interaction between two proteins and uses FFT to find possible binding modes for the given protein.	http://zdock.umassmed.edu/
SwarmDock [21]	Uses particle swarm optimization to find low energy positions and binding orientations. Works in both full blind and restrained modes depending upon the availability of the information on residues. It is developed by Paul A. Bates *et al*	https://bmm.crick.ac.uk/~svc-bmm-swarmdock/
AutoDock [22]	Developed at Arthur J. Olson's Laboratory (The Scripps Research Institute and University of California). Widely known for receptor-ligand docking but also could be used for protein-protein docking.	http://autodock.scripps.edu/

Table 1.2: Widely used tools for computational protein modelling and protein interactions

1.2.5.3 *De novo* design and structural remodelling

Having partial predictive control on the protein function and redefining the functions have driven the field of protein engineering into an era of unprecedented development. However, the total number of existing proteins in nature is finite and the combinations using recombinant engineering also has limitations on what can be achieved. Nature has sampled only a fraction of total possible combinations in the total sequence space [13]. For example, a typical 100 amino acid protein can have about 20^{100} different sequence variations. Given the multiple sizes of each protein, the number of total possible proteins is beyond magnitudes of human interpretation. Nature on the other hand has a little over 10^{12} different proteins. This huge gap forms the drawing canvas for *de novo* protein design [13] Figure 1.5.

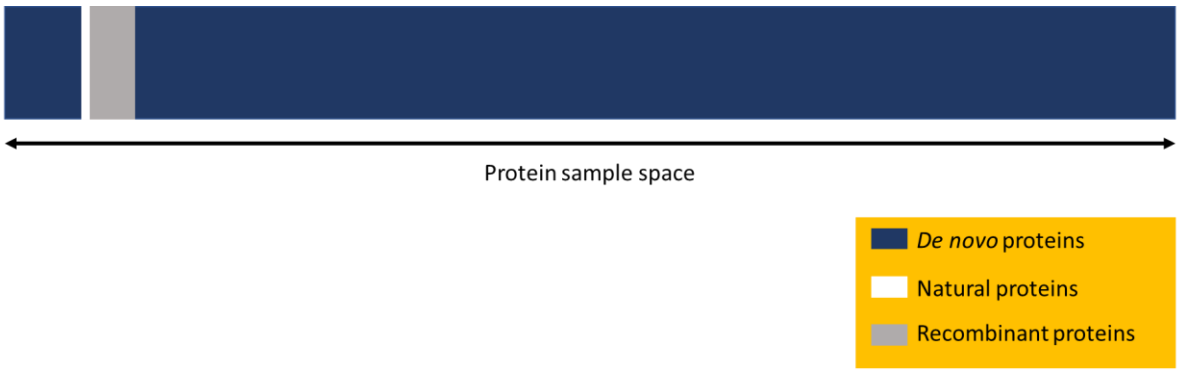


Figure 1.5: Protein sequence space representation. *Natural proteins occupy a minute fraction of the total sequence space. Directed evolution and recombinant technology in recent years have stretched the used sequence space. De novo protein design stands to capture the vast unused sequence space and promises a huge potential [13].*

De novo protein design breaks the evolutionary constraints placed by nature and explores all the mathematically feasible structures in the protein canvas. In theory, *de novo* design is the opposite of protein modelling. Protein modelling asks the question of whether it is possible to predict the protein structure of a given amino acid sequence. While the *de novo* design on the other hand, asks whether it is possible to determine an amino acid sequence that would fold into a specified query structure [4] (Figure 1.6).

Despite the immense potential of rational *de novo* protein design, more than 95% of protein engineering is still being carried out by inserting random mutations and selecting those which confer an advantage [23]. The rational design of proteins falls into two categories - the redesign of existing proteins in a process analogous to directed evolution, and the *de novo* design of completely novel proteins. Protein redesign uses naturally occurring proteins as scaffolds, and then engineers them to introduce desired changes, such as increased stability or new functional properties [24]. This will produce novel proteins, but their origins will be firmly based in the naturally occurring protein fold space. The majority of protein engineering to date has been of this nature. This method is convenient as it provides a protein backbone starting block, particularly if the desired effect represents a minor alteration in the protein's function. This becomes complicated when large numbers of amino acids are altered, since it becomes inevitable that the structure will also be altered. Native proteins are only marginally stable in many cases, so even small sequence changes can lead to dramatic changes such as aggregation or unfolding [25]. Major advances in medical science directly resulting from this degree of protein design include the humanisation of antibodies from other animal species, which entails modifying the wild type antibody to resemble human antibodies while retaining the original function. Two examples are Alemtuzumab [26] and Mepozulimab [27] for the treatment of multiple sclerosis and eosinophilic asthma respectively.

True *de novo* protein design explores the entirety of protein sequence space, guided only by the physical interactions that control protein folding. The scale of possible proteins, once naturally occurring proteins are left behind, is enormous. *De novo* protein design is based on the hypothesis that a protein will always fold into the shape associated with the lowest free energy state allowable by the amino acid sequence. Therefore, if an accurate method for measuring the energy of protein chains is

available, in addition to a method to sample different structures and sequences, it should be possible to identify sequences that fold into novel structures [25]. Once the desired shape has been reached, the stability of the novel protein can be improved by making minor adjustments, maximising the difference in free energy between the desired conformation and alternatives. *De novo* design brings the possibility of producing protein structures with novel functions that never existed in nature. Various researchers have utilised *de novo* design concepts to produce an array of repeat proteins, symmetry aided design of self-assembly and designing interfaces with special affinity towards a target [28]. However, the *de novo* design process relies entirely on computational simulations. This can hinder the accuracy of the predictions by small margins.

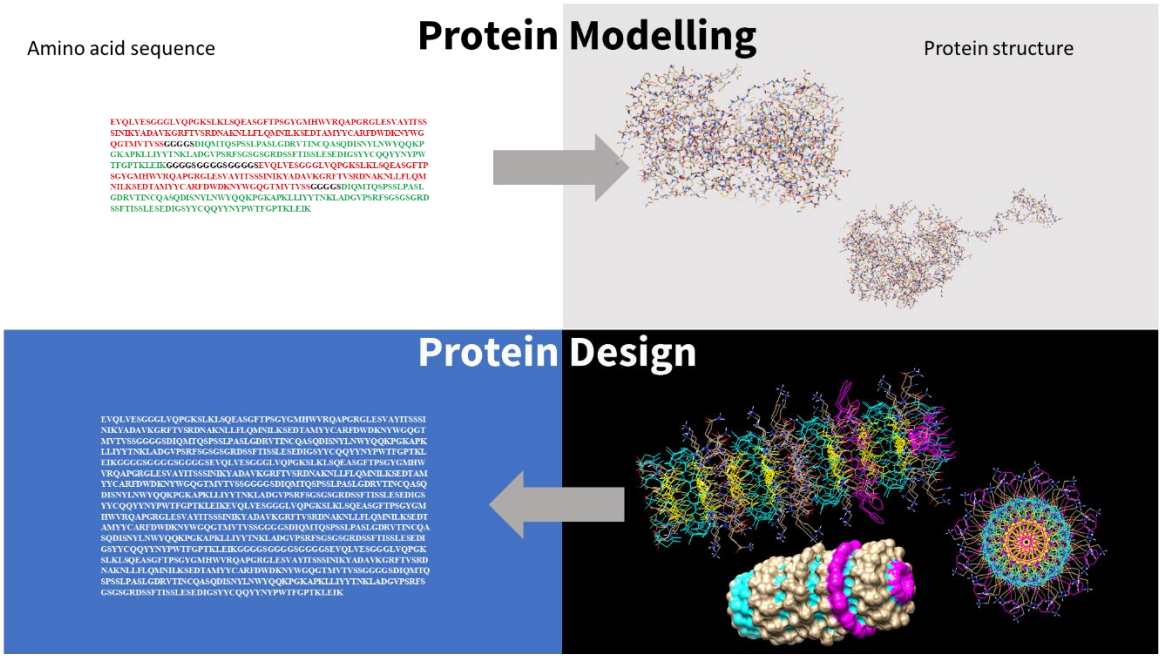


Figure 1.6: Difference between protein modelling and design. *Protein design can be understood as the inverse of protein modeling. In protein design, computational algorithms are used to predict an amino acid sequence for a user-defined structure.*

The principles of *de novo* design could also be used for structural remodelling. In protein engineering, structural modelling is performed when a small region on a protein structure requires a modification. Deleting certain residues, reconstructing backbones, disarming functional regions etc., require changes in amino acids without causing significant changes to the original protein structure. In such cases, *de novo* design principles are used for structural remodelling.

1.2.5.4 Predicting protein-protein interactions

Proteins perform and mediate various life functions by their interactions. Predicting protein interactions is crucial to understand protein function. Acquiring atomic level information of protein interactions from a wet lab study is extremely difficult to achieve. Thus, predicting protein interactions computationally is gaining popularity. Predicting protein interactions is a challenge, even using sophisticated computational algorithms. The range of potential interactions with the surroundings (medium), structural rearrangement associated with binding, flexibility, time scale and other physical factors are some of the well-known hurdles. In computational terminology, inter-molecule binding interaction is termed as docking. Docking is widely used method to study interactions of small molecules. The motion of the molecules post-interaction and structural changes while binding, are not considered. Thus, the rigid body docking falls out of agreement with large proteins. Tools such as RosettaDock have used Monte Carlo minimization-based methods to implement semi-flexible docking, specifically for proteins [29]. However, a universally reliable docking tool for full chain protein-protein interaction is a topic under research. Commonly used computational tools for studying protein interactions are listed in Table 1.2.

1.2.5.5 Visualisation tools aiding protein design

Visualisation of the 3D structure of a protein is an important component in both modelling protein structure (Physical methods and/or computational prediction) and protein design. Visualisation ***(i) For biochemists:*** Provides insights into various protein domains such as hydrophobic regions, active sites and catalytic sites. (ii) ***For evolutionary biologists:*** Visualising and mapping 3D structures aids in studying homology in structure. (iii) ***In drug designing:*** Visualising protein-protein

interactions and protein interactions with small-molecules is key to understand binding. (iv) ***In de novo protein design:*** Visualizing the designed backbone structures and superimposing the designed structures with experimental structures are very commonly performed using the molecular visualisation.

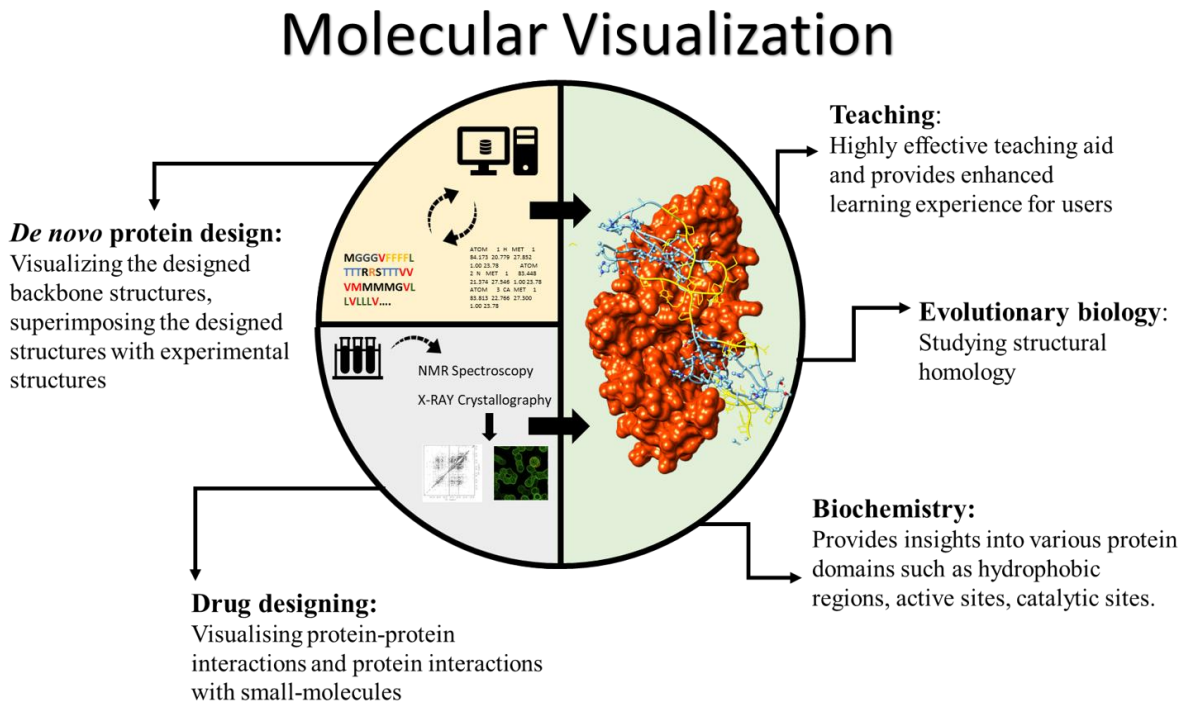


Figure 1.7: Applications of molecular visualisation. Interior graphics: *Connection between computational 'protein modeling and design' and visualisation.*

1.2.6.1 Current tools for visualising proteins

The early 20th century saw a sudden increase in the number of protein structures deposited in protein databases [30] and thus a need for better visualisation tools increased in parallel. Over the years, a wide variety of visualisation tools have been developed and deployed for protein structure visualisation. Pymol [31], VMD [32], Chimera [33] and Rasmol [34] are some examples of widely used standalone applications. Later, web-based applications such as Jmol [35] and iView [36] gained interest in the scientific community. Recently, several mobile-based applications for android and iOS have also been developed by various groups for molecular visualisation. Visualising biomolecules (proteins in particular) in virtual reality has gained wide attention recently [37]. Tools such as ChimeraX [38], BioVR [39], StarCave [40] (cave based) have been developed for visualising the 3D structure of proteins in virtual reality. Although the current technology provides a plethora of functionalities for the user, the potential of molecular visualisation in VR is still a maturing field. Easier navigation in the VR environment, better UI (user interface), faster rendering, and simplified instrumentation are some areas that are expected to see some improvements in the near future. Parallel advances in affordable VR headsets, increasing computational power and graphics, project interesting times ahead.

1.2.7 Commercial landscape of protein design technology

The protein engineering market value is estimated to reach $3.9 billion by 2024 at a CAGR of 12.4% [41]. Biotech industries and Big Pharma have successfully taken the commercial advantage of the ‘build’, ‘test’ and ‘production’ phases of the protein production chain. By expanding the protein canvas, *de novo* design opened an enormous opportunity for business. As a consequence, the *in silico* design and modeling phase gained new commercial interests. Although the ‘build’ and ‘product’ phases still remain as the major money generating stages, the recent surge of start-ups focusing on the design phase is an early indicator of the *in silico* revolution. Figure 1.8 shows various phases of the protein production chain.

Table 1.3 details the services, technology used and information on the commercial stage of recent companies focusing on the ‘design and modelling’ phase of protein

production chain. The association of these protein design companies and their parent institutions with Big Pharma and biotech industries indicates that designer synthetic proteins are approaching the market [42].

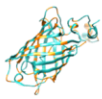
IDEA	DESIGN AND MODELING	BUILD/SCALE-UP	TEST/VALIDATION	PRODUCTS
  	**Sequence and structure modifications** Random mutations • Directed evolution • Adding features by recombinant addition of tags • Improving property improvements (e.g. Stability at higher temperatures) ***De novo* design using AI and machine learning** • Backbone design • Adding novel functions • In silico screening • Custom protein design **Software licensing** • Tools with user friendly GUI for protein modeling and design • Visualisation tools for proteins and protein-protein interactions	**Scale-up and synthesis** • Protein production in large quantities • Protein purification • High throughput screening for the required variant • Hybridoma technology • Chemical conjugation to proteins	**Experimental expertise and services** • Cryo-EM • NMR spectroscopy • X-Ray crystallography • Mass-spec **Services for quality check and property check** • HPLC • Mass-spec • Spectroscopy **Validating function** • *In vitro* testing and profiling • *In vivo* testing	**Processing of material technology and food products** • Enzymes used for food processing • Functional added proteins • Replacement proteins e.g. protein based sweeteners • Biomaterials **Diagnostic and therapeutic** • Antibodies • Biomarkers for disease detection • Reporter kits for laboratory testing • Protein based drugs
 	• **Schrodinger** • **Arzeda** • **ProteinQure** • **Codexis** • **Peptone** • **Cyrus biotechnology** • **Lab Genius** • **Autodesk Biosciences**	• **BioRad** • **Creative Biolabs** • **BioLegend** • **Thermo Fisher Scientific** • **Abcam** • **Codexis** • **Abnova**	• **Cyrus Biotechnology** • **Rockland antibodies and assays** • **Charles river** • **Proteros** • **Saromics** • **Biognosys** • **Creative-proteomics**	• **Amai proteins** • **Neoleukin Therapeutics** • **PvP biologics** • **Abcam** • **Creative Biolabs** • **BioLegend** • **Thermo Fisher Scientific** • **Big Pharma***

Figure 1.8: Various phases of protein production chain and commercial entities at each phase. Top*: Technologies and services involved in the protein production chain.* ***Bottom****: List of companies involved in each respective phase of protein production chain.*

Company Name	Description	Services	Technology	Origin	Commercial Stage*	Funding /Revenue*
Amai proteins[43]	Design of protein-based sweeteners	Hyper-sweet and thermostable designer proteins	Agile Integrative Computational Protein Design	Israel	Seed level **Lead investor:** • Yakumi • The kitchen-FoodTech hub	$850k/-
Arzeda[44]	Computationally engineered and computationally designed *de novo* proteins for agriculture, diagnostics and food-based products	• *De novo* protein design • Enzyme development • Small molecules for therapeutics • Computational services for protein engineering	• AI based prediction models, • Rosetta protein suite	Institute of protein design, University of Washington	**Early Stage Venture** Series A **Lead investor:** Universal materials incubator OS fund	$15.2M/$5 M
Codexis[45]	Engineered enzyme services for biomedical, Food and diagnostic applications	**Software services** **Sage:** Directed evolution **ProSAR:** Makes mathematical models for sequence-function relationships based on experimental data. **MOSAIC:** Evaluates the interactions of multiple mutations	• Machine learning • Neural networks • AI for prediction models	San Francisco, US	Post IPO Equity **Lead Investor** • Casdin capital • EDBI	$87M/$65.9M

		Other services • Protein engineering • Scale-up and Supply Screening kits based on proteins				
Neoleukin Therapeutics[46]	*De novo* – therapeutic proteins for immunological disorders	***De Novo* Protein Design** Custom *de novo* designed proteins for therapeutic applications **The Neoleukin Platform** Proving protein modification, *de novo* reengineering, property modifications etc.	Rosetta suite	Institute of protein design, University of Washington	Acquired by Aquinox Pharmaceuticals	-/-
ProteinQure[47]	Computational platform for designing protein therapeutics	• Protein structure determination • Protein design and • Protein property optimisation services	• Quantum computing • Molecular dynamic simulations • Artificial intelligence and Machine learning models	Toronto	**Seed level** Privately owned company **Lead Investor:** Felicis Ventures	$4.6M/-
Peptone[48]	AI based computational protein assessment and design	• Antibody Design and Optimization • Protein Developability Assessment • Automated Thermostability Engineering • AI — Assisted Survey of "Dark Proteome"	• Computational biophysics • non-linear modeling • statistical approaches • progressive AI • Cloud based technology	London	**Seed level** Privately owned company **Lead Investor:** Founders Factory	$350k/-
PvP biologics[49]	Oral enzyme for Celiac disease	KUMAMAX – an oral enzyme for the treatment of celiac disease	Rosetta suite	**Idea:** iGEM competition	Private company	$35/$3M

				Development: Institute of protein design, University of Washington	**Lead investor:** Takeda Pharmaceutical	
Cyrus Biotechnology[50]	Offers service and solutions for protein design and modeling problems	**Software solutions** Rosetta based protein modeling, design and interaction prediction **Service solutions** • Cryo-EM • Protein design • Protein structure prediction NMR and X-Ray crystallography	**Cyrus Bench:** Rosetta based suite for molecular modeling and design **Cyrus Bench** is an easy-to-use version of various Rosetta based tools packed into one suite.	Institute of protein design, University of Washington	**Early stage venture** Privately-owned company **Lead investor:** Trinity Ventures	$10.4M/$1M
LabGenius[51]	Evolving novel proteins using machine learning driven approaches	**Protein based therapeutics** • Improved targeting • Adding custom properties • Exploring potential drug candidates	• **EVA**- machine learning-driven evolution engine. • Deep-learning neural networks to explore and improve protein properties	London	**Early stage venture** Privately-owned company **Lead investor:** Obvious ventures Lux capital Acequia capital	13.7M/$1M

Table 1: Services, technology used and the commercial stage of recent companies focusing on the 'design and modelling' phase of protein production chain

*Information on commercial stage, lead investors, estimated revenue and investment amounts was obtained from Crunchbase [52].

1.2.8 How far are we from designer proteins

1.2.8.1 Costs of protein production and testing

The advent of synthetic biology has brought down the costs associated with DNA sequencing and analysis dramatically [53,54]. This had a profound effect in lowering research costs which involve studying DNA and technology and products that depend on synthetic DNA. Miniaturised and portable devices for sequencing, handheld alternatives for PCRs, spectrophotometers have enabled the deployment of DNA based technology into commercial settings [55-57]. Protein based technology, on the other hand, is playing a catching-up game, the reliance on large bench top instrumentation such as NMR, X-Ray crystallography, Cryo-EM based methods and traditional testing methods involving expensive reagents such as antibodies and a lack of robust and affordable *in vitro* testing systems are some hurdles to overcome in coming future.

1.2.8.2 Reliability of *in silico* tools

The reliability of outputs of the current *in silico* tools for protein modelling and design depends on several factors, such as protein complexity, sequence and structural homology and the choice of individual algorithms and methods used in the *in silico* workflow. However, on a global scale, community-wide experiments such as Critical Assessment of protein Structure Prediction (CASP), Critical Assessment of PRediction of Interactions (CAPRI) and the Critical Assessment of Functional Annotation (CAFA) have been benchmarking the computational tools for structure prediction, protein interaction prediction and function prediction. This is achieved by ranking the *in silico* tools based on their blind-folded prediction ability of various sets of queries. CASP, CAPRI and CAFA are conducted once every two years. The improvement in the performance shown by the tools in the community-wide experiments assures increasing reliability.

1.2.8.3 Deploying synthetic proteins

The applications of synthetic proteins expand into various sectors such as food-based industries, materials technology, and biomedicine. The advent of *de novo* protein design has only increased the ever-expanding canvas of protein structures. The advances in synthetic biology are promoting lab-based novel scientific concepts to translate into deployable and commercially-viable products. Commercial viability requires transforming a scientific outcome into a deployable product. For example, a protein based diagnostic tool would additionally require an appropriate hardware for testing and a software to analyse and report the results. The deployability factor becomes crucial while these lab-based concepts shift to commercial or consumer-based settings. Interdisciplinary approaches form a bridge connecting a scientific concept and a consumer.

1.2.8.4 Regulations and ethics

Use of purified engineered proteins presents lower environmental, food and drug regulatory barriers than 'live' products featuring genetically engineered DNA, given a typical protein's 'terminal' state. Nonetheless, immunotherapy with engineered antibodies, CAR-T cell technology, Crisper-Cas9, protein-based drugs for autoimmune diseases and Gene Therapy have multiple forms of protein engineering as a key integral component and have always been a topic in scientific ethics and policy making [58-60]. Having control over biological design and the ability to recreate/redesign life has been man's dream since the dawn of genetic engineering [1]. Commercial and ethical barriers provided the necessary tension to contain the scientific creativity within its most useful uses, and this tension between scientific creativity and ethical scepticism is producing a range of products, including protein-based, with capacity to improve the world.

1.3 CONCLUSION

Our ability to design and engineer proteins has advanced considerably over the last decade. In this review, various *in silico* tools that are crucial for synthetic protein design have been highlighted. Special focus was placed on (i) protein structure prediction, (ii) *de novo* protein design and (iii) protein visualisation techniques, all of which are studied in this thesis. The value of integrating multidisciplinary approaches in transforming a scientific concept into a commercially viable product has been discussed.

1.4 REFERENCES

1. Simpson, M.L., *Cell-free synthetic biology: a bottom-up approach to discovery by design.* Mol Syst Biol, 2006. **2**: p. 69.
2. Kuhlman, B. and P. Bradley, *Advances in protein structure prediction and design.* Nat Rev Mol Cell Biol, 2019. **20**(11): p. 681-697.
3. Tiwari, M.K., et al., *Computational approaches for rational design of proteins with novel functionalities.* Comput Struct Biotechnol J, 2012. **2**: p. e201209002.
4. Khoury, G.A., et al., *Protein folding and de novo protein design for biotechnological applications.* Trends Biotechnol, 2014. **32**(2): p. 99-109.
5. Vlieghe, P., et al., *Synthetic therapeutic peptides: science and market*. Drug Discov Today, 2010. **15**(1-2): p. 40-56.
6. Lu, R.M., et al., *Targeted drug delivery systems mediated by a novel Peptide in breast cancer therapy and imaging.* PLoS One, 2013. **8**(6): p. e66128.
7. Craik, D.J., et al., *The future of peptide-based drugs.* Chem Biol Drug Des, 2013. **81**(1): p. 136-47.
8. Vickery, H.B., *The origin of the word protein.* Yale J Biol Med, 1950. **22**(5): p. 387-93.
9. Hartley, H., *Origin of the word 'protein'.* Nature, 1951. **168**(4267): p. 244.
10. Yallapragada, V.V.B., et al., *Function2Form Bridge-Toward synthetic protein holistic performance prediction.* Proteins, 2019.
11. Dorn, M., et al., *Three-dimensional protein structure prediction: Methods and computational strategies.* Comput Biol Chem, 2014. **53PB**: p. 251-276.
12. Setiawan, D., J. Brender, and Y. Zhang, *Recent advances in automated protein design and its future challenges.* Expert Opin Drug Discov, 2018. **13**(7): p. 587-604.
13. Huang, P.S., S.E. Boyken, and D. Baker, *The coming of age of de novo protein design.* Nature, 2016. **537**(7620): p. 320-7.
14. Dishman, A.F. and B.F. Volkman, *Unfolding the Mysteries of Protein Metamorphosis.* ACS Chem Biol, 2018. **13**(6): p. 1438-1446.

15. Deng, H., Y. Jia, and Y. Zhang, *Protein structure prediction.* Int J Mod Phys B, 2018. **32**(18).
16. Yang, J., et al., *The I-TASSER Suite: protein structure and function prediction.* Nat Methods, 2015. **12**(1): p. 7-8.
17. Sali, A. and T.L. Blundell, *Comparative protein modelling by satisfaction of spatial restraints.* J Mol Biol, 1993. **234**(3): p. 779-815.
18. Kallberg, M., et al., *Template-based protein structure modeling using the RaptorX web server.* Nat Protoc, 2012. **7**(8): p. 1511-22.
19. Kozakov, D., et al., *The ClusPro web server for protein-protein docking.* Nat Protoc, 2017. **12**(2): p. 255-278.
20. Pierce, B.G., et al., *ZDOCK server: interactive docking prediction of protein-protein complexes and symmetric multimers.* Bioinformatics, 2014. **30**(12): p. 1771-3.
21. Torchala, M., et al., *SwarmDock: a server for flexible protein-protein docking.* Bioinformatics, 2013. **29**(6): p. 807-9.
22. Morris, G.M., et al., *AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility.* J Comput Chem, 2009. **30**(16): p. 2785-91.
23. Perkel, J.M., *The computational protein designers.* Nature, 2019. **571**(7766): p. 585-587.
24. Wood, C.W., et al., *ISAMBARD: an open-source computational environment for biomolecular analysis, modelling and design.* Bioinformatics, 2017. **33**(19): p. 3043-3050.
25. Huang, P.-S., S.E. Boyken, and D. Baker, *The coming of age of de novo protein design.* Nature, 2016. **537**: p. 320.
26. Havrdova, E., D. Horakova, and I. Kovarova, *Alemtuzumab in the treatment of multiple sclerosis: key clinical trial results and considerations for use.* Therapeutic advances in neurological disorders, 2015. **8**(1): p. 31-45.
27. Poulakos, M.N., et al., *Mepolizumab for the treatment of severe eosinophilic asthma.* Am J Health Syst Pharm, 2017. **74**(13): p. 963-969.
28. Brunette, T.J., et al., *Exploring the repeat protein universe through computational protein design.* Nature, 2015. **528**(7583): p. 580-4.

29. Pedotti, M., et al., *Computational docking of antibody-antigen complexes, opportunities and pitfalls illustrated by influenza hemagglutinin.* Int J Mol Sci, 2011. **12**(1): p. 226-51.
30. PDB, *PDB Statistics: Overall Growth of Released Structures Per Year.* PDB website, 2019.
31. Pymol.org, *Pymol.* 2002.
32. Humphrey, W., A. Dalke, and K. Schulten, *VMD: visual molecular dynamics.* J Mol Graph, 1996. **14**(1): p. 33-8, 27-8.
33. Pettersen, E.F., et al., *UCSF Chimera--a visualization system for exploratory research and analysis.* J Comput Chem, 2004. **25**(13): p. 1605-12.
34. Sayle, R.A. and E.J. Milner-White, *RASMOL: biomolecular graphics for all.* Trends Biochem Sci, 1995. **20**(9): p. 374.
35. Jmol, *Jmol: an open-source Java viewer for chemical structures in 3D.* 2019.
36. Li, H., et al., *iview: an interactive WebGL visualizer for protein-ligand complex.* BMC Bioinformatics, 2014. **15**: p. 56.
37. Goddard, T.D., et al., *Molecular Visualization on the Holodeck.* J Mol Biol, 2018. **430**(21): p. 3982-3996.
38. Goddard, T.D., et al., *UCSF ChimeraX: Meeting modern challenges in visualization and analysis.* Protein Sci, 2018. **27**(1): p. 14-25.
39. Zhang, J.F., et al., *BioVR: a platform for virtual reality assisted biological data integration and visualization.* BMC Bioinformatics, 2019. **20**(1): p. 78.
40. DeFanti, T.A., et al., *The StarCAVE, a third-generation CAVE and virtual reality OptIPortal.* Future Gener. Comput. Syst., 2009. **25**(2): p. 169-178. Markets, M.a., *Protein engineering market size*. 2019. https://www.marketsandmarkets.com/PressReleases/protein-antibody-engineering.asp *accessed on 10th Feb 2020*
42. Amgen, *Amgen and IPD announce strategic research partnership.* 2020. https://wwwext.amgen.com/media/news-releases/2019/06/amgen-and-the-institute-for-protein-design-ipd-at-university-of-washington-announce-unique-strategic-research-partnership/ *accessed on 10th Feb 2020*
43. Proteins, A., 2018. https://www.amaiproteins.com/ *accessed on 10th Feb 2020*

44. Arzeda, *The protein design company, arzeda. https://www.arzeda.com/ accessed on 10th Feb 2020*
45. Codexis, *Codexis company. https://www.codexis.com/ accessed on 10th Feb 2020*
46. Therapeutics, N. https://www.neoleukin.com/ *accessed on 10th Feb 2020*
47. ProteinQure, *ProteinQure. https://www.proteinqure.com/ accessed on 10th Feb 2020*
48. Peptone, *_Peptone. https://peptone.io/ accessed on 10th Feb 2020*
49. biologics, P., *PVPbio. https://www.pvpbio.com/ accessed on 10th Feb 2020*
50. Cyrusbio.com, *Cyrus biotechnology. https://cyrusbio.com/ accessed on 10th Feb 2020*
51. Labgenius, *Labgenius. https://www.labgeni.us/ accessed on 10th Feb 2020*
52. Crunchbase, *Crunchbase, information about companies. https://www.crunchbase.com/home, accessed on 10th Feb 2020*
53. El Karoui, M., M. Hoyos-Flight, and L. Fletcher, *Future Trends in Synthetic Biology-A Report.* Front Bioeng Biotechnol, 2019. **7**: p. 175.
54. Cameron, D.E., C.J. Bashor, and J.J. Collins, *A brief history of synthetic biology.* Nat Rev Microbiol, 2014. **12**(5): p. 381-90.
55. Erlich, Y., *A vision for ubiquitous sequencing.* Genome Res, 2015. **25**(10): p. 1411-6.
56. Mongan, A.E., J.S.B. Tuda, and L.R. Runtuwene, *Portable sequencer in the fight against infectious disease.* J Hum Genet, 2020. **65**(1): p. 35-40.
57. Nanopore, O., *Oxford Nanopore MinION. https://nanoporetech.com/products/minion accessed on 5th May 2020.*
58. Hirsch, F., R. Iphofen, and Z. Koporc, *Ethics assessment in research proposals adopting CRISPR technology.* Biochem Med (Zagreb), 2019. **29**(2): p. 020202.
59. Morrison, M. and S. de Saille, *CRISPR in context: towards a socially responsible debate on embryo editing.* Palgrave Communications, 2019. **5**(1): p. 110.
60. Riva, L. and C. Petrini, *A few ethical issues in translational research for gene and cell therapy.* J Transl Med, 2019. **17**(1): p. 395.

Chapter 2

Design, build and *in vitro* testing of a targeted synthetic protein strategy

Table of Contents

2.1 ABSTRACT....42
2.2 INTRODUCTION....43
2.2.1 Synthetic proteins with minimal regions of antibodies....43
Synthetic antibody fragments....43
2.2.2 Designing a synthetic protein with multiple 'parts'....45
2.2.3 Bacterial Imaging....47
2.2.4 *Staphylococcus aureus*....47
2.2.5 Gaussia Luciferase as a reporter protein....52
2.3 MATERIALS AND METHODS....55
2.3.1 Overview of *in silico* design of synthetic proteins....55
2.3.2 Computational tools used for *in silico* -aided design and validation....55
2.3.2.1 Protein structure modeling....55
2.3.2.2 Superimposing predicted models....55
2.3.2.3 3D visualisation....56
2.3.2.4 Protein-protein interactions....56
2.3.2.5 Theoretical structure validation....56
2.3.2.6 Total hydrophobicity vs Surface hydrophobicity....56
2.3.2.7 Structural remodeling and affinity improvements....57
2.3.3 DNA design....57
2.3.4 Plasmid scale-up....58
2.3.5 Restriction digestion....59
2.3.6 Gibson Assembly....59
2.3.7 Validating cloning using colony PCR and Sanger sequencing....59
2.3.8 *In vitro* transfection....60
2.3.9 Binding assays....60
2.4 RESULTS61
2.4.1 Design elements and design rationale....61
2.4.2 *In silico*-aided screening....66
2.4.3 Test constructs for wet lab testing....68

2.4.4 Protein-Protein Docking....71
2.4.5 Wet-lab experimentation....75
2.4.5.1 Protein production and secretion....75
2.4.5.2 Binding to *S. aureus* ClfA....77
2.4.5.3 Dose response experiments....79
2.4.5.4 Blocking with human fibrinogen....83
2.4.5.5 Optimal performer (S+I+N)....85
2.4.5.6 Nanoluc as the luminescence part....87
2.5 DISCUSSION....89
2.6 CONCLUSION....92
2.7 REFERENCES....93

2.1 ABSTRACT

Background Incorporation of minimal regions of antibodies within engineered proteins presents an attractive strategy to target proteins to specific molecules or cells. Such targeted synthetic proteins have an enormous potential in various diagnostic and therapeutic applications. Designing a synthetic protein involves modeling and testing multiple test variants. A typical synthetic protein with 5 defined subparts can be assembled into 5! i.e. 120 different variants. Current *in silico* tools help identify merits and pitfalls of the design and aid in screening for potential best performers. Such an *in silico* aided screening reduces the costs and labour involved in wet-lab experimentation.

Aims The aim of this work was to (i) computationally design and model multiple variants of a bacterial targeted synthetic protein and screen for potential best performers, and (ii) build and test the synthetic protein test variants using wet-lab assays.

Methods Over 50 different multi-part test constructs targeted to *S. aureus* surface antigen ClfA were modelled and validated using various computational tools to inform and guide downstream wet-lab experiments. For targeting, constructs featured either mono-valent ScFVs or bi-valent mono-specific diabody fragment sequences. Gaussia luciferase (Gluc) was used as a luminescence reporter on all test variants. All test construct variants were subjected to computational screening of predicted functionality. The best predicted performers were appropriately modified to ensure required hydrophobicity, net surface charge, active site exposure and valid 3D structure. After a thorough *in silico* validation, wet-lab studies were conducted to validate protein production and functioning (lumiescence and specific target binding) *in vitro*.

Results Following *in silico* design and analyses, 10 test constructs against ClfA were produced in CHO cells and tested for specific target binding *in vitro*. Based on luminescence readouts, a ScFv-featuring test construct was identified as the best performer, in terms of *S. aureus* specific binding, as evidenced by luminescence-based assays.

Conclusion The outputs of this study were i) a validated *in silico* and wet-lab strategy for design, build and test of targeting synthetic proteins, and ii) a target-specific reporter protein platform for bacterial imaging.

2.2 INTRODUCTION

2.2.1 Synthetic proteins with minimal regions of antibodies

Over the last decade, the number of protein-based imaging and therapeutic applications have grown considerably. Engineering proteins to target specific cells or proteins is a common approach used in protein-based biomedical applications [1]. This targeting requires the binding of proteins to targets of interest. Using antibodies as targeting leads is a well-known strategy. Radioactive labeling of antibodies and chemical conjugations with fluorescent labels and nanoparticles are some of the common ways of exploiting antibodies. However, due to the large size, full antibodies present many disadvantages such as risk of immunogenicity, slow clearance and low tissue penetration [2-4].

Incorporating minimal antibody regions into a synthetic protein as a binding domain is an attractive strategy for targeting a subject of interest. The small size of the minimal regions improves the pharmacokinetic properties and reduces the risk of immunogenicity. A typical synthetic protein consists of multiple subparts with individual functions. The small size of the minimal regions provides great flexibility in design. Depending on the user requirement, the minimal regions can be assembled into multiple variants of antibody fragments such as ScFvs, diabodies, nanobodies etc. The valency and specificity of the synthetic protein could be altered by using multiple minimal regions of antibodies with different targets. Figure 2.1 illustrates a few examples of antibody fragments constructed from the minimal regions of a full IgG antibody.

The functional versatility of proteins and modular aspects of recombinant protein engineering paved the way towards designing proteins with multiple added functionalities. In the last decade, engineering antibodies for customised applications took an unprecedented development [5]. Engineered antibody fragments are fragmented small antibody parts assembled by researchers for diversifying and improving the functionality. With current computational tools and engineering of appropriate genetic elements, antibodies could be engineered to have customised affinity, half-life, valency, minimal toxicity, avidity and other specific biological functions [5, 6]. Examples of such engineered antibody fragments are shown in Figure 2.1.

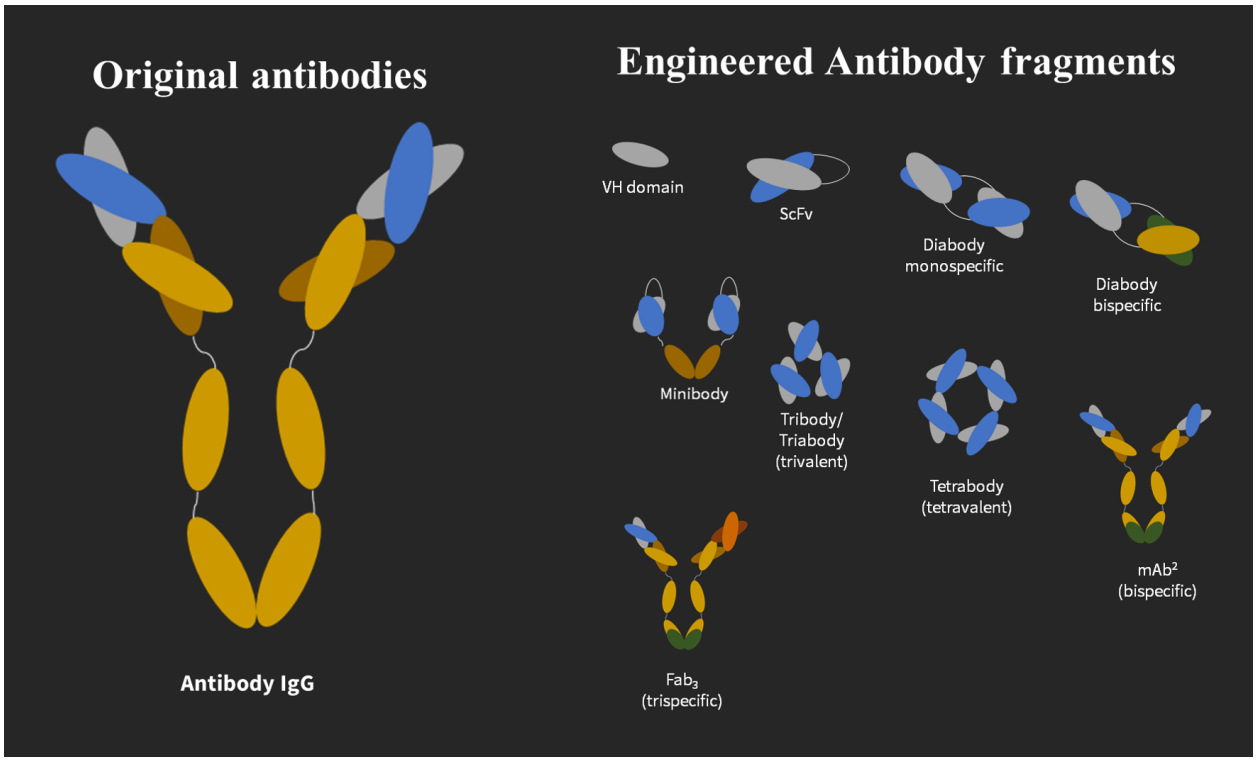


Figure 2.1: Original IgG antibody format and various engineered antibody fragments. The engineered antibody fragments differ in their specificity, stability and valency. VH and VL domains are the minimal regions of the antibody that primarily contribute to binding to the target. The presence of multiple copies of minimal regions increases the valency of the antibody fragment and the presence of minimal regions from two or more different antibodies adds multiple target specificity (shown in different colors). The minimal regions are often connected by small rigid/flexible linkers.

2.2.2 Designing a synthetic protein with multiple ‘parts’

Designing a synthetic protein with a defined function requires building and testing of a range of variants. For any designed protein to function as intended, it is important that the structure is stable, energetically feasible, and supports all the subparts in the intended locations and conformations. In most cases, the synthetic protein is composed of smaller subparts with unique subfunctions. These parts are connected to each other directly or by using small peptide linkers. The subparts may also include additional enhancer elements to improve the stability, solubility and secretion as appropriate. As the number of subparts increases, the choice of subpart types, variants and the number of ways to assemble these also increases (Figure 2.2).

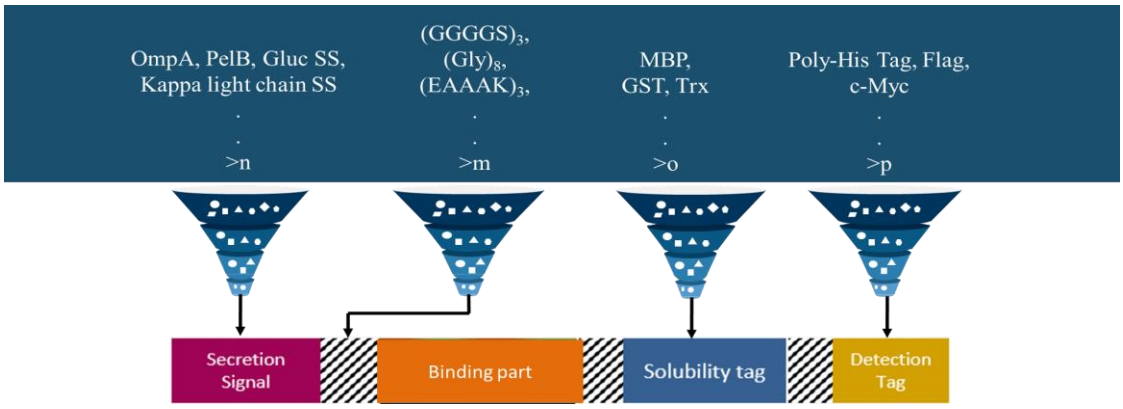


Figure 2.2: Number of variants possible for a synthetic protein with 4 subparts. *For a protein with 4 subparts where there are n different choices available for a secretion signal, m different choices for the linkers, o different choices for the solubility enhancer and p different choices for the detection tag. In most cases the order of the assembly also is a variant in itself.*

For 4 subpart construct, with each subpart belonging to a subset (n,m,o,p), the total possible variants are given in the formula below.

$$[n + m + o + p]!/[(n + m + o + p) - 4]!$$

For an R subparts construct, the equation changes to:

$$[n + m + o + p + \ldots\ldots]!/[(n + m + o + p + \ldots..) - R]!$$

Wet lab synthesis and testing of all variants is a laborious and highly expensive process. Such methods are also not suitable for high throughput applications. Computationally modeling the test variants and *in silico* screening provides a screening rationale and also assists the wet-lab biologist by informing about the pitfalls and merits of an engineered protein, prior to synthesis. The final design variants can also be improved continuously based on the feedback from the computational tools.

2.2.3 Bacterial Imaging

Bacteria present both a beneficial and a detrimental role for various human needs. Understanding the pathogenicity of bacteria is crucial clinically, in preventing and curing infectious diseases. For research purposes, monitoring bacterial trafficking in the body is valuable. Non-invasive detection of specific bacteria also benefits diagnosis and treatment of infectious diseases. Developing a non-invasive imaging strategy to detect bacteria within a living host would benefit various fields including, infectious diseases, gene therapy and cancer.

2.2.4 *Staphylococcus aureus*

S. aureus is a Gram-positive, clinically important pathogen that can cause a range of diseases from superficial skin infections to pneumonia, endocarditis and fatal sepsis in humans [7]. *S. aureus* is often found in skin and nostrils and is a formidable opportunistic type. The ability of *S. aureus* to survive in a variety of locations in the body is an attribute of its range of virulence factors such as toxins, adhesins and proteins that help it to evade the immune system [8, 9]. Understanding the mechanism of *S. aureus* interactions with the host provides insights to develop novel imaging and therapeutic strategies. Biofilm formation and aggregation or microcolony establishment are the two main modes of *S.*

aureus infections [10]. Biofilm based *S. aureus* infections are common in medical devices and other surface-based settings [11]. However, microcolony establishments are increasingly seen in host infections. In microcolony based infections, *S. aureus* aggregates are embedded into the extracellular matrix composed of proteins such as fibrinogen and collagen [10]. *S. aureus* also interacts directly with fibrinogen and thus forms large clusters of cells (Figure 2.3).

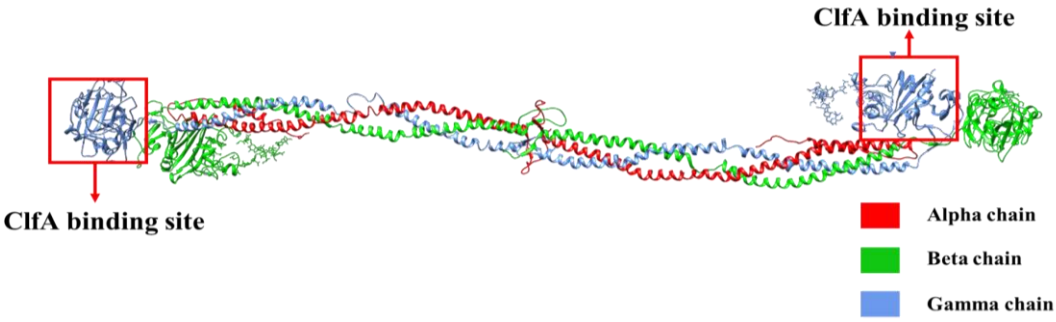


Figure 2.3: Crystal structure of human fibrinogen (PDB ID: 3GHG). *The* α, β *and* γ *chains are depicted in red, green and blue respectively. The MSCRAMM, microbial surface component recognizing adhesive matrix molecules binding site is highlighted in red box.*

This ability to interact with fibrinogen is a hallmark feature of *S. aureus* and is known as clumping (Figure 2.4). The clumping of *S. aureus* is facilitated by its virulence factor ClfA (Clumping factor A) [12]. Targeting the virulence factor ClfA (clumping factor A) is a promising strategy to detect and prevent *S. aureus* based infections.

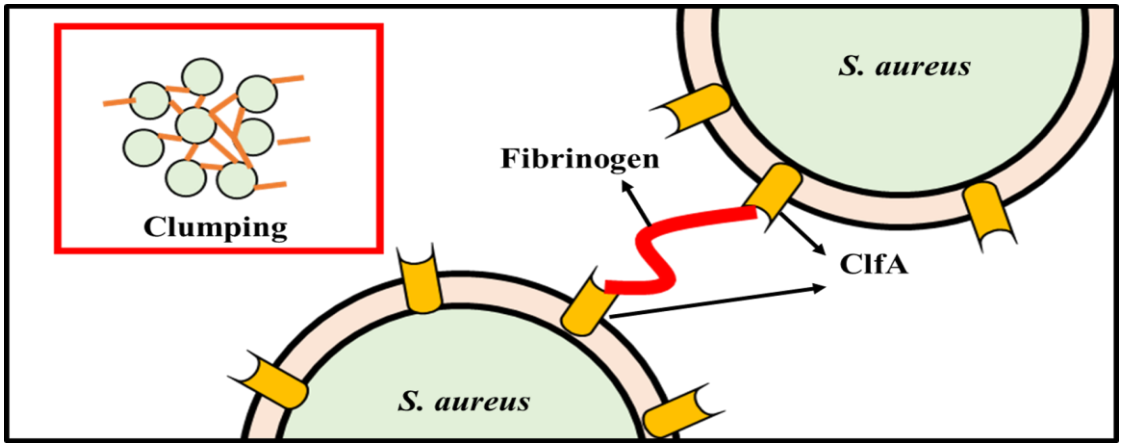


Figure 2.4: Schematic showing clumping of S. aureus. *ClfA facilitates clumping by binding to distal ends of fibrinogen dimer. This allows the dimer to act as a connecting bridge between the bacterial cells.*

2.2.5 Gaussia Luciferase as a reporter protein

Bioluminescence is the liberation of energy in the form of light by a reporter protein which, in most cases, oxidises a substrate molecule in an ATP (or $FMNH_2$) dependent manner [13]. Bioluminescence reactions do not require light absorption/excitation. Bioluminescence depends on an enzyme called luciferase, a substrate, commonly known as luciferin and oxygen. Some reactions also would require, ATP and Mg^{2+} as cofactors for their activity. The term luciferase encompasses all the enzymes that produce light on catalysis. There are a wide variety of luciferases found in nature ranging from fireflies to deep ocean algae. Firefly (Fluc), Click Beetle (CBluc), Renilla (Rluc) and Gaussia (Gluc) are a few of the most widely studied luciferases [14-16]. Amongst all the luciferases mentioned above, Gluc is an ATP independent luciferase. The working mechanism of Gluc is shown in figure 2.5. Due to its small size, ATP independent working, high stability at elevated temperatures, ability to be secreted outside the cells and bright signal, Gluc is suitable to be integrated recombinantly as a reporter into any synthetic construct.

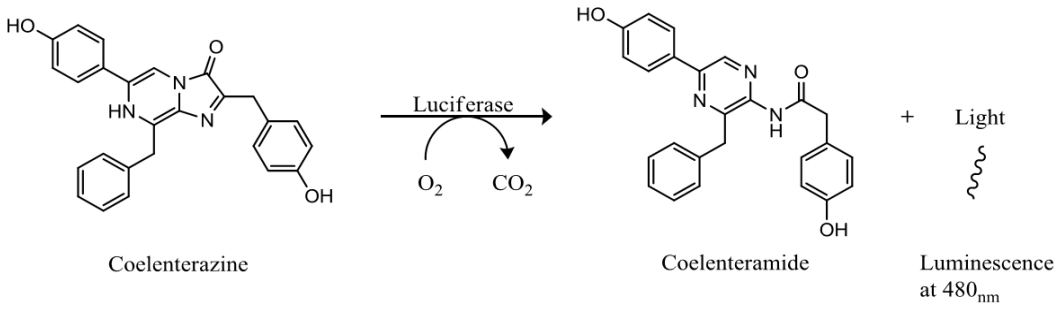


Figure 2.5: Working mechanism of Gaussia luciferase catalysed bioluminescence. *Gaussia luciferase catalyses the conversion of its substrate Coelenterazine to coelenteramide, in presence of oxygen. During this process energy is emitted in the form of light at 480nm (blue).*

The potential of bioluminescence imaging has been explored previously in various disease settings. Bioluminescence has been previously used to study mRNA stability, miRNA expression, studying signaling pathways, post translational modifications, protein-protein interactions, understanding kinetics of proteins, imaging tumors, etc [17]. Bioluminescence represents an efficient and affordable method, as the instrumentation required is relatively inexpensive, low-cost and minimal consumables are required, minimal technical expertise is required to learn the imaging methodology and easy data analysis [18]. Considering the above-mentioned advantages, bioluminescence based Optical Imaging promises a huge potential to be deployed as a novel imaging strategy towards various fields of medicine.

This study demonstrates the proof-of-concept of the design-model-build-test strategy of targeted synthetic proteins, using *S. aureus* surface antigen ClfA as a target.

2.3 MATERIALS AND METHODS

2.3.1 Overview of *in silico* design of synthetic proteins

Synthetic proteins contained multiple subparts with individual subfunctions. Over 200 different variants were made and manually screened for potential best performers. Various *in silico* tools were used during this process and several iterations of each test variant were validated until desired structural conformations were achieved. VH and VL domains of anti ClfA antibody MAb 12-9 was used as the binding domain to target ClfA on *S. aureus*. The amino acid sequence was obtained from a patent application.

http://www.google.ch/patents/US20050287164

Gluc was used as the luminescent imaging part. The test variants differed in their part arrangement order, presence and absence of additional domains. The process workflow and methods are explained in the sections below.

2.3.2 Computational tools used for *in silico* -aided design and validation

2.3.2.1 Protein structure modeling

All the amino acid sequences of the test variants were subjected to protein modeling to predict their 3D conformation. The protein modeling was performed primarily using the I-Tasser protein modeling suite [19]. Both web server and the standalone suite were used to perform modeling of multiple constructs in parallel. C-score from the I-Tasser output was used to determine the best model after each modeling experiment. Higher C-Score indicates higher confidence in the predicted model. The test constructs were also modeled using Rosetta modeling suite to increase the prediction reliability [20, 21].

2.3.2.2 Superimposing predicted models

The models predicted by I-Tasser and Rosetta were superimposed onto each other to calculate the agreement of the relative position of each atom in space. Root Mean Square Distance (RMSD) was used as parameter to judge the agreement between the models predicted by the two modeling tools. R-algorithms, developed in-house at the Tangney lab, were used for this purpose.

2.3.2.3 3D visualisation

Visualising protein structures in 3D is a central element of protein modeling. The alignment of important parts in right conformation was ensured by visually validating each individual test variant. UCSF Chimera was used for all protein visualisation throughout this work [22]. Highlighting various domains and specific sequences in the 3D structure, hydrophobicity and polarity depiction were all carried out using the internal options in Chimera. Chimera was also used to visualise protein-protein interactions after protein docking which was carried in the later stages of this workflow.

2.3.2.4 Protein-protein interactions

Protein-protein interactions were studied using protein docking. AutoDock Vina was used for protein docking [23]. Free energy change (ΔG) was used as an indicator for judging the quality of interactions. The negative sign indicates the energy released during the interaction. Higher numerical values of ΔG reflect stronger interactions. All the bound conformations were visualised in UCSF Chimera to screen for conformations that bind at the active sites (epitope-paratope interaction).

2.3.2.5 Theoretical structure validation

Ramachandran plots (RC plots) were used to validate the theoretical stability of the modeled structures. RC plots inform the percentages of residues that occur in the theoretically favored, allowed and disallowed regions. This information could be used to verify the structural stability of the model to exist in a natural environment.

2.3.2.6 Total hydrophobicity vs Surface hydrophobicity

Hydrophobicity of the 3D models explains the integrity of the conformation in a water based medium. Large sections of hydrophobic residues on the surface of the structure result in structural deformation when dissolved in an aqueous medium. Regular hydrophobicity calculators inform the summative hydrophobicity of the entire protein. However, the hydrophobic residues deprived of external exposure do not contribute to the instability. Surface hydrophobicity was mathematically calculated by identifying the

residues which have over 40 % exposure. The hydrophobicity is then calculated for these surface residues using in-house R algorithms.

2.3.2.7 Structural remodeling and affinity improvements

Improvements to the backbones, linkers and single residue replacements required structural remodeling. This was performed using Rosetta package. In cases where the modification is very small, the specific region is remodeled instead of modeling the whole structure.

2.3.3 DNA design

Following the *in silico* validation of all the test sequence. The finalised constructs were reverse translated into their corresponding DNA sequences using backtranseq feature on EBI website (https://www.ebi.ac.uk/Tools/st/emboss_backtranseq/). The DNA sequences were codon optimised for CHO cells using the codon optimisation tool available on IDT website (https://eu.idtdna.com/codonopt). The final constructs were obtained from Twist Bioscience company. Due to fragment size restriction on DNA synthesis, some of the constructs were synthesised in multiple parts. NEB and SnapGene's Gibson assembly simulators were used to design the homologous arms to facilitate Gibson assembly. Amplification and sequencing primers were designed using Benchling's primer design tool. The primers were cross verified using Primer3Plus. All primers were sourced from Integrated DNA Technologies.

Primer name	Sequence
FullSeqFWd	AATTCAAAGGAGGTACCCACCA
FullSeqREV	AGGTAGATATCGCGGTACCCTTA
StaphDia1aREV	GTG ATA CTA AGG CTT TGA GAA GGT
StaphDia1bFWD	TCT CAA AGC CTT AGT ATC ACT TGT GC
StaphDia3aREV	CAA GCG AGG TAG TTC TTT TGA
StaphDia3bFWD	CAA AAG AAC TAC CTC GCT TGG
StaphDia4aREV	TTT TGT TGA TAC CAT GCC AAA T
StaphDia4bFWD	TTG GCA TGG TAT CAA CAA AAG
StaphDia5aREV	GAG TTC ATC TTC AAA AAG ACC TGT G
StaphDia5bFWD	GTC TTT TTG AAG ATG AAC TCT CTG C

Table 1.1: Primer sequences

2.3.4 Plasmid scale-up

OG176 plasmid (Oxford genetics) with Kanamycin resistance was chosen for producing the synthetic proteins. *E. coli* BL21 cells were used for plasmid scale-up. *E. coli* BL21 Cells were made competent using Cohen et al. 1972 protocol. 100 ng OG176 was mixed with 30 µl competent *E. coli* BL21 cells and were placed on ice for 20 min. The suspension was subjected to heat shock at 42◦Cfor 20 min. The cells were again placed on ice for 2 min and 1 ml of LB broth was added. 100 µl of the transformed cells were plated on LB agar containing 50 µg/ml Kanamycin. The colonies were then subcultured and stored in -80 ◦Cfor further use. For plasmid extraction, overnight subcultures of the transformed bacteria are subjected through Monarch Plasmid miniprep kit (New England Biolabs) protocol.

2.3.5 Restriction digestion

OG176 was digested by restriction enzyme Nco1 HiFI with CutSmart reaction buffer (NEB) at 37 ◦C for 1 h. Manufacturer's protocol was followed to adjust the reaction volumes as per the need. Following the restriction digest, the DNA was purified using PCR purification kit (Qiagen) protocol. In all cases, the restriction digest was verified by Agarose gel electrophoresis and the DNA concentration was determined using a NanoDrop spectrophotometer (Thermofisher).

2.3.6 Gibson Assembly

Gibson Assembly was carried out using the Gibson Assembly master mix described by DG Gibson et al (2009). The plasmid and DNA gene blocks were mixed in 1:3 ratio in a Gibson Assembly master mix and incubated at 50 ◦C. *E. coli* BL21 cells were transformed with the assembled plasmids and plated on LB agar medium.

2.3.7 Validating cloning using colony PCR and Sanger sequencing

The selected colonies were added to NEB PCR master mix with 2.5 µl of corresponding primers. PCR was carried out as per NEB Q5 polymerase PCR protocol. Sanger

sequencing (GATC light-run) was also performed on the selected colonies to confirm the assembly.

2.3.8 *In vitro* transfection

CHO K1 cells were transfected using Turbofect transfection reagent (Thermofisher). Manufacturer's protocol was used to perform transfection. Supernatant from the cells were collected at 24 h and 48 h intervals.

2.3.9 Binding assays

S. aureus TCH959 (naturally bearing ClfA) cells (10^8 cells per sample) were blocked with 5 % BSA for 2 h followed by incubation with supernatant containing each test construct. Cells were washed 3 times and resuspended in PBS. 10 µl of each sample is taken in triplicates into a corning 96 well white plate. 50 µl of Coelenterazine substrate was added to each well and luminescence was measured using Promega GloMax® 96 luminometer.

2.4 RESULTS

2.4.1 Design elements and design rationale

Targeted luminescence is the overall function of these synthetic proteins. Several test variations of the synthetic proteins were made during the design phase. Heavy and light chains of anti-ClfA antibody MAb 12-9, were used as the binding site (T-domain). A monospecific monovalent monobody (single chain variable fragment (ScFv)) and a monospecific bivalent diabody versions were chosen as the binding site variants. The choice of subparts and the design rationale are shown in Table 2. A secretion signal has been placed to allow the protein secretion into the supernatant. This eliminates the need for laborious downstream processing after protein synthesis. GLuc's native secretion peptide was used in all the test constructs. A Flag tag has been added to aid troubleshooting if the luminescence domain doesn't function. GLuc was used as a luminescence reporter. The positioning of Gluc part differs between the test variants. Trx tag was used as a solubility enhancer. The presence and absence of the tag changes between the test variants. Figure 2.6 showed various subparts of an antiClfA synthetic protein.

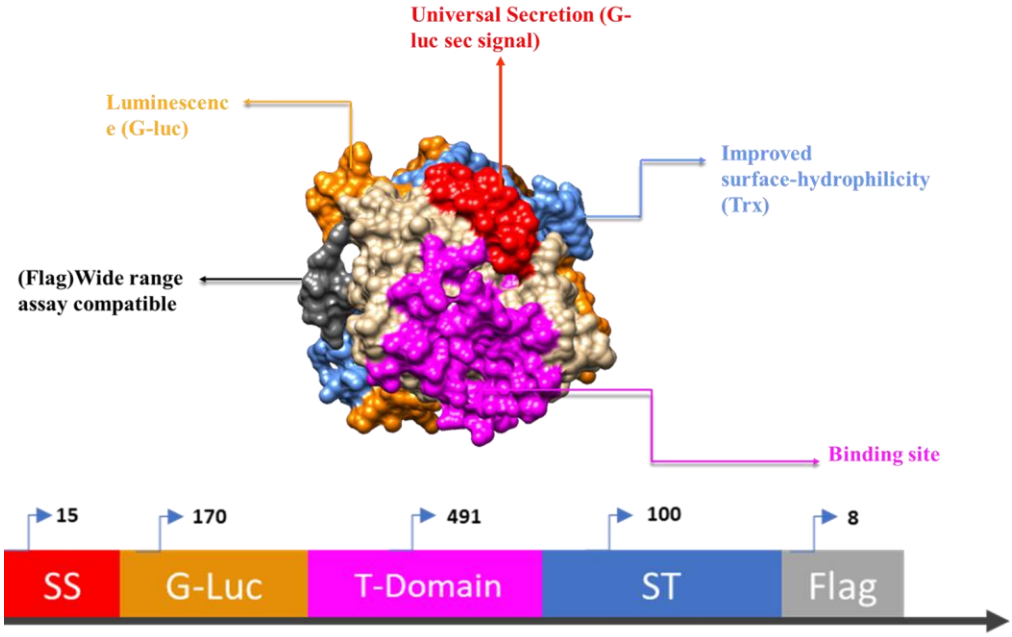


Figure 2.6: 3D model showing different subparts of a synthetic protein. *The Targeting domain (T-domain) coloured in magenta, is either a ScFv version or a diabody version. The bar depiction shows the arrangement order and the amino acid length of each domain. The arrow mark indicates the direction from the N terminus to the C terminus of the protein. (SS = Secretion signal, ST = Solubility tag)*

Design elements	Commonly used	Notes on reviewed elements for the design	Comments on the chosen elements	Design Variants
Secretion Signal Peptide	**GLuc SP, Kappa light chain SS**	**Gluc SP* Eukaryotic**		G-luc signalling peptide is preferred for its greater reliability, greater efficiency.
Linkers	**$(GGGGS)_3$, $(Gly)_8$, $(EAAAK)_3$, Di-Sulfide linkers, PAPAP**	**$(GGGGS)_3$ – Flexible - smaller size of the amino acids provides flexibility, hydrophilic, suitable for ScFv construction** [24-28] **$(EAAAK)_3$ – Rigid – Forms alpha helix structures, keeps the distance between the two domains and to maintain their independent functions Can separate functional domains more effectively than the flexible linkers** [24, 29]	**$(GGGGS)_n$** was chosen where flexible linker was needed. This is rich in hydrophilic amino acids and allows the interaction between the domains due to its flexible nature [24]. Thus, increasing the stability and folding [24, 25, 30]. **$(EAAAK)_n$** is chosen where the domains were supposed to stay rigid with reduced/no interaction between them, increased expression, **Diabody only*** - A combination of the above has been used to put the V_H and V_L of the same chains separate but allowing the two different ScFvs to interact forming the diabody complex.	• Only with flexible linkers varying in number of $(GGGGS)_n$ • Only with rigid linkers with varying in number of $(EAAAK)_n$ • Alternating flexible and rigid linkers between the two variable chains and between the two ScFv fragments • Flexible linker between the VH and VL, with one rigid 3 flexible and one rigid between the ScFv fragments • Without linkers • Linkers between each distinct element such as secretion tag-ScFv-solubility tag-detection tag
Binding domain	**VH fragment, ScFv fragments, Diabody, Minibody, Antibody**	**Smaller have faster blood and renal clearance**		ScFv and Diabody versions were modelled. The test variants differed in the order of arrangements.

Imaging element	**Gluc**		**Gluc** was chosen for its **small size** and that is **secreted out** of the cells	All the design variants have Gluc domain. The arrangement of the domain order differs between the constructs
Solubility Tag	**MBP, GST** **Trx, NusA, SUMO** [31, 32]	• **MBP- Huge size**[33, 34]**, may change fusion protein structure, doesn't require affinity tag, regarded as one of the best solubility tags**[35, 36] • **Trx** – **Small size, highly soluble**, **heat stable,** requires affinity tag[31, 37] **One of the most used Tags** • **SUMO** – **small size, tight and rapid folding soluble structure**[38]**, available for bacterial, yeast, insect and mammalian systems** [39] **Best for N-terminal fusion** [38] **contain His6 Tag**	**SUMO* -** Chosen for its small size	Models with +/- **SUMO** placed **both at N and C terminus** of the POI

Detection/Purification Tag	**Poly-His Tag, Flag, c-Myc** [40, 41]	• **Flag – Short, readily available commercial assays and is more hydrophilic than His** [42] • **His – Most common purification tag, short, commercially available assays, denaturing purification is possible** [43] **not preferred for antibody detection** [43]	**Both Flag and Poly-His** are considered for **computational modelling**	• Positioning of the tag was modelled **both at N and C terminus of the POI**, C terminus might show better results because signal peptide is at N-terminus[32]

Table 1.2: Choice of subparts and design rationale for targeted synthetic proteins against ClfA

2.4.2 *In silico*-aided screening

In silico screening involved rationally eliminating test variants with high chances of failing. For a synthetic protein with multiple subparts, each subpart serves for a defined function. The choice of the individual subpart depends on the overall function intended. In this case of designing proteins targeting *S. aureus* surface antigen ClfA, Table 1.2 provides the decision rationale for selecting individual subparts. This is the first step in funnelling down the sample space of total possible test variants.

Once the subparts are chosen, as mentioned in earlier, a synthetic protein with 'n' number of parts would have n! number of ways into which it could be assembled. All the logically possible assembly combinations have been modelled using I-Tasser. Multiple iterations of test construct variants have been generated while optimising linker types and sizes. During this process test variants were compared and noted for their structure quality, proper exposure of relevant tags and display of the paratope. Protein modeling and visualisation helped screening of several failed constructs. For example, in the model shown in Figure 2.7a the secretion tag is buried deep inside. This could potentially result in improper secretion. By increasing the linker sizes, the model on shown in Figure 2.7b shows an exposed secretion tag. Similar design corrections were made to ensure all the relevant domains are exposed appropriately. Solvent accessibility was used as an empirical measure for assessing proper exposure of a segment on the protein. For example, test variants with less than 60% exposure of an active site were screened out.

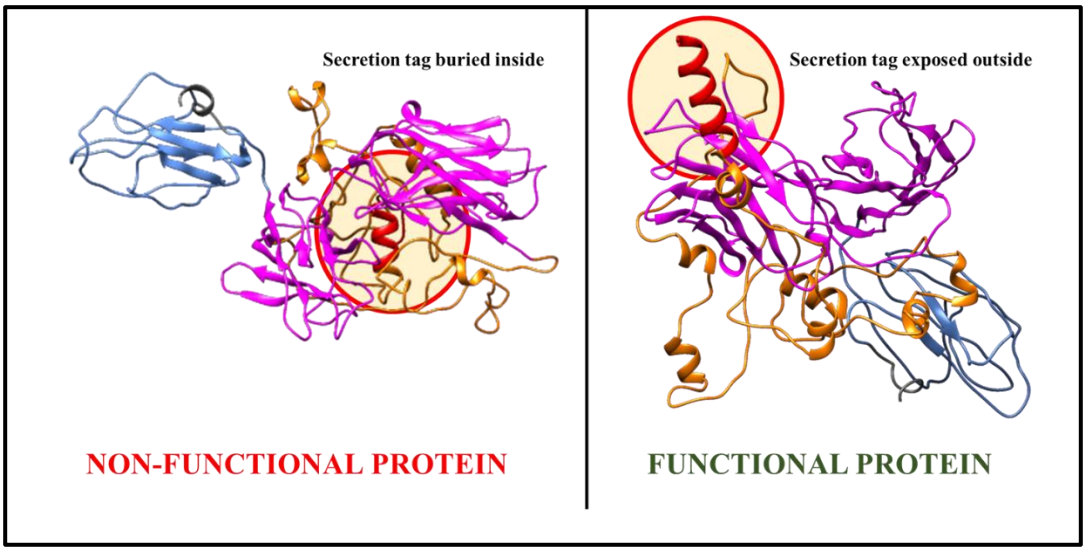
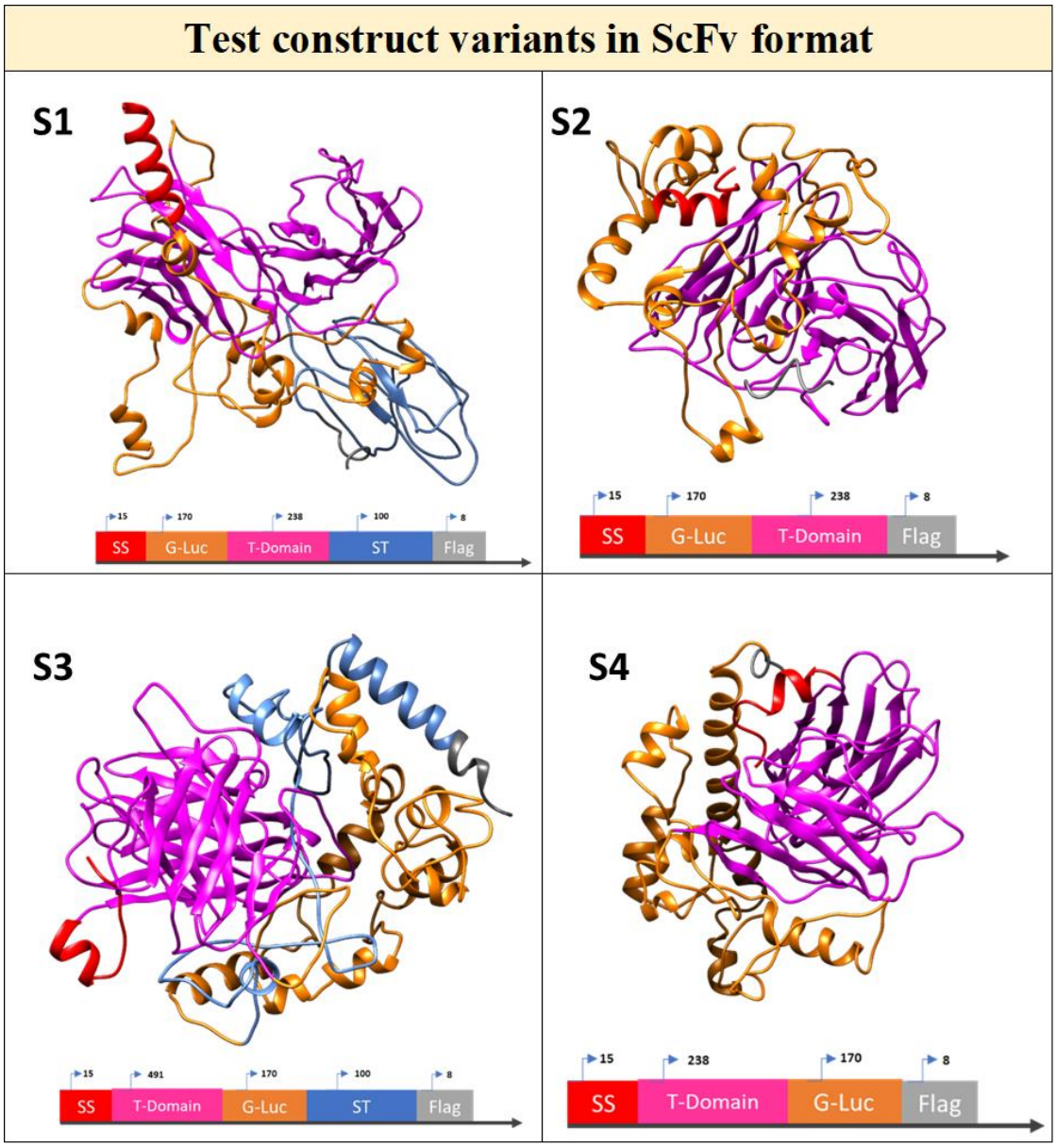


Figure 2.7: Informed screening using protein modelling and visualisation. *(a) Showing the secretion tag in red, buried inside the structure. (b) Increasing the size of the linkers produced a conformation with appropriate exposures of relevant domains. (The above structures are obtained from protein modeling using I-Tasser)*

2.4.3 Test constructs for wet lab testing

4 ScFv variants and 4 diabody variants were screened for wet lab synthesis and testing. 2 test constructs were designed without the T-domain (binding site) which would act as internal negative controls for binding and as a positive control for luminescence. All the final test constructs are shown in Figure 2.8.


Test construct variants in ScFv format
S1
15
170
238
100
8
SS
G-Luc
T-Domain
ST
Flag
S2
15
170
238
8
SS
G-Luc
T-Domain
Flag
S3
15
491
170
100
8
SS
T-Domain
G-Luc
ST
Flag
S4
15
238
170
8
SS
T-Domain
G-Luc
Flag

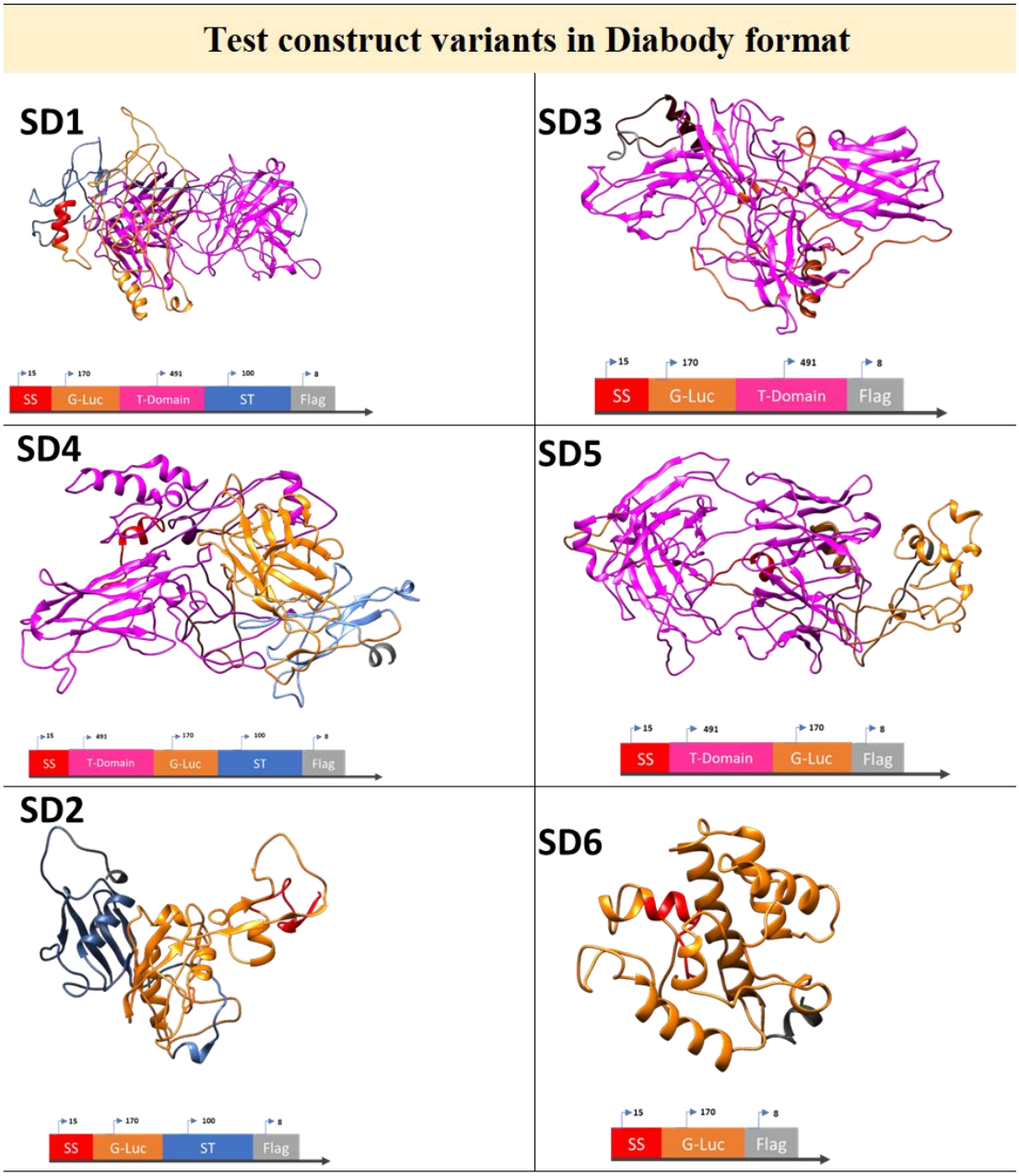


Figure 2.8: Test constructs for wet-lab testing. *SD2 and SD6 lack the binding domain and were designed as internal negative control for binding. Constructs S1, S2, S3 and S4 represent the ScFv version and the constructs SD1, SD3, SD4 and SD5 represent the diabody format. All the structures are obtained by computational modeling using I-Tasser.*

2.4.4 Protein-Protein Docking

Amino acid sequence and structural information of ClfA were obtained from the RCSB PDB server. The structure of ClfA was modeled using I-Tasser (Figure 2.9.)

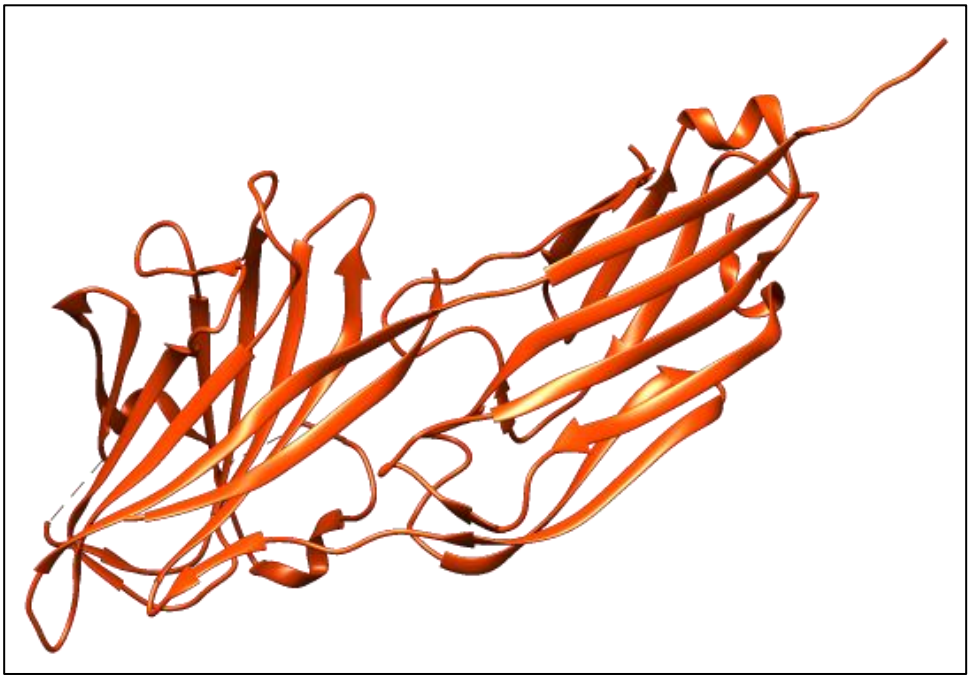


Figure 2.9: 3D structure of ClfA *(PDB ID - 1N67). The minimal binding segment of S. aureus ClfA, containing two similarly folded domains is shown here.*

In silico docking was used to test the interaction of the test constructs with ClfA. The ClfA structure modelled by I-Tasser was used as the receptor. Active site on ClfA was highlighted and the boundaries for binding were defined. Upon docking, the highest free energy conformations were noted. Figure 2.10 shows the ScFv S3 bound to ClfA. The docking results and all the *in silico* data corresponding to the finalised 10 constructs are shown in Table 3.

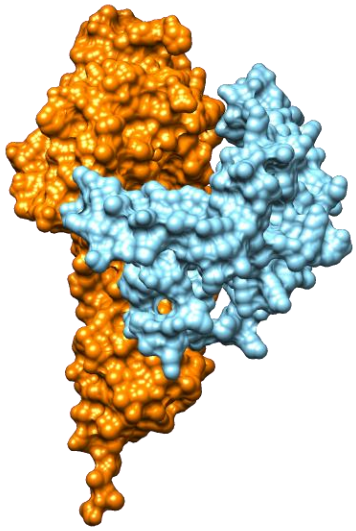


Figure 2.10: S3 ScFv (blue) docked with ClfA (red). *Free energy change for the interaction was found to be (ΔG) -18.3. Protein docking was performed using AutoDock Vina and the assembly was visualised using UCSF Chimera.*

ID	C Score	TM score	RC score (%)	Hydrophobicity (%)	Solvent accessibility of the active site (%)	(ΔG)	Size (kDa)	Instability (%)
SD1	0.045	0.11	74.9	46.11	67.7	-13.6	99.03	38.42
SD2	0.167	0.19	89	45.33	86.6	-11.7	37.53	30.92
SD3	0.182	0.14	80.2	46.88	64.4	-15.2	85.83	35.97
SD6	0.077	0.212	86.9	47.77	73.3	-11.2	24.33	18.42
S1	0.057	0.24	78.1	45.88	60	-15.1	67.67	36.21
S2	0.114	0.16	86.2	47.11	65.5	-15.4	54.47	31.08
S3	0.071	0.162	82.8	46.66	64.4	-18.3	67.67	35.39
S4	0.147	0.205	86.5	47.11	58.8	-16.9	54.47	31.08

Table 1.3: In silico data obtained from various computational tools. *From left to right, C-score (confidence score) from I-Tasser, TM score (template modeling score) as agreement between Rosetta and I-Tasser models, RC score (Ramachandran plot score), solvent accessibility of the active site (paratope regions), Free energy change from docking, size and instability index were tabulated for all the test variants.*

2.4.5 Wet-lab experimentation

2.4.5.1 Protein production and secretion

To validate the production of all the test constructs, luminescence being emitted from CHO cell supernatant was measured. 10 µl of each test construct was taken in a white corning 96 well plate. Supernatant media from a batch of non-transfected cells was used as a negative control. GLuc protein (NanoLight technologies) was used as a positive control. Figure 2.11 shows the luminescence from various constructs. The concentrations of the test constructs were calculated using the standard curve obtained from Gluc protein standards. As expected, both the control proteins SD2 and SD6 were the best produced. In the test constructs with T-domain, ScFv test variant S1 was the highest produced.

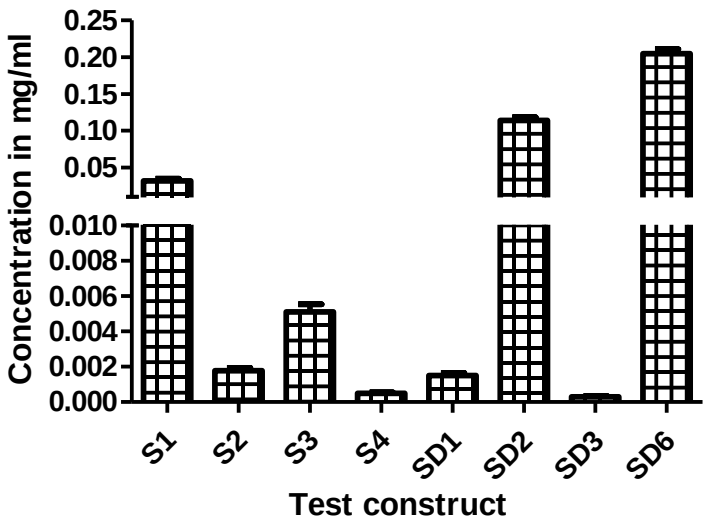


Figure 2.11: Secretion of various test constructs. *Relative luminescence units were corelated to concentration in mg/ml using a standard curve obtained from Gluc protein standards. SD2 and SD6 lack the T-domain and were used as internal controls. All the measurements were taken in triplicates.*

2.4.5.2 Binding to *S. aureus* ClfA

In vitro binding was confirmed by measuring and comparing luminescence signals after binding. 10^8 *S. aureus* cells were treated with 1000 ng of each test construct. Experiments were performed on fresh subcultures of *S. aureus* to avoid the exopolysaccharide formation. Bound luminescence from each test construct is plotted in Figure 2.12. *S. aureus* cells only (without synthetic protein) were considered to measure the background luminescence. The construct SD2 and SD6 had the lowest bound luminescence signal. Both the constructs lack the binding domain and hence cannot bind to ClfA. ScFv S1 showed the highest luminescence signal. The data were plotted again in different graph format to determine the difference between ScFvs vs Diabodies, but no significant correlation was observed. The data set size was also too small for such comparisons. With S1 outperforming all the other test constructs, all the further emphasis was placed in S1.

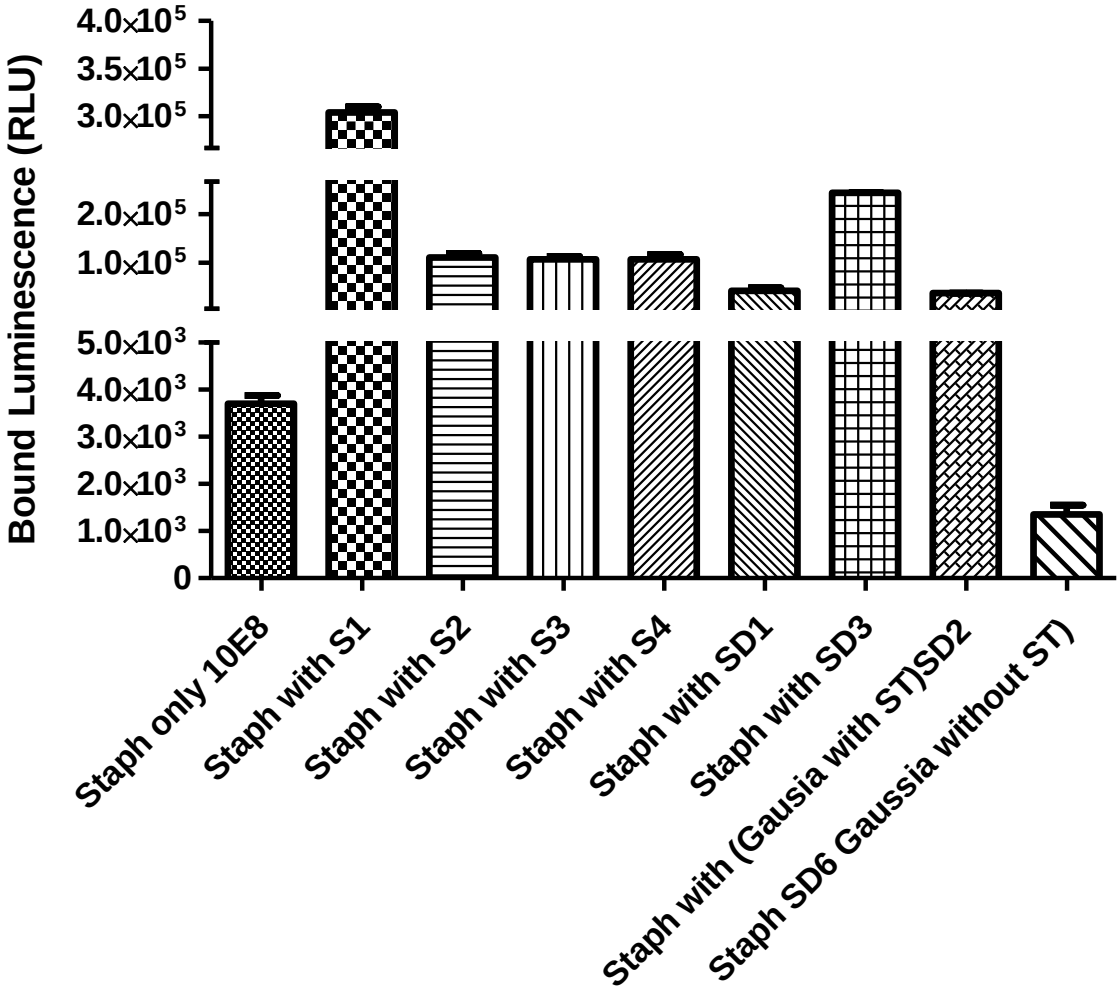


Figure 2.12: Bound luminescence. *Luminescence after binding to ClfA on S. aureus.* 10^8 *S. aureus cells were incubated with 1µg of each test construct for 1 h. Cells without treatment were used as a background. SD2 and SD6 were used as negative controls for binding. S1 (ScFv) emitted highest bound luminescence. All the measurements were taken in triplicates.*

2.4.5.3 Dose response experiments

Dose response experiments were performed using two Gram-positive and two Gram-negative strains. Three different concentrations of the synthetic proteins and three different bacterial concentrations were used during optimisation.

2.4.5.3.1 Synthetic protein dose response

3 different concentrations of the test constructs were tested for their binding to ClfA. 0, 0.50µg and 1µg of test construct were incubated with 10^8 cells of *S. aureus* 959. Bacterial cells were washed 3 times and resuspended in PBS. All the samples in triplicates were placed in a corning 96 well white plate. Luminescence was measured after adding 50µl of coelenterazine. *E. coli* DH5alpha, *Salmonella enterica* Typhimurium *7207* and *Streptococcus agalactae (GBS)* strains were used as negative controls to measure selectivity towards *S. aureus*. Optimisation was carried out on all test constructs using the above mentioned 4 bacterial strains. ScFv S1 construct was again the best performer. The results of the synthetic protein dose response of S1 were shown in Figure 2.13. The selectivity of binding in Figure 2.13b is the ratio between luminescence from S1 bound to S. aureus and *Streptococcus agalactae.* The data in Figure 2.13 clearly shows the selectivity of S1 towards *S. aureus.*

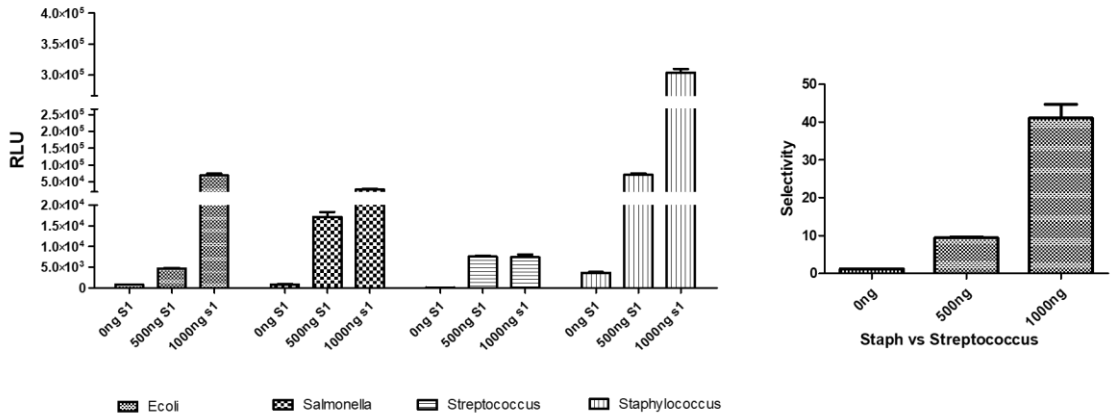


Figure 2.13: Synthetic proteins dose response of S1 with 10^8 bacteria: *a. shows the relative luminescence units from four different bacterial strains treated with 3 different concentrations of S1. b. shows the selectivity of S1 towards S. aureus when compared with Streptococcus (Gram-positive).*

2.4.5.3.2 Bacterial cell dose response

Three different concentrations of four bacterial strains (two gram-positive and two gram-negative) were treated with 1000ng of each test construct. Bacterial cell dose response was carried out on *S. aureus* 959. *E. coli* DH5alpha, *S.* Typhimurium and *S. agalactae (GBS)* strains. Bacterial cells were washed 3 times and resuspended in PBS. All the samples in triplicates were placed in a corning 96 well white plate. Luminescence was measured after adding 50µl of coelenterazine. ScFv S1 showed the highest luminescence signal once again. Figure 2.14 shows the bacterial cell dose response for 1000 ng of S1.

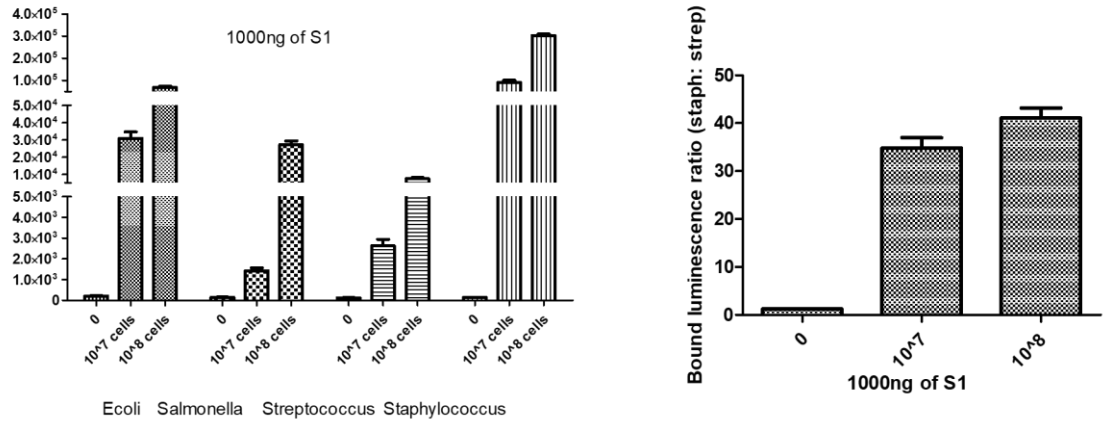


Figure 2.14: Bacterial cell dose response of S1 *(a). Relative luminescence units from four different bacterial strains with 3 different concentrations of bacterial cells. (b) Difference in signal intensity of the luminescence emitted by S1bound to S. aureus when compared with S1 bound to S. agalactae.*

2.4.5.4 Blocking with human fibrinogen

To confirm the binding of S1 specifically to ClfA, the *S. aureus* cells were incubated with different volumes of human fibrinogen to block ClfA. 10^8 S. aureus cells were incubated with 0, 10 and 250 µg of human fibrinogen for one hour at room temperature. The cells were washed 3 times with PBS and then incubated with 10 µl of S1 supernatant, 1 h at room temperature. Bacterial cells were washed 3 times and resuspended in PBS. All the samples in triplicates were placed in a corning 96 well white plate. Luminescence was measured after adding 50 µl of coelenterazine. No binding was expected upon addition of S1. The luminescence readings were plotted in Figure 2.15. As expected, the increase in fibrinogen concentrations led to decrease in luminescence. Fibrinogen successfully ensured restricted access to ClfA. This confirms the specificity of S1 towards ClfA.

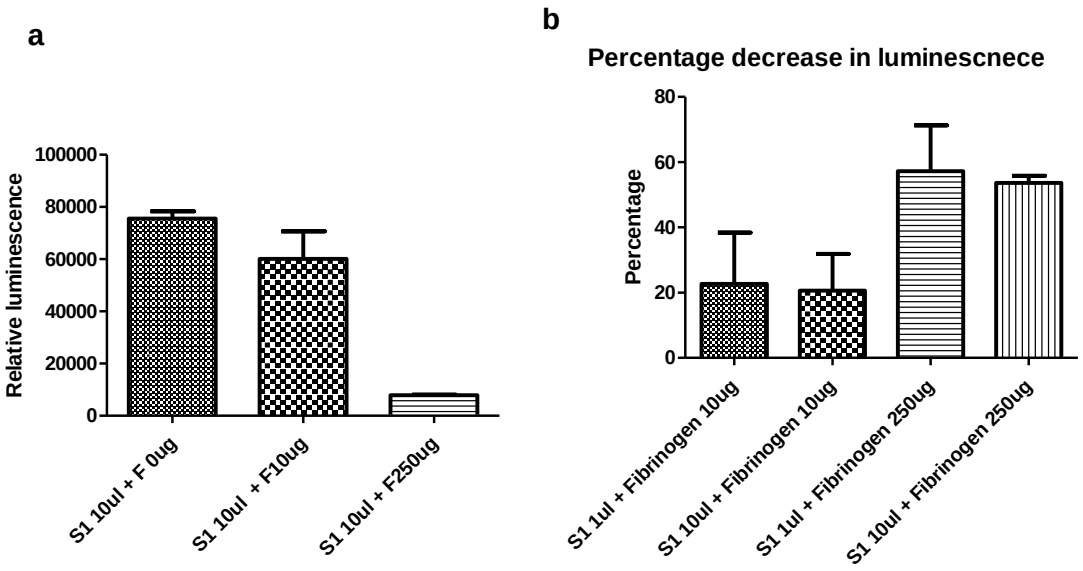


Figure 2.15: Blocking with human fibrinogen. *(a) Shows decreasing luminescence intensity with increase in fibrinogen concentration. (b) showing percentage decrease in signal upon fibrinogen addition. 250 µg of fibrinogen resulted in 68 % decrease in luminescence.*

2.4.5.5 Optimal performer (Selectivity + Intensity +Normalised performance)

After the wet-lab validation of all the 10 test variants, it was intended to carry out an *in vivo* imaging study using the best performing test construct. The choice of the best construct depends on overall performance. Overall performance can be defined as the degree to which the designed protein would perform ultimately on the user defined function(s) (Figure 2.16).

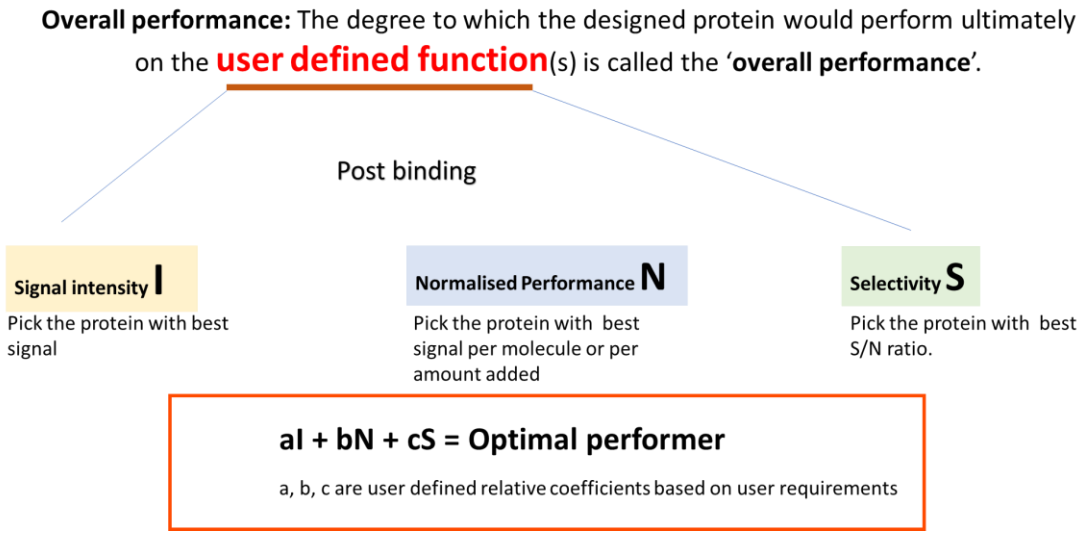


Figure 2.16: Schematic showing the optimal decision process for choosing an optimal performer.

Although all the test constructs performed the user defined function, they differed in their performance quality. Arbitrary numbers were given to the coefficients reflecting the relative importance of each parameter. However, in this case S1 was outperforming all the other test constructs. After choosing S1 as the best performer, keeping in mind the demands of *in vivo* imaging, efforts have been made to improve the signal intensity of S1.

2.4.5.6 Nanoluc as the luminescence part

Previous literature has reported that an engineered luciferase obtained from a deep-sea shrimp (*Oplophorus gracilirostris*) (Nanoluc) has better brightness than Gluc *in vivo [44]*. In an effort to improve the brightness of S1, the Gluc domain was swapped with Nanoluc, and a new Nanoluc version of S1 (NS1) was synthesised. Synthetic protein dose response and bacterial cell dose response for NS1 were examined using *S. aureus* as the target and *E. coli* as a negative control. 20 µl of protein supernatant was added to each sample and incubated for 1 h at room temperature. Luminescence was measured after adding 50 µl of coelenterazine. After 3 subsequent PBS washes, the samples were taken into a 96 well corning white plate. The data for the dose response experiments are plotted in Figure 2.17. However, the brightness (RLU) in *in vitro* experiments of NS1 was found to be significantly lower than Gluc S1.

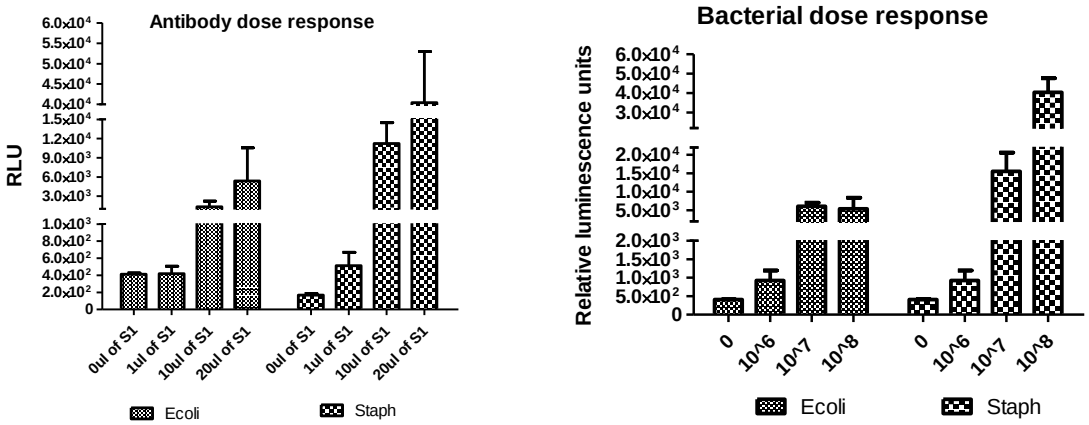


Figure 2.17: Synthetic protein dose response and bacterial cell dose response of NS1 *(a) Synthetic protein dose response was carried out on 10^8 bacterial cells of S. aureus and E. coli. The bacterial cells were treated with varied volumes of the protein supernatants. (b) Bacterial cell dose response was carried out using 0, 10^6, 10^7 and 10^8 cells. The luminescence signal from the S. aureus samples were significantly higher than that of E. coli samples, indicating binding to S. aureus. The dose response experiments showed an increase in signal intensity with respect to protein/cell concentrations.*

2.5 DISCUSSION

In this chapter, a workflow was generated for the 'design, build and test' of synthetic proteins. Various computational tools were used to ensure proper functioning of the subparts. Such an approach validates the structure before wet lab synthesis and testing. Such a computationally informed design strategy would empower wet lab biologist with prediction capabilities. In my cases during the process, modeling helped to realise pitfalls in the design. One such example was shown in Figure 2.7, where the secretion tag was hidden inside. Similar non-functional designs were also spotted with the Flag tag. Both the secretion tag and the Flag tag are located at the extremes of the polypeptide chain. It is crucial for the tags to remain exposed for their full functionality. Without protein modeling, such failed experiments end up in the wet lab validation and cause time and capital wastes.

In this work, over 200 different test variants were designed and modelled. Over 50 test variants were subjected complete computational analysis. This process is analogous to high throughput screening that is observed in wet lab drug design. However, the *in silico* screening only takes a fraction of the time to test the constructs in the wet lab. This work heavily relied on computational tools. However, it must be noted that the accuracy of each computational tool is subjective to each protein. Predictions of synthetic proteins that resemble close similarity to naturally existing proteins have higher confidence levels and accuracy. Community-wide, worldwide studies such as CASP (Critical Assessment of protein Structure Prediction), CAPRI (Critical Assessment of PRediction of Interactions) and CAFA (Critical Assessment of Functional Annotation), organise regular blinded challenges to compare the performance of various computational tools for proteins [45-47]. With artificial intelligence, neural networks, deep learning and increased computational power, the road ahead is promising. The strategy presented here will grow in accuracy as the individual tools get updated.

The aim of this chapter was to develop a targeted report protein specific for *S. aureus*. ClfA was chosen as a test model for the *in silico* aided synthetic protein design strategy. Abundance of literature on the structure of ClfA, ready availability of the amino acid sequences of ClfA and its targeting monoclonal antibody tefibazumab (also known as mAb 12-9 and Aurexis) encouraged the pursuit in this direction. Although, mAb 12-9

recognizes ClfA on *S. aureus* cells with a high affinity the antibody has failed in the clinical trials (no significant advantage over a placebo was observed in relapse of bacteremia) [48]. *S. aureus* employs various virulence and immune evasion factors to survive in the host. This could be one of the prime reasons for the failure of the antibodies targeting a surface antigen like ClfA. Thus, targeting multiple virulence factors could be a strategy to explore [49].

In vitro testing was carried out to validate secretion and functioning (binding and luminescence) of the synthetic proteins. Throughout the work, luminescence was used to inform the protein production, binding and selectivity. The S1 test construct clearly stood as the best performer amongst all other test construct variants. With the encouraging results observed in the dose response assays, S1 was identified as a candidate worthy of examining in a future *in vivo* setting. Although Gluc is a widely used as an imaging module in preclinical research, the issues with signal intensities and tissue absorption still remain. Hence the Gluc was replaced with Nanoluc, to increase the signal intensity. The Nanoluc system has been shown in previous literature, to have a significantly higher brightness and prolonged half-life, when compared to Gluc[50]. *In silico* modelling and computational validation was repeated on NS1 construct, prior to synthesis. In this study, the luminescence signal from NS1 was observed to be lower than its Gluc counterpart. Identifying the reason for this behaviour was beyond the scope of this project. However, as previously studied in literature, the luminescence emitted from Nanoluc lasted significantly longer than luminescence from Gluc. This might be a key feature that would score Nanoluc as a better alternative to Gluc in *in vivo* applications.

The *in vitro* studies in this work, provided a preliminary understanding of the basic functioning of the protein and validated the *in silico* design strategy. The luminescence and binding assays validated the functioning of each subpart on the test construct and its functioning as intended. With the aim of validating protein sizes, integrity and concentration, Western blots using anti-Flag antibodies were performed, but to no avail, despite multiple attempts (data not shown). Similar anti-Flag Western blots were readily achieved in Chapter 5 work. The failure to detect these synthetic proteins by Western blot could be due to some reasons unique to these synthetic proteins. Further wet-lab assays

such as FACS using anti-Flag fluorescent antibody would also be beneficial to reconfirm the binding of the synthetic protein to ClfA.

***In silico* myriad problem** Table 3 shows a summary of numerical evaluation performed using various computational tools. Each individual tool informs the quality of a design parameter. For example, the RC plot provides an RC score, modeling provides the C score etc. These design parameters relate to the functioning of each individual subpart. While screening for the best performer, a holistic overall performance should be considered rather than functioning of individual subparts. In mathematics, this is a common problem observed in cases such as buying a house or buying a car. A decision must be taken on the holistic level rather than a subpart level. There is no tool available to predict the overall performance of test construct. Such a tool would be pivotal in *in silico* -aided screening (see Chapter 3).

2.6 CONCLUSION

Over 50 different multi-part test constructs targeted to *S. aureus* surface antigen ClfA were modelled and validated using various computational tools to inform and guide downstream wet-lab experiments. Following *in silico* design and analyses, 10 test constructs against ClfA were produced in CHO cells and tested for specific target binding *in vitro*. This study demonstrated the proof-of-concept of the design-model-build-test strategy of targeted synthetic proteins, using *S. aureus* surface antigen ClfA as a model target. Finally, a target-specific synthetic protein platform for bacterial imaging has been validated. The results from this chapter encouraged to pursue the *in silico*-aided design in further chapters.

2.7 REFERENCES

1. Adams, G.P. and L.M. Weiner, *Monoclonal antibody therapy of cancer.* Nat Biotechnol, 2005. **23**(9): p. 1147-57.
2. Bates, A. and C.A. Power, *David vs. Goliath: The Structure, Function, and Clinical Prospects of Antibody Fragments.* Antibodies (Basel), 2019. **8**(2).
3. Li, Z., et al., *Influence of molecular size on tissue distribution of antibody fragments.* MAbs, 2016. **8**(1): p. 113-9.
4. Thurber, G.M., M.M. Schmidt, and K.D. Wittrup, *Antibody tumor penetration: transport opposed by systemic and antigen-mediated clearance.* Adv Drug Deliv Rev, 2008. **60**(12): p. 1421-34.
5. Nelson, A.L., *Antibody fragments: hope and hype.* MAbs, 2010. **2**(1): p. 77-83.
6. Olafsen, T. and A.M. Wu, *Antibody vectors for imaging.* Semin Nucl Med, 2010. **40**(3): p. 167-81.
7. Lowy, F.D., *Staphylococcus aureus infections.* N Engl J Med, 1998. **339**(8): p. 520-32.
8. Foster, T.J., et al., *Adhesion, invasion and evasion: the many functions of the surface proteins of Staphylococcus aureus.* Nat Rev Microbiol, 2014. **12**(1): p. 49-62.
9. Thammavongsa, V., et al., *Staphylococcal manipulation of host immune responses.* Nat Rev Microbiol, 2015. **13**(9): p. 529-43.
10. Crosby, H.A., J. Kwiecinski, and A.R. Horswill, *Staphylococcus aureus Aggregation and Coagulation Mechanisms, and Their Function in Host-Pathogen Interactions.* Adv Appl Microbiol, 2016. **96**: p. 1-41.
11. Tong, S.Y., et al., *Staphylococcus aureus infections: epidemiology, pathophysiology, clinical manifestations, and management.* Clin Microbiol Rev, 2015. **28**(3): p. 603-61.
12. Hawiger, J., et al., *Identification of a region of human fibrinogen interacting with staphylococcal clumping factor.* Biochemistry, 1982. **21**(6): p. 1407-13.
13. Chuang, K.-H. and T.-L. Cheng, *Noninvasive Imaging of Reporter Gene Expression and Distribution In Vivo.* Fooyin Journal of Health Sciences, 2010. **2**(1): p. 1-11.

14. Morrissey, D., G.C. O'Sullivan, and M. Tangney, *Tumour targeting with systemically administered bacteria.* Curr Gene Ther, 2010. **10**(1): p. 3-14.
15. Baban, C.K., et al., *Bacteria as vectors for gene therapy of cancer.* Bioeng Bugs, 2010. **1**(6): p. 385-94.
16. Cronin, M., et al., *Orally administered bifidobacteria as vehicles for delivery of agents to systemic tumors.* Mol Ther, 2010. **18**(7): p. 1397-407.
17. Badr, C.E. and B.A. Tannous, *Bioluminescence imaging: progress and applications.* Trends Biotechnol, 2011. **29**(12): p. 624-33.
18. Byrne, W.L., et al., *Use of optical imaging to progress novel therapeutics to the clinic.* J Control Release, 2013. **172**(2): p. 523-34.
19. Yang, J. and Y. Zhang, *Protein Structure and Function Prediction Using I-TASSER.* Curr Protoc Bioinformatics, 2015. **52**: p. 5 8 1-5 8 15.
20. Weitzner, B.D., et al., *Modeling and docking of antibody structures with Rosetta.* Nat Protoc, 2017. **12**(2): p. 401-416.
21. Kaufmann, K.W., et al., *Practically useful: what the Rosetta protein modeling suite can do for you.* Biochemistry, 2010. **49**(14): p. 2987-98.
22. Pettersen, E.F., et al., *UCSF Chimera--a visualization system for exploratory research and analysis.* J Comput Chem, 2004. **25**(13): p. 1605-12.
23. Trott, O. and A.J. Olson, *AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading.* J Comput Chem, 2010. **31**(2): p. 455-61.
24. Chen, X., J.L. Zaro, and W.C. Shen, *Fusion protein linkers: property, design and functionality.* Adv Drug Deliv Rev, 2013. **65**(10): p. 1357-69.
25. Argos, P., *An investigation of oligopeptides linking domains in protein tertiary structures and possible candidates for general gene fusion.* J Mol Biol, 1990. **211**(4): p. 943-58.
26. Huston, J.S., et al., *Protein engineering of antibody binding sites: recovery of specific activity in an anti-digoxin single-chain Fv analogue produced in Escherichia coli.* Proc Natl Acad Sci U S A, 1988. **85**(16): p. 5879-83.

27. Trinh, R., et al., *Optimization of codon pair use within the (GGGGS)3 linker sequence results in enhanced protein expression.* Mol Immunol, 2004. **40**(10): p. 717-22.
28. Alfthan, K., et al., *Properties of a single-chain antibody containing different linker peptides.* Protein Eng, 1995. **8**(7): p. 725-31.
29. Arai, R., et al., *Design of the linkers which effectively separate domains of a bifunctional fusion protein.* Protein Eng, 2001. **14**(8): p. 529-32.
30. Hagemeyer, C.E., et al., *Single-chain antibodies as diagnostic tools and therapeutic agents.* Thromb Haemost, 2009. **101**(6): p. 1012-9.
31. Costa, S., et al., *Fusion tags for protein solubility, purification and immunogenicity in Escherichia coli: the novel Fh8 system.* Front Microbiol, 2014. **5**: p. 63.
32. Rosano, G.L. and E.A. Ceccarelli, *Recombinant protein expression in Escherichia coli: advances and challenges.* Front Microbiol, 2014. **5**: p. 172.
33. Fox, J.D., R.B. Kapust, and D.S. Waugh, *Single amino acid substitutions on the surface of Escherichia coli maltose-binding protein can have a profound impact on the solubility of fusion proteins.* Protein Sci, 2001. **10**(3): p. 622-30.
34. Kapust, R.B. and D.S. Waugh, *Escherichia coli maltose-binding protein is uncommonly effective at promoting the solubility of polypeptides to which it is fused.* Protein Sci, 1999. **8**(8): p. 1668-74.
35. Shih, Y.P., et al., *High-throughput screening of soluble recombinant proteins.* Protein Sci, 2002. **11**(7): p. 1714-9.
36. Kohl, T., et al., *Automated production of recombinant human proteins as resource for proteome research.* Proteome Sci, 2008. **6**: p. 4.
37. LaVallie, E.R., et al., *Thioredoxin as a fusion partner for production of soluble recombinant proteins in Escherichia coli.* Methods Enzymol, 2000. **326**: p. 322-40.
38. Marblestone, J.G., et al., *Comparison of SUMO fusion technology with traditional gene fusion systems: enhanced expression and solubility with SUMO.* Protein Sci, 2006. **15**(1): p. 182-9.

39. Panavas, T., C. Sanders, and T.R. Butt, *SUMO fusion technology for enhanced protein production in prokaryotic and eukaryotic expression systems.* Methods Mol Biol, 2009. **497**: p. 303-17.
40. Lichty, J.J., et al., *Comparison of affinity tags for protein purification.* Protein Expr Purif, 2005. **41**(1): p. 98-105.
41. Kimple, M.E., A.L. Brill, and R.L. Pasker, *Overview of affinity tags for protein purification.* Curr Protoc Protein Sci, 2013. **73**: p. Unit 9 9.
42. Terpe, K., *Overview of tag protein fusions: from molecular and biochemical fundamentals to commercial systems.* Appl Microbiol Biotechnol, 2003. **60**(5): p. 523-33.
43. Bornhorst, J.A. and J.J. Falke, *Purification of proteins using polyhistidine affinity tags.* Methods Enzymol, 2000. **326**: p. 245-54.
44. England, C.G., E.B. Ehlerding, and W. Cai, *NanoLuc: A Small Luciferase Is Brightening Up the Field of Bioluminescence.* Bioconjug Chem, 2016. **27**(5): p. 1175-1187.
45. Moult, J., et al., *Critical assessment of methods of protein structure prediction (CASP)-Round XII.* Proteins, 2018. **86 Suppl 1**: p. 7-15.
46. Lensink, M.F., S. Velankar, and S.J. Wodak, *Modeling protein-protein and protein-peptide complexes: CAPRI 6th edition.* Proteins, 2017. **85**(3): p. 359-377.
47. Dessimoz, C., N. Skunca, and P.D. Thomas, *CAFA and the open world of protein function predictions.* Trends Genet, 2013. **29**(11): p. 609-10.
48. Medscape, *Immunotherapeutic Approaches Against Staphylococcus aureus.* 2011.
49. Romero Pastrana, F., et al., *Human antibody responses against non-covalently cell wall-bound Staphylococcus aureus proteins.* Scientific Reports, 2018. **8**(1): p. 3234.
50. Boute, N., et al., *NanoLuc Luciferase - A Multifunctional Tool for High Throughput Antibody Screening.* Front Pharmacol, 2016. **7**: p. 27.

Chapter 3

F2F Bridge: A synthetic protein holistic performance prediction strategy

Work from this chapter has been published as

V. V. B. Yallapragada, S. P. Walker, C. Devoy, S. Buckley, Y. Flores, M. Tangney, Function2Form Bridge-Toward synthetic protein holistic performance prediction. *Proteins*, (2019). PMID: 31589780

Table of Contents

3.1 ABSTRACT 100

3.2 INTRODUCTION 101

3.2.1 The world of *in silico* aided protein design 101

3.2.2 The overall performance problem: 102

3.2.3 Machine learning methods for proteins 105

3.2.4 A novel mathematical strategy using machine learning approaches. 106

3.3 MATERIALS AND METHODS 107

3.3.1 *In silico* design of test sequences: 107

3.3.2 Data Generation: 108

3.3.3 Wet-lab validation 110

3.3.4 Data generation from wet lab experiments 110

3.3.5 Function2Form Bridge: 112

3.3.6 Generating the OP score: 113

3.3.7 Statistical and machine learning methods: 115

3.4 RESULTS: 117

3.4.1 Predicting the 'Overall Biological Performance' - Unsupervised and unweighted F2F. 117

3.4.2 Applying machine learning methods to F2F bridge 121

2.4.2.1 Improving F2F bridge with supervised and weighted machine learning 121

3.4.2.2 Regression analysis, selection of variables and regularisation 121

3.4.3 LASSO (Least absolute shrinkage and selection operator) regression 122

3.4.4 Secreted luminescence as overall performance. 125

3.4.5 Random Forest regression tree analysis as an alternative performance prediction method. 129

3.5 DISCUSSION 136

3.5.1 Relevance to the laboratory scientist 138

3.5.2 F2F Bridge 138

3.5.3 Lasso driven feature selection and linear models 139

3.5.4 Random forest regression tree directed model 140

3.5.5 F2F bridge - Foundation for overall performance prediction and hurdles ahead 140

3.5.6 Outlook and F2F Bridge V.2.0 141
3.6 Conclusions 142
3.7 References 143

3.1 ABSTRACT

Background Protein engineering and synthetic biology stand to benefit immensely from recent advances in *in silico* tools for structural and functional analyses of proteins. In the context of designing novel proteins, current *in silico* tools inform the user on individual parameters of a query protein, with output scores/metrics unique to each parameter. In reality, proteins feature multiple 'parts'/functions, and modification of a protein aimed at altering a given part, typically has collateral impact on other protein parts. A system for prediction of the combined effect of design parameters on the overall performance of the final protein does not exist.

Aims Function2Form Bridge (F2F-Bridge) aims to address this gap by combining the scores of different design parameters pertaining to the protein being analysed into a single easily interpreted output describing overall performance. The strategy comprises 1. A mathematical strategy combining data from a myriad of *in silico* tools into an OP-score (a singular score informing on a user-defined overall performance); 2. The F2F-Plot, a graphical means of informing the wet-lab biologist holistically on designed construct suitability in the context of multiple parameters, highlighting scope for improvement.

Methods & Results F2F predictive output was compared with wet-lab data from a range of synthetic proteins designed, built and tested for this study. Statistical/machine learning approaches for predicting overall performance, for use alongside the F2F plot, were also examined. Comparisons between wet-lab performance and F2F predictions demonstrated close and reliable correlations.

Conclusion This user-friendly strategy represents a pivotal enabler in increasing accessibility of synthetic protein building and *de novo* protein design.

3.2 INTRODUCTION

3.2.1 The world of *in silico* aided protein design

Proteins are multi-functional biomolecules that perform, mediate and regulate various fundamental functions of life. Over the last 50 years our understanding of proteins and our ability to engineer them has improved exponentially. While proteins find their applications in various fields, biochemical and medical applications have taken the driving seat commercially. Since 1982, when insulin, the first recombinant protein was produced the biotechnological way, the market value for protein-based therapeutics has seen a significant increase, with the market value of bioengineered protein drugs is expected to reach $336.9 billion by 2025 [1]. Nature has sampled only a small fraction of the theoretical combinations of amino acids that are accessible to proteins [2] due to the constraints put in place by evolution. Synthetic protein design represents a vast sea of possible space available to be explored. While multiple industries have the potential to exploit non-natural proteins as components of their products or processes, this potential cannot be fully realised without reliable control over protein design.

Understanding the interplay between structure and function of proteins is pivotal in protein design. Techniques such as NMR spectroscopy and X-Ray Crystallography have revealed the structures of over 100,000 proteins and have aided the establishment of protein databases [3], which in turn have facilitated *in silico* protein structure and function prediction based on the amino acid sequence. The need for a faster and affordable way to predict the putative structure of a protein paved the way for computational protein modelling [4]. *In silico*-aided protein design uses computational strategy for designing and building proteins that perform a specific function(s), in a user defined setting. Computational methods for structure modelling, docking and function prediction provided *in silico* alternatives for screening protein sequences, creating variants of a specific design [5] and building new *de novo* structures [6] . As a consequence of increasing computational power, the power of *in silico* protein structure and function prediction, and analysis tools is expanding rapidly.

3.2.2 The overall performance problem:

Multiple related parameters and the myriad of in silico tools

Computational tools are now available to predict protein structure, active sites, chemical properties and interactions with other proteins [7-11]. In some cases, these tools could also be used to redesign existing proteins [12] or even design entirely new proteins, in the rapidly evolving field of *de novo* design [13-16]. Unfortunately, a side effect of these rapid advances is that, it is becoming difficult to bridge the gap between these advances in computational technology and their originally intended wet-lab applications by biologists. 'Outsourcing' the required *in silico* activity by end users (wetlab biologists) entirely to 'dry-lab' specialists dramatically reduces the potential of *in silico* modelling, 'User empowerment' is key to translating this potential.

Designing a protein involves defining an overall function (Box 1) and associating it with a 3D structure which is coded into an amino acid sequence [17]. In most cases, the overall function of a protein is a combination of several individual sub-functions. To achieve the overall function, fusing different sub-function parts (Box1) has been the most popular strategy to date. In recent years, *de novo* protein design has been used to obtain amino acid sequences which fold into a required 3D structure for a defined function. In both of these cases, several test sequences are generated to validate their performance against the defined overall function. *In silico* protein design aims to find a 'best-fit' test sequence for a defined overall function (Figure 3.1).

Box 1: Definitions and commonly used terms

Overall performance: The degree to which the designed protein would perform the defined Overall Function

Sub-function part: A part of a protein that is responsible for a particular sub-function. A Sub-function part may represent a sequence for a specific known protein or a part of a protein which has a defined distinct function.

Design Parameters: Parameters such as hydrophobicity, solubility, structure, active site exposure etc., influence the overall performance of a protein. While these parameters are tools for studying the nature of existing proteins, the same parameters, in protein design, can be used as controllers to dictate the overall performance of a protein. Such parameters that dictate the overall performance of a designed protein are grouped under an umbrella term called 'design parameters'.

The overall performance prediction problem: The complex relationship between the interconnected design parameters and the overall performance.

The overall performance of a test sequence is a collective functioning of all the individual sub-function parts in concordance. In most cases, the quality of the functioning of the individual sub-function parts is interdependent. Wet-lab validation of each test sequence is time consuming, labour intensive and expensive. Moreover, improving a given sequence by modifying individual subfunction part sequences without a holistic analysis on the overall performance represents an *ad hoc* approach prone to low success rates.

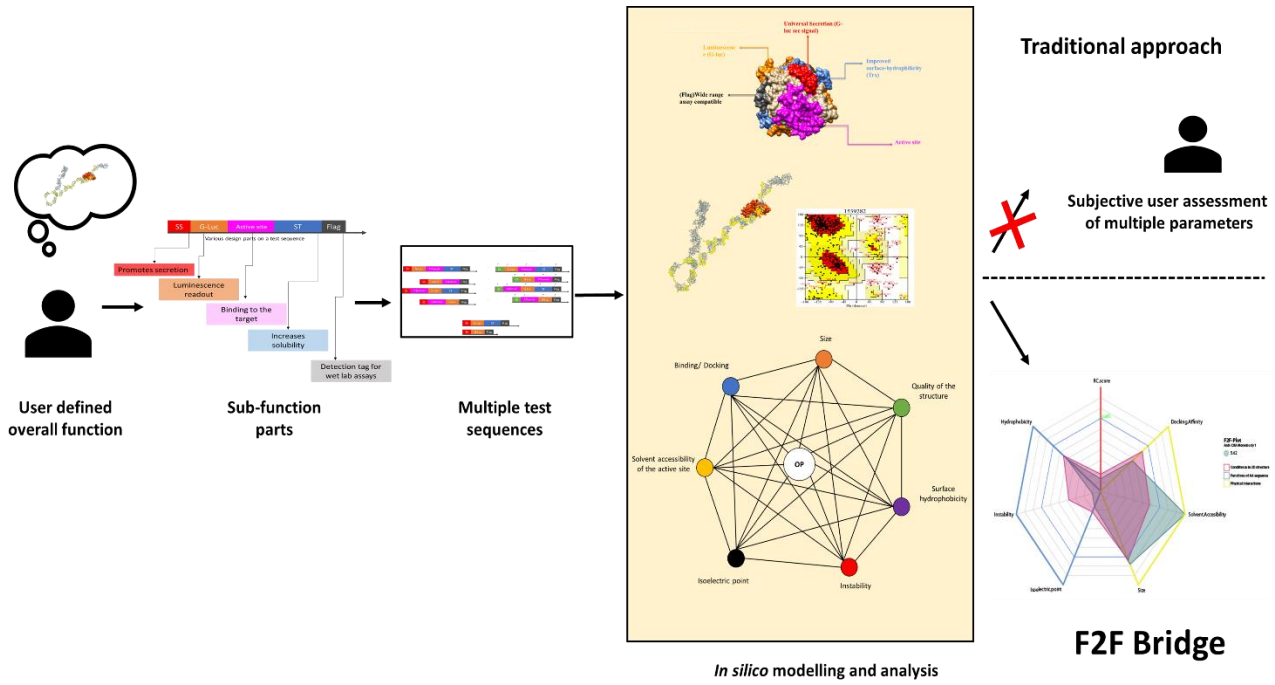


Figure 3.1. The in silico aided protein design process and the overall performance problem. *Multiple test sequences are generated for a user-defined overall function (The overall function is the collective functioning of all the sub-function parts). Sub-function parts and the test sequence build dictate the 3D structure. Test sequences are subjected to in silico modelling and multiple analyses. F2F bridge provides an objective, single metric and a calculated output to aid the selection of an optimal candidate.*

Computational tools are now available to predict protein structure, active sites, chemical properties and interactions with other proteins. These tools score and inform the quality of the individual design parameter using their respective conventional metrics. While these metrics define the quality of the individual parameter in the design, the combined effect of the design parameters (Box 1) on the overall performance is a question yet to be answered. This has been referred as the 'overall performance problem' (Box 1). The overall performance problem is a huge challenge in biology wet-lab experimentation but not completely new in a mathematical perspective.

Recent advances in machine learning and data analyses have solved similar situations. For example, 'selecting a suitable house in a new city' or 'selecting an electronic device that fits my purposes' and 'selecting a car appropriate to my needs and budget'. All these situations are mathematically similar to an extent. The parameters affect the overall performance and the best fit is chosen based on scoring and ranking the potential fits, based on the user's needs. Although the overall performance problem has been addressed in other contexts, in the scientific field of protein design, formulating a mathematical model poses a greater challenge. This is due to (i) the lack of a reporting system for negative results, (ii) time constraints in biological experimentation and (iii) the lack of standardization of measurement units in many areas of biological research.

3.2.3 Machine learning methods for proteins

Our understanding of protein function and the structure-function relationship has been enriched by computational algorithms and *in silico* tools. Machine learning algorithms and strategies have been used for over a decade now to solve sequence-structure-function relationships of proteins [18]. Latest advancements in machine learning approaches and current strategies used to predict the function(s) of a protein are reviewed by Bernardes *et al* [19].

In the lights of *de novo* protein design it is now possible to design proteins with user defined functions and shapes. In such cases, while predicting the overall function of a protein is useful, the extent to which these designed proteins perform their function is a crucial element. Design improvements could be made and measuring the overall performance of proteins provides a rationale for screening for the ideal candidate.

3.2.4 A novel mathematical strategy using machine learning approaches.

In this work, F2F-Bridge is introduced as a novel mathematical strategy aimed at predicting the overall performance of a synthetic protein. Several test sequences were designed for a defined overall function. The individual scores for all the different design parameters pertaining to each test sequence are condensed into a graphical output. The result is a visual and numerical evaluation of the test sequence. The graphical output (F2F-Plot) and the numerical evaluation (OP-score) together form a novel mathematical strategy (F2F-bridge) that scores, ranks and predicts the overall performance of the given set of test sequences. This method combines user input with *in silico* data, to give insights into the predicted overall performance of a test sequence. With the view to eventually developing a robust tool for protein performance prediction, the relationship between *in silico* and laboratory data for test proteins was also examined using two different strategies for feature selection and predictive model building. LASSO and regression-based decision trees implemented with the RandomForest algorithm.

3.3 MATERIALS AND METHODS

3.3.1 *In silico* design of test sequences:

The test synthetic proteins examined were the ClfA- (Chapter 2) and MUC1- (Chapter 4) targeted Gluc synthetic proteins. Each construct was designed to have a luminescent domain, a binding domain, a solubility tag and a secretion signal. All parts are linked in all possible permutations using different rigid and flexible linker sequences [20] (see schematic in Chapter 2 Figure 2.6)**.** Variable heavy and light chain AA sequences from different antibodies were used as the binding domains, from an antibody targeting either cell surface associated epithelial mucin 1 (MUC1; mammalian antigen) and Clumping factor A (ClfA) of *Staphylococcus aureus* (bacterial antigen). Test sequences were designed to bind to their respective target and present luminescence as a readout (bound protein luminescence).

3.3.2 Data Generation:

The different *in silico* features analysed in relation to the overall performance of the test protein, and how they are generated is outlined in Table 3.1. 3D structure prediction, size, instability index and size were obtained by providing the amino acid sequence of the protein as an input to the server. Docking, model quality, active site solvent accessibility, surface hydrophobicity, potential active sites required a .PDB file as an input. In almost all cases (unless specified) the .PDB file is generated by I-Tasser.

Design Parameter	Metric and Scale range	Metric description	Effect on overall performance	Generated:
3D structure	C-score 2 : -5	Confidence in the model and folds predicted	Higher confidence reflects more reliable model	I-Tasser[7]
Docking	ΔG kcal/mol	Gibbs free energy released by reaction	Protein-protein interactions at the active site predicted	Autodock Vina*[21]
Quality of model	RC – score	Proportion of amino acids in different regions based on steric hindrance	High agreement with stereochemistry and free energy reflects stability of structure	Saves server[22]
Active site solvent accessibility	0-9	A measure of the exposure of A residue or group of residues	Depending upon the function, the active site could be exposed to the solvent or 'hidden' inside the core	R (Using I-TASSER output)[23]

Surface Hydrophobicity	-4.5 to +4.5	Each amino acid has a hydrophobicity score between -4.5 and +4.5 as per Kyte Doolittle scale.	Ensuring ideal surface hydrophobicity aids solubility	R (Using I-TASSER output)
Size	kDa	Total weight of the protein	Size forms an important factor if the protein is required to cross/penetrate membranes and biological barriers	ProtParam Hosted by Expasy[24]
Isoelectric point	pH 0 to 14	Point at which molecule carries no net charge	The integrity of the structure of the protein in a setting is influenced by the isoelectric point	ProtParam Hosted by Expasy[24]
Potential active sites	0 to n	Number of potential active sites	Predicting the potential active sites on the designed protein informs on potential off-target effects	Coach Server
Instability index	0 to 100	Half life of protein *in vitro*	Gives an indication of the viability of the protein	ProtParam Hosted by Expasy[24]

Table 3.1: Common in silico tools, and the purpose they serve

3.3.3 Wet-lab validation

Two biological facets were used to assess the effectiveness of the functional prediction strategies – i) binding; ii) secretion.

Sub-function parts on the test sequences include: (i) **Active site:** Heavy and light chains of anti-MUC1 antibody (C595) and anti-ClfA antibody were fused with EAAAK (rigid) and GGGGS (flexible) linkers to obtain Monospecific bivalent diabodies and Monovalent ScFVs (monobodies), (ii) **Secretion signal:** Gaussia luciferase's native secretion signal, (iii) **Solubility enhancer:** SUMO tag, (iv) **Reporter:** Truncated version of GLuc was used as a luminescence reporter. (v) **Detection tag:** Flag peptide was used as a detection tag for downstream assays. See Fig 4.

Presence or absence of certain sub-function parts or their design orientation has a significant effect on the overall performance of the protein and should be accounted carefully in the design phase. In this case, over 50 different amino acid sequences were designed against each target. Of these, 8 variants per target were synthesised for testing in the wetlab. These test sequences vary in (a) (+/-) solubility enhancer, (b) (+/-) and positioning of Active site and (c) the type/format of Active site. All these test sequences were tested for their overall performance. Wet lab data were used to validate and improve the results from the F2F-Bridge. An outline of the laboratory workflow can be seen in Figure 3.2, and a more detailed description on synthesis and build of test sequences can be found in the Supplementary materials.

3.3.4 Data generation from wet lab experiments

Binding assays: As outlined in Chapter 2 and Chapter 4. Briefly, 10^8 *Staphylococcus aureus* TCH959 (naturally bearing *clfA)* or 10^6 MCF7 cells (naturally bearing MUC1) were blocked with 5% BSA for 2 h followed by incubation with supernatant containing each test construct. Cells were washed 3 times and resuspended in PBS. Luminescence was measured using Promega GloMax® 96 luminometer. In this case, since bound luminescence is the overall function, the luminescence readings corresponding to each test sequence are recorded and used for validating and improving F2F Bridge.

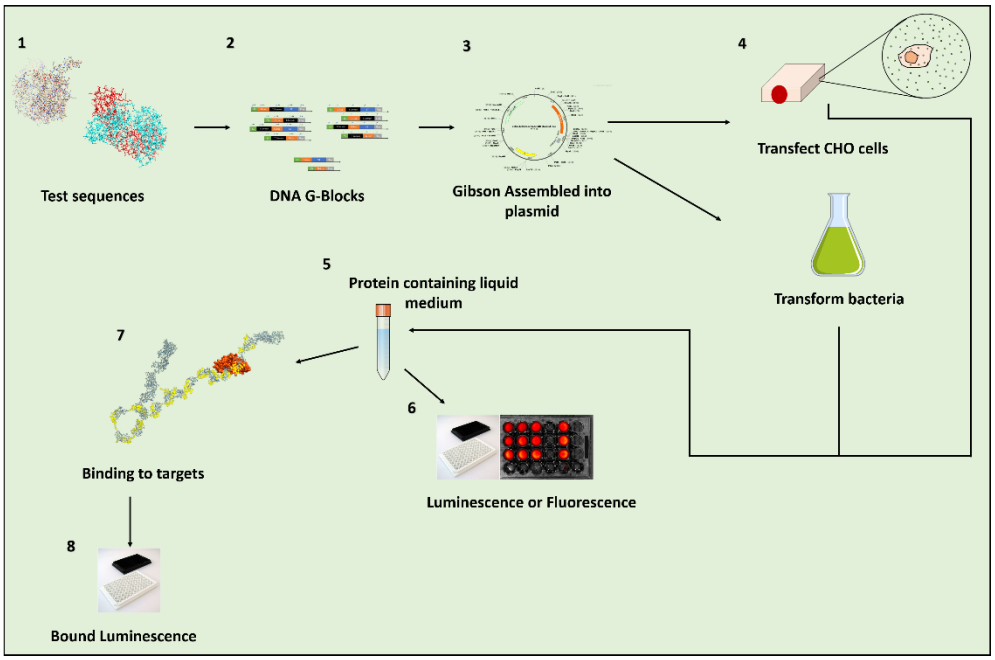


Figure 3.2: Workflow of wet lab validation of the test sequences. *Selected test variants were assembled into a plasmid. CHO cells were transfected with the plasmid containing the test variants. Cell supernatant containing the synthetic protein was collected and used in luminescence assays to confirm secretion and binding.*

3.3.5 Function2Form Bridge:

Function2Form algorithms were written in R programming language with the help of Tangney Lab bioinformatician Sidney Walker. Individual scores from all the respective *in silico* tools mentioned in Table 3.1, are tabulated and the resulting file is given as an input for F2F algorithm. The first row in the table consists of user required input values, which act as a benchmark to which the parameters of the test sequences in future would be compared. Some of these benchmark values such as C-score RC score etc are exactly the same as given by the tool developer Whereas, for parameters such as surface hydrophobicity, size or number of required active sites, the user can input values according to the design requirements. The F2F bridge in this study, uses the scores from 7 different *in silico* parameters but the scope of the tool is not limited only to these. Number and the type of *in silico* parameters used differ based on the defined overall function.

For a particular defined overall function, the choice of *in silico* parameters is based on prior knowledge and literature research manually. Such a manual selection counts all the parameters as equally contributing players. In reality however, the contribution or the effect of a parameter on the overall performance differs case by case. This is understood by considering house picking problem as an analogy. While choosing or buying a house, the locality, distance to work, cost, number of bedrooms etc are all influencing parameters. However, the importance of each parameter differs for every individual. Hence prioritising or weighing the importance of each parameter is necessary.

In this work, machine learning methods such as random forest and lasso regression have been used for feature selection. These methods help weigh the *in silico* parameters by considering their importance towards the overall performance. Wet lab luminescence data was used to train the models. The result of the F2F bridge is a radar plot (F2F plot) and an overall performance score (OP score). The F2F plot is a graphical representation of the benchmark (user defined) values and the scores of the *in silico* parameters corresponding to the test sequences. The area between the two curves indicate the disagreement between the user requirements and the design. In an engineer's eye, this is a region for improvement. The overall performance score (OP score) is a grand average

of the absolute distance between the user defined curve and test sequence curve generated by the F2F bridge. See Figure 3.3.

3.3.6 Generating the OP score:

(i) **The values are converted into a single scale**

$$\boldsymbol{i} = (((\boldsymbol{O} - \boldsymbol{O_min}))/((\boldsymbol{O_max} - \boldsymbol{O_min}))) * (\boldsymbol{N_(max)} - \boldsymbol{N_min}) + \boldsymbol{N_min}$$

Where O is the old range and N is the new range, which in the case of the F2F function is always 0-100.

(ii) **F2F function then iteratively scores each test sequence**

$$OP\ Score\ of\ x = (\Sigma|xi - yi|)/n$$

Where *x* is the test sequence, *y* is the benchmark values, *i* refers to the *ith* observation within the data supplied to the F2F bridge, and *n* is the total number of observations *i.*

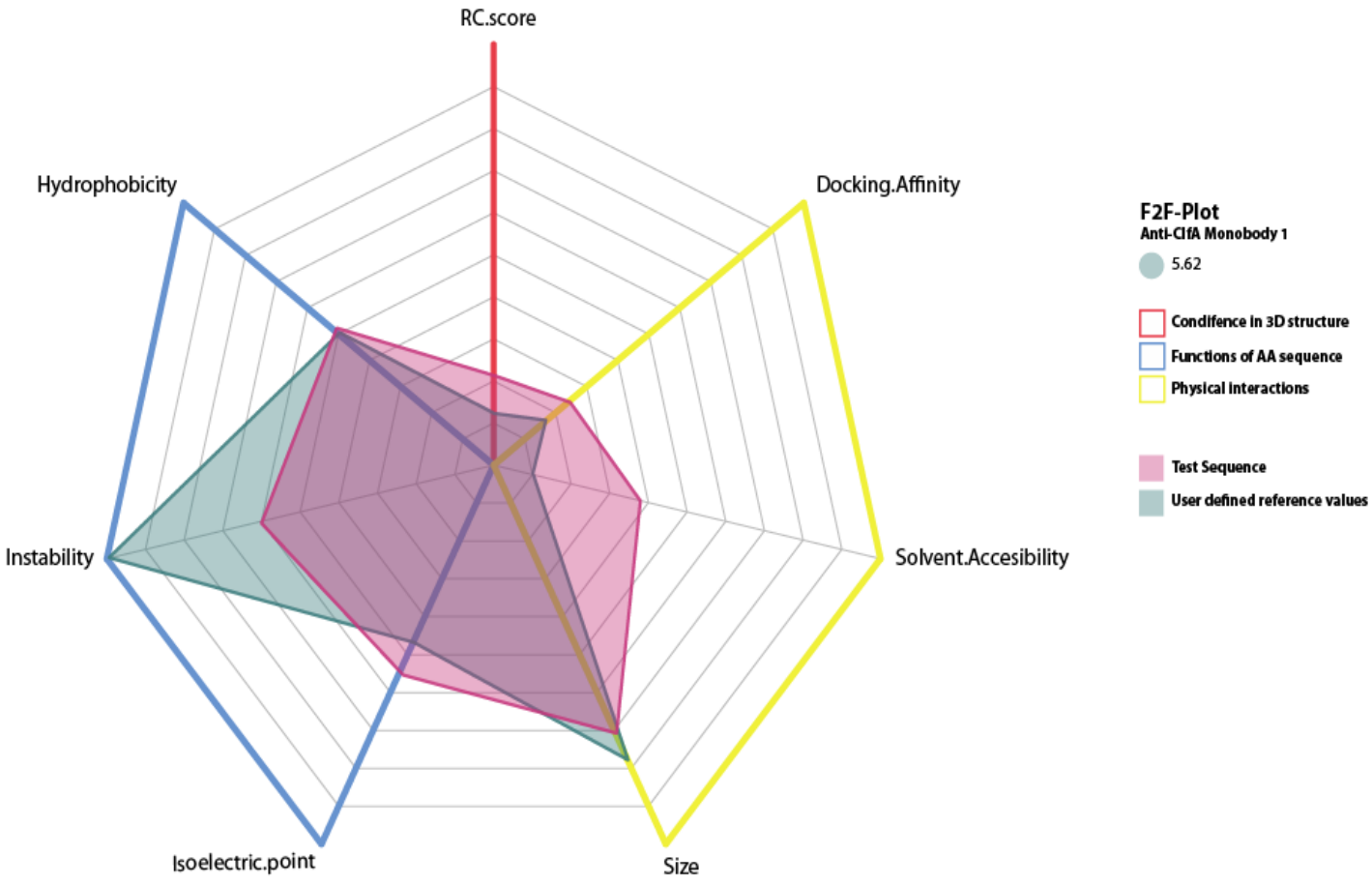


Figure 3.3: *F2F plot is a graphical representation of the merits and pitfalls of the design. The overall performance score is the grand average between the red and blue curves shown in the plot. In a high throughput setting, this could be used to rank the test sequences according to their overall performance.*

3.3.7 Statistical and machine learning methods:

As discussed above, machine learning methods were used to design a system of weights for F2F bridge. The methods used and their working are detailed in Table 3.2.

Statistical analysis: All statistical testing was performed in the base R environment v3.4.3 [25]. The LASSO regression feature selection method was implemented using the Glmnet library v2.0-16 [26], and the Random Forest regression tree analysis was performed using the RandomForest library v4.6-14 [27]. The radar plot within the F2F-bridge function was implemented with the fmsb library, v0.6.3 [28]. Visualisation was carried out using the ggplot2 package,v3.1.1 [29].

Method	Working
Lasso regression	After regularization, the parameters with non-zero coefficients are selected to be part of the final model [30]
Random forest	A variable importance plot is generated using the *in silico* parameters as input, and wet-lab luminescence as an indicator of overall performance.[31]

Table 3.2: Machine learning methods used for Function2Form bridge.

3.4 RESULTS:

3.4.1 Predicting the 'Overall Biological Performance' - Unsupervised and unweighted F2F.

F2F-plot was used to score 16 different test sequences. Figure 3.4 and 3.5 show the F2F-plots and respective OP scores for the test sequences against ClfA and MUC1 respectively. The OP score is inversely proportional to the rank of the test sequence. Lower OP score indicates better predicted overall performance and a high score indicates a poor performance. For example, in Figure 3.4, antiClfA Monobody 2 (pink shaded region) has the lowest OP score and hence predicted to be the best performer. In this case, although the instability index of the test sequence is in disagreement with the desired output, on a holistic level, antiClfA monobody 2 has the lowest disagreement and hence predicted to be the best performer. On the other hand, antiClfA diabody 2 was predicted to be the worst performer based on its high disagreement on instability index, docking affinity and solvent accessibility.

(Note: at this level, the method is still blind folded).

To assess the accuracy of F2F prediction the F2F-prediction of overall biological performance was compared with the laboratory luminescence data. See Table 3.3.

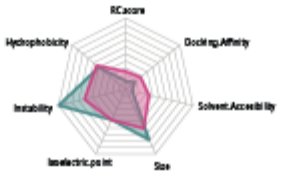
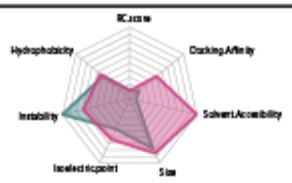
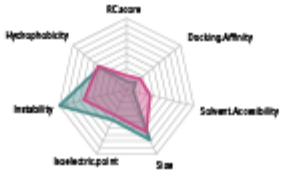
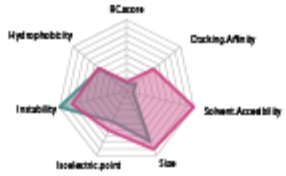
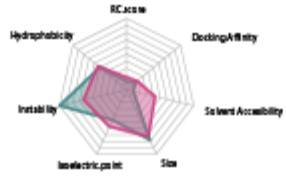
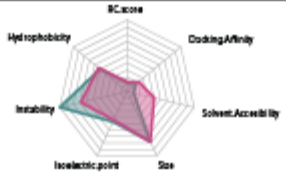
Test Sequence	F2F Result	F2F Score
ClfA Diabody 1		6.88
ClfA Diabody 2		9.48
ClfA Diabody 3		6.03
ClfA Diabody 6		9.27
ClfA Monobody 1		5.62
ClfA Monobody 2		3.62
ClfA Monobody 3		5.29
ClfA Monobody 4		3.8

Figure 3.4: *F2F-bridge output for antiClfA test sequences*

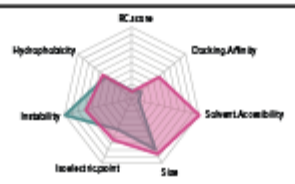
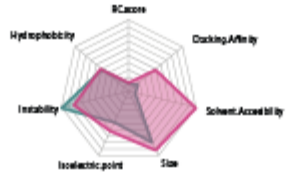
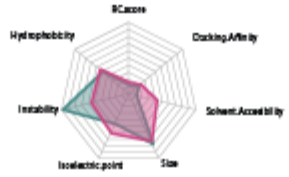
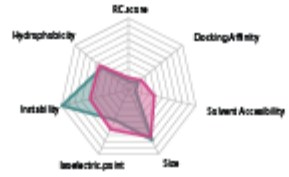
Test Sequence	F2F Result	F2F Score
Muc1 Diabody 1		6.68
Muc1 Diabody 2		9.48
Muc1 Diabody 3		6.9
Muc1 Diabody 6		9.27
Muc1 Monobody 1		5.79
Muc1 Monobody 2		4.69
Muc1 Monobody 3		5.73
Muc1 Monobody 4		5.04

Figure 3.5: *F2F-bridge output for antiMUC1 test sequences*

A

WetLab output	F2F-Plot Score	Test Sequence
2788	4.69	Muc1 Monobody 2
96232	5.04	Muc1 Monobody 4
106712	5.73	Muc1 Monobody 3
4914	5.79	Muc1 Monobody 1
2156	6.68	Muc1 Diabody 1
2436	6.9	Muc1 Diabody 3
1597	9.27	Muc1 Diabody 6
30	9.48	Muc1 Diabody 2

B

WetLab output	F2F-Plot Score	Test Sequence
9302	3.62	ClfA Monobody 2
15179	3.8	ClfA Monobody 4
26519	5.29	ClfA Monobody3
90901	5.62	ClfA Monobody 1
45542	6.03	ClfA Diabody 3
9507	6.88	ClfA Diabody 1
209	9.27	ClfA Diabody 6
1110	9.48	ClfA Diabody 2



Table 3.3: F2F plot vs Wet lab. *Test sequences in both the tables were ranked by their F2F score, and coloured from green (best performing protein), through yellow to red (worst performing protein) for both luminescence and F2F score. (Lowest number = best OP score).*

The results from Table 3.3 indicate a general guide on how F2F bridge predicts the performance and could be used to rank the test sequences based on their performance. However, there was no statistically significant correlation observed between the two panels because of the limited size of the database. Similar analysis was repeated with 'luminescence' as overall performance. There was no significant relationship was observed.

3.4.2 Applying machine learning methods to F2F bridge

2.4.2.1 Improving F2F bridge with supervised and weighted machine learning.

As discussed above the blind method accounts every *in silico* parameter as an equal contributor to the overall performance. In common terms, weighing the influence/effect of each parameter on the overall performance is crucial for an accurate prediction. Machine learning methods, Lasso regression and random forest regression tree analysis were deployed to address this problem with the blind method. 'Selection of variables' and 'regularisation' were introduced to distinguish 'big players' or 'major contributors' from the panel of *in silico* parameters.

Two wet lab outcomes are used as defined overall functions. All the analyses were carried on both 'bound luminescence' and 'secreted luminescence' readouts of the designed test sequences.

3.4.2.2 Regression analysis, selection of variables and regularisation

Applying machine learning methods for biological data is not entirely new and is a rapidly growing field. In the last 10 years, machine learning algorithms have been extensively used for protein structure and function prediction. Introduction of artificial intelligence based strategies have strengthened the field. Recently Google's AI firm 'DeepMind' introduced an algorithm called 'AlphaFold' and this has shown a significant improvement in structure prediction accuracy and gained wide attention. Although the recent advances show promising prospects, in most cases, however, the underlying mechanisms of function prediction and the relationships between the structure, function and the *in silico*

parameters of a protein, are not fully understood. Lack of large biological data to aid the machine learning and the huge number of variables in biological experimentation are a few bottle necks.

To improve the prediction accuracy of F2F bridge, distinguishing the 'big players' and accounting for the contribution of each parameter towards the overall performance is crucial. In mathematical terms (machine learning terminology) this is called feature selection. Linear regression models and decision tree based models are two widely used methods of analyses for feature selection.

In this work, lasso regression (linear regression) and Random forest (decision tree) based methods were used for feature selection and to explore the possibility of providing weights to the *in silico* parameters based on their contribution towards the overall performance.

Note: Regression analysis considers all the *in silico* parameters as independent variables. Which is not the ideal.

3.4.3 LASSO (Least absolute shrinkage and selection operator) regression

Lasso reduces the less important feature's coefficient to '0' and thereby, eliminating some features entirely. This will pick the most important features and hence is called feature selection. This is a very commonly used method in cases where there are high number of features. Figure 3.6 shows the LASSO regression analysis of bound luminescence for the designed test sequences. The features/parameters that deemed to have the maximum effect on each overall function is shown below. For the test sequences against MUC1, the parameters such as Docking Affinity, Iso-electric point, Hydrophobicity and Solvent accessibility deemed to have the maximum influence on the overall performance. In the case of test sequences against ClfA, Hydrophobicity and Instability have shown the maximum influence. However, no linear relationship was detected when the predicted parameters were input into a linear model.

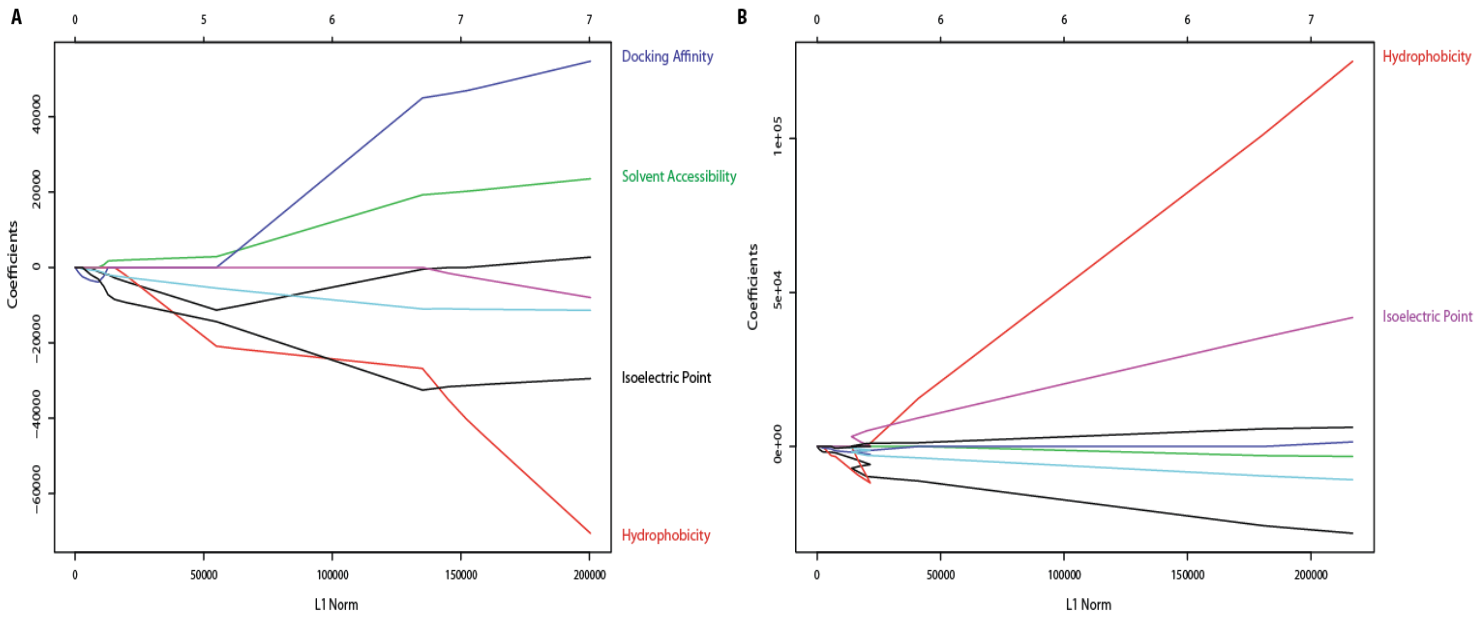


Figure 3.6: Output from LASSO regression analysis for bound luminescence. *Each line in the graph corresponds to an in silico parameter used in the F2F-Bridge. (A) shows results for antiMUC1, and (B) shows results for antiClfA. This shows that Lasso regression analysis was capable of identifying relationships between the in silico parameters, and wet lab obtained luminescence.*

Test sequences	P-Value	Adjusted R Squared
antiClfA	0.507	-0.06
antiMUC1	0.867	-0.6

Table 3.4: Results of multiple regression analysis of features selected by Lasso regression analysis against experimentally determined luminescence.

The lasso regression has successfully (i) detected the potential relationships between the wetlab output (bound protein luminescence) and the predictive features (*in silico* parameters) and (ii) the major contributors, i.e. 'big players' have been identified. However, no predictive model could be constructed. This could be due to (a) the small size of this dataset and (b) too many variables present from a wet lab perspective.
Variables in wet-lab experimentation are one of the major bottlenecks for the application of machine learning techniques. With automation technology providing aid and reducing human intervention this could be brought down.

3.4.4 Secreted luminescence as overall performance.

In the above scenario, 'bound protein luminescence' has been used as a test variable. The two groups of proteins (anti-MUC1 and anti-ClfA), in this case, were considered separately due to their binding towards different targets. However, 'luminescence' only from the secreted proteins could also be used as a test variable. In this case, the two groups (antiClfA and antiMUC1) could be combined into one single dataset and can be directly compared. This doubled the sample size. This has given the freedom to use one set of proteins (anti-ClfA) as a training set and the trained model was used to predict the performance of the other set (anti-MUC1).

In Figure 3.7, Lasso regression function was used to investigate the relationship between the *in silico* parameters and the wetlab output (luminescence due to secretion of anti-ClfA test sequences). The 'big-players' (*in silico* parameters which play a major role in dictating the performance) for luminescence due to secretion were identified by Lasso regression. A linear model was then generated, using the 'big players' identified by Lasso, to examine the degree to which they explained the test variable (luminescence due to secretion). The linear model generated this way explained 84.6% of the variability in the test variable (luminescence due to secretion), with a p-value of 0.004. This gave us the confidence to explore the utility of a lasso dictated linear model as a potential overall performance predictive tool. The experimentally determined performance (luminescence levels) were correlated with the Lasso model predicted performance.The results of correlation and the correlation coefficients of the individual

test sequences against their luminescence are plotted in Figure 3.7 and in Table 3.5. Both anti-ClfA and anti-MUC1 sets showed significant correlations, with Rho values 0.93 and 0.71 respectively.

Note: 'Big players' i.e. *in silico* parameters that have major contribution towards the overall performance, are subjective to the user defined overall function. The 'big players' identified for luminescence due to secretion may not be the same for other overall functions.

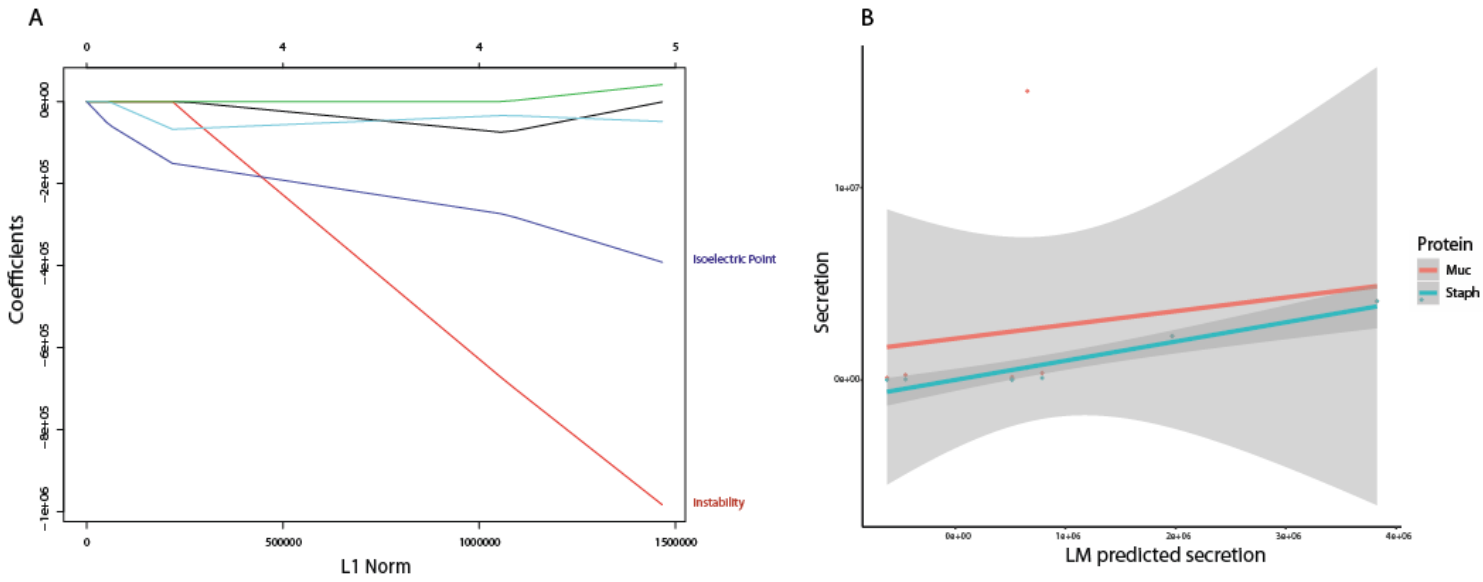


Figure 3.7: (A) ***Results from lasso feature selection (antiClfA test sequences).*** *In this instance, Isoelectric and Instability index appear as big players and are predicted to have the most effect on the test variable. Multiple linear regression was used to test the relationship between the experimental luminescence (wetlab) and features selected by Lasso regression. The model was found to be significant and explained 84.61% of the variability in the test variable and has a p-value of 0.004.* ***(B) Correlation plot of experimental values of test variable vs Lasso directed linear model predicted values of test variable.*** *AntiClfA test sequences (in blue) are used as the training set and AntiMUC1 test sequences (in red) are used as the test set.*

	All	Anti MUC1	Anti ClfA
Rho	0.846	0.71	0.93
P value	0.0009	0.04	0.0006

Table 3.5: Rho (correlation coefficient) and P values of the Lasso directed linear model. *The values are based on the plot between the predicted test luminescence vs experimental luminescence.*

3.4.5 Random Forest regression tree analysis as an alternative performance prediction method.

To increase the versatility of F2F Bridge, random forest regression tree analysis has been used for a non-linear model. Similar methodology to Lasso regression was implemented within random forest. Anti-ClfA test sequences were used as the training set and the anti-MUC1 test sequences were used as the test set. The results from the random forest regression model are shown in Figure 3.8. The model was successfully able to explain 41 % of the variability in the test variable (luminescence from secretion). Table 3.6 shows the correlation values between the random forest algorithm predicted luminescence and experimental luminescence. Significant correlation was observed between the test variable (secreted luminescence) predicted by the random forest and the experimentally determined values. When the same was repeated using the bound luminescence as test variable, no significant results were observed. With bigger datasets for training, this accuracy can only increase. Table 3.7 shows a summary of all the analyses and their respective outcomes.

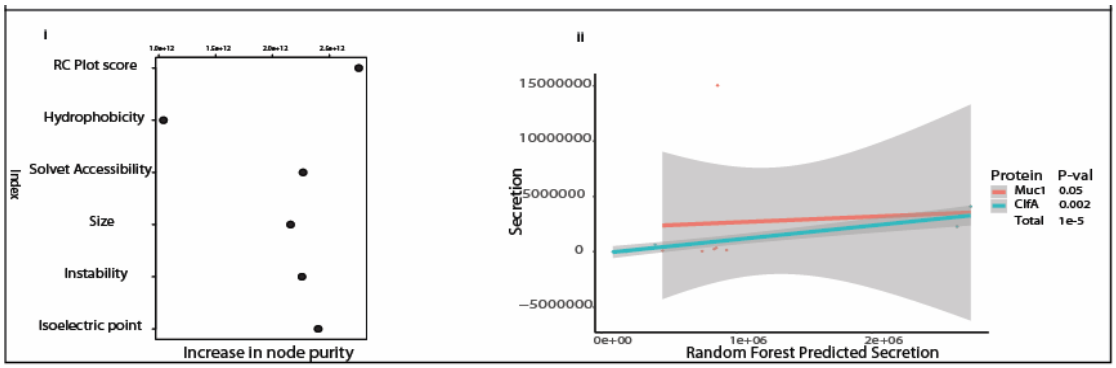


Figure 3.8: Random Forest regression tree analysis summary. *(i) Mean node purity for each predictive feature. Smaller values indicate the high importance of the feature to the model. (ii) The trained model explained 41% of the variability in experimentally determined test variable (secreted luminescence). The model predicted luminescence values were correlated with experimentally determined luminescence values. The model has an overall correlation coefficient of 0.87 and an associated p-value of 1e-05.*

	antiClfA	antiMuc1	Total
P value	0.002	0.05	1e-5
Rho	0.92	0.71	0.87

Table 3.6: *Correlation results from Random forest predicted regression model.*

In silico Test Performed	Biological Performance Tested	Test Sequence	Result
"Blind" F2F Bridge	Binding and Luminescence	antiClfA and antiMUC1	The F2F plot was able to provide a guide for the expected performance of the test sequence when the test sequences were ranked by OP score and by wet lab output, and the accompanying plot was able to inform on how to improve the test sequence. No statistically significant relationship between OP score and wet lab output could be found.

LASSO feature selection and linear model **building**	Binding	antiClfA and antiMUC1	LASSO regression analysis was able to detect discrete patterns in the data, showing Hydrophobicity and Isoelectric point both to have a positive relationship with bound luminescence in antiClfA. In the case of antiMUC1 Docking Affinity and Solvent accessibility were shown to have a positive effect, Isoelectric point and Hydrophobicity a negative one.
Using LASSO regression analysis dictated linear model as a **predictive tool**	Binding	antiClfA and antiMUC1	The models predicted in the above analysis were unable to exaplain any of the variability in the bound luminescence of antiMUC1 or ClfA test sequences.

LASSO feature selection and linear model **building**	Luminescence	antiClfA	LASSO regression analysis was able to detect discrete patterns in the data, a linear regression with solvent accessibility and instability was able to explain 86.4% of the variability in luminescence in the antiClfA samples.
Using LASSO regression analysis dictated linear model as a **predictive tool**	Luminescence	antiClfA and antiMUC1	The model created in the above test was used to predict luminescence values for both antiClfA and antiMUC1. In both cases these predictions showed strong positive correlations with the experimental luminescence values which were statistically significant.
Random Forest regression tree model **building**	Luminescence	antiClfA	A regression tree implemented with randomForest was able to explain ~41% of the variability in the luminescence of antiClfA test sequences.

Using Random Forest regression tree as a **predictive tool**	Luminescence	antiClfA and antiMUC1	The model created in the above test was used to predict luminescence values for antiClfA and antiMUC1 test sequences. In both cases, these predictions showed strong positive correlations with the experimental luminescence values that were statistically significant.

Table 3.7: Summary of all analyses performed in the F2F study

3.5 DISCUSSION

In this study, a novel strategy has been developed to help visualize and score the overall performance of a test sequence (protein). Using machine learning and statistical analyses a mathematical model has been tested with significant prediction accuracy. The resultant, F2F bridge, is a combination of (i) a graphical overview predicting and displaying the strengths and weakness of a test sequence and (ii) an OP score that predicts and ranks the performances of different test sequences. Unlike the current *in silico* tools that inform the quality of only a single individual design parameter, F2F bridge provides a holistic view on a given test sequence. This top down approach of F2F bridge holds a key for informed protein design. F2F bridge could be deployed for both low throughput design and also for high throughput screening. In a low throughput setting, the F2F plot would play a pivotal role in highlighting the 'pitfalls and merits' of the design corresponding to a test sequence. The design could then be improved and F2F plot could be regenerated until satisfactory design is obtained. In a high throughput scenario, multiple designs of test sequences for an overall functions are scored and ranked by F2F bridge.

Scale of use	Outcome
High Throughput	A database of test sequences or extant proteins of known sequence can be queried with the F2F-bridge scoring each test sequence and identifying those most suitable.
Low Throughput	On a protein by protein basis the F2F-bridge provides a graphical overview of the relationship between the features of the test sequence and the optimal values specified by the user, informing the user on how to improve the test sequence.

Table 3.8: High throughput and low throughput applications of F2F-bridge. *Depending on the intended application, F2F-bridge can be used as a tool to screen a large set of test variants by empirical ranking or for improving and fine tuning the existing design by graphically visualising the merits and falls of the model.*

In both cases, F2F bridge is positioned precisely to bridge the gap between the myriad of seemingly abstract *in silico* values and wet lab performance. The simple graphical output and single OP score are easy to interpret with minimal know-how on programming. When combined with wet-lab testing, the patterns existing in the *in silico* data have been successfully used to generate predictive models.

3.5.1 Relevance to the laboratory scientist

The ultimate aim of F2F bridge is to bridge the increasing gap between the wet lab biologists and the powerful *in silico* tools. While scoring and ranking the test sequences based on their predicted overall performance, F2F bridge, mimics wet-lab screening and takes only a fraction of time, cost and expertise required in laboratory screening. This could lead to significant operational savings.

Uses of F2F bridge could be divided into (i) ***wet lab scientists*** to study and analyse the merits and pitfalls of a protein visually using the F2F plot (ii) ***computational biologists*** for informed protein design using automated the redesign and (iii) ***protein based industries*** for screening lead candidates.

3.5.2 F2F Bridge

The blind F2F model (without feature selection, unsupervised and unweighted) showed promising results and an early indication of predicting the performance of a test sequence. This 'unweighted and unsupervised' combination of *in silico* parameters deemed to have an effect on overall biological performance of the test sequences. Wet lab experimentations are labor and time intensive and it is not cost effective to synthesize and test multiple test sequences. Given the ease of implementation, information provided by F2F bridge prior to synthesis is extremely valuable. The OP-score and the accompanying F2F plot highlight the design aspects that diverge from the user requirements and provide caution before wet lab synthesis. The blind model of F2F could be considerably improved with a larger dataset. In the meantime, feature selection and model refinement using machine learning methods has been implemented.

As discussed earlier the blind folded model does not account for big players. However, proteins are multifunctional molecules and for every user defined overall

function, a set of *in silico* parameters would have a bigger influence than others. Thus identifying such big players becomes crucial to improve the accuracy of prediction. The ultimate goal of deploying the machine learning methods was to search for underlying patterns that could help predict the overall performance of the test sequence.

3.5.3 Lasso driven feature selection and linear models

Initially, feature selection was done using Lasso regression to predict the luminescence due binding of the test sequence to its target. In the L1 norm vs coefficients plot, the point at which the predictive feature enters the model and the effect it has on the test variable is very important. Although, big players that affect the bound luminescence were successfully identified, it was impossible to incorporate them into a linear model with statistical significance. As discussed earlier, the small nature of the current database and too many variables in wet lab processes (such as protein expression, binding to target, luminescence etc) were the two main bottlenecks. For this reason and the chance to utilise the increase in sample size, luminescence data from secretion was analysed. In this case, both anti-ClfA and anti-MUC1 datasets could be pooled together. This approach proved to be more successful in terms of improving the prediction accuracy of F2F bridge.

Lasso regression based feature selection was performed again on secreted luminescence as a test variable to explore for patterns between the *in silico* parameters and experimental secreted luminescence. As expected this was more effective than the bound luminescence as test variable. It has been also shown that the linear model can be predicted using this method and a strong correlation was observed between the predicted and experimental test variable values. Once the linear relationship model has been established for anti-ClfA, this was used to successfully predict the test variable (secreted luminescence values) of antiMUC1 test sequences. The predicted luminescence values correlated with experimental luminescence values with good correlation. The antiMUC1 experimental luminescence data originated from a different experiment and they are also a different class of proteins when compared to anti-ClfA set. The fact that the lasso directed model displayed strong predictive power in-spite of all the differences between the two groups, is very encouraging.

3.5.4 Random forest regression tree directed model

Random forest regression tree model was used on 'luminescence due to secretion' as a test variable. The model was successfully able to explain over 41 % of the variability in the experimental secreted luminescence in the anti-ClfA test sequences. The regression tree model predicted test variable values closely correlated with the wet lab experimental values of the test variable. As observed with Lasso regression model, it was highly encouraging that the random forest regression model trained on anti-ClfA test sequences was also able to predict the values of test variable for the anti-MUC1 test sequences with significant correlation with the experimental values of secreted luminescence.

3.5.5 F2F bridge - Foundation for overall performance prediction and hurdles ahead

Using machine learning approaches such as Lasso regression model and the Random Forest regression tree model, F2F showed promising prospects for predicting the overall performance of a test sequence. The workflow and implementation of the prediction model is straightforward computationally but not on the wet lab experimentation. Synthesis/expression, functional assays and quality assessments on proteins is time consuming and is capital intensive. This results in small size of dedicated datasets. Using datasets from multiple sources is also challenging due to the vast variability in wet lab experimentation. Although this seems as a big challenge, recent advancements in synthetic biology provides an inspiring platform for high throughput synthesis and testing of multiple proteins. The current version of F2F lays the foundation for protein performance prediction using *in silico* tools. In the future, with an expanded dataset, the current two models would be re-assessed to confirm the increase in significance in prediction.

The current model also relies on web servers and third party tools for assessing *in silico* parameters of a test sequence. This also means that the accuracy of the individual *in silico* parameter values depends on the corresponding tools developed by various sources. In a future version, the accuracy of every individual web server/tool would be taken into consideration to provide an overall confidence score on the predicted performance. This would prevent error-compounding.

3.5.6 Outlook and F2F Bridge V.2.0

In the current version of F2F bridge, it has been shown that the patterns observed in the *in silico* parameters of proteins could be used to predict significantly accurate overall biological performance and this could play a pivotal role in design optimisation and high throughput screening. Larger datasets could immensely benefit F2F bridge to improve in accuracy. Therefore this raises the need for the establishment of a new community wide data reporting system for *in silico* and wet lab data for proteins. This provides larger training datasets and wide variety of overall biological functions. SourceTracker algorithm used in metagenomic studies which is used to track possible sources of contamination in HTS studies served as standing example and inspiration for the proposed community wide data reporting system. Also, the regulation proposed by journals to deposit structures in PDB prior/after publication provide confidence to the establishment of such a data reporting system. F2F V2.0 would also have a parallel server which would take amino acid sequence as an input and perform all the *in silico* parameter calculations in-house. All the individual errors and compound errors will be taken into consideration and users would be informed with a graphical score. New strategies with applied weights to *in silico* parameters, combinational approach (performing multiple machine learning analyses sequentially for refinement) and suggestion nudges to improve the OP score will be incorporated.

With all the above features F2F bridge forms a novel tool for informed protein design and capitalises on end-user empowerment.

3.6 Conclusions

The design model build test approach promoted by modern synthetic biology stands to benefit immensely from F2F bridge. This becomes an indispensable strategy for a biologist to triage the potential best performers and visualise the merits and falls of the protein. With little further adjustments in V.20 F2F bridge integrates in the DMBT cycle and adds the ‘learn’ step by empowering the end-user (wet lab biologist) with a holistic view on the overall performance of a protein. With a community-based data reporting system and larger datasets, F2F bridge could be tunes to Pareto optimality.

3.7 References

1. Research, T.M. *Bioengineered Protein Drugs Market Worth US$336.9 Billion by 2025: Increasing Drugs in Phase III of Clinical Trials to Boost the Market*. [Market report] 2017 July 25, 2017; Available from: https://www.prnewswire.com/news-releases/bioengineered-protein-drugs-market-worth-us3369-billion-by-2025-increasing-drugs-in-phase-iii-of-clinical-trials-to-boost-the-market-636479663.html.
2. Huang, P.S., S.E. Boyken, and D. Baker, *The coming of age of de novo protein design.* Nature, 2016. **537**(7620): p. 320-7.
3. Rose, P.W., et al., *The RCSB protein data bank: integrative view of protein, gene and 3D structural information.* Nucleic Acids Res, 2017. **45**(D1): p. D271-D281.
4. Prentiss, M.C., et al., *Protein Structure Prediction: The Next Generation.* J Chem Theory Comput, 2006. **2**(3): p. 705-16.
5. Bloom, J.D., et al., *Thermodynamic prediction of protein neutrality.* Proc Natl Acad Sci U S A, 2005. **102**(3): p. 606-11.
6. Woolfson, D.N., et al., *De novo protein design: how do we expand into the universe of possible protein structures?* Curr Opin Struct Biol, 2015. **33**: p. 16-26.
7. Yang, J. and Y. Zhang, *I-TASSER server: new development for protein structure and function predictions.* Nucleic Acids Res, 2015. **43**(W1): p. W174-81.
8. Yang, J. and Y. Zhang, *Protein Structure and Function Prediction Using I-TASSER.* Curr Protoc Bioinformatics, 2015. **52**: p. 5 8 1-15.
9. Kappel, K., Y. Miao, and J.A. McCammon, *Accelerated molecular dynamics simulations of ligand binding to a muscarinic G-protein-coupled receptor.* Q Rev Biophys, 2015. **48**(4): p. 479-87.
10. Li, J., et al., *Transient formation of water-conducting states in membrane transporters.* Proc Natl Acad Sci U S A, 2013. **110**(19): p. 7696-701.
11. Clark, A.J., et al., *Prediction of Protein-Ligand Binding Poses via a Combination of Induced Fit Docking and Metadynamics Simulations.* J Chem Theory Comput, 2016. **12**(6): p. 2990-8.

12. Wood, C.W., et al., *ISAMBARD: an open-source computational environment for biomolecular analysis, modelling and design.* Bioinformatics, 2017. **33**(19): p. 3043-3050.
13. Mitra, P., et al., *An evolution-based approach to De Novo protein design and case study on Mycobacterium tuberculosis.* PLoS Comput Biol, 2013. **9**(10): p. e1003298.
14. Bellows, M.L. and C.A. Floudas, *Computational methods for de novo protein design and its applications to the human immunodeficiency virus 1, purine nucleoside phosphorylase, ubiquitin specific protease 7, and histone demethylases.* Curr Drug Targets, 2010. **11**(3): p. 264-78.
15. Jiang, L., et al., *De novo computational design of retro-aldol enzymes.* Science, 2008. **319**(5868): p. 1387-91.
16. Karanicolas, J., et al., *A de novo protein binding pair by computational design and directed evolution.* Mol Cell, 2011. **42**(2): p. 250-60.
17. Coluzza, I., *Computational protein design: a review.* J Phys Condens Matter, 2017. **29**(14): p. 143001.
18. Bonetta, R. and G. Valentino, *Machine learning techniques for protein function prediction.* Proteins: Structure, Function, and Bioinformatics. **n/a**(n/a).
19. Bernardes, J.S. and C.E. Pedreira, *A review of protein function prediction under machine learning perspective.* Recent Pat Biotechnol, 2013. **7**(2): p. 122-41.
20. Chen, X., J.L. Zaro, and W.C. Shen, *Fusion protein linkers: property, design and functionality.* Adv Drug Deliv Rev, 2013. **65**(10): p. 1357-69.
21. Alhossary, A., et al., *Fast, accurate, and reliable molecular docking with QuickVina 2.* Bioinformatics, 2015. **31**(13): p. 2214-6.
22. 2019;, U.S.s.
23. Fraczkiewicz, R. and W. Braun, *Exact and efficient analytical calculation of the accessible surface areas and their gradients for macromolecules.* Journal of Computational Chemistry, 1998. **19**(3): p. 319-333.
24. Wilkins, M.R., et al., *Protein identification and analysis tools in the ExPASy server.* Methods Mol Biol, 1999. **112**: p. 531-52.

25. Team, R.C., *R: A language and environment for statistical computing.* . R Foundation for Statistical Computing, 2017.
26. Jerome Friedman, T.H., Robert Tibshirani, *Regularization Paths for Generalized Linear Models via Coordinate Descent.* Journal of Statistical Software, 2010. **33**(1): p. 1-22.
27. M.Wiener, A.L.a., *Classification and Regression by randomForest.* R News, 2002. **2**(3): p. 18-22.
28. Nakazawa, M., *Functions for Medical Statistics Book with some Demographic Data*. Pearson Education Japan. 2007.
29. Wickham, H., *ggplot2: Elegant graphics for data analysis.* Springer-Verlag New York, 2009.
30. Tibshirani, R., *Regression shrinkage and selection via the lasso: a retrospective.* Journal of the Royal Statistical Society: Series B (Statistical Methodology), 2011. **73**(3): p. 273-282.
31. Liaw A, W.M.C.a.R.b.R. **Vol 232001**. (R code)

Chapter 4

A novel *in vivo* imaging strategy using synthetic protein engineering

Table of Contents

4.1 ABSTRACT 149

4.2 INTRODUCTION 151

4.2.1 Bioluminescence *in vivo* imaging 151

4.2.2 Synthetic proteins for *in vivo* imaging 152

4.2.3 Tumor associated MUC1 152

4.3 MATERIALS AND METHODS 155

4.3.1 Overview of *in silico* design of synthetic proteins 155

4.3.2 Computational tools used for *in silico* -aided design and validation 155

4.3.2.1 Protein structure modeling 155

4.3.2.2 Superimposing predicted models 155

4.3.2.3 3D visualisation 155

4.3.2.4 Protein-protein interactions 156

4.3.2.5 Theoretical structure validation 156

4.3.2.6 Total hydrophobicity vs Surface hydrophobicity 156

4.3.2.7 Structural remodeling and affinity improvements 156

4.3.3 Wet-lab experimentation methods 156

4.3.3.1 DNA design 156

4.3.3.2 Plasmid scale-up 157

4.3.3.3 Restriction digestion 157

4.3.3.4 Gibson Assembly 157

4.3.3.5 Validating cloning using colony PCR and Sanger sequencing 157

4.3.3.6 *In vitro* transfection 158

4.3.3.7 Binding assays 158

4.3.4 *In vivo* methods 158

4.3.4.1 Animals used in the study 158
4.3.4.2 Bacterial administration to mice 158
4.3.4.3 Systemic administration of synthetic proteins 158
4.3.4.4 Non-invasive *in vivo* imaging 159
4.3.4.5 *In vivo* transfection 159
4.4 RESULTS 160
4.4.1 *In silico* validation and screening 160
4.4.1.1 TA MUC1 antigen 160
4.4.1.2 Design elements and design rationale 162
4.4.1.3 Protein-Protein Docking 164
4.4.2 Wet-lab experimentation 167
4.4.2.1 Validating protein production 167
4.4.2.2 *In vitro* binding to MUC1 169
4.4.2.3 Selecting the best suitable candidate 174
4.4.2.4 Dose response assays 177
4.4.3 *In vivo* imaging 179
4.4.3.1 *In vivo* synthetic protein dose response (M1) 179
4.4.3.3 *In vivo* synthetic protein dose response (S1) 182
4.4.3.5 GlucS1 vs NanolucS1 186
4.4.4 *In vivo* production of synthetic proteins 188
4.5 DISCUSSION 193
4.6 REFERENCES 198

4.1 ABSTRACT

Background Recent advancements in life sciences such as protein engineering place biomedical research in fast and fascinating transformation phase. Translating such novel scientific concepts into a clinical setting requires extensive prior laboratory testing. The lack of efficient methods to track the performance of the therapeutics/diagnostics *in vivo* is a significant barrier and causing hindering their clinical translation. Bioluminescence imaging for is an attractive imaging modality for *in vivo* applications. Targeted synthetic proteins equipped with an optical reporter such as a luciferase could become a valuable tool for *in vivo* optical imaging.

Aims The aim of this study was to develop a novel *in vivo* imaging strategy using *in silico* engineered targeted synthetic luciferase proteins.

Methods In this work, test constructs targeting tumour associated MUC1 were built and tested for their *in vitro* binding to MUC1 antigen-expressing cell lines. Based on the luminescence readouts, the best performer was identified. This, as well as the best performer targeting *S. aureus* ClfA (from Chapter 2), were tested for specific imaging of target cells in *in vivo* murine models.

Results Over 100 different multi-part test constructs targeted to human MUC1 were modelled and validated using various computational tools to inform and guide downstream wet-lab experiments, as per Chapter 2. Gaussia luciferase (Gluc) or nanoluc were used as a luminescence reporters. All test construct variants were subjected to computational screening of predicted functionality. The best predicted performers were appropriately modified to ensure required hydrophobicity, net surface charge, active site exposure and valid 3D structure. Wet-lab studies were conducted to validate MUC1 protein production and functioning (luminescence and specific target binding) *in vitro*.

For both MUC1 and ClfA targeted proteins, *in vivo* luminescence imaging studies involving systemic intravenous (IV) administration of proteins, validated synthetic protein specific accumulation at target cell locations within mice as evidenced by localised luminescence. Dose response studies indicated that luminescence output was both target cell and administered protein quantity related. Upon validation that systemically-administered synthetic proteins functioned as *in vivo* imaging agents, it was

investigated if these proteins could be produced by the mouse *in vivo* to achieve the same effect. *In vivo* transfection of quadriceps with DNA constructs used for synthetic protein production was examined using electroporation and lipofection. However, this DNA strategy proved unsuccessful.

Conclusion This study serves as a proof-of-concept for using targeted reporter proteins for *in vivo* imaging.

4.2 INTRODUCTION

Recent advances in life sciences such as gene-editing [1], protein engineering and the advent of *de novo* designed proteins [2], place biomedical research in a fast and fascinating transformation phase. Translating such novel scientific concepts into a clinical setting requires extensive prior laboratory testing. Traditional laboratory testing methods are time consuming, invasive and involve multiple samplings, and have high associated costs for instrumentation and analysis. The lack of efficient methods to track the performance of the therapeutics/diagnostics *in vivo* is a significant barrier and causing hindrance towards their clinical translation [3].

Testing novel diagnostic and therapeutic strategies requires real-time *in vivo* targeting, tracking and monitoring of various biological events that result due to the intervention. Such real-time *in vivo* tracking is pivotal in understanding complex cellular and systemic functions inside a test subject's body. Optical imaging (OI) represents a simple low-cost solution for real-time *in vivo* monitoring and has contributed significantly to biomedical research [4]. Considering the inexpensive nature of OI and its ability to be applied for high-throughput work, OI is an attractive ionising radiation independent imaging alternative [3].

4.2.1 Bioluminescence *in vivo* imaging

BioLuminescence Imaging (BLI) is an attractive imaging modality for *in vivo* research applications [5, 6]. BLI has a standout advantage when it comes to signal-to-noise ratios. This is due to the negligible background noise when compared to the luminescence signal from the luciferase reaction [7]. Due to the excitation-independent mechanism of luciferases, BLI does not face risks such as photobleaching and phototoxicity which are a major concern in other optical imaging modalities [7]. These properties make BLI highly suitable for *in vivo* imaging. As discussed in Chapter 2, Gluc, Rluc and Fluc are the most widely used luciferases for BLI. Engineered versions of synthetic proteins with luciferase parts have been proven as an attractive imaging strategy [6, 8-12]. However, BLI has faces challenges such as limited depth light penetration due to tissue absorption and scattering.

BLI using synthetic protein also raises questions such as potential for immunogenicity, toxicity and clearance from the system, that act as hurdles to clinical development.

4.2.2 Synthetic proteins for *in vivo* imaging

The number of engineered proteins used in imaging and therapeutic applications has grown significantly in the last 10 years. Engineering synthetic proteins with multiple functions and their applications have been discussed in Chapter 2. In this chapter, I aimed to design synthetic proteins targeting cancer cells, and to assess their potential as *in vivo* imaging agents. MUC1 was chosen as a cancer cell target to test the strategy, along with the bacterial-targeted agent developed in Chapter 2.

.

4.2.3 Tumour associated MUC1

MUCIN-1 (MUC1) is a transmembrane glycoprotein expressed in glandular and luminal epithelial cells of various tissues/organs. MUC1 is long string like structure (200-500 nm long) having a transmembrane and an extracellular domain [13] (Figure 4.1). Both domains are linked together by stable hydrogen bonds. In non-malignant cells, MUC1 acts as a protective layer to underlying epithelia. Upregulation of MUC1 expression is associated with various epithelial cancers [14, 15]. This generated high interest to pursue MUC1 as an oncogenic molecule. MUC1 plays an important role in disease progression involving cancer cell proliferation, metastasis and angiogenesis. Tumour-associated MUC1 differs from MUC1 present on regular cells [15]. The extent of glycosylation is one of the major differences between MUC1 found on regular cells and cancer cells. MUC1 is a heavily glycosylated protein - however, many studies have shown that the MUC1 presented on cancer cells has significantly lower levels of glycosylation [14, 15]. From a therapeutic point-of-view, the loss of glycosylation exposes the protein backbone and provides scope for antibody binding. Many therapeutic strategies to target MUC1 take advantage of the exposed protein due to loss of glycosylation. Given the multifaceted nature in cancer, targeting MUC1 presents a promising strategy in cancer diagnosis and treatment.

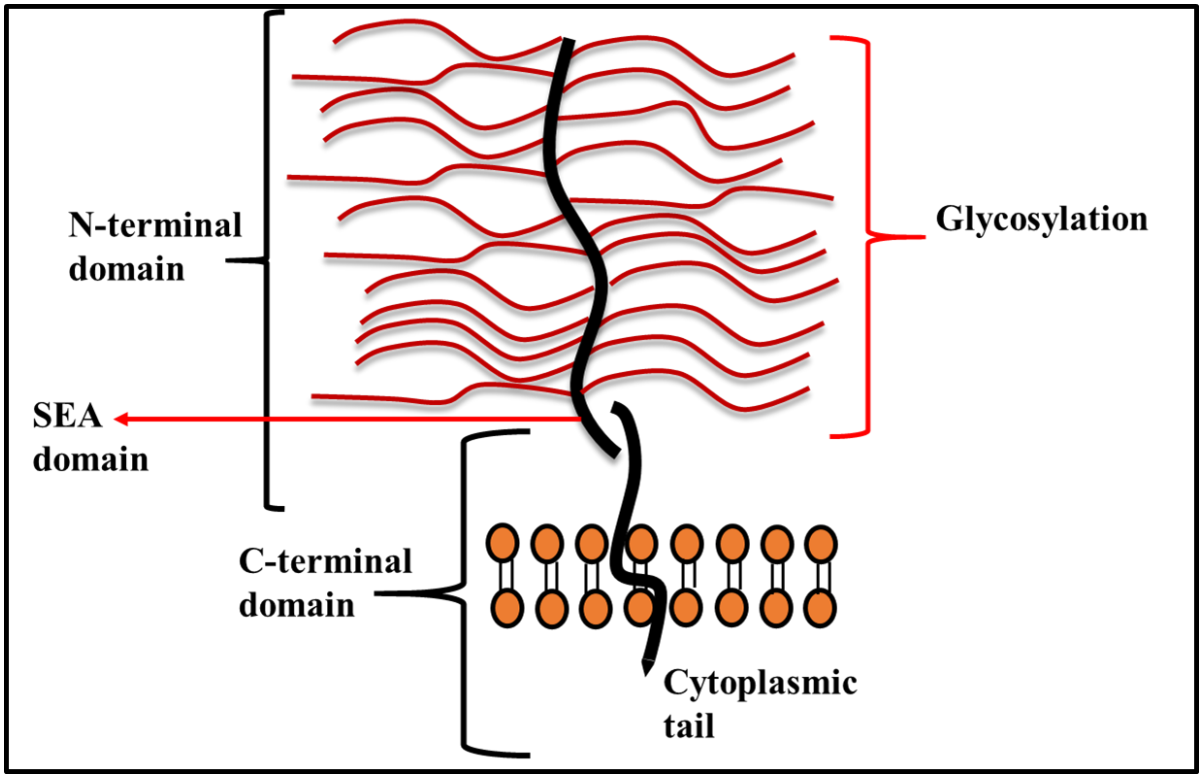


Figure 4.1: Graphical representation of MUC1 structure and various important parts. *The structure of MUC1 consists of a cytoplasmic tail and an extracellular domain. The two domains are linked together by strong hydrogen bonds. SEA (sea urchin sperm protein, enterokinase and agrin, a highly conserved 120 AA module) domain has a proteolytic cleavage site.*

The present work seeks to extend the concept of design-model-build-test of synthetic proteins described in Chapter 2, to validate the *in vivo* functioning of the synthetic protein as an *in vivo* imaging agent using murine models. MUC1 was chosen as a test target due to its high surface expression. The *in vivo* imaging strategy was also validated using S1 test constructs targeting ClfA. The workflow and proof-of-concept of using synthetic designer proteins as *in vivo* imaging agents is described in following sections.

4.3 MATERIALS AND METHODS

4.3.1 Overview of *in silico* design of synthetic proteins

Synthetic proteins targeting hMUC1 contained multiple subparts similar to the synthetic proteins shown in Chapter 2. Over 100 different variants were made and manually screened for potential best performers. Following the successful functioning of the synthetic proteins in Chapter 2, similar subparts and design principles were used for all the test constructs. *In silico* tools described in Chapter 2 were used to make several iterations of each test variant. All the designed structures were validated until desired structural conformations were achieved. VH and VL domains of anti MUC1 ScFv clone C595 were used as the binding domain to target MUC1. The amino acid sequence and the structure for MUC1 antigen was obtained from RCSB PDB. The process workflow and methods are explained in the sections below.

4.3.2 Computational tools used for *in silico* -aided design and validation

All the computational methods followed the same workflow and analyses as shown in Chapter 2 section 2.3.2.

4.3.2.1 Protein structure modeling

Protein modelling was performed primarily using the I-Tasser protein modelling suite [16]. The test constructs were also modelled using Rosetta modelling suite to increase the prediction reliability [17, 18].

4.3.2.2 Superimposing predicted models

The models predicted by I-Tasser and Rosetta were superimposed onto each other and RMSD was calculated using R-algorithms, developed in-house at the Tangney lab.

4.3.2.3 3D visualisation

UCSF Chimera was used for all protein visualisation throughout this work [19]. Chimera was also used to visualise protein-protein interactions after protein docking which was carried in the later stages of this workflow.

4.3.2.4 Protein-protein interactions

Protein-protein interactions were studied using protein docking. AutoDock Vina was used for protein docking [20]. All the bound conformations were visualised in UCSF Chimera to screen for conformations that bind at the active sites (epitope-paratope interaction).

4.3.2.5 Theoretical structure validation

Ramachandran plots (RC plots) were used to validate the theoretical stability of the modeled structures. This information could be used to verify the structural stability of the model to exist in a natural environment.

4.3.2.6 Total hydrophobicity vs Surface hydrophobicity

Surface hydrophobicity was mathematically calculated by identifying the residues which have over 40 % exposure. The hydrophobicity is then calculated for these surface residues using in-house R algorithms.

4.3.2.7 Structural remodelling and affinity improvements

Improvements to the backbones, linkers and single residue replacements required structural remodelling. This was performed using Rosetta package. In cases where the modification is very small, the specific region is remodelled instead of modelling the whole structure.

4.3.3 Wet-lab experimentation methods

4.3.3.1 DNA design

Following the *in silico* validation of all the test sequence. The finalised constructs were reverse translated into their corresponding DNA sequences using backtranseq feature on EBI website (https://www.ebi.ac.uk/Tools/st/emboss_backtranseq/). Codon optimisation was performed using condonopt tool on IDT website (https://eu.idtdna.com/codonopt). The final constructs were obtained from Twist Bioscience company. NEB and SnapGene's Gibson assembly simulators were used to design the homologous arms to facilitate Gibson assembly. Primers were designed using the tools mentioned in Chapter 2 section 2.3.3. All the primers were sourced from Integrated DNA Technologies.

Primer name	Sequence
FullSeqFWd	AATTCAAAGGAGGTACCCACCA
FullSeqREV	AGGTAGATATCGCGGTACCCTTA
MUC1Dia1aREV	ACTGTTGACAATAATAGGTTGCAG
MUC1Dia1bFWD	AACCTATTATTGTCAACAGTGGAGTTC
MUC1Dia3aREV	TCTGGAGATAAAGTGTATTTTTAGCATT
MUC1Dia3bFWD	AAATACACTTTATCTCCAGATGTCCTC
MUC1Dia4aREV	TCTCGATCCCTAGCGCAAT
MUC14bFWD	TATTGCGCTAGGGATCGAGA
MUC1Dia5aREV	TGCCAGGACCCCAGTAATC
MUC1Dia5bFWD	TGATTACTGGGGTCCTGGCAC

4.3.3.2 Plasmid scale-up

OG176 plasmid (Oxford genetics) with Kanamycin resistance was chosen for producing the synthetic proteins. Plasmid scale-up was performed using the protocol described in Chapter 2 section 2.3.4. For plasmid extraction, overnight subcultures of the transformed bacteria are subjected through Monarch Plasmid miniprep kit (New England Biolabs) protocol.

4.3.3.3 Restriction digestion

Refer Chapter 2 section 2.3.5 for detailed methods

4.3.3.4 Gibson Assembly

Refer Chapter 2 section 2.3.6 for detailed methods

4.3.3.5 Validating cloning using colony PCR and Sanger sequencing

Refer Chapter 2 section 2.3.7 for detailed methods

4.3.3.6 *In vitro* transfection

Refer Chapter 2 section 2.3.8 for detailed methods

4.3.3.7 Binding assays

MCF7 (hMUC1-positive) and B16 (hMUC1-negative) cell lines were used to test the binding of synthetic proteins. Cells at different concentrations were blocked with 5 % BSA for 2 h followed by incubation with supernatant containing each test construct. Cells were washed 3 times and resuspended in PBS. 10 µl of each sample is taken in triplicates into a corning 96 well white plate. 50 µl of Coelenterazine substrate was added to each well and luminescence was measured using Promega GloMax® 96 luminometer.

4.3.4 *In vivo* methods

All animal procedures were performed in accordance with Health Products Regulatory Authority (HPRA) ethical guidelines. The project protocols were approved by animal ethics committee at University College Cork and the HPRA. All procedures were performed with care and effort to minimise pain and suffering.

4.3.4.1 Animals used in the study

6-8-week-old female BALB/c mice (weighing about 20 grams) were used for the studies. The animals were obtained from Envigo, UK. All the animals were monitored on a regular basis throughout the experimental period.

4.3.4.2 Bacterial administration to mice

Subcultures of the overnight bacterial strains were made 6 hours before the experiment. Bacteria were grown until an appropriate OD as per CFU requirements. The cells were harvested by centrifugation (3000 rpm for 20 min) and were washed 3 times with PBS. Mice were anaesthetised with Isoflurane throughout the procedure. Both the quadriceps of the mice were shaved to enable better access to the muscle. The 50 µl diluted cultures were administered intramuscularly using a 29G needle syringe.

4.3.4.3 Systemic administration of synthetic proteins

50-100µl of the synthetic protein supernatant were administered to the mice via IV through the lateral tail vein. All IV injections were performed 20-30 min before substrate administration or imaging.

4.3.4.4 Non-invasive *in vivo* imaging

The IVIS Lumina II Imaging system (Perkin Elmer) was used for bioluminescence imaging. All mice were kept under isoflurane-induced anaesthesia all through the imaging period. IVIS living image software was used for image visualisation and analysis.

4.3.4.5 *In vivo* transfection

Lipofection *In vivo* turbofect (Invitrogen, Thermofisher) was used for transfecting mouse quadriceps using lipofection. Manufacturer's protocol was followed to formulate the composition.

Electroporation: Nepagene 21 electroporator was used for this electroporation. Electroporation was carried out on the mouse quadriceps using a plate based electrode setup. The mice were anaesthetised using Ketamine (75 mg/kg) and Medetomidine: (1mg/kg) (IP injection). DNA was injected prior to electroporation (needle size range 26G-30G) at a depth of approximately 5 mm with 50 µl of DNA suspension in water. The injected site is then electroporated (8 pulses, 1000-1300 V/cm for 100 µs).

4.4 RESULTS

4.4.1 *In silico* validation and screening

Detailed workflow of all the *in silico* methods and strategies used here, are described in Chapter 2, (section 2.4).

4.4.1.1 Tumor Associated MUC1 antigen

Tumour associated human MUCIN 1 (MUC1) was chosen as a target to test the imaging strategy as a proof-of-concept. The amino acid sequence for MUC1 was obtained from PDB and the 3D structure was modelled using I-Tasser. Various structurally important domains were highlighted using UCSF Chimera. The antigenic sequence PDTRPAP and the potential binding site for antibodies is depicted in Figure 4.2.

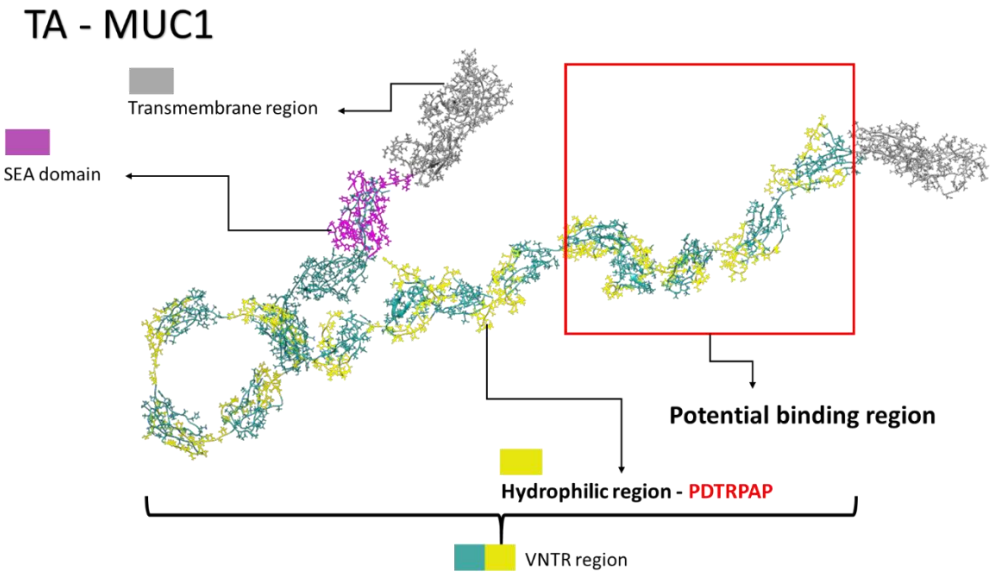


Figure 4.2: 3D structure of MUC1. *Potential binding site for various MUC1 targeting antibodies such as C595, is highlighted in the red box. The VNTR (Variable number tandem repeat) region is highlighted in blue and yellow. The VNTR region is a highly glycosylated and consists of 20 amino acid repeats. Yellow represents the hydrophilic region on MUC1. The structure has been modelled computationally using I-Tasser.*

4.4.1.2 Design elements and design rationale

In silico strategies, as shown in Chapter 2, were used to design synthetic proteins for targeting cancer cells. The amino acid sequence for the binding domain was retrieved from a ScFv clone previously shown to be capable of binding the MUC1 antigen on human breast carcinoma tissues [21]. An outline of different parts of the synthetic protein is shown in Figure 4.3.

Imaging part	Gluc
Binding part	VH and VL domains of anti MUC1 ScFv clone C595
Linkers	**GGGGS** for ScFv, **EAAAKGGGGSEAAAK** for connected two ScFvs
Detection tag	Flag
Solubility tag	SUMO

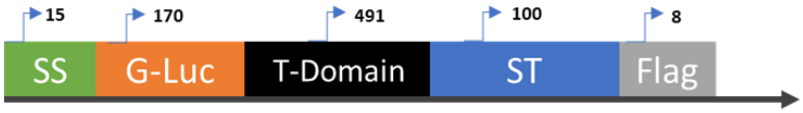


Figure 4.3: Different subparts of the test construct. *Test variants differ in the presence or absence of each subpart and their arrangement. A monobody (ScFv) version and a diabody version were chosen as two variants of T-Domain.*

Over 100 different test variants were designed and screened for best performers. After multiple iterations of redesign and remodelling, 8 different test constructs were selected for wet-lab testing. The 3D structure and the construct schematics of the selected ScFv and Diabody test variants are shown in Figure 4.4.

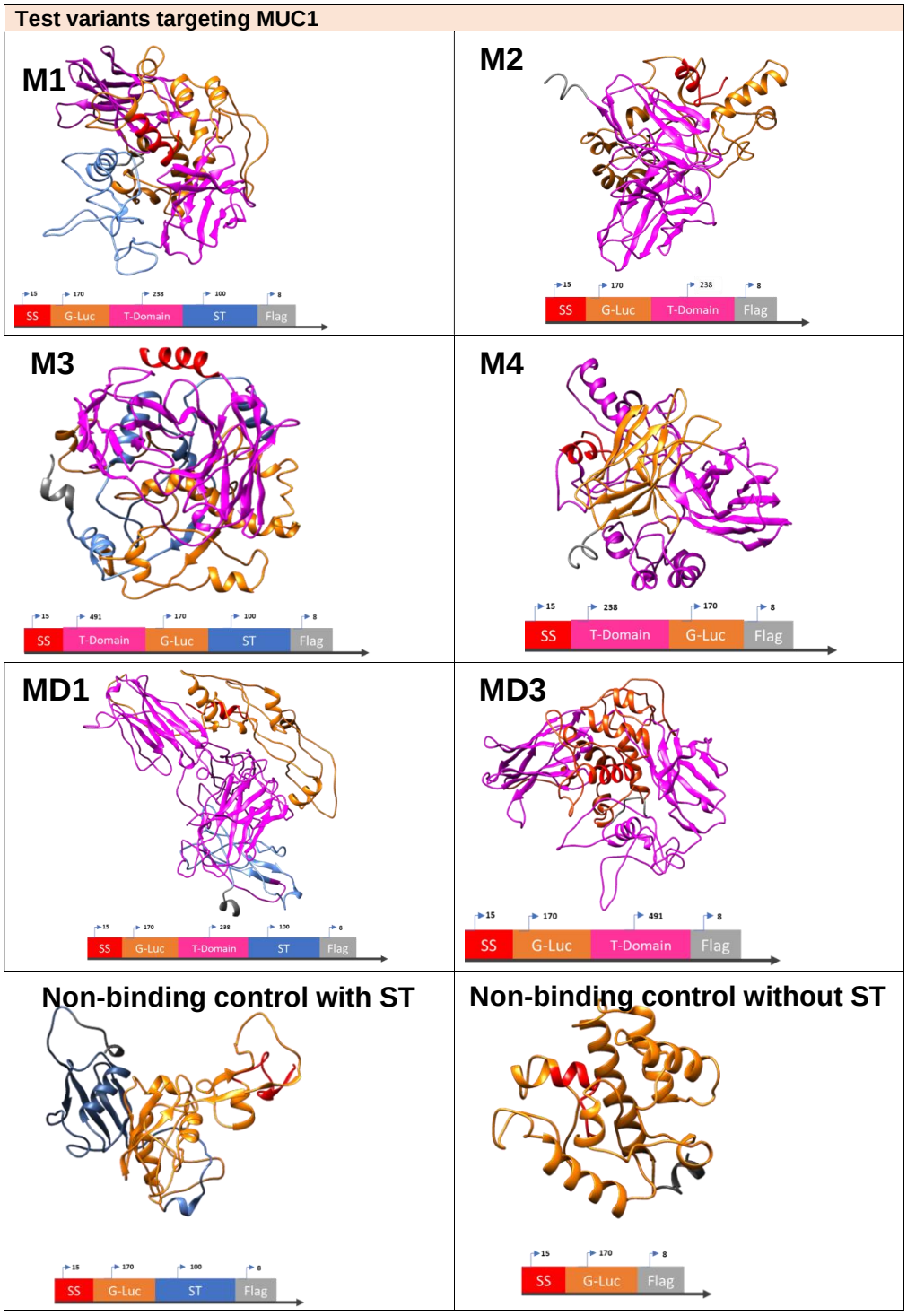


Figure 4.4: 3D structure and the construct schematics of the selected ScFv and Diabody test variants. *MD2 and MD6 lack the binding domain and were designed as internal negative control for binding. Constructs M1, M2, M3 and M4 represent the ScFv version and the constructs MD1, and MD3 represent the diabody format.*

4.4.1.3 Protein-Protein Docking

Protein docking was performed on all the test constructs to test the *in silico* binding affinity. MUC1 structure modelled by I-Tasser was used as the receptor. MUC1 is a large protein and performing full length docking was computationally-intensive. The VNTR region where C595 binds was selected and the docking was performed in this restricted region. Upon docking, 8 different potential binding conformations for each test construct were visualised using UCSF Chimer and the best conformation for each test construct was selected based on free energy. Figure 4.5 shows the monobody M3 bound to MUC1. The docking results and all the *in silico* data corresponding to the finalised test constructs are shown in Table 4.1.

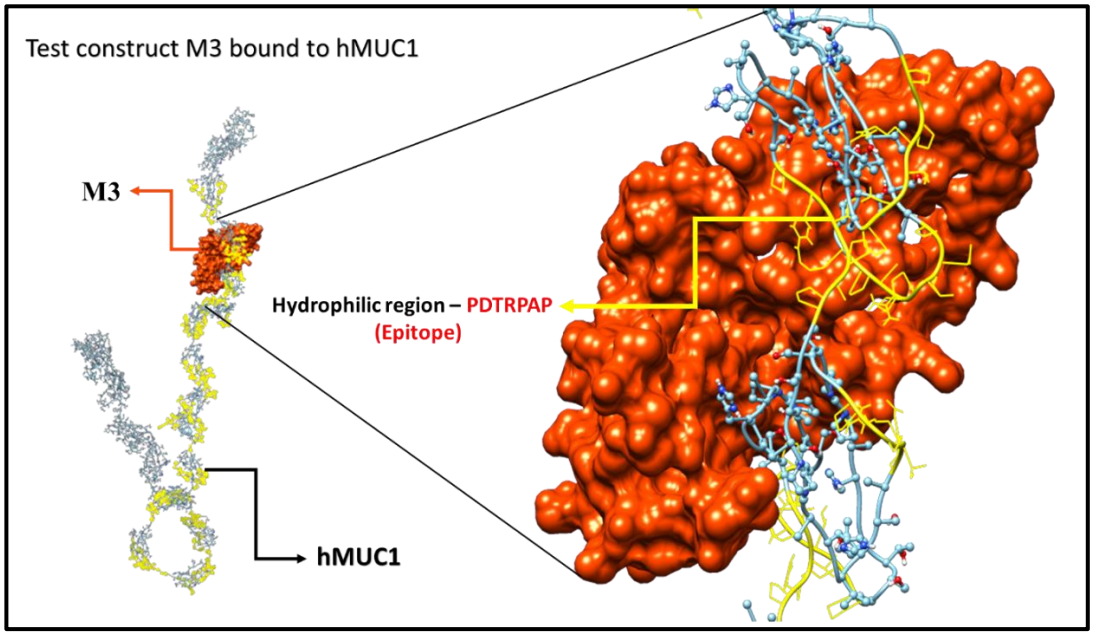


Figure 4.5: Test construct M3 (ScFv variant) bound to MUC1. *The whole chain of MUC1 structure is shown on the left. The zoomed-in section shows the Hydrophilic PDTRPAP region (Epitope), where the synthetic protein bind. Both the structures shown here were computationally modelled using I-Tasser.*

ID	C Score	TM score	RC score (%)	Hydrophobicity (%)	Solvent accessibility of the active site (%)	Docking affinity (ΔG)	Size
M1	0.04	20.45	84.3	44.787	59.888	-17.5	66
M2	0.17	21.44	88.4	45.777	60	-18	53.3
M3	0.08	18	83.5	44.888	58.888	-16.7	66.5
M4	0.22	4	75.5	45.194	61.509	-13.4	83.46
MD1	0.07	5.54	84.3	44.888	59.444	-17.4	66.5
MD3	0.0014	19	87.3	45.274	64.960	-18.6	37.3
MD2	0.16	19	89	45.333	65	-17.7	37.53
MD6	0.077	21.29	86.9	47.777	65.555	-15.2	24.33

Table 4.1: In silico data of selected test constructs, obtained from various computational tools. *From left to right, C-score (confidence score) from I-Tasser, TM score (template modeling score) as agreement between Rosetta and I-Tasser models, RC score (Ramachandran plot score), solvent accessibility of the active site (paratope regions), Free energy change from docking, size and instability index were tabulated for all the test variants.*

4.4.2 Wet-lab experimentation

General assay methods are described in Chapter 2, section 2.4.5. All assays were performed in triplicate.

4.4.2.1 Validating protein production

Following thorough *in silico* validation the selected test constructs were tested in wet-lab studies for their functioning. Transfected cell supernatant, containing test proteins, was collected 48 h post transfection and a series of luminescence assays have been conducted to test the protein production. Figure 4.6 shows the luminescence from all the selected constructs. In this case, test construct M1 was the best produced.

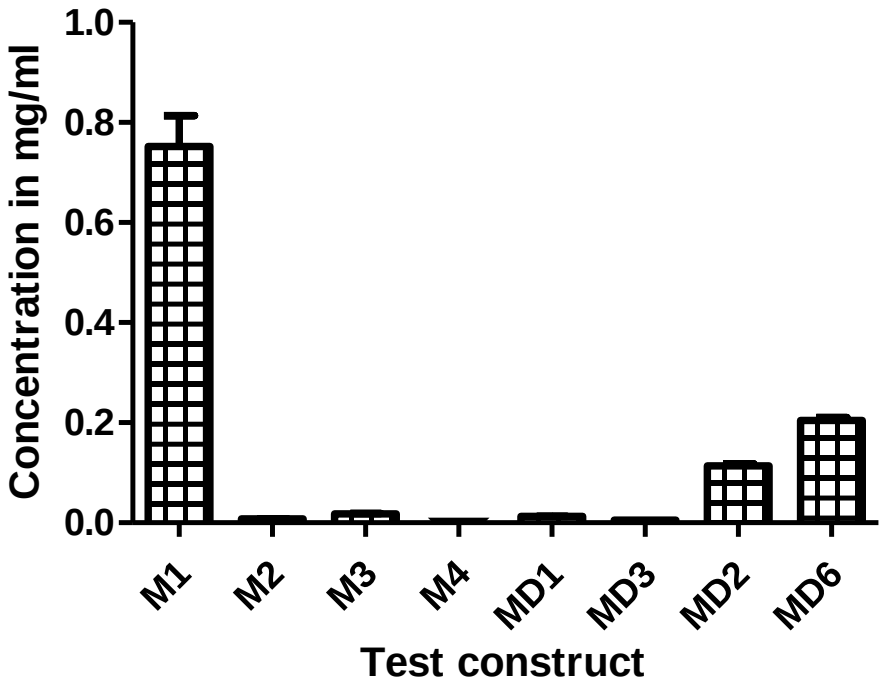


Figure 4.6: Secretion of various test proteins. *Relative luminescence units were correlated to concentration in mg/ml using a standard curve obtained from Gluc protein standards. MD2 and MD6 lack the T-domain and were used as internal controls. All assays were performed in biological triplicates.*

4.4.2.2 *In vitro* binding to MUC1

In vitro binding was confirmed by measuring and comparing luminescence signals after binding. MCF7 were chosen as the MUC1 positive cell line and B16 cell line was chosen as MUC1 negative cell line. The choice of cell lines was based on previous literature and data from human protein atlas and expression atlas [14, 22, 23]. Both the cells were treated with 10 µl of each test construct for 1 h, at room temperature. The cells were washed 3 times using PBS and 10 µl of the sample is taken into a corning 96 well white plate. Bound luminescence from each test construct is plotted in Figure 4.7. No significant correlation was observed when ScFv variants and Diabody variants were placed into groups. It was deemed that the data set size was also too small for such comparisons.

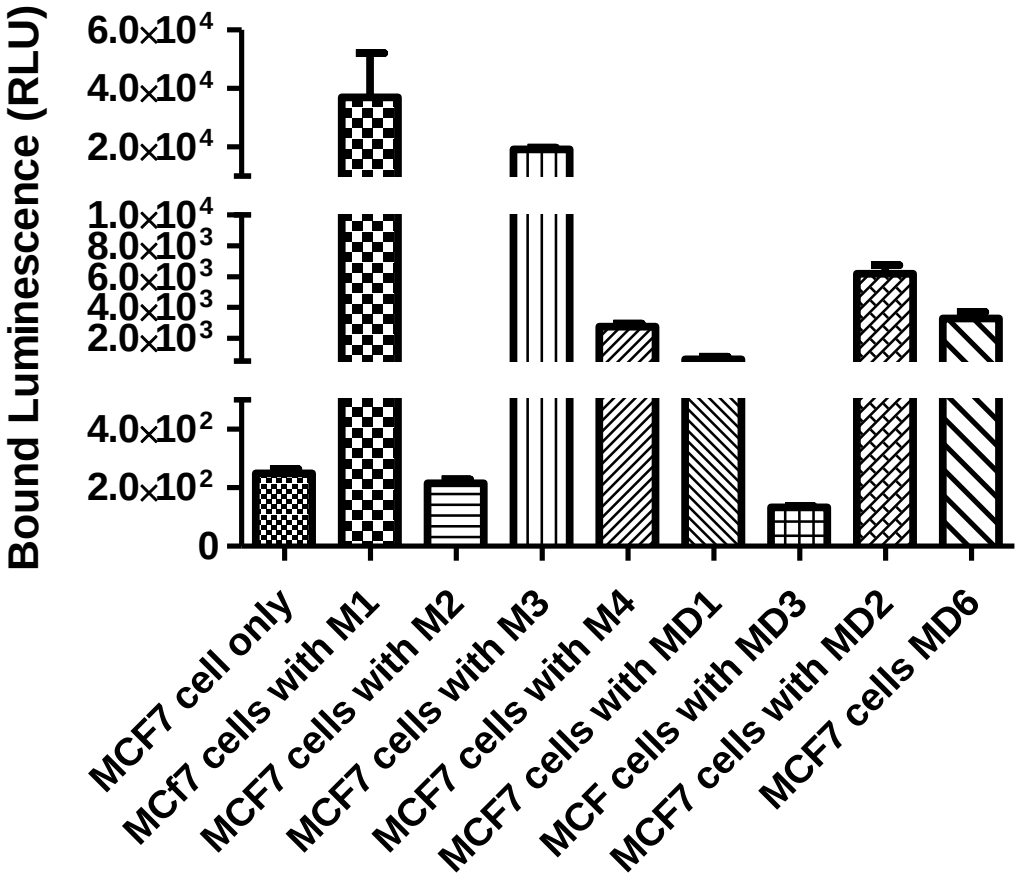


Figure 4.7: Bound luminescence of test variants to MUC1. *Luminescence after binding to MUC1 on MCF7 cells.* 10^6 *MCF7 cells were incubated with 10 µl of each test construct for 1 h. Test construct M1 emitted highest bound luminescence followed by M3.*

Figure 4.8 shows the bound luminescence per µg of synthetic protein. Luminescence emitted from each test construct (non-bound) was measured in parallel and bound luminescence normalised to 'amount of synthetic protein added (in µg)' was calculated mathematically. In this case, M3 and M4 outperformed all the other test constructs.

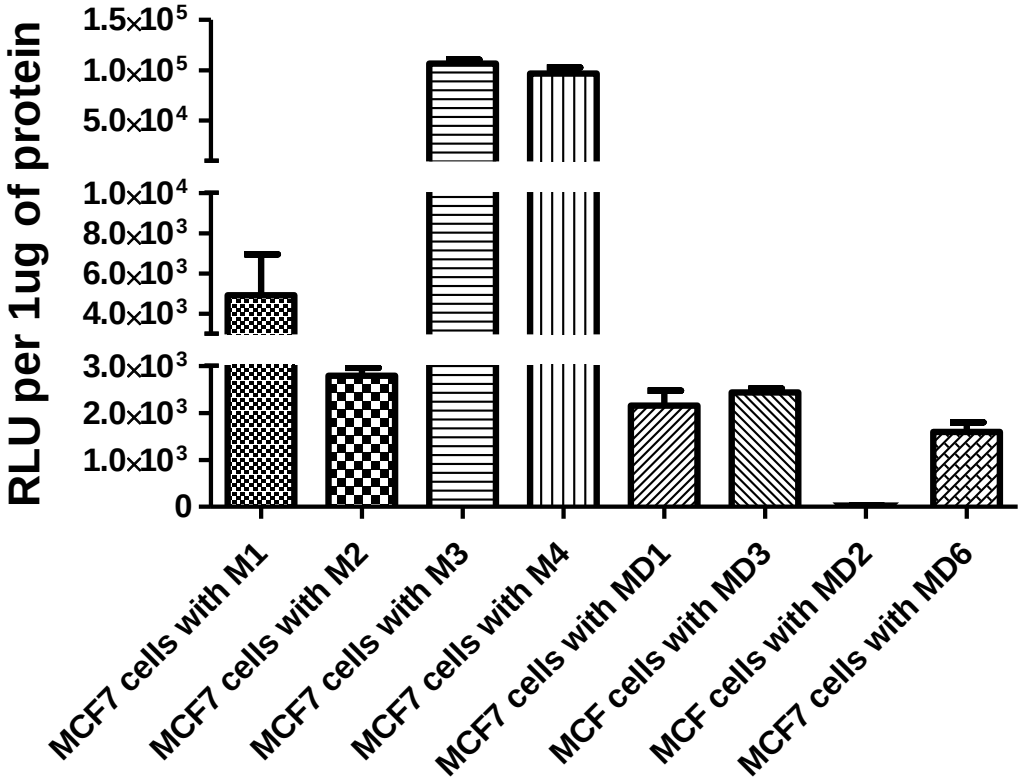


Figure 4.8: Bound luminescence normalised to protein quantity. *Test constructs M3 and M4 outperformed M1, which was producing highest bound luminescence in the previous non-normalised analysis.*

Selectivity of the synthetic proteins towards MUC1 positive cell lines was calculated by calculating ratio between bound luminescence (per amount of protein added) from MCF7 cells and B16 cells. Figure 4.9 shows the selectivity of all the test constructs per 1µg of protein added. Test constructs M1 and M3 presented the highest selectivity for MUC1.

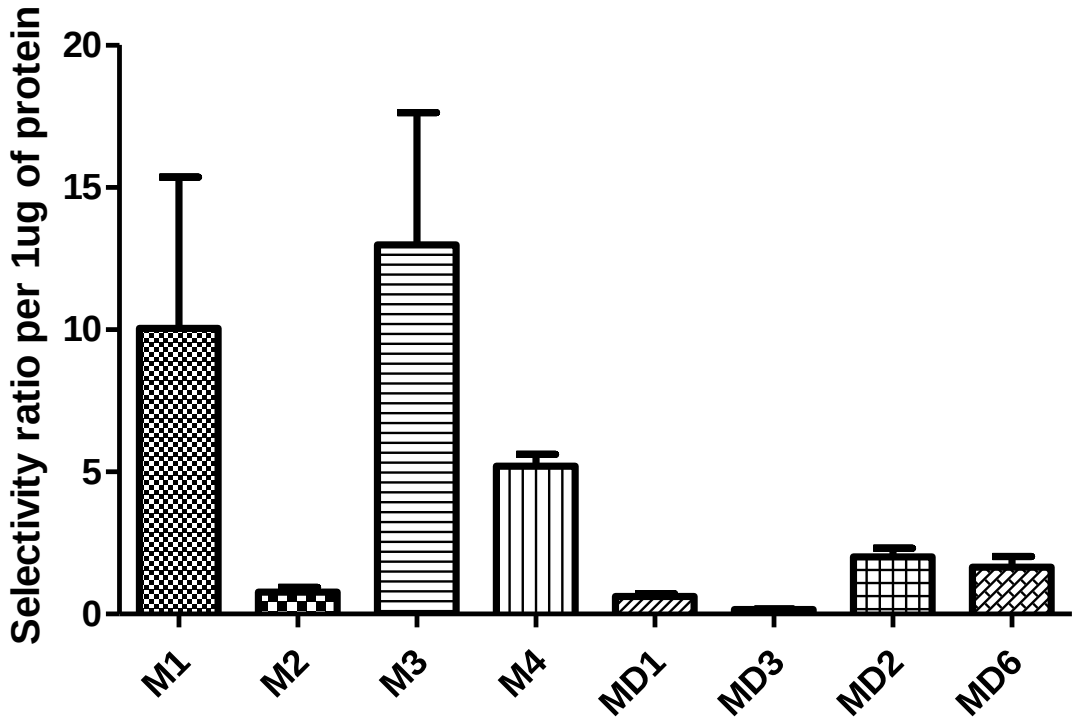


Figure 4.9: Selectivity of the synthetic proteins towards MUC1 positive cell lines. *M1 and M2 showed highest selectivity followed by M4. All the other test constructs have selectivity lesser than the non-binding controls MD2 and MD6.*

4.4.2.3 Selecting the best suitable candidate

The *in vitro* binding assays presented an interesting scenario while selecting for a best performing test construct. M1 showed the highest signal intensity, M3 showed the highest bound luminescence per amount of protein added, and M1 and M3 showed high selectivity. The test constructs are ranked in Table 4.2, based on each wet-lab binding analysis strategy.

Signal intensity	Normalised performance	Selectivity
M1	M3	M3
M3	M4	M1
M4	M1	M4
MD1	M2	M2
M2	MD3	MD1
MD3	MD1	MD3

Table 4.2: Test proteins ranked based on their performance in different wet-lab binding analysis strategies.

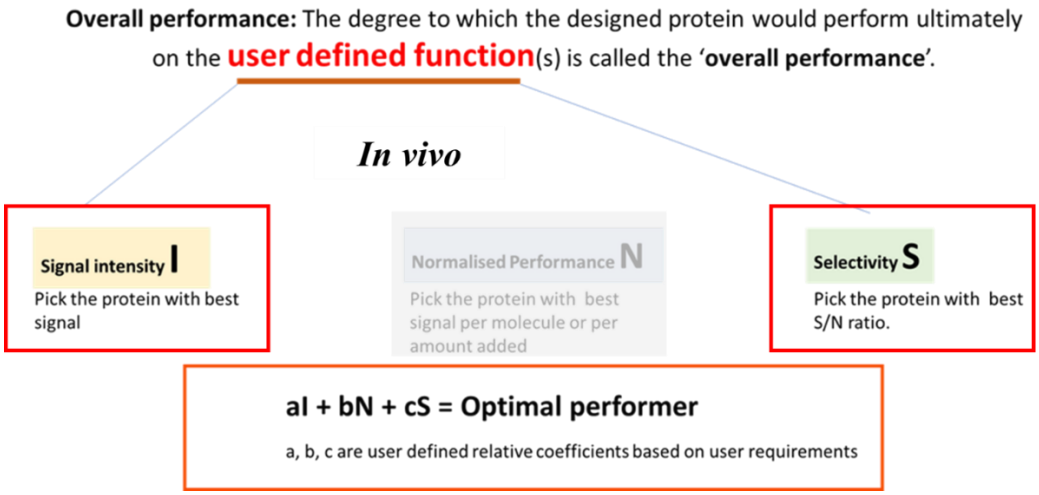


Figure 4.10: Selecting the best suitable candidate. *Signal intensity and selectivity were deemed to be the most important factors in further in vivo studies.*

M1 presented the highest bound luminescence and high selectivity. For preliminary *in vivo* testing, high signal intensity and high selectivity stand as the most important factors (Figure 4.10). Based on this hypothesis, M1 was selected as the best suitable candidate for further studies and all the further emphasis was placed in M1.

4.4.2.4 Dose response assays

Dose response experiments were performed using two MCF7 and B16 cells. Three different concentrations (by volume) of the synthetic proteins and three cell concentrations were used. Cells were treated with each synthetic protein for 1 h, at room temperature. Cells were washed 3 times and resuspended in PBS. All the samples, in triplicate, were placed in a Corning 96-well white plate. Luminescence was measured after adding 50 µl of coelenterazine. Cell dose response and synthetic protein dose response using test construct M1 is shown in Figure 4.11. In both instances, the bound luminescence from MCF7 cells was higher than bound luminescence from B16 cells.

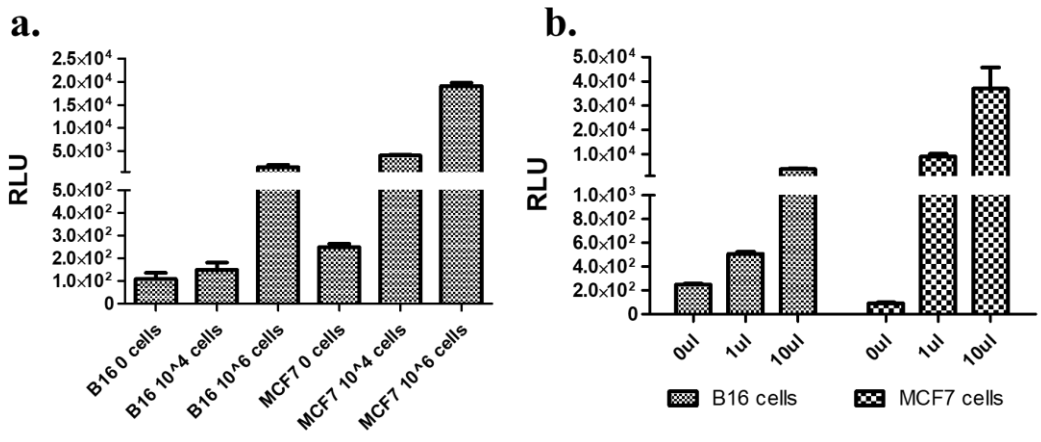


Figure 4.11: Cell and synthetic protein dose response of M1 (a) Cell dose response: *0,10^4, 10^6 cells of both B16 and MCF7 cell lines were treated with 10 µl of M1.* ***(b) Synthetic protein dose response:*** *0, 1 µl, 10 µl of M1 have been added to 10^6 cells of both the cell lines. As expected, the luminescence increased with the increase in number cells/synthetic protein concentration.*

4.4.3 *In vivo* imaging

Following successful *in vitro* validation, the synthetic proteins, targeting MUC1 and *S. aureus* ClfA, were tested in murine models. It should be noted that significant 'trial and error' optimisation was required to identify the appropriate time intervals for imaging, and multiple *in vivo* studies were performed in advance to optimise for the below shown studies (data not shown).

4.4.3.1 *In vivo* synthetic protein dose response (M1)

M1 in three different concentrations, was administered systemically by IV injection via lateral tail vein. In all cases, the synthetic proteins were diluted in PBS and reconstituted to a 100 µl volume. Subcutaneous injections of 10^7 MCF7 and B16 cells were given on each side of the mouse. 20 min after administering the cancer cells, 50 µl of coelenterazine was injected into the cell locations. Mice were anaesthetised with Isofluorane and were subjected to imaging. All experiments were performed in triplicate.

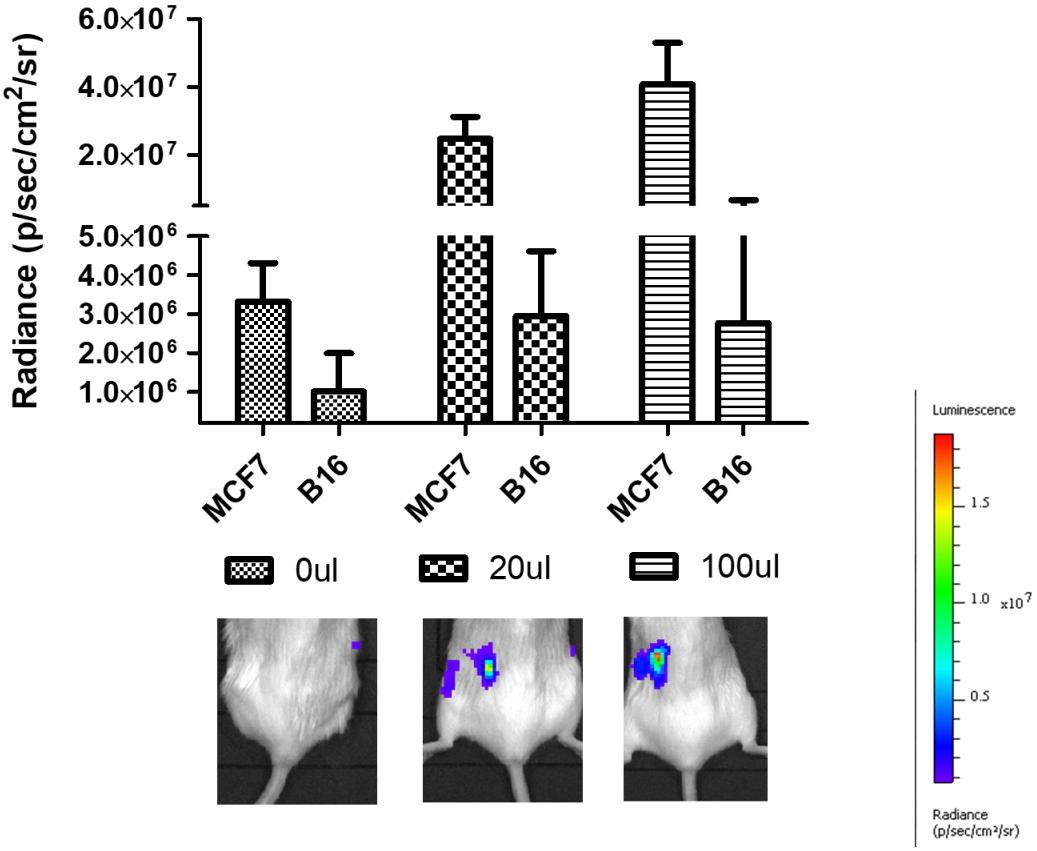


Figure 4.12: M1 dose response and target specificity in vivo: *An increase in luminescence was observed with respect to the increase in synthetic protein concentration. Luminescence from left subcutaneous pocket, with MCF7 cells, produced a significantly (p 0.0087) higher signal than the right side, with B16 cells.*

4.4.3.2 *In vivo* cell dose response (M1)

100 µl of M1 was administered intravenously via lateral tail vain to all the mice. Mice were injected subcutaneously with different concentrations of cancer cells (0, 10^5 and 10^7cells). 20 min after administering the cancer cells, 50 µl of coelenterazine was injected into the cell locations. Mice were anaesthetised and subjected to imaging. All the experiments were performed in triplicate.

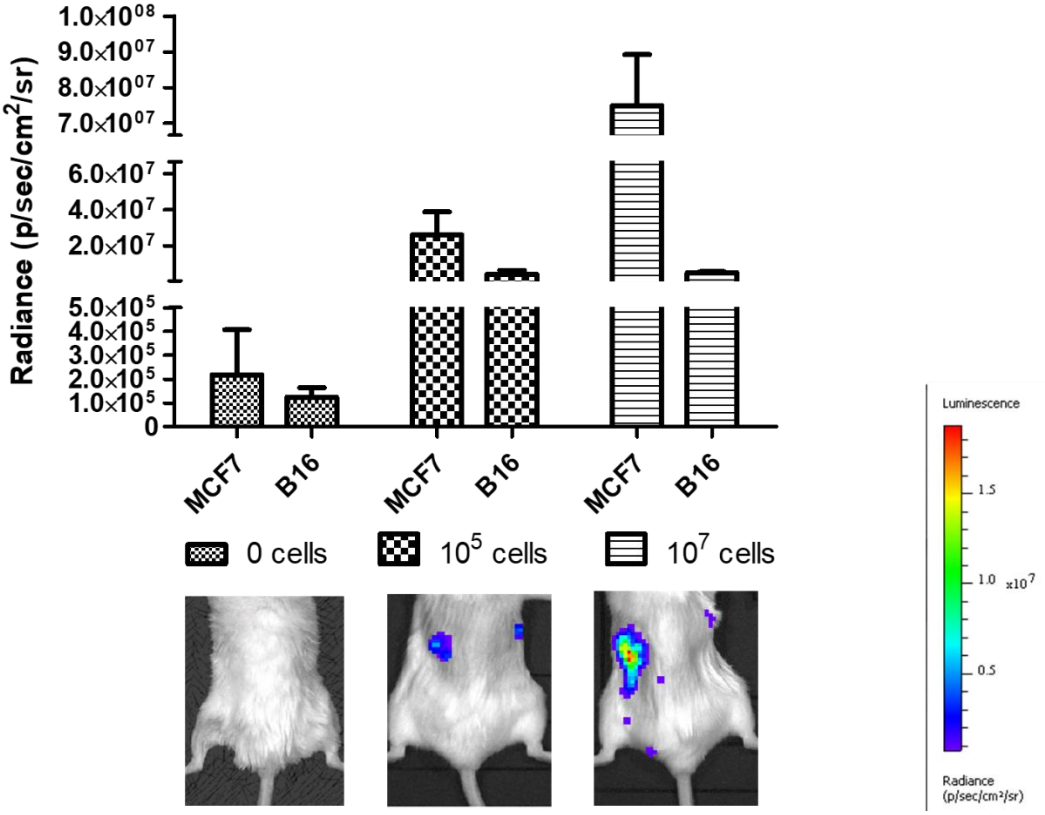


Figure 4.13: In vivo cell dose response. *An increase in luminescence was observed in relation to the increase in number of cells. Luminescence from left subcutaneous pocket, with MCF7 cells, produced a higher signal than the right side, with B16 cells.*

4.4.3.3 *In vivo* synthetic protein dose response (S1), *S. aureus* targeting synthetic protein

S1 in three different concentrations, was administered systemically by IV injection via lateral tail vein. In all cases, the synthetic proteins were diluted in PBS and reconstituted to a 100 µl volume. Mice were anaesthetised and 10^7 *S. aureus* cells diluted in PBS (in 50 µl) were administered intramuscularly on the left quadricep. 20 min after the IM injection, 50 µl of coelenterazine was injected into both the quadriceps. Mice were anaesthetised with isoflurane and were subjected to imaging. All experiments were performed in triplicate.

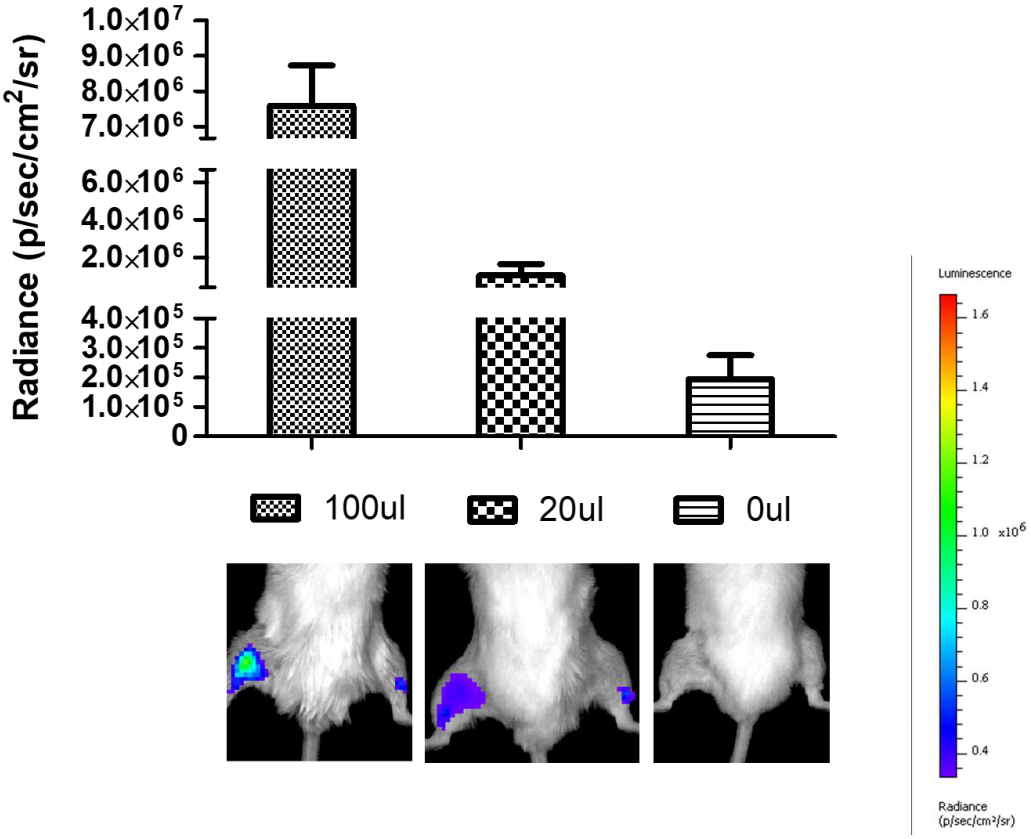


Figure 4.14: S1 dose response in vivo *100 µl, 20 µl and 0 µl of S1 was administered systemically via tail vein. 10^7 S. aureus cells were injected (IM) on the left quadriceps of mice. An increase in luminescence was observed with the increase in concentration of S1.*

4.4.3.4 *In vivo* bacterial cell dose response (S1)

100 µl of S1 was administered intravenously via lateral tail vein to all mice. Mice were anaesthetised with isoflurane and different concentrations of *S. aureus* cells (0, 10^5 and 10^7 cells) diluted in PBS (in 50 µl) were administered intramuscularly to the left quadriceps. 20 min after administering the bacterial cells, 50 µl of coelenterazine was injected into the cell locations. Mice were anaesthetised and subjected to imaging (Figure 4.15). A dose response was not evident in this case. While both 10^5 and 10^7 bacteria groups displayed higher luminescence than no bacteria, the 10^7 group did not produce higher luminescence than the 10^5 group. Furthermore, some off-target luminescence was observed in the right quadriceps of mice injected with 10^5 cells, perhaps due to basal level of coelenterazine reacting with mouse blood.

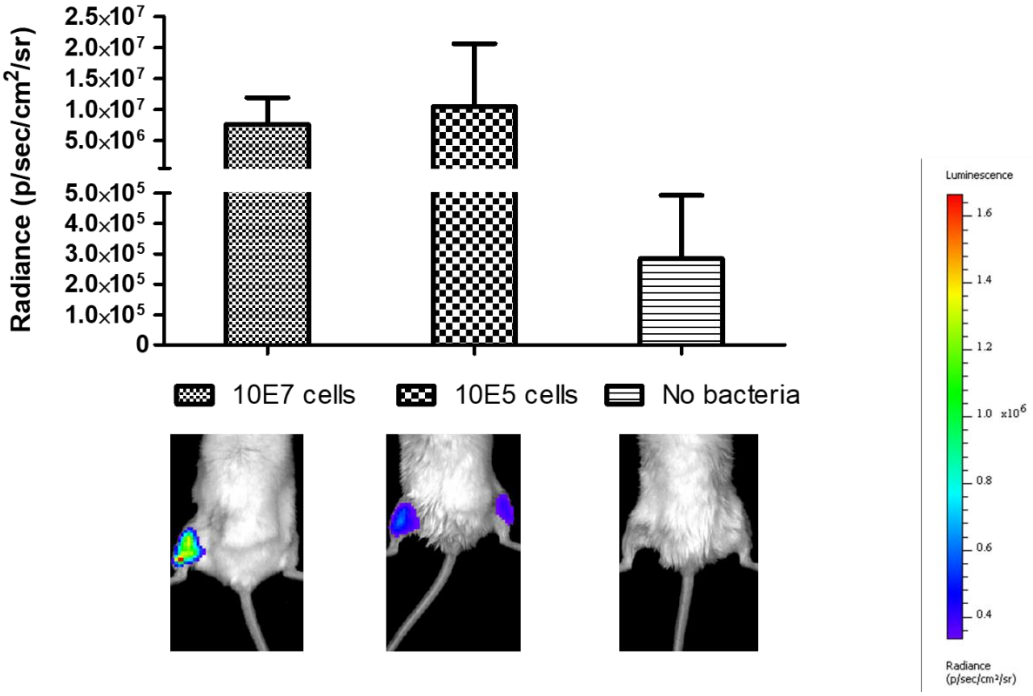


Figure 4.15: In vivo bacterial cell dose response study. *100 µl of S1 was administered systemically via tail vein. 0, 10^5 and 10^7 S. aureus cells were injected (IM) on the left quadriceps of mice. In this case the luminescence from 10^5 cells was slightly higher than luminescence from 10^7 cells (n=2). Some off-target luminescence was observed in the right quadriceps of mice injected with 10^5 cells.*

4.4.3.5 GlucS1 vs NanolucS1

As discussed in Chapter 2, NanoLuc was deemed likely to have a significant advantage over Gluc for *in vivo* imaging. Previous literature shows that NanoLuc is brighter and has improved half-life over Gluc. To test this *in vivo,* 100 µl of GlucS1 and NanolucS1 were administered systemically to mice by IV injection via lateral tail vein. Mice were anaesthetised and 10^7 *S. aureus* cells diluted in PBS (in 50 µl) was administered intramuscularly to the right quadriceps. 20 min after IM injection, 50 µl of coelenterazine (GLuc) or furamazine (NanoLuc) was injected into both left and right quadriceps. Mice were anaesthetised and subjected to imaging. Mice were imaged for 2 h at various time points. Figure 4.16 shows the signal intensities of GlucS1 and NanoLuc S1 with respect to time. This study was performed in triplicate. NanoLucS1 was brighter than GlucS1 throughout the experiment and produced a significantly higher signal even after 100 min ($p < 0.0001$).

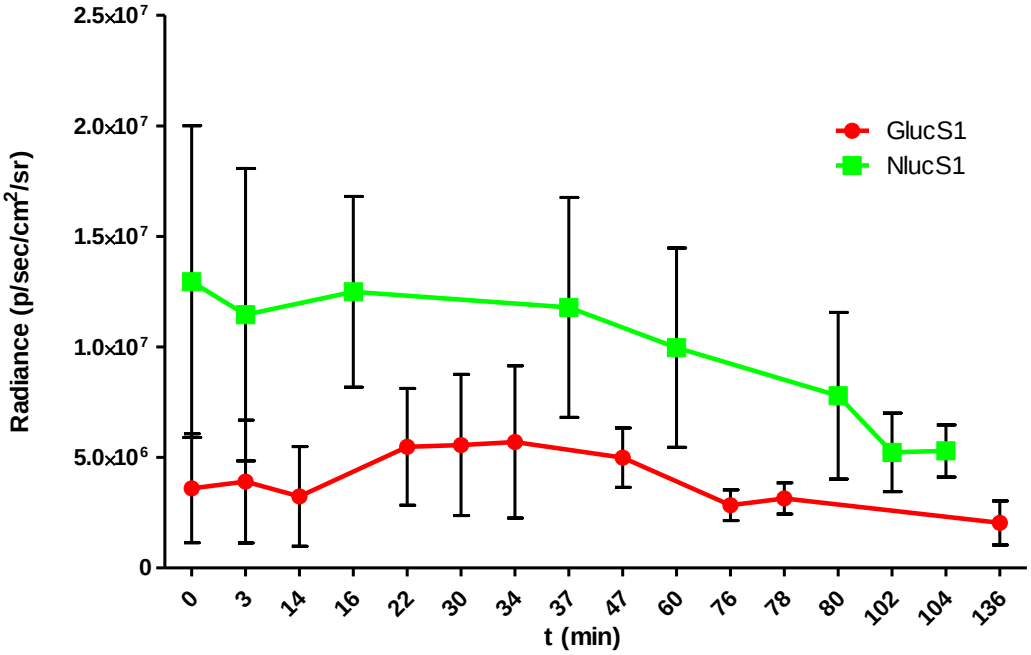


Figure 4.16: GlucS1 vs NanoLucS1. *10^7 S. aureus cells were injected into the right quadripceps of mice. 100µl of test protein was injected via tail vein. NanoLucS1 (shown in green) showed a significantly higher signal intensity throughout the time course* (p <0.0001)**.**

4.4.4 *In vivo* production of synthetic proteins

Following validation that systemically administered proteins function as imaging agents *in vivo*, it was investigated if these proteins could be produced by the mouse *in vivo* to achieve the same effect. A schematic of this concept is shown in Figure 4.17.

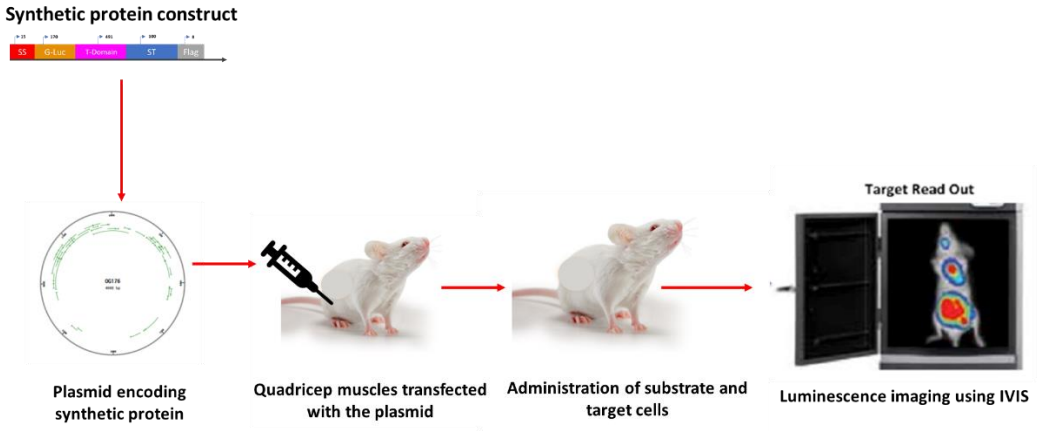


Figure 4.17 Schematic of in vivo production of targeted synthetic proteins: *The concept involves administration of plasmid DNA to mouse quadriceps to induce systemic production of the targeting luciferase. The systemically-circulating protein binds to the specified target, facilitating BLI of the target cells.*

Quadriceps of mice were transfected using electroporation or lipofection with DNA encoding NanolucS1 or Fluc (as a control for transfection). After 72 h, 10^7 *S. aureus* cells were injected into the left quadriceps of mice. 100 µl of furamazine (NanoLuc) or luciferin (FLuc) was injected by IV via lateral tail vein. Mice were imaged after anaesthetising with isoflurane. For lipofection groups, no luminescence was observed in any mouse, including the FLuc transfection control group, indicating insufficient DNA transfection to produce luminescent protein (data not shown). This was repeated, producing the same result.

For electroporation groups, high-level, quadriceps-localised luminescence was evident in the group transfected with Fluc, indicating successful DNA transfection and localised intracellular luminescent protein production (FLuc is not secreted). However, no luminescence was detected in any mice transfected with Nanoluc S1, either at the quadriceps site of transfection, or distally at target cell sites, indicating insufficient transfection and/or protein production for detection (Figure 4.17).

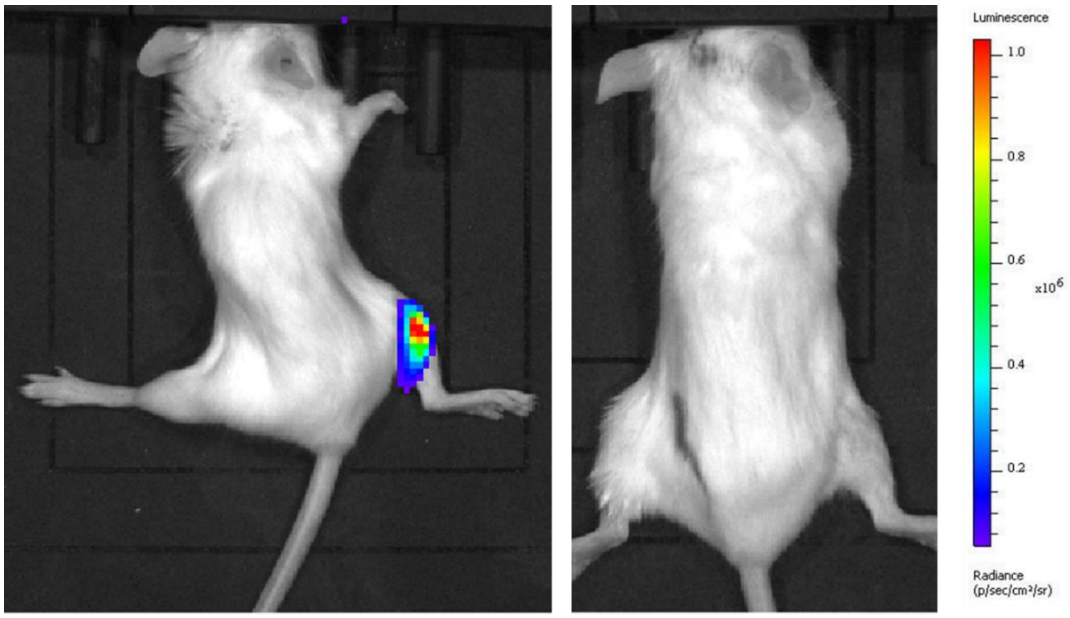


Figure 4.17: Transfection of mice quadriceps for in vivo protein production. *(a) Right quadricep was electroporated with plasmid encoding Fluc. (b) Right quadricep was electroporated with plasmid encoding NanolucS1. No luminescence was observed in the Nanoluc mice group.*

While the *in vivo* production strategy proved unsuccessful in this study, it is possible that further optimisation of various parameters might bring this concept to reality.

4.5 DISCUSSION

This work presents use of synthetic targeted proteins as an *in vivo* imaging strategy. Two different targets, MUC1 and ClfA, were chosen for the study. However, the *in silico* strategies described could be adopted for various other targets. Synthetic proteins were built to bind to hMUC1, using the *in silico* aided-design strategy described in Chapter 2. Design rationale and choice of subparts were adopted from Chapter 2, to have high resemblance. Over 100 different test variants were designed and modelled with the aim of obtaining the optimal structural conformations. The binding domain was built by incorporating minimal regions of C595 ScFv. Immense care was taken to ensure maximum exposure of the active sites. Over 10 different linker type and size combinations were sampled. Designing a diabody format required additional care while choosing linkers due to its bigger size. Gluc was chosen as the imaging part due to its small size and, due to its ATP-independence, ability to function outside the cellular environment where ATP is scarce.

In vitro testing was carried out to validate secretion and functioning (binding and luminescence) of the synthetic proteins. Throughout the work, luminescence was used to inform the protein production, binding and selectivity. The goal of the *in vitro* assays was to find a best suitable candidate for use in *in vivo* imaging. The bound-luminescence data presented a challenging and interesting case while selecting for the best candidate. M1 showed the highest luminescence per volume, M3 showed highest luminescence per protein weight and M1 and M3 showed similar selectivity for the target. Solving this issue required the application of 'optimal performer' logic that was described in Chapter 2. Detecting bound-luminescence *in vivo* requires high selectivity for the targeted cells. M1 and M3 both showed appreciable selectivity. Luminescence per volume of M3 was noticeably lower than M1. The low signal intensity of M3 would play a detrimental role for the overall goal (balance between high signal and high target binding). This solidified the argument of selecting of M1 over M3. The optimal performer logic helped to guide the screening process based on the user-defined overall function. In future, adding

arbitrary values to the coefficients of S, N and I, would provide an empirical basis while screening for the best performer.

The *in vitro* studies provided a preliminary understanding of the basic functioning of the protein. However, questions regarding the structural integrity and stability would have been answered through more wet-lab validation. In this context, Western blots were performed using anti-Flag antibodies, as per Chapter 5. However, after multiple attempts with both ClfA and MUC1 proteins, this was unsuccessful (data not shown). These data would have been beneficial to confirm the size, integrity and concentration of the synthetic proteins. Similar anti-Flag Western blots were readily performed in Chapter 5 to show the size and validate protein production. The failure to detect synthetic proteins on Western blots could be due to some features unique to these synthetic proteins.

The ability of the synthetic protein to circulate systemically throughout the body and localise to a specific target to produce an imageable luminescence signal acts as a proof-of-concept for *in silico* aided synthetic protein design of targeted proteins for *in vivo* use (therapeutic, imaging etc). *In vivo* studies using systemically-administered M1-Gluc indicated localisation of this protein to target cells. In this study, +/- MUC1 cell lines (MCF7 and B16) instead of solid tumor models due to quick turnover time of experiments. However, significant 'trial and error' optimisation was required to identify the appropriate time intervals for imaging. Although the synthetic proteins displayed strong localisation, in the case of S1, use of off-target cells would have further validated the target-specific binding. Further assays such as FACS would also reconfirm the binding. In both the *in vivo* studies using M1 and S1, administering the mice with a non-targeting binding antibody would have been beneficial to rule out the chances of accidental accumulation. Further, using solid tumor models and a complete dose response experiments showing the signal saturation would solidify the confirmation of binding. As the study proceeded, it was determined that the short *in vivo* half-life of the Gluc-colanterizine system may represent a limiting factor in performing these studies. Literature on Nanoluc guided towards testing of Nanoluc and its substrate as the imaging subpart in one of the test constructs. In Chapter 2, a variant of S1 with Nanoluc was designed and tested *in vitro*. These assays didn't show any noticeable *in vitro* advantage when Gluc was replaced with Nanoluc. However, *in vivo*, the NanolucS1 construct showed a significant improvement

in signal intensity (Figure 4.16). It is to be noted that both Coelenterazine and Furamazine have been tested as a substrate against NanolucS1 and in this study Coelenterazine was shown to be brighter when compared to Furamazine.

Following the findings that systemic administration of Nanoluc S1 validated the functioning of the synthetic proteins as *in vivo* imaging agents, studies were conducted to investigate if these proteins could be produced by the mouse *in vivo* to achieve the same effect. Transfecting quadriceps muscle was hypothesised to serve as a continuous protein producing reservoir [24] (Figure 4.17). This *in vivo* production strategy proved unsuccessful in this study. This could be due to a number of reasons such as (i) low protein production, (ii) poor secretion, (iii) immune response to sustained protein production etc. With further studies on the *in vivo* properties of the synthetic proteins and optimising design parameters, it might be possible to bring this concept to reality. In this study, plasmid with a moderate strength constitutive promoter has been used. Moderate strength plasmid was chosen to avoid the chance of toxicity of the produced protein. Considering the failure to obtain a signal from the transfected quadricep as well, further experiments using plasmids with stronger promoters could be explored.

Systemic administration of the various synthetic proteins showed no toxic or any adverse reaction during this study. Ethical constraints at the time of experiments directed towards the use of BalbC instead of nude mice. Although the successful *in vivo* targeting and imaging validated the original intended function, further experiments are required to test the pharmacokinetics, pharmacodynamics and long-term toxicity of the synthetic proteins. For instance, *in silico* tools could be used to identify and eliminate commonly found immunogenic motifs from the designed construct. *In vitro* synthesis *in vivo* testing of multiple test variants would be useful to find the least immunogenic candidate.

Throughout the study, mathematical and computational approaches were used at various stages of design, model, build and test. Including the concept of Function2Form bridge, detailed in Chapter 3, the data from the wet-lab studies could be used to train the machine learning models of F2F bridge. This adds the learn step to the ‘design-model-build-test’ approach of modern synthetic biology. However, predicting *in vivo* performance requires additional data from murine models such as their pharmacokinetics,

pharmacodynamics, immunogenicity and various assays facilitating the translation of the *in vivo* work to clinical stage.

In future, F2F bridge would be an integral component in *in silico* aided design and screening of synthetic proteins.

4.6 CONCLUSION

In this study, multiple variants of synthetic proteins targeted to human MUC1 were modelled and validated using various computational tools to inform and guide downstream wet-lab experiments, as per Chapter 2. Wet-lab studies were conducted to validate MUC1 protein production and functioning (luminescence and specific target binding) *in vitro*. For both MUC1 and ClfA targeted proteins, *in vivo* luminescence imaging studies involving systemic intravenous (IV) administration of proteins, validated synthetic protein specific accumulation at target cell locations within mice as evidenced by localised luminescence. This study serves as a proof-of-concept for using targeted reporter proteins for *in vivo* imaging.

4.6 REFERENCES

1. Hsu, P.D., E.S. Lander, and F. Zhang, *Development and applications of CRISPR-Cas9 for genome engineering.* Cell, 2014. **157**(6): p. 1262-78.
2. Huang, P.S., S.E. Boyken, and D. Baker, *The coming of age of de novo protein design.* Nature, 2016. **537**(7620): p. 320-7.
3. Byrne, W.L., et al., *Use of optical imaging to progress novel therapeutics to the clinic.* J Control Release, 2013. **172**(2): p. 523-34.
4. Tangney, M. and K.P. Francis, *In vivo optical imaging in gene & cell therapy.* Curr Gene Ther, 2012. **12**(1): p. 2-11.
5. Contag, C.H. and M.H. Bachmann, *Advances in in vivo bioluminescence imaging of gene expression.* Annu Rev Biomed Eng, 2002. **4**: p. 235-60.
6. Badr, C.E. and B.A. Tannous, *Bioluminescence imaging: progress and applications.* Trends Biotechnol, 2011. **29**(12): p. 624-33.
7. Tung, J.K., et al., *Bioluminescence imaging in live cells and animals.* Neurophotonics, 2016. **3**(2): p. 025001.
8. Luker, K.E., et al., *In vivo imaging of ligand receptor binding with Gaussia luciferase complementation.* Nat Med, 2011. **18**(1): p. 172-7.
9. Suzuki, T., et al., *Real-time bioluminescence imaging of a protein secretory pathway in living mammalian cells using Gaussia luciferase.* FEBS Lett, 2007. **581**(24): p. 4551-6.
10. van Rijn, S., et al., *Functional multiplex reporter assay using tagged Gaussia luciferase.* Sci Rep, 2013. **3**: p. 1046.
11. Venisnik, K.M., et al., *Bifunctional antibody-Renilla luciferase fusion protein for in vivo optical detection of tumors.* Protein Eng Des Sel, 2006. **19**(10): p. 453-60.
12. Luker, K.E. and G.D. Luker, *Bioluminescence imaging of reporter mice for studies of infection and inflammation.* Antiviral Res, 2010. **86**(1): p. 93-100.
13. Hattrup, C.L. and S.J. Gendler, *Structure and function of the cell surface (tethered) mucins.* Annu Rev Physiol, 2008. **70**: p. 431-57.

14. Walsh, M.D., et al., *Heterogeneity of MUC1 expression by human breast carcinoma cell lines in vivo and in vitro.* Breast Cancer Res Treat, 1999. **58**(3): p. 255-66.
15. Nath, S. and P. Mukherjee, *MUC1: a multifaceted oncoprotein with a key role in cancer progression.* Trends Mol Med, 2014. **20**(6): p. 332-42.
16. Yang, J. and Y. Zhang, *Protein Structure and Function Prediction Using I-TASSER.* Curr Protoc Bioinformatics, 2015. **52**: p. 5 8 1-5 8 15.
17. Weitzner, B.D., et al., *Modeling and docking of antibody structures with Rosetta.* Nat Protoc, 2017. **12**(2): p. 401-416.
18. Kaufmann, K.W., et al., *Practically useful: what the Rosetta protein modeling suite can do for you.* Biochemistry, 2010. **49**(14): p. 2987-98.
19. Pettersen, E.F., et al., *UCSF Chimera--a visualization system for exploratory research and analysis.* J Comput Chem, 2004. **25**(13): p. 1605-12.
20. Trott, O. and A.J. Olson, *AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading.* J Comput Chem, 2010. **31**(2): p. 455-61.
21. Denton, G., et al., *Production and characterization of a recombinant anti-MUC1 scFv reactive with human carcinomas.* Br J Cancer, 1997. **76**(5): p. 614-21.
22. Uhlen, M., et al., *Proteomics. Tissue-based map of the human proteome.* Science, 2015. **347**(6220): p. 1260419.
23. Papatheodorou, I., et al., *Expression Atlas: gene and protein expression across multiple studies and organisms.* Nucleic Acids Res, 2018. **46**(D1): p. D246-D251.
24. Morrissey, D., et al., *Control and augmentation of long-term plasmid transgene expression in vivo in murine muscle tissue and ex vivo in patient mesenchymal tissue.* J Biomed Biotechnol, 2012. **2012**: p. 379845.

Chapter 5

In silico aided-engineering of self-assembling protein cages

Table of Contents

5.1 ABSTRACT 203

5.2 INTRODUCTION 204

5.2.1 Interface design and potential applications of self-assembly 204

5.2.2 Computational protein-protein interaction design 205

5.2.2.1 Knowledge based methods vs protein docking/assembly 205

5.2.3 Understanding protein cages 206

5.2.3.1 Assembly mechanics and encapsulation kinetics 211

5.2.4 Visualising self-assembly 211

5.3 MATERIALS AND METHODS 213

5.3.1 *In silico* experimental methods 213

5.3.1.1 Design overview 213

5.3.1.2 Construct design 215

5.3.1.3 *In silico* tools used for design, modeling and validation 215

5.3.1.4 DNA design and synthesis 215

5.3.2 Wet-lab methods 216

5.3.2.1 Plasmid amplification and extraction 217

5.3.2.2 Restriction digestion 217

5.3.2.3 Gibson Assembly 217

5.3.2.4 Validating cloning using colony PCR and Sanger sequencing 217

5.3.2.5 Inducing protein expression 218

5.3.2.6 Analysing protein expression 218

5.3.2.7 Fluorescence assays and validating FlAsH binding. 219

5.4 RESULTS 220

5.4.1 *In silico* design and modelling of self-assembling cage 220

5.4.1.1 *In silico* validation of engineered cysteines 222

5.4.1.2 *In silico* validation of cage bio-assembly and functioning of inserted cysteines 226

5.4.1.3 *In silico* validation of integrity of the modelled protein structures..... 229

5.4.1.4 F2F plots to visualise overall performance..... 233

5.4.2 Wet-lab experimentation..... 235

5.4.2.1 Validating protein production..... 235

5.4.2.2 Self-assembly verification strategy..... 237

5.4.2.3 Whole-cell FlAsH-EDT2 based fluorescence assays confirming oligomerisation..... 239

5.5 DISCUSSION..... 243

5.6 REFERENCES..... 245

5.1 ABSTRACT

Background Self-assembling protein cages are abundant in nature. Viruses, bacterial microcompartments, ferritins and heat shock proteins are some examples of these highly organised protein structures. The spontaneous assembling and disassembling of such proteins presents promising applications in targeted release and encapsulation of drugs. Assembly mechanics of these self-assembling proteins rely on the composition of their interactive protein interfaces. Engineering protein interfaces helps to understand protein interactions and aid real-time monitoring of spontaneous self-assembly to visualise the encapsulation and release mechanics.

Aims The aim of this work was to (i) use *in silico* methods to engineer self-assembling monomers and (ii) to develop a wet lab method to validate the self-assembly using fluorescence.

Methods For *in silico*-aided protein design, computational tools such as I-TASSER and Rosetta were used for 3D structure modeling and UCSF Chimera for protein visualisation and identifying appropriate cysteine insertion locations to enable employment of a FLAsH-EDT2 fluorescence assay to report peptide interaction. DNA constructs were generated for protein production in *E. coli.* Protein production was confirmed by Western blot. Self-assembly was examined using a whole-cell fluorescence assay.

Results All proteins were successfully produced and confirmed by Western Blot. Whole-cell fluorescence assays provided the evidence that supports the interaction between FlAsH-EDT2 and the engineered bi-partite cysteine residues, indicating oligomerisation of monomer proteins.

Conclusion In this work, proof-of-concept of a novel method to visualise self-assembly is introduced, by computationally inserting bi-partite cysteine residues at the protein interactive interfaces.

5.2 INTRODUCTION

5.2.1 Interface design and potential applications of self-assembly

Proteins are multifunctional building blocks of life, that play various fundamental functions in all living organisms. Naturally existing proteins are involved in various biological functions such as regulatory and sensory functions, immune responses, mobility and structural stability [1, 2]. The ability of these biomolecules to assemble into multiple conformational variants and possess interactive interfaces, provides for their multifunctional nature [3]. In some higher order interactions, protein monomers self-assemble into large oligomeric complexes, giving rise to unique structural cage-like complexes. Comprehending the complexity of biological design and redesigning it into reliable, predictable and useful system is an underlying aim of modern-day synthetic biology. Naturally occurring proteins represent only a tiny fraction of the mathematically possible protein canvas and natural proteins are poorly spread in the total sequence space.

The introduction of *de novo* principles to protein design provides the opportunity to explore countless novel structures outside the realms of natural evolutionary constraints [4]. Unlike protein structure prediction, where computational methods predict the possible 3D structures corresponding to a sequence of amino acids, *de novo* protein design predicts a sequence that folds into a user defined 3D structure. With the rise in *de novo* designed proteins, self-assembly is gaining new interests and forms a powerful blueprint for bottom-up design of various synthetic proteins for a wide range of applications such as bio-therapeutics, protein-based diagnostics, biosensing and biomaterials [4, 5].

Intermolecular forces such as hydrophobic interactions, hydrogen bonding and ion pairing etc, are the key players for protein self-assembly. These forces when combined with structural complementarity at the interface of an interacting protein generate molecular complexes [6]. Most biological systems rely on these protein-protein interactions in order to carry out various cellular functions. Thus, understanding the fundamentals of protein-protein interactions (PPIs) is the first step for creating artificial protein complexes [7]. In recent years, significant scientific interest in protein engineering has been focusing on targeted therapeutics, symmetry guided PPIs and constructing multimeric assemblies for encapsulations [8].

5.2.2 Computational protein-protein interaction design

Achieving desired affinity at a precise location and specificity is still under research and no single method to date can design an interacting interface with total accuracy. Knowledge-based methods and protein docking assembly are two mainly used principles in designing protein-protein interactions. In most cases, a combination of both sequence design and protein docking algorithms are used to identify structures and sequences that promote binding with further optimisation [7].

5.2.2.1 Knowledge based methods vs protein docking/assembly

Knowledge-based methods rely on existing experimental data and statistical modes to extract and interpret the patterns in the given protein structures. These homology based, data driven methods map the interfacial residues and commonly found structures onto the structure and composition of the query protein to predict binding [9]. *Hot spot centric design* is the most commonly used approach for designing new interfaces from the homology driven statistical models. In this approach, the hot-spot residues occurring on the naturally existing protein complexes are grafted on the target interface. This is followed by computational optimisation to optimise for steric hindrance and side chain optimisation [7].

Docking based approaches use 3D models to search for conformations with high surface complementarity and low free energy [9]. This method is also referred to as “dock and optimisation” [10]. The predicted interface with lowest energy is first established by computationally comparing the free energy changes at all possible conformations. In most cases the structures and the interface composition are refined to achieve ideal binding efficiencies[7].

Although both these methods differ in their core principles and approaches, they can be used sequentially or in parallel to validate the predicted interface model and to refine the existing models.

5.2.3 Understanding protein cages

Protein cages can be defined as hollow, three-dimensional oligomeric protein structures built from self-assembly of constituent monomers. This hollow cage like structure gives them the potential to encapsulate and control the release of a molecule of interest by genetic or chemical conjugation [11]. Protein cages are highly organised structures and are abundantly found in nature. Viral capsids, bacterial microcompartments, HSPs and ferritins are a few examples. Ion storage, catalysis and packing nucleic acids are some common roles of protein cages [12]. Naturally existing protein cages acts as a structural and functional guide for the development of synthetic protein cages. Their ability to spontaneously self-assemble has driven the ideas tailoring novel synthetic cages, with novel functional architectures, for biosensing and targeted therapeutic release.

Table 5.1: Engineered Protein Cages with biomedical applications

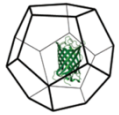
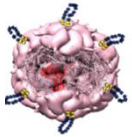
Table 5.1: Engineered Protein Cages for Biomedical Applications						
Classification of protein cage type	*Applications*	*Engineered protein interface/surface*	*Effects*	*Modifications*	*Protein cage structure*	*References*
Non-viral cages						
Aquifex aelicus lumazine synthase bacterial microcompartments (AaLS)	Diagnostic imaging and drug delivery system	Protein interior surface	Encapsulation of molecular cargo by charge complementarity	Chemical and genetic modification		(Seebeck *et al.*, 2006)[13]
Thermotoga maritima derived bacterial nanoparticle.	Delivery of drugs, anti-tumour therapy and imagining agents	Protein interior and exterior surface	Encapsulation of active molecular cargo by chemical conjugation and cell specific targeting	Genetic and chemical modification		(Moon *et al.*, 2014)[14]

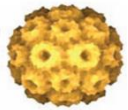
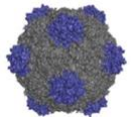
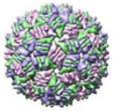
Table 5.1 (contd): Engineered Protein Cages for Biomedical Applications						
Classification of protein cage type	*Applications*	*Engineered protein interface/surface*	*Effects*	*Modifications*	*Protein cage structure*	*References*
Viral cages						
Cowpea Chlorotic Mottle Virus (CCMV)	Diagnostic imaging and vaccine development	Protein cage exterior and interior surface	MRI contrast agent , target ligand binding and inorganic nanoparticle synthesis	Genetic and chemical modification		(Flenniken *et al.*, 2009)[16]
Cowpea Mosaic Virus (CPMV)	Vaccine development, anti-tumour therapy, and imaging applications	Exterior surface of the protein cage	Site specific ligand interaction, selective cell-targeting and bio-imaging through conjugation with fluorescent probes	Chemical and genetic modification		(Lee and Wang, 2006)[17]
Bacteriophage MS2	Viral based delivery vector for specific cell targeting, imaging applications and microRNA delivery	Interior and exterior surface	Degradable interior linkage for therapeutic cargo and specific cell uptake.	Bioconjugation, chemical and genetic modification		(Fu and Li, 2016)[18] (Bhaskar and Lim, 2017)[19]

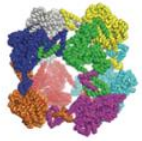
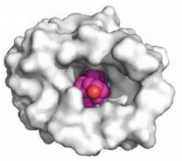
Table 5.1 (contd): Engineered Protein Cages for Biomedical Applications						
Classification of protein cage type	*Applications*	*Engineered protein interface/surface*	*Effects*	*Modifications*	*Protein cage structure*	*References*
Tobacco Mosaic Virus (TMV)	Vaccine development and injectable biologics	Exterior surface of protein cage	Induce activation of immune cells and biological ligand display	Chemical modification		(Smith *et al.*, 2006)[2
Synthetic proteins						
Symmetrical nanohedral protein cage	Proof of concept. Possible applications in cell targeting and imaging	Designed cage protein based on dimeric and trimeric protein conjugation with helical linker	Directed protein self - assembly	Recombinant conjugation		(Padilla, Colovos and Yeates, 2001)[21] (Sciore *et al.*, 2016)[22]
Porous cube protein	Proof of concept. Possible application in cargo encapsulation and drug delivery	Computationally designed cubic protein based on cage between natural dimeric and trimeric protein interfaces, conjugated to a helical linker	Directed protein self-assembly	Recombinant conjugation		(Lai *et al.*, 2014)[23]

Table 5.1 (contd): Engineered Protein Cages for Biomedical Applications						
Classification of protein cage type	*Applications*	*Engineered protein interface/surface*	*Effects*	*Modifications*	*Protein cage structure*	*References*
De novo Protein Cages						
Digoxigenin binding protein (DIG 10.3)	Drug overdose therapeutic	Computationally designed protein exterior and interior surface	Site-specific ligand binding and ligand encapsulation.	De novo designed		(Tinberg *et al.*, 2013)[15]

5.2.3.1 Assembly mechanics and encapsulation kinetics

In nature, assembly mechanisms can be broadly divided into two main types based on whether or not they need cargo intervention for self-assembly [24]. In those cases which assemble without any cargo, the self-assembly relies on subunit-subunit interactions. This type of assembly construction initiates with nucleation and growth [25]. Cargo independent assemblies could be found in cowpea chlorotic mottle virus (CCMV). While, the cages that need cargo for self-assembly rely on scaffolding proteins and packing machinery. In such cases, the assembly is achieved by electrostatic interactions between the monomers proteins and negatively charged DNA/RNA [25]. Post assembly encapsulation is commonly achieved either by post cage by loading the cargo through environmental changes (such as pH or temperature) or by electrostatic interactions. Cargo can also be encapsulated during the cage assembly by molecular recognition in which the cargo is attached to the proteins by structural/electrostatic linking.

5.2.4 Visualising self-assembly

While design and synthesis of self-assembling protein cages has been well documented, the visualisation of the cage architectures remains expensive. This often relies on physical methods such as Cryo-EM, NMR or X-ray crystallography. Although these methods help visualise the final 3D conformation of the assembly, the costs and efforts associated with them are high. In recent years, protein assemblies and protein interactions were monitored by labelling the proteins with appropriate optical tags. The optical readout (Fluorescence/luminescence/colorimetric readouts) from these proteins is analysed using an end-point assay. Such an end-point analysis assay would not effectively monitor the process of self-assembly and has too many experimental variables.

FlAsH-EDT2 is an organoarsenic compound used in bioanalytical research as a fluorescence reporter for tagging various proteins in cells. The structure of FlAsH-EDT2 contains 1,3,2-dithiarsolane substituents on a fluorescein core. FlAsH-EDT2 becomes fluorescent upon binding to proximal tetra-cysteines. FlAsH on its own is non-fluorescent when bound to EDT2. When FlAsH-EDT2 binds to a tetra-cysteine

motif, EDT2 is displaced and the compound becomes highly fluorescent. Due to its small size and membrane permeability, it has unique advantages over existing protein-based fluorescent tags.

In this work, bi-partite cysteine residues were computationally engineered into protein cage self-assembly interfaces and a FlaSH-EDT2 based fluorescence strategy was developed to visualise self-assembly and continuously monitor the protein-protein interactions. During self-assembly, the proximity of the bi-partite cysteines initiates the FlAsH-EDT2 labelling which produces a fluorescence readout. This fluorescence readout indirectly indicates the protein-protein interaction. This addition of the bi-partite cysteine residues and FlaSH-EDT2 mediated fluorescence strategy provided the scope to continuously monitor the self-assembly.

5.3 MATERIALS AND METHODS

5.3.1 *In silico* experimental methods

5.3.1.1 Design overview

Previous literature on self-assembling cages acted as a starting point for proof-of-concept and to validate the FlAsH-EDT2 based visualisation strategy. An existing cage strategy from Padilla *et al* [21] was adopted for the symmetry driven self-assembly (Figure 5.2). A dimeric subunit and a trimeric subunit were linked together with a rigid 9 amino acid helical linker to form a megamonomer. Oligomerisation is achieved when the dimeric subunit and the trimeric subunits find their identical copies to drive the cage self-assembly.

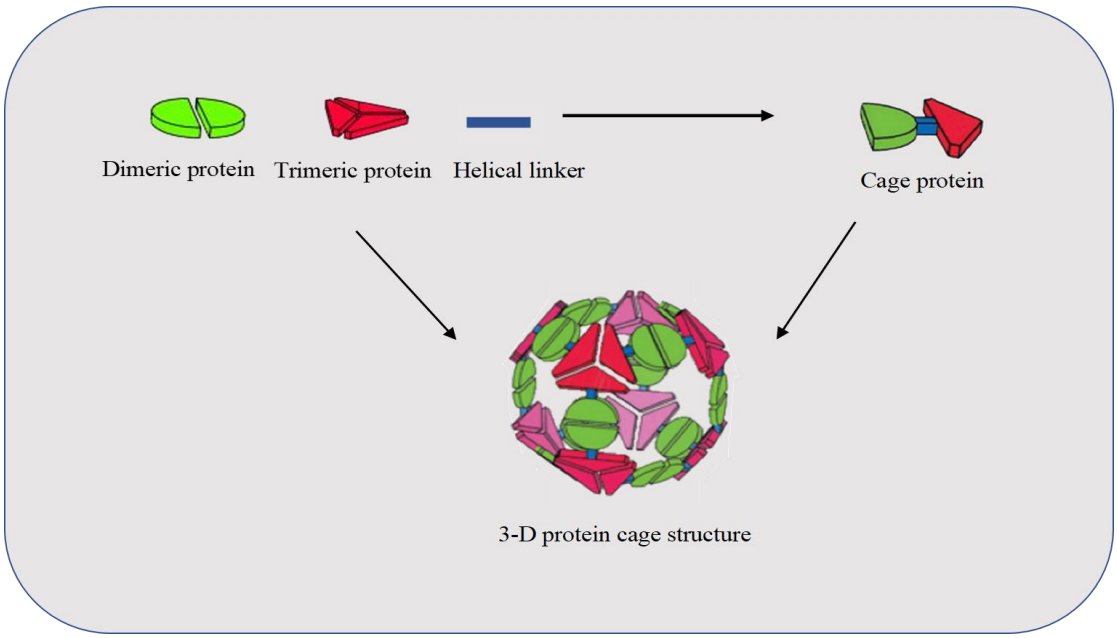


Figure 5.2: Cage self-assembly model. *Illustration adapted from Padilla et al (2001) [21]. The dimeric and trimeric subunits were linked together with a helical linker. The resultant Mega Monomer (cage protein) spontaneously assembles into a cage like structure through symmetry aided interactions.*

5.3.1.2 Construct design

For testing the proposed bi-partite cysteine insertion, the first step was to find monomers capable of self-assembly. In this context, protein subunits PDB ID 1AA7 (dimer subunit) and 1BRO (trimer subunit) were chosen based on previous literature on self-assembly and their stability. The 9 amino acid helical linker KALEAQKQK was chosen based on the literature [21].

5.3.1.3 *In silico* tools used for design, modeling and validation

Table 5.2 lists the *in silico* tools used for designing assemblies with cysteine inserts.

Tool	Purpose
Rosetta suite	Redesign and modeling
I-Tasser [26]	Protein modeling
R-package and in-house algorithms	Evaluating conformational integrity and measuring agreement between models using RMSD scoring.
COACH (web server)	Active site prediction
PROCHECK web server	Generating Ramachandran plot
Chimera [27]	Protein 3D visualisation

Table 5.2: *In silico tools used for computationally inserting cysteine residues and validating protein models.*

5.3.1.4 DNA design and synthesis

The final amino acid sequences, after a thorough *in silico* validation, were reverse translated into their corresponding DNA sequences. The DNA sequences were codon optimised for *E. coli* BL21 using the codon optimisation tool available on IDT website (https://eu.idtdna.com/codonopt). NEB and SnapGene's Gibson assembly simulators were used to design the homologous arms to facilitate Gibson assembly. Amplification and sequencing primers were designed using Benchling's primer design tool. The

primers were cross verified using Primer3Plus. All construct DNA sequences and primers were sourced from Integrated DNA Technologies.

Primer Name	Sequence
1AA7 FWD	actttaataaggagatatacATGCAGAAATTATTGACAGAGGTTG
1AA7 RVS	tgctcagcggtggcagcagcTTATTTATCGTCATCATCTTTGTAATCT
IBRO FWD	actttaataaggagatatacATGCCTTTTATCACAGTTGGG
IBRO RVS	tgctcagcggtggcagcagcTTATTTATCGTCATCGTCTTTGTAATC
Cage FWD	actttaataaggagatatacATGCCTTTTATAACAGTCGGGC
Cage RVS	tgctcagcggtggcagcagcTTATTTATCGTCGTCGTCTTTGTA
Gaussia FWD	actttaataaggagatataccATGATGGAAGCCAAACCC
Gaussia RVS	tgctcagcggtggcagcagcgcagcATCCTGGACAACATTTGGC
SEQ Nco1 FWD	GAAGAGGCATAAATTCCGTC
SEQ AVR2 RVS	CGACCATCAAGCATTTTATC

Table 5.3: Sequences of all amplification and sequencing primers used.

5.3.2 Wet-lab methods

5.3.2.1 Plasmid amplification and extraction

pRSFduet-1 (Novagen) dual expression vector with Kanamycin resistance was used for protein expression. *E. coli* BL21 was used for plasmid amplification. *E. coli* BL21 cells were made competent using Cohen et al. 1972 protocol. 100 ng plasmid DNA was mixed with 30 µl competent cells and placed on ice for 20 min. The suspension was subjected to heat shock at 42 ˚C for 20 min. The cells were placed in ice for 2 min and 1 ml LB broth added. 100 µl transformed cells were plated on LB agar containing 50 ug/ml Kanamycin. Colonies were then subcultured and stored in -80 ˚C for further use. For plasmid extraction, overnight subcultures of the transformed bacteria were subjected to the Monarch Plasmid miniprep kit (New England Biolabs) protocol. The extracted DNA was stored at -20 ˚C in situations where it wasn't used immediately.

5.3.2.2 Restriction digestion

pRSFduet-1 was digested by restriction enzymes Nco1 and AvrII with CutSmart reaction buffer (NEB) at 37 ˚C for 1 h. The manufacturer's protocol was followed to adjust the reaction volumes as per the need. Following the restriction digest, the DNA was purified using the PCR purification kit (Qiagen) protocol. In all cases, the restriction digest was verified by Agarose gel electrophoresis and the DNA concentration was determined using NanoDrop (Thermofisher).

5.3.2.3 Gibson Assembly

Gibson Assembly was carried out using the Gibson Assembly master mix described by DG Gibson *et al* (2009). The plasmid and DNA gene blocks were mixed in 1:3 ratio in a Gibson Assembly master mix and incubated at 50 ˚C. *E. coli BL21* cells were transformed with the assembled plasmids and plated on LB agar medium containing the appropriate antibiotic. Gibson Assembly was confirmed by colony PCR and sanger sequencing.

5.3.2.4 Validating cloning using colony PCR and Sanger sequencing

The selected colonies were added to NEB PCR master mix with 2.5 µl of corresponding primers. PCR was carried out as per NEB Q5 polymerase PCR protocol. Sanger sequencing (GATC light-run) was also performed by GATC (Eurofinsgenomics) on the selected colonies to validate assembly.

5.3.2.5 Inducing protein expression

Transformed cells were suspended in 20 mL LB broth containing 30 mg/mL Kanamycin. 1:500 sub cultures were made from the overnight stock and grown to 0.6 OD600. The subcultures were induced with 0.5 mM Isopropyl ß-D-thiogalactoside (IPTG) left overnight in a shaking incubator at 25 ˚C. Cells were harvested by centrifugation at 4000 rpm for 20 min and were subjected to lysis. BugBuster lysis buffer, supplemented with Lysonase reagent (Novagen) and a cocktail of protease inhibitors (cOmplete) was used for cell lysis. Lysis was carried out at room temperature for 20 min. The lysate was then clarified by centrifuging at 10,000 rpm for 20 min at 4 ˚C.

5.3.2.6 Analysing protein expression

Total protein concentration was determined by standard Bradford assay. Protein standards were prepared using 1 mg/ml Bovine Serum Albumin. 200 µl of Protein Assay Reagent (Bio-Rad) was used for the analysis. BSA standards were made from 0-60 ug/ul. 1 µl test protein sample was mixed with 799 µl MilliQ water and 200 µl Protein Assay Reagent (Bio-Rad). Samples were loaded onto a 96 well plate and absorbance was measured using a FLUOstar OMEGA (BMG) plate reader.

To validate protein expression, Coomassie blue staining was performed on all samples. 1 mg of lysate was diluted with LDS (NuPAGE, Invitrogen) to form a 1X solution. The lysate was boiled for 5 min at 95 ˚C. 10-15 µl of the sample is loaded into a 4-10 % Bis-Tris gel (1 mm, 15 well) along with precision plus protein ladder (Bio-Rad) and electrophoresis was carried out at 100 V for 90 min. After the run completion, the gel was washed and fixed in 50 % methanol + 10 % acetic acid solution for 15 min. After subsequent washing with water, Coomassie blue (Eazy Blue, Bio-Rad) was added to gel and stained for 1 h. The gel was washed 3 times with water and results analysed using ImageLab 5.2.1 (Bio-Rad).

For Western blotting, the electrophoresis was carried out as above and the proteins were transferred onto a nitrocellulose membrane (Bio-rad Midi) using rapid transfer for 7 min. The nitrocellulose membrane was blocked with Blocking Buffer PBS

(Odyssey) for 1 h and incubated overnight at 4 ˚C with anti-FLAG antibody (Sigma, F1804). After washing the membrane 3 times with TBST, the membrane was incubated with anti-mouse green secondary antibody for 1 h in the dark at room temperature. The membrane was scanned using LICOR Odyssey infrared imaging system and the image analysis was performed using Odyssey imaging software (LI-COR). FLAG-BAP protein was used for standards to normalise protein concentrations for downstream experiments. The integrated intensity of the Western blot bands were used for relative quantification on the Li-Cor Odyssey software.

5.3.2.7 Fluorescence assays and validating FlAsH binding.

Binding assays with cell lysates. 600 µl of 0.5 mg/ml of lysate was incubated with FlAsH-EDT2 (0.1 µg) in the dark (FlAsH is photosensitive). Lysates were washed with PBS and transferred in triplicates to a black 96 well plate (Titude Vision plate). Fluorescence was measured on an Infinite M200 Multimode plate reader (TECAN). Fluorescence was also measured using the IVIS Lumina II Imaging system (Perkin Elmer) with filters, Ex 485nm and Em 535nm. Living image (Perkin Elmer) software and Microsoft excel were used for analysing the data.

Binding assays using whole cells.

Subculture cells were washed and resuspended in PBS and normalised to OD 0.3. Post washing, cells were incubated in FlAsH Buffer (100 mM Tris HCl, 100 mM NaCl, 1 mM EDTA, 1 mM βME, pH 7.8) for 1 h at room temperature. 0.5 µg FlAsH-EDT2 was added and the samples were incubated in the dark for 2 h. The unbound FlAsH-EDT2 was removed by washing with FlAsH buffer for 3 times. The samples were loaded in triplicated into a black plate (Titude Vision plate) and read in triplicates using an Infinite M200 Multimode Plate reader. Microsoft excel was used for analysing data.

5.4 RESULTS

5.4.1 *In silico* design and modelling of self-assembling cage

The amino acid sequences for the dimeric protein M1 matrix protein of influenza virus (PDB ID 1AA7) and the trimeric protein Bromoperoxidase (PDB ID code 1BRO) were fetched from PDB and a 9-residue helical linker was inserted to connect them together. The total protein was termed as MegaMonomer. The 3D structure was modelled using I-Tasser. This MegaMonomer would self-assemble into 12-mer, 550 kDa protein cage. Insertion of sequences for cysteines was carried out using Rosetta design. The protein-protein interaction interface was identified for all necessary combinations using computational bio-assemblies.

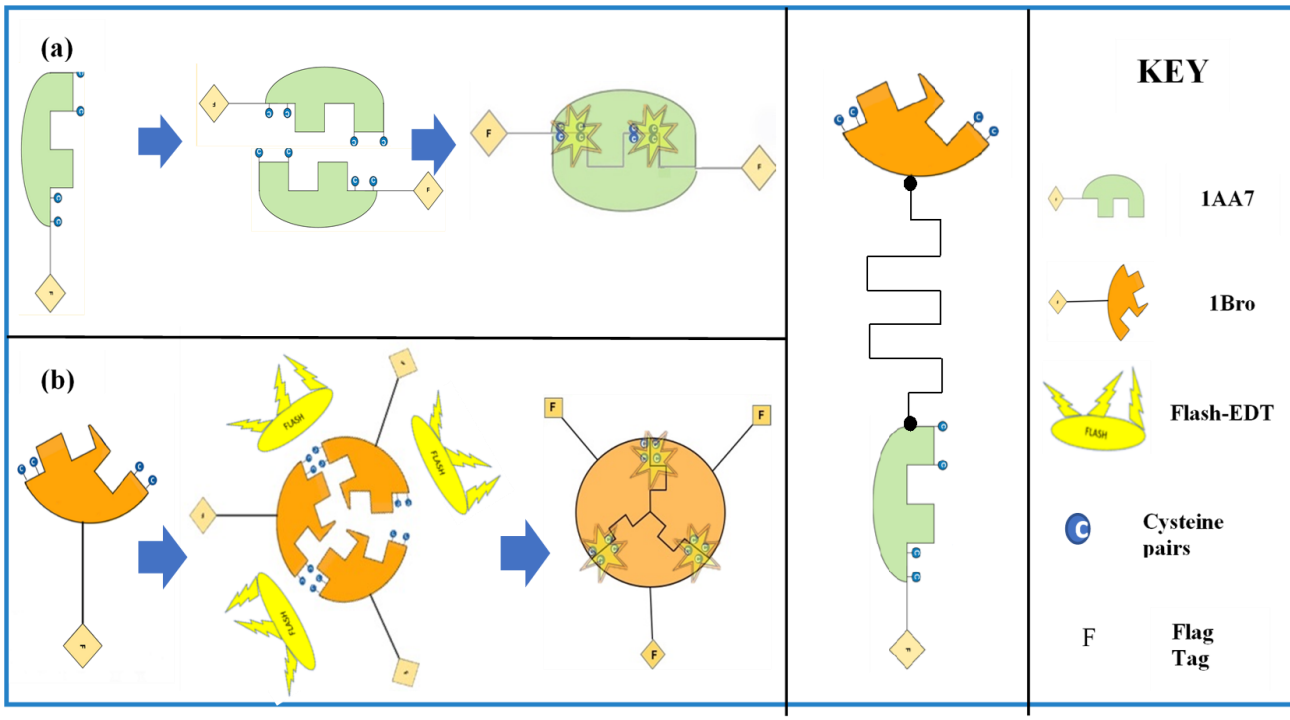


Figure 5.3: Protein construct map for the 3 oligomerization subunits. *(a) Dimeric subunit with engineered cysteine residues. (b) Trimeric subunit with cysteine insertions. (c) MegaMonomer, showing the dimeric and trimeric subunits linked via a 9 amino acid helical linker.*

5.4.1.1 *In silico* validation of engineered cysteines

For cysteine insertions, on all three proteins (Dimer, Trimer and the MegaMonomer) amino acids at the interface which are most energetically favourable and minimal structural importance were identified to ensure structural integrity. The dimeric protein assembles in a head-tail fashion and thus required addition of 4 cysteines. Residues 372 (Gly 372) and 373 (Asn 373) were identified as the head side spots for cysteine insertion. Residues 418 (Tyr 418) and 419 (Asn 419) were identified as the optimal tail side spots. This insertion of 4 cysteine residues on head-tail configuration ensures 2 FlAsH interactions as shown in Figure 5.4. On the trimeric protein subunit, residues 156 (Asp 156) and 157 (Asp 157) as well as 179 (Ile 179) and 180 (Ser 180) were identified as optimal for cysteine insertion. This ensures 2 FlAsH binding sites on the dimer and 3 binding sites on the trimer and 30 binding sites on fully assembled cage.

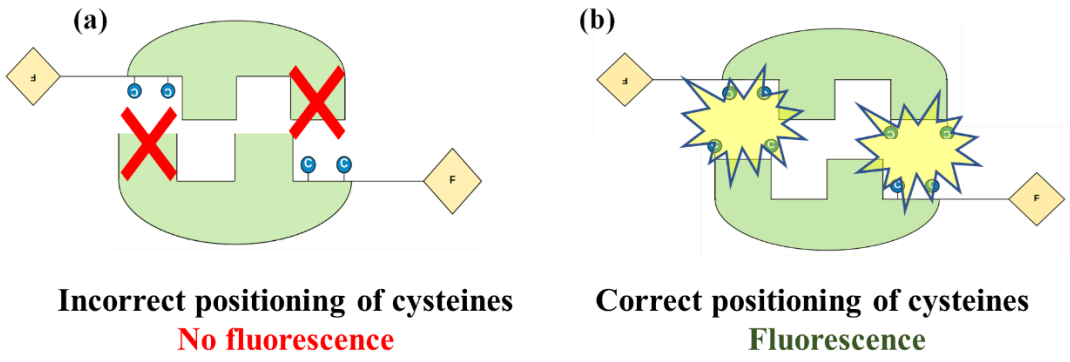


Figure 5.4: Illustration showing head-tail assembly of the dimer and cysteine insertions *(a)* No fluorescence*: The absence of bipartite cysteine pairs during head-tail dimerization would have no FlAsH binding. (b) 4 cysteines were engineered on each subunit to ensure bipartite pairs when assembled in head-tail conformation.*

The distance between the cysteine residues is crucial in terms of avoiding unintended binding false positives. FlAsH-EDT 2 recognizes proximally located bipartite cysteine pairs and emits a fluorescent signal. If the cysteines on one subunit were placed in close proximity, this would result in unintended flash binding (false positives). Initially, the two cysteine pairs on the dimer were separated by 10.6 Angstroms (Å). This was feared to result in a false validation. The distance between the two pairs was increased to 16 Å by inserting amino acids with minimal influence on the total structure (Figure 5.5). Similarly, it was ensured that the cysteine pairs on the trimeric protein were separated by 21 amino acid residues.

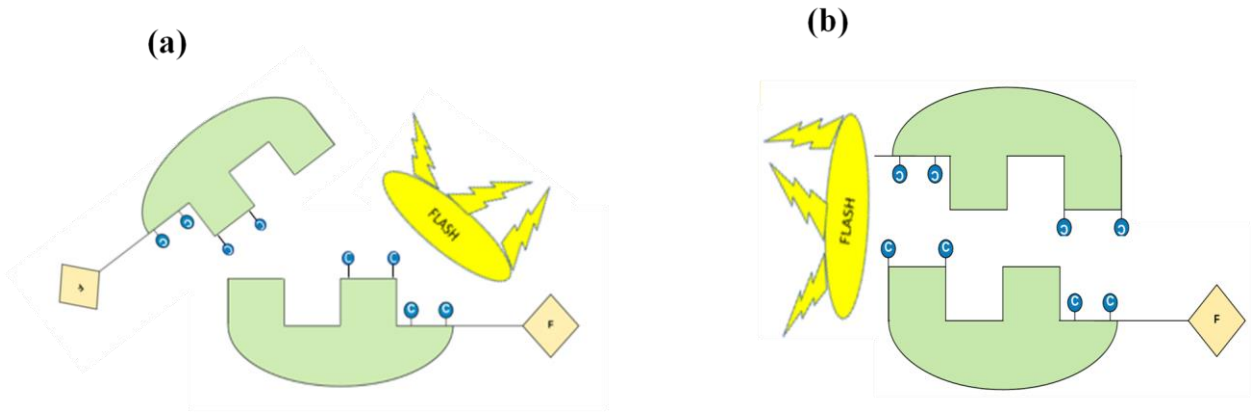


Figure 5.5: Illustration explaining the importance of distance between the bipartite cysteine pairs. *(a) FlAsH-EDT recognising cysteine residues on the same subunit due to close proximity. (b) FlAsH-EDT recognising the bipartite cysteine pairs on the interface, after increasing the distance between the cysteine residues on the same subunit.*

5.4.1.2 *In silico* validation of cage bio-assembly and functioning of inserted cysteines

However, after intensive assembly modelling, it was discovered that incorporated cysteines were not presented on the external surface and were folding into the concave curvature of the protein cage. This might hinder with FlAsH-EDT binding. See Figure 5.6.

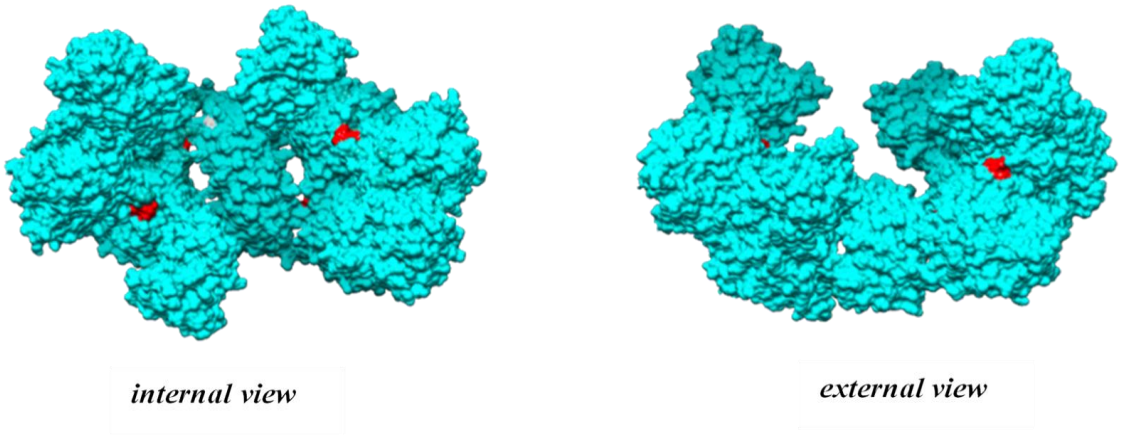


Figure 5.6: *Protein cage assembly has been divided in half to show the internal (top view) and external (side view) angles. The inserted cysteines are depicted in red.*

The cysteine residues were redesigned, and multiple 3D models were generated to ensure the improved cysteine placement on the external curvature. On the dimeric subunit, residues 358 (Leu) and 359 (Gln) as well as 364 (Arg) and 365 (Phe) were replaced with cysteines. On the trimeric subunit residues 158 (Tyr) and 159 (Ala) as well as 181 (Glu) and (Glu) were replaced with cysteines. Following cysteine insertion, all proteins were remodelled using both I-Tasser and Rosetta. The 3D structures were visualised using Chimera to ensure proper bio-assembly. RMSD scores were calculated to inform on deviations from the original unmodified protein structures (Figure 5.7).

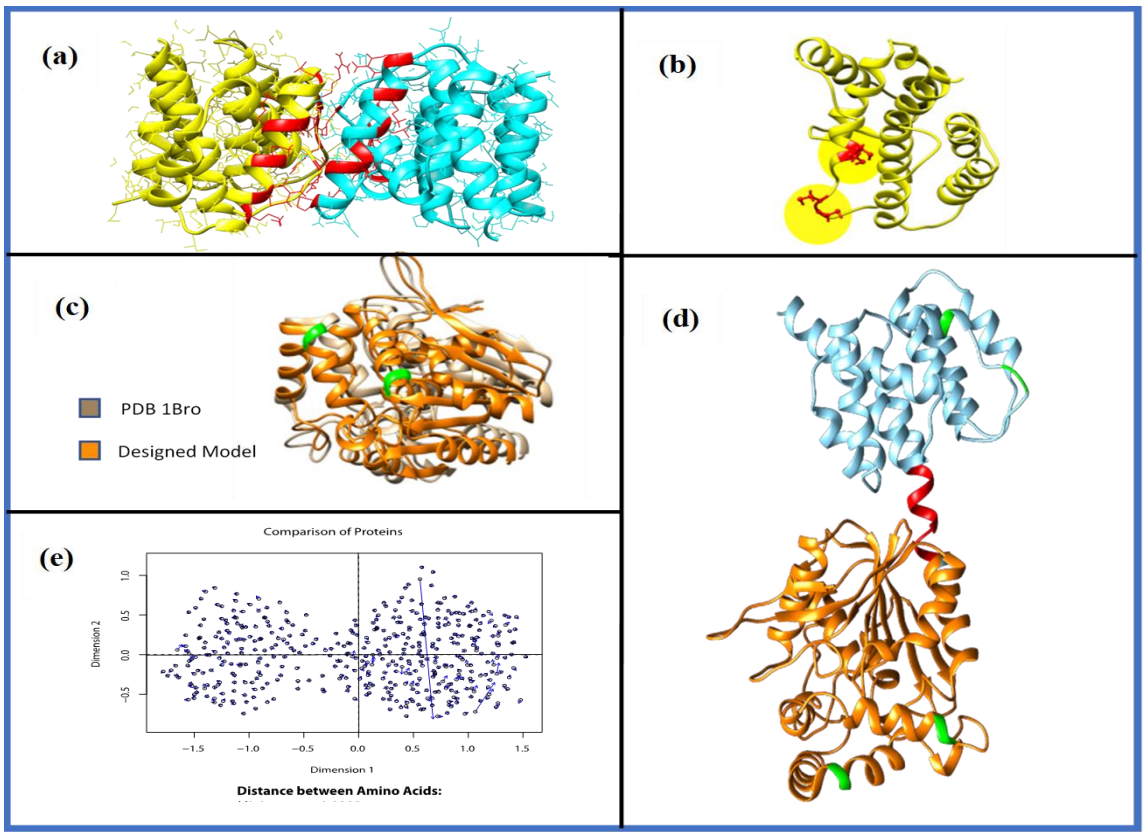


Figure 5.7: 3D models of dimer, trimeric and MegaMonomer subunits. *(a) The dimer subunits in assembly, showing the inserted cysteines (in red) at their binding interface. (b) Showing cysteine insertion (red) on a single dimeric protein subunit. (c) Trimeric protein with cysteine insertions superimposed on the unmodified protein. (d) MegaMonomer with cysteine inserted dimeric (blue) and trimeric (orange) subunits connected with a rigid helical linker (red). (e) RMSD of atomic distances of the MegaMonomer with cysteine insertion superimposed on the MegaMonomer without cysteine insertions.*

5.4.1.3 *In silico* validation of integrity of the modelled protein structures

Ramachandran plots were generated for all protein constructs to theoretically validate the folding integrity. It was observed that 90 % of all amino acid residues in all three proteins fall in the most energetically favourable region and a maximum of 0.5 % in

the disallowed region. Table 5.4 summarises the Ramachandran plot scores for all three protein constructs.

The Ramachandran plot for the MegaMonomer is shown in Figure 5.8. The computational models and evaluation provided sufficient confidence in the final structures to proceed to the wetlab experimentation.

Modelled Protein	Residues in favored regions	Residues in allowed regions	Residues in disallowed region
Dimer (1AA7)	90.4 %	9.6 %	0.0 %
Trimer (1BRO)	89.6 %	9.6 %	0.0 %
Protein cage	90.3 %	9.2 %	0.5 %

Table 5.4: Summary of Ramachandran plots of all three proteins.

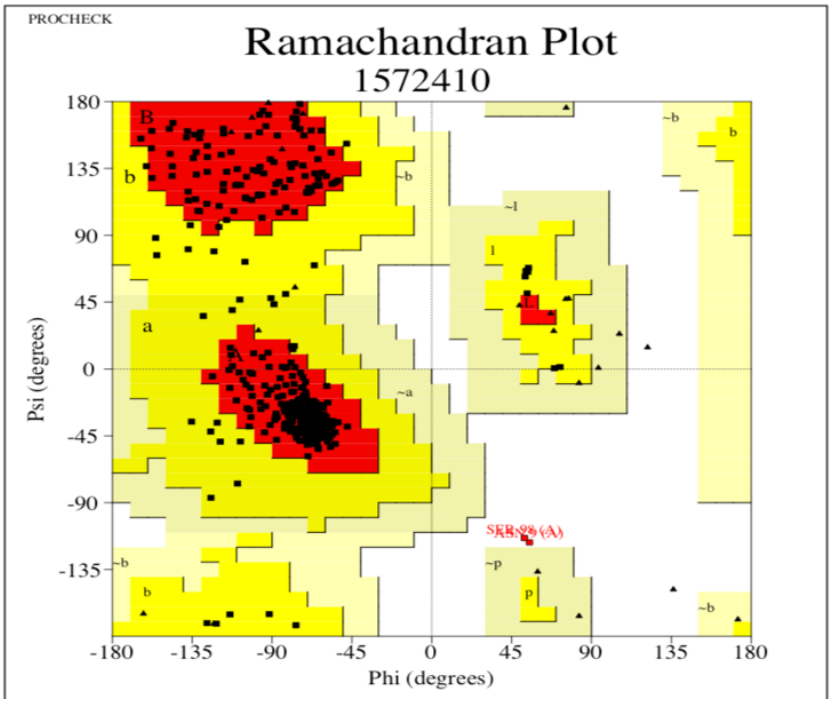


Figure 5.8: Ramachandran plot generated for MegaMonomer. *The red, yellow and light yellow regions represent the favoured, allowed and disallowed regions respectively. Black dots represent the amino acid residues. The ψ (Psi) and φ (Phi) torsional angles are represented on the X and Y axis respectively.*

5.4.1.4 F2F plots to visualise overall performance

F2F plots, as discussed in Chapter 3, were generated to have a holistic look on the overall performance of each designed construct. Solvent accessibility of the active sites (cysteine residues), PI, Grand hydrophobicity average, Instability index and Ramachandran Score (RC score) were used as plot axes. OP score was calculated as shown in Chapter 3.

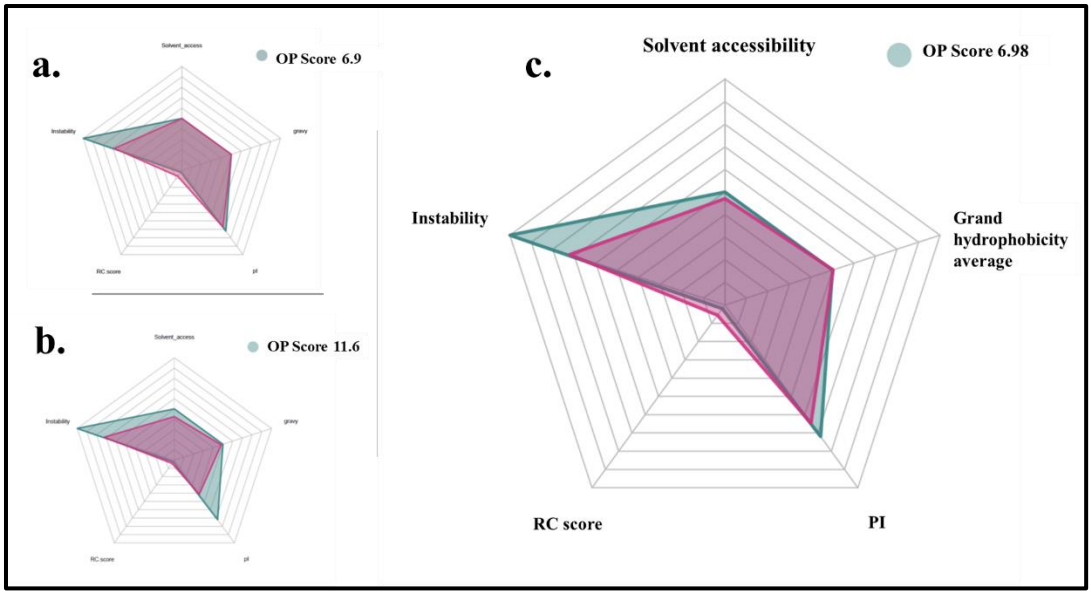


Figure 5.9: F2F plots for the dimer, trimer, and cage monomers. *(a) F2F plot for trimer monomer, (b) F2F plot for dimer monomer (c) Plot for cage monomer (MegaMonomer). Lower OP score indicates high agreement between the designed construct parameters and desired in silico parameters. Trimer monomer showed the highest OP score and was predicted to perform the best. The dimer scored the highest and hence was predicted to have a relatively low overall performance.*

5.4.2 Wet-lab experimentation

5.4.2.1 Validating protein production

Post-transformation, bacterial subcultures were induced for protein expression, using two different concentrations of IPTG (1 mM and 0.5 mM). Cells were harvested and the total protein concentration was determined using Bradford Assay. The concentrations of all protein samples were normalised to 5 μg/μl. A Western blot was carried out to validate and quantify the proteins of interest. FLAG-BAP was used as a positive control and as a standard to quantify the proteins of interest (Figure 5.10).

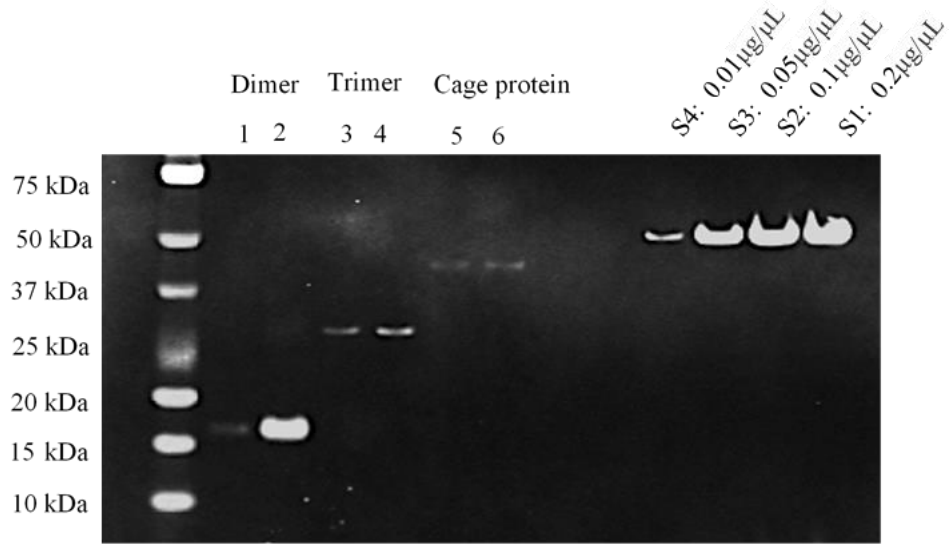
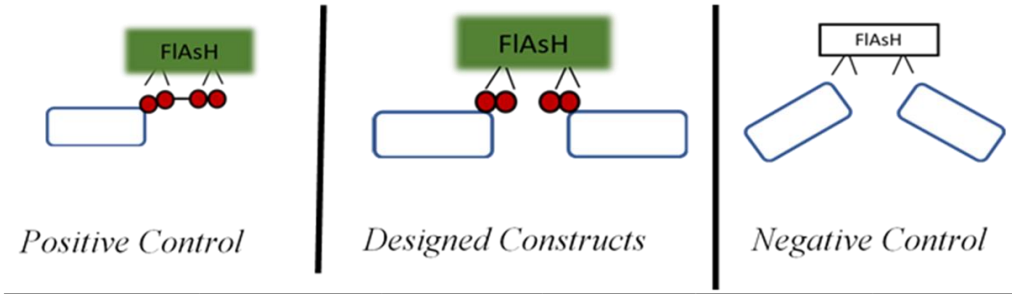


Figure 5.10: Western blot validating protein production. *Lanes 1,3 and 5 contain the proteins induced with 1mM IPTG and lanes 2, 4 and 6 contain the proteins induced with 0.5 mM IPTG. The 4 lanes on the extreme right show FLAG-BAP standards in increasing concentrations.*

5.4.2.2 Self-assembly verification strategy

To validate the FlAsH-EDT2 binding mediated oligomerisation, a series of fluorescence assays were carried out in controlled *in vitro* conditions. In all the assays, a small stable protein (GLuc) with an engineered FlAsH-EDT2 binding tetra-cysteine motif (CC-PG-CC) was designed and used as a positive control for FlAsH-EDT2 binding upon producing in *E. coli* in parallel with the test proteins. The rationale of the wetlab experimentation is illustrated in Figure 5.11.



Assay	Dimer	Trimer	Cage	Positive Control	Negative Control
Before FlAsH	✕	✕	✕	✕	✕
After FlAsH					✕
Denaturation	✕	✕	✕		✕

Figure 5.11: Wet-lab strategy for confirming self-assembly. *The positive control protein has a tetra-cysteine tag that binds to FlAsH-EDT and emits fluorescence independent of binding. The test proteins interface engineered bipartite cysteines and their fluorescence is self-assembly dependent. WT cells with no modification were used as negative control and there would be no fluorescence observed.*

5.4.2.3 Whole-cell FlAsH-EDT2 based fluorescence assays confirming oligomerisation

Whole-cell fluorescence assays were carried out to validate self-assembly. All the bacterial cultures were maintained at 0.3 OD before harvesting for the assays. Intense care was taken to remove unintended autofluorescence. Figure 5.12 shows the outcome of the fluorescence-based FlAsH-EDT2 binding assays. As expected, the control protein generated the maximum fluorescence and the negative control had a minimum fluorescent signal. When compared with negative control, a two-fold increase in fluorescence was observed in the dimeric protein assembly, and a four-fold increase was observed in the trimeric assembly and the cage assembly. This provided a clear indication of protein-protein interactions and the response of the FlAsH-EDT2 towards proximally placed bi-partite cysteines. The graphs representing the data in both signal to noise ratios and as percentage increase in fluorescence are shown in Figure 5.12.

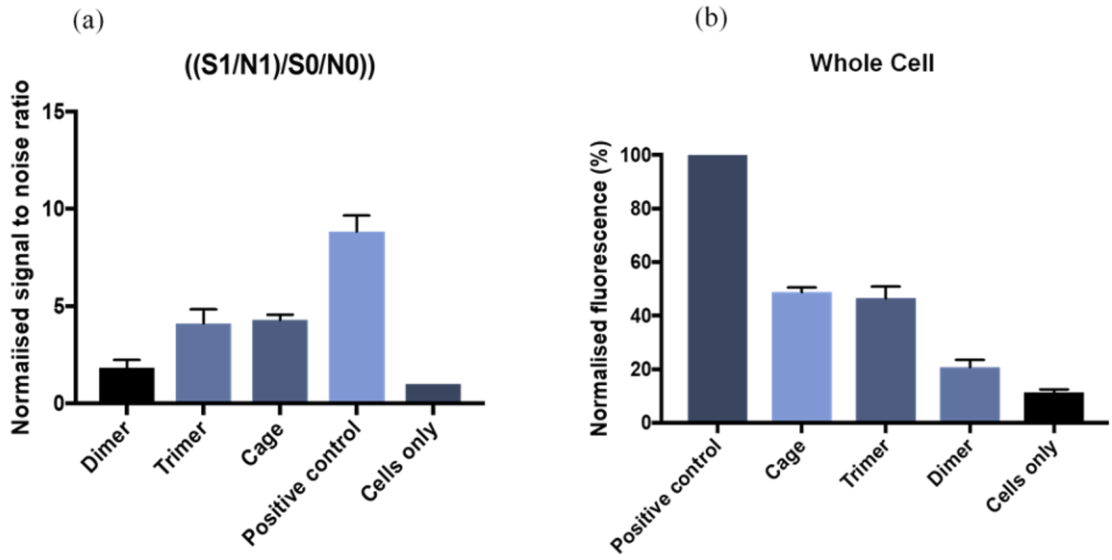


Figure 5.12: Confirmation of oligomerisation using FLaSH-EDT binding. *Equal quantities of bacterial cells were harvested simultaneously and subjected to FLaSH-EDT assay. (a) Fluorescence in signal to noise ratios. Cells without any modification were used as background noise. All samples were prepared in triplicate. All proteins with engineered bi-partite cysteines produced higher fluorescence than the negative control. (b) Percent fluorescence output was normalized to the positive control (GLuc with tetracysteine tag). The cage assembly produced a signal close to 48 % of the positive control, while the trimer and dimer assemblies produced 46 % and 21 % of the signal produced by the positive control respectively.*

Initially, the cage assembly was expected to produce higher fluorescence than the dimer and trimer assemblies. However, observing the data in Figure 5.12, this was not the case. In order to account for fluorescence per molecule, the data from Figure 5.12 was used to calculate the relative fluorescence units per FlAsH binding site on each protein (Figure 5.13).

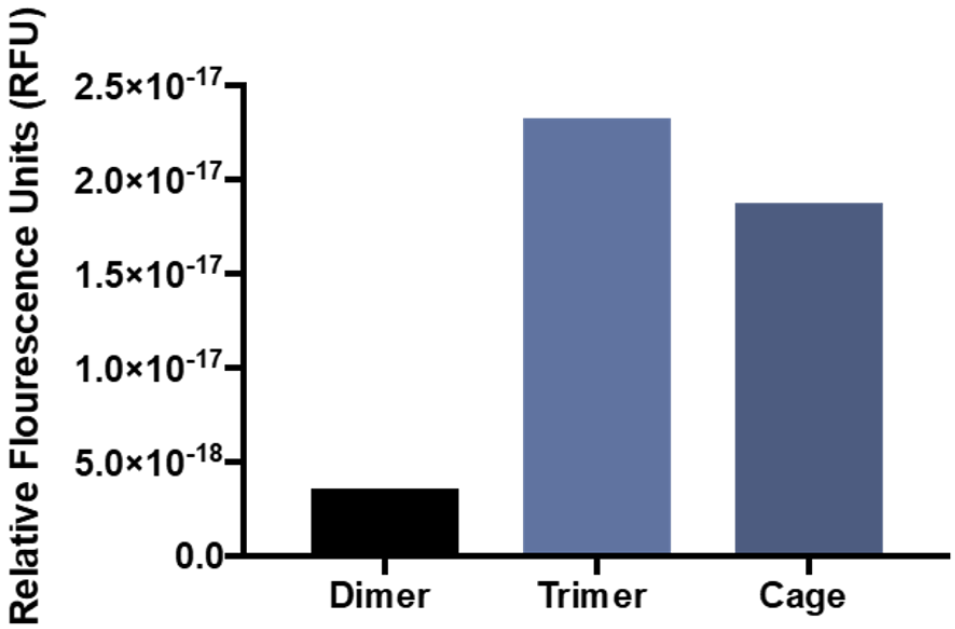


Figure 5.13: Relative fluorescence units per FlAsH binding site. *All engineered proteins were expected to mediate similar fluorescence per FlAsH binding site. However, this was not observed in experimental validation. The Trimer and the cage assemblies produced similar (p=0.02) magnitude of fluorescence signal, but the dimer assembly produced a significantly (p = 0.0008) lower signal.*

5.5 DISCUSSION

The current work utilised computational modelling and design approaches to engineer bipartite cysteines into the interface of the self-assembling proteins. Although the concept of using FlAsH-EDT to confirm protein-protein interaction or self-assembly has been done before, all such methods used bi-partite cysteines as a peptide tag at the N or C terminus of the protein. Although peptide tags are a validated as good reporters, complicated assemblies such as protein cages require proper exposure of the cysteines on the surface, at all stages of the self-assembly. The strategy used in this work, provides the design advantage of choosing the positioning of the cysteines. This is highly useful to eliminate false positive reporting due to proximally close non-assembly conformations.

Multiple iterations of *In silic*o redesigning was performed using Rosetta to ensure structural stability and right conformational positioning of the cysteines. The data provided a clear indication that FlAsH-mediated fluorescence could be successfully deployed to monitor protein-protein interactions. As designed, the FlAsH biarsenical complex was able to validate the self-assembly in all three test proteins.

Although the protein-protein interactions were confirmed, full validation of self-assembly of complex structures poses further challenges. It was observed that the dimeric protein (M1 matrix protein of influenza virus) was produced in significantly lower quantities when compared with the other proteins in the wet-lab studies which also affected the cage MegaMonomers production. This could be due to M1 viral capsid protein resulting in protein aggregation as mentioned in previous studies [28]. The F2F plot for the dimeric protein also predicted a low performance. The ability of the cages to assemble into intermediate oligomers is one of such complications. It is also important to outline that the sample sources (cell lysates vs whole cells), cell concentrations (OD before harvesting) and buffer conditions would be expected to have significant influences on the study outcomes. The fluorescence assays were initially conducted on clarified cell lysates. However, this type of assay resulted in high background noise. Using whole-cells provided the scope to wash the unbound FlAsH-EDT by light centrifugation. FlAsH-EDT complex also has a tendency to recognise natural cysteine motifs presented on natural cellular proteins and resulting in non-specific binding. To avoid the non-specific FlAsH-EDT binding, the samples

were incubated with FlAsH buffer that contains 2ME and EDT which improved the selectivity of the FlAsH-EDT complex [29]. With stringent experimental conditions (such as with the addition of competing thiols) it was clearly shown that FlAsH binding could be improved with enhanced signal intensities. The mathematical approach to represent data in terms of fluorescence per FlAsH binding site did not reflect the expected results. The variations in the wet lab assays demand further experimental validation to confirm the oligomerisation states.

However, further protein denaturing experiments are required to confirm the oligomerisation dependence of FlAsH-EDT2 based fluorescence. Techniques such as size exclusion chromatography (SEC) can be deployed to confirm the oligomerisation states. Additionally, a Cryo-EM image would solidify the evidence of cage assembly. Also, further experiments are needed to test the release kinetics of the cage. The results from this work are proof-of-concept to validate the functioning of the engineered bi-partite cysteines and the imaging concept using FlAsH-EDT2. Without such further validation, the data on the current synthetic cage are insufficient for translation into a full application.

With the increasing market for protein-based products, most academic and industrial settings would rely on monitoring proteins for their quality, interactions and stability. The concept of engineering bipartite cysteine residues, introduced in this work, is a proof-of-concept and building platform with an immense potential in biomedical technology. In the future, with the expanding interest in *de novo* protein design, this approach would benefit the visualisation of combinatorial assembly of various protein structures. This could be used to monitor various enzymatic and biochemical behaviours involving proteins.

5.6 REFERENCES

1. Tiwari, M.K., et al., *Computational approaches for rational design of proteins with novel functionalities.* Comput Struct Biotechnol J, 2012. **2**: p. e201209002.
2. Setiawan, D., J. Brender, and Y. Zhang, *Recent advances in automated protein design and its future challenges.* Expert Opin Drug Discov, 2018. **13**(7): p. 587-604.
3. Lai, Y.T., et al., *Structure and flexibility of nanoscale protein cages designed by symmetric self-assembly.* J Am Chem Soc, 2013. **135**(20): p. 7738-43.
4. Huang, P.S., S.E. Boyken, and D. Baker, *The coming of age of de novo protein design.* Nature, 2016. **537**(7620): p. 320-7.
5. Lapenta, F., et al., *Coiled coil protein origami: from modular design principles towards biotechnological applications.* Chem Soc Rev, 2018. **47**(10): p. 3530-3542.
6. Yagi, S., et al., *De novo design of protein-protein interactions through modification of inter-molecular helix-helix interface residues.* Biochim Biophys Acta, 2016. **1864**(5): p. 479-87.
7. Yagi, S., S. Akanuma, and A. Yamagishi, *Creation of artificial protein-protein interactions using alpha-helices as interfaces.* Biophys Rev, 2018. **10**(2): p. 411-420.
8. Samish, I., et al., *Theoretical and Computational Protein Design.* Annual Review of Physical Chemistry, 2011. **62**(1): p. 129-149.
9. Xue, L.C., et al., *Computational prediction of protein interfaces: A review of data driven methods.* FEBS Lett, 2015. **589**(23): p. 3516-26.
10. Smith, G.R. and M.J. Sternberg, *Prediction of protein-protein interactions by docking methods.* Curr Opin Struct Biol, 2002. **12**(1): p. 28-35.
11. Sasaki, E., et al., *Structure and assembly of scalable porous protein cages.* Nature Communications, 2017. **8**(1): p. 14663.
12. Putri, R.M., J.J. Cornelissen, and M.S. Koay, *Self-assembled cage-like protein structures.* Chemphyschem, 2015. **16**(5): p. 911-8.
13. Seebeck, F.P., et al., *A simple tagging system for protein encapsulation.* J Am Chem Soc, 2006. **128**(14): p. 4516-7.

14. Moon, H., et al., *Developing genetically engineered encapsulin protein cage nanoparticles as a targeted delivery nanoplatform.* Biomacromolecules, 2014. **15**(10): p. 3794-801.
15. Tinberg, C.E., et al., *Computational design of ligand-binding proteins with high affinity and selectivity.* Nature, 2013. **501**(7466): p. 212-216.
16. Flenniken, M.L., et al., *A library of protein cage architectures as nanomaterials.* Curr Top Microbiol Immunol, 2009. **327**: p. 71-93.
17. Lee, L.A. and Q. Wang, *Adaptations of nanoscale viruses and other protein cages for medical applications.* Nanomedicine : nanotechnology, biology, and medicine, 2006. **2**(3): p. 137-149.
18. Fu, Y. and J. Li, *A novel delivery platform based on Bacteriophage MS2 virus-like particles.* Virus research, 2016. **211**: p. 9-16.
19. Bhaskar, S. and S. Lim, *Engineering protein nanocages as carriers for biomedical applications.* NPG Asia Materials, 2017. **9**(4): p. e371-e371.
20. Smith, M.L., et al., *Modified Tobacco mosaic virus particles as scaffolds for display of protein antigens for vaccine applications.* Virology, 2006. **348**(2): p. 475-488.
21. Padilla, J.E., C. Colovos, and T.O. Yeates, *Nanohedra: using symmetry to design self assembling protein cages, layers, crystals, and filaments.* Proc Natl Acad Sci U S A, 2001. **98**(5): p. 2217-21.
22. Sciore, A., et al., *Flexible, symmetry-directed approach to assembling protein cages.* Proceedings of the National Academy of Sciences, 2016. **113**(31): p. 8681-8686.
23. Lai, Y.T., et al., *Structure of a designed protein cage that self-assembles into a highly porous cube.* Nat Chem, 2014. **6**(12): p. 1065-71.
24. Aumiller, W.M., M. Uchida, and T. Douglas, *Protein cage assembly across multiple length scales.* Chem Soc Rev, 2018. **47**(10): p. 3433-3469.
25. McManus, J.J., et al., *The physics of protein self-assembly.* Current Opinion in Colloid & Interface Science, 2016. **22**: p. 73-79.
26. Yang, J. and Y. Zhang, *I-TASSER server: new development for protein structure and function predictions.* Nucleic Acids Res, 2015. **43**(W1): p. W174-81.
27. Pettersen, E.F., et al., *UCSF Chimera--a visualization system for exploratory research and analysis.* J Comput Chem, 2004. **25**(13): p. 1605-12.

28. Noton, S.L., et al., *Identification of the domains of the influenza A virus M1 matrix protein required for NP binding, oligomerization and incorporation into virions.* J Gen Virol, 2007. **88**(Pt 8): p. 2280-90.
29. Krishnan, B. and L.M. Gierasch, *Cross-strand split tetra-Cys motifs as structure sensors in a beta-sheet protein.* Chem Biol, 2008. **15**(10): p. 1104-15.

Chapter 6

Development and validation of a novel miniaturised optical imaging device

A version of this chapter has been published as

Venkata V. B. Yallapragada, Uday Gowda, David Wong, Liam O'Faolain, Mark Tangney, and Ganga C. R. Devarapu. ODX: A Fitness Tracker-Based Device for Continuous Bacterial Growth Monitoring. Analytical Chemistry 2019 91 (19), 12329-12335. DOI: 10.1021/acs.analchem.9b02628

Impact Factor 6.35

Table of Contents

6.1 ABSTRACT 250
6.2 Introduction 251
6.2.1 Designing and deployability in synthetic biology 251
6.2.2 Fitness tracker-based handheld devices for bacterial growth monitoring. .. 251
6.3 MATERIALS AND METHODS 254
Hardware modifications were assisted by colleagues; Dr. Ganga Chinna Rao Devarapu and Uday Gowda in the Cork Institute of Technology. 254
6.3.1 Optical and mechanical design of ODX device and choice of materials 254
6.3.2 Fitness tracker 254
6.3.3 LED 256
6.3.4 3D printed enclosure 258
6.3.5 Smartphone app development 258
6.3.6 Bacterial culture 260
6.4 RESULTS 261
6.4.1 Deploying ODX hardware 261
6.4.1.1 Calibration of ODX device 261
6.4.2 Continuous bacterial growth monitoring 264
6.4.3 Adapting ODX for fluorescence 266
6.5 DISCUSSION 269
6.6 CONCLUSION 271
6.7 REFERENCES 272

6.1 ABSTRACT

Background Current optical devices to monitor biological phenomena such as bacterial growth and biofluorescence are bulky, expensive and remain benchtop-based. These factors can prevent real-time measurements in several cases, such as where biological processes are occurring in shaking incubators. Technologies such as miniaturised electronics, smartphone apps and 3D printing provide a rapid and low-cost platform to design, build and test hardware and software for such biological processes.

Aims The aim of this work was to utilise enabling technology such as 3D printing, miniaturisation and smartphone-assisted electronics to provide a deployable solution towards continuous monitoring of bacterial growth.

Methods A handheld fitness tracker-based device (ODX) was developed for this purpose. Multiple protypes of designed hardware were used to optimise the form and functioning of the device. The final device was calibrated and tested using 3 different bacterial genera. The accuracy and reliability were compared with a benchtop spectrophotometer. A basic web-based app builder tool was used to build a smart-phone app to record the data from the device.

Results The data from the growth curves validated the functioning of ODX. The accuracy tests carried out while calibrating the device proved ODX as accurate as a standard benchtop spectrophotometer. With smartphone-aided wireless data logging and other advantages such as portability and the ability to be taken into shaking incubators, ODX has various advantages over traditional instruments.

6.2 Introduction

6.2.1 Designing and deployability in synthetic biology

With the advent of synthetic biology and *de novo* design, protein-based systems stand to dominate the commercial landscape of life sciences. Proteins find their applications in food-based industries, materials technology, and medicine and healthcare. This versatility of protein-based applications promises an increasing potential for commercialisation. This potential for commercialisation is complemented by recent advances in synthetic biology. Cheap DNA synthesis and sequencing, recombinant technology, computational tools, miniaturised electronics, automation and robotics have driven synthetic biology into a golden era. Modern synthetic biology highlights the importance of multidisciplinary approaches in designing biological solutions and aims to accelerate the pace of biological research by integrating various scientific disciplines [1]. Commercial viability requires transforming a scientific outcome into a deployable product. For example, a protein based diagnostic tool would additionally require an appropriate hardware for testing and a software to analyse and report the results. Technologies such as miniaturised electronics, smartphone apps and 3D printing provide a platform bed to design and test multiple variations of the product until its final form.

6.2.2 Fitness tracker-based handheld devices for bacterial growth monitoring.

Bacteria are present a commercially viable, cheap and easily scalable platform for protein expression [2]. For centuries, food processing and fermentation industries have driven the commercial markets of bacterial based products. The advent of recombinant technology paved the way for various engineered enzymes and novel protein-based therapeutics [3-5]. In all those mentioned above, clinical, scientific, and commercial settings, monitoring of the population, and growth kinetics of bacteria plays a crucial role [6]. Each species and strain of bacterium has unique growth kinetics [7]. These growth kinetics depend on various parameters such as oxygen availability, temperature, medium in which the bacteria are grown, pH, culture vessel, the volume of the culture etc. Working with these microorganisms typically requires continuous monitoring of growth patterns. In many cases, microbiologists monitor the microbial

growth regularly to ensure that the population does not exceed pre-set thresholds or to maintain the population at a particular level.

Traditionally, several methods such as plate counting [8], direct counting [9], biomass measurement [9, 10], and light scattering have been used to measure bacterial growth. At present, optical density measurement based on the scattering of light from individual bacterial cells remains the gold standard. While the last few years have seen tremendous evolution of spectrophotometers, most of these devices are costly, bulky and remain benchtop based and so cannot be taken inside the incubator for real-time monitoring of OD measurements [11]. Furthermore, the usage of these instruments requires significant user interaction with the analyte and lack both versatility and flexibility due to their large form factor. These spectrophotometers also come with penalty of high labour costs and introduce contamination risks.

Recent advances in electronic miniaturization have paved the way for various types of wrist-worn low-cost fitness trackers [12]. These commercially available fitness trackers typically track, monitor and analyse various activities such as physical movement, sleep, and heartbeat rate, facilitated by various sensors such as an accelerometer, heart rate monitor, ECG, GPS etc. The output of these sensors is processed and stored by a small but powerful microprocessor. Fitness trackers transmit the result of the activities directly to the built-in OLED screen as well as to smartphones via Bluetooth. Despite having many sophisticated sensors, a powerful microprocessor, and highly miniaturized design, these fitness trackers are priced as low as $10 on the consumer market. Of all the sensors in fitness trackers, the optical heart rate sensor is particularly interesting [13]. The heart rate sensor consists of LEDs, one or more photodiodes, the essential components required of an OD/colorimetric/fluorescence monitor. Modifying these cheap fitness trackers using miniature electronics and 3D printing technology could provide handheld OD and fluorescence monitoring solutions with several advantages and added benefits over the existing benchtop devices. Monitoring bacterial growth and protein expression stand to benefit immensely from miniaturised handheld optical devices.

Considering the broad need for continuous remote bacterial growth monitoring, a fitness tracker-based handheld device, ODX, was developed to monitor

continuous bacterial growth (Figure 6.1). Leveraging the capability of ODX, its potential to measure fluorescence is also described.

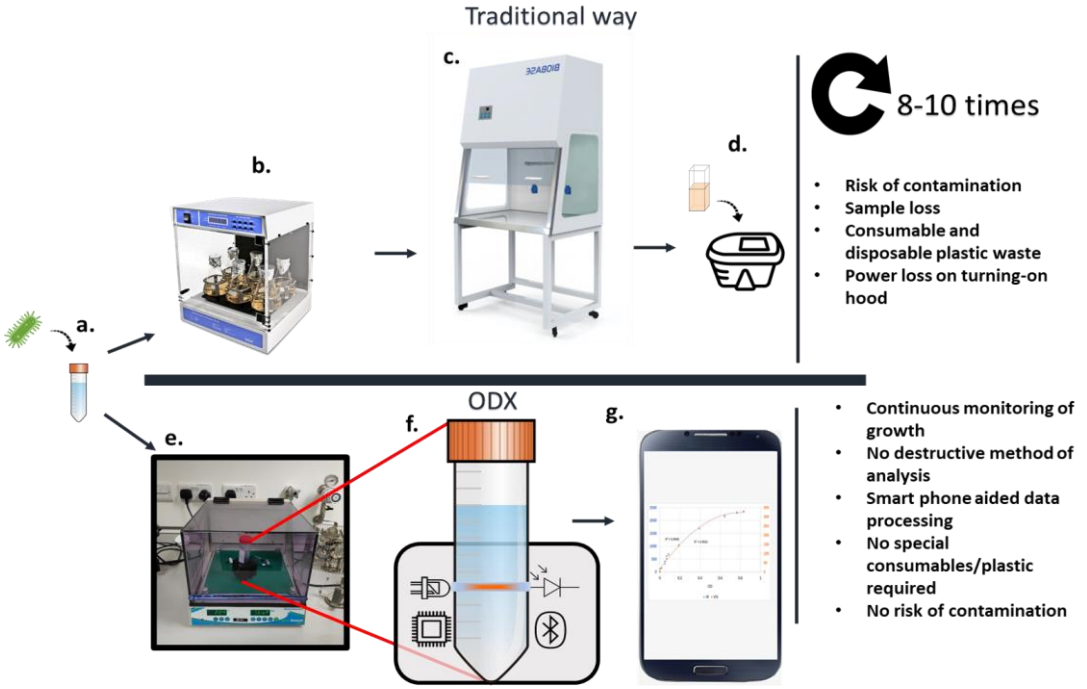


Figure 6.1: A schematic comparing the traditional method to the ODX device-based method for continuous optical bacterial growth monitoring. *(a) The tube containing appropriate culture medium inoculated with bacterial cells. Traditional (b) Bacterial cultures in a shaking incubator (c) Laminar airflow chamber (all the biological sampling is done inside a laminar airflow chamber to reduce the potential risk of contamination) (d) Culture sample collected in a cuvette and OD measured using a commercial benchtop spectrophotometer. ODX (e) The sample is inserted into the ODX device and placed in a shaking incubator. (f) Basic components of ODX. (g) Data collected via Bluetooth are processed on ODX smartphone app and readouts displayed.*

6.3 MATERIALS AND METHODS

Hardware modifications were assisted by colleagues; Dr. Ganga Chinna Rao Devarapu and Uday Gowda in the Cork Institute of Technology.

6.3.1 Optical and mechanical design of ODX device and choice of materials

The ODX hardware consists of the following parts: 1. A generic fitness tracker 2. The 3D printed enclosure 3. An orange LED. 4. A voltage regulator and a current regulator.

6.3.2 Fitness tracker

An ID107HR branded (Shenzhen DO Intelligent Technology Ltd) fitness tracker was chosen for the work presented in this article as it is inexpensive ($10 to $25) and widely available through online retailers. More importantly, it contains an nRF51822 microprocessor from Nordic Semiconductors Ltd (Figure 6.2a), which has well-documented open-source firmware development tools for modifications.

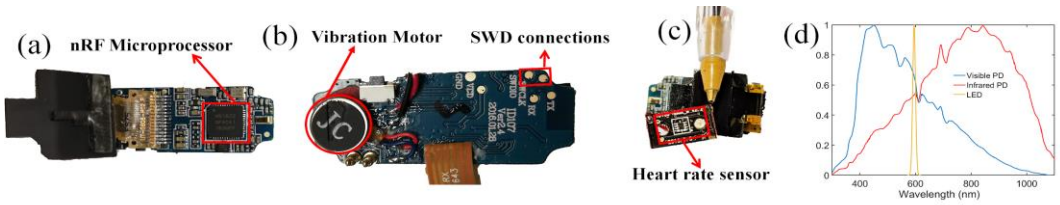


Figure 6.2 The internal circuitry of the ID107HR fitness tracker. *(a) Rear view of the circuit board showing SWD programming connections to upload the firmware into the fitness tracker (b) Front view of the circuit board showing nRF51822 microcontroller and OLED display. (c) Heart rate sensor consisting of visible and IR photodiodes. (d) The spectral response of the visible (blue line) and IR (red line) photodiodes of the heart rate sensor (Si1143) [22] in the ID107HR fitness sensor. The emission spectrum of the LED is represented by the orange line*.

The heart rate sensor (Si1143, Silicon Labs Ltd) in the ID107HR fitness tracker has two photodiodes (Figure 6.2c): one to cover the visible spectral range and the other to cover the infrared (IR) spectral range as indicated with blue and red lines respectively in Figure 6.2d. Raw readings from both PDs can be accessed using the modified firmware. These two PDs provide two complementary measurements for each optical density measurement of bacteria, thus resulting in more accurate OD values than OD meter designs that have only a single PD.

6.3.3 LED

The optical density of bacteria is usually measured at a wavelength of 600 nm as most bacteria and growth media in which bacteria are incubated known to have negligible absorption at that wavelength [14]. Therefore, an orange colored LED (C503B-AAN-CY0B0251, CREE) with peak emission at a wavelength of 596 nm was chosen as the light source for the ODX device as it has low power consumption, small size, low weight, high robustness, and acceptable monochromaticity.

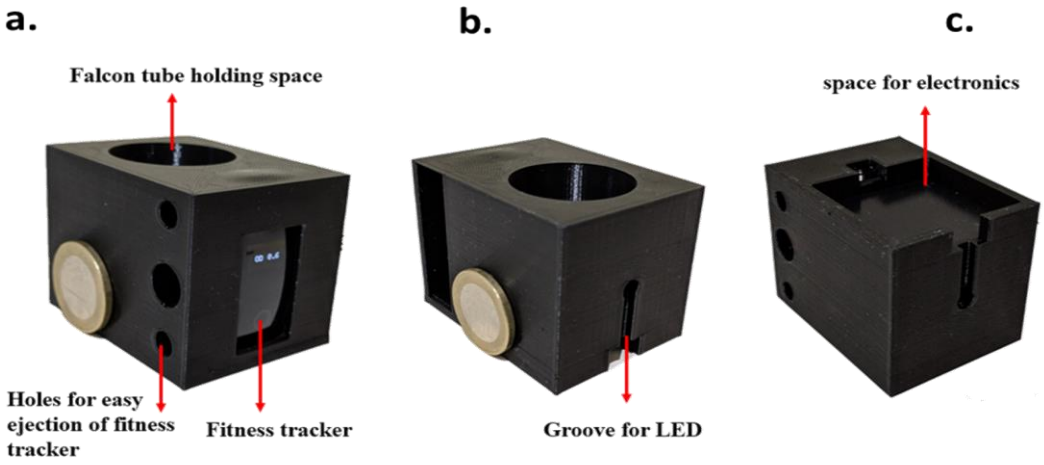


Figure 6.3 3D printed enclosure of ODX device *(a) Front view showing the holding space for a culture tube (sample holder), (b) Rear view showing the groove for an orange LED. (c) Bottom view showing the space for LED controlling circuit. A one-euro coin was placed for size comparison to ODX device.*

6.3.4 3D printed enclosure

The enclosure for the ODX device was designed with an open source parametric CAD software (OpenSCAD version 2015.03-2). The CAD design of ODX has provisions for holding the fitness tracker, a culture tube (Figure 6.3a), an orange LED (Figure 6.3b) and the additional circuitry powering the LED (Figure 6.3c). The CAD design of the ODX enclosure was fabricated using a 3D printer (Ultimaker 3) with black coloured Polylactic Acid (PLA) material.

6.3.5 Smartphone app development

An Android app was specifically developed using an open source platform (App Inventor) to transfer the data from the ODX device to a smartphone via Bluetooth. The app then processes the data and displays the OD on the screen. Moreover, the app displays the growth of bacteria graphically in terms of OD and saves the data in a text file inside the smartphone's internal memory for further analysis. However, the primary role of the app was to let the users create alerts informing them when a bacterial culture reaches the required growth stage. The working mechanism of the firmware and the app are shown schematically in Figure 6.4a and Figure 6.4b, respectively, while Figure 6.4c shows the user interface of the ODX app.

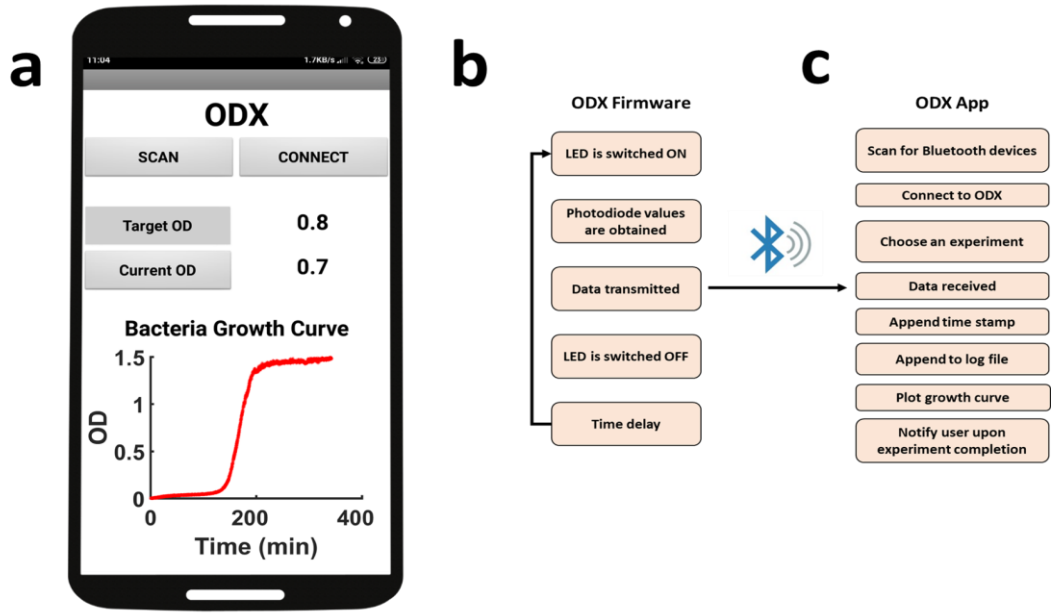


Figure 6.4 Functional workflow of ODX device *(a) Firmware (b) Smartphone app (c) ODX app user interface.*

6.3.6 Bacterial culture

E. coli Nissle was grown on Luria Broth agar plates and 25 ml inocula made in LB broth from single colonies, before shaking at 37 °C overnight. *S. aureus* (ATCC 25923) was cultured on Trypsin soy broth (TSB) agar plates and 25 ml inocula made in TSB broth from single colonies, before shaking at 30 °C overnight. *S. agalactiae* was grown on Trypsin soy broth (TSB) agar plates and 25 ml inoculum made in TSB broth from single colonies. Bacteria were sub-cultured in 50 ml falcon tubes by adding 100 µl overnight bacterial culture to 30 ml of fresh broth. For batch measurements, bacterial cultures were diluted, and OD readings given by ODX device were recorded. For continuous measurement, the falcon tube containing freshly inoculated broth was inserted into the ODX device. The measurements were logged into a text file, which was later used for optimizing the ODX app. In all the cases, OD was cross verified by a standard spectrophotometer (Eppendorf BioPhotometer) in biological and technical replicates. Continuous growth monitoring was performed by seeding the bacteria into Luria-Bertani broth (LB) and monitoring the growth pattern over 10 h.

6.4 RESULTS

6.4.1 Deploying ODX hardware

6.4.1.1 Calibration of ODX device

To convert the ODX device output into optical density values, it was necessary to determine the empirical relationship between the 'Visible and IR' photodiode values of the heart rate sensor and the OD values given by a benchtop spectrophotometer. For this purpose, ODX was calibrated using three bacterial samples. Overnight cultures of *E. coli* Bl21, *S. aureus* and *S. agalactiae* were diluted to five different concentrations. Each sample was measured using the ODX as well as a traditional benchtop spectrophotometer and these results are plotted in Figure 6.5. A logarithmic function is fitted to these data sets following the Beer-Lambert law [15]. The quality of fit (R^2) obtained for the visible and IR photodiodes were above 0.9 for all the three bacterial solutions, indicating the excellent quality of fit and showing the accuracy of above 96% (assuming that the benchtop spectrometer has minimal error). For each of the bacterial solution, logarithmic functions of the photodiodes are shown in Table 6. 1.

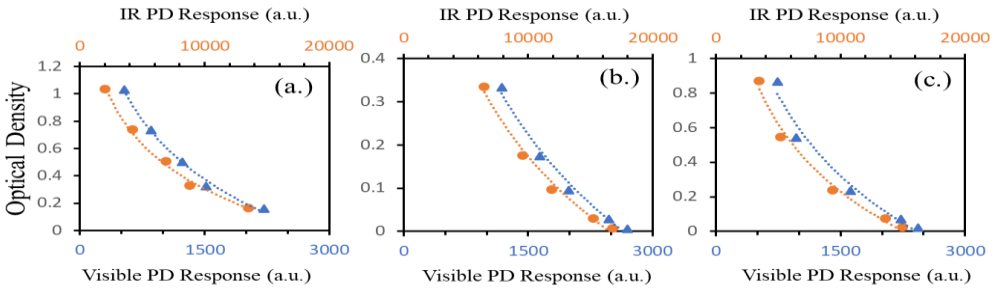


Figure 6.5 Response of the IR photodiode (blue dots, top x-axis) and Visible photodiode (orange dots, bottom x-axis) for different optical density solutions (y-axis). *ODX was calibrated using three different bacterial strains (a) Escherichia coli Nissle, (b) Streptococcus agalactiae and (c) Staphylococcus aureus. The respective coloured lines represent the logarithmic trend lines.*

Bacteria	Polynomial regression function	R^2
Escherichia coli	OD_{IR} = -0.47 ln[IR] + 4.6351	0.9923
	OD_{VIS} = -0.625 ln[Vis]+4.9436	0.9944
Streptococcus agalactae	OD_{IR} = -0.344 ln[IR] +3.3413	0.9903
	OD_{VIS} = -0.396 ln[Vis] + 3.1196	0.9865
Staphylococcus aureus	OD_{IR} = -0.563 ln[IR] +5.4173	0.9878
	OD_{VIS} = -0.674 ln[Vis] +5.2519	0.9737

Table 6. 1 Calibration of the ODX using bacterial samples. *Best fit functions and corresponding R^2 for the IR, and Visible photo diode values as a function of OD.*

These logarithmic functions were used to obtain the optical density corresponding to each photodiode (OD_{IR} and OD_{vis}). The final optical density of a bacterial solution is obtained by calculating the weighted average of corresponding individual optical densities of IR and Vis photodiodes, i.e. OD = 0.5 (OD_{IR}+OD_{vis}). Accordingly, these logarithmic calibrating functions are programmed into the persistent memory of the fitness trackers microprocessor, so that raw readings of the photodiodes will directly output the OD values.

6.4.2 Continuous bacterial growth monitoring

The performance of the ODX device to continuously monitor bacterial growth was evaluated with three bacterial strains *E. coli, S. aureus* and *S. agalactiae*. Bacteria were inoculated in a 50ml tube containing growth media. The tube was inserted into the ODX device and placed in a shaking incubator. OD readings corresponding to the bacterial strains were collected wirelessly via Bluetooth-enabled smartphone and recorded using the ODX Android app. OD readings were measured every 8 sec and data extracted to plot the data representing the growth curves of the three bacterial strains. The measurement periods for these organisms were approximately 10h, allowing the collection of complete growth dynamics data. The results for batch-wise bacterial growth monitoring is shown in Figure 6.6. The three major growth phases of bacterial growth, the lag, log and stationary, were clearly evident in Figure 6.6.

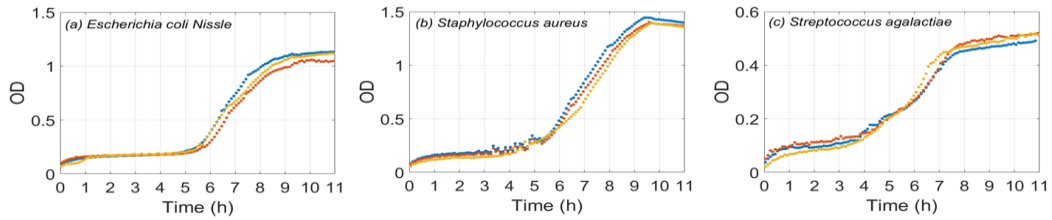


Figure 6.6 Growth curves of (a) Escherichia coli, (b) Staphylococcus aureus and (c) Streptococcus agalactiae obtained with the ODX device. *In all cases, ODX was placed in a shaking incubator and the OD measurements (once every 8 sec) were recorded on a Bluetooth enabled smartphone. The resultant data are plotted after averaging the values for every 400 sec to avoid the slight variations in the OD values resulting from the shaking of tubes inside the incubators.*

In all experiments, data from ODX were transmitted continuously to the ODX app via Bluetooth and throughout the process, it could be accessed in real-time. This real-time access to bacterial growth phase and optical density would be a highly valuable asset for fine-tuning the process efficiency in biotechnology industries.

6.4.3 Adapting ODX for fluorescence

As shown in Figure 6.2d, the photodiodes on the fitness tracker present in ODX have the capability to capture a wide range of emission spectra in the visible and IR wavelength regions. This allowed the adaptation of ODX into a fluorescence monitoring device. A second LED, whose emission wavelength matches the excitation wavelength of FlAsH-EDT2 (see Chapter 5), was engineered in ODX at the side interface to provide access to side illumination. An emission filter was placed in front of the diode (Figure 6.7). The software of the device was re-programmed to alternatively switch between the OD LED and Fluorescence LED, with an interval of 8 sec. This ensures the capturing of two different photodiode outputs in parallel. Such a dual LED system could perform both OD and fluorescence monitoring simultaneously.

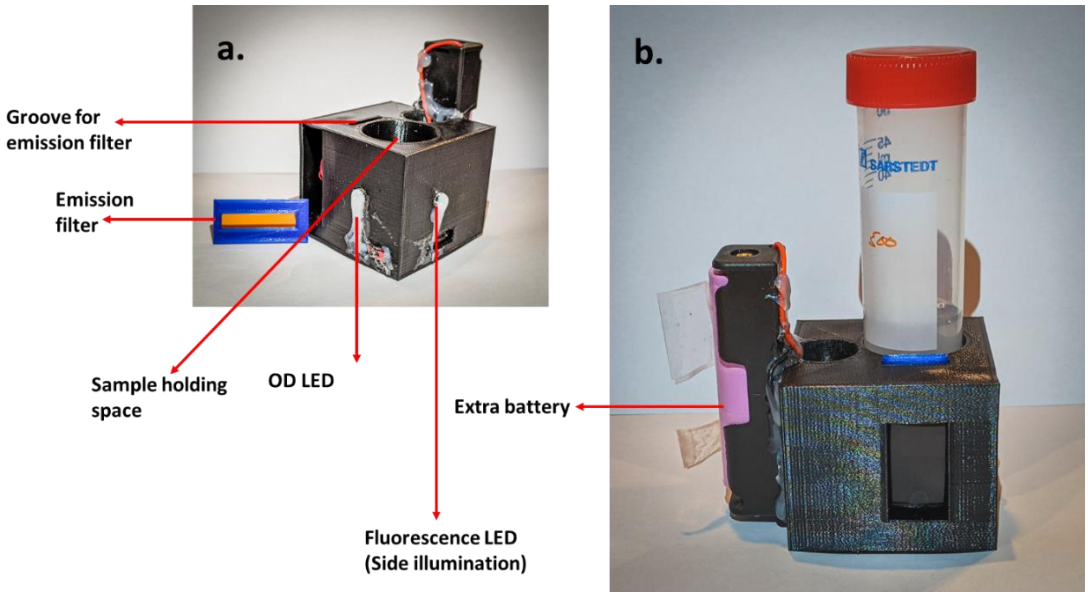


Figure 6.7: Adapting ODX for fluorescence. *(a) An additional groove for an emission filter was placed in front of the photodiodes. The LED for fluorescence excitation is placed on the side. (b) The build looks highly similar to ODX. An extra battery is added to ensure prolonged measurements even while continuously utilising both the LEDs.*

The fluorescence capabilities were tested with Polystyrene Microspheres, 10 µm, yellow-green fluorescent (505/515) FluoSpheres™ (ThermoFisher Scientific F58836). Serial dilutions were made up to 10 different concentrations. All measurements were made in the dark. Although a noticeable photodiode response was lodged, the validation was unsuccessful as no linearity was observed in the data (data not shown). Further optimisation of excitation light intensities, integration times, and the positioning of the LEDs needs to be carried out to bring this device to a fully functional state.

6.5 DISCUSSION

Monitoring of bacterial growth represents a staple process in every bacterial or biological engineering focused laboratory. Plotting the growth characteristics of various bacteria is important to harvest cells for protein production and to study the effects of various test substances on bacterial growth. The traditional method to measure OD and monitor bacterial growth is a time-consuming process, uses a huge amount of plastic consumables and adds risk of contamination. In this study, it has been shown that the ODX device has successfully overcome most of the problems and challenges posed by the traditional OD measuring methods. Bacteria are extensively used in research and industry. In most cases where bacteria are used as bio factories, the products (proteins, secondary metabolites etc.) are produced in the log phase of the bacterial growth. In such settings, the bacterial OD is maintained between 0.4 and 0.8. For cloning DNA, bacterial OD of between 0.6 and 0.8 is preferred. ODX showed similar lower limit of detection to the benchtop spectrophotometer (as low as 0.01 OD, approx. 10^6 CFU). Thus, ODX maintains the same levels of quality standards as the benchtop spectrophotometers.

By combining a generic fitness tracker and a smartphone-aided data reporting system, ODX forms a complete continuous bacterial monitoring system. In this study, ODX has been tested on three different bacterial strains and their growth was monitored continuously for over 10h. The resultant growth curves are shown in Figure 6.6 resemble the typical bacterial growth curves [26]. ODX device presents several advantages over the existing commercial and DIY spectrophotometers. Since the ODX is ultra-portable, it could be used in various biological settings such as shaking incubators, anaerobic incubators or sterile laminar airflow chambers, thus eliminating the potential chances of contamination. ODX could also be used as a regular benchtop spectrophotometer as it can display the OD values on the OLED Screen without requiring a smartphone or a computer. ODX works with a range of standard sample containers. The current device was tested using a standard cuvette and a generic test-tube (data not shown), thus eliminating the need for specific consumables.

One of the key aspects of ODX is the availability of low-cost fitness trackers, that makes it affordable. Today, fitness trackers are available at a retail price of less than $10. These fitness trackers have all the electronic components required to make

OD meters. The same components, when bought individually, are the main contributors to the high costs of the currently proposed prototypes.

Although the current work explores the potential of ODX primarily in an academic lab setting, the scope of ODX is not only limited to academic labs. Biotech industries such as the recombinant protein production industries, fermentation industries, dairy, and food-based industries, use turbidity and optical density monitoring for both batch and continuous quality monitoring. The portability and modularity to adopt in many settings expand the potential of ODX into any biotech industry setting. The ability to continuously log the data with unmanned supervision (via Bluetooth) reduces the risk of data fraud and data loss. This data log file could later be used for retrospective inspections [27]. Using wireless Bluetooth based systems avoids sophisticated wiring systems and decreases physical maintenance costs in the industry. OD methods are also widely used in clinical labs and hospitals for various blood, urine, and other body fluid analyses [28], [29]. The ODX could be deployed as an NPD (Near patient diagnostics) which reduces the burden on personnel on the health sector. The data logging system could form a very helpful feature for patients who require regular monitoring of body samples. This eliminates the manual errors in clinics where analyses are still done by physical examination by a staff member.

In this work, the ability of ODX to continuously monitor the bacterial growth is shown, this is not possible with the current benchtop spectrophotometers. With all the features such as portability, versatility and the customisability ODX can be a valuable tool for monitoring bacteria in a wide range of academic and industrial settings. Continuous fluorescence monitoring capabilities were added to ODX with the aim of aiding confirmation of self-assembly of test constructs in Chapter 5. It was important to measure the bacterial growth continuously to validate the instrument as well as to provide a baseline for reference. This however, proved unsuccessful. In future, further hardware optimisation will be made to bring fluorescence features to ODX. This addition of fluorescence capability to such a handheld device would increase the value of the invention and benefit wet-lab experimentation.

6.6 CONCLUSION

This study explores the scope of introducing miniaturised optical devices for biological experiments. Handheld devices such as ODX, bring deployability and reduce reliance on high-cost, bulky lab equipment. In this work, the end goal of continuous monitoring of bacterial growth was facilitated through various interdisciplinary resources. Enabling technology such as 3D printing and miniaturisation have provided the ease of design-model-build-test of multiple hardware and software prototypes. Modern day biomedical science demands novel concepts with deployable technology to assist their translation into user-based settings. The approach taken in this chapter uses principles of electronics, material design, wet-lab assays and optics, working in tandem to deliver on the holistic goal.

This approach stands as a unique example to demonstrate a strategy that would guide novel scientific concepts into deployable and commercially viable products.

6.7 REFERENCES

1. Shapira, P., S. Kwon, and J. Youtie, *Tracking the emergence of synthetic biology.* Scientometrics, 2017. **112**(3): p. 1439-1469.
2. Du, J., Z. Shao, and H. Zhao, *Engineering microbial factories for synthesis of value-added products.* J Ind Microbiol Biotechnol, 2011. **38**(8): p. 873-90.
3. Kamionka, M., *Engineering of therapeutic proteins production in Escherichia coli.* Curr Pharm Biotechnol, 2011. **12**(2): p. 268-74.
4. Flores Bueso, Y., P. Lehouritis, and M. Tangney, *In situ biomolecule production by bacteria; a synthetic biology approach to medicine.* J Control Release, 2018. **275**: p. 217-228.
5. Lehouritis, P., G. Hogan, and M. Tangney, *Designer bacteria as intratumoural enzyme biofactories.* Adv Drug Deliv Rev, 2017. **118**: p. 8-23.
6. Hall, B.G., et al., *Growth rates made easy.* Mol Biol Evol, 2014. **31**(1): p. 232-8.
7. Hibbing, M.E., et al., *Bacterial competition: surviving and thriving in the microbial jungle.* Nat Rev Microbiol, 2010. **8**(1): p. 15-25.
8. Herigstad, B., M. Hamilton, and J. Heersink, *How to optimize the drop plate method for enumerating bacteria.* J Microbiol Methods, 2001. **44**(2): p. 121-9.
9. Kogure, K., U. Simidu, and N. Taga, *A tentative direct microscopic method for counting living marine bacteria.* Can J Microbiol, 1979. **25**(3): p. 415-20.
10. Lee, S. and J.A. Fuhrman, *Relationships between Biovolume and Biomass of Naturally Derived Marine Bacterioplankton.* Appl Environ Microbiol, 1987. **53**(6): p. 1298-303.
11. Koch, A.L., *Turbidity measurements of bacterial cultures in some available commercial instruments.* Anal Biochem, 1970. **38**(1): p. 252-9.
12. Koydemir, H.C. and A. Ozcan, *Wearable and Implantable Sensors for Biomedical Applications.* Annu Rev Anal Chem (Palo Alto Calif), 2018. **11**(1): p. 127-146.
13. Henriksen, A., et al., *Using Fitness Trackers and Smartwatches to Measure Physical Activity in Research: Analysis of Consumer Wrist-Worn Wearables.* J Med Internet Res, 2018. **20**(3): p. e110.
14. Shibata, K., A.A. Benson, and M. Calvin, *The absorption spectra of suspensions of living micro-organisms.* Biochim Biophys Acta, 1954. **15**(4): p. 461-70.
15. Kutschera, A. and J.J. Lamb, *Cost-Effective Live Cell Density Determination of Liquid Cultured Microorganisms.* Curr Microbiol, 2018. **75**(2): p. 231-236.

DISCUSSION AND FUTURE PROSPECTS

The work demonstrated in this thesis provides a proof-of-concept of *in silico*-aided design-model-build-test strategies for synthetic proteins. Various computational tools were used to ensure proper functioning of the subparts of all the synthetic proteins. Such a computationally informed design strategy would empower wet-lab biologist with prediction capabilities. 100s of different test variants of each synthetic protein were designed and modelled. Data from computational tools was used to screen for potential best performers. This process is analogous to high throughput screening that is observed in wet-lab drug design. However, the *in silico* screening only takes a fraction of the time to test the constructs in the wet-lab. The design and modelling process in this thesis heavily relied on computational tools. However, it must be noted that the accuracy of each computational tool is subjective to each protein. Predictions of synthetic proteins that resemble close similarity to naturally existing proteins have higher confidence levels and accuracy. Community-wide, worldwide studies such as CASP (Critical Assessment of protein Structure Prediction), CAPRI (Critical Assessment of PRediction of Interactions) and CAFA (Critical Assessment of Functional Annotation), organise regular blinded challenges to compare the performance of various computational tools for proteins [1-3]. With artificial intelligence, neural networks, deep learning and increased computational power, the road ahead is promising. The strategy presented here will grow in accuracy as the individual tools get updated.

Development of targeted synthetic protein specific for S. aureus. Chapter 2 aimed to target ClfA of *S. aureus.* The basic functioning of the synthetic protein was validated using wet lab luminescence assays. Based on wet-lab data, the best performer was identified. The S1 test construct clearly stood as the best performer amongst all other test construct variants. With the encouraging results observed in the dose response assays, S1 was identified as a candidate worthy of examining in a future *in vivo* setting. Previous literature on the Nanoluc system indicated higher brightness and prolonged half-life, when compared to Gluc. Hence the Gluc was replaced with Nanoluc, to increase the signal intensity. This proved to be helpful later in Chapter 4. The technology developed would benefit from future, further validation work involving

FACS using anti-Flag fluorescent antibody and Western blottting to validate protein sizes, integrity and concentration.

The in silico myriad problem and F2F bridge. As discussed in Chapter 2, the design parameter metrics define the quality of the individual parameter in the design and present an overwhelming amount and variety of data which are practically impossible to comprehend. Chapter 3 brings in mathematical concepts of machine learning to help predict the combined effect of the design parameters on the overall performance of a synthetic protein. F2F bridge is a novel way to visualize and score the overall performance of a test sequence. F2F bridge is a very useful tool for informed protein design, and may be deployed for low throughput design or high throughput screening. In a low throughput setting, the F2F plot would play a pivotal role in highlighting the 'pitfalls and merits' of the design corresponding to a test sequence. The design could then be improved and F2F plot could be regenerated until satisfactory design is obtained. In a high throughput scenario, multiple designs of test sequences for an overall function are scored and ranked by F2F bridge.

Using machine learning approaches such as Lasso regression and the Random Forest regression tree models, F2F showed promising prospects for predicting the overall performance of a test sequence. The workflow and implementation of the prediction model is straightforward computationally compared with wet-lab experimentation. Synthesis/expression, functional assays and quality assessments of proteins is time consuming and is capital intensive. This results in a small size of dedicated datasets. Using datasets from multiple sources is also challenging due to the vast variability in wet-lab experimentation. Although this seems as a big challenge, recent advances in synthetic biology provide an inspiring platform for high throughput synthesis and testing of multiple proteins. Future work on F2F would involve training the mathematical model with larger datasets. Larger datasets would ensure increases in significance in prediction. The current model also relies on web servers and third party tools for assessing *in silico* parameters of a test sequence. This also means that the accuracy of the individual *in silico* parameter values depends on the corresponding tools developed by various sources. In a future version, the accuracy of every individual web server/tool would be taken into consideration to provide an overall confidence score on the predicted performance. This would prevent error-compounding. With little further adjustments in V.20, F2F bridge integrates in

the DMBT cycle and adds the 'learn' step by empowering the end-user (wet-lab biologist) with a holistic view on the overall performance of a protein.

In vivo imaging using synthetic proteins. The technical know-how provided by F2F bridge and the results from Chapter 2, encouraged the building of synthetic protein for *in vivo* cell imaging. *In silico* design strategies from Chapter 2 were adapted to target MUC1. *In vitro* testing was carried out to validate secretion and functioning (binding and luminescence) of the all the synthetic proteins. Based on wet-lab validation and the 'optimal performer' logic, discussed in Chapter 4, M1 was chosen as the best candidate for *in vivo* imaging. The ability of the synthetic protein to circulate systemically throughout the body and localise to a specific target to produce an imageable luminescence signal acts as a proof-of-concept for *in silico*-aided synthetic protein design of targeted proteins for *in vivo* use (therapeutic, imaging etc). *In vivo* studies using systemically-administered M1-Gluc and S1-Gluc indicated localisation of this protein to target cells. In Chapter 2, a variant of S1 with Nanoluc was designed and tested *in vitro*. These assays didn't show any noticeable *in vitro* advantage when Gluc was replaced with Nanoluc. However, *in vivo*, the NanolucS1 construct showed a significant improvement in signal intensity. In the future, similar to the synthetic proteins in Chapter 2, FACS using anti-Flag fluorescent antibody and Western blot could be performed to validate protein sizes, integrity and concentration. Furthermore, *in vivo* work would involve administering the mice with a non-targeting version of the proteins, which would have been beneficial to rule out the chances of accidental accumulation. Timepoint optimisation and pharmacokinetic studies of the protein would also benefit understanding the concepts such as bioavailability.

In silico aided interface engineering of synthetic proteins. The capabilities of computational tools in protein design were exploited in both Chapter 2 and Chapter 4. In Chapter 5, *in silico*-aided protein engineering was used to engineer the protein interfaces to help visualise protein-protein interactions during self-assembly. A novel reporting strategy was introduced, by incorporating cysteine residues at the interaction interface of monomeric proteins of a self-assembling protein cage. The data from the wet-lab fluorescence assays provided a clear indication that FlAsH-mediated fluorescence could be successfully deployed to monitor protein-protein interactions. As designed, the FlAsH biarsenical complex was able to validate the self-assembly in

all three test proteins. However, further protein denaturing experiments are required to confirm the oligomerisation dependence of FlAsH-EDT2 based fluorescence. Techniques such as size exclusion chromatography (SEC) can be deployed to confirm the oligomerisation states. Also, further experiments are needed to test the release kinetics of the cage. The results from this work are proof-of-concept to validate the functioning of the engineered bi-partide cysteines and the reportitng concept using FlAsH-EDT2. Without such further validation, the data on the current synthetic cage are insufficient for translation into a full application.

Handheld devices for biomedical applications In Chapter 6, the scope of introducing miniaturised optical devices for biological experiments was explored. A novel handheld device, ODX, for monitoring continuous bacterial growth, with prospects of measuring biofluorescence was developed. The device was tested using different bacteria and showed accuracy levels similar to a standard benchtop spectrophotometer. Handheld devices such as ODX, bring deployability and reduce reliance on high-cost, bulky lab equipment. In this work, the end goal of continuous monitoring of bacterial growth was facilitated through various interdisciplinary resources. Enabling technology such as 3D printing and miniaturisation have provided the ease of design-model-build-test of multiple hardware and software prototypes. Modern day biomedical science demands novel concepts with deployable technology to assist their translation into user-based settings. The approach taken in this chapter uses principles of electronics, material design, wet-lab assays and optics, working in tandem to deliver on the holistic goal. ODX stands as a unique example to demonstrate a strategy that would guide novel scientific concepts into deployable and commercially viable products.

The goal of this thesis was to bridge interdisciplinary approaches in science to (i) aid laborious wet-lab experimentation and (ii) transform the novel biomedical concepts into deployable products. Chapters 2, 4 and 5 exploit the computational tools available today, to design and validate imaging strategies using synthetic proteins. Chapters 3 and 6, on the other hand act as enabling technology that improves the ease and pace of the research. The marriage between these two goals has resulted in outcomes that have various future applications in biomedical science.

References

1. Moult, J., et al., *Critical assessment of methods of protein structure prediction (CASP)-Round XII.* Proteins, 2018. **86 Suppl 1**: p. 7-15.
2. Lensink, M.F., S. Velankar, and S.J. Wodak, *Modeling protein-protein and protein-peptide complexes: CAPRI 6th edition.* Proteins, 2017. **85**(3): p. 359-377.
3. Dessimoz, C., N. Skunca, and P.D. Thomas, *CAFA and the open world of protein function predictions.* Trends Genet, 2013. **29**(11): p. 609-10.