THE UNIVERSITY *of* EDINBURGH

Data-driven evaluation of designed proteins using structural features, machine learning and cell-free expression systems

Michael James Stam

Doctor of Philosophy

Centre for Doctoral Training in Biomedical Artificial Intelligence

School of Informatics

University of Edinburgh

2024

Abstract

Proteins are the biological molecules that perform almost all the biochemical work that is necessary for life. Native proteins have a vast array of functionality as catalysts, materials, signalling molecules and more. They also have applications outside of their natural context as therapeutics, sensors, and industrial feedstocks. *De novo* protein design aims to find new protein sequences with useful properties, that can be used to solve challenges across scientific areas. Unfortunately, protein design has several limitations, including high failure rates, challenges in designing towards specific functions, and many design methods are inaccessible to non experts. This PhD project has three major research outputs which aim to address some of the limitations of protein design. Firstly, the DEsigned STRucture Evaluation ServiceS (DE-STRESS) web server was developed, which generates a set of physico-chemical properties for protein structural models, in order to evaluate designs. DE-STRESS includes functionality which allows users to design towards functions, and the web server was developed to be responsive and user friendly. Secondly, analysis was performed which demonstrated that the DE-STRESS features were predictive of *in vivo* protein production levels, and that they varied systematically across half a million predicted structures from 48 organisms, to such an extent that the tree of life could be reconstructed. This first result is significant as it provides evidence that DE-STRESS is valuable for ranking protein designs, and the second result suggests that the properties of proteins are optimised to their unique chemical environment, which could be used to develop more robust design methodologies. Finally, a method for screening designs in *E.coli* cell-free systems was developed, which will be used to explore the relationship between the DE-STRESS structural features and failure reasons of designed proteins. The insights gained from this work will be used to screen designs to avoid some of the common reasons for failure. Overall, the results from this PhD show how structural features of proteins, combined with machine learning methods and cell-free systems, can be used to increase the reliability and accessibility of protein design, so that it can be become a vital tool for researchers, in solving challenges across medicine, agriculture, energy and beyond.

Lay Abstract

Proteins are tiny molecular machines that are critical for life to exist. They have a large variety of roles such as protecting the body from bacteria and viruses, carrying messages and even creating different types of materials. Amino acids are the building blocks that are combined to create proteins. Currently, nature has only explored a small number of the various ways that these building blocks can be put together. This means that there is an opportunity to explore different methods to combine these amino acids and to design new proteins. Designed proteins can be used for applications in medicine, for example, new drugs to treat cancer and vaccines to protect us from viruses. Additionally, designed proteins can be used in other areas, such as agriculture, to help increase the production of crops, and in the environment, for recycling plastic waste into materials we need. However, creating new proteins is a challenging task, with many designs failing when tested in labs, and protein design methods are very difficult to use for non experts. Therefore, the work in this PhD project aims to address some of the limitations of protein design, in order to make the design process easier and more reliable. Firstly, a user friendly website called DE-STRESS was developed, which provides information to scientists about their protein designs, to help them decide which proteins to test in the lab, and to help them design towards specific applications. Secondly, analysis was performed which showed that the information from DE-STRESS is related to how well a design can be produced in the lab, and DE-STRESS could distinguish between proteins from different organisms, such as, humans, plants and bacteria. These insights could be important for picking which proteins to test in the lab, and for improving the success rate of protein design. Finally, an experimental method was developed which could help understand some of the reasons why designs fail. The results from this lab work could be combined with information from DE-STRESS, in order to help avoid common reasons for failure. Overall, the work in this PhD has focused on methods to help increase the reliability and accessibility of protein design, so that it can be become a vital tool for scientists in solving challenges across many scientific areas.

Acknowledgements

To begin with, I would like to thank my PhD supervisors Dr Christopher Wells Wood, Dr Diego Oyarzún and Dr Nadanai Laohakunakorn, for their incredible support, guidance and patience throughout my PhD. Dr Christopher Wells Wood has an unlimited knowledge on everything to do with protein design, molecular biology and software engineering, and always had time to talk through ideas. Dr Diego Oyarzún and Dr Nadanai Laohakunakorn provided a huge amount of support on the machine learning methods used in this project and the experimental work, respectively, and helped me address any challenges in my work. Next, I would like to thank all the members of the Wells Wood Research Group, the Biomolecular Control Group and Synthetic Biophysical Systems Group, who I have really enjoyed working with and collaborating on projects together. Surendra Yadav and Alex Perkins patiently taught me how to work in a lab environment, how to use a pipette and how to perform my first cell-free experiments. Dr Sahan Liyanagedera provided a lot of support in my last year of lab experiments, in making high performing crude cell-free systems, cloning DNA with Gibson assembly and how to be efficient in my lab work. Additionally, I would like to thank Dr Ian Simpson, Ekaterina Churkina, Isabelle Hanlon, and the rest of the UKRI Centre for Doctoral Training in Biomedical AI for their help and guidance over the last four years. Finally, a special mention to my parents, my sister Emma Stam, Sameer Dhumale, Patrick Corsar and the rest of my friends, and my girlfriend Lauren DeLong, for listening to me constantly talk about proteins, being incredibly patient with me when I've been working a lot, and offering me a huge amount of support during this PhD project.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(*Michael James Stam*)

Table of Contents

Chapter 1

Introduction

To begin with, as the work in this thesis is focused upon protein design, this chapter will provide some background and motivation for this field, along with a literature review, in order to demonstrate the contribution of the work completed in this PhD project. Firstly, a section is provided that introduces DNA, RNA and proteins, along with the central dogma of molecular biology, which describes how information is exchanged between these molecules in living systems. The following section will cover the basics of proteins and their roles in nature, how proteins are synthesised, and some background behind protein folding, protein structures and structure prediction. After this, a section will discuss the immense size of protein sequence space and the tiny proportion that has been explored by nature. Next, the field of protein design will be discussed, including its motivation and history, current methods for designing proteins and recent advancements in this field, which have involved the increasing use of machine learning techniques. Additionally, several limitations and challenges of protein design will be detailed, which have restricted protein design from becoming more widely used by researchers across scientific areas. Finally, a section will discuss the contributions of this PhD project to the field of protein design, and how it aims to address some of these limitations by using structural features, machine learning and cell-free expression systems to evaluate designed proteins, and to make protein design more reliable and accessible to researchers.

1.1 The fundamentals of molecular biology

All living systems are comprised of fundamental units called **cells** [1]. The structures and functions of these cells have led to the incredible diversity in the organisms we see

in the world, including, humans, chimpanzees, trees, yeast, flowers, bacteria, and many more [1]. Cells contain a variety of biological molecules that form sub-structures and take part in complex physical interactions, which allow the cells to function properly, therefore, allowing an organism to function properly [1]. Life as we know it today, would not exist without three critical biological molecules: **Deoxyribonucleic acid (DNA)**, **Ribonucleic acid (RNA)**, and **proteins**. Each of these molecules has a unique structure and are polymers [2]. DNA consists of 4 different nucleotide bases: adenine (A), cytosine (C), guanine (G) and thymine (T), and RNA has the same bases, except thymine (T) is replaced by uracil (U) [2]. In contrast to this, proteins consist of 20 different **amino acids** [3], which are discussed in detail in section 1.2.2. DNA and RNA are two types of genetic material known as **nucleic acids**, which are best known for storing the instructions, or **genes**, in order to create proteins. DNA is more stable than RNA [4] and acts as a cell's library for these instructions, while RNA acts as a temporary messenger by copying the instructions from DNA and delivering them to other areas in the cell, where these instructions are then used to synthesise proteins. RNA is also known to have a variety of functional roles [5]; however, this is out of the scope of this thesis. Proteins have a huge variety of different structures, properties and important functions in cells [2], which are all explored in section 1.2. Over the last 60 years, rapid advancements in the field of molecular biology have increased our understanding of these molecules, their functions, and their importance for life.

Initially, it was widely believed that proteins, rather than nucleic acids, were the molecules that carried genetic material [6]. However, in 1943, Avery and McCarty, made a groundbreaking discovery, demonstrating for the first time that DNA was the biological molecule that contained genetic information [6]. After this, the central dogma of molecular biology, still in use today, was introduced by Francis Crick in 1957, and described how information is exchanged between DNA, RNA, and protein sequences [7]. Figure 1.1 displays a simplified version of the central dogma of molecular biology, which shows information passing from DNA to RNA through a process called **transcription**, information passing from RNA to protein through a process called **translation**, and information passing from DNA to DNA through **DNA replication** [7; 2]. This depiction is slightly simplified as there are some special cases where information can pass from RNA to RNA, RNA to DNA, and DNA to proteins [7]. However, in general, DNA is the primary source of information inside a cell, which is then used to make RNA, and then RNA is used to make proteins. Notably, the entire forward process from DNA to proteins is also known as **gene expression** [1]. This

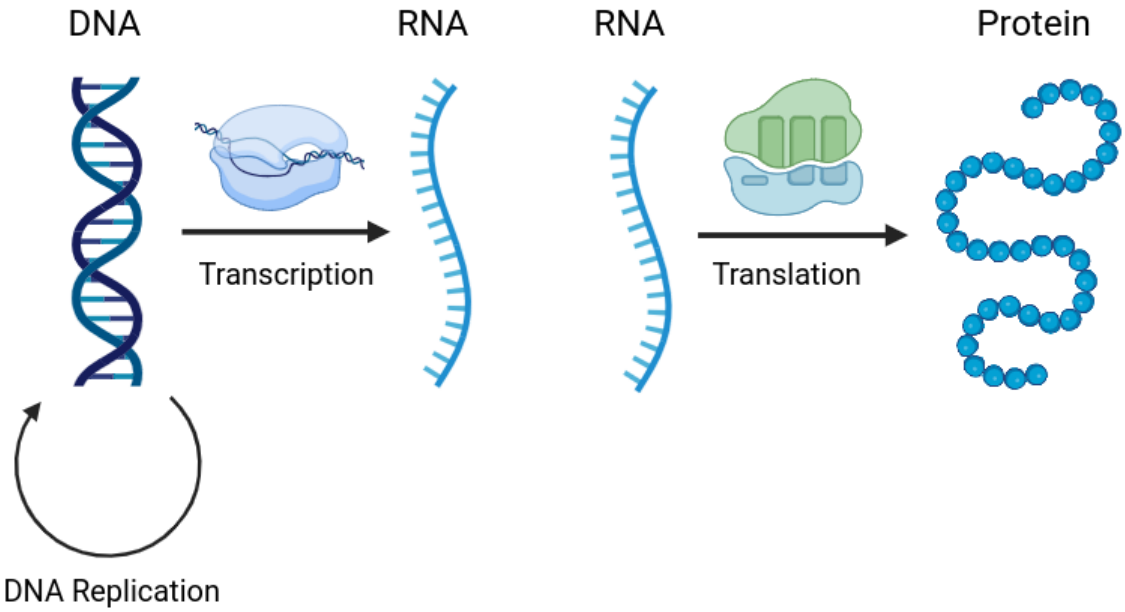

Figure 1.1: A simplified version of the central dogma of molecular biology. Information passes from DNA to DNA through DNA replication, from DNA to RNA through transcription, and from RNA to protein through translation. However, information cannot flow backwards from proteins to DNA or RNA. Created with BioRender.com.

theory states that, on the cellular level, information can flow between nucleic acids and nucleic acids, as well as between nucleic acids and protein sequences. However, it can not flow backwards from proteins sequences to nucleic acids [7]. This is partially because DNA and RNA contain information in a series of **codons**, which are consecutive frames of three nucleotides [1]. Each codon encodes for a single amino acid, and there are 64 possible codons in total, due to the fact there are 4 nucleotide bases (A, T, C and G) and 3 positions in each codon. As there are only 20 natural amino acids [3], this means that multiple codons encode for the same amino acid, which is the main reason for this loss of information. Section 1.2.2 explores the degeneracy of the codons and the genetic code in more detail.

Overall, cells are the fundamental units that make up all living systems and are responsible for the huge array of different organisms present in the world, while DNA, RNA and proteins, are critical for many of the characteristics and behaviour of these cells. In this thesis, the main focus will be on proteins, due to the incredible diversity of functions they perform in nature and the many applications they have outside of their natural context.

1.2 The fundamentals of proteins

1.2.1 Proteins and their functions in nature

Proteins are the biological molecules that perform the majority of the biochemical work that is necessary for life. These molecules make up about 60% of the organic molecules inside a cell, and they are crucial for the majority of functions of cells [2]. In nature, there are a variety of different proteins with diverse functions, such as structural proteins, enzymes, regulatory proteins, transport proteins, and many more [8]. All these different proteins form the fundamental components for life, and they contribute to the complex behaviours, characteristics, and appearances observed across organisms in this world.

To begin with, proteins contribute to many structures within cells. Examples of these structures are microtubules, which are long, hollow, cylindrical structures that are found in some cells, and are used for transporting proteins and other molecules to different areas of the cell [9]. These subcellular structures are made from proteins called tubulin [9], and an example of one is shown in figure 1.2. In addition to this, actin is another structural protein which helps to form the cellular skeleton, or cytoskeleton, which is vital for cells to maintain their shapes [10]. Other examples of structural proteins, such as keratin and collagen, are particularly characteristic to certain tissues. For example, keratin is important for the structure of hair in humans and wool in sheep [11], and collagen is important for the elastic structure of skin [12].

Additionally, many proteins have functional rather than structural roles. Enzymes, for instance, are proteins that catalyse chemical reactions [2]. Ribulose-1,5-bisphosphate carboxylase/oxygenase (rubisco) is an enzyme that is vital for capturing carbon dioxide from the air, and performing photosynthesis in plants and other organisms to create sugar and oxygen [13]. As this enzyme is very inefficient, organisms have evolved to produce a large amount of this protein, which means rubisco is one of the most abundant proteins in the world [14]. Another enzyme that is crucial in nature is RNA polymerase, which can be seen in figure 1.2. RNA polymerase is the enzyme responsible for transcription, as mentioned in the previous section, which means that this protein reads DNA in order to synthesise RNA [15]. In addition to this, other examples of enzymes in nature include: DNA helicases which are responsible for the unwinding of the DNA double helix for DNA replication [16], proteases which break the bonds between amino acids for protein degradation [17], and amylase, which is an enzyme that breaks apart starch into simple sugars in order to digest it [18]. While enzymes allow

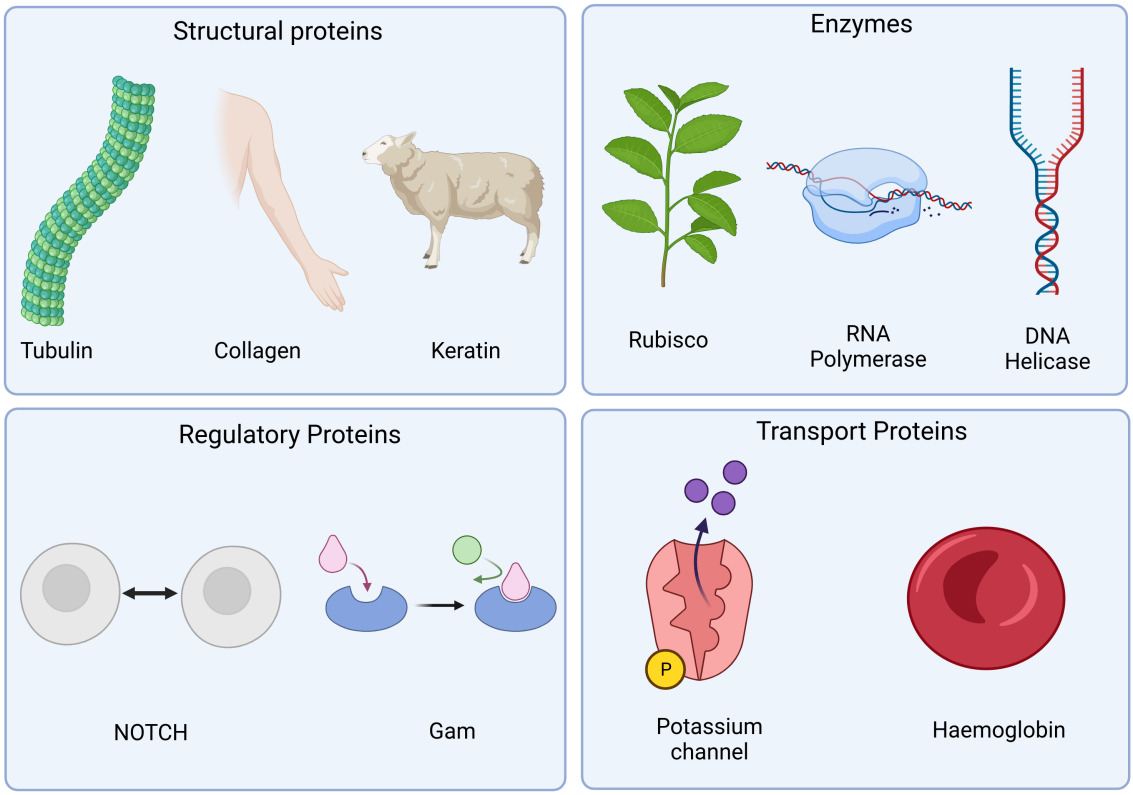

Figure 1.2: Examples of the wide variety of roles that proteins have in nature. These include, but are not limited to, structural proteins, enzymes, regulatory proteins and transport proteins. Created with BioRender.com.

actions to be taken within the cell, these actions and their effects must be moderated inside the cell, which is performed by another type of protein.

Regulatory proteins are another type of protein that is significant in nature, as they regulate the expression of genes and the activity of other molecules, such as proteins [2]. One example of regulatory proteins is the p53 protein, which regulates transcription by binding to DNA [19]. The inactivation of this protein is critical for the development of tumours, abnormal cell growths, and so p53 is crucial for protecting cells from becoming cancerous [19]. The NOTCH transmembrane receptor proteins are also regulatory, and they are part of a signaling pathway for multicellular organisms, such as ourselves [20]. These proteins are important for regulating essential cellular processes, such as the differentiation of cells into various types or roles within a tissue [21] and programmed cell death (apoptosis) [22]. In addition to this, some regulatory proteins directly counter the effects of other proteins. For example, the Gam protein, which was isolated from the λ bacteriophage, a virus for bacterial cells, is an enzyme inhibitor and binds onto the RecBC DNase protein, in order to stop the repair and

degradation of DNA [23; 24], as depicted in figure 1.2.

Finally, transport proteins are a type of protein which move different molecules across cell membranes or around the body [2]. Haemoglobin is one of the most famous transport proteins, which is responsible for transporting oxygen from the lungs to tissues, and transporting carbon dioxide from the tissues to the lungs, in the bodies of vertebrates [25]. Other examples of transport proteins are glucose transporters from the Glut (SLC2A) family of transmembrane transporter proteins, which transport glucose, a basic sugar, across cell membranes for the cell to use for energy [26]. Additionally, potassium channels, which are shown in figure 1.2, are large protein complexes that are situated in cell membranes, and they allow the transport of potassium ions in and out of the cell [27]. Overall, proteins are crucial for many of the biological processes and functions that happen in cells, and they have an incredibly diverse set of functions from catalysing reactions, transporting molecules, regulating biological processes, providing structural support and many more.

1.2.2 Amino acids as the building blocks of proteins

As stated in section 1.1, proteins are made from linear chains of amino acids, which are joined together with an amide (or peptide) bond [3]. An amino acid is a molecule that contains both amino groups and carboxyl groups [28], and the amino and carboxyl groups are attached to the C_α carbon, which is shown in the chemical structure in

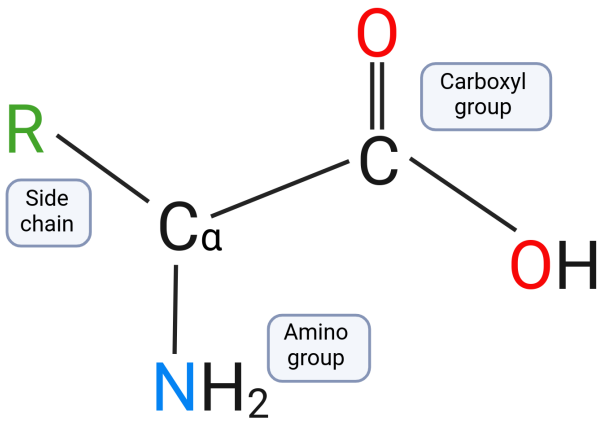

Figure 1.3: The chemical structure of an amino acid, with the amino group, carboxyl group and the side chain labelled. C represents carbon, O represents oxygen, N represents Nitrogen, H represents Hydrogen, and R represents an identifying side chain. Created with BioRender.com.

figure 1.3. In total, there are 20 different standard amino acids in nature, the identities of which are distinguished by the **side chain** of the amino acid, denoted by *R* in the chemical structure in figure 1.3 [29].

Amino Acid	**Three letter symbol**	**One letter symbol**	**Average relative abundance in proteins (%) [3]**
Alanine	Ala	A	9.0
Leucine	Leu	L	7.5
Glycine	Gly	G	7.5
Serine	Ser	S	7.1
Lysine	Lys	K	7.0
Valine	Val	V	6.9
Glutamate	Glu	E	6.2
Threonine	Thr	T	6.0
Aspartate	Asp	D	5.5
Arginine	Arg	R	4.7
Isoleucine	Ile	I	4.6
Proline	Pro	P	4.6
Asparagine	Asn	N	4.4
Glutamine	Gln	Q	3.9
Phenylalanine	Phe	F	3.5
Tyrosine	Tyr	Y	3.5
Cysteine	Cys	C	2.8
Histidine	His	H	2.1
Methionine	Met	M	1.7
Tryptophan	Trp	W	1.1

Table 1.1: The 20 standard amino acids with their common three letter and one letter symbols, along with their average relative abundance in proteins [3].

Table 1.1 shows the full list of the 20 standard amino acids, along with their common three and one letter abbreviations, and also their relative abundance in known proteins [3]. From table 1.1, we can see that alanine, leucine, glycine, and serine are some of the most abundant amino acids in proteins, while histidine, methionine and tryptophan are much rarer. The names of these amino acids can tell us something about the properties of the amino acids or where they were first isolated from. For

example, arginine forms a well defined silver salt and comes from the Latin word for silver (*argentum*), arginine was first isolated from asparagus, glycine tastes sweet and comes from the Greek word for sweet (*glykys*), and proline contains a chemical structure called a pyrrolidine ring [30]. In general, these amino acids have quite distinct properties due to their different side chains. However, they can be grouped into a few different categories.

Figure 1.4 shows the 20 standard amino acids, grouped into five categories: hydrophobic, uncharged polar, special, positive charge, and negative charge [29]. These properties of the amino acids affect how they interact with each other as well as their surroundings. Ultimately, these complex physical interactions between different amino acids and other molecules, determine the three-dimensional structure, or fold, of the protein [31], which is discussed in detail in section 1.2.3 For instance, hydrophobic amino acids such as alanine, valine and leucine, repel water and tend to be buried in

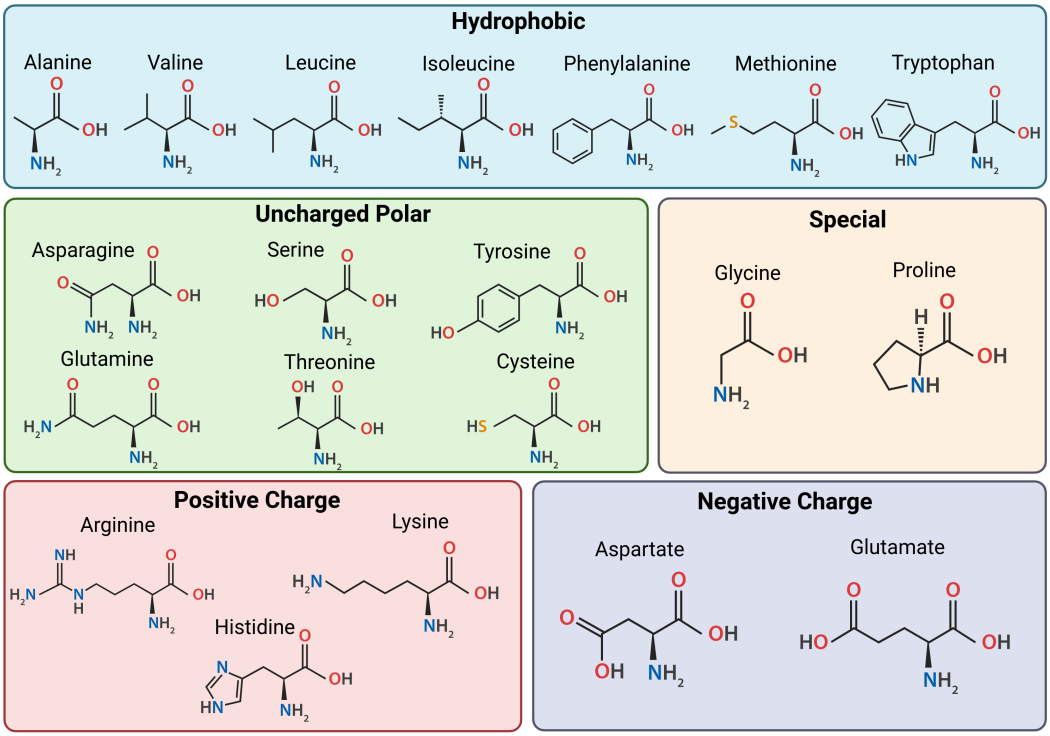

Figure 1.4: The chemical structures of the 20 standard amino acids, grouped into five categories: hydrophobic, uncharged polar, special, positive charge, and negative charge [29]. The convention for these chemical structures is that the carbon atoms, represented by vertices between lines, are not labelled, hydrogen atoms connected to carbon atoms are hidden for simplicity. The letter S depicts a sulphur atom. Created with BioRender.com.

the core of a protein [29], while uncharged polar amino acids such as asparagine, glutamine and threonine, are attracted to water and are generally found on the surface of proteins [29]. Figure 1.4 shows that aspartate and glutamate are both negatively charged, and arginine, lysine and histidine are positively charged. However, histidine can also be neutral [29]. In addition to this, glycine is classed as special as it has no charge, it is neither hydrophobic or polar, and, because it is so small, it can adopt unusual dihedral angles which increases the flexibility of the protein [29]. Finally, proline is also classed as special as the ring in its side chain joins with the amine group, which makes this side chain very rigid, bulky, and reduces the flexibility of the protein [29].

As there are 20 different amino acids and only four nucleotides in DNA and RNA,

UU	**UC**	**UA**	**UG**
UUU (Phe)	UCU (Ser)	UAU (Tyr)	UGU (Cys)
UUC (Phe)	UCC (Ser)	UAC (Tyr)	UGC (Cys)
UUA (Leu)	UCA (Ser)	UAA (Stop)	UGA (Stop)
UUG (Leu)	UCG (Ser)	UAG (Stop)	UGG (Trp)
CU	**CC**	**CA**	**CG**
CUU (Leu)	CCU (Pro)	CAU (His)	CGU (Arg)
CUC (Leu)	CCC (Pro)	CAC (His)	CGC (Arg)
CUA (Leu)	CCA (Pro)	CAA (Gln)	CGA (Arg)
CUG (Leu)	CCG (Pro)	CAG (Gln)	CGG (Arg)
AU	**AC**	**AA**	**AG**
AUU (Ile)	ACU (Thr)	AAU (Asn)	AGU (Ser)
AUC (Ile)	ACC (Thr)	AAC (Asn)	AGC (Ser)
AUA (Ile)	ACA (Thr)	AAA (Lys)	AGA (Arg)
AUG (Met)	ACG (Thr)	AAG (Lys)	AGG (Arg)
GU	**GC**	**GA**	**GG**
GUU (Val)	GCU (Ala)	GAU (Asp)	GGU (Gly)
GUC (Val)	GCC (Ala)	GAC (Asp)	GGC (Gly)
GUA (Val)	GCA (Ala)	GAA (Glu)	GGA (Gly)
GUG (Val)	GCG (Ala)	GAG (Glu)	GGG (Gly)

Table 1.2: The genetic code for translation from RNA codons to amino acid sequences. The nucleotide bases are introduced in section 1.1, and the three letter amino acid codes from table 1.1 are displayed here.

amino acids are encoded by combinations of three nucleotide bases, otherwise known as codons [1]. Table 1.2 shows the genetic code, which displays the different codons, along with the corresponding amino acid that they encode [1]. The codon AUG, typically known a start codon because it initiates translation, encodes for a methionine. However, there are some other codons that can initialise translation as well, such as GUG and CUG [32]. On the other hand, the codons UAA, UAG and UGA are all stop codons, and they terminate translation. One major observation from table 1.2, is the degeneracy in the genetic code, as we have multiple codons encoding for the same amino acid, such as UCU, UCC, UCA and UCG for serine, and GGU, GGC, GGA and GGG for glycine. This degeneracy in the genetic code makes it more robust to mutations, or changes to the sequence. Even if mutations occur in the nucleotide sequence (e.g., UCC to UCA), they may not affect the translated amino acid sequence, and therefore the resulting protein [33]. In addition to this, as shown in figure 1.4, several of the 20 standard amino acids share core properties, indicating that sometimes, amino acids could be substituted for others, without affecting the overall function of the protein [29]. Consequently, it means there are many nucleotide sequences that can translate into a particular amino acid sequence, and there are many amino acid sequences that can have a particular function. In contrast to this, despite this degeneracy, it has been shown that these synonymous codon substitutions can affect the folding pathway of the protein, and ultimately the final structure and function [34].

1.2.3 Protein folding

Protein folding is the process of how an amino acid sequence spontaneously folds into a globular protein, after it has been synthesised by a ribosome (the cellular machinery that performs translation) [31], which is illustrated in figure 1.5. Christian Anfinsen performed experiments with the enzyme ribonuclease, showing that after denaturing (unfolding) with chemicals and renaturing (re-folding) the enzyme by removing such chemicals, it was still active [35]. From this, he concluded that the amino acid sequence spontaneously folds into its active structure and that the native structure of a protein, in its normal environment, is determined by its amino acid sequence [31]. Christian Anfinsen was the first to propose the “thermodynamic hypothesis”, which states that the three-dimensional structure of a native protein, in its normal environment, is the one in which the Gibbs free energy of the whole system is the lowest [31]. The Gibbs

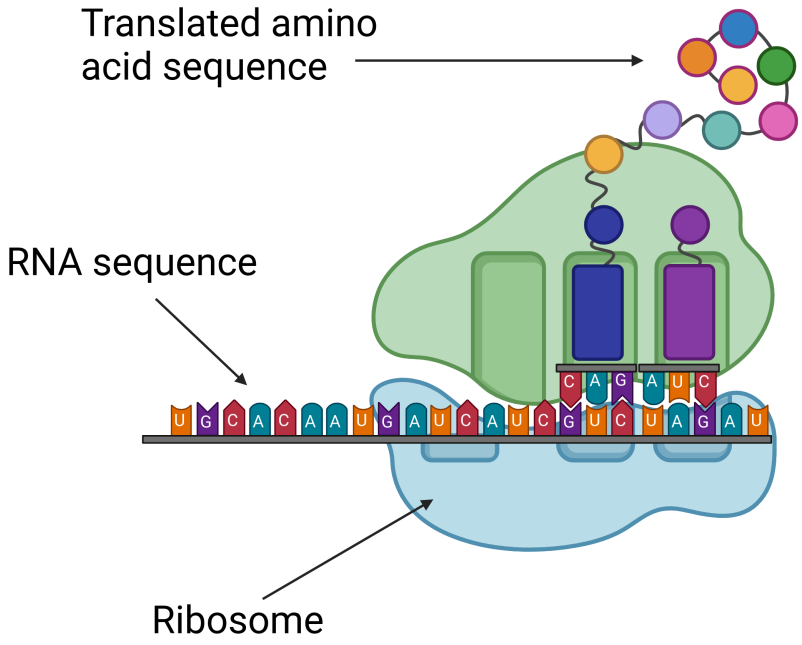

Figure 1.5: An amino acid sequence folding into a globular protein after being synthesised by a ribosome. This process is called translation. Created with BioRender.com.

free energy of a system, commonly denoted by G, is defined as

$$G = H - TS$$

where H is the enthalpy, T is the temperature and S is the entropy of the system [36]. The change in Gibbs free energy of a system, denoted by ΔG, is defined as

$$\Delta G = \Delta H - T\Delta S$$

and if ΔG is negative, this means that a physical or chemical process is favourable and it will happen spontaneously [36]. However, subsequent research showed that, in a complex environment such as inside a cell, there are other factors that can affect folding such as chaperones, which help proteins fold into their native state and reduce the probability of them folding into non-functional structures [37].

The amino acid properties shown in figure 1.4 and discussed in section 1.2.2, are responsible for the interactions between different amino acids and other molecules, and how an amino acid sequence folds. Firstly, the hydrophobic effect describes the repulsion between the hydrophobic amino acids in the sequence and the water-rich environment and is a major factor influencing the stability of the protein structure [38]. This interaction causes the amino acid sequence to fold into a more compact structure in which hydrophobic residues are buried, thereby reducing the interactions between

these residues and water, and as a result increasing the entropy of the water [38]. Another type of interaction that influences protein folding is the Van der Waals interactions between atoms, which are weak attractive and repulsive interactions caused by the movement of electrons around atoms [39]. Van der Waals forces are dependent on the distance between the two atoms [38], and although theses forces are fairly weak, there are many across the whole length of the amino acid sequence, which can have a significant impact on protein folding [40]. In contrast, hydrogen bonds are a particularly strong type of interaction that occur when a hydrogen atom is shared between two different atoms [40]. Hydrogen bonds have been shown to have a large impact on stabilising proteins during folding [41], and they are especially important for the burial of hydrophobic residues in the core of the protein [42; 43]. In addition to this, electrostatic interactions also contribute toward protein folding and help to stabilise protein structures. These interactions occur between charged amino acids, and they can form salt bridges if opposite charged residues are within 5 Å distance to one another. These salt bridges can contribute a lot more to the stability of the protein when they are buried, rather than on the surface [41]. However, if these charged residues are buried and are non-hydrogen bonded, they can contribute negatively to the stability of the protein structure [44].

Van der Waals interactions, hydrogen bonds, and electrostatics are examples of non-covalent interactions which are relatively weak interactions [3]. However, there are also covalent interactions, which are strong bonds where atoms share electrons [3],

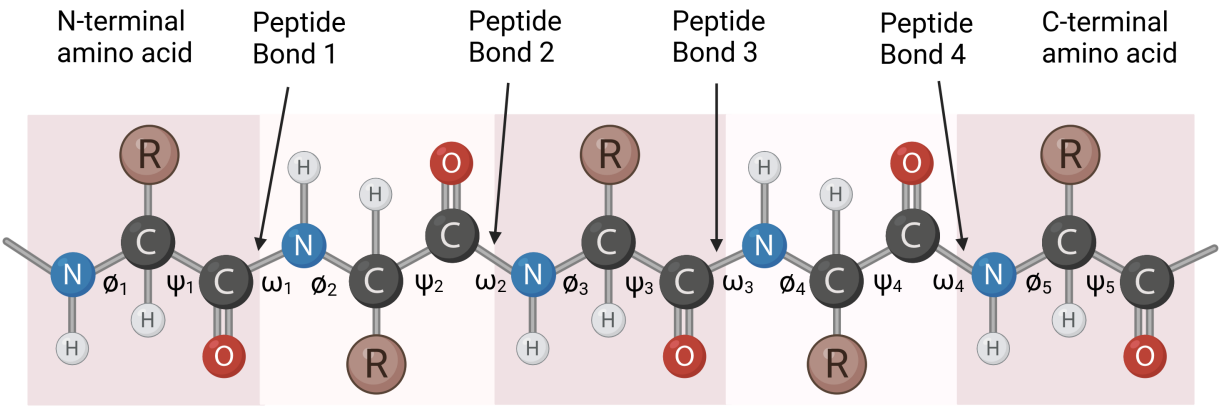

Figure 1.6: A polypeptide consisting of 5 amino acids with the peptide bonds, the N-terminal amino acid, the C-terminal amino acid, and the backbone torsion (dihedral) angles labelled. The side chains of the amino acids are denoted by R. Created with BioRender.com.

that have a large contribution in protein folding. Firstly, the peptide bonds between different amino acids in the sequence can have a large effect on how the protein can fold [3]. Figure 1.6 shows an example of how 5 amino acids are joined together into a polypeptide chain, and the peptide bonds between the amino acids have been labelled. A peptide bond forms between the carbon in the carboxyl group of one amino acid (C-terminus) and the nitrogen atom in the amine group of another (N-terminus) [3]. As this peptide bond forms, a water molecule is formed as a byproduct [3]. This polypeptide chain shown in figure 1.6, excluding the side chains of the amino acids, is typically known as the **backbone** of the protein structure. As the amino acid side chains interact with each other and their environment, the backbone is folded into a three-dimensional structure. Figure 1.6 also shows the locations of all the backbone torsion (dihedral) angles, labelled ϕ_i and ψ_i for $i = 1$ to 5 amino acids as well as ω_j for $j = 1$ to 4 peptide bond linkages [29]. The peptide bond restricts the flexibility of the amino acid sequence, which can lead to steric hindrance, due to the relatively large size of some of the amino acids [29]. As a result of this, molecules can clash with each other while the protein is folding, limiting the scope of conformations the dihedral angles can adopt [29]. In addition to this, as mentioned in section 1.2.2, proline reduces the flexibility of the main chain, as its ring joins back onto the backbone of the protein, which restricts the ϕ torsion angle to $-60^\circ \pm 20^\circ$ [29]. Glycine only has a single hydrogen atom for its side chain, which means that it allows the dihedral angles to adopt a range of conformations that are not possible for the other amino acids, thus increasing the flexibility of the protein [29]. Overall, these factors impact how the amino acid sequence can fold.

Finally, disulfide bonds are another type of covalent bond that can help to stabilise proteins, and these can occur between two thiol groups, where a thiol group consists of a sulfur atom bonded to a hydrogen atom [3]. As shown in figure 1.4, cysteine has a thiol group, and the thiol groups between two cysteines can be connected, in order to form disulfide bonds (disulfide bridges) [3]. These bonds are the only covalent bonds found between nonadjacent amino acids in proteins [3]. They are post translational modifications, and they are much more frequent in eukaryotic organisms compared to prokaryotic organisms [45]. **Eukaryotic organisms** are organisms that have cells with a membrane-bound nucleus, such as mammals, fungi and and insects, whereas **prokaryotic organisms** do not have a nucleus or other membrane-bound organelles, and these organisms include bacteria and archaea [1]. As these disulfide bonds are formed after the protein has been synthesised, they tend to give stability to otherwise

properly folded proteins, and they can help to cross-link and stabilise different chains of a protein [29]. For example, they have been shown to be important for the stability of human IgG1 antibody CH3 domains [46].

In general, there are many factors that can influence the folding of a protein into its native state, including the hydrophobic effect, non-covalent interactions such as Van der Waals interactions, electrostatics and hydrogen bonds, and covalent interactions such as peptide and disulfide bonds. Usually a combination of these factors contribute to the stability of a protein, and a lot of these can be determined from the amino acid sequence.

1.2.4 Protein structures

In general, the structures of proteins can be described in four different levels: primary, secondary, tertiary and quaternary structures [3]. Figure 1.7 shows examples for these different types of structures. Firstly, the primary structure of a protein is given by the amino acid sequence, which is represented in figure 1.7 as a sequence of circles, with the one letter code given for each amino acid (table 1.1). This representation is a simplification of the amino acid sequence displayed in figure 1.6, which shows a molecular "ball and stick" representation of the amino acids and how they join together.

The secondary structure of a protein describes the repetitive conformations in the backbone of the folded protein [3]. The most common secondary structural conformation observed in proteins is the alpha helix [29], and figure 1.7 shows an example of a right handed alpha helix. The handedness of an alpha helix describes the way the helix twists [3], and in proteins, the right handed alpha helix is the most frequent. However, while left handed alpha helices are rarer, they can still be found in natural proteins [29; 48]. In addition to this, the alpha helix is very stable, which might explain why it is the most abundant in proteins [29]. Another common secondary structure observed in proteins is the beta sheet, and figure 1.7 shows an example of an anti-parallel beta sheet, where the N-terminus and C-terminus of adjacent strands run in opposite directions [3]. In contrast to this, a parallel beta sheet has the N-terminus and C-terminus of adjacent strands running in the same direction, and both of these types of beta sheet are abundant in proteins [3]. Additionally, for the alpha helix and beta sheets displayed in figure 1.7, we can see how the "ball and stick" representation of the amino acids, shown in figure 1.6, is folded into these secondary structures. Ramachandran noticed in 1972 that the dihedral angles of proteins can only adopt certain conformations due

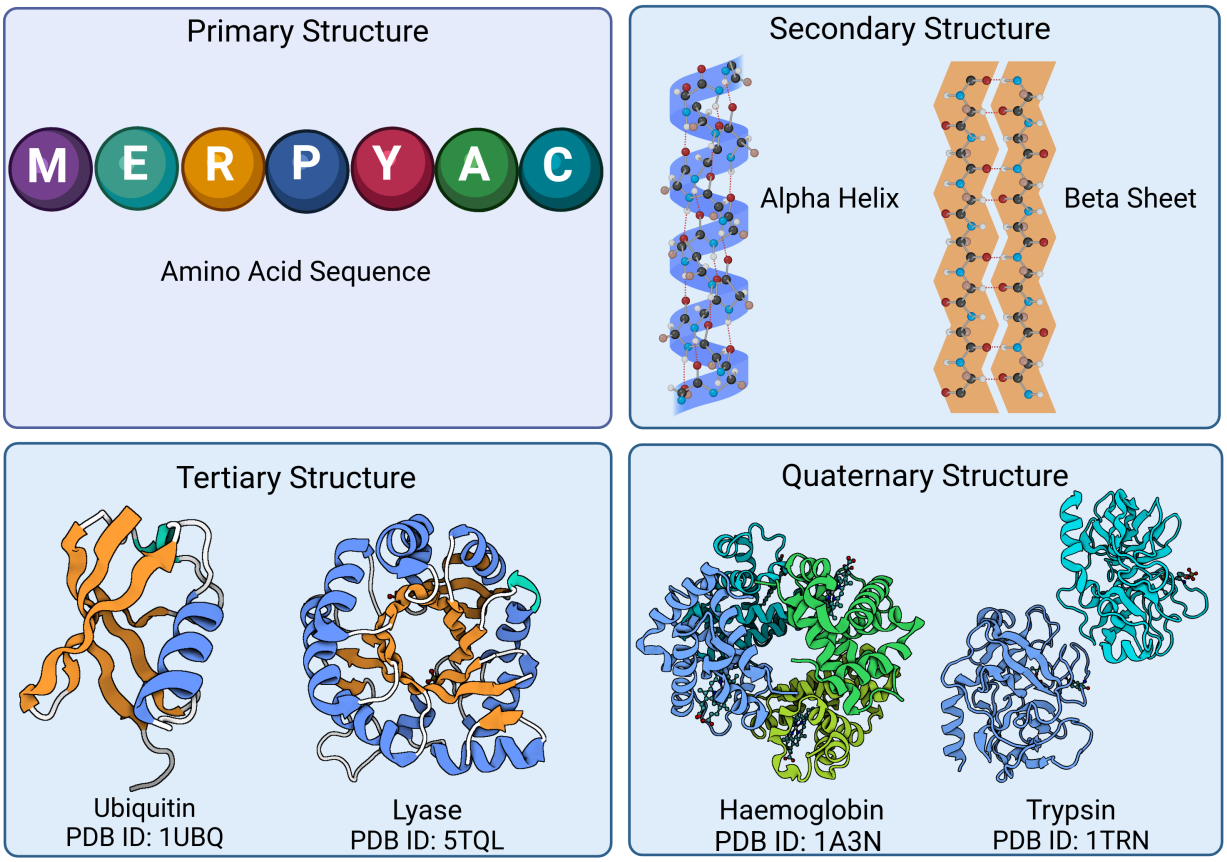

Figure 1.7: The different types of protein structure: primary, secondary, tertiary and quaternary. PDB IDs are included for the structures taken from the Protein Data Bank (PDB) [47]. Created with BioRender.com.

to steric hindrance, which was mentioned in section 1.2.3, and these conformations can be visualised as a **Ramachandran plot** [49]. Figure 1.8 shows a Ramachandran plot for the SARS-CoV-2 spike glycoprotein, with PDB ID 6VXX from the Protein Data Bank (PDB) [47]. In this plot, the ϕ and ψ dihedral angles are plotted against each other for each amino acid in the protein structure, which are represented by the yellow dots, and the blue regions on the plot represent the allowed conformations of these two dihedral angles [49]. From this plot, we can see the regions of space that correspond to different secondary structures, such as beta sheets and both left and right handed alpha helices, and the large majority of points lie in the blue regions of space.

Although alpha helices and beta sheets are the most common type of secondary structure in proteins, there are other types of secondary structure that exist in nature. For example, the 3_{10}-helix is a type of helix that has 3 residues per turn, and is much tighter than the alpha helix [29], which has 3.6 residues per turn [50]. The π-helix

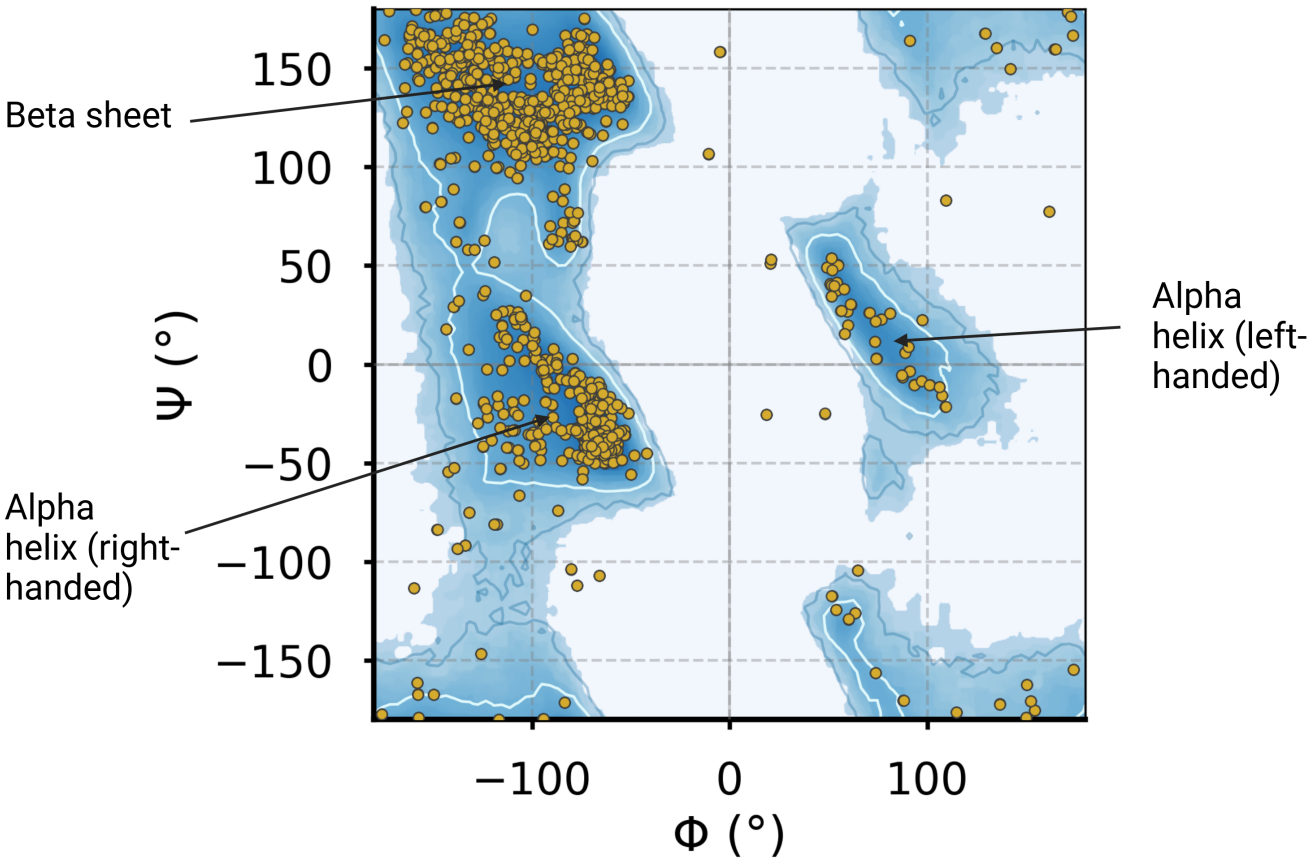

Figure 1.8: Ramachandran plot for the SARS-CoV-2 spike glycoprotein. The PDB ID for this structure is 6VXX from the Protein Data Bank (PDB) [47]. The Python package https://github.com/Joseph-Ellaway/Ramachandran_Plotter was used to create this Ramachandran plot.

has 4.4 residues per turn, and is slightly wider than an alpha helix [50]. Both of these helices are only found in short regions in a protein, and they are not as common as alpha helices [50]. However, the π-helix has been found in more proteins than was originally thought [51]. Despite there being various different types of secondary structure observed in proteins, less than half of a protein's backbone is arranged in a defined secondary structure [3]. The rest of the protein backbone is non repetitive and these regions are generally called random coil or loop conformations [3]. Many intrinsically disordered proteins, which are proteins that do not have a well-defined and stable three-dimensional structure [52], exist in nature. However, these proteins are much more common in eukaryotic organisms than prokaryotic organisms [53].

Following on from this, the tertiary structure of a protein is the full three-dimensional

arrangement of all the atoms in the protein [3]. These structures can be represented in the same "ball and stick" representation that was shown in figure 1.6, or they can be shown as a "cartoon" representation, which is a simplified representation of the structure based on its secondary structures. Figure 1.7 shows two examples of these cartoon representations for different structures, ubiquitin with PDB ID 1UBQ from the Protein Data Bank (PDB) [47], and a lyase with PDB ID 5TQL, with both of these representations coloured by secondary structure. Both of these structural representations consist of mainly alpha helices and beta sheets, in addition to random coils (loops), which are represented as the white strands. Additionally, these three-dimensional structures have compact regions called **folds** or **domains**, that are independently folding units of the protein, and they usually have some distinct function [54]. The structures shown in the tertiary structure section of figure 1.7 are two distinct types of protein domain: ubiquitin is its own domain and the lyase is an example of a TIM barrel domain [55]. Furthermore, most proteins in nature are actually made up of multiple domains [54], and there have been a lot of attempts to classify proteins based on their secondary structure and proteins domains [56; 57; 58].

Finally, the last level of protein structure is quaternary, which describes proteins that have more than one polypeptide chain [3]. The individual chains are called **subunits**, and the quaternary structure describes how these subunits are arranged together into oligomers [3]. Figure 1.7 shows two examples of quaternary structures: haemoglobin with PDB ID 1A3N and trypsin with PDB ID 1TRN, with both of these structures coloured by their different chains. From these structures, we can see that haemoglobin has four chains, which is called a tetramer; however, trypsin only has two chains, which is known as a dimer [3]. In addition to this, protein quaternary structures are closely related to protein-protein interactions, and can be crucial in understanding the function of many proteins [59].

1.2.5 Protein sequence and structure databases

As the amino acid sequence of proteins and the three-dimensional structure of proteins, are extremely useful for inferring the function of proteins, a large amount of amino acid sequence and protein structure data has been collected over the last few decades [60; 47]. UniProt is a protein sequence database which currently has around 250 million protein sequences, along with functional annotations [60]. This large amount of protein sequence data has been possible due to incredible advancements in high-

throughput DNA and RNA sequencing methods [61], as around 99% of the sequences in UniProt are either predicted from DNA and RNA sequences, or inferred from homology [60]. On the other hand, the Protein Data Bank (PDB) has around 215,000 experimentally-determined protein structures [47], that have been found using methods such as X-ray crystallography [62], Nuclear Magnetic Resonance (NMR) [63] and Cryogenic Electron Microscopy (CryoEM) [64]. In addition to this, due to the recent advancements in protein structure prediction, with AlphaFold2 [65] and ESMFold [66], there are now databases with large amounts of highly accurate predicted protein structures [67; 66; 68].

1.2.6 Protein structure prediction

Although there are experimental methods, such as X-ray crystallography, that can be used to determine the tertiary/quaternary structures of proteins, these methods are extremely slow and laborious [69]. This can be very limiting for studies concerned with the function or structure of proteins, including areas such as antibody/drug design. Protein structure prediction methods aim to use computational techniques to predict the tertiary/quaternary structures of proteins from their amino acid sequences (structural models), as depicted in figure 1.9.

Protein structure prediction methods can largely be grouped into two main categories: template based methods and free modeling or Ab-Initio methods [70; 71]. Ab-Initio protein structure prediction methods, such as, ROSETTA [72] and FRAGFOLD [73], are based on the "thermodynamic hypothesis" described in section 1.2.3, which assumes that the native protein structure has the lowest Gibbs free energy [74]. These methods begin by generating a number of possible structural models, sometimes by using fragments of structures from the PDB, and then an energy scoring function to estimate the Gibbs free energy, in order to rank these conformations [70; 74]. In addition to this, physico-chemical properties of the structures, such as hydrogen bonding, contact potential energy, PDB-derived secondary structure proportions, and interactions involved in folding, are taken into account when evaluating possible structural models [70].

On the other hand, template based methods use structures of known proteins, that have similar sequences to the target sequence, in order to predict structures [75]. Homology modelling is a specific category of template based methods and assumes that two amino acid sequences that are highly similar have similar structures [70], and

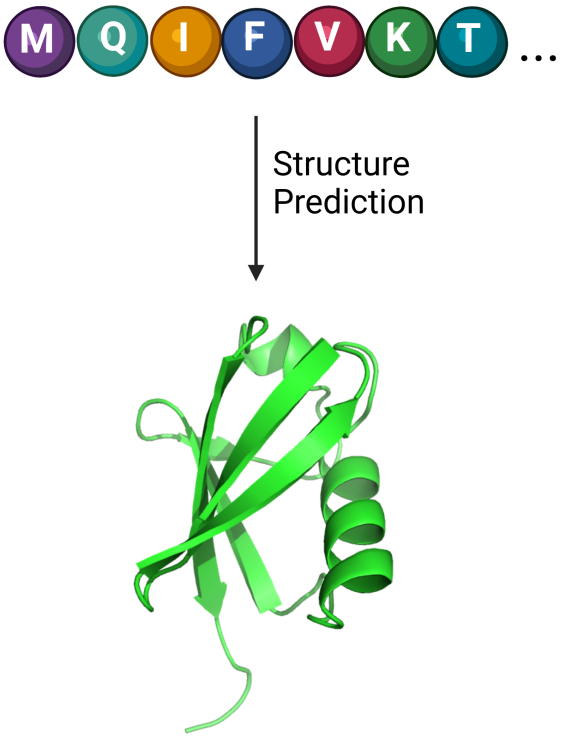

Figure 1.9: Protein structure prediction methods predict the tertiary/quaternary structure of proteins from their amino acid sequences. The example structure is shown in its cartoon representation. Created with BioRender.com.

MODELLER is an example of a structure prediction method that uses homology modelling [76; 77]. These methods align a target amino acid sequence to a sequence where the structure is known, and then the similarities with the template are used to generate a structural model for the target sequence [70]. After this, energy minimisation is used to finalise the structural models, and then they are evaluated using a Ramachandran plot, as shown in figure 1.8 [70].

In 2018, DeepMind introduced AlphaFold (version1) at the 13th edition of CASP (Critical Assessment of Structure Prediction) which is a biannual competition for evaluating protein structure prediction methods, and AlphaFold won the Free Modeling category of the competition [78; 79]. AlphaFold built on previous protein structure prediction methods, and used a deep learning based method called a convolutional neural network [80] trained on structures from the PDB [78]. Neural networks are types of machine learning methods that use connected units called neurons, in order to process data and learn complex patterns [81]. These neurons are organised in layers, with each layer performing different transformations to the input data [81], and con-

volutional neural networks (CNN) are a specific type of neural network, that have had incredible success in areas such as computer vision and natural language processing [82]. In addition to this, deep learning refers to the use of neural networks with many different layers, and has become widely popular across different applications, due to the high levels of performance and the accessibility to large amounts of data [83].

The inputs to the CNN in AlphaFold, included features that were extracted from a multiple sequence alignment (MSA) of the target sequence, to other amino acid sequences that have an evolutionary relationship [78]. MSAs are methods in bioinformatics that are used to compare the similarity of a set of biological sequences [84]. After this, the MSA features were used to predict the distances between residues, as distograms [70; 78]. Next, these distograms were then used to predict backbone torsion distributions, and ultimately then used to obtain a structural model [70; 78]. Although AlphaFold won the Free Modeling category, it could only predict a small number of proteins, without a template, at experimental accuracy [71].

However, in 2020, AlphaFold2 [65] was entered into the 14th edition of CASP, and it achieved unprecedented accuracy in predicting protein structures, with $2/3$ of the predicted structures from AlphaFold2 being competitive to experimental accuracy [85]. AlphaFold2 uses a completely different architecture to AlphaFold, as it uses transformer based networks, which are a type of deep learning technique that use a concept called attention [86; 65]. Rather than handcrafted features from an MSA being used as an input, AlphaFold2 uses the raw MSA along with proteins with similar structures to the input sequence, that are used as a template (pair representations) [65; 70]. This information is then passed into a module called an EvoFormer, which has attention and non-attention based components, and allows information to be exchanged between the MSA and pair representations, in order to enable direct reasoning about the spatial and evolutionary relationships [65]. After this, a structure module takes the outputs from the Evoformer module, and also uses attention based modules, in order to predict a structure for the input sequence [65]. In addition to this, AlphaFold2 provides a confidence score called predicted local distance difference test (pLDDT) score, which ranges between 0 and 100, and provides a level of confidence around each residue in the structural model [65].

Although structure prediction methods are incredibly useful to researchers exploring the properties and functions of proteins, they do have some limitations that can restrict how they can be used. One major limitation is that they do not perform well for predicting the structure of disordered protein regions and for ligand binding, which

is mainly due to the lack of experimental data [71; 87]. Additionally, many structure prediction methods, such as AlphaFold2, are highly dependent on MSAs of related sequences in order to accurately predict structures, and their performance can drop considerably for orphan sequences which do not have any related sequences [70]. The MSA step is also very computationally expensive, especially for large proteins, which means that these methods can take a long time to run [70]. In order to address these issues, there have been structure prediction methods such as ESMFold [66] and OmegaFold [88] that have replaced the MSA step with language models, and can predict structures in a fraction of the time. However, these methods are currently not as accurate as AlphaFold2 [71]. Furthermore, protein structure prediction methods fail at predicting rare structural conformations which can be extremely important in health and disease, and they cannot predict the impact of post translational modifications [87].

Overall, AlphaFold2 has triggered a revolution in protein structure prediction, with many more methods being published afterwards that also achieve high levels of accuracy, such as, RosettaFold [89], ESMFold [66] and OmegaFold [88]. Despite the limitations of protein structure prediction, these methods have a had a huge impact on the field as they allow researchers to explore protein structures in a way that has never been possible before. This has helped to advance our understanding of the role of proteins in nature, and even design novel proteins for applications across a wide range of scientific areas [90; 70].

1.3 Nature has only sampled a tiny fraction of possible protein sequences

In section 1.2, it was discussed how there is an incredible variety of natural proteins, along with a diverse set of functions, properties and levels of structural complexity. However, despite all of this, nature has only sampled a tiny fraction of the possible proteins that could exist [91]. A lot of this section is covered in Huang et al. "The coming age of *de novo* protein design" review paper [91]. An estimate of the amount of unique protein sequences in nature is roughly 5×10^{10}, which is negligible compared to the vast space of possible sequences [92]. Across the 250 million protein sequences in UniProt, the average sequence length is 351 amino acids [60], and as there are 20 standard amino acids (table 1.4), this means there are 20^{351} possible amino acid sequences of this length. This is an incomprehensibly large number of possible protein

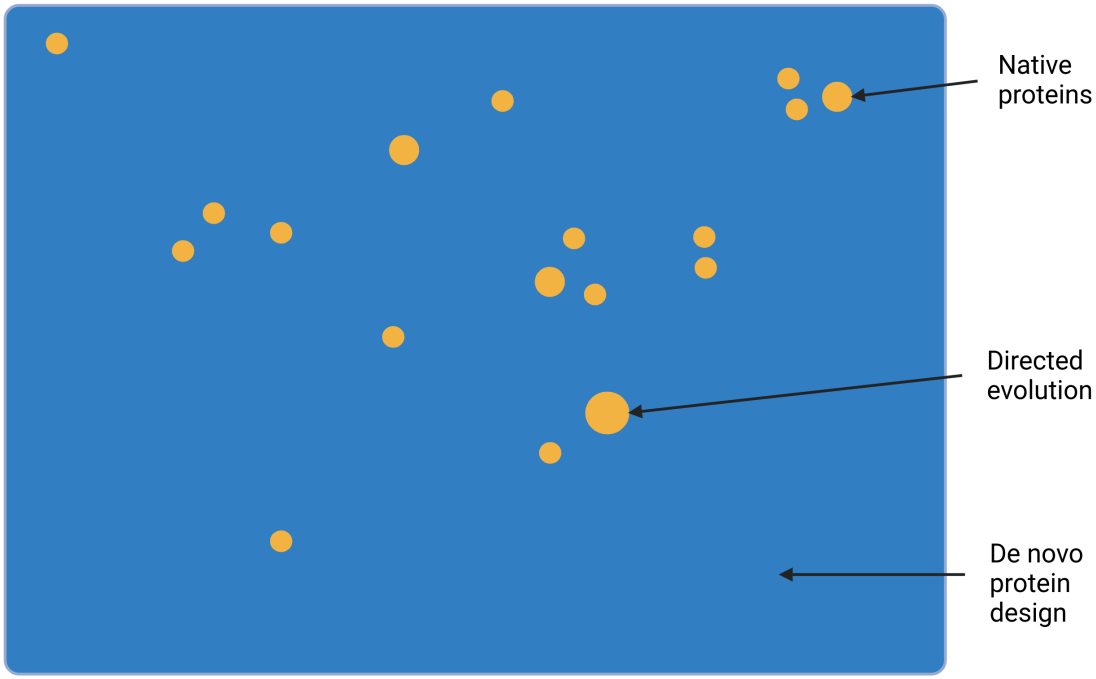

Figure 1.10: A representation of the sequence space that has been explored by nature. The blue space represents all the possible amino acid sequences, while the yellow circles represent native proteins. Directed evolution can only sample the sequence space around native proteins, whereas new regions of sequence space can be explored by *de novo* protein design. This figure has been adapted from figure 1a from Huang et al. [91]. Created with BioRender.com.

sequences, which is much larger than the 10^{24} estimate, for the number of stars in the observable universe from the European Space Agency [93]. Furthermore, there has been a lot of work on extending the genetic code and introducing non canonical amino acids, which can expand the range of functions and properties of proteins [94; 95]. This addition of non canonical amino acids, dramatically increases the size of this sequence space even more.

Figure 1.10 shows a representation of the amount of sequence space that has been explored by nature, and it was adapted from figure 1a from Huang et al. [91]. The blue space represents all the possible sequences with the 20 standard amino acids, and the yellow circles represent clusters of native proteins. Evolution has explored this sequence space by making small changes to existing proteins in nature [96], and from the size of the possible sequences that could exist, it is clear that all possible sequences that could have function have not been explored [96]. This is supported by the fact

that function has been observed from random protein sequence libraries, that are distinct from anything found in nature [97]. In addition to this, functional proteins are not uniformly distributed across sequence space, and are clustered into tight protein families [98; 99], which are represented by the yellow circles in figure 1.10. Protein sequences have also been shown to evolve at different rates, based on factors such as their expression levels [100], structure [101] and function [102]. Therefore, this suggests it is likely that large areas of sequence space have not been explored by evolution, and some of these undiscovered protein sequences will have novel properties and functions. In order to discover novel proteins with enhanced properties and functions, a lot of research has focused on modifying existing proteins, using methods such as directed evolution [103]. These methods have had huge success in improving the function and properties of different proteins, such as enzymes [104]. However, these methods are limited to explore regions of sequence space around native proteins, and cannot explore completely new areas [91]. On the other hand, *de novo* protein design aims to design novel proteins from first principles [105], in order to explore new regions of sequence space, and discover useful proteins for applications across scientific areas [91].

1.4 Protein design

Protein design is a broad field that encompasses many different techniques, which aim to create proteins with novel structures and functions [105; 106; 91; 107]. This can include rational protein design methods, which modify existing native proteins to improve or alter their function [108; 109], and *de novo* protein design techniques that design novel proteins from first principles [105; 106; 91; 107]. Within the field of *de novo* protein design, there are a lot of different sub categories, such as manual protein design [105], physics based computational protein design [110] and over the last few years, the advent of machine/deep learning based protein design [111; 112]. This section will provide a brief overview of the field of protein design, mainly focusing on *de novo* protein design methods, and starting with applications of designed proteins in different areas of science. After this, a section will describe computational protein design, along with the difference between sequence and backbone design. Next, physics based computational protein design methods will be detailed along with various examples of *de novo* proteins that have been designed using these methods. Finally, recent advancements in deep learning based protein design methods will be described, along with the impact these methods are having on the field.

1.4.1 Applications of designed proteins

In nature, proteins have a vast array of different functionality as materials, catalysts, signalling molecules and transport systems, which was discussed in detail in section 1.2.1. However, proteins also have applications outside of their natural context, with many examples in areas such as agriculture, biotechnology and medicine. In section 1.3, the immense size of protein sequence space was discussed, along with the tiny fraction that nature has sampled from this space. As a result of this, there is a huge opportunity to design proteins, with novel properties and functions, in order to solve problems across various scientific areas.

To begin with, protein design has a huge potential to help address challenges in agriculture. Currently, major challenges in agriculture involve feeding the growing global population in a sustainable way, and to ensure the security of food against the threat of climate change [113]. One way that protein design can help address these challenges, is by increasing the nutritional quality of proteins in plants, by replacing plant proteins with designed proteins, that have a more desirable amino acid composition [114]. For sustainability reasons, meat proteins are becoming more discouraged; however, plant proteins generally don't contain all the essential amino acids needed for humans [114]. Therefore, protein design can help to create more sustainable protein sources that are nutritionally complete [114]. Another way that protein design can help solve challenges in agriculture, is for the bio containment of genetically modified organisms (GMOs) [115]. GMOs offer an enormous opportunity to increase the yield and reduce the cost of crops, in order to ensure the security of food for the world population; however, there are concerns about unexpected consequences for the environment with the use of GMOs [116]. Essential enzymes in GMOs have been computationally redesigned, so that they are reliant on synthetic metabolites, which results in GMOs that are easier to isolate from natural ecosystems [115]. Furthermore, protein design could be used in the future to redesign photosynthesis, which is the biological process that plants, and some other organisms, perform to convert solar energy and carbon dioxide into oxygen and sugars [117]. Some of the enzymes involved in photosynthesis, such as rubisco, which was discussed in section 1.2.1, are extremely inefficient and wasteful [13]. For rubisco, this is mainly due to the fact it evolved before the great oxygenation event, and nowadays oxygen competes with carbon dioxide for binding, which impacts the efficiency of the enzyme [118]. Although redesigning photosynthesis is a very difficult task, protein design could help contribute to making

photosynthesis more efficient, which would improve the yield of crops and help meet the growing global demand for food [117].

Another area where protein design has a lot of potential is in biotechnology. In nature, proteins are key components of many sensory pathways, such as the Notch1 protein [20] which was mentioned in section 1.2.1, and this ability can be leveraged, to design proteins that are capable of detecting a range of different substances. Quijano-Rubio et al. used protein design to develop various biosensors, for the detection of molecules such as the HER2 receptor, Botulinum neurotoxin B, an anti-hepatitis B virus antibody and Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) spike protein [119], which are relevant for applications in medicine. These biosensors were able to detect these molecules with high sensitivity, which shows the potential of protein design for making clinically relevant diagnostic tools [119]. Additionally, protein design has been used to develop biosensors for the monitoring and visualisation of auxin, which is an important regulatory molecule in plants [120]. Auxin is critical for plant growth, and has been shown to regulate factors such as cell division, growth, differentiation and responses to light and gravity [121]. The developed biosensor allows real-time monitoring of auxin in plant cells, in order to understand how auxin concentrations vary throughout a plant life cycle [120]. Another application of protein design in biotechnology, is the *de novo* design of nanopores for single molecule detection [122]. Nanopores are a type of membrane protein which can capture and identify a single molecule of interest [123], and have been used for DNA sequencing, where they offer a cheap and rapid alternative to existing sequencing methods [124; 125]. *De novo* protein design can be used to develop nanopores with different pore sizes and chemical properties, so they can detect a wider variety of molecules [122]. One other application of protein design in biotechnology, is the design of a protein conjugate pair, that can fuse together by forming a rapid covalent bond [126]. The SpyCatcher/SpyTag system was adapted from a protein found in the bacteria *S. pyogenes*, which has a domain containing a spontaneous covalent bond [126]. After splitting this domain, the resulting protein fragments were rationally designed, to obtain a peptide (SpyTag) that forms a covalent bond to the protein partner (SpyCatcher) in minutes [126]. This system has applications in biotechnology, for example, developing biomaterials for tissue engineering and regeneration [127], and developing effective bacteriophage treatments for antibiotic resistant bacterial infections, by decorating bacteriophages with patient-specific proteins to suppress their immune response against the bacteriophage [128].

Medicine is one other area where designed proteins can be applied to solve a range

of different problems. In 2019/2020 the outbreak of SARS-CoV-2 caused worldwide panic, and a lot of research focused on ways to treat this novel respiratory illness [129]. One example of these treatments is the *de novo* design of SARS-CoV-2 miniprotein inhibitors, that could potentially be delivered into the nose, and provide prophylactic protection and therapeutic benefit for treatment of early infection [130]. These miniprotein inhibitors were designed to bind onto the spike protein of SARS-CoV-2, to prevent viral entry into cells [130]. This type of treatment could be really beneficial for individuals who come into frequent contact with infected people, such as healthcare workers [130]. Another application of protein design in medicine is for the development of vaccines. The *de novo* design of proteins can be used to mimic a viral epitope outside the context of the native protein, in order to induce neutralising antibodies [131; 132]. Once these proteins are designed, they could then be used to develop vaccines for these viruses [131; 132]. One other application of *de novo* protein design in medicine, is for the development of novel treatments for genetic diseases, by editing the DNA of a person, which is known as their genome [133; 134]. CRISPR-Cas9 is a bacterial immune system that can be repurposed for editing DNA, in order to treat genetic diseases [133; 134]. The discovery of CRISPR-Cas9 has revolutionised genome engineering and recently the UK became the first country to approve a therapy based on this system, for the treatment of sickle-cell disease and β-thalassaemia [135]. However, this system has some limitations and can suffer from off target effects, therefore protein design could be used to develop proteins that have greater precision, in order to develop more effective treatments [136].

All of these applications show that proteins have a huge potential to solve problems across a wide range of different scientific areas. Although, there has been incredible technological advancements over the last 100 years, machines developed by humans cannot compete with the precision of proteins at such a small scale, and they cannot be produced by self-assembly [137]. In addition to this, only a small percentage of the possible amino acid sequences have been explored by nature [137]. This means there is a vast number of protein sequences with interesting and useful properties, that have not been explored in the natural world [137; 106; 105].

1.4.2 Computational protein design

As a protein's function is largely determined by its structure [138], many computational protein design methods design towards a target structure or backbone [107].

This backbone can be a native protein structure from the PDB, excluding the side chains of the amino acids, or different methods can be used to design a *de novo* backbone [139]. After obtaining a target backbone, the next step is to design an amino acid sequence that will fold into the target backbone [139]. Therefore, computational protein design has two coupled problems, **backbone design** and **sequence design**. Figure 1.11 demonstrates this by showing an example of a backbone of a protein (PDB ID: 1UBQ), which is obtained using backbone design methods, and then sequence design methods are used to find the amino acid identities for a specific backbone.

A number of different methods can be used to design *de novo* backbones, in order to create novel proteins with various structures and functions. One method involves combining fragments of native protein structures from the PDB and different protein

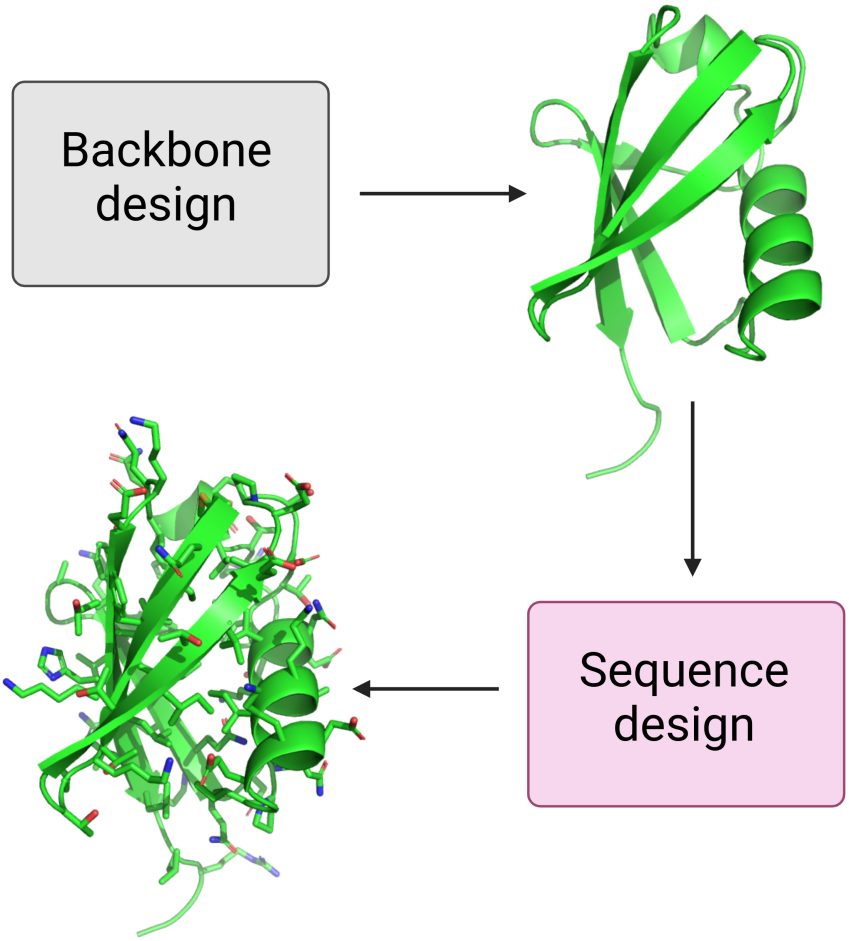

Figure 1.11: Backbone design creates a backbone or scaffold for a protein, while sequence design creates a sequence of amino acids for a given target backbone. Created with BioRender.com.

folds, in order to design new backbones [140; 141], while other methods uses parametric models to design new backbone structures [142; 143]. Furthermore, recently there has been a focus on using deep learning based methods, such as diffusion models, which have been widely used in computer vision and had a lot of success with generative modeling [144]. An example of a diffusion model for protein backbone design is RFDiffusion, which has shown high performance for applications such as, protein binder design and enzyme active site scaffolding [145].

Over the years, there has been a lot of different methods developed for sequence design as well. One method involves treating sequence design as an optimisation problem [146; 147; 148], where side chains are sampled for a template backbone, and a combi-

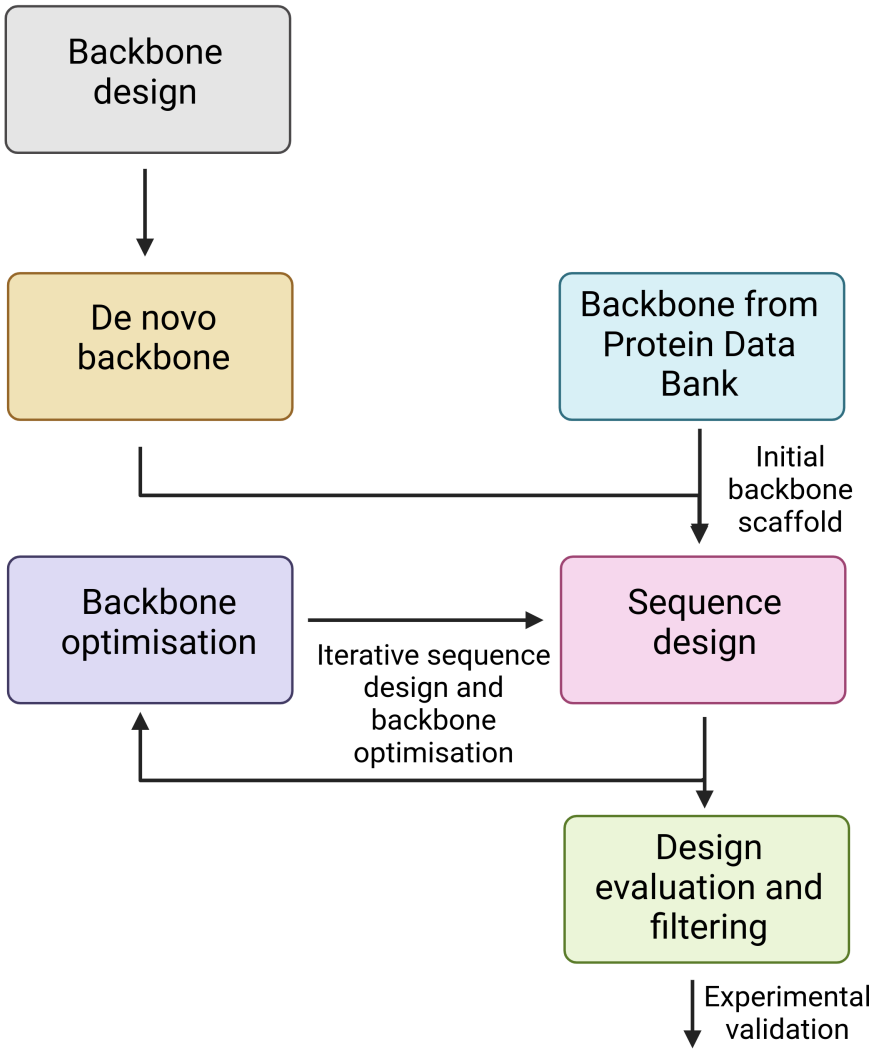

Figure 1.12: A general computational protein design pipeline. This figure was adapted from figure 1 in MacDonald et al. [139]. Created with BioRender.com.

nation of the side chains are found by minimising the energy of a structure, which is discussed further in section 1.4.3. Other methods allow for flexibility of the backbone during sequence design, and iterate between side chain and backbone optimisation [139; 146; 149]. Furthermore, recently deep learning methods have become very popular for sequence design as well. These methods train deep learning methods on large amounts of protein sequence and structure data, and then they can be used to sample amino acid sequences. There are many examples in the literature; however, a few examples include TIMED [150], ProteinMPNN [151], ProtGPT2 [152] and Frame2Seq [153], and these are discussed in more detail in section 1.4.4.

Figure 1.12 shows an general overview of a computational protein design pipeline. In general, computational protein design pipelines begin with template protein backbones from the PDB or *de novo* designed backbones, which provide the main structure of the protein [139]. After this, different side chains are sampled to design amino acid sequences for the template backbones, and design evaluation methods are used to rank these sequences [139]. Following on from this, some methods perform cycles of backbone optimisation along with the sequence optimisation, in order to greatly increase the sequence space that can be explored [139]. Finally, the sequences that are ranked the best are usually selected for experimental testing [146].

1.4.3 Physics based methods for protein design

Physics based protein design methods approximate the interactions between different molecules in protein sequences and their environment, in order to design proteins from first principles [110]. The approach of physics based design methods is based on Christian Anfinsen's hypothesis, that protein sequences fold into their lowest energy states [154], and was discussed in section 1.2.3. In order to do this, these methods use energy scoring functions, which calculate the energy of a protein sequence by approximating the contributions of different interactions to the stability of the protein [146]. These interactions include, but are not limited to, hydrogen bonds, electrostatics, Van der Waals interactions and disulfide bonds, which were discussed in section 1.2.3. Energy scoring functions are used in these design methods, to distinguish correct designs from incorrect designs, and these scoring functions are generally minimised to obtain sequences that are as stable as possible [146]. A lot of these energy scoring functions, use energy terms that are derived from statistics or parameterised from experimentally-determined protein structures or small molecule data [110; 155].

Stephen L. Mayo developed the first physics based computational design method with successful experimental validation in the 1990s [156]. This method was used to design a novel sequence for the ββα motif of a zinc finger protein fold, and a protein sequence alignment tool called BLAST [157] showed that this designed sequence had very low identity to any known sequence [156]. In order to do this, a scoring function, along with a search algorithm and stereochemical constraints, were used to find an optimal sequence, from a large library of possible amino acid sequences for a template backbone structure [156]. Additionally, NMR was used to obtain a structure for this designed sequence, and it was found to be consistent with the design target structure [156].

One of the most popular and well established physics based protein design methods is Rosetta, which was initially developed in the 1990s and has a whole suite of tools to tackle problems across structural biology [158]. The RosettaDesign software was used to design and validate Top 7, which was the first *de novo* protein fold [106; 140]. This method constructed a protein scaffold using fragments from the Protein Data Bank (PDB) [47], and then used a Monte Carlo search protocol and energy function, to design sequences for this backbone [140]. After this, the design method cycled between sequence design and backbone optimisation, with the backbone optimisation step identifying the lowest free energy backbone conformation for a fixed amino acid sequence [140]. Finally, the best designs were selected after these cycles of sequence design and backbone optimisation, and then experimentally validated in the lab [140]. Top7 was found to be highly stable and an experimentally-determined x-ray crystal structure of Top7 was extremely similar to the predicted structure, with a root mean square deviation (RMSD) equal to 1.2 Å [140].

Nowadays, Rosetta offers a whole range of different tools, including *de novo* protein structure prediction, protein-protein docking and protein-ligand docking [158]. Top7 was an incredible achievement in *de novo* protein design, and showed the potential of computational protein design in accurately designing proteins with novel structures. However, this protein was not designed for any particular function [140], and since this, Rosetta has been used to design a large amount of *de novo* proteins with different functions, such as, immunogens for the development of vaccines [131], antibodies [159], and different enzymes [160; 161; 162].

Although, the Rosetta software suite has had a huge impact on the field of protein design, it is not the only set of tools that can be used for protein design. There have been many different energy functions developed for estimating the energy of a protein

sequence, and guiding the protein design process. EvoEF2 is another energy scoring function that was developed for *de novo* protein design [163]. Similar to Rosetta, EvoEF2 is based on physico-chemical descriptors such as Van der Waals interactions and hydrogen bonding. However, EvoEF2 used sequence recapitulation for parameter optimisation [163], while Rosetta used small-molecule thermodynamic data, and high-resolution structural features to optimise the parameters [155]. EvoEF2 has been used to design SARS-CoV-2 spike proteins for vaccine design [164], and for the development of antibodies against the SARS-CoV-2 spike protein [165]. In addition to this, BUDE FF (Bristol University Docking Engine Force Field) [166; 167] is another energy function that can be used for *de novo* protein design. The BUDE FF models each atom, except for hydrogen, as a sphere with a specific radius and hydrophobicity or hydrophilicity potential, and the energy function's parameters were derived from experimental data [166; 167]. This energy function has been used to design a highly efficient and thermostable *de novo* enzyme [168], and for the design of water soluble α-helical barrels [169].

1.4.4 Machine learning based methods for protein design

Although physics based protein design methods have been incredibly successful, and there are many examples of functional proteins designed with these methods (section 1.4.3), recently there has been a major shift towards machine/deep learning approaches for protein design. The major reasons for this are the large amount of protein sequence and structural data that is now available through the PDB [47] and Uniprot [60], the incredible advancements in deep learning methods over the last few years, and deep learning based methods offer increased performance at a lower computational cost than physics based methods as well. In addition to this, AlphaFold2 [65] achieved unprecedented success in protein structure prediction at CASP14, as discussed in section 1.2.6, and this contributed to the increase in interest in the field of protein design, from the machine learning community. The general computational pipeline for these methods is still fairly similar to figure 1.12; however, the methods that are used for backbone and sequence design are being replaced by deep learning methods [111].

One of the major advancements in deep learning that has had a massive impact on the field of protein design, is the development of the transformer deep learning architecture [86]. Neural networks and deep learning were briefly introduced in section 1.2.6, and transformers are a specific type of deep learning technique that use a mecha-

nism called self-attention, in order to learn contextual relationships between positions in sequences [86]. These methods have had a huge impact on the field of Natural Language Processing (NLP), with large language models being built using a form of transformer architecture, such as ChatGPT from OpenAI, and becoming widely popular with millions of people around the world [170; 171].

As amino acid sequences can be represented as character strings, where each amino acid is represented by their one letter code (sections 1.2.2 and 1.2.4), language models have also been widely applied to protein sequence data. One example of a language model for protein design is ProtGPT2, which used a transformer-based architecture to rapidly generate protein sequences [152]. The protein sequences generated from ProtGPT2 can be conditioned towards a specific protein family, fold or function, and these sequences were shown to be globular and distantly related to native protein sequences [152]. Another example of a language model applied to protein design is ProGen, which was able to generate *de novo* lysozymes that had similar catalytic activity to native lysozymes, with some of the designs having sequence identities as low as 31.4% [172]. MetaAI also developed a protein sequence language model (ESM-1b) that was trained only on amino acid sequence data, and was shown to contain information about biological properties, such as, secondary structure, contacts and biological activity [173]. In addition to this, a later version of this language model (ESM-2) was used for protein structure prediction [66]. This language model replaced the MSA step that was used in protein structure methods such as AlphaFold2, which significantly decreased the amount of time it took for prediction, and this method still achieved relatively high performance [66]. Furthermore, the language model ESM-2 was used for designing *de novo* proteins as well [174].

Although language models have been shown to have a huge potential for protein design, and can be applied to a diverse set of problems, they are not the only type of deep learning method used for protein sequence design. ProteinMPNN uses a different type of neural network architecture called a Message Passing Neural Network (MPNN), which encodes a template backbone structure as a graph, and predicts proteins in an autoregressive manner from the N-terminus to C-terminus [151]. This means that it predicts each residue in the sequence one by one, and takes into account the previously predicted residues [151]. On the other hand, the protein sequence design method TIMED takes a protein backbone as input, and splits it into voxels, which are small cubes of space around each amino acid position in the backbone [150]. After this, a Convolutional Neural Network (CNN) is then used to predict the identity of each

amino acid [150]. In addition to this, Microsoft Research developed a diffusion model, called EvoDiff [175], for protein sequence design, where diffusion models are another type of neural network architecture that have shown considerable success in computer vision [144]. Diffusion models gradually corrupt the input data with Gaussian noise, and then a model is trained to reverse this process [144]. The resulting model learns the underlying probability distribution of the input data, and then can be used to generate new samples [144]. Microsoft Research trained EvoDiff on 45 million protein sequences and it was shown to generate diverse sequences, that cover natural sequence and functional space [175].

Diffusion models have also been applied to protein backbone design as well as sequence design. One example is RFDiffusion which fine tuned the RosettaFold structure prediction method on protein structure denoising tasks, in order to obtain a generative model that could sample new protein backbones [89; 145]. RFDiffusion was experimentally validated by designing *de novo* proteins to bind onto influenza haemagglutinin, and then CryoEM showed that the designed binder in complex with influenza haemagglutinin, was nearly identical to the design model [145]. Another example of a diffusion model for protein backbone design is ProteinSGM which is trained to generate four matrices that describe a protein's backbone [176]. After this sequence design is performed with fixed backbone Rosetta FastDesign [177], and circular dichroism spectroscopy was used to validate some proteins experimentally, which showed similar secondary structure compositions between the designs and experimental data [176]. Furthermore, diffusion models have also been developed for joint backbone and sequence design, such as Chroma, which was built by Generate Biomedicines [178]. Chroma can be conditioned to steer the generated proteins towards specific structures, properties and functions, and two crystal structures were obtained that were highly consistent with their design models [178].

1.5 Limitations of protein design methods

Rapid advances in protein design methods over the last few years, have allowed the design of increasingly more complex proteins and have provided protein designers with more control over the design process. Despite these rapid advances and the wide range of tools available, there are still a lot of limitations that make protein design inaccessible for a lot of researchers. Addressing some of these limitations would help to improve the reliability and efficiency of these methods, resulting in more wide spread

use of protein design techniques across different scientific areas.

To begin with, one major limitation of protein design methods, is that designed proteins still suffer from high failure rates [91; 146]. There are many different ways that protein designs can fail, such as, low expression, misfolding, aggregation and lack of function [91]. This means that generally a large number of computational designs need to be generated, and subsequently a lot of sequences tested in the lab, in order to find proteins that can be successfully expressed, with the correct structure and function. Pan and Kortemme calculated success rates for a number of different *de novo* protein design studies and presented the results in a plot, which is shown in figure 1.13 [146]. The size of each circle represents the amount of designs and the success rates show the proportion of designs in a particular study that, folded correctly, had their experimental structure solved, were functional, and were functional and helical [146]. From figure 1.13, we can see that, although there are some studies that achieve high proportions of folded designs, there is a huge amount of variation across studies, and there are many studies that have low numbers of correctly folded designs. The

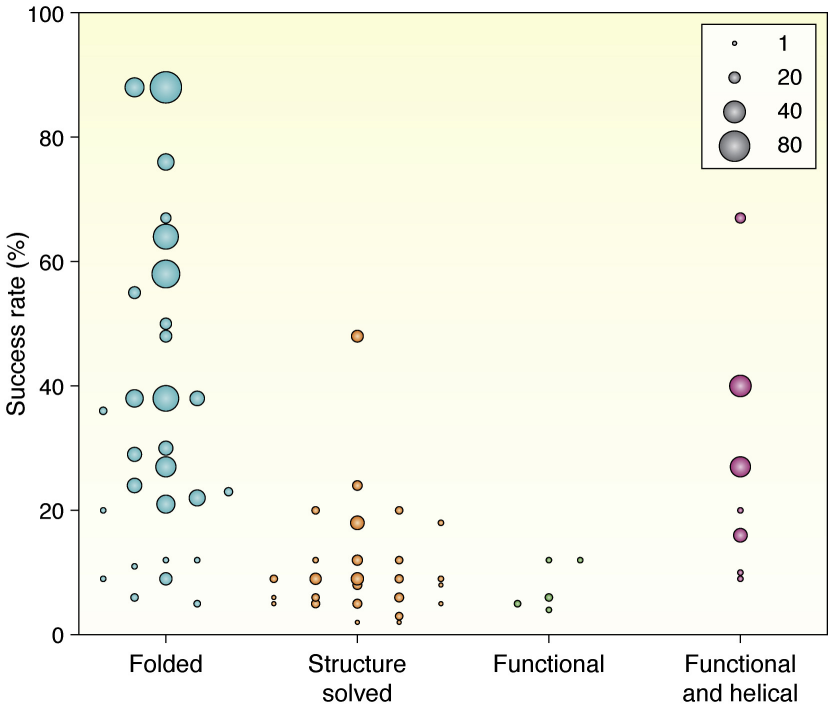

Figure 1.13: Success rates by different protein design studies, where the success rate is the proportion of protein designs that either, were correctly folded, had their experimental structure solved, were functional, and were functional and helical. In addition to this, the size of the circles indicates the amount of designs. This figure was taken from Figure 7 in Pan and Kortemme [146], which has a creative commons CC-BY licence.

majority of studies achieve low success rates for solving experimental structures for these designs, and have even lower success rates for achieving the desired function. However, Pan and Kortemme make a distinction with helical designed proteins as the success rates are higher than other proteins, due to robust design methodologies for helical bundles [146]. Furthermore, the low success rate of designs is not helped by the lack of published data including negative results, such as, designs that have failed due to low expression, misfolding, aggregation or lack of function. This type of data is equally valuable as data on successful designs, as it could help protein designers avoid common reasons for failure. In general, these results show that protein design methodologies are still very unreliable and have low success rates for designing functional *de novo* proteins. As a result of this, protein design can be very expensive and inefficient, which limits it as a tool that can be used by researchers.

Another major limitation of protein design methods, is that they still require a large amount of expertise in order to use for designing *de novo* proteins. The majority of computational tools for protein design, including both physics and deep learning based methods, are only accessible through a Command Line Interface (CLI) or through Python packages [145; 151; 152; 158; 163; 176; 178]. This means that these computational protein design methods are not very accessible and are limited to people who have a computational background. Although, there are some methods that are available through easy to use Google Colab notebooks now, such as RFDiffusion [145], and TIMED has a user friendly web server for sequence design [150]. By making protein design tools more user friendly so that they require less computational expertise in order to use them, a larger number of labs around the world will be able to use these methods and help advance the field.

One other limitation of current protein design methods, is that it is difficult to design proteins with well defined functions. Proteins can be designed to have certain properties, such as binding to a specific target motif however, the full behaviour and dynamics of the designed protein is usually not well understood [106]. One reason for this is that the "function" of a protein is usually ambiguous, as it is common for native proteins to have multiple functions [179]. Another reason is that a lot of protein functions are difficult to measure quantitatively, and for the functions that can be measured, such as binding affinity [180; 181] and enzyme activity [182; 183], there can be a lot of variation across measurements from different labs. There are sets of functional annotations for proteins, such as, Enzyme Commission (EC) numbers [184] and the Gene Ontology (GO) [185]; however, these categories can be fairly broad and

do not include any quantitative measurements for protein function, which can provide insights into more granular details, for example, the relative strength of a binding interaction and the rate of catalysis of an enzyme. From figure 1.13, we can see that protein designers are generally better at designing towards structures and folds rather than designing towards function, which could be the result of having large amounts of quality protein structure data and a lack of quantitative functional data. In addition to this, many methods design towards a target structure, and design methods could potentially benefit from having more information about the target properties and function in the requirements of their designs, in order to improve the success rate of designing towards specific functions. Furthermore, another way to improve the success rate of *de novo* functional proteins, could be to collect more high quality data sets with quantitative measurements of protein functions, which could be leveraged by deep learning based methods for designing functional proteins.

Finally, one more limitation of protein design is that the majority of designed proteins are produced in cell-based systems such as *E. coli* [91; 186]. These organisms are extremely versatile and have been able to produce a large variety of different proteins [187; 188; 189]. However, it is unreasonable to assume that these organisms could produce all the possible *de novo* proteins that we want to produce in the future. It could be the case that, by using these organisms for producing *de novo* proteins, we are limiting ourselves to specific areas of sequence space (figure 1.10). This is because designed proteins could be toxic to these organisms, or designs may need a completely different environment in order to fold. Cell-free systems are alternative methods that can produce proteins in a controlled environment outside of cells [190; 191; 192; 193], and these methods are discussed in detail in chapter 4, along with their advantages and disadvantages over cell-based systems. These systems offer increased control over the production of proteins, as the expression environment can be directly modified in order to tailor it for a specific protein [194], and these systems do not have the restraint of sustaining life, so they can be used for producing proteins that are toxic to cells [193]. On the other hand, these systems still use proteins extracted from organisms such as *E. coli*, in order to perform transcription and translation [191; 195; 196]. In the future, we may need completely *de novo* expression systems, with *de novo* designed proteins for transcription and translation, and tailored expression environments, in order to produce *de novo* proteins in far off areas of sequence space. If this were possible, we may start to think about designing custom expression systems to produce a target protein, rather than designing proteins that can be produced by a specific expression system.

1.6 The contribution of this work to the field of protein design

As discussed in this chapter, proteins in nature have an incredible array of different functions, including but not limited to, materials, catalysts, signalling molecules, and transport systems. Despite the incredible diversity in protein functions, nature has only sampled a tiny fraction of the possible proteins that could exist. Protein design aims to explore new regions of sequence space, in order to design novel proteins, that can solve problems across different scientific areas, such as, medicine, agriculture and biotechnology. The field of protein design has made huge advancements recently, with many deep learning based methods being developed, providing more control over the design process, and facilitating the design of increasingly more complex proteins. However, there are some major limitations of current methods that limit the potential of protein design. These limitations include the high failure rate of designed proteins, the difficulty of designing towards functions, and the high levels of expertise that is currently needed, in order to use computational protein design tools.

The work completed in this PhD project focused on developing methods for evaluating designed proteins, in order to help address some of these limitations of protein design. In the computational protein design pipeline in figure 1.12, this work fits into the design evaluation and filtering step, which is the last step before protein designs are tested experimentally. There are three major research outputs from this PhD project, which contribute to helping address the high failure rate of designed proteins, the difficulty with designing towards properties and functions, and improving the accessibility of computational tools for protein design.

Firstly, a user friendly web server called DE-STRESS was developed, which calculates a range of different physico-chemical properties for protein designs, in order to computationally evaluate designed proteins before testing them in the lab. The DE-STRESS webserver can be accessed at https://pragmaticproteindesign.bio.ed.ac.uk/de-stress/ and it was published in Protein Engineering, Design and Selection (M. J. Stam and C. W. Wood [197]). DE-STRESS offers reference set and specifications functionality, which respectively, allows the users to compare the properties of their designs against a set of known proteins, and allows them to filter designs based on certain requirements for the properties. This functionality in DE-STRESS aims to make it easier for users to design towards specific properties and functions, while in general, DE-STRESS aims to help reduce the failure rate of designed proteins, and intends to

be accessible and easy to use for a wide range of users. The work completed to develop DE-STRESS is discussed in detail in chapter 2, along with two protein re-design projects, where DE-STRESS was used to evaluate designs, which were generated from a sequence design method called TIMED [150] developed in our lab.

Secondly, analysis was performed that demonstrated that the physico-chemical properties from DE-STRESS are predictive of *in vivo* properties of proteins, such as protein production and host organism, and this work is presented in a pre-print on BioRxiv (M. J. Stam, D. A. Oyarzún, N. Laohakunakorn & C. W. Wood [198]). Predicting *in vivo* protein production could be useful for ranking designs for experimental validation, as designs predicted to be low producing could be excluded from testing in the lab. In addition to this, systematic variation was discovered in these physico-chemical properties across large amounts of predicted protein structures from 48 organisms, to such an extent that the tree of life could be reconstructed. This is the first time to our knowledge, that this has been demonstrated, and suggests that properties of proteins might have evolved to their unique molecular environment, which could be an important factor to consider when designing novel proteins. Both of these results are discussed in chapter 3 and they could help with addressing the high failure rate of designed proteins.

Finally, initial experimental protocols were developed for measuring protein production levels of designed proteins in *E. coli* cell-free systems, using fluorescent protein labelling [199]. In the future, this protocol will be used for high-throughput experiments on a library of protein designs, in order obtain protein production data. This data will be used to explore the relationship between the DE-STRESS metrics and protein production even further, in order to identify ways that we can filter out flawed designs before testing them in the lab. Furthermore, these cell-free systems could help with understanding some of the reasons why designed proteins fail, with the aim of avoiding these reasons in the future, and incorporating these insights into design methods. The completed experimental work is described in chapter 4, along with the next steps that could be taken, in order to help address the high failure rate of protein designs.

Chapter 2

DE-STRESS: High quality structural features for evaluating designs

This chapter introduces DEsigned STRucture Evaluation ServiceS (DE-STRESS), which is a user-friendly web server for evaluating structural models of designed and engineered proteins. DE-STRESS calculates a set of high quality structural features, capturing various physico-chemical properties of protein structures, to allow protein designers to rank designs before taking them into the lab. Firstly, some background and a literature review will be provided on computational methods for evaluating protein design candidates. After this, a detailed description of the DE-STRESS web server and the metrics it provides for evaluating protein designs will be given. In addition to this, a headless version of this software was developed, which can be ran from the Command Line Interface (CLI), enabling these metrics to be calculated at scale. Following on from this, the DE-STRESS metrics are shown to distinguish between native structures and folding decoys, which are models of proteins that have atomic coordinates very close to the native structures, suggesting these metrics are useful for ranking designs. Finally, the role of DE-STRESS in re-designing two proteins, a protease and a rubisco sub-unit, will be described, which outlines an example of how DE-STRESS could be incorporated into protein design pipelines in the future.

2.1 Background and motivation

The field of *de novo* protein design has a huge potential to develop novel proteins, in order to solve problems across a variety of scientific areas, such as agriculture [114; 115; 117], medicine [130; 131; 132; 136] and biotechnology [119; 120; 122; 126]

(section 1.4.1). Unfortunately, as described in section 1.5, protein design has a few major limitations, such as, the high failure rate of designs, the difficulty designing towards properties and functions, and the majority of design methods are not that accessible to non experts. A lot of computational methods have focused on scoring the folding and stability of the protein, calculating geometric properties of the structure and assessing solubility and aggregation propensity, in order to rank designs for experimental testing. Although these methods have been incredibly useful in the field of protein design over the years, the failure rate remains high [91; 200; 146]. Currently, there is a huge opportunity to leverage the growing amount of protein sequence and structural data [60; 47], in order to build new and improved data-driven evaluation techniques, that improve the reliability and success rate of designs.

2.1.1 Energy scoring functions

One of the major reasons that contributes to the failure of a protein design, is the instability of the structure and misfolding [91]. Energy scoring functions have been developed to approximate the Gibb's free energy of the folded state of a protein sequence, based on Christian Anfinsen's "thermodynamic hypothesis", that protein sequences fold into their lowest energy state [154]. This was discussed in detail in section 1.2.3. These functions capture the physical interactions that govern how proteins fold and interact with other molecules, and they play a central role in discriminating correct from incorrect designs in physics based protein design algorithms [146], which were explored in section 1.4.3. Protein sequences that have a lower energy score are favoured over higher energy scores, as these are predicted to be more native-like and stable.

Physics based energy scoring functions model the physico-chemical interactions between the atoms in the protein structure and its environment, in order to model how it will fold [110]. The majority of these energy functions are linear combinations of energy terms, which are mathematical models of the physics that control protein folding and stability [201]. Some of these energy terms include: Van der Waals and electrostatic interactions, solvation, hydrogen and disulfide bonds, backbone and side-chain torsional preferences and reference energies for different amino acids [201; 163; 166; 167; 202; 203]. A number of these different energy scoring functions were discussed in section 1.4.3, including, Rosetta [201], EvoEF2 [163] and BUDE [166; 167], and they have been used to design a range of different *de novo* proteins [131; 160; 164; 165; 168; 169]. In addition to this, there are other physics based

energy functions which are mainly used for molecular dynamic simulations [202; 203], but have also been used for protein design [204].

Knowledge based statistical potentials/energy functions are derived from databases of known proteins, in contrast to modelling the interactions between atoms in physics based energy functions [205]. These methods capture properties such as amino acid side chain orientations, solvation energy, hydrogen bonding, electrostatics potentials and more from analysing natural protein structures [205]. Before the Rosetta energy function moved to physics based, there was a version that was knowledge based [206], derived from the Protein Data Bank (PDB) [47]. Since this, there have been many more knowledge based statistical potentials developed such as DFIRE [207; 208], GOAP [209], CUPSAT [210] and MAESTRO [211].

Recently, there has been a shift towards developing energy functions that are deep learning/neural network based [212; 213; 214; 215; 216]. Rather than defining a linear combination of energy terms, which measure different properties of proteins, these methods train different neural networks to learn representations of protein sequences, which can then be used to predict stability. In addition to this, some studies have started to use root mean square deviation (RMSD) of the AlphaFold2 [65] structural models, against the intended target structure, along with the predicted local distance difference test (pLDDT), to rank designs. This approach has been shown to be effective for ranking β barrel protein designs [217].

2.1.2 Solubility and aggregation propensity

Another major reason why protein designs fail in the lab, is that the proteins are insoluble and aggregate when being produced [91], which is why protein solubility is another vital metric for evaluating designs. There are various computational tools which evaluate solubility and aggregation propensity, based on either the protein sequence or structure, in order to allow users to rank sequences before testing them experimentally [218; 219; 220; 221; 222; 223; 224]. A lot of these methods focus on solubility in *E. coli* as this is one of the most widely used organisms for expressing proteins [91].

AGGRESCAN [218] and PASTA [220; 221] are both sequence-based, solubility prediction methods and compare short sequence stretches, in order to find aggregation prone regions. AGGRESCAN was developed by using an amino acid aggregation-propensity scale, which was derived from *in vivo* experiments [225] and on the assumption that short and specific sequence stretches control protein aggregation [218].

The method uses the experimental data and calculates the average aggregation propensity for regions of the sequence, by using a sliding window to identify "hot spots" for aggregation. On the other hand, PASTA uses a pairwise energy function based on the propensities of two residues to be found within a beta sheet, facing one another on neighbouring strands [220; 221]. AGGRESCAN3D builds on AGGRESCAN by considering structural information, such as amino acid exposure to solvent and distance to other residues, along with the intrinsic aggregation propensity for each amino acid [219; 226].

In addition to computational methods that identify aggregation prone segments in protein sequences, there are also a lot of methods that predict the overall solubility of the protein [222; 224]. Protein-Sol predicts solubility of protein sequences in *E. coli* but was trained on data from cell-free expression systems. In addition to this, Protein-Sol uses a linear model, which combines 35 different features extracted from the protein sequence, that includes amino acid compositions along with other sequence features such as charge, length and β-strand propensity [222]. On the other hand, NetSolP predicts solubility and usability for protein purification of a protein sequence in *E. coli*, by using a deep learning transformer model [224; 227]. Furthermore, 5-fold cross validation results showed that NetSolP outperformed existing techniques and seemed to generalise better [224].

2.1.3 Geometric methods

Another way to evaluate protein designs before testing them in the lab is by evaluating geometric properties of the protein structure. This has become increasingly easier recently, due to the development of highly accurate structure prediction methods such as AlphaFold2 [65] and ESMFold [66]. Over the years, there has been a lot of methods developed to analyse cavities and pores in a protein structure, including CASTp [228; 229; 230], CICLOP [231], and CAVER [232; 233; 234]. CASTp is an intuitive web sever that has been well used for identifying voids and surface pockets, for a protein structure of interest. This method uses techniques from computational geometry, to find and characterise these voids and pockets, with the total area, volume and details of the atoms that are involved in their formation [228]. Similarly, CICLOP was developed to find inner cavities of protein structures however, it voxelises the structure and uses search algorithms to find empty nodes, in order to find these cavities [231]. This method provides similar information to CASTp however, it also provides extra

physico-chemical properties of these cavities, such as their hydrophobicity. In addition to this, CAVER uses geometry based methods to identify tunnels, pores and channels in protein structures as well [233], and can even be applied to large molecular dynamics simulations [232]. CAVER also provides a user friendly web server [232] and can be used in PyMol [235] and VMD [236], which allows users to visualise these tunnels, pores and channels. All three of these methods can be used for functional annotation of proteins, as cavities, tunnels and pores can be important for binding and enzyme function. Although, additionally, these metrics are important for understanding whether *de novo* proteins can be produced *in vivo*, as large cavities have been associated with low protein production for some proteins [159]. Furthermore, packing density [237] and hydrophobic fitness [238] are two other geometric metrics, that can be used to evaluate the packing quality of protein structures. The packing density of an non hydrogen atom, is defined as the number of non-hydrogen atoms within a specified radius, for example a radius of 7Å[237], and the hydrophobic fitness of a protein calculates the packing quality, by taking into account the hydrophobicity of residues [238]. Efficient implementations of these algorithms are available in the ISAMBARD Python package [239], which can be ran rapidly for a set of protein structures to assess their packing quality.

2.1.4 How can we make protein design more reliable and accessible?

In order to address the low success rate of protein designs, there is a need for a set of high quality metrics, which can be used to evaluate designs during the design process. There are methods that exist to rank structural models [240; 241; 242; 243]; however, some of these methods require a large number of models to be generated, or a large number of features from other tools, in order to be accurate [244]. However, ProteinTools is an example of an easy to use, intuitive web server that can be used to analyse protein structures, and calculates a number of different factors of proteins, such as, hydrophobic clusters, hydrogen-bond networks, salt bridges and contact maps [245]. Energy functions, geometric descriptors, and aggregation propensity/solubility measures, are examples of the types of metrics that are generally used in the protein design process. However, these metrics are typically only available as command line tools, which can be very restrictive for non technical users, and usually only a few of these metrics are included in the design process. As these metrics capture different

properties of structures, protein designers could potentially benefit from using a large number of these metrics in their design process. If the metrics were available in a user friendly tool, then this could encourage wider use across the protein design community. In addition to this, comparing a set of high quality metrics for a target structure, against a set of structures with similar target properties and functions, could help with the difficulty of designing towards specific properties and functions. By addressing these limitations, protein design would become more efficient, reliable and accessible to research labs.

This chapter introduces DEsigned STRucture Evaluation ServiceS (DE-STRESS) [197], which is a user-friendly web server for evaluating structural models of designed and engineered proteins. The web server aims to provide the user with as much information as possible about their designs, before taking them into the lab to characterise experimentally. A glossary of the structural features that are included in the DE-STRESS web server is shown in appendix A, along with a description of what each metric is calculating. Novel functionality of reference sets and requirement specifications, allow users to compare their designs against other sets of structures, and filter designs for those that meet certain conditions. Overall, DE-STRESS aims to address some of these limitations of protein design, so that protein design can become more widely used as a methodology.

2.2 Methods

This section will provide a detailed description of DEsigned STRucture Evaluation ServiceS (DE-STRESS), including an overview of the software architecture for the web server, along with an outline of the various physico-chemical metrics that are calculated for an input structure. These metrics consist of energy scoring functions, geometric properties, non-covalent interactions, aggregation propensities and other basic properties, such as amino acid and secondary structure composition. In addition to this, the reference sets and specification functionality will be explored, which allow users to compare the metrics for a specific design against a set of reference proteins, and filter the designs for those that fit user-defined requirements. After this, a description of the headless version of the DE-STRESS software will be provided, which can be executed from the Command Line Interface (CLI) and uses multiprocessing so that it can be ran across a large number of protein structures. Finally, the methodology for how DE-STRESS was used to distinguish between native and decoy structures, and for

how it was used to help re-design a protease and the small sub-unit of rubisco, is also discussed.

2.2.1 DE-STRESS web server

The DE-STRESS web server is composed of a simple and intuitive interface, written in Elm and JavaScript, and a backend web stack written in Python, consisting of Gunicorn, Flask, GraphQL and PostgreSQL. The web server was containerised using Docker to make it easier for people to host their own instances of the web server. Although DE-STRESS does not store any data from the user, people may prefer to host local instances of DE-STRESS to ensure the privacy of their designs. The diagram in figure 2.1 provides an overview of the web server architecture. Users can upload structural models to the web server in a PDB file format [246], which is a well established file format in the field. It contains the 3D coordinates of the atoms in the structure and also includes information about the chemistry of the protein, details about data collection, structure refinement and structural descriptors [246].

2.2.1.1 DE-STRESS metrics

After a user has uploaded a structure to the DE-STRESS web server, an analysis module is run on the PDB file, in order to generate a number of different metrics from the structure. Various external software packages are used in DE-STRESS to calculate these metrics. Firstly, basic information about the design is generated from the ISAMBARD Python package [247], such as amino acid composition, isoelectric point and implementations of geometric descriptors such as packing density [248] and hydrophobic fitness [238]. Secondly, the energy scoring functions BUDE FF [166; 167], DFIRE2 [207], EvoEF2 [163] and Rosetta [155] are calculated for designs that are uploaded to the web server. These functions approximate the energy of a protein sequence and capture a lot of information about the interactions involved in protein folding. Finally, Aggrescan3D [219] is used to calculate an aggregation propensity score for the structure. Protein designers generally want to avoid structures that have a high probability of aggregating, as this can cause loss of function and other undesired effects. A detailed glossary of the different metrics, how they are implemented and the conventions for use, is included in the DE-STRESS web server and also in appendix A.

Once the metrics have been calculated for a PDB file, the results are displayed on

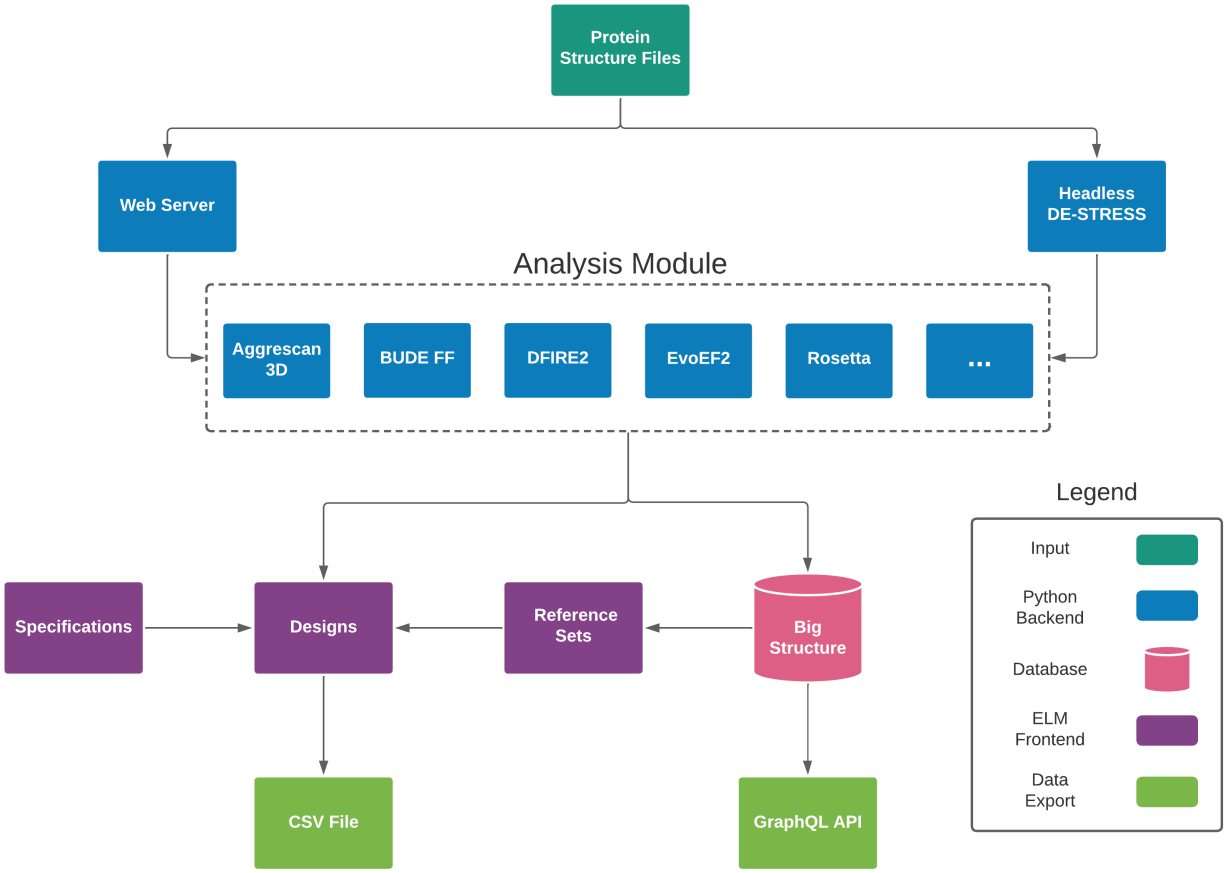

Figure 2.1: Overview of the DE-STRESS software architecture.

the designs page on the web server. Users can click on the individual designs to see detailed information for a structural model, or they can compare the uploaded structures with the overview plots. Visualisations of the structural models are shown for each design through the use of the NGL JavaScript library [249; 250], and secondary structure assignment is also displayed using DSSP [251; 252]. This page provides users with a wealth of information regarding their design, which would have taken a lot of effort and time to implement themselves. Also, a lot of the software included in DE-STRESS use command line tools or Python libraries, which makes them inaccessible to non technical users. Therefore, DE-STRESS provides an intuitive and responsive web server for people to access this information.

The data generated for uploaded designs can be exported as a CSV file and downloaded from the web server. This allows users to explore the designs further, and train machine learning models with the downloaded data. Using this functionality, protein designs can be uploaded to DE-STRESS, metrics will then be generated for the de-

signs, and then they can be exported as a CSV file to create a data set of features for the designs.

2.2.1.2 Reference sets

In order to contextualise the values of the structural features, novel functionality offered by DE-STRESS allows users to compare their designs against a reference set of known protein structures. Reference sets can be used to compare the physico-chemical properties of designed proteins, against the properties of proteins which possess desirable traits. This may include proteins with a target fold, favourable properties such as high solubility and expressibility, or a desired function, for example, binding to a particular target or catalysing a reaction. In addition to this, reference sets provide a way to validate whether designed proteins have improved properties compared to a set of known proteins, such as, predicted stability or solubility. The main aim of providing this functionality in DE-STRESS, is to help users design towards particular properties and functions, and to ultimately reduce the failure rate of designed proteins.

The reference sets page on the web server, provides users the choice of two predefined sets of high quality crystal structures, and the ability to define custom sets of structures by specifying the PDB IDs. The reference sets are generated from the Big Structure database, which contains pre-calculated DE-STRESS metrics for 82,010 structures from the PDB. The rest of the structures on the PDB were excluded for reasons such as not containing protein, having formatting errors in the PDB file, or the files were too large to run through DE-STRESS in a reasonable time frame. Big Structure was created using the database generation script, which passes a list of PDB structures through the analysis module and creates the PostgreSQL database. This database can be queried through a graphQL API. Functionality was also added to DE-STRESS to create reference sets from uploaded designs, in case users want to use structures that have not been published yet. In the future, reference sets could be expanded to allow users to compare against predicted protein structures from the AlphaFoldDB [67] and the ESM Metagenomic Atlas [66] as well.

2.2.1.3 Requirement specifications

One other feature that has been added to DE-STRESS is the ability to set requirement specifications. This functionality allows users to define the intent of their protein design. On the Specifications page users can define complex rules with boolean

operators, in order to capture the properties required from their design, and automatically filter their designs by this requirement. Currently, these rules involve setting constraints on the values of the DE-STRESS metrics, such as energy scoring function values. These rules could be informed by using reference sets of structures with similar properties that are required by the designed protein. For example, if users are designing a new antibody structure, rules could be set based on the values of packing density, a combination of energy values and aggregation propensity of a set of antibody structures. Requirement specifications will be expanded in the future to capture different types of requirements, and to help users design towards particular properties and functions.

2.2.1.4 Availability

DE-STRESS is available for non-commercial use, without registration, through the following url: https://pragmaticproteindesign.bio.ed.ac.uk/de-stress/. The source code is available in the git repository https://github.com/wells-wood-research/de-stress, and the data that was used to generate the reference sets is available through a graphQL API, with the following url https://pragmaticproteindesign.bio.ed.ac.uk/big-structure/graphql.

2.2.2 Headless DE-STRESS

Although the DE-STRESS web server has been developed to be responsive and easy to use, it has a few restrictions which are in place to ensure the stability of the web server. These restrictions include: only proteins of 500 residues or less can be uploaded, only 30 files at one time can be uploaded, and there is a maximum computational run time of 20 seconds to calculate all the DE-STRESS metrics. However, a headless version of DE-STRESS was developed, which does not have any of these limitations. Headless DE-STRESS can be ran locally from the CLI and uses multiprocessing to calculate the metrics rapidly, for a given input path of PDB files. The user can adjust the settings for headless DE-STRESS, by updating variables in a .env file, which controls the computational time (seconds) that the DE-STRESS metrics are allowed to run, how many PDB files are in a batch, and how many processors should be used. The full set of input PDB files can be split up into batches to avoid running out of memory during run time. In addition to this, Docker is used to run the analysis module container shown in figure 2.1, without the Elm front end. This functionality, allows users to calculate the DE-STRESS metrics for much larger protein structures, and for a greater number of

structures. In this PhD project, headless DE-STRESS was used to calculate physicochemical properties for 180,000 experimentally-determined structures from the Protein Data Bank (PDB) [47] and 560,000 structures from the AlphaFold DB [67]. This work is discussed in detail in chapter 3.

2.2.3 Decoy Analysis

This section describes the decoy analysis, which was performed to show how the DE-STRESS metrics can be used to distinguish between a set of native or experimentally-determined protein structures, and their folding decoys. Firstly, a data set was constructed using 3DRobot [253] and the DE-STRESS metrics were then calculated for these structures. After this, the features were scaled and preprocessed, and principal component analysis (PCA) was performed on the DE-STRESS metrics. The code for this analysis is available at https://github.com/wells-wood-research/stam-m-wood-c-de-stress-2021.

2.2.3.1 Creating the Data Set

The data set used for the decoy analysis was created using the 3DRobot_set generated by 3DRobot [253]. A sample of 9 experimentally-determined structures, along with 360 randomly sampled decoys (40 per structure), were selected. The decoy structures in the 3DRobot_set have a root mean square deviation (RMSD), ranging from 0 to 12$\mathring{A}$ from the experimentally-determined structure. RMSD is used to compare the atomic coordinates between different protein structures therefore, the random sampling was done to ensure a spread across this range.

Additional crystal structures for the 9 experimentally-determined structures were downloaded from the PDB and added to the data set. These additional crystallographic structures that were included in the analysis were found using the "Find similar assemblies" search functionality on the PDB. Table 2.1 shows the PDB IDs and chains for the structures that were selected from the PDB, and the experimentally determined structures from 3DRobot_set. Once the structures had been selected, the DE-STRESS web server was used to generate metrics and the resulting data was downloaded as CSV files. Finally, these CSV files were joined together to obtain the full data set.

2.2.3.2 Feature Selection and Scaling

Firstly, before performing principal component analysis (PCA), a number of data pre-processing steps were performed. As the energy scoring function values are dependent on the size of the protein, they were divided by the number of residues in the structure. After this, a number of variables were removed from the data set. Features such as amino acid composition, mass, num_residues and isoelectric_point were constant for structures from the same PDB ID. This analysis is concerned with features that can distinguish between the decoys and experimentally-determined structures, so that's why these metrics were excluded. Other metrics that were excluded were the categorical features design name, pdb_id and structure group, evoef2_interD total and rosetta_yhh_planarity were excluded as they were constant across the data set, and the decoy or native flag was also excluded. Two other metrics evoef2_ref_total and rosetta_dslf_fa13 were excluded because they were discrete variables and aggrescan3d_total_value was excluded as we already have aggrescan3d_avg_value in the data set which is normalised for the size of the structure.

After excluding these metrics, the remaining features were scaled so that the values were between 0 and 1. This is an important step because PCA is extremely sensitive to the magnitude of the values, and higher magnitude features could skew the analysis.

Experimentally-determined structure	**Additional structures**
1N8V A	1N8U A
1ZI8 A	1ZIC A, 1ZIX A, 1ZI9 A
2HS1 A	2HS2 A, 3S53 A
2XOD A	2X2P A
3CHB D	1PZK D, 1PZJ D
3LDC A	3OUS A, 3R65 A, 3LDD A
3NJN A	3NJH A, 3NJM A
3WCQ A	3AB5 A
3WDC A	3WDE A, 3WDD A

Table 2.1: PDB IDs and chain IDs for the experimentally-determined structures included in this analysis, along with the additional crystallographic structures that were downloaded from the PDB.

The features were scaled with min-max scaling rather than using standardisation, as some of the features were not normally distributed.

2.2.3.3 Principal component analysis

Principal component analysis (PCA) [254] was used to explore and visualise how the DE-STRESS metrics varied across the scaled DE-STRESS metrics of the 9 native structures, the additional experimentally-determined structures from the PDB, and the 360 decoy structures from 3DRobot. PCA was performed with 10 different components and the variance explained by each component was calculated. In addition to this, scatter plots of the top two principal components were plotted and split out by the 9 different native structures, and then the top contributors to each principal component were found. PCA was chosen for this analysis because it is a simple method with only one hyperparameter (number of principal components), which makes it less likely to overfit to small data sets. Furthermore, PCA is explainable as it provides the relative feature contributions to each principal component, which is not true for other methods such as Uniform Manifold Approximation and Projection (UMAP) [255].

2.2.4 Protein re-design projects

This section describes the work completed for redesigning the Tobacco Etch Virus (TEV) protease and the small sub-unit of ribulose-1,5-bisphosphate carboxylase/oxygenase (rubisco) enzyme. To begin with, some background will be provided encompassing these two proteins, along with motivation for why we're interested in re-designing them. After this, the re-design of the rubisco sub-unit will be detailed first, followed by the re-design of the TEV protease.

For both of these design projects, a number of different sequences were generated for the target structures, using the sequence design method TIMED-design [150]. In order to do this, the empty backbones were extracted from the target structures (by removing all side-chain information), and they were voxelised into discrete areas of space, called frames, using aposteriori [150]. Following on from this, TIMED-design takes these discretised backbones as an input, and predicts the side chain identities based on the structure. Various different versions of TIMED were used for this re-design work, including TIMED, TIMED Deep, TIMED Rotamer, TIMED Rotamer Deep, TIMED Charge and TIMED Polar. TIMED and TIMED Deep were trained to predict amino acid identities but have different neural network architectures, TIMED

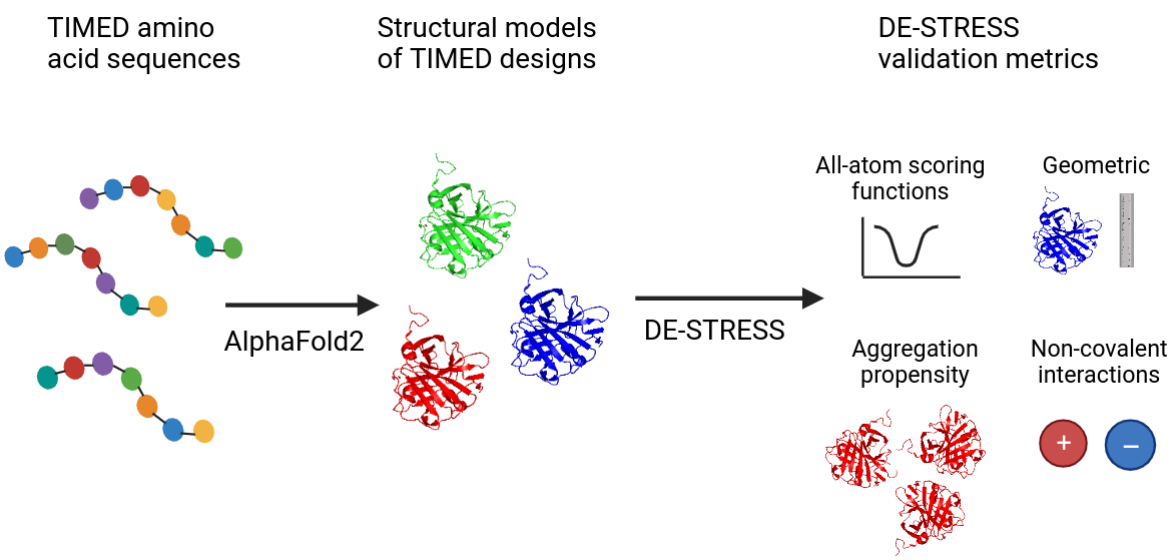

Figure 2.2: Computational pipeline for evaluating TIMED sequences using AlphaFold2 structural models and the DE-STRESS web server.

Rotamer and TIMED Rotamer Deep are similar but instead of amino acid identity, they were trained to predict the identity and orientation (rotamer identities), and finally TIMED Charge and TIMED Polar incorporate more information about the charge and hydrophobicity of residues during training. Once all the sequences had been designed, AlphaFold2 [65] was used to generate structural models for these sequences, and DE-STRESS was used to evaluate these designs against experimentally-determined structures. Finally, the top ranked sequences based on these DE-STRESS metrics, were selected and sent to collaborators for experimental validation. Figure 2.2 shows an illustration of this computational pipeline for the TEV protease designs. The sequence design and AlphaFold2 part of this work was completed by Leonardo Castorina in the Wells Wood lab, while the rest of the work in evaluating the designs with the DE-STRESS web server was completed by myself. In addition to this, for the rubisco small sub-unit designs, Jack O'Shea performed molecular simulations on the top ranked designs, to understand whether they would assemble with the main rubisco complex.

2.2.4.1 Background and motivation

Ribulose-1,5-bisphosphate carboxylase/oxygenase (rubisco) is one of the most abundant enzymes in the world, with it being estimated that there is roughly 5kg of rubisco, for every person on earth [256]. On top of that, rubisco is incredibly important, as it plays a vital role in carbon dioxide fixation during photosynthesis [13], where carbon

dioxide and water are converted into oxygen and carbohydrates by light energy. Despite all of this, rubisco is very inefficient and slow, with large amounts of the enzyme needed to maintain reasonable rates of photosynthesis [257], which has made it a target for protein design, with the aim of increasing the yield of crops [258] and securing climate and food security [259]. Most forms of rubisco in plants, algae, cyanobacteria and proteobacteria are hexadecameric with 8 large sub-units and 8 small sub-units [260], and the small sub-unit has been shown to increase the specificity and carboxylation efficiency of rubisco [261]. In this project, the small sub-unit of the *Arabidopsis thaliana* rubisco was chosen for re-design because of this purpose, and we worked with Dr Manajit Hayer-Hartl, a collaborator in at the Max Planck Institute of Biochemistry, for experimentally validating these *de novo* designs in *E. coli*, following the same approach as outlined in Aigner et al. [262]. Figure 2.3 shows the hexadecameric structure of *Arabidopsis thaliana* rubisco (PDB ID: 5iu0) and the asymmetric unit of the same structure, with the small sub-units shown in magenta.

Another design target that DE-STRESS was applied to, was the Tobacco Etch Virus (TEV) protease, which is a plant virus and is widely used for various applications in biotechnology [263]. One of these major applications is to remove affinity tags that are added in order to express and purify proteins [264], as these tags can interfere with the function and activity of the proteins. Therefore, the TEV protease is extremely useful to remove these tags after purification, before using the purified protein for any applications [264]. In addition to this, the TEV protease can be used to develop split biosensor assays, that can allow the monitoring of protein-protein interactions [265], and to study membrane receptor activation [266]. One issue with the TEV protease is that it is quite difficult to produce due to poor solubility and it is prone to aggregation [267]. These reasons make the TEV protease an interesting design candidate, and in this project we designed different TEV protease sequences, and collaborated with Professor Lynne Regan at the University of Edinburgh to collect experimental data. Figure 2.4 shows the AlphaFold2 structural models of the TEV protease with PDB ID 1LVB, and additional proteins that are highly related, such as the tobacco vein mottling virus (TVMV) protease, with PDB ID 3MMG.

2.2.4.2 Rubisco small sub-unit redesign

This section describes the work completed for re-designing the small sub-unit of ribulose-1,5-bisphosphate carboxylase/oxygenase (rubisco). Firstly, the small rubsico sub-unit structure was taken from an experimentally-determined structure of *Arabidopsis thaliana*

rubisco, with PDB ID 5iu0 [268]. The asymmetric unit in this PDB file has two of these small sub-units in different places in the structure, with chain IDs 5iu0I and 5iu0J. In addition to this, the full rubisco structure has eight copies of the small sub-unit, with chain IDs 5iu0C, 5iu0H, 5iu0L, 5iu0M, 5iu0O, 5iu0W, 5iu0X and 5iu0Y. All of the different TIMED models were used to generate sequences for these 10 rubisco small sub-unit structures. As the output of TIMED provides a probability distribution over the different amino acid identities, the amino acid with the highest probability was chosen at each position. Following this procedure, a total of 60 sequences were designed, one for each combination of target structure and TIMED model.

Once all the sequences had been designed, AlphaFold2 was used to generate 5 structural models for each sequence. Headless DE-STRESS was then ran to calculate physico-chemical properties for all of these structural models (300 in total), as well as the AlphaFold2 structural models for the original target structures. The reason for this was to ensure the comparison between the designs and the reference small sub-unit structures, was not influenced by any inherent variance between crystal structures and Alphafold2 structural models. Following on from this, the data set of DE-STRESS features was processed to remove metrics only capturing sequence information, those with low variance or constant across the data set, and highly correlated features. Prin-

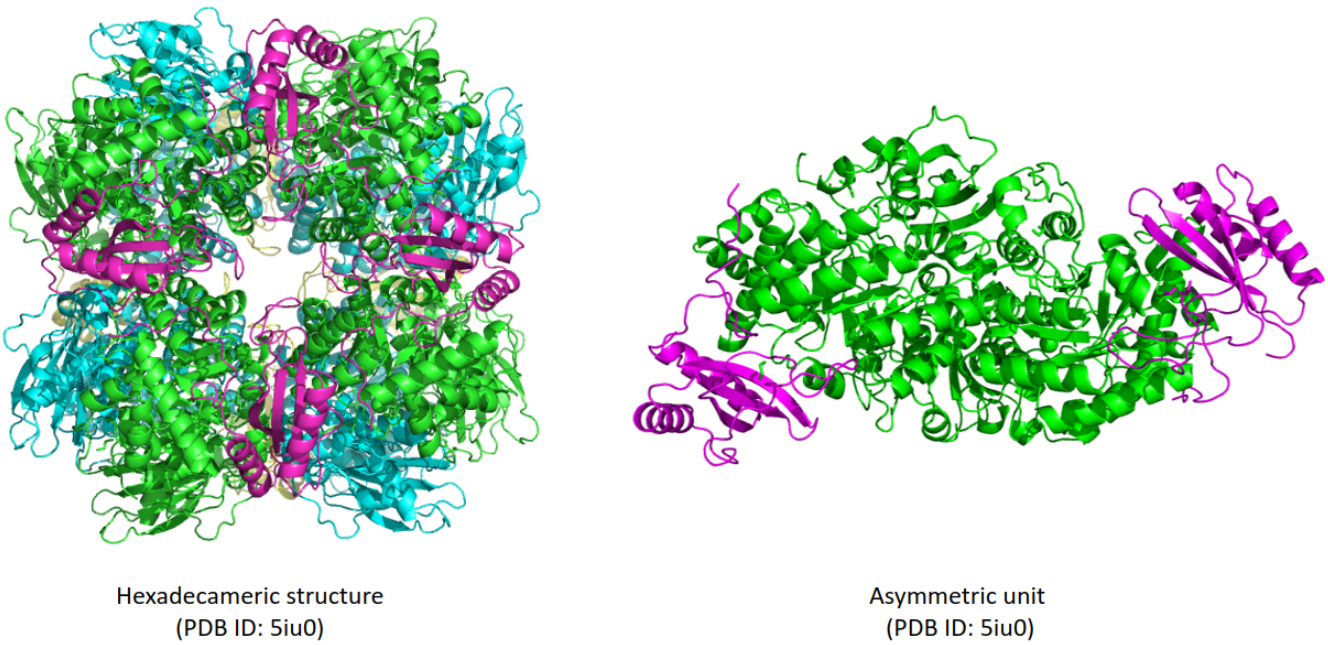

Figure 2.3: The hexadecameric and asymmetric unit of the *Arabidopsis thaliana* rubisco with PDB ID 5iu0. The small sub-units are shown in magenta in both structures.

cipal Component Analysis (PCA) was applied to the processed data set of over 300 structural models and their physico-chemical features, and the top two principal components were plotted on a scatter plot. The average principal component values were computed for each of the 60 designs and then the euclidean distance was calculated between each design and the reference structures, across 10 principal components, in order to rank the designs. In addition to this, molecular dynamics simulations were performed for the top ranked small sub-unit designs, to investigate whether they were likely to assemble with the main rubisco complex. BUDE Alanine Scan (BAlaS) [269] was ran on different frames from the molecular dynamics simulations to explore the interaction energy between the designed rubisco small sub-units and the reference large sub-unit. This analysis confirmed that the top ranked designs had similar BAlaS predicted binding to the reference structure. Finally, the top ranked sequences were sent to Dr Manajit Hayer-Hartl, a collaborator at the Max Planck Institute of Biochemistry, for experimental validation in the lab.

2.2.4.3 Protease redesign

Similarly, for the protease re-design project, the same pipeline was followed as outlined in the rubisco small sub-unit work in section 2.2.4.2. Although for the protease re-design, there were two target structures that were used to design sequences, instead of ten. These were the experimentally-determined structures for the Tobacco Etch Virus (TEV) protease (PDB ID 1LVMA) and the tobacco vein mottling virus (TVMV) protease (PDB ID 3MMGA). Figure 2.4 shows AlphaFold2 structural models of these proteases, along with the structural models of additional proteases that are highly related, including the turnip mosaic virus (TuMV_Q, TuMV_J) and the plum pox virus (PPV), which were not used as design targets for TIMED, but were used as reference structures for comparison. In addition to this, we performed this design pipeline twice, once for completely *de novo* sequences, and another time where residues that were conserved across all proteases, including TEV, TVMV, TuMV_Q, TuMV_J and PPV, were fixed, and only the other residues were re-designed with TIMED. In a similar fashion to the rubisco small sub-unit re-design work, all of the TIMED models were used to generate sequences for these two different backbones, TEV and TVMV, either with the conserved residues or completely *de novo*. This resulted in a total of 24 designed protease sequences, that were processed in the same way as the rubisco small sub-unit sequences, with Alphafold2 structural models being generated, the DESTRESS metrics calculated, and then PCA was used to evaluate the sequences against

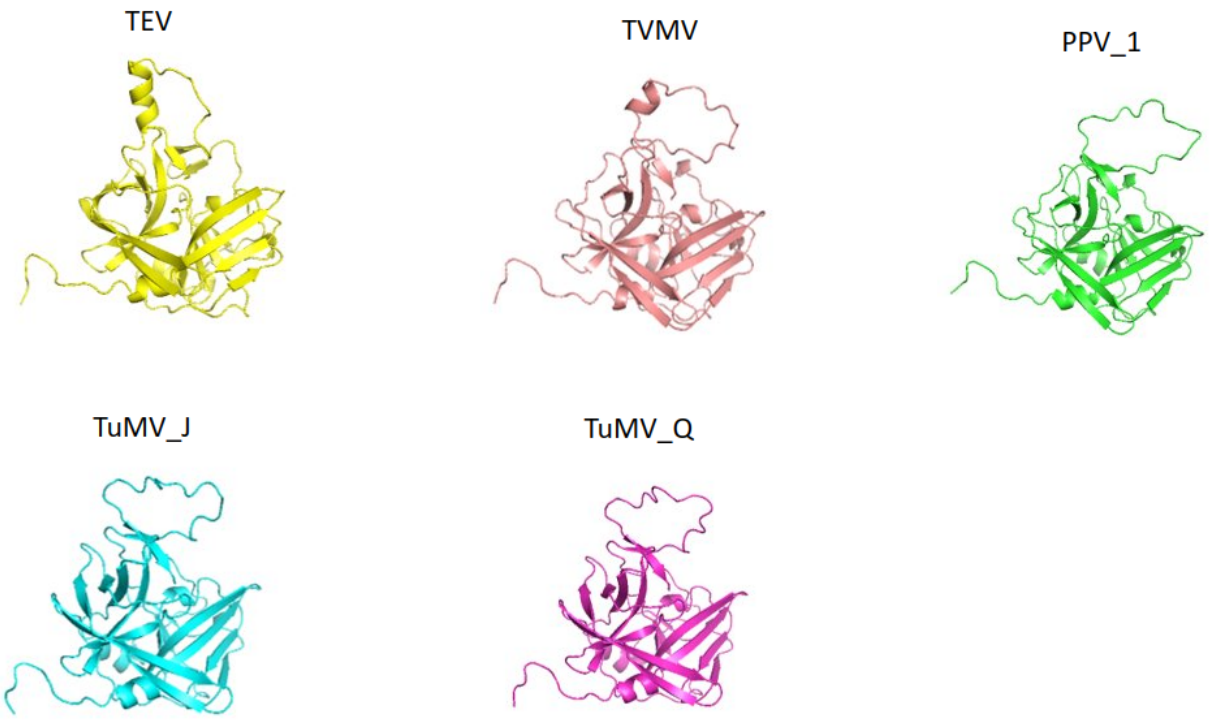

Figure 2.4: AlphaFold2 structural models of tobacco etch virus (TEV) protease and other related proteases including, the tobacco vein mottling virus (TVMV), the turnip mosaic virus (TuMV), and the plum pox virus (PPV).

the reference structures.

2.3 Results

This section details the results from applying the DE-STRESS web server to various applications, in order to demonstrate how it can be used for designing proteins. Firstly, the DE-STRESS metrics were calculated for a set of native/experimentally-determined structures along with their folding decoys. Using PCA it was shown that these metrics could distinguish between the native structures and their decoys. In addition to this, DE-STRESS was used to rank TIMED designs for the rubisco small sub-unit and for a TEV protease. Similarly, PCA was used to find lower dimensional representations for these DE-STRESS metrics, and then the euclidean distance was used to rank the designs, based on their distance from reference structures, across 10 principal components.

2.3.1 Decoy analysis

Firstly, before performing PCA, the variance explained by different numbers of principal components was calculated. The first 2 principal components explain 60% of the variance in the data set and 90% of the variance is explained with 8 components. Therefore, the first two components capture the majority of the variance in the data set and should provide a useful representation of the feature space.

After exploring the different numbers of components, the first two principal com-

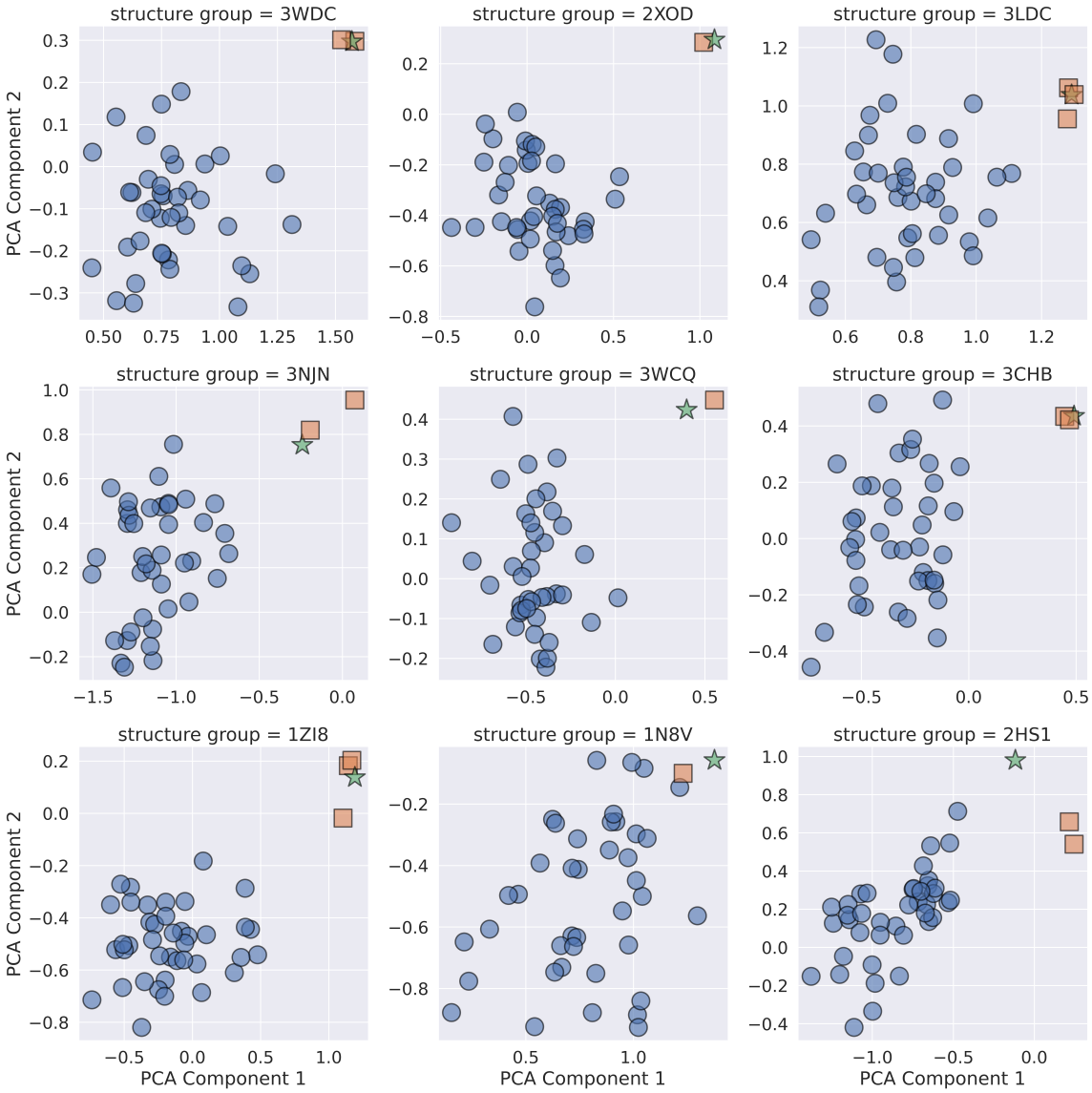

Figure 2.5: Individual scatter plots for each of the experimentally determined structures (green stars), along with their decoy structures (blue circles) and additional crystallographic structures (orange squares).

Top contributors to PC1	Importance	Top contributors to PC2	Importance
rosetta_hbond_sr_bb	0.34	aggrescan3d_avg_value	0.33
rosetta_hbond_lr_bb	0.33	budeff_steric	0.32
budeff_total	0.32	rosetta_fa_intra_sol_xover4	0.31
evoef2_intraR total	0.29	aggrescan3d_max_value	0.31
budeff_charge	0.25	aggrescan3d_min_value	0.28
dfire2_total	0.25	packing_density	0.25

Table 2.2: Top contributors to principal components 1 and 2 along with their importance values/contribution to the principal components.

ponents for each of the 9 experimentally-determined structures, were plotted against each other, along with their decoys and additional crystallographic structures. This is shown in figure 2.5, with the decoy structures represented with blue circles, the experimentally-determined structures as green stars, and the additional structures shown as orange squares. Figure 2.5 shows that for each PDB ID, the experimentally-determined structure, is consistently in the top right corner of the chart, and they generally have the largest values for both principal components 1 and 2, when compared to the decoys. Moreover, the additional crystallographic structures are close to the experimentally-determined structures from the 3DRobot_set. One conclusion that can be made from this observation, is that this analysis is robust to small variations in the experimentally determined structure.

Furthermore, it was investigated which features contributed to the first two principal components. From figure 2.5, it can be seen that both principal components are important in separating out the decoys from the experimentally-determined structures. Table 2.2 shows the top 6 contributors to each of the components, ordered by relative contribution. This analysis shows that short and long range hydrogen bonds in the backbone (rosetta_hbond_sr_bb, rosetta_hbond_lr_bb) and aggregation propensity values (aggrescan3d_avg_value) are some of the most important features that contribute to these principal components. This could suggest that the decoy structural models have differences in the hydrogen bonds and surface properties, which can help distinguish them from the experimentally-determined structures. Another observation, is that a number of values from all the different energy functions (Rosetta, EvoEF2, DFIRE2, BUDE FF) are all in the top 6 contributors to one of these components. Usually, only

one of these energy functions is used in the protein design process, which could suggest that there may be a benefit in using a range of different energy functions. In addition to this, packing density and aggregation propensity are also contributors to the top principal components, which suggests that geometric features and solubility/aggregation propensity measures, are also important, as well as the energy functions.

2.3.2 Protease re-design

The decoy analysis results in the previous section suggest that the DE-STRESS metrics could be useful for ranking designed proteins, as they were able to distinguish between native proteins and folding decoys. Therefore, by following a similar method to the one outlined in section 2.3.1, we should be able to identify high quality designs. For this protease re-design project, we are interested in selecting for high quality designs that are stable and fold into the target structure, but are more soluble than the native proteases. In addition to this, we are interested in designs that have some variance across their physico-chemical properties, which could help us find proteases which have slightly different functions as well, for example, different cut sites. In order to do this, a sequence design method from our lab called TIMED [150] was used to re-design the TEV protease and then DE-STRESS was used to evaluate the designs. A set of conserved sequences were designed, where the conserved residues across the TEV, TVMV, TuMV Q, TuMV J and PPV sequences were kept, and a set of designs were made that were completely *de novo*, for the TVMV and TEV target structures. A reference set of the native proteases, TEV, TVMV, TuMV Q, TuMV J and PPV, was used for comparison as well.

After obtaining the conserved and *de novo* TIMED sequences, for the different combinations of target structures and TIMED models, AlphaFold2 was used to obtain 5 different structural models for each sequence, and then DE-STRESS was ran to obtain physico-chemical properties. PCA was then applied to these DE-STRESS metrics, and the average principal component values were taken across the different AlphaFold2 structural models, for each of these designed and reference sequences. The top two principal components explained 47% of the variance in the original data set while, the top 10 principal components explained 90%. Figure 2.6 shows a scatter plot of principal components 1 and 2, with each colour representing the different TIMED models and the reference structures, and the shapes representing the conserved, *de novo* and reference sequences. In addition to this, table 2.3 shows the top DE-STRESS

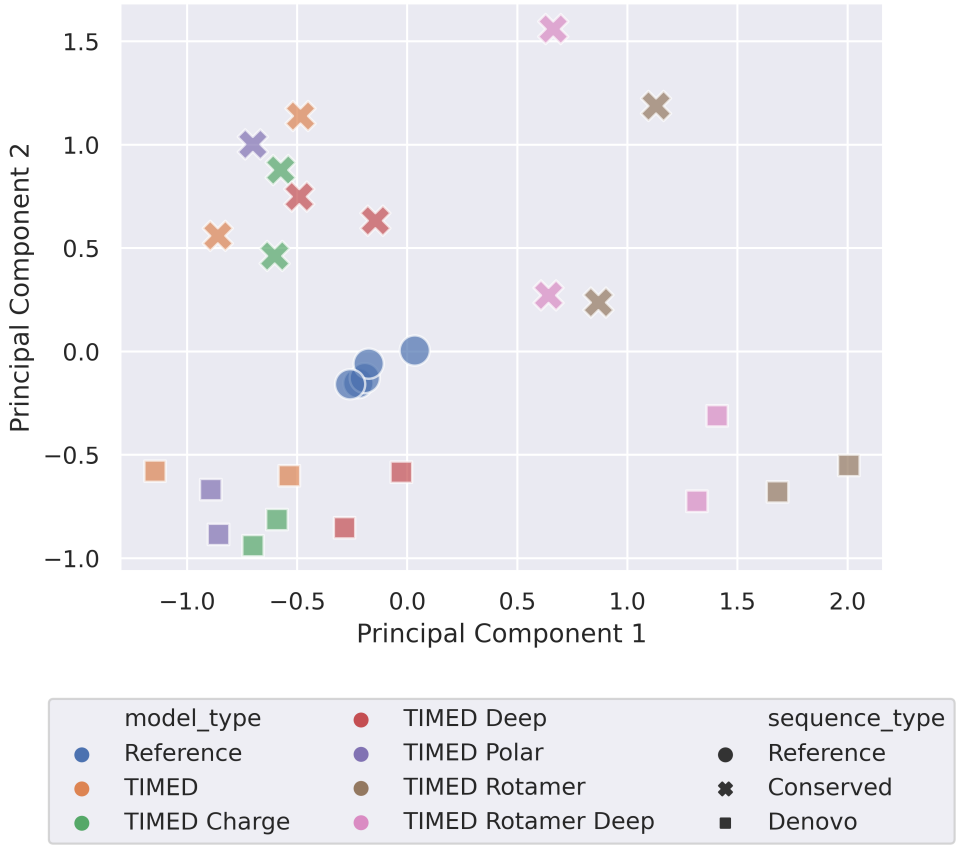

Figure 2.6: Scatter plot of the average values of principal components 1 and 2 across 5 AlphaFold2 models of the TIMED designs. The different colours represent the TIMED or reference sequences, and the shapes represent the conserved, *de novo* and reference sequences.

metrics that contributed to principal components 1 and 2. One observation from figure 2.6 is that the reference structures for TEV, TVMV, TuMV_Q, TuMV_J and PPV, all cluster together towards the middle of the plot. This is encouraging, as we know these proteases have similar functions, and they are shown to have similar physico-chemical properties across this PCA space.

Another observation, is that all the conserved designs have higher values for principal component 2 compared to the reference proteases, and the *de novo* designs all have lower values for principal component 2. Table 2.3 shows that the major contributors to principal component 2, include energy values such as rosetta_total, evoef2_total, evoef2_interS_total and rosetta_omega, and packing_density. In addition to this, figure 2.7 shows bar charts for the average values of these DE-STRESS metrics by sequence type, along with error bars displaying the standard deviation. From this figure, we

can see that overall, the average values for the *de novo* designs are comparable to the reference proteases. In fact, the *de novo* designs have the lowest rosetta_total score on average, which suggests that these designs could be even more stable than the reference proteases. In contrast to this, the conserved sequences have higher average values for all the energy metrics and a lower packing density than both the *de novo* and reference proteases. The rosetta_omega is a backbone dependent penalty for unfavourable omega dihedral angles, which could suggest that the fixing of some residues, has led to steric collisions between atoms for these designs.

One other observation from the scatter plot in figure 2.6, is that the TIMED Rotamer and TIMED Rotamer Deep designs, have much higher values for principal component 1 compared to the reference and the other TIMED designs. Table 2.3 shows that the composition of different amino acids, including alanine, glutamine and valine are some of the major contributors to PC1, along with aggregation propensity and rosetta energy metrics. From figure 2.8, we can see that the sequences designed with TIMED Rotamer and TIMED Rotamer Deep have much higher proportions of alanine, at 10% and 8% respectively, compared to 2-3% for the rest of the TIMED and reference sequences. Additionally, both of these rotamer model sequences have much lower proportions of glutamine, with roughly 1% glutamine in their sequences compared to over 4% in the reference sequence. Furthermore, these sequences also have higher aggrescan3d aggregation propensity scores, at -0.5 compared to -0.8 for the reference sequence, and lower rosetta solvation energy scores, at 0.150 compared to 0.175 for the reference sequence. These differences between the various TIMED models could be due to biases in the models, that cause over or under prediction of certain amino acids.

Top contributors to PC1	**Importance**	**Top contributors to PC2**	**Importance**
composition_ALA	0.30	evoef2_total	0.33
rosetta_fa_intra_sol_xover4	0.29	evoef2_interS_total	0.33
rosetta_fa_intra_rep	0.28	rosetta_omega	0.29
aggrescan3d_avg_value	0.23	rosetta_total	0.26
composition_GLN	0.22	packing_density	0.25
composition_VAL	0.21	rosetta_rama_prepro	0.24

Table 2.3: Top contributors to principal components 1 and 2 and importance values for the PCA analysis shown in figure 2.6

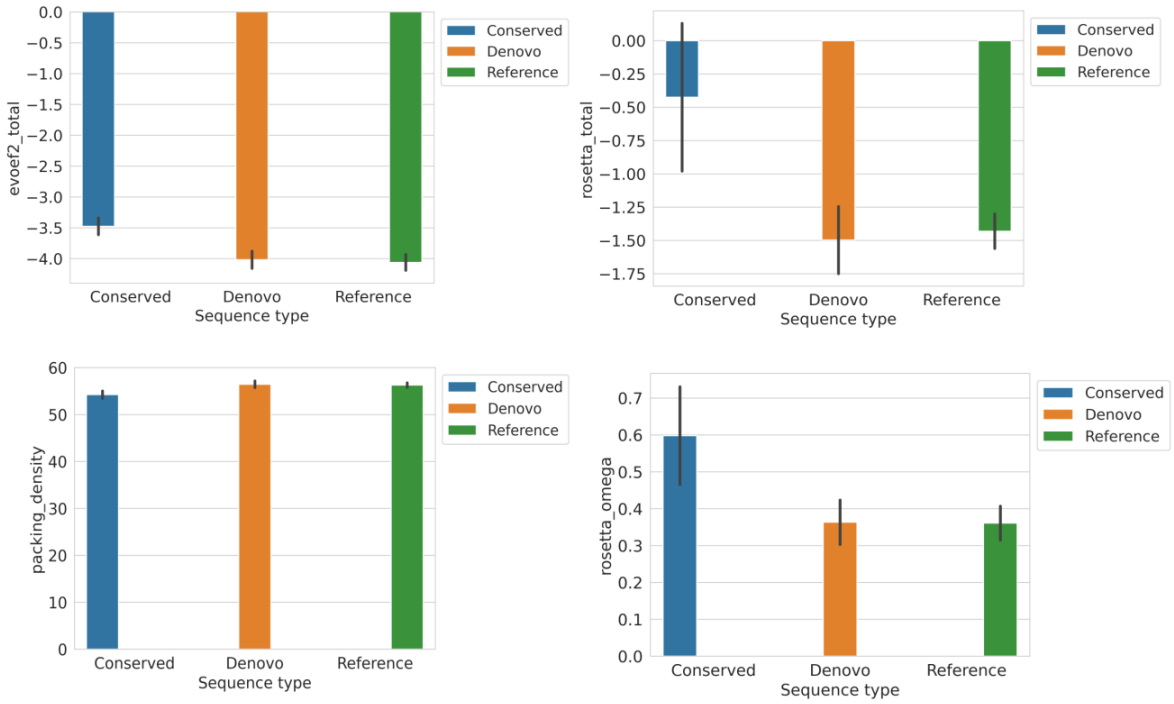

Figure 2.7: Bar plots of the average evoef2_total, rosetta_total, packing_density and rosetta_omega by sequence type, with error bars showing the standard deviation. These metrics are the major contributors to principal component 2.

As both the TIMED Rotamer and TIMED Rotamer Deep sequences have significantly more alanine and less glutamine compared to the rest of the sequences, this could be the reason why some of the other metrics, like rosetta solvation energy and aggrescan3d aggregation propensity, are also quite different. Overall, the major factors that are driving variance across the DE-STRESS metrics, for these protease designs, are issues in the stability and torsion angles of the conserved sequences, and differences in the prediction of certain residues for the TIMED Rotamer and TIMED Rotamer Deep models.

Additionally, the TIMED sequences were ranked by calculating the euclidean distance across 10 principal components against the reference proteases, as these 10 principal components explained 90% of the variance of the original data set. In figure 2.6, only the top two principal components were shown which explain 47% of the variance across the data set. It was decided to use 10 principal components for ranking the designs, as they explain the majority of the variance across the data set. The RMSDs of the AlphaFold2 structural models were calculated against the target structures, to understand how well the designed sequences were likely to fold to the desired structure.

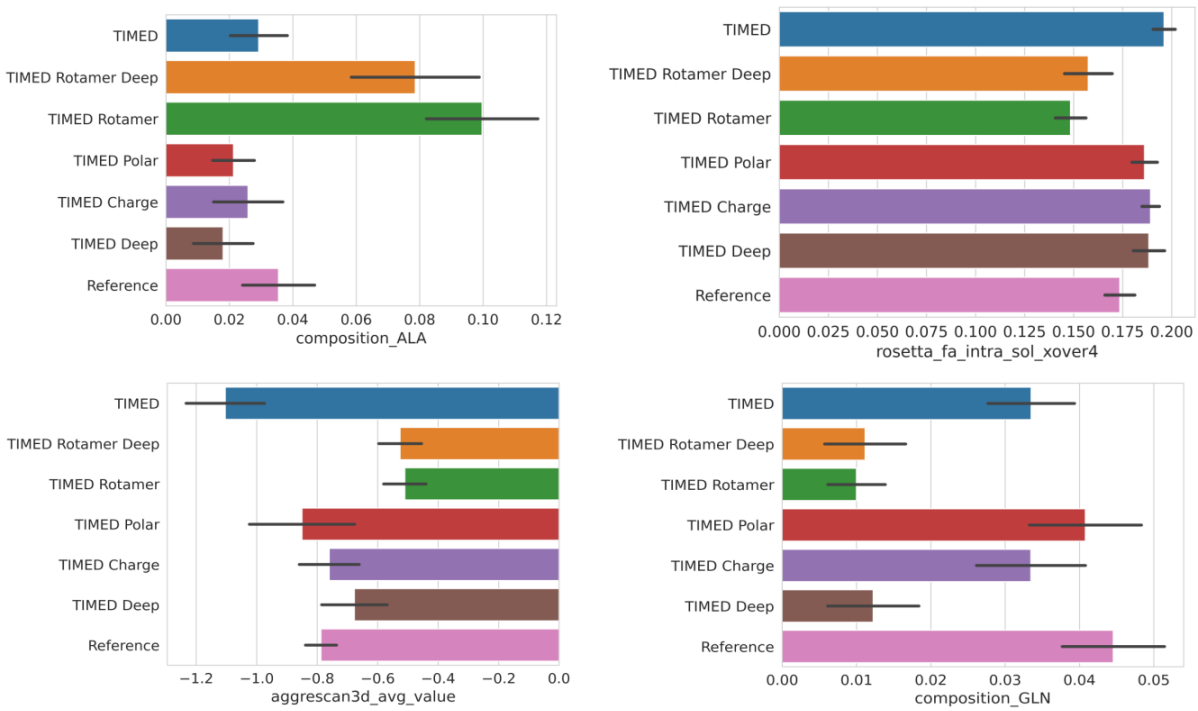

Figure 2.8: Bar plots of the average composition_ALA, rosetta_fa_intra_sol_xover4, aggrescan3d_avg_value and composition_GLN by model type, with error bars showing the standard deviation. These metrics are the major contributors to principal component 1.

Table 2.4 shows the chosen sequences that were sent to our collaborator for experimental testing, along with the PCA 10D distance (euclidean distance across 10 principal components from the reference sequences), RMSD, rosetta total score normalised by number of residues, NetSolP solubility and usability scores (expressibility combined with a measure of how well they can be purified) measures in E.coli [224]. NetSolP uses the ESM-1b language model from Facebook AI [173], which was trained on 250 million sequences, in order to predict solubility and usability of protein sequences in *E. coli*. Both of these NetSolP scores range from 0 to 1, with 0 meaning very low solubility/usability and 1 meaning very high solubility/usability. It was decided to choose a few conserved and *de novo* sequences for both backbones, and some sequences that had the PCA 10D distance close to the reference proteases, and some with much further distances. This was done in order to obtain a range of sequences with different physico-chemical properties. In addition to this, a multiple sequence alignment (MSA) was performed to make sure that these sequences were not too similar, and there was a good amount of variation across the sequence space. One major observation from this table of sequences, is that the RMSD values for TVMV sequences are

Model	**Backbone**	**Sequence Type**	**PCA 10D Distance**	**RMSD**	**Rosetta Score (REU/AA)**	**NetSolP Solubility**	**NetSolP Usability**
TIMED Charge	TVMV	Conserved	1.37	1.47	-0.76	0.42	0.30
Timed Rotamer Deep	TVMV	Conserved	1.50	1.47	-1.01	0.36	0.24
TIMED Deep	TVMV	De novo	2.01	1.45	-1.56	0.48	0.32
TIMED Rotamer Deep	TVMV	De novo	2.05	1.66	-1.69	0.46	0.24
TIMED Deep	TEV	Conserved	1.88	1.68	-0.71	0.36	0.26
TIMED Charge	TEV	Conserved	1.91	1.80	-0.54	0.36	0.25
TIMED Charge	TEV	De novo	2.01	1.61	-1.67	0.44	0.34
TIMED	TEV	De novo	2.24	1.60	-1.21	0.55	0.41
TIMED Rotamer	TEV	Conserved	2.32	2.27	-0.29	0.38	0.27

Table 2.4: Table of the chosen sequences for experimental testing, along with the euclidean distance across 10 principal components from the reference sequences, the root mean square deviation (RMSD) against the target structure, the normalised rosetta total score, NetSolP solubility and usability scores. The PCA 10D distance, RMSD and rosetta scores are averages across AF2 structural models, while the NetSolP scores are calculated from the sequence.

all roughly the same, except for the TIMED Rotamer Deep *de novo* sequence, which has a slightly higher RMSD value of 1.66. In contrast to this, the PCA 10D distance varies more across these sequences and appears to distinguish the TIMED Charge conserved, TIMED Rotamer Deep conserved and TIMED Deep *de novo* sequences better

than RMSD. The same is true for the TEV sequences with the PCA 10D distance varying more than RMSD across these sequences, which could suggest that this distance measure provides much more information for ranking designs. In addition to this, the PCA 10D distance ranks some sequences as further away from the reference sequences while RMSD ranks them as closer, such as the TIMED *de novo* sequence for the TEV backbone, which has a PCA 10D distance of 2.24 but has the smallest RMSD value of 1.60 for this backbone. As RMSD only considers the structure, then this could suggest that the PCA 10D distance is ranking sequences further away based on additional properties and not only the structure.

Furthermore, by comparing the rosetta total and the NetSolP scores, we see that the *de novo* designs have a lower rosetta score (more favourable) and higher solubility and usability scores (more favourable), than the conserved sequences. Although, the conserved sequences, are shown to be closest to the reference sequences with the PCA 10D distance, despite the lower predicted solubility and less favourable rosetta scores, which makes sense as the conserved residues across the reference sequences were fixed in these designs. As the active site is fixed in these sequences, they could be ranked closer to the reference sequences in the PCA 10D distance measure, as they have similar properties and potentially a similar function. However, these results suggest that we cannot treat the designs with lower PCA 10D distance scores as "better", as even though the *de novo* sequences are shown to be further away using this measure, they could end up being better designs due to the lower rosetta scores and higher predicted solubility and usability in *E. coli*. Overall, this PCA 10D distance measure is useful for making sure we select a set of sequences with a range of different physico-chemical properties, which is difficult to do only using RMSD and rosetta energy scores.

2.3.3 Rubisco small sub-unit re-design

Following on from re-designing the TEV protease, the same pipeline was used to re-design the small sub-unit of rubisco from the *Arabidopsis thaliana* plant. For this re-design project, we are interested in obtaining a set of high quality designs, that are predicted to fold to the target structure, are stable, and are predicted to bind onto the main rubisco sub-unit. In addition to this, we want designs that have some variance across their physico-chemical properties, such as charge, as we want to find variants of the small sub-unit that could potentially increase the efficiency of rubisco. However, even if we find designs that reduce the efficiency of rubisco, then this will provide us

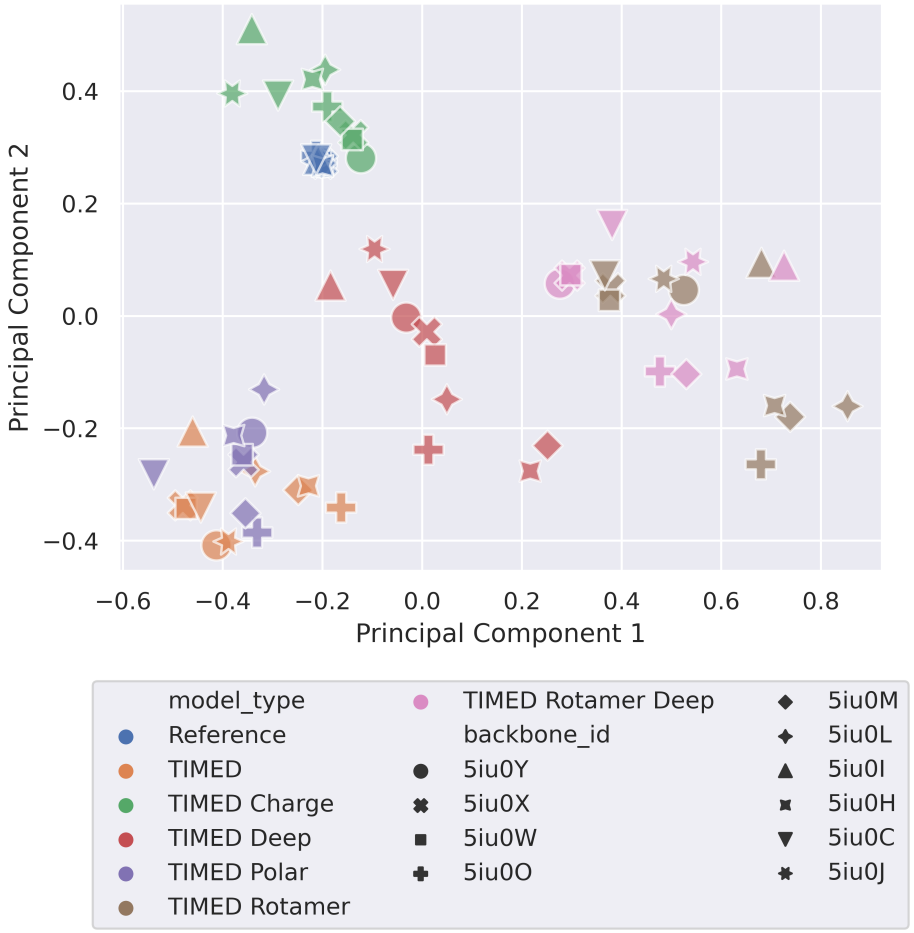

Figure 2.9: Scatter plot of the average values of principal components 1 and 2 across 5 AlphaFold2 models of the TIMED designs. The different colours represent the TIMED or reference sequences, and the shapes represent the different backbone targets.

vital information about the function of the small sub-unit, and its role in the overall function of rubisco.

Similar to the protease re-design project, PCA was applied to the DE-STRESS metrics of the AlphaFold2 structural models for the TIMED and reference sequences. Figure 2.9 shows a scatter plot of the top two principal components, with the different colours representing the TIMED and reference sequences, and the various shapes representing the different target backbones. An average of the principal components has been taken across the 5 different AlphaFold2 structural models for each sequence. The top two principal components explain around 46% of the variance from the original data set, while the top ten principal components explain 90% of the variance. In addition to this, table 2.5 shows the top DE-STRESS metrics that contribute to principal components 1 and 2.

One observation from figure 2.9 is that the reference rubisco small sub-units with the different backbones, are all clustered together. In this case, these reference rubisco sub-units actually have the same sequence, therefore this is expected. The backbone structures are slightly different due to where these sub-units are in the hexadecemeric complex, and the crystallisation process. One major observation is that the TIMED

Top contributors to PC1	Importance	Top contributors to PC2	Importance
charge	0.53	charge	0.57
rosetta_fa_sol	0.39	rosetta_hbond_sr_bb	0.45
evoef2_ref_total	0.32	rosetta_fa_intra_sol_xover4	0.29
rosetta_fa_intra_sol_xover4	0.31	aggrescan3d_max_value	0.27
rosetta_fa_intra_rep	0.29	ss_prop_hbonded_turn	0.27
aggrescan3d_max_value	0.27	rosetta_fa_elec	0.21

Table 2.5: Top contributors to principal components 1 and 2 and importance values for the PCA analysis shown in figure 2.9

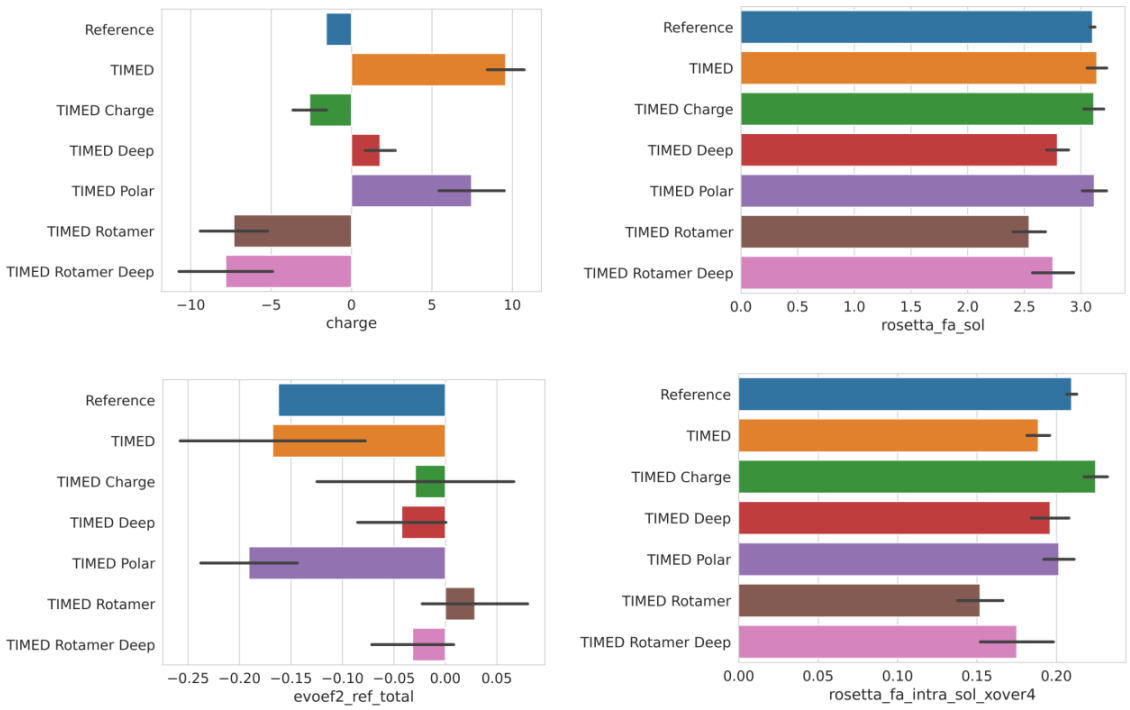

Figure 2.10: Bar plots of the average evoef2_total, rosetta_total, packing_density and rosetta_omega by sequence type, with error bars showing the standard deviation. These metrics are the major contributors to principal component 1.

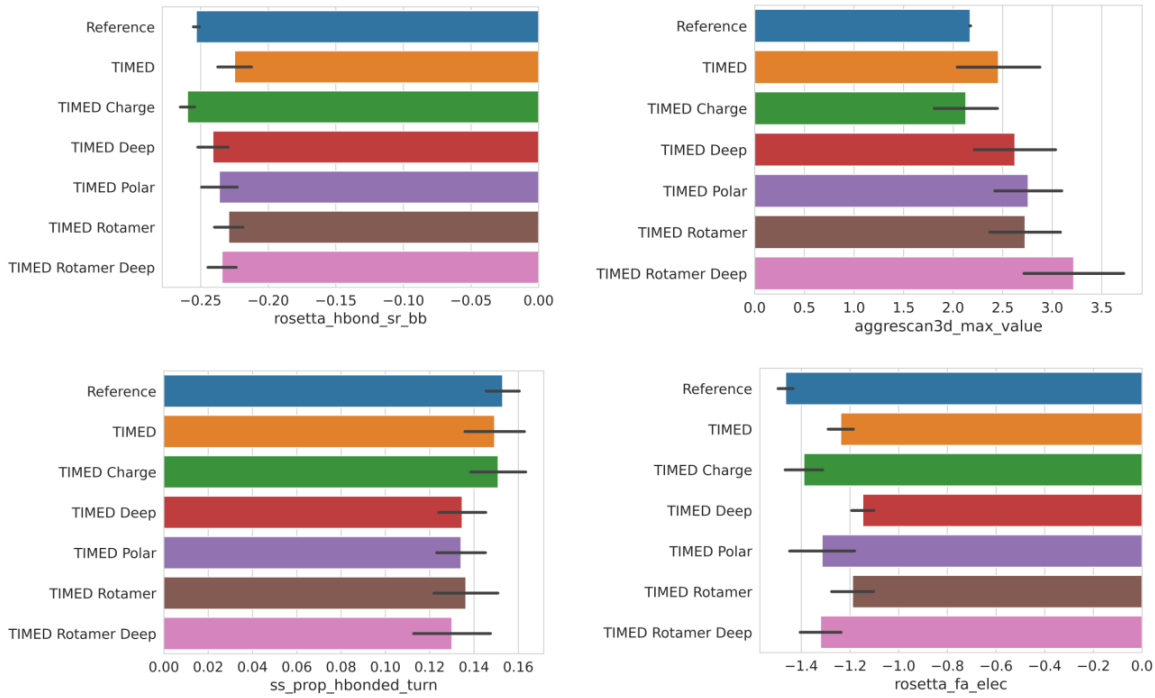

Figure 2.11: Bar plots of the average composition_ALA, rosetta_fa_intra_sol_xover4, aggrescan3d_avg_value and composition_GLN by model type, with error bars showing the standard deviation. These metrics are the major contributors to principal component 2.

Charge designs are all clustered around the reference sequences, which suggest these designs could have very similar properties to the native rubisco small sub-units. Some of the TIMED Deep sequences are also fairly close however, the rest of the designs are pretty spread out across the space. From table 2.5, we can see that charge is actually the largest contributor to both principal components and is driving most of the difference between these designs. In addition to this, rosetta energy terms capturing solvation energy and aggregation propensity also contribute to both principal components, while some additonal Rosetta and EvoEF2 energy terms contribute to principal component 1, and some metrics capturing hydrogen bonds and electrostatics also contribute to principal component 2. Figures 2.10 and 2.11, both show how the average value of these metrics, vary across the different TIMED models. The top left plot on figure 2.10 shows that the reference sequences have an overall negative charge and that TIMED charge is fairly close, although slightly more negative. All of the other models are quite far apart, with TIMED Rotamer and TIMED Rotamer Deep having much more negative charges, and TIMED, TIMED Deep and TIMED Polar having large positive charges. In terms of the two solvation energy terms, rosetta_fa_sol and

rosetta_fa_intra_sol_xover4, TIMED Charge is also fairly comparable to the reference sequences, and most of the other designs have a more favourable solvation energy. In contrast to this, the TIMED Charge designs are very different to the reference sequences in the evoef2_ref_total energy value, which is amino acid specific and captures the energy of proteins in their unfolded state [163]. Most of the the models have a large variance in this metric and only TIMED and TIMED Polar have an average value that is more favourable than the reference. In addition to this, TIMED Charge is shown to be comparable to the reference sequences in figure 2.11 across the rosetta short range hydrogen bond energy on the backbone, the Rosetta electrostatics energy and Aggrescan3d aggregation propensity max value, while the rest of the other designs are more unfavourable. Overall, these results suggest that the main differences between these sequences is the total charge, along with a few other differences in energy values and aggregation propensity, and that TIMED Charge and some of the TIMED Deep designs, are the closest to the reference sequences across this space.

Finally, similar to the protease re-design project, the euclidean distance across 10 principal components between the TIMED and the reference proteases, was used to rank the designs. However, in this case 20 designs were chosen to send to our collaborator for experimental testing. Table 2.6 shows these selected sequences with the models and backbones used, the PCA 10D distance, RMSD, Rosetta total score (averaged by number of residues), and the NetSolP solubility and usability scores. The first observation is that all of the RMSD values are below 1, which means that the AlphaFold2 predicted structural models are very consistent with the target structures and all the Rosetta total scores are negative, which means that these sequences are predicted to fold. In addition to this, sequences were sampled across this PCA space, with different PCA 10D distances, to ensure that we have a set of designs with different physico-chemical properties. The NetSolP solubility scores range quite a lot across these sequences, ranging from 0.47 to 0.86, with the reference sequences having a NetSOLP solubility score of 0.87 and usability score of 0.71. TIMED Charge 5iu0J, which is the highest ranked design in terms of PCA 10D distance, has a high NetSolP solubility score of 0.78 although, it has a low usability score of 0.31. Across the rest of the designs, we do have some sequences with high solubility and relatively high usability scores, including TIMED Deep 5iu0I and TIMED Deep 5iu0M, although most of the designs do have lower scores than the reference sequence. Furthermore, molecular simulations were performed by Jack O'Shea on some of these selected designs, such as the TIMED Charge sequences, to ensure that they assembled with the full rubisco

Model	**Backbone**	**PCA 10D Distance**	**RMSD**	**Rosetta Score (REU/AA)**	**NetSolP Solubility**	**NetSolP Usability**
TIMED Charge	5iu0J	0.42	0.71	-0.63	0.78	0.31
TIMED Deep	5iu0J	0.74	0.86	-0.73	0.79	0.48
TIMED	5iu0J	0.78	0.77	-1.00	0.87	0.45
TIMED Rotamer	5iu0J	0.81	0.97	-0.99	0.47	0.19
TIMED Rotamer Deep	5iu0J	0.97	0.54	-1.24	0.84	0.48
TIMED Charge	5iu0C	0.48	0.77	-0.91	0.69	0.22
TIMED Rotamer	5iu0C	0.68	0.75	-0.73	0.65	0.25
TIMED Deep	5iu0C	0.69	0.60	-0.58	0.74	0.41
TIMED Polar	5iu0L	0.59	0.66	-0.51	0.81	0.41
TIMED	5iu0L	0.68	0.93	-0.52	0.81	0.39
TIMED Rotamer Deep	5iu0L	0.79	0.56	-1.31	0.71	0.30
TIMED Rotamer	5iu0L	1.25	0.79	-1.01	0.80	0.23
TIMED Charge	5iu0M	0.64	0.60	-0.80	0.50	0.36
TIMED Deep	5iu0M	0.85	0.51	-0.62	0.81	0.55
TIMED Charge	5iu0Y	0.70	0.73	-0.82	0.53	0.44
TIMED	5iu0Y	0.82	0.74	-0.56	0.81	0.33
TIMED Deep	5iu0I	0.73	0.59	-0.53	0.85	0.64
TIMED	5iu0I	0.73	0.99	-0.05	0.82	0.41
TIMED Polar	5iu0X	0.80	0.81	-0.51	0.86	0.41
TIMED Rotamer Deep	5iu0H	0.97	0.63	-0.95	0.69	0.35

Table 2.6: Table of the chosen sequences for experimental testing, along with the euclidean distance across 10 principal components from the reference sequences, the root mean square deviation (RMSD) against the target structure, the averaged total rosetta score over number of residues and NetSolP solubility and usability scores. The PCA 10D distance, RMSD and rosetta scores are averages across AF2 structural models, while the NetSolP scores are calculated from the sequence.

complex. Overall, these rubisco small sub-unit designs are predicted to fold to the correct structure and have a range of different physico-chemical properties; however, some of these sequences are predicted to not be as soluble in *E. coli*. However, these results are consistent with the protease results, which show that the PCA 10D distance metric, is useful for selecting designs with a range of properties, which is difficult to do using only RMSD and Rosetta energy scores.

2.4 Discussion

Over the last few years, incredible advancements have been made in protein structure prediction [65; 270; 66; 89] and sequence design methods [150; 151; 271; 272], which has largely been driven by the use of machine learning and the large amounts of protein sequence and structure data available [47; 67; 60]. Generally, these methods demonstrate that the field is becoming fairly proficient at designing towards a target structure however, despite these advancements, the failure rate of protein designs remains high and it is still extremely difficult to design towards properties and functions [200]. In this field, a lot of different computational metrics have been developed to rank designs, such as, energy scoring functions [163; 155], geometric properties [248; 238], solubility and aggregation propensity measures [219; 224; 222], and even function prediction [273; 274; 275]; however, the failure rate of protein designs has still remained very high. There is a huge opportunity to address this high failure rate, in order to make protein design more reliable and cheaper, so that researchers can use designed proteins to solve problems across various scientific areas [91].

To begin with, this chapter introduces the DE-STRESS web server [197], which aims to address some of these issues by generating a set of high quality metrics for structural models of proteins, which are becoming increasingly easier to obtain, due to the development of methods such as AlphaFold2 [65] and ESMFold [66]. These metrics include energy scoring functions, geometric properties, aggregation propensity measures, metrics capturing non-covalent interactions, and amino acid and secondary structure composition. Generally, only one or two of these metrics are used in the design pipeline of novel proteins, and so with the development of DE-STRESS, we propose generating a large set of these high quality metrics for protein designs, that can all be used to compare designs against a set of reference proteins. The DE-STRESS web server allows users to do this with the reference sets functionality, where PDB IDs can be provided to generate a custom reference set from the Protein Data Bank (PDB)

[47], or from uploaded designs. In addition to this, the specifications functionality can also be used to automatically filter designs by certain properties, such as a negative Rosetta energy score or a positive total charge for the protein, for example. Overall, DE-STRESS attempts to extract as much information as possible from structural models of designed proteins, and combines this with the functionality of reference sets and specifications, in order to help design towards functions and address this high failure rate of protein designs.

Another reason for the development of DE-STRESS, was to make protein design more accessible, by providing a user-friendly and intuitive tool, that people can use to evaluate designs. Most of the metrics that are incorporated into DE-STRESS can only be ran from the Command Line Interface (CLI) or Python libraries, which makes these evaluation metrics inaccessible to non technical users. However, DE-STRESS now provides an approach where PDB files can be uploaded to the web server, which calculates all of these metrics, and then the data can be explored on the web server or downloaded as a CSV file by the user. This allows the users to perform their own analysis on the DE-STRESS metrics, in order to gain insight about their designed proteins and rank them for experimental validation. For technical users, headless DE-STRESS was developed, so that the DE-STRESS metrics can be calculated for a large set of PDB files, from the CLI rather than through the user interface. As headless DE-STRESS allows the user to change settings, such as the number of CPUs used and the time limit for calculating metrics, this allows the DE-STRESS metrics to be ran at scale across large structural data sets such as the PDB [47] and the AlphaFoldDB [67]. Furthermore, headless DE-STRESS could be incorporated into protein design pipelines, where sequence design methods are used to generate sequences, structure prediction methods are used to obtain structural models, and then headless DE-STRESS is used to generate a set of physico-chemical properties for evaluating the designs. By providing both the user-friendly web server and headless DE-STRESS, we aim for DE-STRESS to be used by non experts and seasoned protein designers, and to be easily incorporated into protein design pipelines.

After developing the DE-STRESS web server, analysis was performed on a set of experimentally-determined structures and their folding decoys, which were generated using 3DRobot [253]. The DE-STRESS metrics were calculated for all of these structures and then PCA was used, to understand how these structures compared across these metrics. PCA was chosen for this analysis because it is a simple method with only one hyperparameter and it is explainable. However, one disadvantage is that it

assumes a linear relationship between features which may not be true. Other techniques such as UMAP [255] can capture non-linear relationships across features and therefore they could also have been used for these analyses. However, these techniques have many hyperparameters which can dramatically impact the results of the models, and they are not explainable. The explainability of PCA was important for this analysis as we needed to understand which DE-STRESS features contributed the most to the variance across this data set, to better inform the design of novel proteins. Furthermore, PCA was chosen for other analyses in this PhD thesis, for example in chapter 3, for the same reasons.

Figure 2.5 shows the results from this PCA analysis and we can see clearly that the experimentally-determined structures, cluster away from the decoy structures, and generally have higher values for both principal components 1 and 2. Table 2.2 shows that the top contributors to principal component 1 include rosetta hydrogen bond energy terms and other energy terms from BUDE FF, EvoEF2 and DFIRE2. On the other hand, the top contributors to principal component 2 include Aggrescan3d aggregation propensity metrics, Rosetta and BUDE FF energy terms, and packing density. These results suggest that the decoy structures have differences in the hydrogen bonds and surface properties, which can help distinguish them from the experimentally-determined structures. Furthermore, a range of different energy functions, geometric features and aggregation propensity measures, were found to be important for distinguishing these decoy structures from the experimentally determined structures. In general, only one of these energy functions is used to evaluate designs, so this result suggests that it could be beneficial to include a range of metrics for designing novel proteins.

Although the DE-STRESS metrics were shown to be useful for distinguishing experimentally-determined structures from folding decoys, they still needed to be tested on how well they evaluated designs for experimental testing. In order to do this, we used these DE-STRESS metrics to evaluate protein designs for two different re-design projects, targeting the TEV protease and rubisco small sub-unit. Firstly, the different TIMED sequence design models [150] were used to generate a set of amino acid sequences for the different target backbones. After this, AlphaFold2 [65] was used to obtain structural models for these sequences, these models were then relaxed using energy minimisation with Amber [276; 277], and finally DE-STRESS was used to calculate a set of physico-chemical properties for these designs, and for a reference set of known proteins. For both of these re-design projects, PCA was used to understand

the main metrics that varied across the designs and reference structures, and then the euclidean distance was used across 10 principal components, between the designed proteins and reference proteins, in order to evaluate the designs.

The major factors that varied across the protease designs, were found to be issues in the stability and torsion angles of the conserved sequences. This was probably caused by the fact that some of the residues were fixed in these sequences, and the other residues were changed using TIMED, without considering these fixed residues. In addition to this, the TIMED Rotamer and TIMED Rotamer Deep models, appeared to predict amino acid identities in different proportions to the other models, which also drove a lot of the variance across these protease sequences. In contrast to this, the main factors that varied across the rubisco small sub-unit designs, were the total charge of the sequences, solvation, electrostatics and solvation energy values, and aggregation propensity measures. By generating a lot of different metrics for these designs and then using PCA for dimensionality reduction, we were very quickly able to understand the main factors that varied across these designed sequences and how they compared to the reference sequences. These projects demonstrate how DE-STRESS can be used to gain a wealth of information about designed proteins, and how this could form part of a robust design methodology which is explainable and simple to implement.

Another observation from these results, is that the RMSD values were pretty similar across designs, while the PCA 10D distance varied quite a lot. In general, sequence design methods are evaluated by using a structure prediction method, such as AlphaFold2 or ESMFold, and calculating the RMSD for the predicted structure against the target structure, in addition to calculating the native sequence recovery [271; 278]. However, the results here show that there is a huge amount of information, about the physico-chemical properties, that is ignored using this approach, which could be crucial for the success of a design. Therefore, these results suggest that it might be better to use a scoring method, that considers a large amount of metrics capturing physico-chemical properties about the design, rather than only RMSD.

Other approaches have used the predicted local distance difference test (pLDDT), along with the RMSD to the target structure, in order to rank designs [217], and many studies use a single energy scoring functions to evaluate designs [279; 280]. However, DE-STRESS offers an alternative approach that provides much more information about the properties of these designs. The PCA 10D distance was used in both of these projects, to sample designs with a range of different properties, and were still predicted to fold to the correct structure. For both of these protein re-design projects, the

chosen designs were sent to collaborators for testing in the lab, and these experimental results will be used to understand how well DE-STRESS can be used to evaluate designs. In the future, we will sample a large number of sequences from the highest performing TIMED model, rather than selecting the top predicted amino acid at each position, from different TIMED models. This will provide us with a larger set of designs, which will hopefully capture a range of physico-chemical properties. The same approach could be then be used to evaluate designs, but we could also explore different dimensionality reduction methods and other approaches to ranking the designs, rather than using the euclidean distance across the PCA space. Overall, these projects outlined a way that DE-STRESS can be used, along with sequence design and structure prediction methods, to evaluate protein designs for different applications, and there is evidence that this approach could be more useful than only using RMSD or a single energy scoring function. However, we will need to obtain experimental results before we can understand how well DE-STRESS can be used to evaluate designs, and how useful it is for reducing the failure rate of the protein design process.

2.5 Next steps

Although the DE-STRESS web server includes a range of different metrics covering, energy scoring functions, geometric properties, non-covalent interactions and aggregation propensities of proteins, there are a huge number of other metrics that could be added. Two of these metrics that will be incorporated into DE-STRESS, are the NetSolP solubility and usability scores [224], that have been shown to be useful in evaluating how well protein sequences are soluble and can be expressed in *E. coli*. Both of these metrics were used for evaluating the protease and rubisco small subunit designs however, they have not been added into the DE-STRESS web server yet. Metrics capturing more detailed information about cavities and pores in the protein structures, will also be added into DE-STRESS. CICLOP [231] provides rich granular information about the size, hydrophobicity and charge of cavities, which could be extremely useful for protein designers for evaluating designed proteins, in addition to packing density and hydrophobic fitness, which are more global descriptors of the protein structure. Furthermore, metrics such as contact order and the distribution of charged residues across the sequence, could also be explored and incorporated into the web server. Correlation coefficients and mutual information scores, could be used to understand whether these new metrics add any additional information over the cur-

rent metrics, by calculating them across the PDB. On the other hand, there are risks with adding more features, especially for training models on small data sets, as models trained with many more features than data points, are susceptible to overfitting [281]. However, this can also be avoided by using feature selection methods and properly exploring and understanding the data sets before training any models. By expanding the set of DE-STRESS metrics calculated for designed proteins, we aim to extract as much information as possible from the structural models, to help better inform which designs to test experimentally.

One of the major limitations of DE-STRESS at the moment, is that a few of the software packages that are included, such as the Rosetta energy metrics, have to be installed separately, if users want to install a local version of DE-STRESS on their own machines. In addition to this, Rosetta is free for academics but companies have to purchase a commercial licence before using it, which currently limits who can use DE-STRESS. This will be addressed by creating a new version of DE-STRESS, that does not include Rosetta and any other metrics that require commercial licences, which will allow DE-STRESS to be used by a larger group of people, and will help to simplify the installation process. These changes will make DE-STRESS easier to incorporate into protein design pipelines, and to become a useful tool for academic and industrial users.

Both of the protein re-design projects detailed in this chapter, show examples of how DE-STRESS can be incorporated into protein design pipelines, together with sequence design and structure prediction methods. However, currently these tools have to be ran separately with a few manual steps. Therefore, one of the next steps for DE-STRESS, is to incorporate it into an easy to use computational pipeline with sequence design methods such as TIMED [150] and ProteinMPNN [151], and ESMFold [66] or OmegaFold [88] for structure prediction. Users will be able to upload a target structure and the sequence design methods will generate a set of amino acid sequences. After this, ESMFold or OmegaFold, will be used to obtain structural models, as these methods are very simple to install and fast to run. Next, the DE-STRESS metrics will be ran across these structural models, and all of the data including the designed sequences, structural models and DE-STRESS metrics, will be available for download by the user. The developed pipeline will be easy to use, very fast to run and will allow the user to select different sequence design models and structure prediction methods. Overall, this pipeline would extend the capabilities of DE-STRESS to include the full protein design process, with the aim of making the process of designing proteins a lot easier and

more accessible.

Although making protein design more accessible to a wider range of labs would have huge benefits for the field, one critical issue that needs to be considered is biosecurity, and how to prevent these tools from being misused. AI based methods have revolutionised areas such as drug discovery in recent years; however, there is increasing focus on the potential for these tools to be misused, to produce toxic compounds or biological weapons [282]. In order to prevent this misuse for protein design, there is a need for more systematic screening and logging at DNA synthesis companies [283]; however, protein design methods will also need to incorporate safe guards, to limit the potential of users knowingly or unknowingly designing harmful proteins in the first place. This will require collaboration from the entire protein design community, government and international agencies and DNA synthesis companies, to realistically address these issues. In the future, we will aim to incorporate safe guards into our design methods, to limit the potential for misuse of our tools, and to allow the incredible benefits of protein design to be realised, while minimising the potential for harm.

Finally, further work should be performed to understand the sensitivity of the DE-STRESS metrics to different protein structure prediction methods (for example AlphaFold, ESMFold and OmegaFold) and across structural models from a particular method. This is important to understand as the choice of structure prediction method could impact the DE-STRESS metrics and the results from any downstream analysis. Additionally, the sensitivity of these metrics across structural models from a particular method could be explored, to understand which DE-STRESS metrics are robust or sensitive to small variations in the protein structure. This type of analysis could also be performed by conducting molecular dynamics simulations on a structure and analysing how the DE-STRESS metrics vary over time. Overall, this work will help us understand the impact of different structure prediction methods on the DE-STRESS metrics and also how robust these metrics are to small variations in the predicted structures.

2.6 Conclusion

In conclusion, this chapter has introduced a user friendly web server called DE-STRESS, which provides high quality physico-chemical metrics for evaluating designed proteins. DE-STRESS incorporates novel functionality, such as reference sets and specifications, to help users design towards specific properties and functions. Additionally, by developing an easy to use and intuitive web server, as well as a headless version

that can be ran through the command line interface, we have enabled both non experts and experienced protein designers, to leverage these metrics for evaluating designs. To understand whether these metrics were useful for design, they were applied to a set of experimentally-determined structures and a set of folding decoys. These metrics were shown to consistently separate the experimentally-determined structures away from the decoys, and we found that a range of different metrics including, energy functions, geometric properties and aggregation propensity measures, were useful for distinguishing them. Generally, only one of these is used for evaluation, and these results suggest that we could benefit from using a wider range of metrics for ranking designs.

After this, DE-STRESS was used to evaluate a set of protease and rubisco small sub-unit designs before testing these designs in the lab. Using the simple dimensionality reduction method, principal component analysis, we were able to rapidly understand the main factors that varied across the DE-STRESS metrics, for designed proteins and a reference set of known proteins. In addition, this method was shown to provide a lot more information than the root mean square deviation (RMSD) of the predicted and target structures, which may make this more suitable for evaluating designs. However, we are still waiting for experimental data to properly validate how well DE-STRESS can evaluate designs before taking them into the lab, and if this method could reduce the failure rate of designs. In the future, additional metrics will be incorporated into DE-STRESS to expand the properties that it can generate for designs, and further improvements will be made to the web server so that it can be widely used by both academia and industry. Furthermore, moving forward, DE-STRESS will be developed into an easy to use computational protein design pipeline, that includes sequence design models and structure prediction methods as well, with the aim of making protein design more accessible and to reduce the failure rate of designs. Although, in the future, it will be critical to work together with the whole protein design community, to incorporate safe guards into our design tools to reduce the risk of them being misused to develop harmful proteins.

Chapter 3

AlphaFold structural features predict antibody production and phylogenetics

Rapid advancements in protein structure prediction methods have ushered in a new era of abundant and highly accurate structural data. Leveraging these huge data sets can provide greater insight into the biological properties and functions of proteins, which is vital for understanding the role of proteins in nature and creating new proteins. In this chapter, large structural data sets were utilised to determine whether features derived solely from predicted structures, could be used to understand *in vivo* properties of proteins. By analysing physico-chemical features from DE-STRESS for structural models of 192 designed, single chain variable fragment (scFv) antibodies, it was demonstrated that these properties were predictive of *in vivo* protein production. In addition to this, these properties were calculated for half a million AlphaFold2 models of proteins, and it was shown there was systematic variation between the properties of proteins from different organisms, to such an extent that the tree of life spontaneously arose out of these data. Due to the high degree of functional constraint around the chemistry of proteins, this result is surprising and suggests that properties of proteins may be optimised to their unique molecular environment. In the future, design methodologies could be developed that consider more information about the environment of the designed protein, which could help to reduce the failure rate of designs. However, these results have only be shown for predicted protein structures, and further work needs to be performed to validate these findings and to understand whether it would be useful for design.

3.1 Background and motivation

3.1.1 The advent of large protein structural data sets

The development of highly accurate protein structure prediction methods, such as AlphaFold [78; 65] and ESMFold [66] has resulted in unprecedented amounts of protein structural data becoming available to researchers. Protein structure prediction methods are explained in detail in section 1.2.6. The AlphaFold DB [67] has over 200 million predicted protein structures, covering the majority of the UniProt database [60] and a huge variety of organisms. Complementary to the AlphaFold DB is the ESM Metagenomic Atlas [66] which used ESMFold to obtain predicted structures for 772 million metagenomic proteins, and the MIP database [68] which used Rosetta [158] and DMPFold [284] to predict the structures of 200,000 metagenomic proteins across the microbial tree of life. By leveraging these large structural data sets, researchers can now study the biological properties and functions of proteins at a scale that has never been possible before. Furthermore, the insight gained from exploring these data sets is important for understanding the role of proteins in nature and designing novel proteins to address challenges across various scientific areas.

Recent studies have used these large structural data sets for a broad range of applications including; uncovering new protein families and folds across natural protein structures [285], identifying and prioritising drug targets [286], augmenting training data with predicted structures to learn inverse folding for protein sequence design [287], and using protein structural information to improve phylogenetic trees [288]. These studies show a glimpse of the potential that these data sets have in the fields of structural biology, protein design, medicine, and evolutionary biology, and they demonstrate that we are at the beginning of a completely new field of science [289].

3.1.2 Leveraging these data sets for understanding *in vivo* properties of proteins

In this project, we performed a large-scale analysis of predicted protein structures to determine if physico-chemical descriptors of these structural models were predictive of *in vivo* properties. To explore this, we calculated a set of model-derived physico-chemical properties using our structural model evaluation server DE-STRESS [197]. This programme calculates various structural descriptors using a range of software including measures of packing density, hydrogen bonding quality, aggregation propen-

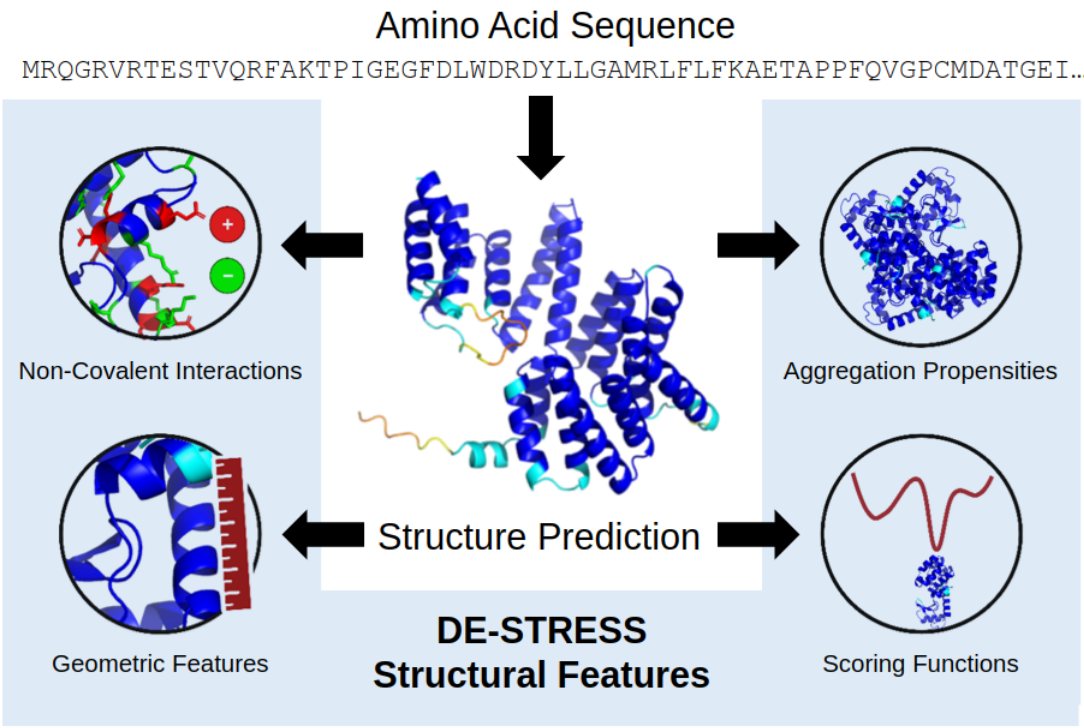

Figure 3.1: Overview of the computational pipeline detailing how AlphaFold2 and DE-STRESS are used to extract model-derived physico-chemical properties from proteins. Created with BioRender.com.

sity, isoelectric point, statistical potentials and many more (figure 3.1). We determined these properties for 192 designed single chain variable fragment (scFv) antibodies from the Fleishman lab [159]. Unsupervised and supervised learning methods were used to demonstrate that these properties were predictive of *in vivo* protein production. Next, we applied similar methods at scale for 564,446 AlphaFold2 (AF2) structural models from 48 model organisms [67] to gain an understanding of how these properties varied more broadly. Finally, we found that there are systematic differences in the model-derived properties between organisms, to such an extent that eukaryotic and prokaryotic organisms can easily be distinguished, and these properties can be used to reconstruct the tree of life. This work is presented in a pre-print on BioRxiv [198].

3.2 Methods

Firstly, this section provides a description of the AlphaFold DB, Fleishman scFv and SAbDab scFv data sets that were used for the analysis in this chapter. After this, details

are provided about how these data sets were prepared and processed for downstream analysis. Finally, the steps taken for training models to predict scFv protein production levels, performing the large scale analysis of protein properties across the AlphaFold DB, and exploring how these properties vary across organisms, are all described.

3.2.1 Datasets

3.2.1.1 Fleishman and SAbDab (PDB) scFv data sets

The Fleishman data set consists of 192 single chain variable fragment (scFv) amino acid sequences, along with a yeast display protein production measure [159]. Five unrelaxed structural models of the scFv sequences, were obtained using AlphaFold2 (AF2) [65], and the highest ranked model was selected for each design. In addition to this, a set of non-redundant, experimentally determined scFvs were obtained from the Structural Antibody Database (SAbDab) [290; 291]. This database contains all the antibody structures from the Protein Data Bank (PDB) [47], and it was used to search for scFvs that had a maximum sequence identity of 90%, were not in a complex with an antigen, and included both heavy and light chain regions. After obtaining this data set of 41 experimentally determined scFvs, AF2 was then used to generate unrelaxed, structural models for these scFvs, in order to be consistent with the Fleishman data set. This is because there could be inherent differences between the PDB structures and the AF2 structural models, from biases in the AF2 structure prediction. This is important as this could confound our analysis of the model-derived properties and scFv protein production.

3.2.1.2 AlphaFold DB and PDB data sets

AF2 structural models for model organisms and global health proteomes, were obtained from the AlphaFold Protein Structural Database [67]. In total, there were 564,446 structural models for 48 organisms, including animals, archaea, bacteria, fungi, plants and protozoans, which were downloaded from https://alphafold.ebi.ac.uk/download as PDB files. Table S11 in the supplementary materials shows the full list of 48 organisms, along with the corresponding volume of protein structural models in the data set. For the PDB data set, a total of 189,942 experimentally determined, protein structures were downloaded from the Protein Data Bank [47] as PDB files. This data set is up to date as of July 2020.

3.2.2 Data preparation

3.2.2.1 Fleishman and SAbDab scFv data sets

Our model evaluation software DE-STRESS [197], was used to generate physicochemical properties from the structural models in the Fleishman and SAbDab scFv data sets, and these were used as the features throughout this analysis. These properties included a range of all-atom scoring functions [155; 163; 166; 167; 207], geometric metrics such as packing density [248; 247] and hydrophobic fitness [238; 247], aggregation propensity [219; 226], isoelectric point, and amino acid and secondary structure composition. All these metrics were generated from structural models of folded proteins and did not include any information on DNA or mRNA sequences.

Author reported protein production values, found from performing yeast display experiments on the scFv designs [159] were split into three equal classes: low, medium and high. These three classes were informed by analysing the distribution of protein production values across the whole data set. Histograms of these values showed peaks at the low and high protein production levels, and lower volumes around the medium production levels. After this, these protein production classes were used for stratified sampling, in order to create a 75% training set and 25% test set.

Several different data processing, feature selection and scaling methods were applied to the training set to prepare it for analysis and model training. All-atom scoring functions and hydrophobic fitness values were normalised by the number of residues in each design, as sequence length impacts the magnitude of these energy value features. In addition to this, features that were constant or had the same value for more than 75% of the samples in the training set, were identified and removed, and the data set was scaled using the standard, robust and minmax scaling methods. Highly correlated features were determined using the spearman correlation coefficient [292] (0.7 or higher), and these features were removed sequentially until no correlated features remained in the data set. All these steps were repeated, including, and excluding the amino acid composition metrics, and with two different methods of feature selection. One of these methods involved calculating the mutual information score [293] of each feature against the protein production classes, and retaining those features that had a score greater than the mean mutual information score of all features. In addition to this, a random forest [294] was used with the number of estimators set to 1000 and a balanced class weight. Finally, the scalers used, and features removed from the training set, were also applied to the SAbDab and test scFv data sets as well.

3.2.2.2 AlphaFold DB and PDB data sets

Firstly, 564,446 PDB files were downloaded from the AlphaFold DB, and the DE-STRESS metrics were calculated, using the headless version of the DE-STRESS software. Features that contained greater than 5% missing values were dropped from the dataset, along with amino acid and secondary structure composition metrics. Once these features had been dropped, any rows that had missing values for the rest of the features were removed, which resulted in 564,432 structural models left in the data set. Next, the Uniprot API [60] was used to extract additional information about each of the protein structures in the data set, including the organism that the protein originates from. This information was then joined onto the data set of DE-STRESS metrics by Uniprot id, and any duplicates in the data set were removed. After this, constant features were dropped, the data set was scaled with the minmax, standard and robust methods, and highly correlated features were removed, in the same way as the Fleishman scFv data set. This data set was used for the analysis in figure 3 however, for the organism clustering in figure 4 and 5, we also excluded homologous sequences in order to remove bias. MMseqs2 [295] easy cluster with min_seq_id set to 0.3, was used to remove these sequences. After this, the output was used to filter the data set to 387,810 non-redundant proteins. In addition to this, low quality AF2 structural models were excluded, by removing models with average pLDDT $< 70\%$. which resulted in 241,134 structural models in the organism clustering. The final set of DE-STRESS metrics used in this data set were; isoelectric_point, budeff_charge, evoef2_ref_total, rosetta_lk_ball_wtd, rosetta_fa_intra_sol_xover4, rosetta_hbond_lr_bb, rosetta_hbond_sr_bb, rosetta_hbond_sc, rosetta_fa_dun, aggrescan3d_avg_value, aggrescan3d_min_value, aggrescan3d_max_value. All these different DE-STRESS metrics are detailed in appendix A.

After this, 189,942 structures from the PDB were processed in the same way as the AF2 structural models, except the MMseqs2 step was not used in this case. As the PDB data set is smaller than the AF2 data set, a threshold of 20% was used to remove features that had a lot of missing values. After removing rows that still had missing values remaining across the features, there were 165,293 structures left in the data set. Designed proteins were identified and labelled in this data set using a curated list [107], and the rest of the PDB structures were labelled as “native”. The final set of DE-STRESS metrics used in this data set were; isoelectric_point, packing_density, evoef2_ref_total, rosetta_fa_elec, rosetta_fa_sol, rosetta_lk_ball_wtd,

rosetta_fa_intra_sol_xover4, rosetta_hbond_bb_sc,
rosetta_rama_prepro, rosetta_p_aa_pp, rosetta_fa_dun, aggrescan3d_avg_value, aggrescan3d_min_value, aggrescan3d_max_value (also detailed in appendix A).

3.2.3 Predicting protein production from model-derived properties

Principal component analysis (PCA) [254] was used to explore whether the model-derived properties could distinguish scFv designs with low and high protein production. This was performed on the different training sets, with the minmax, standard and robust scaling methods, including and excluding amino acid composition metrics, and using a random forest [294] and mutual information [293] for feature selection. The features that contributed to the principal components (PCs) were calculated, along with the variance explained by each component. The AF2 structural models of the 41 SAbDab scFvs were also included in this analysis, to see how they compared to the designed scFvs across the model-derived properties. After performing PCA, stratified 5-fold cross validation, with 10 random splits, was used to train and validate naive bayes classifiers [81] on the different training sets. These models were evaluated with weighted precision, recall, one-vs-rest multiclass ROC curves, and confusion matrices. Finally, these models were also evaluated on the 25% test set with the same metrics.

3.2.4 Large-scale analysis of model-derived properties

PCA was applied to the data sets of 564,432 AF2 structural models and 165,293 PDB structures, scaled using the minmax and robust scaling methods. Scatter plots of PC1 and PC2 were created, and a sample of proteins was labelled around these spaces. In addition to this, the main contributors to PC1 and PC2 were calculated, and plots were created to show how metrics such as secondary structure composition, isoelectric point, packing density and aggregation propensity, varied across this space. This was shown across the scatter plots of PC1 and PC2, but also as cumulative histograms.

3.2.5 Exploring the relationships between organisms

For all three scaling methods, the average model-derived properties were computed for the 48 organisms, and then PCA and clustering methods were applied to these data sets. K-means [296] was performed with 100 random initialisations, with the number of clusters ranging between 2 and 20 clusters. The adjusted rand index [297] was

used to compare the clustering labels against labels indicating whether the organism was eukaryotic or prokaryotic. Hierarchical agglomerative clustering [298] was then performed on the average model-derived properties of each organism, with the Euclidean distance metric and four different linkage methods: single, average, complete and ward. Dendrograms were created for each of these clusterings and then compared to a tree of the 48 organisms created from the common tree tool from the NCBI Taxonomy Browser [299]. The clustering information distance metric [300] was used to compare each of the dendrograms of the 48 organisms against the NCBI tree. Finally, the Interactive Tree Of Life online tool [301] was used to display these trees and to sort the trees by the number of leaf nodes, with the fewest at the bottom and the most at the top.

3.3 Results

This section describes the main results found in this chapter. To begin with, physico-chemical properties from DE-STRESS were generated for a set of designed scFv AlphaFold2 structural models, and they were shown to be predictive of *in vivo* protein production levels. After this, these properties were generated for 564,432 AlphaFold2 structural models from 48 different organisms, and also 165,293 experimentally determined structures from the PDB, and secondary structure was shown to be one of the major factors that varied across these protein structures. In addition to this, for the experimentally-determined structures, designed proteins were shown to be different across these properties compared to native proteins. Finally, these properties were shown to distinguish between eukaryotic and prokaryotic organisms, and they could even be used to reconstruct the tree of life.

3.3.1 Model-derived properties can be used to predict protein production levels

To understand the relationship between the DE-STRESS physico-chemical properties and protein production, a set of designed scFv antibodies were analysed, which had been experimentally characterised in Baran et. al [159]. Figure 3.2 A shows the 4m5.3 reference scFv, that was used in this study, along with one of the high producing designed scFvs with design ID 5ins05. This set of proteins was of interest as, while

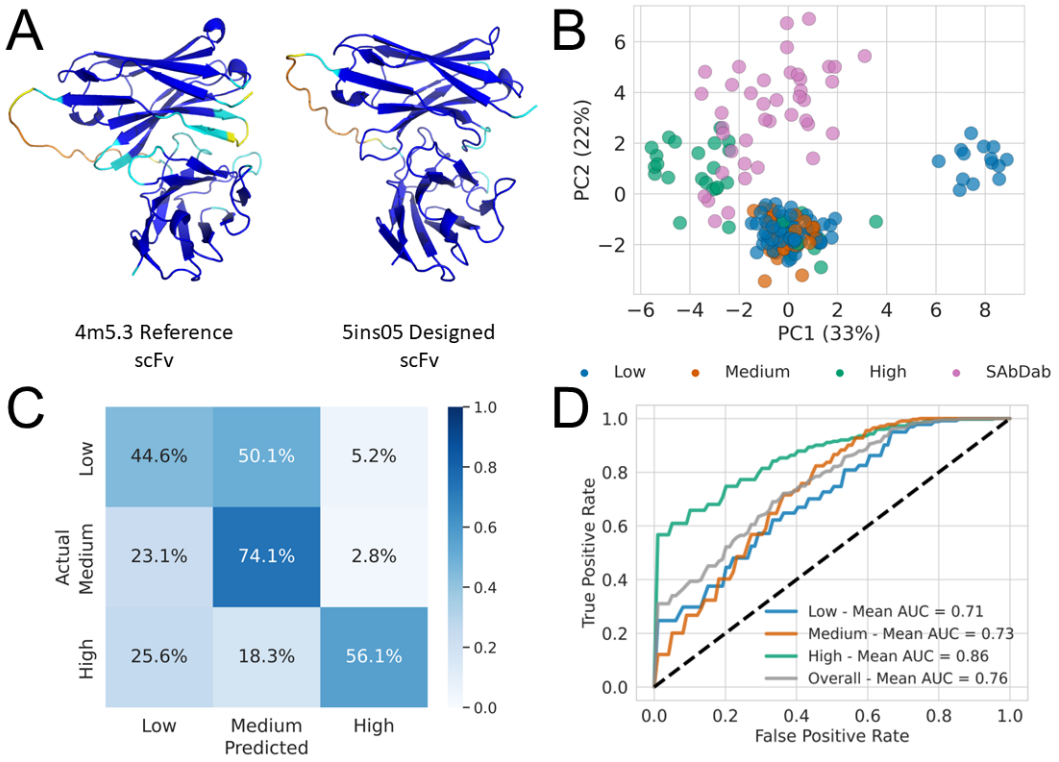

Figure 3.2: A) AlphaFold2 structural models of the reference scFv and a designed scFv from the Fleishman dataset. B) A plot of PC1 and PC2 for the scFv designs and 41 experimentally determined scFvs, along with the variance explained. C) A confusion matrix for the best classifier. D) ROC curves split out by protein production level, and an overall ROC curve, for the best classifier.

they were generated using the same basic design method, they showed vastly different levels of protein production experimentally [159]. AlphaFold2 [65] was used to generate predicted structures of these proteins, and then DE-STRESS [197] was used to calculate a set of structural descriptors. After generating this data set, dimensionality reduction was performed using principal component analysis (PCA) [254], and the major contributors to the top two principal components (PC1 and PC2) are shown in table 3.1. Plotting PC1 and PC2 against each other, shows multiple clusters that separate out protein production into low, low/medium, and high-level clusters for the designed antibodies (figure 3.2 B). The structural predictions for unrelated PDB scFvs, obtained from the Structural Antibody Database (SAbDab) [290; 291], cluster together with the highly produced designs, indicating that they share features that are related to this property. PC1 and PC2 explain over 50% of the variance in the data set, with aggregation propensity metrics [219; 226] and lysine, glycine and aspartate composition

Top contributors to PC1	**Top contributors to PC2**
aggrescan3d_min_value	composition_GLY
composition_LYS	composition_GLN
composition_GLY	composition_VAL
composition_ASP	composition_TYR
composition_PRO	composition_PRO
composition_GLN	rosetta_hbond_bb_sc
aggrescan3d_avg_value	composition_THR
composition_THR	composition_LYS
composition_GLU	hydrophobic_fitness
rosetta_hbond_bb_sc	composition_HIS

Table 3.1: Top 10 contributors to PC1 and PC2 for PCA space in figure 3.2 B. For this PCA space the robust scaling method was used, amino acid composition metrics were included, and the mutual information score was used to select features. The glossary of DE-STRESS metrics is in appendix A.

contributing the most to PC1, while glycine, glutamine, valine, tyrosine and proline composition, contributing the most to PC2. Figure 3.3 repeats the PCA analysis with the amino acid composition metrics excluded from the data set. Similarly, clusters of high and low producing designs are still observed, and the scFv structures from the SAbDab cluster together with the high producing designed scFvs. In addition to this, table 3.2 shows that aggregation propensity metrics are still major contributors to PC1 and PC2, however, there are also Rosetta energy metrics [155] capturing, backbone torsion preferences, hydrogen bonds, solvation energy and a background dependent penalty for omega dihedral angles, that contribute to these principal components.

After exploring these data sets with PCA, the levels of protein production were classified using simple naive bayes classifiers [81] (figure 3.2 C/D). The best classifier had mean Receiver Operator Characteristic (ROC) Area Under Curve (AUC) of 0.76, a mean weighted precision of 0.63 and mean weighted recall of 0.54, across the repetitions of 5-fold cross validation. In addition to this, the rest of the classifiers fitted across the folds, still performed well with mean ROC AUCs ranging from 0.73 to 0.76 (table B.2). The top performing model correctly classifies 56% of high producing designs and incorrectly predicts the rest as low or medium producing. In contrast to this,

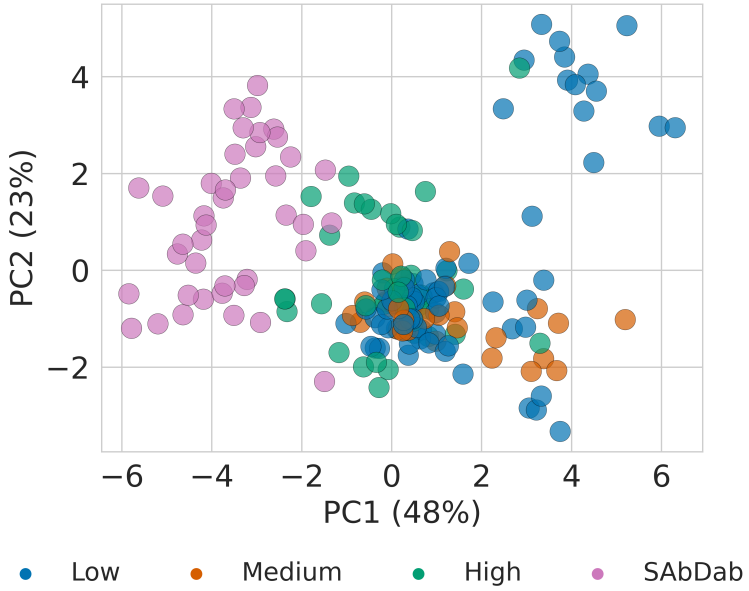

Figure 3.3: A plot of PC1 and PC2 for the scFv designs and 41 experimentally determined scFvs, along with the variance explained. For this PCA space the robust scaling method was used, amino acid composition metrics were excluded, and the mutual information score was used to select features.

Top contributors to PC1	**Top contributors to PC2**
rosetta_rama_prepro	aggrescan3d_min_value
aggrescan3d_min_value	rosetta_rama_prepro
rosetta_omega	aggrescan3d_avg_value
rosetta_hbond_bb_sc	rosetta_fa_sol
aggrescan3d_avg_value	rosetta_omega
rosetta_fa_atr	rosetta_p_aa_pp
evoef2_total	evoef2_total
rosetta_fa_sol	rosetta_fa_atr
aggrescan3d_max_value	aggrescan3d_max_value
rosetta_p_aa_pp	rosetta_hbond_bb_sc

Table 3.2: Top 10 contributors to PC1 and PC2 for PCA space in figure 3.3. For this PCA space the robust scaling method was used, amino acid composition metrics were excluded, and the mutual information score was used to select features. The glossary of DE-STRESS metrics is in appendix A.

very few of the low and medium producing designs are misclassified as high producing, a feature that would be useful when triaging protein designs to be characterised in the lab. Next, the performance of this classifier was evaluated on the test set, and it had weighted ROC AUC of 0.78, weighted precision of 0.68 and weighted recall of 0.60, showing that it generalises well to unseen data. Table B.1 in the supplementary materials shows all the validation metrics on the test sets. Considering that the success rates of designed proteins are very low (see section 1.5), this type of generalisation to unseen scFv sequences would be very valuable to protein designers, when ranking designs for experimental characterisation. Furthermore, features that were important for the prediction accuracy included certain amino acid composition metrics, hydrophobic fitness [238], aggregation propensity scores [219; 226], and Rosetta energy scores capturing solvation energy [155] (table 3.3). When excluding amino acid composition

Including amino acid composition	**Excluding amino acid composition**
composition_GLN	aggrescan3d_max_value
hydrophobic_fitness	rosetta_fa_sol
composition_VAL	aggrescan3d_avg_value
composition_ASP	rosetta_fa_atr
composition_GLU	rosetta_hbond_bb_sc
composition_PRO	aggrescan3d_min_value
composition_GLY	evoef2_total
aggrescan3d_max_value	rosetta_fa_intra_sol_xover
composition_LEU	rosetta_fa_elec
aggrescan3d_avg_value	rosetta_fa_dun
composition_HIS	
rosetta_fa_intra_sol_xover4	
rosetta_fa_sol	
composition_THR	

Table 3.3: Random forest selected features including and excluding amino acid composition metrics. The same features were found for the minmax, standard and robust scaling methods. Similar features were also found by using the mutual information score, which is shown in table B.3. The glossary of DE-STRESS metrics is in appendix A.

metrics, aggregation propensity and Rosetta solvation energy scores were still found to be important, in addition to other Rosetta energy values capturing van der waals forces, electrostatics, dunbrack rotamer preferences and an EvoEF2 total energy score [163]. Finally, similar features were also selected by using the mutual information score, rather than a random forest, which is shown in table B.3.

3.3.2 Large-scale analysis of model-derived properties performed across half a million predicted protein structures

Following on from the analysis of the small antibody data set, we decided to apply these properties to a large data set of structural models from the AlphaFold DB, to gain greater insight into how the physico-chemical properties varied between protein models. We calculated DE-STRESS metrics for 564,446 AF2 structural models, excluding amino acid composition metrics to be sure that any systematic variation was the result of structural features, and performed PCA on these properties. PC1 and PC2 accounted for over 50% of the variance from the original data set (figure 3.4 A). One observation from this scatter plot is that there were clear differences in secondary structure across this space, with α-helix rich proteins to the top of the plot, β-rich proteins to the bottom right and disordered proteins to the bottom left, while proteins in the middle had a mix of secondary structure types (figure 3.4 A/B). The main contributors to PC1 were consistent with secondary structure, including long and short-range hydrogen bond energy values and aggregation propensity scores (table 3.4). Aside from secondary structure, features associated with isoelectric point also varied across

Top contributors to PC1	**Top contributors to PC2**
isoelectric_point	rosetta_hbond_sr_bb
rosetta_hbond_sc	isoelectric_point
rosetta_hbond_lr_bb	rosetta_hbond_lr_bb
rosetta_hbond_sr_bb	rosetta_fa_dun
aggrescan3d_max_value	rosetta_hbond_sc

Table 3.4: Top 10 contributors to PC1 and PC2 for the AF2 structural model PCA space in figure 3.4 A. For this PCA space the minmax scaling method was used and amino acid composition metrics were excluded. The glossary of DE-STRESS metrics is in appendix A.

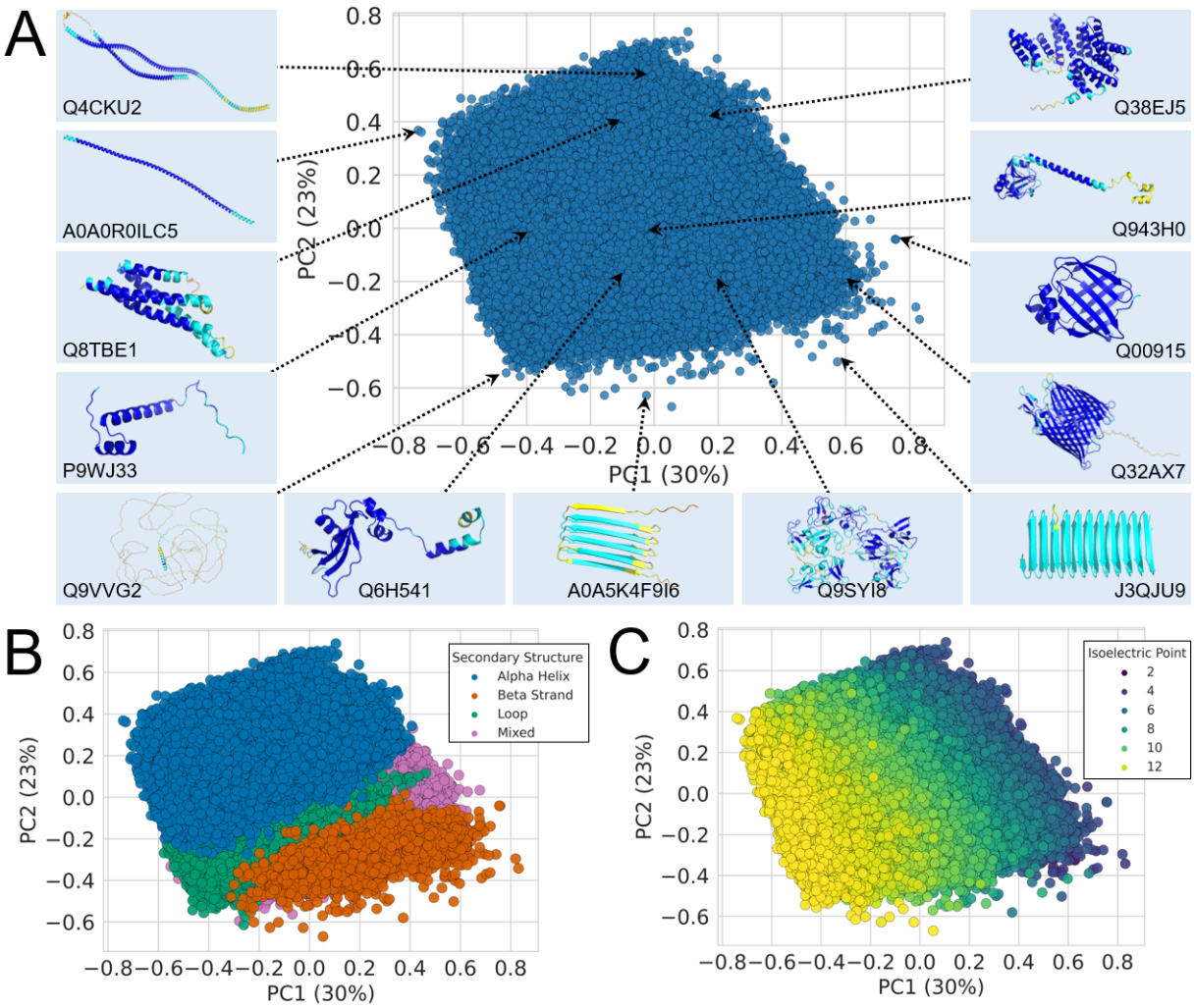

Figure 3.4: Large-scale analysis of structural features of AF2 models. Amino acid composition metrics are excluded from this analysis. A) A plot of PC1 and PC2 for 564,446 AF2 structural models from 48 model organisms, with a sample of proteins labelled around the PCA space, along with their Uniprot ID and the cartoon representation of the structure. B) The same PCA space as A) but coloured by secondary structure group. C) The same PCA space as A) but coloured by isoelectric point.

the space (figure 3.4 C), with long and short-range hydrogen bond energy, isoelectric point and rotamer energy as the main contributors to PC2 (table 3.4). Cumulative histograms for secondary structure, isoelectric point, packing density and aggregation propensity across PC1 and PC2 demonstrate these relationships clearly (figure 3.5). The y-axis of each cumulative histogram represents the proportion of data points that are less than or equal to the value in the x-axis, and they are useful to understand the shift in distribution between different sub groups in the data set.

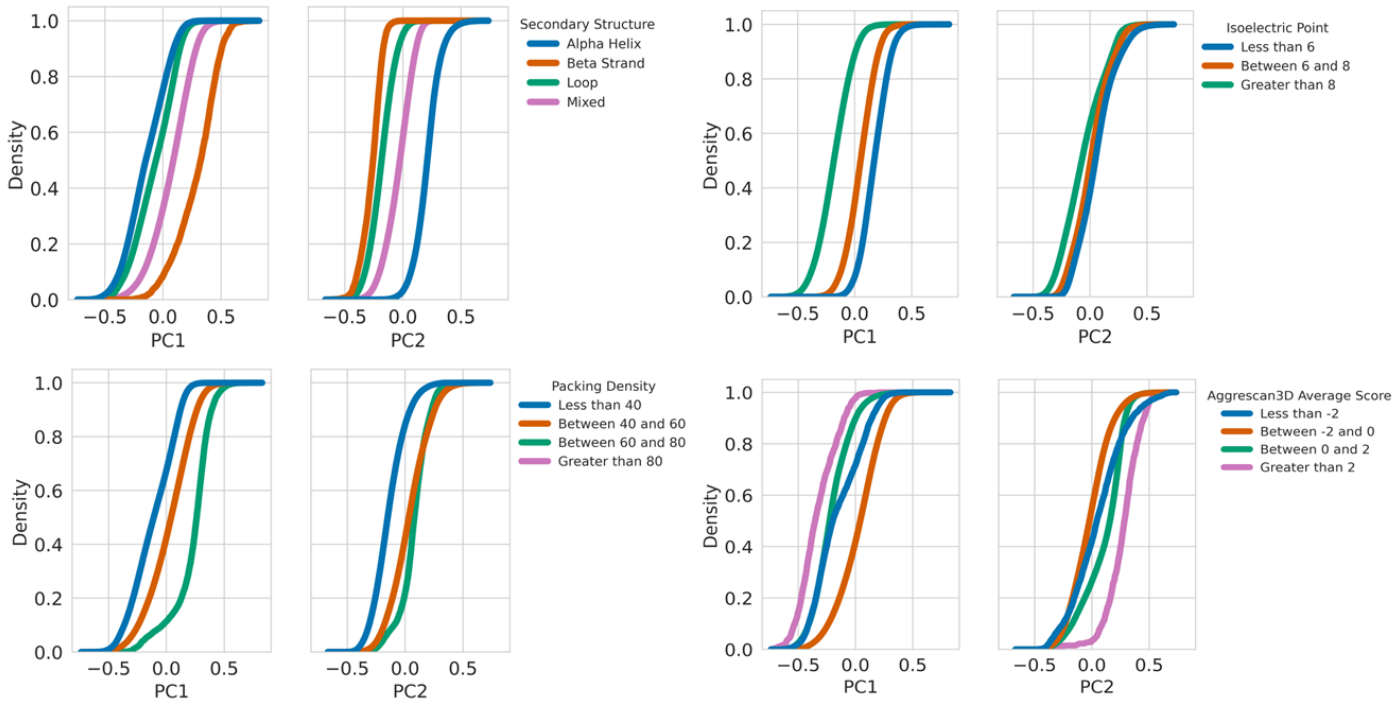

Figure 3.5: Cumulative histograms of different metrics, across PC1 and PC2, for the physico-chemical properties of 564,442 AF2 structural models in figure 3.4 A. The y-axis of each cumulative histogram represents the proportion of data points that are less than or equal to the value in the x-axis. The minmax scaling method was used for this PCA space.

Top contributors to PC1	**Top contributors to PC2**
rosetta_hbond_lr_bb	isoelectric_point
isoelectric_point	rosetta_hbond_lr_bb
rosetta_hbond_bb_sc	rosetta_hbond_bb_sc
rosetta_hbond_sc	rosetta_fa_sol
packing_density	rosetta_lk_ball_wtd

Table 3.5: Top 5 contributors to PC1 and PC2 for PDB PCA space in figure 3.6. For this PCA space the minmax scaling method was used, and amino acid composition metrics were excluded. The glossary of DE-STRESS metrics is in appendix A.

As these relationships were observed across half a million AF2 predicted structures, the analysis was repeated on 160,000 experimentally-determined structures from the Protein Data Bank (PDB) [47], to ensure that the same relationships were still

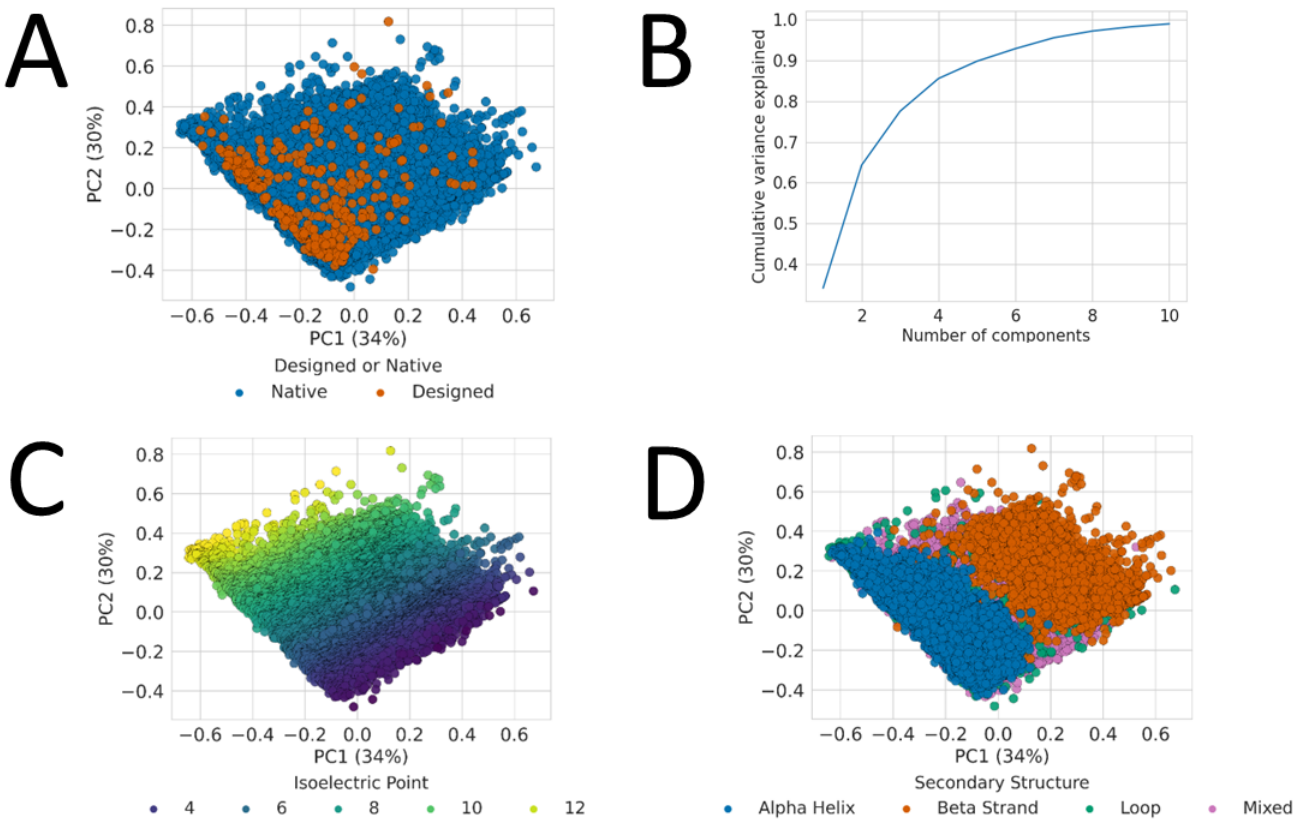

Figure 3.6: A), C), D) shows PC1 and PC2 for the physico-chemical properties of the PDB and how different metrics vary across this space, while B) shows the cumulative variance explained by number of components. For this space the minmax scaling method was used, and amino acid composition metrics were excluded.

found. Figure 3.6 shows the results from applying PCA to the DE-STRESS metrics of the PDB structures, while table 3.5 shows the top 5 contributors to PC1 and PC2, and figure 3.7 shows the cumulative histograms of different DE-STRESS metrics across these principal components. These results are consistent with the relationships found on the AF2 predicted structures, with secondary structure and isoelectric point being the major factors that vary across these experimentally-determined structures. In addition to this, figure 3.6 A shows that designed proteins, which were labelled using a curated list [107], are skewed across PC1 and PC2. From figure 3.6 D and figure 3.7, we can see that these designed proteins are skewed towards the alpha helical region of the PCA space. Furthermore, this analysis was also repeated using the robust scaling method rather than the minmax scaling method, and the results are shown in figures C.1, C.2, C.3, C.4 and tables C.1, C.2 in the appendices. Factors relating to secondary structure composition, such as the Rosetta hydrogen bond energy terms, are

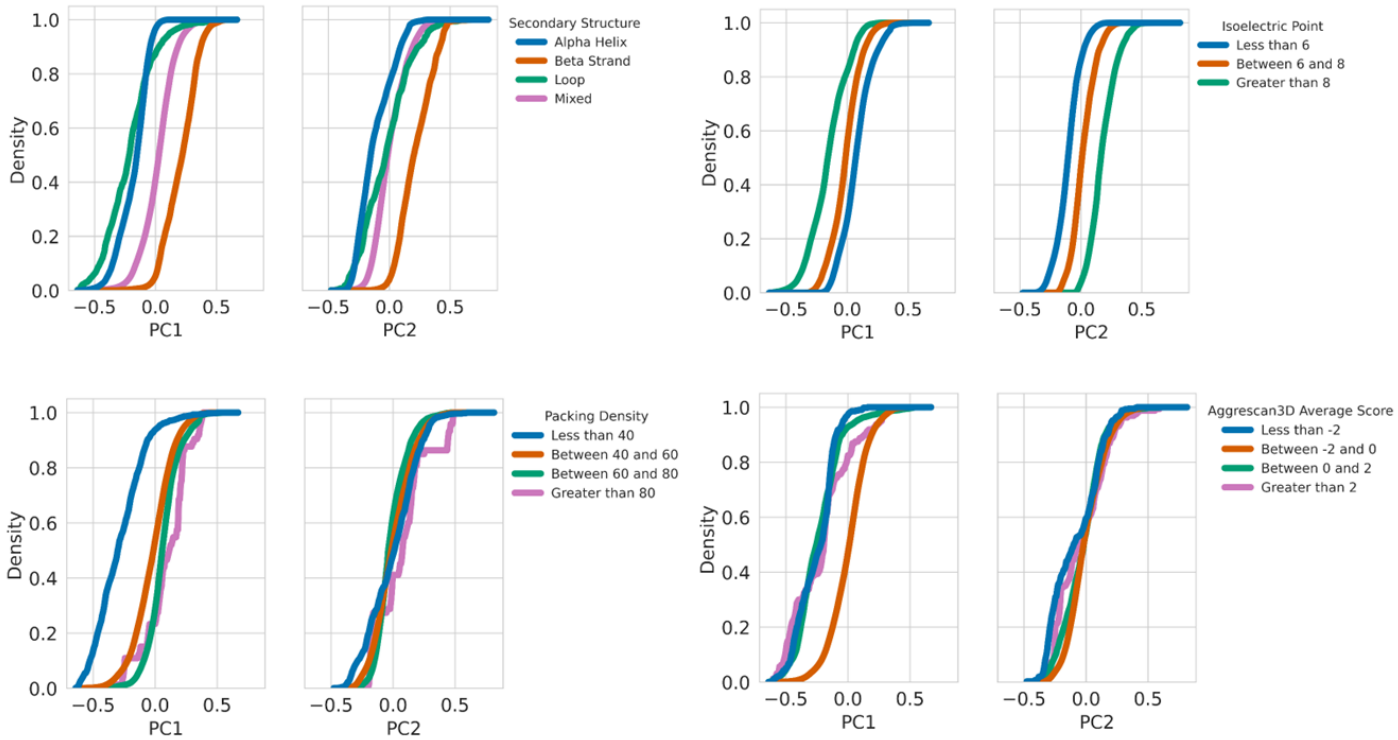

Figure 3.7: Cumulative histograms of different metrics, across PC1 and PC2, for the physico-chemical properties of PDB structures in figure S6. The y-axis of each cumulative histogram represents the proportion of data points that are less than or equal to the value in the x-axis. The minmax scaling method was used for this PCA space.

still contributors to the top principal components for both the PDB and AF2 DB, and from figures C.2 and C.4, we still see separation of secondary structure across these PCA spaces. However, isoelectric point does not appear to be a major contributor across these principal components and instead, Aggrescan3D aggregation propensity metrics and Rosetta energy scores capturing Dunbrack rotamer preferences, contribute the most to the variance across these PCA spaces. Finally, figure C.3 shows us that we still see a separation of the designed proteins in the PDB from native proteins, which is consistent with the findings in figure 3.6, where the minmax scaling method was used.

3.3.3 Model-derived properties distinguish eukaryotic and prokaryotic organisms

After observing that simple metrics, such as secondary structure composition, were major factors that varied across large structural data sets, it was then explored whether

there was systematic variation in the physico-chemical properties of proteins between organisms. Our hypothesis was that the average properties of proteins might vary by organism, and this could be an important consideration while designing or engineering novel proteins. This analysis was performed on the full data set and a culled data set, which excluded homologous proteins for each organism using MMSeqs2 [295]. This culled data set was created to remove potential bias as a result of redundancy. In addition to this, we excluded low quality AF2 structural models, with average pLDDT score $< 70\%$. Table E.1 in the appendices shows the number of proteins by organism, that were downloaded from AlphaFold DB, the volumes after using MMSeq2 to

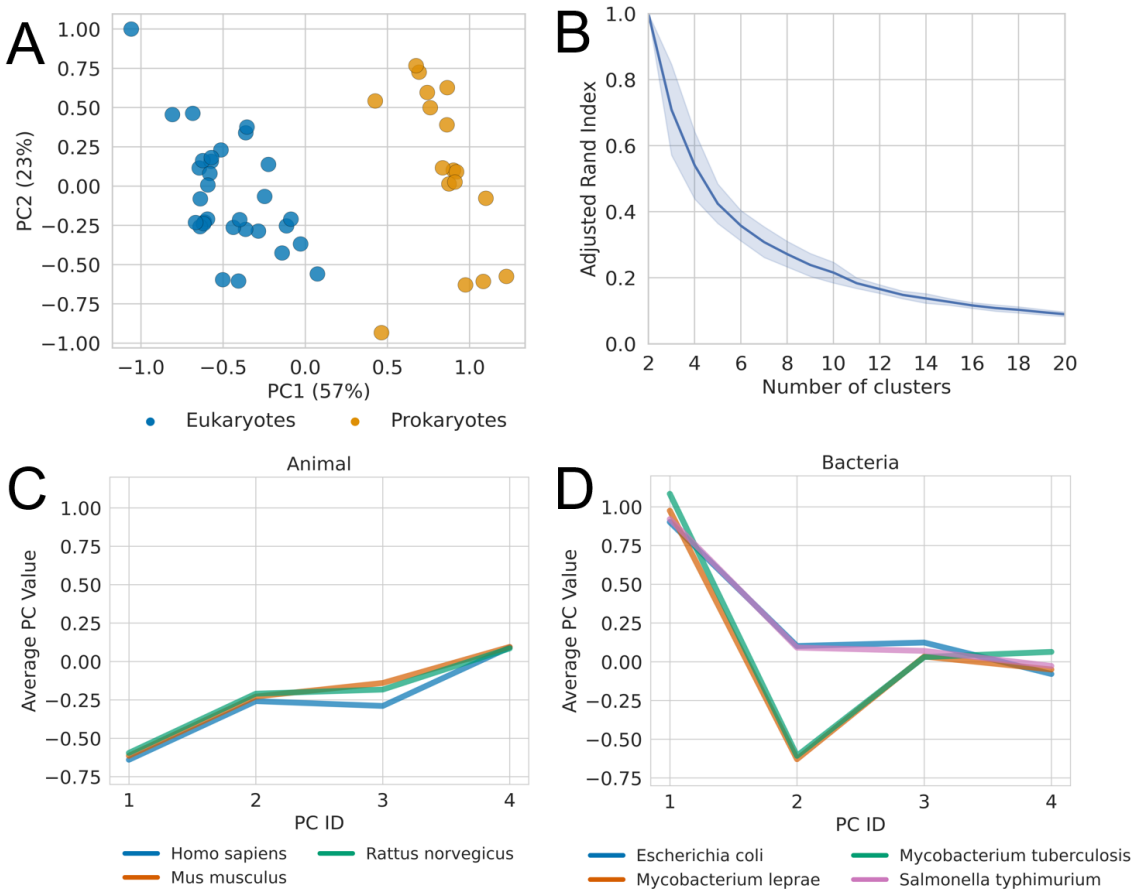

Figure 3.8: A) PCA analysis of the average model-derived properties for each organism, along with the variance explained. B) The mean and standard deviation of the adjusted rand index against the eukaryote and prokaryote groups, for 100 random initialisations of K-means, and different numbers of clusters. C) The average value of 4 principal components with a 95% confidence interval for 3 different animals. D) The average value of 4 principal components with a 95% confidence interval for 4 different bacteria.

remove redundant proteins, and the volumes after removing low quality models.

For each of these data sets, we calculated the DE-STRESS metrics for all proteins, once again excluding amino acid composition, and applied PCA to the average properties per organism (figure 3.8 A). From this plot it is clear eukaryotic and prokaryotic proteins are distinct in this space. We also performed similar analysis by applying K-means clustering [296] to the average model-derived properties and found that 2 clusters have the highest mean adjusted rand index [297], supporting this conclusion (figure 3.8 B). The main contributors to PC1 included Rosetta energy terms capturing rotamer preferences, electrostatics, and Aggrescan3D aggregation propensity scores, while Rosetta energy terms capturing solvation energy and electrostatics, were found to be the main contributors to PC2 (table 3.6). The standard scaling method was used for this analysis; however, similar results were found using the minmax and robust scaling methods shown in figure D.1 in the appendices. In addition to this, the top contributors to the principal components for minmax and robust scalers are shown in tables D.2 and D.1 in the appendices.

As we determined that eukaryotes and prokaryotes can be distinguished from the average properties of their proteins, we decided to explore this phenomenon at the organism level. Figure 3.8 C&D show the results of calculating the average PC values for different organisms, across the animal and bacteria kingdoms. Four PCs were used for this analysis, as together, they explained over 95% of the variance across the average model-derived properties. *M. musculus*, *R. norvegicus* and *H. sapiens* are shown to have extremely similar profiles across this 4D PCA space. Figure 3.8 D shows a 4-dimensional PCA space for different bacteria, which have greater variance between

Top contributors to PC1	**Top contributors to PC2**
rosetta_fa_dun	rosetta_fa_intra_sol_xover4
rosetta_fa_elec	rosetta_fa_elec
aggrescan3d_min_value	rosetta_lk_ball_wtd
aggrescan3d_max_value	rosetta_fa_dun
rosetta_hbond_bb_sc	rosetta_hbond_sr_bb

Table 3.6: Top 5 contributors to PC1 and PC2 of the average model-derived properties for each organism using the standard scaling method. The glossary of DE-STRESS metrics is in appendix A.

the average PC values for some of the bacteria. *E. coli* and *S. typhimurium* have very similar profiles, and this is also true for *M. leprae* and *M. tuberculosis*; however, there is a large variance between these two groups across this space.

3.3.4 Reconstructing the tree of life from model-derived properties

Following on from the observation that these physico-chemical properties can be used to distinguish between individual organisms, a larger scale analysis was performed across all 48 organisms in the data set. Hierarchical agglomerative clustering [298] was performed on the average model-derived properties of the organism's proteomes, using different scaling and linkage methods. It was found that the clustering results largely recreated the tree of life and captured relationships observed at a DNA sequence level from structural properties of protein models alone (figure 3.9). After obtaining these trees, we compared them to a reference tree [299], using the clustering information distance [300]. For context, a distance of 0 is an exact match and the expected value for 1,000 randomly generated trees with 48 leaf nodes was found to be 0.89 [300]. The highest ranked tree had a clustering information distance of 0.43; however, the other trees had comparable distances which are shown in table 3.7.

Qualitatively, there are broad similarities between the trees, such as the separation of prokaryotes and eukaryotes, consistent with figure 3.8 A&B, as well as details that make sense, such as clusters composed of similar organisms: the mammals *R. norvegicus*, *M. musculus* and *H. sapiens*; the fungi *P. lutzii* and *A. capsulatus*; the crop plants *Z. mays*, *O. sativa* and *G. max*; and the pathogenic enteric bacteria *E. coli* and *S. typhimurium*. However, there are also interesting differences. For example, the NCBI taxonomic tree has three main clusters, consisting of eukaryotes, bacteria and archaea, while tree created from structural properties, the archaea *M. jannaschii* is grouped together with bacteria, despite this organism being a thermophilic methanogen [302] that experiences a very different cellular environment. Furthermore, in the structural properties tree, the bacteria *M. ulcerans* is in a cluster on its own, separated from the rest of the bacteria. *M. ulcerans* is a pathogenic bacteria, and unlike the other bacteria in this analysis, produces a polyketide toxin called mycolactone and has evolved to live in a restricted environment [303], which could explain why it is separated from the other bacteria. Another difference observed in the structural properties tree is that the protozoan *D. discoideum*, groups together with the fungi, rather than the other proto-

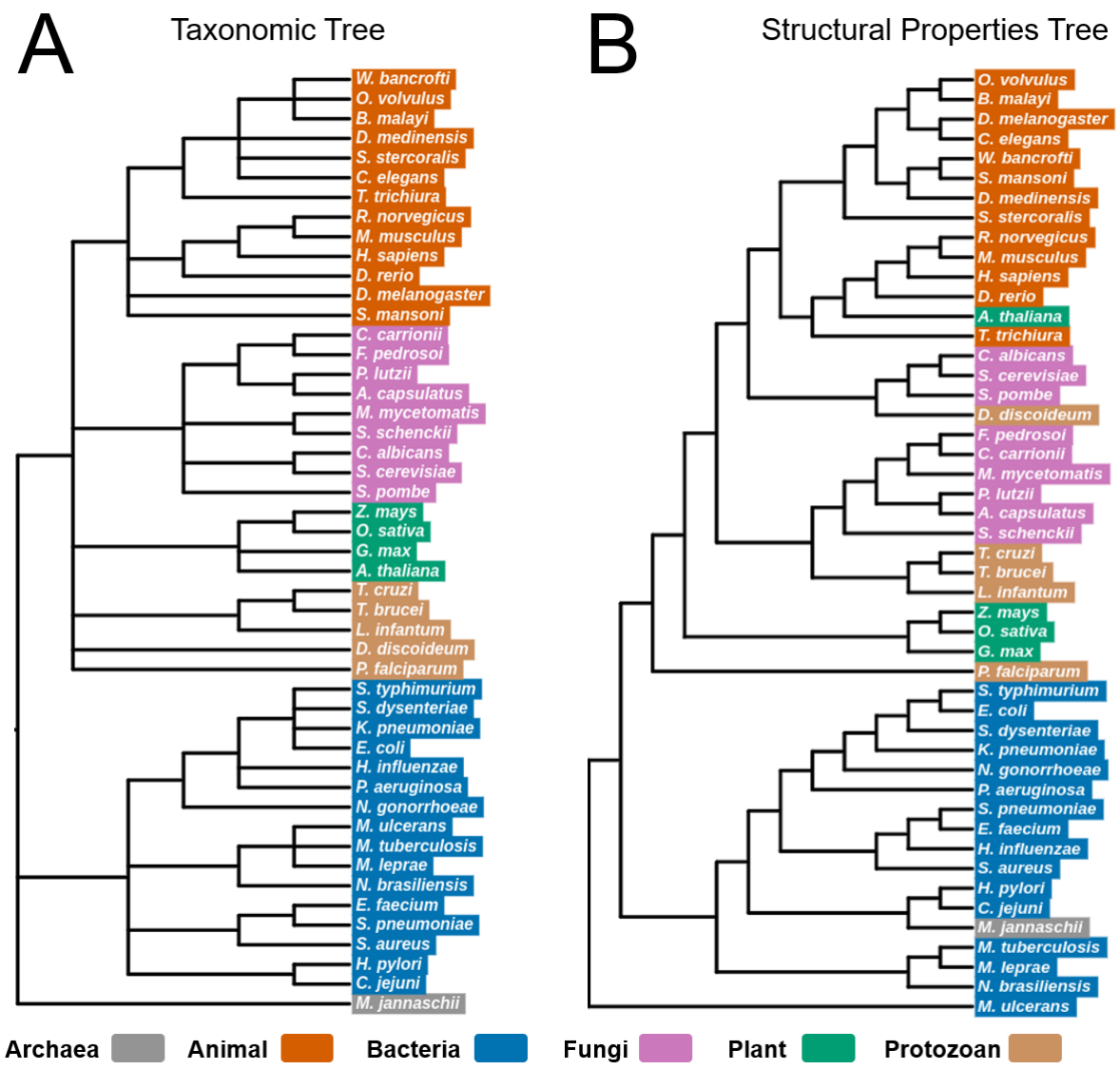

Figure 3.9: A) A tree of the 48 organisms created from the Common Tree tool from the NCBI Taxonomy Browser, which is based on a diverse set of phylogenetic information. B) The highest ranked tree created from hierarchical clustering on the average model-derived properties of each organism.

zoans. *D. discoideum* has a complex life cycle, that begins as a single-celled amoeba, which transforms into a multi-cellular slug, and finally it becomes a fruiting body that releases spores [304]. Releasing spores is a common trait of fungi, and so this could suggest *D. discoideum* clusters closer to fungi because proteins related to sporulation display similar physico-chemical properties. Curiously, one other difference observed from this tree is that *A. thaliana* groups together with animals such as *D. rerio* and *T.*

Scaling Method	Linkage Method	Clustering Information Distance to NCBI Taxonomic Tree
Minmax	Average	0.43
Standard	Average	0.43
Minmax	Complete	0.43
Minmax	Ward	0.44
Standard	Ward	0.46
Robust	Ward	0.47
Robust	Average	0.47
Robust	Complete	0.47
Standard	Complete	0.48
Minmax	Single	0.49
Standard	Single	0.53
Robust	Single	0.53

Table 3.7: Clustering information distance to NCBI Taxonomic Tree for different Scaling and Linkage Methods.

trichiura. It is not entirely obvious why this could be the case; however, *A. thaliana* does have a much smaller proteome and a shorter life cycle than the other plants included in this analysis [305; 306], and it could be possible that the differences observed in the protein physico-chemical properties of these organisms, could be related to some of these behavioural differences in the plants. Finally, from figure 3.9 we can see that there is a difference in the depth of hierarchy between the two trees, with the structural properties tree having a greater depth than the NCBI taxonomic tree. This is probably due to the difference in the data used to build these phylogenetic trees, as DE-STRESS metrics were used for the structural properties tree and DNA sequences for the NCBI taxonomic tree. Despite this difference, the primary insight from this comparison is that there is substantial consistency in how the organisms are grouped across both trees.

3.4 Discussion

With the development of novel structure prediction algorithms, we now have access to structural data that was previously unavailable, enabling us to apply a range of data

analysis techniques to gain new insights into protein structure and function. In this work, it has been demonstrated that structural properties of proteins are useful on small datasets and can be predictive of protein production. This was shown for a set of designed scFv antibodies, and using only simple methods that required little to no fitting of hyperparameters. Key structural features identified in this data-driven approach, aligned with factors that the authors explicitly discussed as being linked with low levels of expression, during the experimental characterisation of the scFv designs, such as cavities in the core of the structure (hydrophobic_fitness), unpaired buried charges (rosetta_fa_elec) and loss of long range hydrogen bonds (rosetta_hbond_bb_sc) [159]. In addition to this, aggregation propensity of designs (aggrescan3d_avg_value, aggrescan3d_min_value, aggrescan3d_max_value), were also shown to be important for separating out low and high producing designs, which makes sense as aggregation is a common failure reason for low producing designs [91]. Furthermore, these results appear to generalise beyond the initial set of designed scFv antibodies to other scFvs that have been experimentally characterised from the SAbDab, and so this could form the basis of a method to optimise protein production for applications in biotechnology.

After exploring how these physico-chemical properties varied across a small set of proteins, they were then used across large structural data sets, to uncover different kinds of relationships. Our analysis of over half a million AlphaFold2 (AF2) structural models from 48 different organisms, demonstrated that the biggest difference between proteins related to secondary structure. This supports decades of efforts to classify proteins by fold, such as the CATH [56] and SCOP [57; 58] databases. These relationships were also found when analysing 160,000 experimentally determined structures from the Protein Data Bank (PDB), which suggests that these properties vary across protein structures in general, rather than just AF2 structural models. Additionally, it was found that designed proteins in the PDB have a different distribution across these physico-chemical properties compared to native proteins, which is partly due to secondary structure composition. These designed proteins were found to be skewed towards the alpha-helical region of the PCA space, using both the minmax and robust scaling methods. This makes sense, as a lot of designed proteins in the PDB are alpha helical coiled coils, due to a lot of focus on these proteins over the years and they have been used as models for computational and experimental studies on protein folding, engineering and design [107]. However, this observation that designed proteins appear different to native proteins, across these physico-chemical properties, could be useful for improving design methods, in order to obtain more “native-like” protein designs.

Following on from this, the average properties of proteins by organism were analysed, to determine if there was systematic variation in these properties. The first result showed that eukaryotic and prokaryotic organisms were trivially separable, which is likely to be the result of known differences between eukaryotic and prokaryotic proteins, such as differences in the level of disordered proteins [307], multi-domain proteins [308] or isoelectric point [309]. Eukaryotic and prokaryotic organisms are mainly separated across PC1, which has Rosetta energy terms capturing rotamer preferences, electrostatics and Aggrescan3D aggregation propensity scores, as the main contributors. Previous research has found statistically significant differences in rotamer preferences between trans-membrane and soluble proteins [310]; therefore, these results could suggest that this is the same for proteins from different organisms as well.

Lastly, we found that even individual organisms can be distinguished by the average model-derived properties of their proteins. Initially, this was shown in figure 4C&D for a subset of animal and bacteria organisms, as we see that *S. typhimurium* and *E. coli* are both separated from *M. leprae* and *M. tuberculosis* across PC2. Rosetta energy terms capturing solvation energy and electrostatics, are the major factors that contribute to PC2, which could be an indicator of the difference in environments that these organisms have evolved in. After observing these relationships, we then explored this across all 48 organisms in the data set using hierarchical agglomerative clustering. Remarkably, the trees created from the average model-derived properties of proteins, are largely consistent with a tree generated from diverse phylogenetic information from NCBI. This is surprising, as the chemistry of proteins is functionally constrained to a much greater extent than DNA sequences [311].

As this type of analysis has only become possible over the last few years, this is, to our knowledge, the first time these relationships have been demonstrated. Since amino acid and secondary structure composition metrics were removed from the data set for the large-scale clustering analysis, the clustering is performed only on structural properties of the proteins. These results might indicate that the properties of proteins are optimised to the chemical environment of the organism, something that has been largely ignored in the fields of synthetic biology, protein engineering and protein design. Examples of this include, expressing monoclonal antibodies in Chinese hamster ovary (CHO) cells for pharmaceutical use [312], or assembling novel metabolic pathways [313]. This observation could lead to the development of more robust design methodologies that incorporate this information and increase the reliability of protein design in the future.

3.5 Next Steps

To begin with, the analysis of the scFv sequences and protein production levels, could be extended by exploring a larger data set of designs, different measures of protein production and more complex models for prediction. Naive bayes and PCA have the advantage of being simple models with no hyperparameters, which means that the chance of overfitting to the data set is a lot less than other more complex models. However, if a larger data set of designs is obtained, then further methods could be used to capture non-linear and more complex patterns, between the DE-STRESS physico-chemical properties and protein production levels. Additionally, the protein production measures that were used in this project were from yeast display, so the results could be different for other expression systems such as in *E. coli* or Chinese Hamster Ovary (CHO) cells. By exploring different expression systems, we might be able to discover factors that are predictive of protein production and expression in general, rather than in a particular organism, and potentially we could identify designs that are fundamentally flawed and would fail to be produced in any expression system. Furthermore, as only scFv sequences were considered in this analysis, we cannot make any conclusions about how the DE-STRESS physico-chemical properties relate to the protein production levels of other types of proteins. Future work could explore a diverse set of proteins, in order to understand if there are any common factors that affect the production levels of these different proteins. This additional work would make this analysis more robust, and could provide models that are useful for ranking designs for experimental validation.

In relation to the analysis of these properties across large structural data sets, further work could be done to extend this analysis to look at protein function. The work in this chapter showed how these properties varied across two large structural data sets, and how they varied across 48 different organisms; however, exploring how these properties vary by protein function, could be very useful for design purposes. Functional annotations could be joined onto the 500,000 AF2 structural models and the 160,000 PDB structures from UniProt [314], or predictive models such as CLEAN [315] could be used to predict enzyme function. Similar analysis could be performed, to understand how well function is clustered across this data set, or classifiers could be built to predict function based on the DE-STRESS physico-chemical properties, in order to understand which properties are predictive. This analysis could be extremely useful for design, as these physico-chemical properties can then be used for

ranking de novo protein designs for testing in the lab. Moreover, the PCA spaces that were fitted to the physico-chemical properties of the AF2 structural models and the PDB, could be used for mapping de novo designs into these spaces, and comparing the designs to a reference set of proteins. As these spaces capture a huge amount of information about the properties of proteins, this could provide us with a method for evaluating designs. On the other hand, PCA only captures linear relationships, so it could be worthwhile exploring other methods such as Uniform Manifold Approximation and Projection (UMAP) [255] or Gaussian Process Linear Variable Models (GPLVMs) [316], to learn latent spaces that capture more complex relationships between the properties of proteins. Overall, this additional work could help understand how these physico-chemical properties are related to protein function, and how they can be used to rank protein designs for taking into the lab.

Another future direction that this work could take, is exploring how these physico-chemical properties are related to different groups of proteins. The CATH classes provide a hierarchy of protein structures based on factors such as secondary structure [56]. The results in this chapter showed that across large structural data sets, secondary structure is one of the major factors that varies, which agrees with the top level of the CATH hierarchy, which are the classes, Mainly Alpha, Mainly Beta, Alpha Beta, Few Secondary Structures and Special [56]. However, in future work we could join on the CATH classes to the PDB and AlphaFold DB data sets, and filter by specific CATH classes, in order to see how these physico-chemical properties vary, when secondary structure is fixed. For example, we could select all Alpha Beta structures from these data sets and use dimensionality reduction methods like PCA on the DE-STRESS metrics, to see if the proteins will cluster by fold, as this is the next level in the CATH hierachy. After this, we could select a particular fold such as a TIM barrel and repeat the same analysis, to understand whether the proteins would then group by enzyme substrate specificity. By using the CATH classes in addition to the DE-STRESS metrics for these large structural data sets, we could gain more insights into how these factors vary across different subsets of proteins. This could be extremely useful for understanding which factors are important for different types of proteins, which could better inform design methodologies.

Finally, the organism clustering results, which showed that we could reconstruct the tree of life from only properties of proteins, could be explored further in future work. One improvement to the hierarchical clustering analysis could be calculating the intra and inter cluster distances to understand whether the difference is statistically

significant and to ensure that the results are robust. Additionally, in this chapter we showed that there was systematic variation in the properties of proteins, across a large amount of proteins from 48 different organisms from the AlphaFold DB; however, this could also be explored for a specific protein family, such as immunoglobulins, or even a single protein that is present across a large number of organisms. As a result of this, we could explore whether we observe the same systematic variation in properties across protein families or even single proteins. In addition to this, the same analysis could be performed by subcellular location, by joining this label onto the PDB and AlphaFold DB data sets from UniProt, and performing a similar analysis to the organism clustering. This would allow us to understand how these physico-chemical properties vary across different areas in the cell, such as the cytoplasm, membrane and the nucleus, which could be very useful for designing proteins for functions in these areas. Furthermore, sequence design methods could be trained with the host organism or subcellular location included in the labelled data, or reference sets could be created of proteins from a particular organism or sub cellular location, in order to include information about the molecular environment of the protein into the design methodology. By including this information into the design of novel proteins, this could potentially help to reduce the failure rate and improve the reliability of protein designs.

3.6 Conclusion

In conclusion, this chapter explored how physico-chemical properties, calculated from predicted protein structures, can be used to understand *in vivo* properties of proteins. Firstly, the DE-STRESS physico-chemical properties were calculated for AlphaFold2 structural models of a set of 192 designed scFv sequences, and these properties were shown to be predictive of *in vivo* protein production. Features that were predictive of protein production, such as packing quality of the proteins and energy functions capturing electrostatics and hydrogen bond energy, agreed with author reported reasons of failure for these designs, in addition to aggregation propensity metrics, where protein aggregation is a common reason for the failed production of proteins. In addition to this, these relationships appeared to generalise to unrelated experimentally-determined scFv structures. After exploring these properties on a small data set of scFv designs, we then explored how they varied across a large data set of 500,000 AlphaFold2 structural models, and 160,000 experimentally-determined structures from the PDB. DE-

STRESS structural features associated with secondary structure were found to be some of the major factors that varied across both of these large data sets, which agrees with decades of work to categorise proteins by fold. Across the physico-chemical properties of the PDB, designed proteins were shown to look different to native protein structures, which could provide insights for improving design methodologies to make more "native-like" proteins. Furthermore, systematic variation was found in the properties of proteins from 48 different organisms, to such an extent that the tree of life could be recreated from these properties. To our knowledge, this is the first time that this relationship has been shown, and it could not have been demonstrated until the recent advancements in structure prediction, and the large amounts of protein structural data that is now available as a result of this. This result suggests that the properties of proteins are optimised to their unique molecular environment, which has been largely ignored in the fields of protein engineering and design. Therefore, this could be used to develop more robust design methodologies in the future, that incorporate information about the environment of the designed protein, to ultimately lower the failure rate of protein design and make it more accessible as a tool for researchers. Future work could explore expanding the results from predicting the protein production levels of scFv designs to different proteins and expression systems, exploring how these physico-chemical properties of proteins are related to function across large structural data sets, and finally incorporating information about the host organism, or subcellular location of proteins, into design methodologies, to see if this will improve the success rate of designs.

Chapter 4

Cell-free expression systems for producing designed scFvs

Although the majority of this thesis so far has focused on computational methods for evaluating designed proteins, this chapter changes focus to exploring experimental methods, for helping to reduce the cost and failure rate of protein design. Cell-free expression systems offer a unique opportunity to understand the factors that affect the production of proteins, in a controlled environment and without the constraint of sustaining life. This makes them suitable for exploring some of the reasons why designed proteins fail to be produced, if these reasons are related to transcription and translation. In this project, cell-free systems were set up to produce single chain variable fragment (scFv) antibody designs, in order to collect expression data and gain an understanding of the factors that affect the production of protein designs. As a result of this work, a protocol was developed for collecting scFv protein production levels in *E. coli* cell-free systems. In the future, this protocol will be used for high-throughput screening of a library of scFv designs, using acoustic liquid-handling robots at the Edinburgh Genome Foundry, in order to collect an expression data set for these designed proteins. Following on from this, models could be trained to predict this *E. coli* cell-free protein production measure, using the DE-STRESS metrics of the designed scFv structural models as features. These models could then be used to rank *de novo* scFv designs for experimental testing, and these could be validated in the same cell-free systems. Finally, the composition of these reactions for low producing and high producing designs, could be investigated further using techniques such as mass spectrometry, to gain an understanding of some of the reasons why designs fail, in order to improve the reliability and accessibility of protein design.

4.1 Background and motivation

To begin with, this section provides an overview of cell-based methods, which are the traditional and most widely used methods for protein production. After this, cell-free systems will be introduced, with a description of how they are created, and the advantages and disadvantages, over traditional cell-based methods for protein production. Following on from this, there will be a small section detailing single chain variable antibody fragments (scFvs), as these proteins were used for the experiments in this chapter. Finally, a literature review will be presented on cell-free systems for protein production, and a section on how cell-free systems could be used to reduce the failure rate of protein design.

4.1.1 Cell-based systems for protein production

Generally, most proteins that are produced in industry or academic labs, are produced in cell-based systems, where a gene is constructed that encodes for the protein of interest, placed in cells such as bacteria, yeast, insect, mammalian, or even plant cells, and then the native transcription and translation machinery is used to produce the protein [186; 317; 318; 312; 319]. These methods are used because they are relatively cheap, they can produce large quantities of the protein of interest, and they can produce a huge variety of different proteins [320; 321]. Figure 4.1 shows an overview of how these cell-based systems are used for protein production. Firstly, the DNA construct that encodes for the protein of interest is designed, along with the necessary promoters, terminators and ribosome binding sites, that can be recognised by the transcription/translation machinery in the cell [322]. In addition to this, the DNA sequences are codon optimised, as different organisms have different bias in their codon usage, which can have a huge impact on protein production [323]. After this, the DNA constructs are introduced into the cytoplasm of the cells through a molecular biology technique called a transformation, the cells are then used to produce the protein of interest, and the protein is purified from the cells. Overall, there are a large amount of therapeutics, such as monoclonal antibodies and vaccines [324; 325], and other biologics that are used in industrial processes, such as proteases and amylases [326], which are produced in high quantities in cell-based systems.

Although these cell-based systems are extremely useful for producing a wide range of different proteins, these systems also have a number of big disadvantages. One major disadvantage of these methods, is that they generally suffer from a lot of variability

in protein production levels [327; 328]. Some of this variation in production levels is caused by external factors, such as the growth media used [329] and technical failures and changes in the environment during production [330]; however, there is even general instability in the cell-lines that causes a huge amount of variation [331]. As a result, this variability in protein production levels from cell-based systems, can cause major issues for companies producing biologics. For example, Tharmalingam et al. explored the genotypic and phenotypic diversity in subclones of a Chinese Hamster Ovary (CHO) clonal cell line, and showed that protein production levels in sub clones varied as much as $\pm 40\%$ (figure 4A in Tharmalingam et al. [332]). This is significant as CHO cells are widely used for producing therapeutics, such as monoclonal antibodies and vaccines [312], and this level of variability in protein production levels can cause increased costs and supply shortages [333]. Another disadvantage of using these systems for the production of designed proteins, is that due to a large amount of

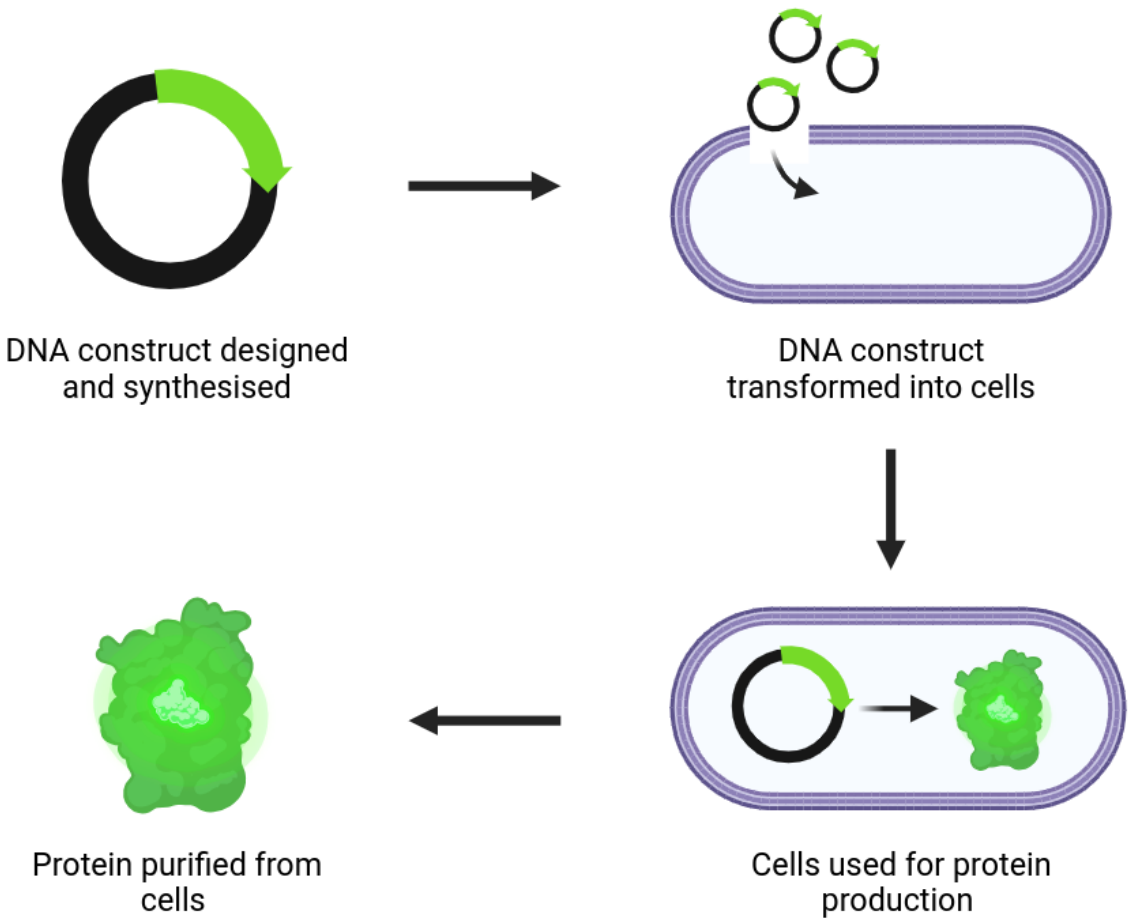

Figure 4.1: An overview of how cell-based systems are used for protein production. Created with BioRender.com.

background noise in living cells [334], it can be extremely difficult to understand why designed proteins fail to be produced. In addition to this, there can be many reasons that designed proteins fail, such as aggregation, misfolding, and even toxicity of the designed protein to the cells producing them [137], which may all require different solutions to avoid these failure reasons. As these systems are the main methods for producing *de novo* proteins, and the majority of designs fail to be produced in these systems [137], there is a need for a greater understanding of the main reasons why they fail to be produced.

4.1.2 Cell-free systems for protein production

In contrast to cell-based systems, cell-free systems offer a flexible way to study protein expression, and provide a unique way to control the expression environment [190; 191; 192; 193]. These cell-free systems include crude cellular extracts [191], or reconstituted systems of enzymes such as the PURE system [195; 196], for the activation of

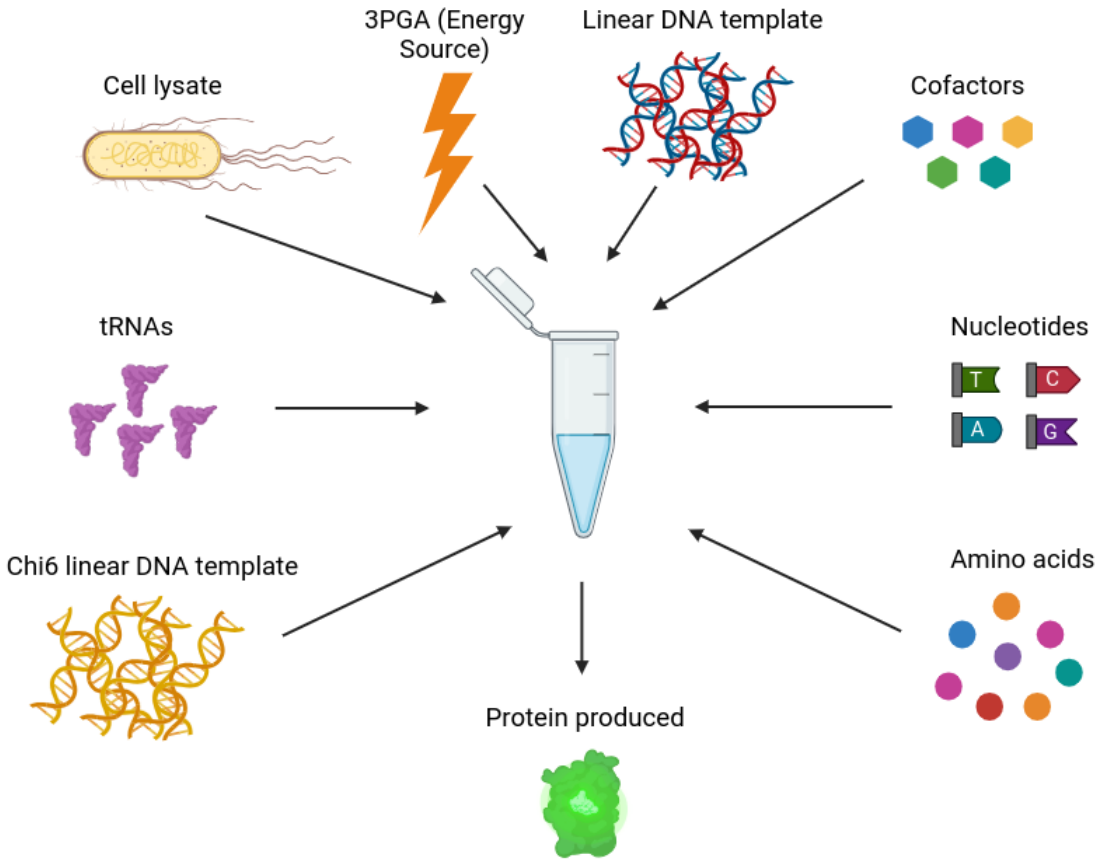

Figure 4.2: An overview of a crude lysate cell-free system for protein production. Created with BioRender.com.

transcription and translation, without the need for intact and living cells [193]. In order to make the cell lysates used for the crude lysate cell-free systems, cells are generally grown in media for a period of time, lysed open using methods such as sonication, and then the soluble portion of this lysate is extracted, which contains the components needed for transcription and translation [191]. Figure 4.2 shows an overview of how crude lysate cell-free systems can be used for protein production. Firstly, cell lysates from *E. coli*, *S. cerevisiae* or another organism, are used to obtain the transcription and translation machinery that is needed for protein production, such as the ribosomes and polymerase. In addition to this, the reaction is supplemented with the other components that are needed for transcription and translation, such as extra nucleotides, amino acids, cofactors, tRNAs, and an energy source (3PGA), in order to achieve as much protein as possible. A lot of these components are usually grouped together into an "energy solution" when assembling reactions [335], and tables 4.8 and 4.9 show the full list of components for the energy solutions used in this project. Following on from this, the DNA template of the protein of interest, which can either be linear or a plasmid, is added into the reaction as well. As there are DNases (proteins that can degrade linear DNA such as RecBCD) present in the cell lysate, components such as chi6 linear DNA template [336] or proteins that inhibit the DNase activity such as GamS [337], can be added into the reaction to stop the degradation of a linear DNA template of interest. Finally, additional components such as T7 RNA polymerase, which can be used to express sequences with T7 promoters [338], or components that are needed for post translational modifications, such as phosphorylation and glycosylation [193], can also be included in the cell-free reactions.

One major benefit of cell-free systems over cell-based methods, is that they do not have the constraint of sustaining life, and the resources of the system can be used for protein production only [193]. Cells perform a wide range of different biological processes, for example reproduction, which use up resources and introduce background noise when using these systems to produce proteins. In contrast to this, cell-free systems offer an alternative method for protein synthesis, that allows much more control over the sources of variation [339], which can lead to reproducible levels of protein produced for a particular cell-free system. As a result of this, cell-free systems provide a better method for comparing the production levels of different proteins, without the confounding factors of different growth rates of cells, and sources of variation from other cellular processes. Another benefit of cell-free systems over cell based methods for protein production, is that cell-free offers an incredible amount of control over the

expression environment [190]. As the cell-free reactions are performed without the barriers of cell walls, this allows controlled modification of the reaction conditions [194], direct measurement of factors such as transcription and translation [340], and even the combined production and functional analysis of proteins [341]. Consequently, cell-free is better suited for performing detailed analysis on transcription, translation, and for understanding some of the reasons why protein designs fail to be produced. However, this could only be used for understanding failure reasons that are associated with transcription and translation, and would not help us investigate other reasons for failure, such as the toxicity of proteins to cells.

One more advantage of cell-free systems over cell-based methods, is that they are suitable for high-throughput screening of designed proteins, as they can be performed in very small volumes and the results can be obtained relatively fast [342]. In order to do this, acoustic liquid-handling robots or microfluidics can be used to assemble a large number of cell-free reactions, for testing a library of designed proteins [343] or genetic parts for genetic circuit development [344]. This can significantly reduce the time for screening different designed proteins for a specific application and can generate a large amount of data. However, unfortunately, a major disadvantage of cell-free systems over cell-based systems, is that they suffer from relatively low yields of the amount of total protein produced, in comparison to cell-based methods [190]. On the other hand, the yield is also dependent on the protein produced as well, as some proteins are easier to produce than others in both cell-based and cell-free methods, and additionally some proteins that are toxic to cells, can have higher yields in cell-free than cell-based methods [345]. In general, this relatively low yield of cell-free systems limits their usability for some applications; however, when screening designed proteins, we are only interested in relative yield or function, and therefore cell-free is still well suited for screening libraries of protein designs. In contrast to this, the yield of cell-free reactions is increasingly being improved, and some proteins have been produced in high yields from cell-free systems [346; 347]. Furthermore, cell-free systems have been developed that can even assemble bacteriophage viruses [128], which shows how cell-free systems are being applied to larger proteins and protein complexes.

4.1.3 Single chain variable fragments (scFvs)

Antibodies, or immunoglobulins, are a specific type of protein that are used by the immune system to neutralise foreign objects, known as antigens, in the body. These

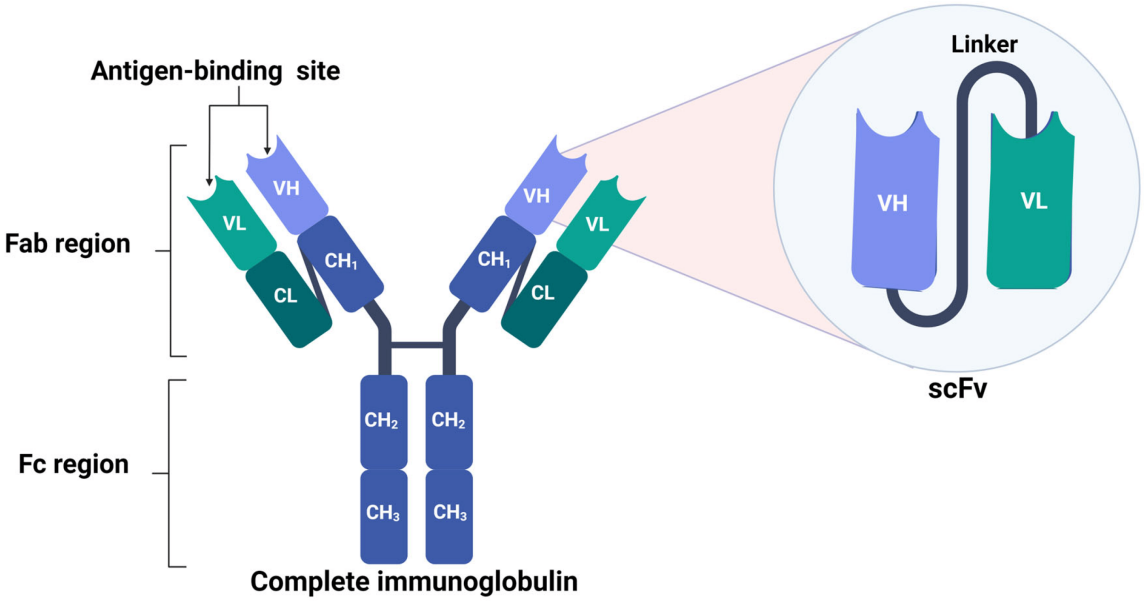

Figure 4.3: A complete immunoglobulin shown on the left and a scFv shown on the right. The scFv is constructed from the variable heavy (VH) and variable light (VL) regions from the complete immunoglobulin, and joined together with a short linker. This figure is taken from Rodríguez-Nava et al. [348], which has a Creative Commons Attribution (CC BY) license.

proteins have a large number of applications and can be used to treat many diseases such as cancer [349; 350]. In 2019, monoclonal antibodies accounted for 9 of the top 20 therapeutics by sales, with cumulative earnings of $75 billion that year [351]. Figure 4.3 shows an image of an immunoglobulin on the left and a single chain variable fragment (scFv), which is an engineered protein that is made up from fragments of the complete immunoglobulin. Structurally, immunoglobulins consist of two main regions, the fragment crystallizable region (Fc) and the fragmented antigen binding region (Fab) [348], which are both shown in figure 4.3. These two regions are made up from two polypeptide chains, known as the light and heavy chains (shown in green and blue in figure 4.3 respectively). In addition to this, these light and heavy chains have variable regions (VL and VH) and also constant regions, CL for the light chain and CH_1, CH_2 and CH_3 for the heavy chain [348]. In order to construct a scFv, the VH and VL variable regions, which are responsible for binding to an antigen, are linked together with a short linker sequence [348]. To begin with, scFvs are of interest to researchers, as they can be useful for understanding properties and functions of larger monoclonal antibodies [352]; however, designed scFvs can also be used for other ap-

plications, such as in immunochromatographic strips for the detection of toxins [353], and for cancer treatments [354]. Understanding the reasons why designed scFvs fail to express, and being able to predict successful expression from the scFv design, would be extremely useful for developing these antibody based therapeutics and diagnostics. Gaining this understanding, could help to make the design process more efficient, faster and reliable in the future.

4.1.4 Reducing the failure rate of designed proteins

Currently, designed proteins have a high failure rate, which limits the reliability and accessibility of protein design to researchers. As described above, cell-free systems offer a powerful and flexible way for understanding protein production, and for studying the factors that affect transcription and translation. In addition to this, they provide a level of control over the expression environment, that is not possible in cell-based methods for protein production. For these reasons, cell-free systems offer a unique opportunity, to gain an understanding of why the majority of designed proteins fail, with the aim of using this insight to inform future design methods.

In this project, *E. coli* cell-free systems were used to produce a set of designed scFv antibodies [159], in order to collect expression data. This set of designed scFvs was chosen because some of the designs had very low expression in yeast, while some had very high levels, which makes this library suitable for studying the factors, that might cause low protein production levels. In addition to this, as described above, scFvs have a range of applications as therapeutics and diagnostics, which means that finding out the main factors that cause low production levels, could help make these designed proteins easier and cheaper to produce for these applications. Furthermore, there have been many studies that have produced scFvs in both eukaryotic and prokaryotic cell-free systems [355; 356; 357; 358; 359]. In order to to produce these scFv proteins in an *E. coli* cell-free system, different strains of *E. coli*, such as SHuffle and Rosetta gami, are used to make the cell lysate, as these strains are better for producing eukaryotic proteins, such as antibodies [357; 359].

To begin with, a method for collecting expression data from this set of scFv designs was developed, which involved attaching a fluorescent protein, such as Green Fluorescent Protein (GFP) or mCherry, to the scFv designs, and measuring the fluorescence levels of the cell-free system over time using a plate reader. Using fluorescent proteins in this way, is a well used technique for high-throughput measurement of expression

levels [199]. The majority of this work in this chapter, involved setting up the *E. coli* cell-free systems and optimising the signal to noise ratio of the expression levels, for a single scFv design. This is important as when performing cell-free reaction using *E. coli* lysate, there is background fluorescence from the cell-free reaction components [360], which could mask the signal from the scFv and fluorescent protein complex. As a result of this work, this protocol can be used, in the future, for high through-put screening to collect expression data for a larger set of designed proteins, and to explore some of the main reasons of failure for these designed proteins.

4.2 Methods

This section details the experimental methods that were used in this PhD project, for performing cell-free reactions on scFv antibody sequences. Firstly, an overview of the experimental set-up will be described, along with the library of designed scFv sequences that were used in this project. Following on from this, a section will list the different template linear DNA sequences that were used to set up the cell-free reactions. Finally, the standard lab protocols for preparing DNA, cloning DNA constructs into plasmids, preparing the components of the cell-free systems, and performing the cell-free reactions, are described.

4.2.1 Overview of experimental set-up

The library of Fleishman scFv designs used in this project, is based on the set of 192 sequences designed by Baran et al. [159], where they experimentally validated the scFv sequences using yeast display. In order to set-up the yeast display experiments, the 4m5.3 anti-fluorescein scFv (with PDB ID 1X9Q), was used as a test construct in this study, and similarly this sequence was also used as a control for the cell-free systems in this project. Fusing the scFv designs to a fluorescent protein, such as a mutant of Green Fluorescent Protein (deGFP) which is optimised for cell-free reactions [361], or the red fluorescent protein mCherry [362], allows us to monitor the expression of the scFv sequences. This is a well used technique for high-throughput fluorescence based screening for protein expression [199], and fluorescence can be measured quantitatively by using a plate reader, in Relative Fluorescence Units (RFUs), which can allow us to infer levels of expression and protein production.

4.2.2 Template linear DNA sequences

In this project, various template linear DNA sequences were used, in order to set-up the cell-free systems and to collect expression data for the 4m5.3 scFv design. Figure 4.4 shows an overview of how these sequences were constructed, while table H.1 in the appendices provides the full linear DNA sequences (promoter, coding domain and terminator regions), along with the origin of these sequences. In addition to this, table H.2 shows the promoter, terminator and lac sequences used in this project and table H.3 shows the translated amino acid sequences for deGFP, mCherry, 4m5.3 scFv and the linker sequence.

Figure 4.4 shows two GFP sequences with different promoters, p70a-deGFP and T7p14-deGFP [363], which were used as positive controls in this project. The p70a promoter is a promoter specific to *E. coli* sigma factor 70 [364], while the T7 promoter is an efficient bacteriophage promoter [338] which requires T7 RNA polymerase to be induced in *E. coli* cells during growth [191]. Both of these GFP sequences were used to ensure the cell-free systems were functioning correctly, while testing new constructs.

Next, figure 4.4 shows the different 4m5.3 scFv DNA constructs that were used in this project. Initially, the scFv sequence was ordered as p70a-4m5.3-deGFP, which uses the native *E. coli* promoter and is linked to deGFP by a short Glycine-Serine and Proline-Alanine (GSPA) linker sequence (GGGGSPAPAPP) [365]. The amino acid sequence of the 4m5.3 scFv design was taken from the Baran et. al [159] supplementary materials and linked to the amino acid sequence of deGFP by using a short linker sequence. Before obtaining expression data for this sequence, it had to be reverse translated into a DNA sequence. However, due to the degeneracy of codons, there are many DNA sequences that could be translated into the same amino acid sequence [366]. In addition to this, the choice of DNA sequence depends on the organism that is being used to express the sequences, as different organisms have different codon usage bias. This is extremely important to consider as DNA sequences that contain codons which are rare in an organism, will be much more difficult to express in that particular organism. As a result of this, the Integrated DNA Technologies (IDT) website, https://eu.idtdna.com/pages/tools/codon-optimization-tool, was used to reverse translate the amino acid sequence, and it was codon optimised for *E. coli* B. Finally, this construct was ordered as a linear DNA sequence by ordering a gBlock from IDT.

After testing the p70-4m5.3-deGFP in cell-free reactions, the 4m5.3 scFv construct was ordered as a clonal gene, with a T7 promoter in the pET24(+) plasmid from Twist

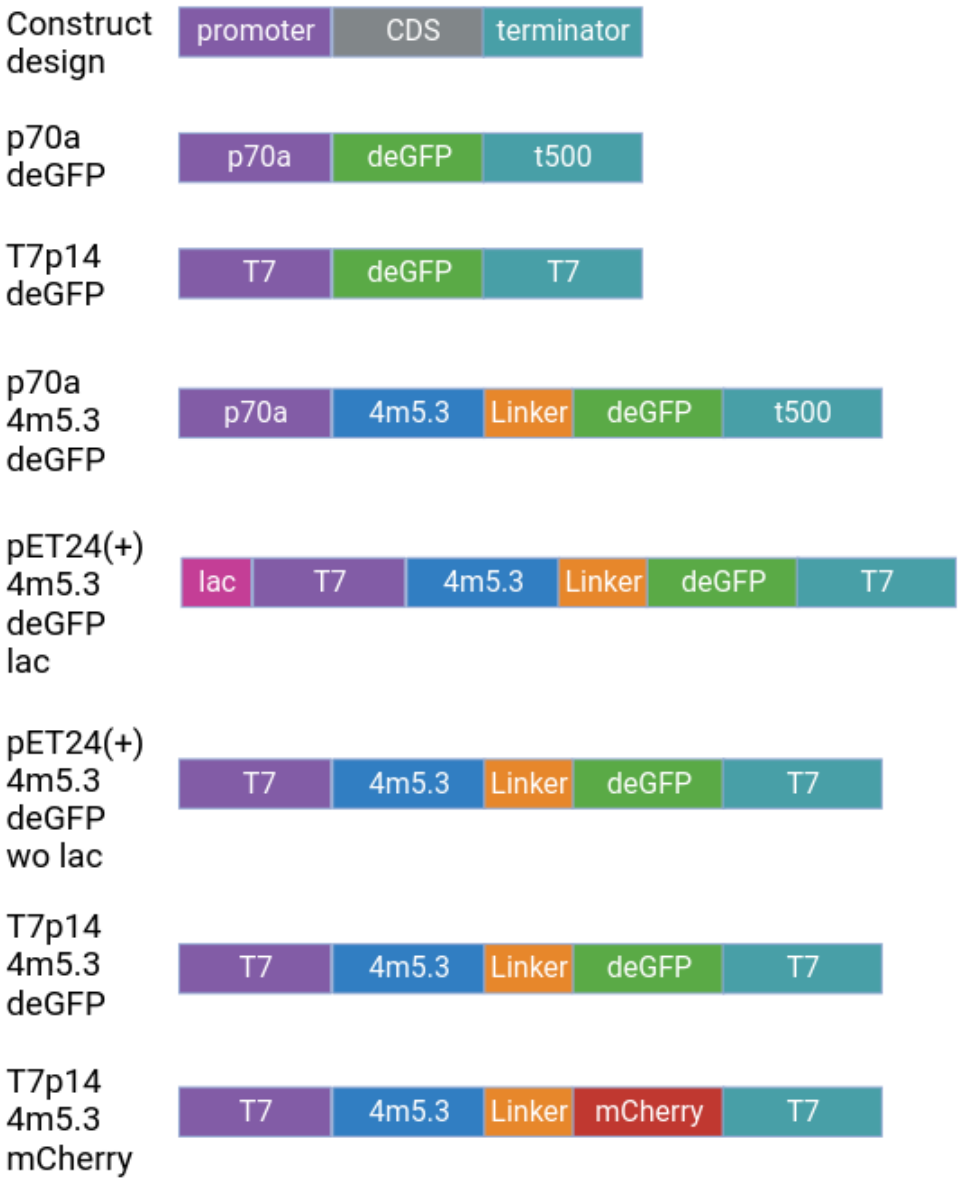

Figure 4.4: Overview of the different template linear DNA constructs used in this project, along with the promoter and terminator regions. The linker that was used in this project was a short Glycine-Serine and Proline-Alanine (GSPA) linker sequence (GGGGSPA-PAPP).

Bioscience. Ordering the Fleishman scFv library as gBlocks from IDT would have been prohibitively expensive; however, ordering the library as clonal genes from Twist, allowed us to order more of the Fleishman scFv sequences. The T7 promoter was chosen for these DNA constructs as using an *E. coli* native p70 promoter, would have caused issues in the production of the clonal genes for Twist. In addition to this, a lac operator [367] was also added to the sequence, which provides further control for

Twist during the production of these clonal genes. As we were unsure whether this lac operator would interfere with the cell-free reactions, the DNA constructs were ordered with and without the lac operator. Figure 4.4 shows these two DNA constructs as pET24(+)-4m5.3-deGFP-lac which has the lac operator and pET24(+)-4m5.3-deGFP-wo-lac, which is the exact same construct without the lac operator. Both of these sequences have the same linker to deGFP; however, Twist's codon optimisation tool, https://www.twistbioscience.com/faq/using-your-twist-account/what-does-twist-codon-optimization-tool-do was used to reverse translate these sequences.

Finally, Gibson assembly was used to clone the 4m5.3-deGFP DNA construct from the pET24(+)-4m5.3-deGFP-wo-lac plasmid, into the T7p14 plasmid with deGFP and mCherry. The T7p14-deGFP and T7p14-holin-mCherry plasmids were obtained from myTXTL plasmids by Sahan B. W. Liyanagedera. Figure 4.4 shows both of these DNA constructs as T7p14-4m5.3-deGFP and T7p14-4m5.3-mCherry. The details of how the cloning was performed is described in section 4.2.3.11, and table H.3 in the appendices shows the amino acid sequences for the 4m5.3 scFv design, the linker and deGFP, while the various DNA sequences of these constructs, with the promoter and terminator sequences, are shown in table H.1.

4.2.3 Standard molecular biology protocols

This section details the standard molecular biology protocols that were used through this experimental work. A full list of the chemicals and materials used is shown in the appendices in table F.1, a full list of the molecular biology kits in table G.1, linear DNA sequences in table H.1, promoter and terminator sequences in table H.2, primers in table H.4, and finally buffers and media in tables I.1, I.2 and I.3.

4.2.3.1 Polymerase Chain Reaction (PCR)

Polymerase Chain Reaction (PCR) with NEB Phusion high–fidelity polymerase, was used to amplify template DNA constructs, to obtain high concentrations of DNA for cell-free reactions. Tables 4.1 and 4.2 show the standard Phusion PCR reaction set-up and thermocycler protocol for the p70a-deGFP construct. For the other constructs used in this PhD project, the main factors in these protocols that changed were the forward and reverse primers, the annealing temperature, the extension time and the number of cycles. The full list of primers along with their annealing temperatures are shown in table H.4, while the optimal PCR settings and primers, for each of the

Component	**Reaction Volume** (µL)	**Stock Concentration**	**Reaction Concentration**
Water	31.125		
Phusion HF buffer	10	5x	1x
dNTPs	5	2 mM	0.2 mM
DMSO	2.5	100%	5%
Forward primer	0.25	50 µM	0.25 µM
Reverse primer	0.25	50 µM	0.25 µM
Template DNA	0.5	2 ng/µL	1 ng
Phusion HF polymerase	0.375	2000 U/mL	0.75 U/50µL

Table 4.1: Standard set-up for a 50 µL p70a-deGFP Phusion PCR reaction

Stage	**Temperature** (°C)	**Time (s)**	**Cycles**
1 (Initial annealing)	98	120	1
2 (Denaturation)	98	5	35
3 (Annealing)	68	30	35
4 (Extension)	72	44	35
5 (Final extension)	72	300	1

Table 4.2: Standard protocol for a 50 µL p70a-deGFP Phusion PCR reaction

different DNA constructs used in this project, are shown in table 4.3. For some of the PCRs performed in this project, the dNTPs, Phusion polymerase and Phusion HF buffer, were ordered as a Phusion HF mastermix, and this was used instead of adding the individual components.

4.2.3.2 Gel electrophoresis

After performing PCR, agarose gel electrophoresis was used on a 5 µL sample of the PCR mixture, to ensure that the reaction had worked properly. Firstly, 40 mL of TBE 1x solution was put into a flask with 0.48 g of agarose. This was dissolved by putting the flask in the microwave for 45 seconds, and once the flask has cooled down, 4 µL of

DNA Name	Forward primer	Reverse primer	Annealing temperature (°C)	Cycles	Extension time (s)
p70a-deGFP	ZF10	NL32	68	35	44
T7p14-deGFP	T7FWD250	T7REV250	61	35	60
p70a-4m5.3-deGFP	NL001FWD	NL001REV	72	31	60
pET24(+)-4m5.3-deGFP-lac	MS001	MS002	60	31	130
pET24(+)-4m5.3-deGFP-wo-lac	MS001	MS002	60	31	130
T7p14-4m5.3-deGFP	T7FWD250	T7REV250	61	31	130
T7p14-4m5.3-mCherry	T7FWD250	T7REV250	61	31	130

Table 4.3: Primers, annealing temperatures, cycles and extension times that worked for the different template linear DNA constructs. All other factors was the same as shown in tables 4.1 and 4.2.

SybrSAFE DNA gel stain was added. Next, the solution was poured into a casting tray and left to set for 30 minutes, and a gel well comb was put into the casting tray to make wells in the gel. Once the gel had set, it was placed into the buffer chamber inside the BioRad electrophoresis machine, and 5 µL of PCR sample was added to a well, along with 1 µL of loading dye. In addition to this, 6 µL of 1kb HyperLadder or GeneRuler was then added to a separate well, which was used to compare the bands that were obtained after running gel electrophoresis. In general, the gel electrophoresis was ran at 110 volts for 50 minutes, with TBE 1x as the running buffer. Images of the gel under UV light were taken to compare the length of the DNA segments in the PCR mixture, to the known length of the DNA template, using a 1kb HyperLadder or GeneRuler.

4.2.3.3 PCR Clean-up

Purified DNA was obtained from the remaining PCR product, by performing a PCR clean up protocol using the Zymo DNA Clean and Concentrator kit (table G.1 in the appendices). To begin with, a fresh filtration and collection tube were taken from the Zymo kit, with the filtration tube sitting inside the collection tube. Binding buffer and PCR sample were added into the filtration tube, in a 5:1 ratio, and this was mixed

well by pipetting up and down. This was then placed in the microcentrifuge and spun at 2000g for 5 minutes. Afterwards, the liquid in the collection tube was discarded, 200 µL of wash buffer was added directly to the filtration tube, and then this was spun at 10,000g for 1 minute. This wash step was repeated once more, and then next a dry spin was performed at 10,000g for 1 minute. The samples were left for 15 minutes with their caps open, to ensure there wasn't any ethanol remaining in the sample, which was crucial for obtaining high quality DNA, with no contaminants, that could be used in cell-free reactions. Finally, 30 µL of Milli-Q water was heated on a plate at 72 °C, and added to the filtration tube, which was incubated at room temperature for 5 minutes. This was then spun in the centrifuge at 16,000g for 3 minutes, and then a Nanodrop machine was used to measure the concentration and the quality of the final DNA sample.

4.2.3.4 Overnight cell cultures

Overnight cultures were set up by adding 5 mL of Luria-Bertani (LB) media to each culture tube, along with antibiotic (mainly 5 µL of ampicillin was used at 100 µg mL^{-1} concentration), and the culture was inoculated with cells taken from a glycerol stock, generally using a sterile inoculation loop. These culture tubes were then placed in an incubator overnight, at 37 °C and 300RPM. The next day these cell cultures could be used to inoculate larger cell cultures, or used for minipreps in order to obtain plasmid DNA.

4.2.3.5 Minipreps

In order to prepare plasmid DNA, minipreps were performed using the GeneJET plasmid miniprep kit (table G.1 in the appendices) on 5 mL overnight cell cultures. If the cultures appeared cloudy in the morning, then they were processed using the GeneJET plasmid miniprep kit to obtain purified plasmid DNA. Firstly, the cells were pelleted in the centrifuge, the supernatent was discarded, and then the pelleted cells were suspended in 250 µL of the resuspension solution in a microcentrifuge tube. After this, 250 µL of lysis solution was added to the tube and inverted 6 times, until the solution became viscous, then 350 µL of neutralisation solution was added and mixed thoroughly by inverting 6 times. Next, the tube was spun in the centrifuge for 5 minutes, the supernatent was transferred to a GeneJET spin column, and then the column was processed in a similar way to the PCR cleanup protocol (section 4.2.3.3). The column

was centrifuged for 1 minute and the flow through was discarded. Then 500 µL of wash solution was added to the column and it was centrifuged for 60 seconds, with the flow through also being discarded. This wash step was repeated, then a dry spin was performed, and the columns were placed in a new 1.5 mL microcentrifuge tube with the caps open, to allow any remaining ethanol to evaporate. Finally, 30 µL of elution buffer from the GeneJET kit was heated on a plate at 72 °C, and added to the filtration tube, which was incubated at room temperature for 5 minutes. This was then spun in the centrifuge, and then a Nanodrop machine was used to measure the concentration and the quality of the final plasmid DNA sample. After obtaining the purified plasmid, PCR was then used to double check the plasmid contained the correct DNA segment of interest, by using primers that are known to amplify the DNA segment, and comparing the length of the PCR product on a gel to the known length.

4.2.3.6 DNA sequencing

Sanger sequencing [368] was used to sequence plasmids in this project, using the company GeneWiz. Firstly, minipreps (section 4.2.3.5) were performed to to obtain a sample of the plasmid that was at least at 100 ng/µL concentration. The plasmid sample, along with primers, were prepared following the GeneWiz sample preparation guidelines https://www.genewiz.com/en-GB/Public/Services/Sanger-Sequencing in 1.5 mL eppendorf tubes. After obtaining the sequencing results, a multiple sequence alignment (MSA) was performed using Clustal Omega [369] of the different sequencing reads, against the expected sequence.

4.2.3.7 Making agar plates with ampicillin

Agar was heated up in the microwave at 50% power for 21 minutes. After this, 50 µL of ampicillin at 100 µg/mL concentration, was added to 50ml of agar under a flame, and mixed carefully to avoid bubbles. Next, 25ml of the agar was poured into each plate, and the lids were left off until they cooled down. These plates were then labelled and stored in fridge.

4.2.3.8 Making glycerol stocks

Firstly, overnight cultures were performed using 5 mL of LB media, 5 µL of ampicillin at 100 µg/mL concentration, and a scraping of cells. These cultures were put in the incubator at 37 °C overnight. The next day, glycerol stocks were made by adding

500 µL of culture to a cryo tube, along with 500 µL of 50% glycerol. These tubes were then snap frozen with liquid nitrogen, and then stored in the −70 °C freezer for long term storage of the cells.

4.2.3.9 Performing chemical transformations

NEB DH5 alpha competent cells were taken from the −70 °C freezer and thawed on ice. After this, 2 µL of plasmid and 2ul of positive control plasmid, were added to competent cells, mixed thoroughly and then kept on ice for 30 minutes. The cells were then heat shocked at 42 °C for 30 seconds on a heatblock. Next, 900 µL of room temperature SOC media was added to both the sample plasmid and the positive control tubes, and they were placed in the incubator at 37 °C for 45 minutes. The cells were then pelleted in the centrifuge at 16,000RPM for 3 minutes, 900 µL of SOC media was removed from both tubes, and then the cells were re-suspended in the remaining media. Finally, a sterile inoculation loop was then used to streak the cells onto plates, and they were then put in the incubator at 37 °C. This was all performed under a flame to ensure a sterile working environment, so that the plates wouldn't be contaminated.

4.2.3.10 Colony PCR

Firstly, a scraping from a colony on a plate was taken and put into 20 µL of LB media in an eppendorf. This was then placed in an incubator for 3 hours to allow the cells to grow. After this, 10 µL of this sample was boiled at 98 °C for 10 minutes on a heat block, and then the cells were pelleted by putting them in the centrifuge at top speed for 3 minutes. A PCR reaction was then performed, using 2 µL of the supernatent as the template DNA and using primers that amplify the DNA segment of interest. Finally, gel images were used to see if the colony contains the DNA segment of interest, and the remaining 10 µL of the colony sample was used to inoculate another culture, which was used for making glycerol stocks and minipreps.

4.2.3.11 Gibson assembly cloning

In this project, Gibson assembly was used to clone the coding region of the 4m5.3 scFv antibody fragment, from the pET24(+)-4m5.3-deGFP-wo-lac plasmid obtained from Twist, into the T7p14-deGFP and T7p14-holin-mCherry plasmids. The T7p14-holin-mCherry plasmid was the only T7p14 plasmid we had in the lab with mCherry in it, where holin is a bacteriophage virus protein used in the lytic cycle [370]. However,

this was only used as a backbone for Gibson assembly, and primers were designed to only amplify the T7p14 vector along with the mCherry DNA. This section details the general Gibson assembly protocol, that was used to create both the T7p14-4m5.3-deGFP and T7p14-4m5.3-mCherry constructs.

Firstly, primers were designed to amplify the backbone sequence of the target plasmid, which included everything in the plasmid except the coding region. Additional primers were designed to amplify the insert sequence, which is the coding region of the sequence that is being cloned into the target backbone. However, the primers for the insert sequence also had overhangs that would anneal onto the target backbone. The full list of primers used for Gibson assembly is shown in table H.4 in the appendices.

After this, separate PCR reactions were performed to amplify the backbone and insert sequences using these primers. The same PCR set-up and protocol shown in

DNA Name	**Forward primer**	**Reverse primer**	**Annealing temperature** (°C)	**Extension time (s)**
T7p14-deGFP backbone	MS007	MS008	61	130
T7p14-deGFP insert 1	MS003	MS004	61	60
T7p14-deGFP insert 2	MS005	MS006	61	60
T7p14-mCherry GSPA backbone	MS012	MS008	63	165
T7p14-mCherry GS TEV backbone	MS011	MS008	61	165
T7p14-mCherry GSPA insert	MS005	MS010	61	30
T7p14-mCherry GS TEV insert	MS005	MS009	61	30

Table 4.4: Primers, annealing temperatures and extension times that worked for the different DNA constructs. All other factors was the same as shown in tables 4.1 and 4.2.

tables 4.1 and 4.2 was used for these PCR reactions; however, the primers used along with their annealing temperatures and extension times are shown in table 4.4. Once the backbone and insert sequences had been amplified, and the amplified sequences were shown to have the expected length from using gel electrophoresis, Dpn1 digestion was performed to remove any of the remaining template plasmid. The Dpn1 digestion step was important in order to limit the possibility that colonies have the original template plasmid rather than the new plasmid, when performing transformations later. Dpn1 digestion was performed with the combined 100 µL of two PCR reactions, 6 µL of milliQ water, 12 µL of Dpn1 buffer and 2 µL of Dpn1 enzyme. This was performed using the New England BioLabs (NEB) Dpn1 digestion kit (table G.1 in the appendices). After assembling the Dpn1 reaction together, it was put on the heatblock at 37 °C for 65 minutes, and then on the heatblock at 80 °C for 20 minutes to deactivate the enzyme. Gel electrophoresis was then performed to check that the length of the amplified DNA constructs were correct and then these samples were then purified using the PCR protocol described in section 4.2.3.3.

Next, Gibson assembly was performed with the NEB Gibson assembly master mix with 0.5 µL of backbone, 1.5 µL of insert, 10 µL of NEB Gibson assembly master mix and 8 µL of de-ionised water. This was then put in the ThermoCycler at 50 °C for 15 minutes. After this, 2 µL of the Gibson assembly mixture was used to perform a transformation into NEB DH5 alpha competent cells, as described in section 4.2.3.9. These were streaked onto an agar plate with ampicillin and left in the incubator overnight at 37 °C. The next day 5 colonies were picked from the plate and put into 20 µL of LB media. Colony PCRs were performed on 10 µL of the sample as described in section 4.2.3.10, while the remaining 10 µL of the sample was used to set-up overnight cultures of these colonies. Finally, minipreps (section 4.2.3.5) were performed on the colonies that appeared to have the correct plasmid from the colony PCRs.

4.2.4 Preparing and performing cell-free reactions

This section will describe how the main components of the cell-free reactions, the crude lysate, energy solution and Chi6 linear DNA, were prepared and then it will describe how the cell-free reactions were performed.

4.2.4.1 Crude lysate systems

Throughout this project, various crude lysate systems were used and they were made in slightly different ways. Table 4.5 details the names of these lysate systems, along with the *E. coli* strain, protocol used, and any plasmids that were induced during growth. The first protocol was developed by Alex Perkins (AP Protocol) and was adapted from Sun et al. 2013 [335]. This protocol was used to make the first lysate in table 4.5, with ID AP RGami. After this, another protocol developed by Sahan Liyanagedera (SL AutoSonic 3.0), based on his work in [128], was used to make the lysates with IDs MS RGami T7 and MS RGami T7 pRARE. Table 4.6 details the steps taken in the AP crude lysate protocol that was used to make the AP RGami crude lysate.

Lysate ID	***E. coli* strain**	**Protocol**	**Induced plasmids or prophages**
AP RGami	Rosetta gami 2 (DE3)	AP Protocol (Adapted from [335])	
MS RGami T7	Rosetta gami 2 (DE3) pAD LyseR	SL AutoSonic 3.0 [128]	DE3 prophage (T7 RNA Polymerase)
MS RGami T7 pRARE	Rosetta gami 2 (DE3) pAD LyseR	SL AutoSonic 3.0 [128]	DE3 prophage (T7 RNA Polymerase), pRARE plasmid

Table 4.5: Table of the different crude lysates used in this project, along with the *E. coli* strain and the protocol used, and a list of additional plasmids or prophages that were induced during growth.

Step	**Description**
1 (Preparation)	A 10 mL culture of Rosetta gami 2 (DE3) cells was left in the incubator overnight at 37 °C and 220RPM, and the S30A and S30B buffers were prepared as described in Sun et al. 2013 [335].

Step	Description
2 (Main growth)	The overnight culture was used to inoculate a larger 1 L culture of LB media, in a 2 L baffled flask, and this was left in the incubator at 37 °C and 220RPM. The cells were harvested when the OD600, measured on the nanodrop machine, reached 2.7 and then was placed on ice.
3 (Pelleting cells)	After this, a large floor centrifuge (Thermo scientific, Sorvall Lynx 4000 centrifuge) was used to pellet the cells, and the LB media was discarded.
4 (Wash step x2)	Next, the pelleted cells were re-suspended in 100 mL of S30A buffer, and then pelleted again in the centrifuge. The S30A buffer was then discarded, the wash step was repeated, and the cells were flash frozen in liquid nitrogen and stored in the −70 °C freezer.
5 (Re-suspending cells)	The next day, the cells were thawed on ice and re-suspended in 10 mL of S30A buffer and 20 µL of DTT.
6 (Sonication)	The cells were placed in a box packed full of ice to keep them cold, and then sonicated with a Fisherbrand 120 Sonicator and probe, at 70% amplitude, 20 seconds on: 20 seconds off, for 1 minute 40 seconds in total. After this, they were pelleted again at 12,000g and the supernatent was extracted.
7 (Run-off)	The lysate is placed in the incubator at 37 °C and 220RPM for 1 hour 30 minutes.
8 (Dialysis)	The sample is placed into a 10k MWCO dialysis casette in a beaker of 1 L of S30B buffer and 2 mL of 1 M DTT and left overnight. The following morning, the sample was taken out and the supernatent was extracted.
9 (Aliquoting)	Finally, the lysate is aliquoted out into small PCR tubes, flash frozen with liquid nitrogen and then stored in the −70 °C freezer.

Table 4.6: Crude cell-free lysate AP protocol which is adapted from Sun et al. [335].

Step	Description
1 (Preparation)	A mini culture of 5 mL of 2xYTP media (recipe described in table I.1 in the appendices) with 5 µL of 100 µg/mL ampicillin and a scraping of Rosetta gami 2 DE3 pAD LyseR from an overnight agar plate at 37 °C and 220RPM for 8 hours. Afterwards, 100 µL of mini culture was used to inoculate a 50 mL midi culture of 2xYTP media and 50 µL of 100 µg/mL ampicillin which was also left at 37 °C and 220RPM for 8 hours. S30A and S30B buffers were prepared as described in Sun et al. 2013 [335].
2 (Main growth)	The next day, the midi culture was used to inoculate 3 flasks of 400 mL 2xYTP media cultures so that the OD600 would be 0.1, and 400 µL of 100 µg/mL ampicillin was also added. These cultures were left in the incubator at 37 °C and 220RPM and harvested once the OD600 values reached 2. In order to induce the production of T7 RNA polymerase, 1mM IPTG was added to the cultures once the OD600 reached 0.6.
3 (Pelleting cells)	After this, the large floor centrifuge was used to pellet the cells, and the LB media was discarded.
4 (Wash step x2)	Next, each of the pelleted cells were re-suspended in 100 mL of S30A buffer, and then pelleted again in the centrifuge. The S30A buffer was then discarded and the wash step was repeated.
5 (Re-suspending cells)	The cell pellet was weighed and re-suspended with 2.5 mL S30A buffer for every 1 g pellet, and split into falcon tubes with 5 mL of re-suspended pellet in each. After this, the pellets were flash frozen in liquid nitrogen and put in the −70 °C freezer. By flash freezing the cells, the pAD LyseR plasmid is expressed to produce phage lambda endolysin, which helps to lyse the cells [371].
6 (Lysozyme)	The next day, the cell pellets were thawed in a water bath at room temperature for 30 minutes and then 1 mg/mL of lysozyme was added to each falcon tube. These tubes were left on ice for 2 hours and vortexed for 10 seconds every 30 minutes.

Step	Description
7 (Sonication)	Afterwards, the Fisherbrand 120 Sonicator and probe was used with 10 seconds on 10 seconds off, 70% amplitude and 1500J per 5 mL of re-suspended cell pellet, and afterwards they were vortexed vigorously for 2-3 minutes. After this, they were pelleted again at 12,000g and the supernatent was extracted.
8 (Run-off)	The lysate is placed in the incubator at 37 °C and 220RPM for 1 hour 30 minutes.
9 (Dialysis)	The sample is placed into a 10k MWCO dialysis casette in a beaker of 1 L of S30B buffer and 2 mL of 1 M DTT and left overnight. The following morning, the sample was taken out and the supernatent was extracted.
10 (Bradford assay and concentration)	A Bradford assay (table G.1) was performed to find out the protein concentration of this lysate. Then it was concentrated to 50 mg/mL using an Amicon Ultra 3K falcon column.
11 (Aliquoting)	Finally, the lysate was aliquoted out into PCR tubes, flash frozen with liquid nitrogen, and then stored in the −70 °C freezer

Table 4.7: Crude cell-free lysate SL AutoSonic 3.0 protocol [128].

Table 4.7 describes the SL AutoSonic 3.0 protocol [128] protocol that was used to make the MS RGami T7 and the MS RGami T7 pRARE lysates. However, for MS RGami T7 pRARE, a few changes were made. In the main growth step, the cells were grown with chloramphenicol as well as ampicillin, to ensure the cells retain the pRARE plasmid (AddGene #84650) which expresses rare tRNAs in *E. coli* and helps expressing proteins such as scFv antibody fragments. In addition to this, for the sonication step, the cells were sonicated in a 50 mL glass beaker at 10 seconds on, 10 seconds off 70% amplitude x 12, followed by 10 seconds at 100% amplitude x 1. For the concentration step, the MS RGami T7 pRARE lysate was concentrated to 35 mg/mL.

4.2.4.2 Preparing chi6 linear DNA

Firstly, the Chi6 primer sequences (table H.4 in the appendices), ordered from IDT, were resuspended in water at 1 mM. After this, 15 µL of each primer stock was added into a PCR tube, making a total reaction volume of 30 µL. This solution was then heated to 95 °C, and cooled 1 °C per minute until room temperature was reached in

a PCR machine. Finally, this 500 µM solution of Chi6 was then aliquoted into 2.5 µL aliquots in new PCR tubes and stored in the −20 °C freezer for further use.

4.2.4.3 Energy solution

Similar to the crude lysate systems, in this project there were two different methods of preparing the energy solution used in the cell-free systems. The first method combined everything into one solution that was used in addition to the crude lysate and DNA template to perform cell-free reactions. On the other hand, the other method left amino

Component	**Stock concentration (mM)**	**Final concentration (mM) (4x)**	**Volume added** (µL)
HEPES	2,000.00	200.00	100.00
tRNA	43.75 mg/mL	0.80 mg/mL	18.3
ATP	100.00	6.00	60.00
GTP	100.00	6.00	60.00
CTP	100.00	3.60	36.00
UTP	100.00	3.60	36.00
CoA	65.00	1.04	16.00
NAD	175.00	1.32	7.50
cAMP	650.00	3.00	4.60
Folinic acid	33.90	0.27	8.00
Spermidine	1,000.00	4.00	4.00
3PGA	1,400.00	120.00	85.70
PEG-8k	50.00%	8.00%	160.00
Mg glutamate	1,000.00	42.00	42.00
K glutamate	6,000.00	400.00	66.70
DTT	1,000.00	1.00	1.00
Amino acids (except tyrosine)	50.00	6.00	120.00
Tyrosine	50.00	3.00	60.00
Water			114.2

Table 4.8: This table details the different components, along with volumes and concentrations, for making 1.000 µL of 4x concentrated ES1.

Component	**Stock concentration (mM)**	**Final concentration (mM) (14x)**	**Volume added** (µL)
HEPES	2,000.00	700.00	350.00
tRNA	43.75 mg/mL	2.80 mg/mL	64.00
ATP	500.00	21.00	42.00
GTP	500.00	21.00	42.00
CTP	500.00	12.60	25.20
UTP	500.00	12.60	25.20
CoA	65.00	3.64	56.00
NAD	175.00	4.62	26.40
cAMP	650.00	10.50	16.15
Folinic acid	33.90	0.95	28.02
Spermidine	1,000.00	14.00	14.00
3PGA	1,400.00	420.00	300.00
Water			11.02

Table 4.9: This table details the different components, along with volumes and concentrations, for making 1.000 µL of 14x concentrated ES2.

acids, PEG-8K, Mg glutamate, K glutmate and DTT out of this solution, and they were added in separately in to the cell-free reactions.

Table 4.8 shows how the first energy solution (ES1) was assembled with the volumes and concentrations of all the individual components. This energy solution is 4x concentrated, therefore 5 µL would be added to a 20 µL cell-free reaction. In contrast to this, table 4.9 shows how second energy solution (ES2) was assembled, along with the volumes and concentrations of the individual components. This energy solution is 14x concentrated, therefore 1.43 µL would be added to a 20 µL cell-free reaction.

4.2.4.4 Performing cell-free reactions

After the crude lysate systems and energy solutions had been prepared, cell-free reactions could be performed, along with the addition of a few extra components. One of these additional components was Chi6 DNA, which is double-stranded DNA containing χ sites, which preferentially binds to nucleases that are contained in the *E. coli* lysate, and therefore stabilises the linear DNA construct used in the reaction [336].

Component	**Stock concentration**	**Final concentration**	**Volume added** (µL)
Energy Solution #1	4x	1x	27.00
S30A buffer	4x	1x	27.00
Chi6	50 µM	2 µM	4.32
Crude lysate	4x	1x	27.00
Total			**85.32**

Table 4.10: Cell-free reaction set up #1: Standard master mix assembly for 9 10 µL cell-free reactions with 20% extra volume added per component.

Component	**Stock concentration**	**Final concentration**	**Volume added** (µL)
Energy Solution #2	14x	1x	7.71
Amino Acids	6 mM	1.5 mM	27.00
PEG 8K	40%	2%	5.40
Mg glutamate	200 mM	5 mM	2.70
K glutamate	2000 mM	60 mM	3.24
DTT	100 mM	1.5 mM	1.62
Maltose	250 mM	15 mM	6.48
Chi6	50 µM	5 µM	10.80
Crude Lysate	50 mg/mL	10 mg/mL	21.60

Table 4.11: Cell-free reaction set up #2: Standard master mix assembly for 9 10 µL cell-free reactions with 20% extra volume added per component.

Another component was S30A buffer which was only added into the reaction when ES1 was used. Also, when ES2 was used, amino acids, PEG-8K, Mg glutamate, K glutmate and DTT were added separately into the reaction and maltose was added into these reactions as well.

Firstly, all components, except the template DNA and water, were assembled into a master mix with an extra 20% added to the volume, in order to account for waste and pipetting errors. The crude lysate was left to last and all of these were kept on ice during assembly; however, there was a short incubation period where the master

mix was left for 10 minutes at room temperature, to allow the nucleases from the crude lysate to bind to the chi6 DNA. Generally, cell-free reactions were performed in triplicate, and positive and negative controls would be performed, along with the template DNA of interest. Table 4.10 shows the first approach to assembling cell-free reactions, called cell-free reaction set up #1, which uses ES1, and shows an example for setting up 10 µL reactions. After assembling this master mix, it was mixed thoroughly by vortexing, and 7.9 µL of master mix was added to each well in a nunc 384-well plate, along with 10 nM of template DNA and the remaining volume as water. For example, for a 100.40 nM stock solution of p70a-deGFP template DNA, 1.0 µL of DNA template and 1.1 µL of water were added to the well, along with the 7.9 µL of master mix. Once all the reactions had been assembled, the plate would be put in the plate spinner to make sure all the liquid is at the bottom of the well and 35 µL of Chill-out liquid wax was added on top of each reaction. In addition to this, a nunc 384 well plate plastic seal was added on top of the plate to ensure the reactions didn't evaporate, and then the plate was put in the plate reader to measure fluorescence in Relative Fluorescence Units (RFU), over 12 hours and at 37 °C. For measuring GFP fluorescence, generally the wavelengths Excitation: 485, Emission: 520 were used, and for mCherry the wavelengths Excitation: 587, Emission: 610 were used. Different gain settings were used throughout the project, depending on what construct was being expressed; however, mCherry generally needed a higher gain setting.

In contrast to this, table 4.11 shows an example protocol to assemble 10 µL cell-free reactions using ES2. For the same 100.40 nM stock solution of p70a-deGFP template DNA, 1.0 µL of DNA template and 1.1 µL of water were added to the well, along with the 8.01 µL of this master mix. After this, the same protocol outlined for cell-free reaction set up #1, was followed to perform the cell-free reactions using the plate reader.

4.3 Results

4.3.1 AlphaFold2 prediction of 4m5.3-deGFP and 4m5.3-mCherry structures

Firstly, AlphaFold2 [65] was used to obtain structural models of the amino acid sequence of the 4m5.3 scFv antibody sequence, joined to deGFP and mCherry, with a short GSPA linker sequence. These amino acid sequences are shown in table H.3 in the

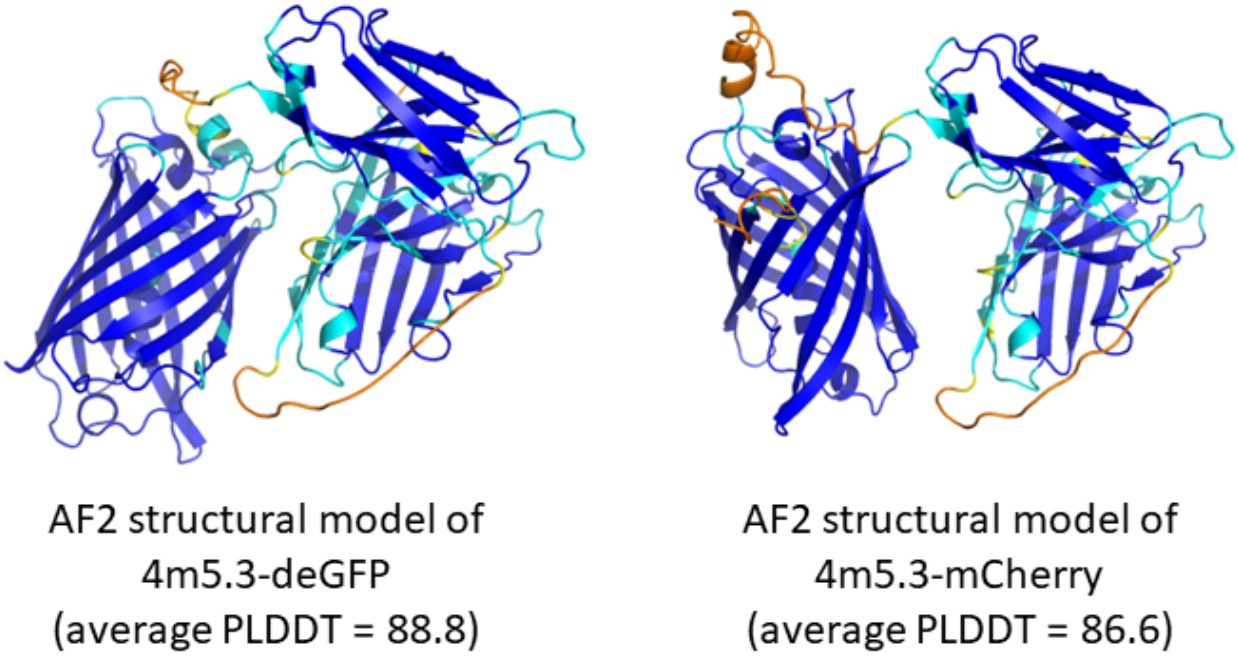

Figure 4.5: Alphafold2 structural models of the 4m5.3 scFv antibody fragment attached to deGFP and mCherry with a short linker, coloured by PLDDT score. Dark blue shows highly confident regions of the structural model, whereas lighter blue, yellow and orange regions show less confident areas of the structural model.

appendices and ColabFold [276] was used to run AlphaFold2. Figure 4.5 shows the highest ranked structural model for 4m5.3-deGFP and 4m5.3-mCherry, which shows that the scFv and deGFP/mCherry domains are predicted to fold correctly. The highest ranked structural models for 4m5.3-deGFP and 4m5.3-mCherry, have average predicted local distance difference test (pLDDT) scores of 88.8 and 86.6 respectively, which shows that these predictions are very confident. The only areas in these structural models that have low pLDDT scores, are the linker regions, between the scFv and deGFP, and between the light and heavy chains within the scFv antibody fragment. This makes sense as these parts of the structure are expected to be more flexible and to move around freely. Overall, these results suggest that fusing the scFv antibody fragments to a fluorescent protein, such as deGFP or mCherry, could be used as a method to measure the expression of the scFv antibody fragments. This is because, as the scFv antibody fragment and deGFP/mCherry are predicted to fold correctly, this suggests that fusing the fluorescent proteins will not interfere with the folding of the scFv.

4.3.2 Initial testing for p70a-4m5.3-deGFP construct

After AlphaFold2 showed that the 4m5.3-deGFP and 4m5.3-mCherry sequences were predicted to fold correctly, the p70a-4m5.3-deGFP construct (figure 4.4 and table H.1 in the appendices) was ordered as a gBlock from IDT. PCR with Phusion polymerase was used to amplify the DNA construct using the PCR protocol shown in table 4.3 (section 4.2.3.1). Gel electrophoresis was then performed on a 5 µL sample and compared to a 1kb HyperLadder, to determine if the DNA is of the correct molecular weight (section 4.2.3.2). Figure 4.6 shows the gel image obtained, and we can see that the band for the PCR sample in well 2 is slightly above the 2kbp line in the HyperLadder, which is expected as the length of the construct is 2,144bp. This sample was then processed using the PCR clean-up protocol (section 4.2.3.3) to obtain purified DNA that could be used in cell-free reactions.

Cell-free reactions were performed with the AP RGami crude lysate system, which

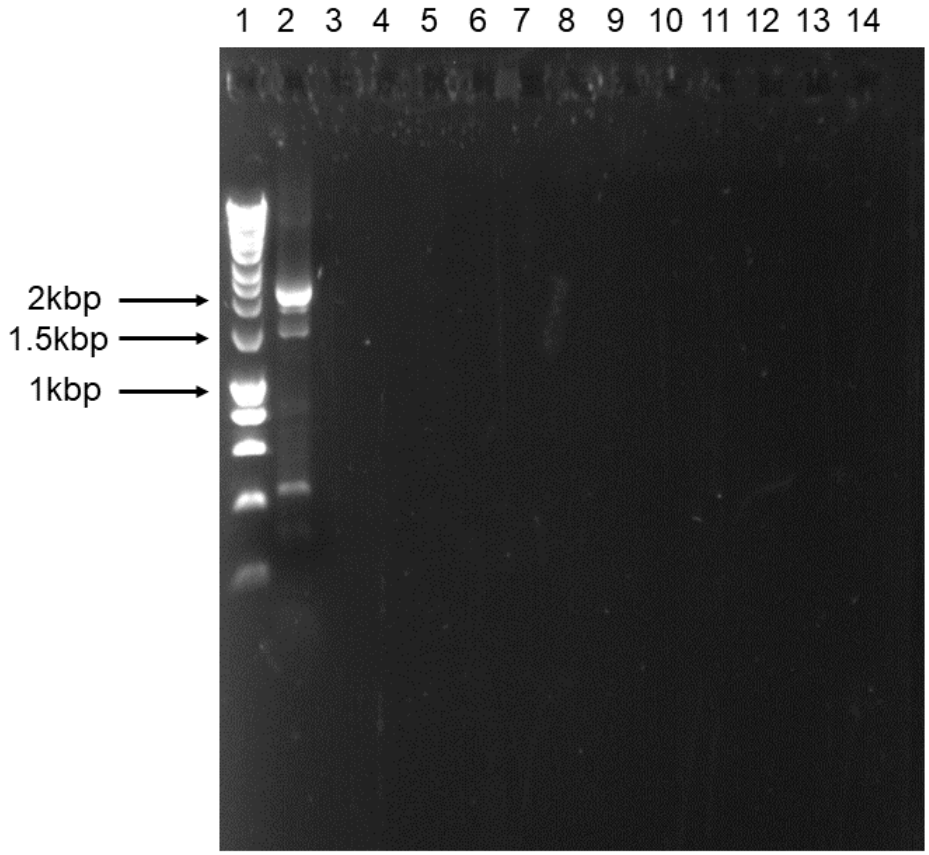

Figure 4.6: PCR gel image of the p70a-4m5.3-deGFP linear DNA construct that was ordered from IDT as a gBlock. The HyperLadder is shown in well 1 and the p70a-4m5.3-deGFP 5 µL PCR sample is shown in well 2.

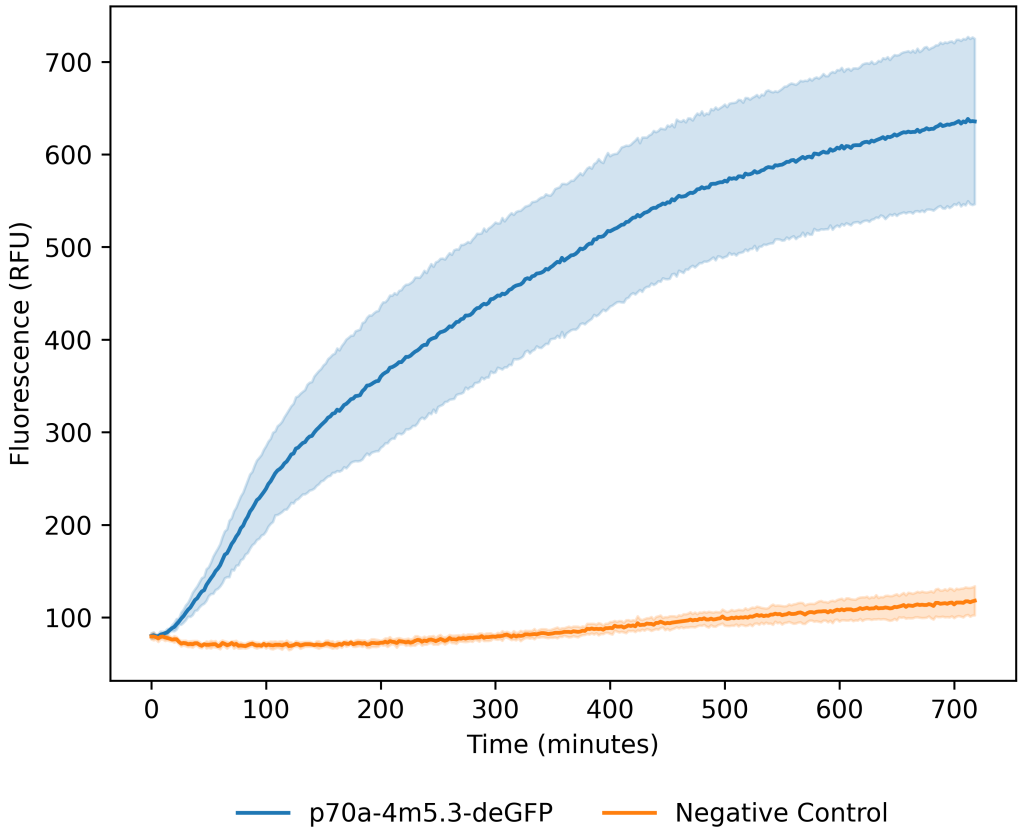

Figure 4.7: Cell-free reactions performed in triplicate on the p70a-4m5.3-deGFP DNA template and a negative control, which has the exact same cell-free components except template DNA. These reactions used the AP RGami crude lysate, energy solution #1 and the cell-free set-up #1. The mean RFU values and standard deviation have been plotted.

was a lysate made by Alex Perkins using the AP Protocol (tables 4.5 and 4.6), ES1 (table 4.8) and the cell-free set-up #1 (table 4.10). This was performed as a quick test to see if we could get a signal from the p70a-4m5.3-deGFP linear construct. The cell-free reactions were performed for 12 hours, at 37 °C, at gain 800 on the BMG Labtech Omega plate reader, and in triplicate. Also, a negative control was included, which had the exact same cell-free components; however, the template p70a-4m5.3-deGFP template DNA was excluded. Figure 4.7 shows the results of performing cell-free reactions for p70a-4m5.3-deGFP, along with a negative control. This plot shows a clear signal from the p70a-4m5.3-deGFP reactions (blue), as the RFUs are a lot higher than the RFUs for the negative control (orange). Therefore, this provides evidence that this 4m5.3 antibody construct can be expressed in *E. coli* cell-free systems.

4.3.3 Creating a batch of lysate with T7 RNA polymerase

Although the p70a-4m5.3-deGFP construct was shown to be successfully expressed in the AP RGami lysate, the cost of ordering the library of scFv sequences from IDT as gBlocks was too prohibitive for this project. As a result of this, it was decided to order this library from Twist Bioscience as clonal genes instead. However, Twist could not manufacture the original p70a-4m5.3-deGFP construct due to the p70a promoter, so this construct was ordered as pET24(+)-4m5.3-deGFP-lac and pET24(+)-4m5.3-deGFP-wo-lac, as described in section 4.2.2. These sequences had T7 promoters and therefore required a crude lysate system with T7 RNA polymerase, which meant the MS RGami lysate could not be used. In order to produce these T7 constructs, a batch of T7 RNA polymerase was made using the SL AutoSonic 3.0 protocol (table 4.7), as this protocol had been shown to achieve high performing crude lysate systems, which could even assemble bacteriophage [128].

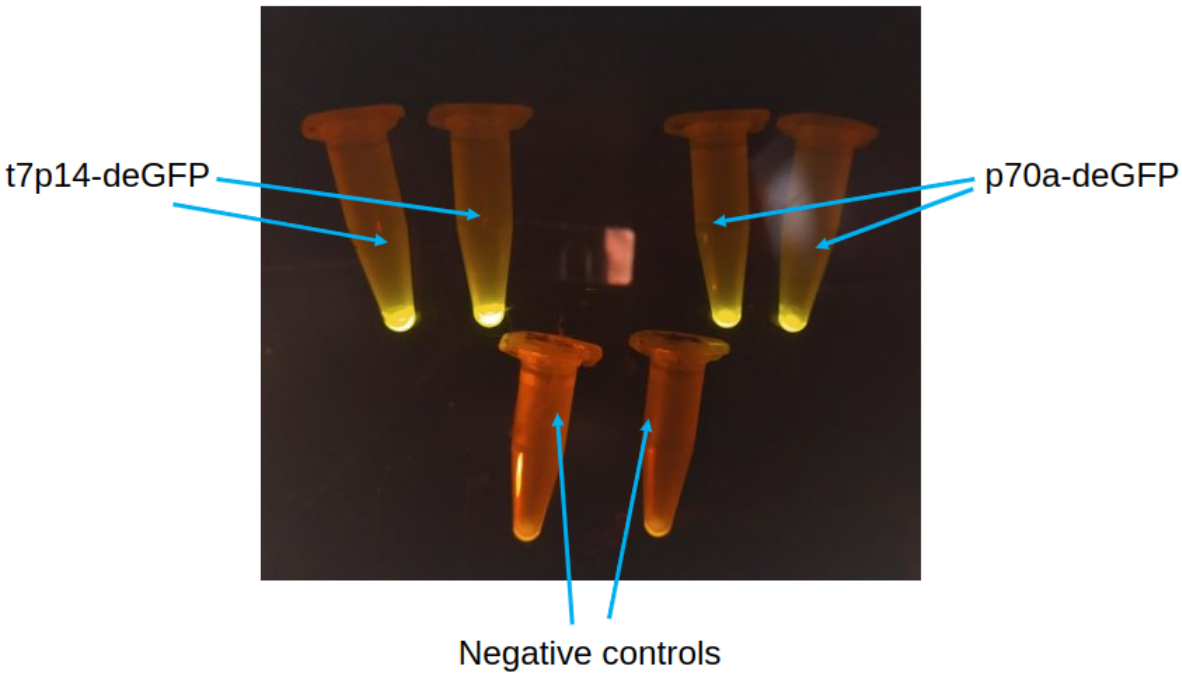

Figure 4.8: Testing MS RGami T7 lysate batch with 20 µL cell free reactions in eppendorfs and viewing fluorescence under UV light. The two top left eppendorfs are for T7p14-deGFP, the two top right eppendorfs are for p70a-deGFP and the bottom two eppendorfs are negative controls, which included the exact same cell-free components as the other reactions, but the template DNA was replaced with water.

After creating this MS RGami T7 lysate, the performance was tested with the p70a-deGFP and T7p14-deGFP constructs. Firstly, 20 µL cell-free reactions were performed in eppendorfs and placed in the ThermoMixer at 29 °C overnight. These reactions were performed with an energy solution created using the ES2 protocol (table 4.9) by Sahan Liyanagedera (SL ES2), and they were performed in triplicate. Figure 4.8 shows 2 reactions for each construct, in eppendorf tubes and under UV light. The top left reactions show were performed with the T7p14-deGFP construct, the top right constructs on the p70a-deGFP construct and the bottom two reactions were negative controls. The negative controls included the exact same cell-free components as the other reactions, but the template DNA was replaced by water. These results show that this MS RGami T7 lysate is highly functional for producing both p70a and T7 constructs, as the eppendorf tubes for these constructs show fluorescence, in comparison to the negative control eppendorf tubes which don't show any fluorescence. However, in order to obtain quantitative values for the fluorescence observed from these tubes, 10X and 20X dilutions were made of the eppendorf reactions at 10 µL, for end point fluorescence measurements on the Biotek plate reader. This also included a straight 10 µL sample

Sample ID	**Measurement 1 (RFU)**	**Measurement 2 (RFU)**	**Measurement 3 (RFU)**	**Mean (RFU)**
T7p14-deGFP 10X	38,860	38,860	42,660	$40,127 \pm 1,267$
T7p14-deGFP 20X	41,460	43,860	43,720	$43,013 \pm 778$
T7p14-deGFP Neat	27,399			27,399
p70a-deGFP 10X	33,260	33,620	37,640	$34,840 \pm 1,404$
p70a-deGFP 20X	29,960	31,180	30,280	$30,473 \pm 365$
p70a-deGFP Neat	11,622			11,622

Table 4.12: End point plate reader measurements, shown in relative fluorescence units (RFU), for T7p14-deGFP and p70a-deGFP cell free reactions at gain 50.

from the eppendorfs as a "neat" measurement, and these end point measurements were ran at gain 50.

Table 4.12 shows the results from the end point measurements for these cell-free reactions. The RFUs for the 10X dilutions were multiplied by a factor of 10, and RFUs for the 20X dilutions were multiplied by a factor of 20, so that all the values could be compared. From this table, we can see that the p70a-deGFP neat measurement of 11,622 RFU is an outlier compared to the 10X and 20X dilution values of 34,840 RFU and 30,473 RFU respectively. The same is true for the T7p14-deGFP neat value of 27,399 RFU, compared to the 10X and 20X dilution values of 40,127 RFU and 43,013 RFU, therefore both of these neat measurements were excluded. After excluding the neat measurements, the average for T7p14-deGFP across all the dilutions is $41,570 \pm 927$ RFU, and the average for p70a-deGFP across all the dilutions is $32,657 \pm 1,172$ RFU. This tells us that the T7p14-deGFP had higher expression than the p70a-deGFP. Overall, these results from table 4.12 and figure 4.8, show that the MS RGami T7 crude lysate is functional, and can be used to express proteins with both p70a and T7 promoters.

4.3.4 Testing pET24(+)-4m5.3-deGFP constructs

Now that we had a fully functional crude lysate with T7 RNA polymerase, the pET24(+)-4m5.3-deGFP constructs from Twist could be tested. The only difference between these sequences is that pET24(+)-4m5.3-deGFP-lac has a lac operator, and pET24(+)4m5.3-deGFP-wo-lac does not have a lac operator. The lac operator is regulatory site on the plasmid, which a lac repressor binds onto and blocks the RNA polymerase from binding onto the promoter region and starting transcription [367]. However, when lactose is present, it binds onto the lac repressor and stops it binding onto the lac operator, which consequently allows RNA polymerase to start transcribing the gene [367]. Using lac operators provides increased control over transcription, which makes it easier for companies, such as Twist, in their cloning pipelines. However, there was concern that the lac operator could impact the expression of the construct in cell free reactions, so this is why it was ordered without the lac operator as well.

Firstly, PCR was performed on the pET24(+)-4m5.3-deGFP-lac and pET24(+)-4m5.3-deGFP-wo-lac mini-prepped DNA stocks received from Twist, in order to obtain amplified linear DNA for the cell-free reactions (section 4.2.3.1). This was performed with Phusion polymerase with the primers and PCR settings shown in table

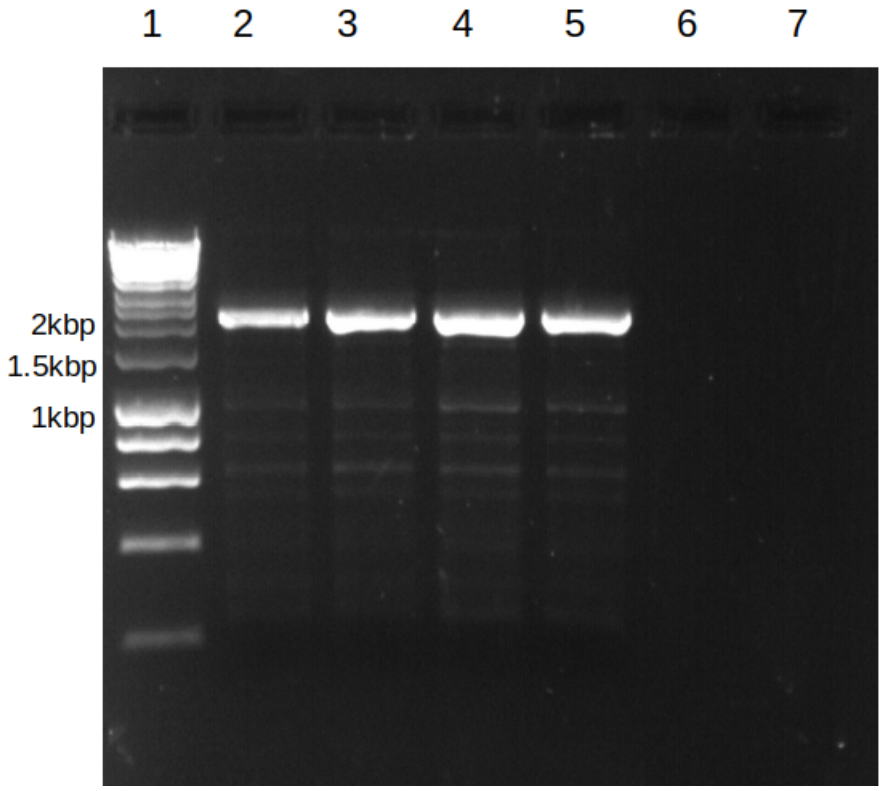

Figure 4.9: PCR gel image of the Twist pET24(+)-4m5.3-deGFP-lac and pET24(+)-4m5.3-deGFP-wo-lac constructs that have been amplified from the mini-prepped plasmids. The second well has the hyper ladder, the third and fourth wells have pET24(+)-4m5.3-deGFP-lac PCR samples and fifth and sixth wells have the pET24(+)-4m5.3-deGFP-wo-lac PCR samples.

4.3. PCR was performed for two 50 µL reactions and then gel electrophoresis was used on 5 µL samples, to see if the correct band had been obtained (section 4.2.3.2). Figure 4.9 shows four strong bands on the gel image, which are around the 2kbp line on the hyper ladder in the second well. Both of these constructs have sequence lengths around 2kbp, therefore this gel image shows that the DNA is of the correct molecular weight.

After purifying the DNA from the PCR samples using the PCR clean-up protocol (section 4.2.3.3), 10 µL cell-free reactions were performed to test whether these constructs could be expressed. Positive controls of T7p14-deGFP and p70a-deGFP were included in the reactions, along with negative controls which included the exact same cell-free components but the template DNA was replaced with water, to ensure the reactions were set up correctly. All of these cell free reactions were performed using the cell-free reaction set up #2, the MS RGami T7 lysate, in triplicate, with the tempera-

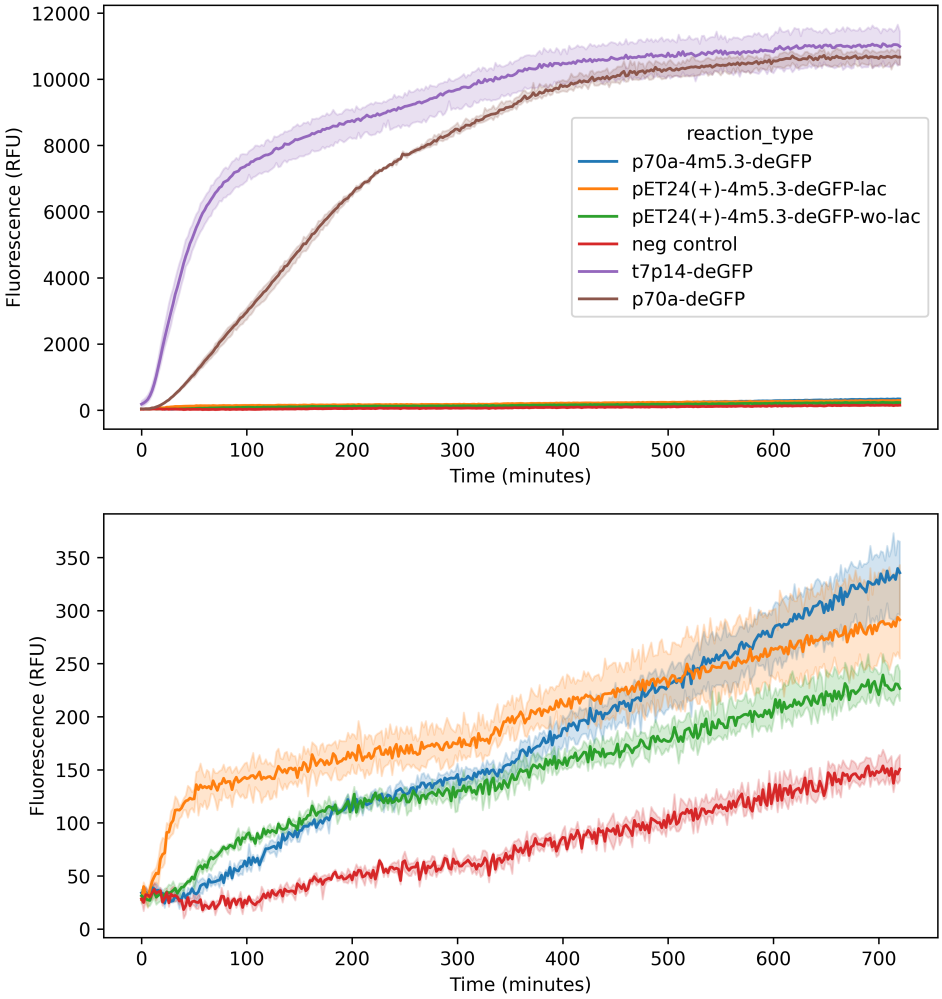

Figure 4.10: Cell-free reactions performed for T7p14-deGFP, p70a-deGFP, p70a-4m5.4-deGFP, pET24(+)-4m5.3-deGFP-lac and pET24(+)-4m5.3-deGFP-wo-lac, along with a negative control which included the exact same cell-free components but the template DNA was replaced with water.

ture set to 29 °C, and ran for 12 hours. In addition to this, the Biotek plate reader and gain 50 were used for these fluorescence measurements.

Figure 4.10 shows the results from the cell-free reactions measured on the plate reader. The top chart on figure 4.10 shows all the constructs and the bottom chart is the same, except it excludes the T7p14-deGFP and p70a-deGFP reactions. This is because both these constructs have very high RFU values, and the other constructs are indistinguishable when they are included in the same plot. The first major observation from these plots, is that the reactions for T7p14-deGFP and p70a-deGFP, shown in purple and brown respectively, were highly expressed, as they have very large RFU values

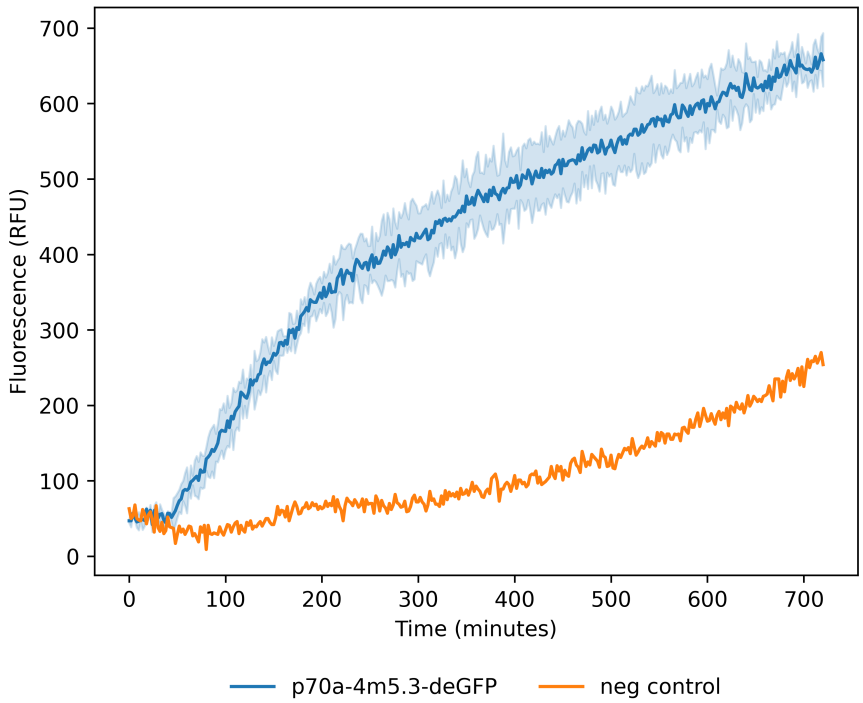

Figure 4.11: Cell-free reactions performed on p70a-4m5.3-deGFP with the MS RGami T7 lysate, SL ES2 energy solution, and a fresh batch of template DNA. A negative control includes the exact same cell-free components except the template DNA is replaced with water.

compared to the negative control shown in red. This means that the reactions were set up correctly and have been able to express constructs with both promoters. The second major observation from these charts is that the scFv constructs; p70a-4m5.3-deGFP, pET24(+)-4m5.3-deGFP-lac and pET24(+)-4m5.3-deGFP-wo-lac, have extremely low expression. These reactions show RFUs that are slightly higher than the negative control, but not by much at all. This means that we cannot be certain that these constructs have even expressed.

Another observation is that even though the T7p14-deGFP and p70a-deGFP constructs were expressed to a similar level, with RFUs roughly around 10,000, they have very different expression curves. The T7p14-deGFP reaction shows a very high rate of expression initially, and then this starts to slow down after 50 minutes. On the other hand, the p70a-deGFP has a much slower rate of expression compared to the T7p14-deGFP reaction; however, it only starts to slow down after 200 minutes. This makes sense as the T7 promoter has a faster transcription rate compared to the p70 promoter, and potentially the reason for the rate of expression slowing down and flattening off,

is due to the depletion of resources. This is because both reactions start to slow down after reaching around 6,000 RFUs, even though the T7p14-deGFP reaches this after 50 minutes, and the p70a-deGFP reaches this level after 200 minutes.

Finally, one other observation from figure 4.10 is that p70a-4m5.3-deGFP had very low expression, even though we managed to obtain a strong signal with this construct using the AP RGami lysate in figure 4.7. PCR was performed again on this construct to obtain a new batch of template DNA, as the DNA template used for the reactions in figure 4.10 was fairly old. These reactions were performed in exactly the same way as in figure 4.10; however, only duplicate reactions were performed for the p70a-4m5.3-deGFP and one reaction for the negative control. Figure 4.11 shows the results from these cell-free reactions, and we see a clear signal from the p70a-4m5.3-deGFP reaction, compared to the negative control. This suggests that the reason for the low expression in the previous experiment was down to the template DNA. On the other hand, the results for pET24(+)-4m5.3-deGFP-lac and pET24(+)-4m5.3-deGFP-wo-lac were repeated multiple times and these consistently had low expression. This suggests that there was an issue with the pET24(+) constructs that was causing them to have low expression. The two main factors that could be causing this issue were the pET24(+) vector itself or the Twist codon optimisation that was used for this DNA sequence.

4.3.5 Cloning 4m5.3 construct into T7p14-deGFP and T7p14-mCherry

Due to the low expression of the pET24(+)-4m5.3-deGFP-lac and pET24(+)-4m5.3-deGFP-wo-lac constructs, it was decided to use Gibson assembly to clone the coding region of 4m5.3 into the T7p14-deGFP and T7p14-mCherry plasmids. This T7p14 plasmid has a T7 promoter and is well used in cell-free systems [364]. Section 4.2.3.11 explains the general procedure of Gibson assembly and it describes the primers that were used for amplifying the different backbone and the insert sequences.

Firstly, before cloning the coding region of 4m5.3-deGFP into the T7p14-deGFP plasmid, the insert and backbone sequences were amplified using PCR (section 4.2.3.1). The PCR protocol, primers and annealing temperatures used, are shown in table 4.3. Two different sets of primers were used for the insert sequence, in case one didn't work. Gel electrophoresis (section 4.2.3.2) was performed on the PCR samples and the gel image is shown in figure 4.12. The lengths of the backbone and insert sequences were 3,445bp and 1,476bp respectively, and from figure 4.12 we can see that the band for the backbone is slightly below the 3.5kb band in the HyperLadder, and

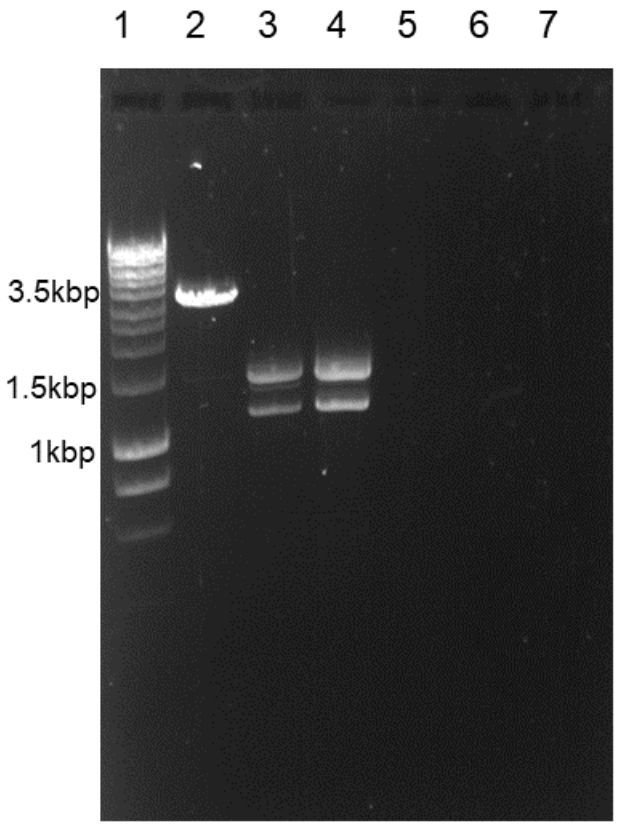

Figure 4.12: PCR gel image of the backbone and insert sequences for cloning into T7p14-deGFP. The HyperLadder is shown in well 2, the backbone PCR sample in well 3, and the insert PCR samples in wells 4 and 5.

the insert sequences both have a band around 1.5kb. Unfortunately, there was another shorter DNA segment that had been amplified for both the insert sequences, as well as the correct sequence. It was decided to proceed with the Gibson assembly using this backbone and the insert sequence in well 4 of the gel in figure 4.12, as colony PCRs will be performed afterwards.

Following on from conducting the Gibson assembly reaction with the backbone and insert sequences, the new plasmid was transformed into NEB DH5 alpha competent cells (section 4.2.3.9). The next morning, colonies were picked the from the plate and colony PCR was performed (section 4.2.3.10), in order to identify which colony has the correct plasmid. The first gel image on the left in figure 4.13 shows the results from performing colony PCRs on 5 different colonies using the primers to amplify the T7p14-deGFP construct in table 4.3. Colonies 3 and 5 appear to have the correct plasmid as the bands from the gel image are slightly higher than 2kbp, and the length of the the linear T7p14-4m5.3-deGFP is around 2,200bp. After this, minipreps

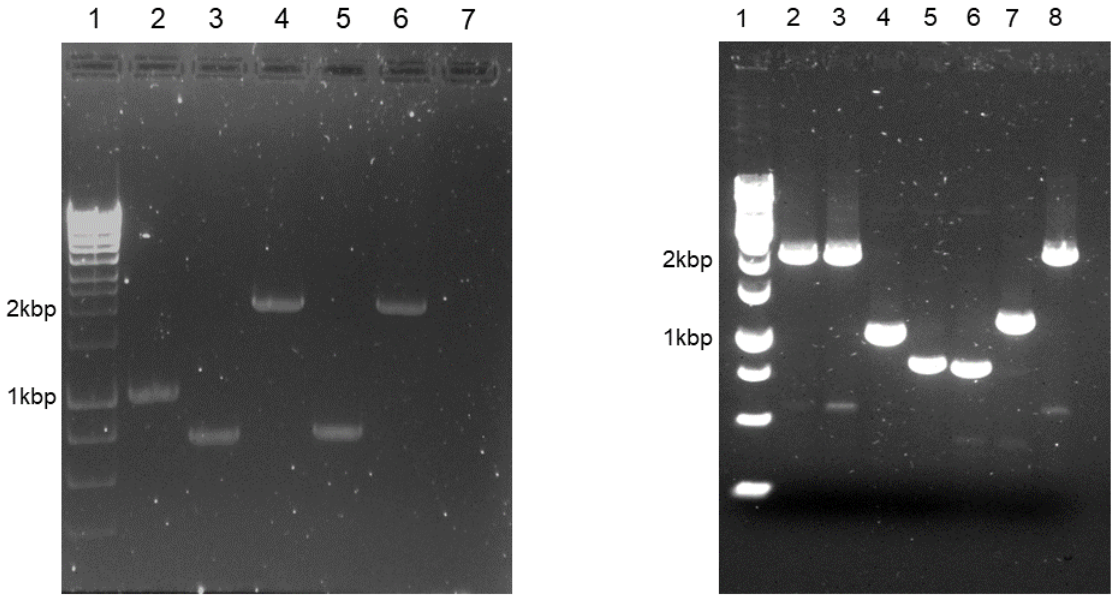

Figure 4.13: The PCR gel image on the left shows the result of colony PCRs for 5 different colonies. HyperLadder is in well 1, colony 1 in well 2, colony 2 in well 3, colony 3 in well 4, colony 4 in well 5 and colony 5 in well 6. On the other hand, the gel image on the right shows PCRs performed on miniprepped plasmids extracted from each colony. HyperLadder is in well 1, colony 3 in well 2, colony 5 in well 3, colony 6 in well 4, colony 7 in well 5, colony 8 in well 6, colony 9 in well 7 and colony 10 in well 8.

were performed (section 4.2.3.5) on colonies 3 and 5, and also 5 additional colonies, colonies 6, 7, 8, 9 and 10. PCRs were then repeated, using the primers and settings for the T7p14-deGFP construct in table 4.3, and the gel image on the right of figure 4.13 shows these results. As expected, we get the correct band for colonies 3 and 5, and additionally we found that colony 10 also had the correct band. These miniprepped samples were kept and glycerol stocks were made for colonies 3, 5 and 10.

In a similar fashion, this protocol was repeated to clone the 4m5.3 antibody DNA sequence into the T7p14-holin-mCherry plasmid. This was performed in order to investigate whether fusing with mCherry gave us a better signal to noise ratio than fusing to GFP, as the crude lysate and energy solution always have a background fluorescence which is captured by the negative controls. For the previous T7p14-deGFP cloning work for the full 4m5.3-deGFP sequence was cloned from pET24(+)-4m5.3-deGFP-wo-lac into the T7p14-deGFP plasmid however, for this cloning work, only the 4m5.3 antibody sequence and the short GSPA linker were cloned into T7p14-holin-mCherry. As this plasmid already contained a holin DNA sequence linked to

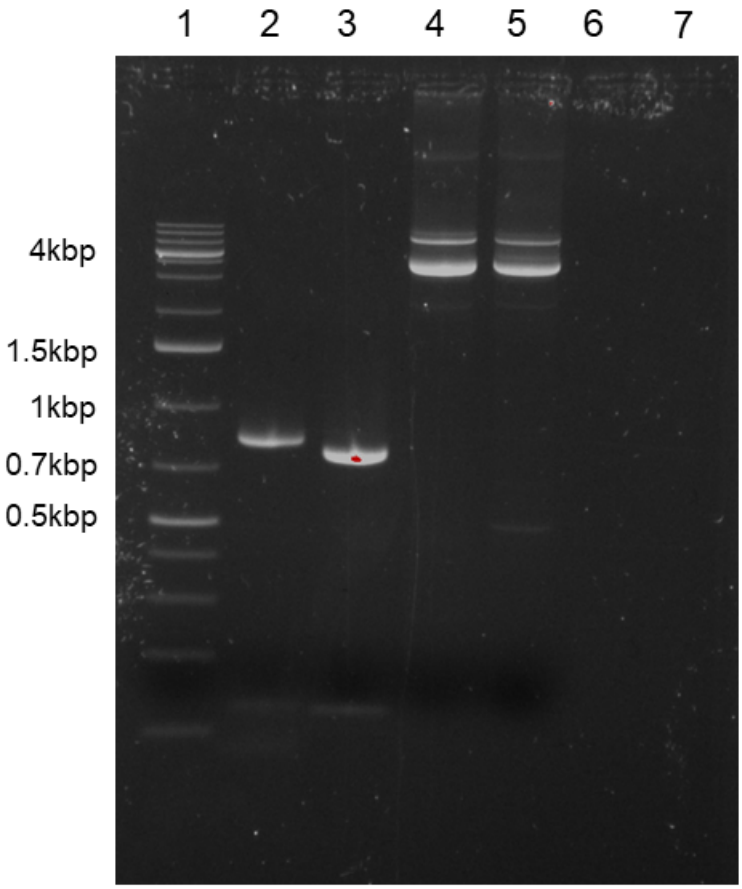

Figure 4.14: PCR gel image of the backbone and insert sequences for cloning into T7p14-holin-mCherry. The GeneRuler 1kb plus is shown in well 1, the GSPA insert in well 2, the GS TEV insert in well 3, the GSPA backbone in well 4 and GS TEV backbone in well 5.

mCherry, the cloning was also repeated by only cloning the 4m5.3 sequence, and using the Glycine-Serine linker with a TEV cut site (GS TEV), that was already present in the T7p14-holin-mCherry sequence. The insert sequence that used the GSPA linker was called the GSPA insert, and the insert sequence that did not include the GSPA linker, was called the GS TEV insert. The PCR protocol, primers and annealing temperatures used, are shown in table 4.4, and the gel image is shown in figure 4.14. As the lengths of the GSPA and GS TEV insert sequences are 774bp and 741bp, and their corresponding backbone sequences have lengths of 4,216bp and 4,225bp respectively, we can see that the DNA is of the correct molecular weight from figure 4.14. However, the gel image also shows some non-specificity for the backbone sequences. The same procedure that was followed for the T7p14-deGFP cloning work was repeated, in order to perform the Gibson assembly, transformation and for picking colonies. Af-

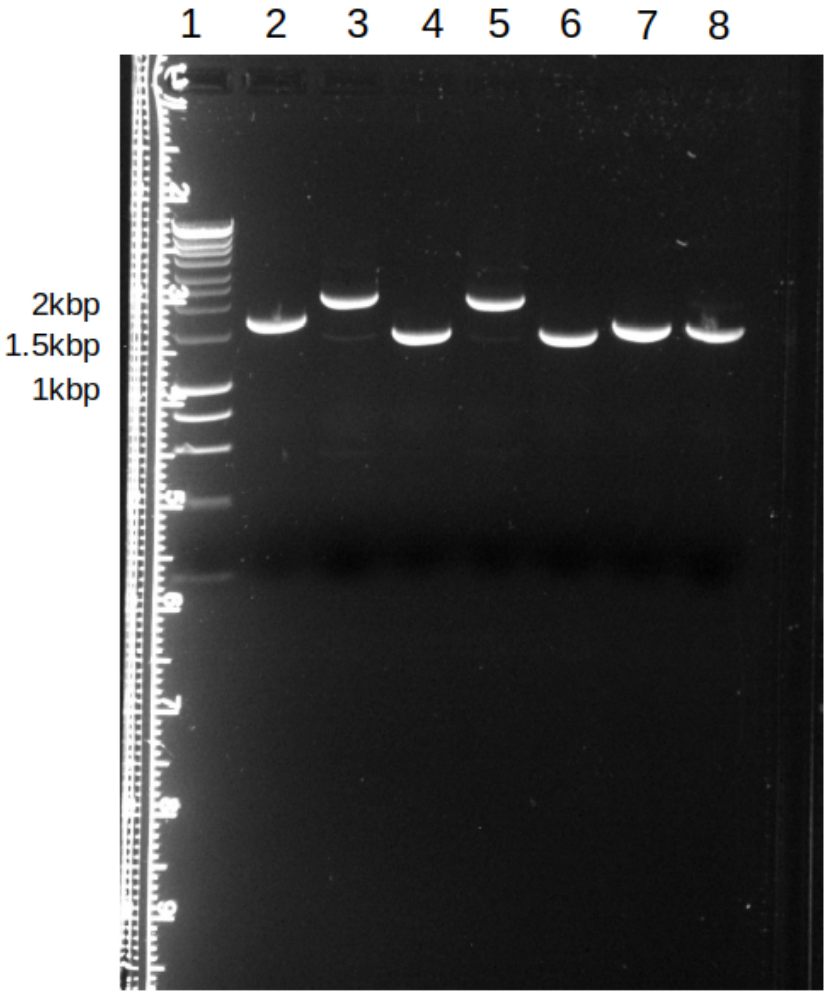

Figure 4.15: The PCR gel image shows PCRs performed on miniprepped plasmids extracted from each colony. HyperLadder 1kb is in well 1, GSPA insert colony 1 in well 2, GSPA insert colony 3 in well 3, GSPA insert colony 4 in well 4, GSPA insert colony 5 in well 5, GS TEV insert colony 2 in well 6, GS TEV insert colony 3 in well 7 and GS TEV insert colony 5 in well 8.

ter minipreps were performed for the chosen colonies, with the GSPA and the GS TEV linkers, PCRs were then repeated, using the primers and settings for the T7p14-mCherry construct in table 4.4. Figure 4.15 shows the gel image for these PCR reactions, and as the T7p14-4m5.3-mCherry construct length is around 2,200bp, we can see from this gel that GSPA insert colony 3 and GSPA insert colony 5 have the correct molecular weight. Unfortunately, none of the chosen colonies had the correct molecular weight for the GS TEV insert. Following on from this, glycerol stocks (section 4.2.3.8) were made for GSPA insert colony 3 and GSPA insert colony 5 and they were

stored in the −70 °C freezer.

4.3.6 Testing T7p14-deGFP and T7p14-mCherry constructs

Firstly, cell-free reactions were performed on the T7p14-4m5.3-deGFP construct, in order to understand whether it could be expressed. These reactions were performed with the MS RGami T7 lysate (table 4.7), along with an energy solution prepared following the ES2 protocol (table 4.9), and cell-free reaction set up #2. Time series and endpoint reactions were performed at 29 °C and gain 50 on the Biotek plate reader. For the time series data, 15 µL reactions were performed in the plate reader over 12 hours. In contrast to this, for the end point data, 20 µL reactions were performed in epdendorfs in the ThermoMixer, and then end point measurements were performed on the plate reader. Figure 4.16 shows the results from these cell-free reactions, with the 15 µL time series reactions on the top and 20 µL end point measurements on the bottom. Both of these plots show a clear signal from the T7p14-4m5.3-deGFP construct compared to the negative control, which is comparable to the signal observed from the p70a-4m5.3-deGFP construct, in the same MS RGami T7 crude lysate system.

After this, cell-free reactions were performed on the T7p14-4m5.3-mCherry construct with the MS RGami T7 lysate, along with an energy solution prepared following the ES2 protocol. The plate reader was set to gain 70, 29 °C, and these reactions were performed in triplicate. The top plot in figure 4.17 shows the results from these cell-free reactions, and we can see that the signal for the T7p14-4m5.3-mCherry construct, roughly 10 times as high as the negative control. However, the T7p14-4m5.3-mCherry signal is very noisy which suggests, we could benefit from increasing the gain on the plate reader. These cell-free reactions were repeated at gain 120 and ran for 24 hours to ensure that the mCherry signal had enough time to mature. The results from these experiments are shown in the bottom plot in figure 4.17, and we can see a much better signal to noise ratio between the T7p14-4m5.3-mCherry signal and the negative control, which is the best achieved so far. As the average RFU across all replicates for the end point of T7p14-4m5.3-mCherry is 3,501, and the average RFU for all replicates for the end point of negative control is 109, we can see that we have a signal which is 32 times higher than the noise. Following on from this, the sequence for T7p14-4m5.3-mCherry was shown to be correct after sending if off for Sanger sequencing [368] with GeneWiz (section 4.2.3.6), and then a set of 29 scFv sequences were ordered in this T7p14 plasmid, fused to mCherry, from Twist Bioscience.

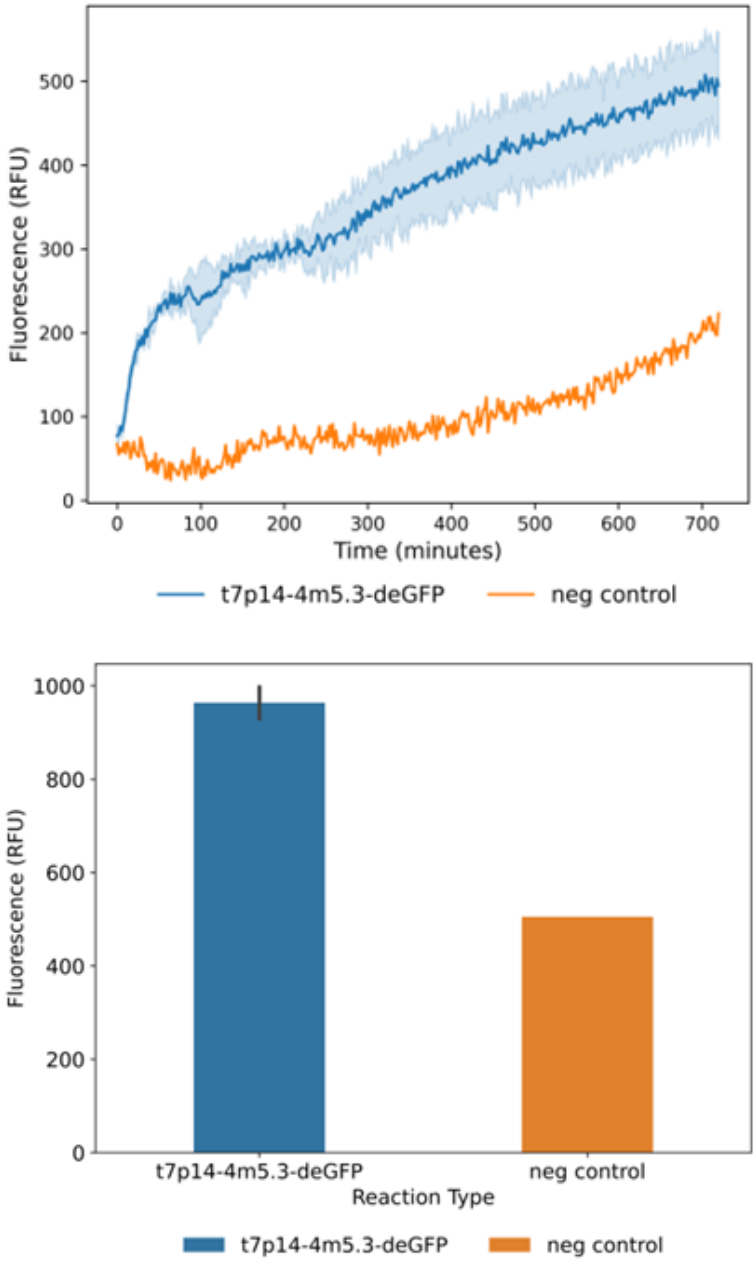

Figure 4.16: The top plot shows the results from performing $15\,\mu\mathrm{L}$ cell-free reactions on the T7p14-4m5.3-deGFP construct and a negative control at gain 50, $29\,^{\circ}\mathrm{C}$ and for 12 hours. The bottom plot shows the results from performing $20\,\mu\mathrm{L}$ cell-free reactions for the same constructs in eppendorfs in a ThermoMixer at $29\,^{\circ}\mathrm{C}$. End point measurements were then taken on the plate reader at gain 50. Negative control includes the same cell-free components, but the template DNA is replaced with water.

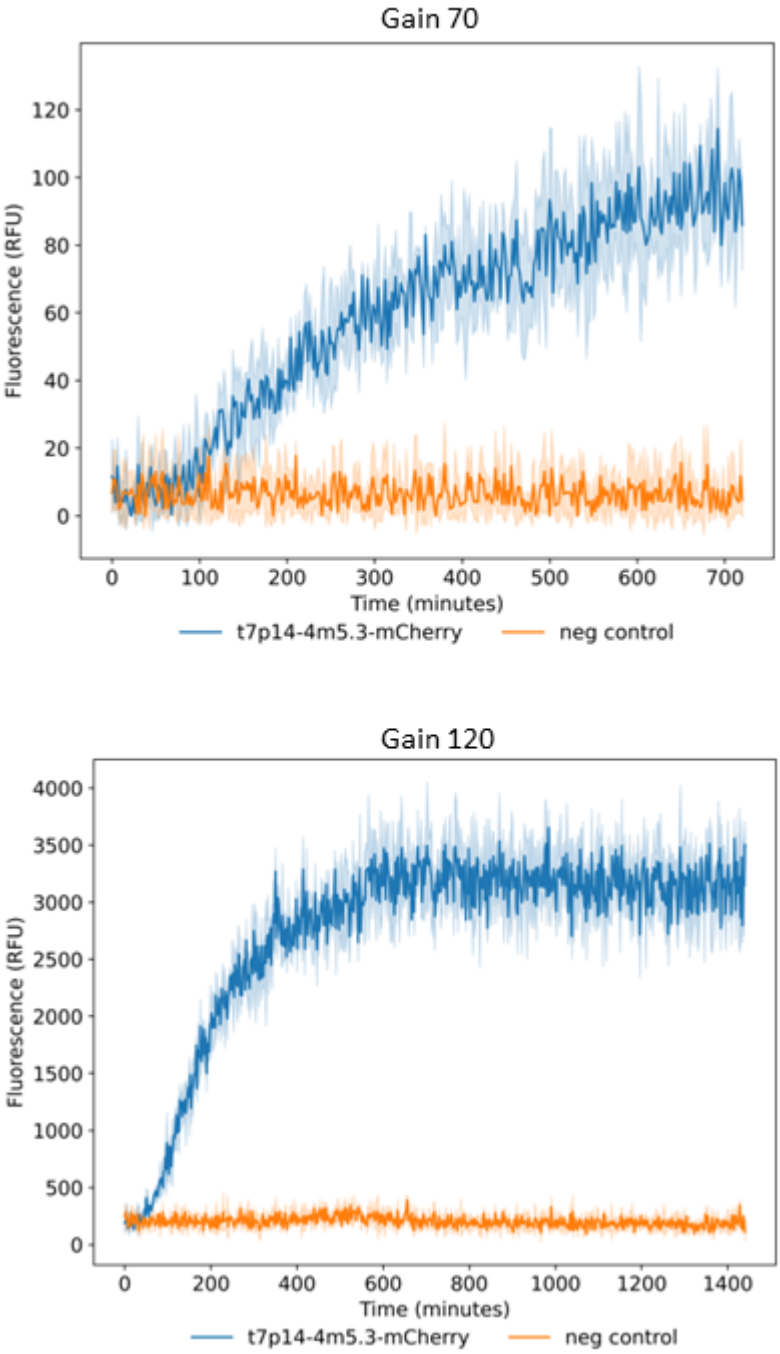

Figure 4.17: This top plot shows the results from performing $15\,\mu\mathrm{L}$ cell-free reactions on the T7p14-4m5.3-mCherry construct and a negative control at gain 70, $29\,^{\circ}\mathrm{C}$ and for 12 hours. The bottom plot shows the same reactions repeated but at gain 120. Negative control includes the same cell-free components, but the template DNA is replaced with water.

Design Name	Design Cycle	Target Antigen	Yeast Surface Expression Levels (% cells exhibiting higher fluorescence levels than negative control) [159]
4m5.3 (test)		Fluorescein	72.1
1ins01	1	Insulin	13.8
1ins02	1	Insulin	0.20
1ins06	1	Insulin	4.60
2ins19	2	Insulin	36.1
2ins23	2	Insulin	33.9
2ins25	2	Insulin	7.50
2ins26	2	Insulin	21.4
2acp05	2	Acyl-carrier protein	73.2
2acp10	2	Acyl-carrier protein	8.50
2acp16	2	Acyl-carrier protein	73.8
2acp17	2	Acyl-carrier protein	65.0
3ins03	3	Insulin	35.5
3ins06	3	Insulin	29.0
3ins20	3	Insulin	53.6
3ins28	3	Insulin	21.6
3ins30	3	Insulin	43.1
3acp04	3	Acyl-carrier protein	5.70
3acp15	3	Acyl-carrier protein	15.0
3acp20	3	Acyl-carrier protein	11.0
3acp25	3	Acyl-carrier protein	33.7
3acp28	3	Acyl-carrier protein	45.6
4ins02	4	Insulin	59.8
4ins14	4	Insulin	27.5
5ins01	5	Insulin	75.6
5ins14	5	Insulin	80.3
5ins16	5	Insulin	72.5
5acp04	5	Acyl-carrier protein	78.9
5acp05	5	Acyl-carrier protein	80.9

Continued on next page

Table 4.13 – *Continued from previous page*

Design Name	**Design Cycle**	**Target Antigen**	**Yeast Surface Expression Levels (% cells exhibiting higher fluorescence levels than negative control) [159]**

Table 4.13: A set of 29 scFv sequences from Baran et al. [159], that were ordered in the T7p14 plasmid and fused to mCherry. Stratified sampling across the target antigen (insulin and *M. tuberculosis* acyl-carrier protein) and yeast display expression levels was performed to obtain this sample.

Table 4.13 shows the ID of each scFv sequence from Baran et al. [159] that was ordered, along with the design cycle, the target antigen and the expression levels in yeast display. The sequences were sampled across the different target antigens and yeast display expression levels, and the 4m5.3 test sequence was also ordered as a control. This set of sequences will be used in future experiments, with the Edinburgh Genome Foundry, to collect additional cell-free expression data.

4.4 Discussion

Cell-free systems have a number of advantages over cell-based methods for protein production, such as the level of control they offer for understanding the factors that influence transcription and translation [190; 191; 192; 193]. This means that they offer a unique opportunity to explore some of the reasons why designed proteins fail to be produced, if these reasons are related to transcription and translation. The experimental work in this chapter, focused on setting up *E. coli* cell-free systems, to collect expression data from a set of designed single chain variable fragment antibodies (scFvs), from the Fleishman lab, which had varying levels of *in vivo* expression [159]. In order to collect these expression levels, fluorescent protein labelling was used with deGFP and mCherry sequences. A major advantage of fluorescent protein labelling is that it allows the high-throughput measurement of expression levels, and can be used to collect a lot of data in a relatively short amount of time [199]. This technique is more suitable for screening larger sets of sequences, in comparison to low-throughput methods such as Western blots [372]. In addition to this, the coding region for the fluorescent protein is placed downstream from the coding region of the protein of interest, which means that

if fluorescence is observed, then we can conclude that the protein of interest must have been translated. However, a disadvantage of this method for screening protein production levels, is that it does not tell us anything about whether the protein of interest has folded correctly, or whether it is functional. This is because it is quite possible that the scFv protein misfolds while the fluorescent protein folds correctly, and we would still observe fluorescence.

Before performing any experimental work, AlphaFold2 [65] was used to obtain structural models, for the amino acid sequence of the test 4m5.3 scFv linked to deGFP. This was performed to check whether the two proteins were predicted to fold independently and not misfold. Both of these structural models show that the scFv and the fluorescent proteins were predicted to fold correctly, and that these models are highly confident, with average predicted local distance difference test (pLDDT) scores of 88.8 for 4m5.3-deGFP and 86.6 for 4m5.3-mCherry. These results suggest that these fluorescent proteins could be successfully linked to the designed scFvs in order to measure expression levels; however, although these structural models are highly confident, they do not replace experimental validation [373]. In addition to this, as AlphaFold2 was trained on experimentally-determined structures from the PDB [65; 47], it is possible that it could be biased towards predicting folded structures for these sequences.

The main output of the experimental work completed in this chapter, is a protocol for screening the protein production levels of scFv sequences in *E. coli* cell-free systems. This protocol will be used in the future to collect a data set of protein production levels, and it could potentially be applied to other proteins as well. The Rosetta-gami 2 *E. coli* strain was used to create the crude lysates used in the cell-free systems, as this strain allows for enhanced disulfide bond formation and is better for expressing eukaryotic proteins, which contain codons rarely used in standard *E. coli* strains [374]. In addition to this, crude lysate cell-free systems have been created from Rosetta-gami *E. coli* strains before, and have been shown to work for scFvs [359]. In order to set up this protocol for measuring the production levels of proteins, a range of DNA constructs for the test 4m5.3 scFv sequence, with different promoters and fluorescent proteins, were used. Various tests were performed on these sequences in cell-free systems, with the aim of optimising the signal to noise ratio of the protein production levels. Cell-free systems generally have a background fluorescence, which is measured by performing negative controls, and this helps us understand whether the signal from our test sequence is higher than this background noise. As this protocol will be used to screen a larger set of sequences, we wanted the signal to noise ratio to be as high as possi-

ble, so that it will allow us to distinguish between the different sequences. Overall, it was found that the T7p14-4m5.3-mCherry construct obtained the best signal to noise ratio (33:1) of any of the constructs that we tested in this project, and therefore, it was decided to use the T7p14-mCherry plasmid for screening the larger set of scFv sequences. However, more optimisation could have been performed on the cell-free systems, such as optimising the concentrations of magnesium glutamate, potassium glutamate and other components in the energy solution, as these can have a large effect on the performance of the cell-free systems [364]. Therefore, it is possible that with more optimisation in the future, we could increase this signal to noise ratio even more.

4.5 Next steps

As the T7p14-4m5.3-mCherry construct achieved a signal that was 33 times higher than the negative control in the *E. coli* cell-free system (figure 4.17), which was the highest achieved from the constructs used in this project, a larger set of 29 designed scFvs sequences was ordered from Twist Bioscience in this plasmid. These sequences were sampled across the different design cycles and yeast display expression levels found in the Baran et al. study [159], with the aim of using these to obtain an informative *E. coli* cell-free expression data set. After obtaining these sequences, the Edinburgh Genome Foundry will be used to perform cell-free reactions using acoustic liquid-handling robots. Multiple *E. coli* cell-free lysates will be made, so that we can collect biological replicates as well as technical replicates for each construct, as cell-free systems can have a large batch to batch variation [375]. Generally, the concentrations of magnesium glutamate, potassium glutamate and other components of the energy solution, need to be optimised for each batch of lysate [364], which was not performed for this experimental work. These optimisations will be performed for future crude lysate systems, and this could help to increase the expression, along with the signal to noise ratio, of the scFv constructs. Overall, this additional work will result in a robust *E. coli* cell-free expression data set for these designed scFvs, which could help to provide insight into some of the reasons why designed proteins fail to be produced.

Following on from this, the measure of protein production collected from these experiments, will be used as a predictor variable to train models for predicting protein production. Additionally, the DE-STRESS metrics [197] of the designed scFv AlphaFold2 structural models [65], will be used as features for predicting these protein

production levels. These models will then be compared to models trained to predict the yeast display protein production measure from Baran et al. [159], in order to understand if there are any differences and similarities, in the factors that are predictive of protein production. After this, we will explore whether there are any designed scFvs that fail to be produced in both expression systems, and whether any of the DE-STRESS features can identify these designs. On the other hand, we will also explore whether there are designs that fail in one expression system, but are successfully produced in the other. Both of these questions could help us understand some of the factors that influence protein production, and if there are some designs that will fail in multiple different expression systems. In addition to this, the developed models could be used to rank *de novo* scFvs, that are generated from a sequence design method such as TIMED [150]. These designs could then be tested in the same cell-free systems to validate them experimentally, and to see whether these models have helped to reduce the failure rate of designs. On the whole, the insights gained from analysing this cell-free expression data, could be incredibly important for understanding the factors that can cause low protein production, and these insights could potentially help develop more reliable design methods in the future.

Finally, after collecting this *E. coli* cell-free expression data for the set of designed scFvs, we will perform additional experimental work with the aim of understanding some of the factors that contribute to the low production of designed proteins. As cell-free systems will be used for this analysis, we can only investigate factors associated with transcription and translation, such as, truncation, aggregation, resource depletion and misfolding of the proteins, and not other reasons, for example, the toxicity of proteins to cells. Truncations could be identified using SDS page gels [376] or mass spectrometry analysis using isotopically labeled amino acids [377] and aggregates could be identified using size exclusion chromatography [378]. Furthermore, high-performance liquid chromatography (HPLC) could be used to measure the concentrations of NTPs and amino acids in the cell-free systems, to understand whether the resources of the system are being depleted [379]. These experiments would be performed for samples of low and high producing scFv designs, in these cell-free systems, in order to compare which factors are important. As a result of these additional experiments, we would gain further insight into some of the main contributors to low protein production, for this set of designed scFvs, which potentially could be useful for understanding the variables that affect the low production rates of designed proteins in general. Once we have gained some insight into the factors that cause low protein production, we could incor-

porate this insight into protein design methods, with the aim of avoiding these failure reasons and increasing the success rate of designs.

4.6 Conclusion

In conclusion, the main result from the experimental work in this chapter, is the development of a method to screen scFv designs for protein production levels, using *E. coli* cell-free systems and fluorescent proteins. The initial test construct that was ordered, p70a-4m5.3-deGFP, was successfully expressed in cell-free systems; however, the p70a promoter could not be used for scaling up these experiments, as it would have caused problems for Twist Bioscience's production pipeline. After this, constructs were ordered from Twist Bioscience in the pET24(+) plasmid, which has a T7 promoter and can be manufactured by their production pipeline. However, both of these constructs, pET24(+)-4m5.3-deGFP-lac and pET24(+)-4m5.3-deGFP-wo-lac, had very low expression in the cell-free systems. In order to address this issue, Gibson assembly was used to clone the coding region of pET24(+)-4m5.3-deGFP-wo-lac into the T7p14-deGFP plasmid, as this is well used in cell-free reactions and was shown to express well in our cell-free systems. The T7p14-4m5.3-deGFP construct was shown to express in the cell-free reactions; however, it still had fairly low expression and the signal was only double the negative control.

Following on from this, Gibson assembly was used to create the T7p14-4m5.3-mCherry construct, which showed a higher signal to noise ratio of 10:1 in the cell-free reactions, with a gain setting of 70 on the plate reader. By increasing the gain setting to 120, we obtained a much higher signal to noise ratio of 33:1, which was the best achieved in this project. These results show that the best vector to use for ordering a larger set of scFv designs was the T7p14-mCherry plasmid, as it was able to be expressed in cell-free reactions, Twist Bioscience was able to produce it, and fusing to mCherry gave a lot less background noise than deGFP, when measuring fluorescence. Therefore, a set of 29 designed scFv sequences were ordered in this vector, which will be used to collect additional cell-free expression data. Finally, in future work, this cell-free expression data will be used to train models to predict protein production, with the DE-STRESS metrics of the scFv AlphaFold2 structural models being used as features. These models will then be compared to models trained to predict the yeast display protein production measure from Baran et al. [159], and further experimental work will be performed to understand the factors causing low protein production, such

as truncation, aggregation, resource depletion and misfolding of the protein. Overall, this work will provide insight into some of the reasons why designed proteins fail to be produced, with the aim of using these insights in the future, to improve the design process, and make protein design more reliable and accessible.

Chapter 5

Conclusions and future perspectives

In conclusion, the work in this PhD thesis has focused on developing methods to address some of the current limitations of protein design, in order to make it cheaper, more reliable and accessible for researchers. There are three main outputs from this PhD project that have been covered in chapters 2, 3 and 4, and these outputs aim to help address the high failure rate of designed proteins, the difficulty with designing towards properties and functions, and the high level of expertise needed to use the majority of protein design computational tools.

Firstly, the DE-STRESS webserver was developed [197], which calculates a set of structural features capturing physico-chemical properties of designed proteins, so that users can rank their designs before testing them experimentally in the lab. This work is described in detail in chapter 2. DE-STRESS offers novel functionality through reference sets and specifications, which provide context for the DE-STRESS metrics, and help users design towards properties and functions. Additionally, analysis was performed that showed DE-STRESS could distinguish between native and decoy structures, with a range of different metrics being important for this, which suggests that DE-STRESS could be useful for ranking designs. Two protein redesign projects were also performed using a sequence design method from our lab called TIMED [150], and DE-STRESS was used to rank designs for experimental testing. However, we still need to validate if DE-STRESS can reduce the failure rate of designs once we obtain experimental results. Furthermore, in the future, additional metrics will be incorporated into DE-STRESS, and it will be developed into a user friendly pipeline, along with sequence design and structure prediction methods, to provide an accessible protein design pipeline for a wide range of people.

Secondly, chapter 3 describes analysis that was performed on the DE-STRESS

structural features for different sets of proteins. These properties were shown to be predictive of *in vivo* protein production levels for a set of designed proteins. This result is significant as it provides evidence that DE-STRESS could be useful for ranking designed proteins; however, this will have to be extended to larger data sets of designs, additional measures of protein production, and different classes of proteins as well. After this, these metrics were ran across 160,000 proteins from the Protein Data Bank (PDB) [47], and 500,000 predicted structures from the AlphaFold DB [67]. One observation from these data sets was that the physico-chemical properties of designed proteins, were skewed compared to those of native proteins, which could be important for improving the success rate of designs, by designing proteins that look more "native like". Additionally, these properties were found to vary across 500,000 predicted structures from 48 organisms, to such a degree that the tree of life could be reconstructed, even with sequence information excluded. To our knowledge, this is first time this has been shown, and implies that the properties of proteins have evolved to their unique molecular environment. This result indicates that design methods may benefit from incorporating information about the intended environment of the designed protein, such as the organism or sub-cellular location. However, further work needs to be performed to validate these results experimentally, to investigate it for different groups of proteins, and to explore whether incorporating this information into design methods could help reduce the failure rate of designs, and to design towards functions and properties.

Finally, chapter 4 describes initial experimental work that was performed to set up *E. coli* cell-free reactions, for measuring the protein production levels of a set of proteins designed by Baran et al. [159], with fluorescent protein labelling. Various DNA constructs were tested in cell-free systems for a test design, with different promoters and fluorescent proteins, and the T7p14 plasmid along with mCherry gave the best signal to noise ratio. After this, 29 of these designed proteins were ordered in the T7p14 plasmid, and attached to mCherry with a short linker. In future work, high-throughput cell-free reactions will be performed with the Edinburgh Genome Foundry on these sequences, in order to collect a small data set of *E. coli* cell-free protein production levels. This data set will be compared to the yeast display protein production levels, that were collected for this set of designed proteins in Baran et al. [159], in order to understand whether the same DE-STRESS structural features are predictive of both protein production levels. After this, the insight gained from analysing this data set will be used to design *de novo* proteins, which will then be validated in the same cell-free systems. A major advantage of fluorescent protein labelling is its suitability for

high-throughput screening; however, a limitation is that it does not tell us anything about whether the protein of interest is correctly folded. On the other hand, using cell-free reactions allows us to explore some of the main reasons of failure (if they are related to transcription and translation) in the future for these designed proteins, using techniques such as mass spectrometry, which could be used to inform design methods to help avoid the most common reasons for failure.

Overall, the work presented in this PhD thesis shows how structural features of proteins, machine learning and cell-free expression systems can be used to help address some of the limitations of protein design, so that it can become more widely used as a technique. Protein design has already been shown to have a huge potential across areas such as medicine, agriculture, and biotechnology, and recently machine/deep learning has transformed the field, providing protein design techniques with increased performance, that are less computationally expensive than physics based methods. If these current limitations can be addressed, then protein design can become cheaper, more reliable and more accessible to researchers, which will cause a revolution across different scientific areas, where custom proteins can be designed to solve a wide range of complex problems.

Appendix A

Glossary of DE-STRESS metrics

Programme name	Description	Command Used	Citations
Aggrescan3D 2.0 (v1.0.2)	Aggregation Propensity	`aggrescan-ipdb_file_path-wo utput-D10-v4`	Kuriata et al. (2019). Aggrescan3D standalone package for structure-based prediction of protein aggregation properties. Bioinformatics 35, 3834–3835.

Continued on next page

Table A.1 – *Continued from previous page*

Programme name	**Description**	**Command Used**	**Citations**

BUDE (v1.0.0)	Energy Function	`import ampal`
		`import budeff`
		`design = ampal.load_pdb(pdb_string, path=False)`
		`budeff.get_internal_energy(design)`
			McIntosh-Smith et al. (2012). Benchmarking Energy Efficiency, Power Costs and Carbon Emissions on Heterogeneous Systems. The Computer Journal 55, 192–205. McIntosh-Smith et al. (2015). High performance in silico virtual drug screening on many-core processors. The International Journal of High Performance Computing Applications 29, 119–134.

DFIRE2-pair	Energy Function	`calene dfire_pair.lib pdb_file_path`	Yang et al. (2008). Ab initio folding of terminal segments with secondary structures reveals the fine difference between two closely related all-atom statistical energy functions. Protein Science 17, 1212–1219.

Continued on next page

Table A.1 – *Continued from previous page*

Programme name	**Description**	**Command Used**	**Citations**

DSSP (v2.0.4)	Secondary Structure Assignment	`import isambard.evaluation as ev`
		`import ampal`
		`design=ampal.load_pdb (pdb_string, path=False)`
		`ev.tag_dssp_data(design)`
			Kabsch et al. (1983). Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 22, 2577–2637. Touw et al. (2015). A series of PDB-related databanks for everyday needs. Nucleic Acids Research 43, D364–D368.

EvoEF2 (EvoEF v2)	Energy Function	`EvoEF2 --command = Compute Stability --pdb = pdb_file_path`	Huang et al. (2020). EvoEF2: accurate and fast energy function for computational protein design. Bioinformatics 36, 1135–1142.

Continued on next page

Table A.1 – *Continued from previous page*

Programme name	**Description**	**Command Used**	**Citations**
Hydrophobic Fitness (ISAMBARD v2.3.1)	Energy Function	`import isambard. evaluation as ev import ampal design=ampal. load_pdb (pdb_string, path=False) ev.calculate_ hydrophobic_ fitness(design)`	Huang et al. (1995). Recognizing native folds by the arrangement of hydrophobic and polar residues. J Mol Biol 252, 709–720. Wood et al. (2017). ISAMBARD: an open-source computational environment for biomolecular analysis, modelling and design. Bioinformatics 33, 3043–3050.

Continued on next page

Table A.1 – *Continued from previous page*

Programme name	**Description**	**Command Used**	**Citations**
Packing Density (ISAMBARD v2.3.1)	Geometric Analysis	`import isambard. evaluation as ev import ampal design=ampal. load_pdb (pdb_string, path=False) ev.tag_packing_ density(design) mean_ packing_density = np.mean ([a.tags ["packing density"] for a in design.get_ atoms() if a.element != "H"])`	Weiss (2007). On the interrelationship between atomic displacement parameters (ADPs) and coordinates in protein structures. Acta Crystallogr D Biol Crystallogr 63, 1235–1242. Wood et al. (2017). ISAMBARD: an open-source computational environment for biomolecular analysis, modelling and design. Bioinformatics 33, 3043–3050.

Continued on next page

Table A.1 – *Continued from previous page*

Programme name	**Description**	**Command Used**	**Citations**
Rosetta (ref2015)	Energy Function	`rosetta_src_2020.08.61146_bundle/main/source/bin/score_jd2.linuxgccrelease -in:file:spdb_file_path -ignore_unrecognized_res -scorefile_format json`	Alford et al. (2017). The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design. J. Chem. Theory Comput. 13, 3031–3048.

Table A.1: This table details the different programmes used in DE-STRESS, along with the command used to run them and references.

Metric Name	**Metric Description**	**Variable Name in CSV Output**
Total Score	This value is a global indicator of the aggregation propensity/solubility of the protein structure. It depends on the protein size. It allows assessing changes in solubility promoted by amino acid substitutions in a particular protein structure. The more negative the value, the higher the global solubility.	aggrescan3d_total_value

Continued on next page

Table A.2 – *Continued from previous page*

Metric Name	**Metric Description**	**Variable Name in CSV Output**
Average Score	This value is a normalized indicator of the aggregation propensity/solubility of the protein structure. It allows comparing the solubility of different protein structures. It also allows assessing changes in solubility promoted by amino acid substitutions in a particular protein structure. The more negative the value, the higher the normalized solubility.	aggrescan3d_avg_value
Minimum Score	This is the value of the most soluble residue in the structural context.	aggrescan3d_min_value
Maximum Score	This is the value of the most aggregation-prone residue in the structural context.	aggrescan3d_max_value

Table A.2: This table provides a description of the metrics from the Aggrescan3D 2.0 programme and their variable names in the CSV output file.

Metric Name	**Metric Description**	**Variable Name in CSV Output**
Total Energy	This value is the total BUDE force field energy. It is the sum of the steric, desolvation, and charge components.	budeff_total
Steric Energy	This value is the steric component of the BUDE force field energy. It is calculated with a simplified Leonard-Jones potential. It is softer than the steric component of many other force fields.	budeff_steric
Desolvation Energy	This value is the desolvation component of the BUDE force field energy.	budeff_desolvation
Charge Energy	This value is the charge component of the BUDE force field energy.	budeff_charge

Table A.3: This table provides a description of the metrics from the BUDE programme and their variable names in the CSV output file.

Metric Name	Metric Description	Variable Name in CSV Output
Total Energy	This value is the total DFIRE2 energy. This is the only field that is returned from running DFIRE2 on a PDB file.	dfire2_total

Table A.4: This table provides a description of the metrics from the DFIRE2 programme and their variable names in the CSV output file.

Metric Name	Metric Description	Variable Name in CSV Output
H	α-helix	ss_prop_alpha_helix
B	Isolated β-bridge	ss_prop_beta_bridge
E	Extended β-strand	ss_prop_beta_strand
G	3-10 helix	ss_prop_3_10_helix
I	π-helix	ss_prop_pi_helix
T	Hydrogen-bonded turn	ss_prop_hbonded_turn
S	Bend	ss_prop_bend
-	Loop	ss_prop_loop

Table A.5: This table provides a description of the metrics from the DSSP programme and their variable names in the CSV output file.

Metric Name	Metric Description	Variable Name in CSV Output
Total Energy	This value is the total EvoEF2 energy. It is the sum of the reference, intra residue, inter residue - same chain, and inter residue - different chains energy values. In the EvoEF2 output, this field is called Total.	evoef2_total
Reference Energy	This value is the total reference energy. This value is not included in the EvoEF2 output and is calculated in DE-STRESS.	evoef2_ref_total
Intra Residue Energy	This value is the total energy for intra residue interactions. This value is not included in the EvoEF2 output and is calculated in DE-STRESS.	evoef2_intraR_total
Inter Residue - Same Chain Energy	This value is the total energy for inter residue interactions in the same chain. This value is not included in the EvoEF2 output and is calculated in DE-STRESS.	evoef2_interS_total
Inter Residue - Different Chains Energy	This value is the total energy for inter residue interactions in different chains. This value is not included in the EvoEF2 output and is calculated in DE-STRESS.	evoef2_interD_total

Table A.6: This table provides a description of the metrics from the EvoEF2 programme and their variable names in the CSV output file.

Metric Name	Metric Description	Variable Name in CSV Output
Charge	This value is the total charge of the protein sequence.	charge
Hydrophobic Fitness	This value is an efficient centroid-based method for calculating the packing quality of the protein structure. For this method C, F, I, L, M, V, W and Y are considered hydrophobic.	hydrophobic_fitness

Continued on next page

Table A.7 – *Continued from previous page*

Metric Name	**Metric Description**	**Variable Name in CSV Output**
Isoelectric Point	This value is the pH of a solution at which the net charge of the protein becomes zero.	isoelectric_point
Packing Density	This value is a the mean packing density of the protein structure, where the packing density of a non-hydrogen atom, is defined as the number of non-hydrogen atoms within a specified radius, for example a radius of 7Å.	packing_density

Table A.7: This table provides a description of the metrics from the ISAMBARD programme and their variable names in the CSV output file.

Metric Name	**Metric Description**	**Variable Name in CSV Output**
Total Energy	This value is the total Rosetta energy. It is a weighted sum of the different Rosetta energy values. In the Rosetta `score.sc` output file, this value is called `total_score`.	rosetta_total
Reference	This value is the reference energy for the different amino acids. In the Rosetta `score.sc` output file, this value is called `ref`.	Not included in CSV output
VDW Attractive	This value is the attractive energy between two atoms on different residues separated by distance, d. In the Rosetta `score.sc` output file, this value is called `fa_atr`.	rosetta_fa_atr
VDW Repulsive	This value is the repulsive energy between two atoms on different residues separated by distance, d. In the Rosetta `score.sc` output file, this value is called `fa_rep`.	rosetta_fa_rep

Continued on next page

Table A.8 – *Continued from previous page*

Metric Name	**Metric Description**	**Variable Name in CSV Output**
VDW Repulsive Intra Residue	This value is the repulsive energy between two atoms on the same residue separated by distance, d. In the Rosetta `score.sc` output file, this value is called `fa_intra_rep`.	rosetta_fa_intra_rep
Electrostatics	This value is the energy of interaction between two non-bonded charged atoms separated by distance, d. In the Rosetta `score.sc` output file, this value is called `fa_elec`.	rosetta_fa_elec
Solvation Isotropic	This value is the Gaussian exclusion implicit solvation energy between protein atoms in different residues. In the Rosetta `score.sc` output file, this value is called `fa_sol`.	rosetta_fa_sol
Solvation Anisotropic Polar Atoms	This value is the orientation-dependent solvation of polar atoms assuming ideal water geometry. In the Rosetta `score.sc` output file, this value is called `lk_ball_wtd`.	rosetta_lk_ball_wtd
Solvation Isotropic Intra Residue	This value is the Gaussian exclusion implicit solvation energy between protein atoms in the same residue. In the Rosetta `score.sc` output file, this value is called `fa_sol_intraR`.	rosetta_fa_intra_sol_xover4
HB Long Range Backbone	This value is the energy of long-range hydrogen bonds. In the Rosetta `score.sc` output file, this value is called `hbond_lr_bb`.	rosetta_hbond_lr_bb
HB Short Range Backbone	This value is the energy of short-range hydrogen bonds. In the Rosetta `score.sc` output file, this value is called `hbond_sr_bb`.	rosetta_hbond_sr_bb
HB Backbone Sidechain	This value is the energy of backbone-side chain hydrogen bonds. In the Rosetta `score.sc` output file, this value is called `hbond_bb_sc`.	rosetta_hbond_bb_sc

Continued on next page

Table A.8 – *Continued from previous page*

Metric Name	**Metric Description**	**Variable Name in CSV Output**
HB Sidechain Sidechain	This value is the energy of side chain-side chain hydrogen bonds. In the Rosetta `score.sc` output file, this value is called `hbond_sc`.	rosetta_hbond_sc
Disulfide Bridges	This value is the energy of disulfide bridges. In the Rosetta `score.sc` output file, this value is called `dslf_fa13`.	rosetta_dslf_fa13
Backbone Torsion Preference	This value is the probability of backbone ϕ, ψ angles given the amino acid type. In the Rosetta `score.sc` output file, this value is called `rama_prepro`.	rosetta_rama_prepro
Amino Acid Propensity	This value is the probability of amino acid identity given the backbone ϕ, ψ angles. In the Rosetta `score.sc` output file, this value is called `p_aa_pp`.	rosetta_p_aa_pp
Dunbrack Rotamer	This value is the probability that a chosen rotamer is native-like given backbone ϕ, ψ angles. In the Rosetta `score.sc` output file, this value is called `fa_dun`.	rosetta_fa_dun
Omega Penalty	This value is a backbone-dependent penalty for cis ω dihedrals that deviate from 0° and trans ω dihedrals that deviate from 180°. In the Rosetta `score.sc` output file, this value is called `omega`.	rosetta_omega
Open Proline Penalty	This value is a penalty for an open proline ring and proline ω bonding energy. In the Rosetta `score.sc` output file, this value is called `pro_close`.	rosetta_pro_close
Tyrosine χ_3 Dihedral Angle Penalty	This value is a sinusoidal penalty for non-planar tyrosine χ_3 dihedral angle. In the Rosetta `score.sc` output file, this value is called `yhh_planarity`.	rosetta_yhh_planarity

Continued on next page

Table A.8 – *Continued from previous page*

Metric Name	**Metric Description**	**Variable Name in CSV Output**

Table A.8: This table provides a description of the metrics from the Rosetta programme and their variable names in the CSV output file.

Appendix B

Predicting scFv protein production

Scaling Method	Amino Acid Composition Metrics	Feature Selection Method	Test ROC AUC	Test Precision	Test Recall
Standard	Included	Random Forest	0.785	0.683	0.604
Robust	Included	Random Forest	0.785	0.683	0.604
Minmax	Included	Random Forest	0.785	0.683	0.604
Standard	Included	Mutual Information	0.765	0.728	0.604
Robust	Included	Mutual Information	0.765	0.728	0.604
Minmax	Included	Mutual Information	0.765	0.728	0.604
Standard	Excluded	Mutual Information	0.757	0.562	0.521
Robust	Excluded	Mutual Information	0.757	0.562	0.521
Minmax	Excluded	Mutual Information	0.757	0.562	0.521
Standard	Excluded	Random Forest	0.733	0.531	0.500
Robust	Excluded	Random Forest	0.733	0.531	0.500
Minmax	Excluded	Random Forest	0.733	0.531	0.500

Table B.1: Validation metrics for the Naive Bayes models on the different test sets. These metrics are sorted by the mean ROC AUC score.

Scaling Method	Amino Acid Composition Metrics	Feature Selection Method	Mean Cross Validation ROC AUC	Mean Cross Validation Precision	Mean Cross Validation Recall
Minmax	Included	Random Forest	0.762	0.635	0.545
Standard	Included	Random Forest	0.758	0.628	0.535
Robust	Included	Random Forest	0.757	0.625	0.538
Standard	Excluded	Random Forest	0.750	0.624	0.536
Minmax	Excluded	Random Forest	0.746	0.623	0.532
Robust	Excluded	Random Forest	0.744	0.626	0.527
Robust	Included	Mutual Information	0.743	0.647	0.518
Minmax	Excluded	Mutual Information	0.742	0.612	0.539
Standard	Excluded	Mutual Information	0.741	0.604	0.536
Robust	Excluded	Mutual Information	0.740	0.610	0.542
Standard	Included	Mutual Information	0.735	0.646	0.525
Minmax	Included	Mutual Information	0.734	0.643	0.523

Table B.2: Averaged validation metrics for the naive bayes models fitted using 10 repetitions of 5-fold validation on different training sets. These metrics are sorted by the mean ROC AUC score.

Including amino acid composition	Excluding amino acid composition
composition_HIS	aggrescan3d_max_value
composition_GLN	evoef2_total
composition_GLU	aggrescan3d_min_value
hydrophobic_fitness	aggrescan3d_avg_value
composition_PRO	rosetta_hbond_bb_sc
composition_LYS	rosetta_omega

Continued on next page

Table B.3 – *Continued from previous page*

Including amino acid composition	**Excluding amino acid composition**
composition_VAL	rosetta_fa_atr
composition_ASP	rosetta_rama_prepro
aggrescan3d_max_value	rosetta_p_aa_pp
composition_GLY	rosetta_fa_sol
aggrescan3d_min_value	
composition_PHE	
aggrescan3d_avg_value	
composition_THR	
composition_TYR	
rosetta_hbond_bb_sc	
composition_ILE	
composition_TRP	
composition_LEU	

Table B.3: Mutual information selected features including and excluding amino acid composition metrics. The same features were found for the different scaling methods. The glossary of DE-STRESS metrics is in appendix A.

Appendix C

Large scale analysis of physico-chemical properties

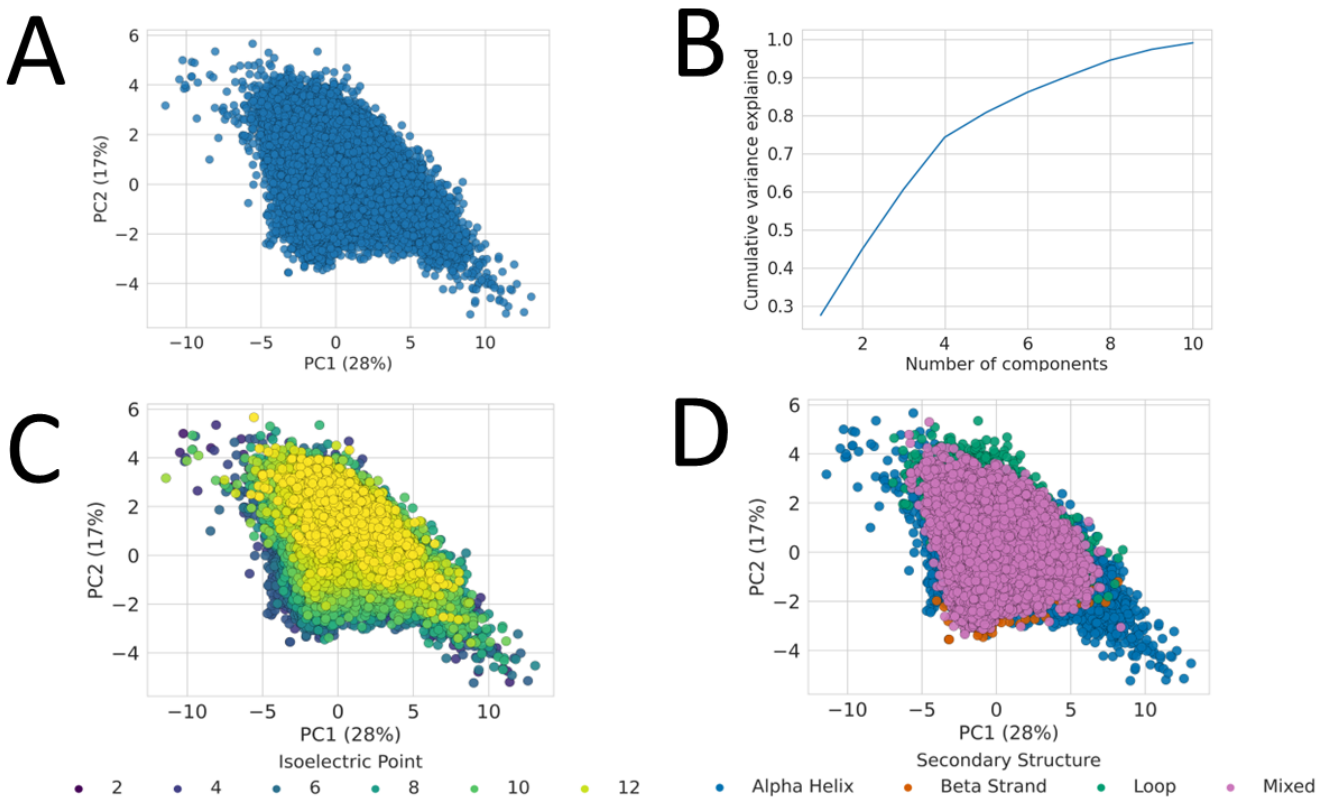

Figure C.1: A), C), D) shows PC1 and PC2 for the physico-chemical properties of the 564,442 AF2 structural models and how different metrics vary across this space, while B) shows the cumulative variance explained by number of components. For this space the robust scaling method was used, and amino acid composition metrics were excluded.

Top contributors to PC1	Top contributors to PC2
aggrescan3d_avg_value	rosetta_fa_dun
aggrescan3d_min_value	rosetta_hbond_sc
rosetta_fa_intra_sol_xover4	aggrescan3d_min_value
aggrescan3d_max_value	rosetta_fa_intra_sol_xover4
budeff_charge	rosetta_hbond_lr_bb

Table C.1: Top 5 contributors to PC1 and PC2 for AF2 structural model PCA space in figure C.1. For this PCA space the robust scaling method was used, and amino acid composition metrics were excluded. The glossary of DE-STRESS metrics is in appendix A.

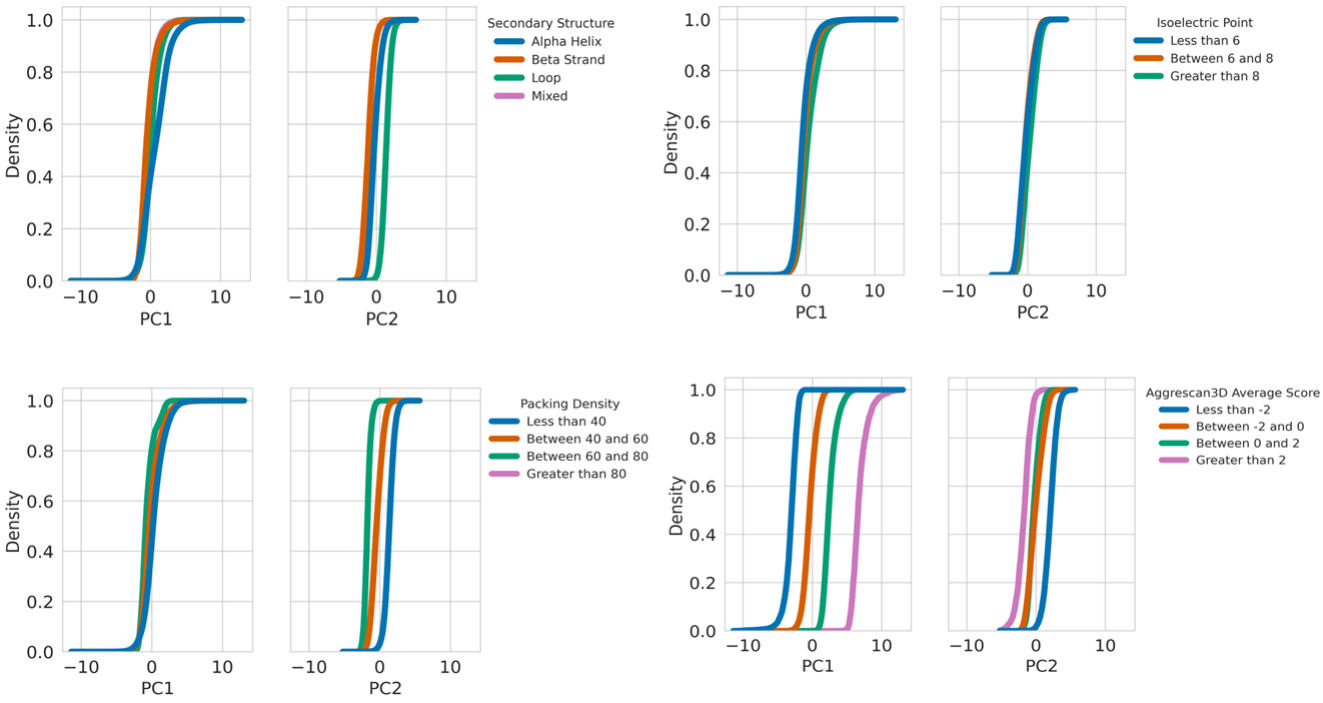

Figure C.2: Cumulative histograms of different metrics, across PC1 and PC2, for the physico-chemical properties of 564,442 AF2 structural models in figure C.1. The robust scaling method was used for this PCA space.

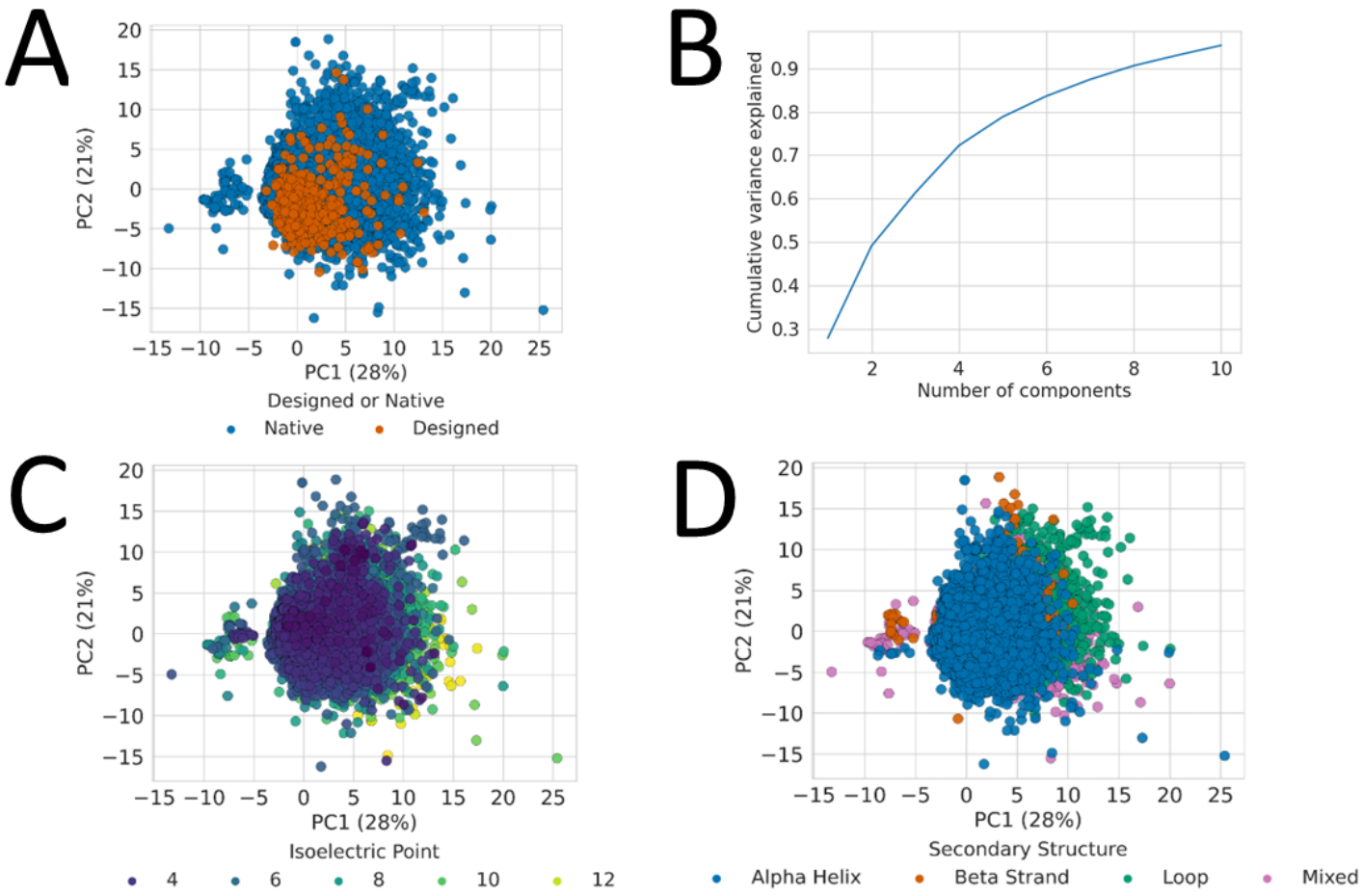

Figure C.3: A), C), D) shows PC1 and PC2 for the physico-chemical properties of the PDB and how different metrics vary across this space, while B) shows the cumulative variance explained by number of components. For this space the robust scaling method was used, and amino acid composition metrics were excluded.

Top contributors to PC1	**Top contributors to PC2**
rosetta_fa_dun	aggrescan3d_avg_value
packing_density	rosetta_fa_intra_sol_xover4
evoef2_intraR_total	aggrescan3d_min_value
rosetta_hbond_bb_sc	aggrescan3d_max_value
rosetta_hbond_sc	rosetta_fa_sol

Table C.2: Top 5 contributors to PC1 and PC2 for PDB PCA space in figure C.3. For this PCA space the robust scaling method was used, and amino acid composition metrics were excluded. The glossary of DE-STRESS metrics is in appendix A.

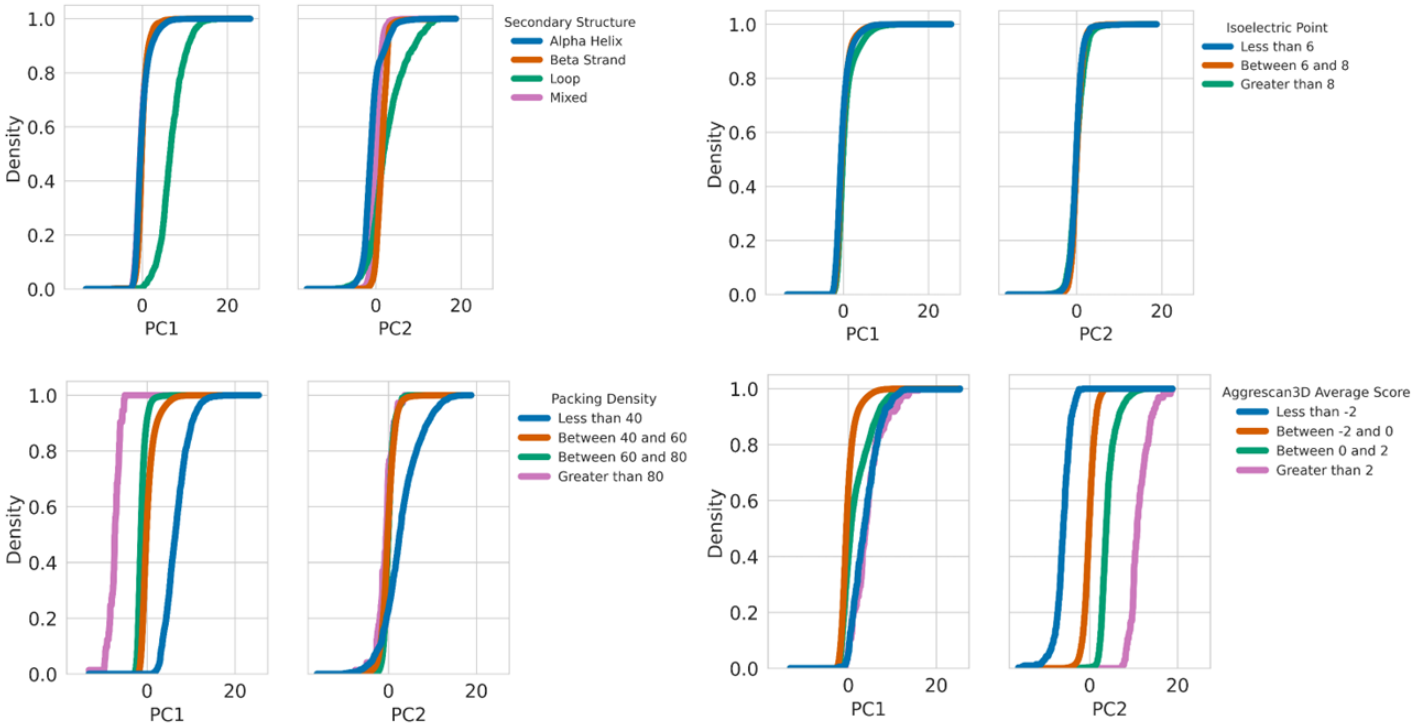

Figure C.4: Cumulative histograms of different metrics, across PC1 and PC2, for the physico-chemical properties of PDB structures in figure C.3. The robust scaling method was used for this PCA space.

Appendix D

Protein properties distinguish eukaryotic and prokaryotic organisms

Top contributors to PC1	**Top contributors to PC2**
rosetta_fa_dun	rosetta_fa_intra_sol_xover4
aggrescan3d_min_value	rosetta_fa_elec
rosetta_fa_elec	rosetta_lk_ball_wtd
aggrescan3d_max_value	rosetta_hbond_sr_bb
budeff_charge	rosetta_fa_dun

Table D.1: Top 5 contributors to PC1 and PC2 of the average model-derived properties for each organism using the robust scaling method. (DE-STRESS glossary in appendix A.)

Top contributors to PC1	**Top contributors to PC2**
rosetta_fa_elec	rosetta_fa_elec
rosetta_fa_dun	isoelectric_point
aggrescan3d_max_value	rosetta_fa_intra_sol_xover4
rosetta_hbond_sc	rosetta_hbond_sr_bb
rosetta_hbond_bb_sc	rosetta_lk_ball_wtd

Table D.2: Top 5 contributors to PC1 and PC2 of the average properties for each organism using the minmax scaling method. (DE-STRESS glossary in appendix A.)

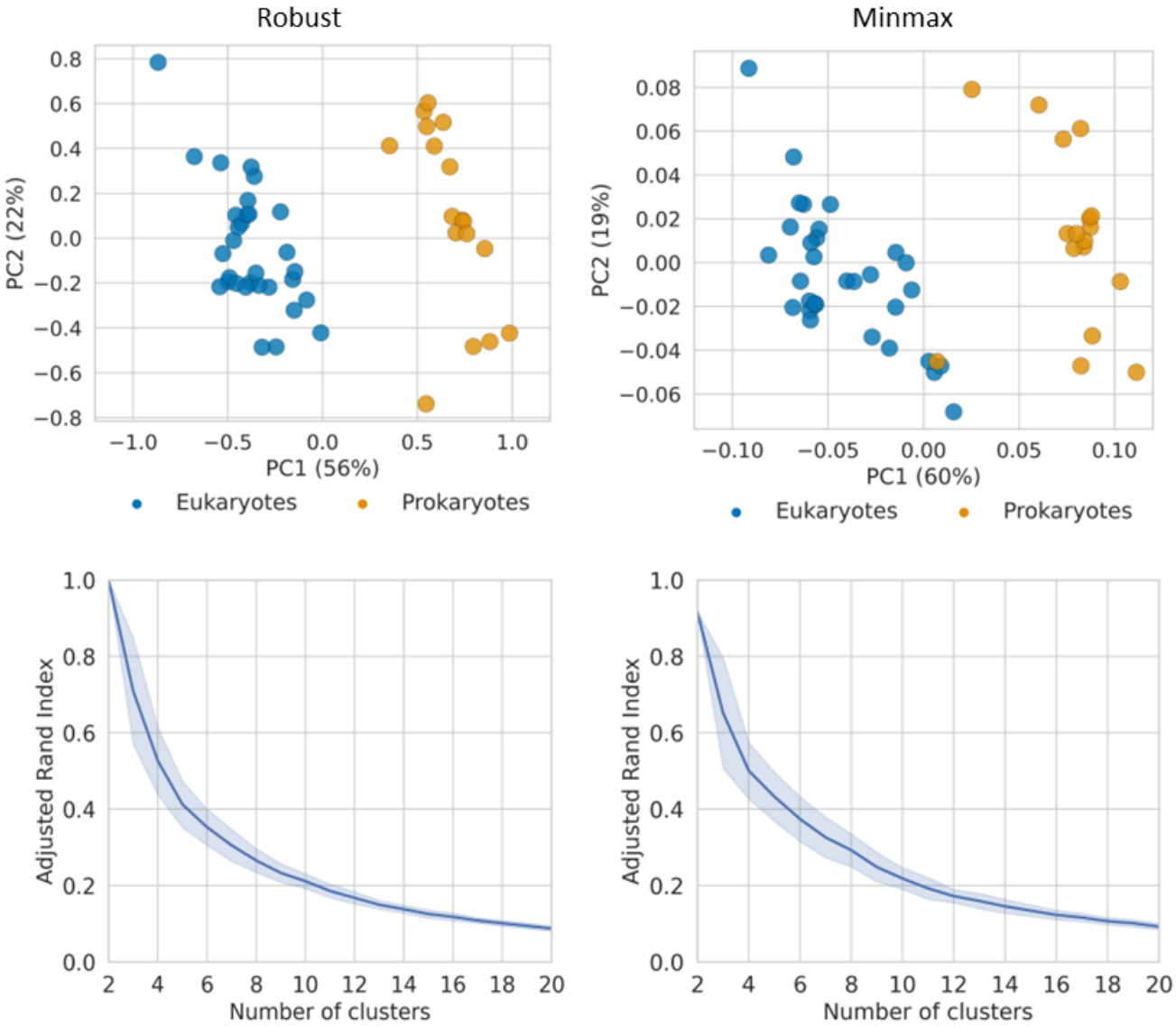

Figure D.1: The average properties for each organism across PC1 and PC2, along with the variance explained for different scaling methods. Also, the mean and standard deviation of the adjusted rand index against the eukaryote and prokaryote groups, for 100 random initialisations of K-means, and different numbers of clusters.

Appendix E

Reconstructing the tree of life

Organism name	**Volume of structures**	**Volume of non-redundant structures**	**Volume of high-quality models (Avg pLDDT $\geq$ 70%)**
Ajellomyces capsulatus	9,199	8,666	4,575
Arabidopsis thaliana	27,427	14,455	8,503
Brugia malayi	8,731	6,876	4,311
Caenorhabditis elegans	19,671	13,948	8,655
Campylobacter jejuni	1,620	1,536	1,499
Candida albicans	5,974	5,325	3,770
Cladophialophora carrionii	11,170	9,825	6,192
Danio rerio	24,656	13,802	9,382
Dictyostelium discoideum	12,622	10,129	5,396
Dracunculus medinensis	10,834	9,864	6,391
Drosophila melanogaster	13,455	10,853	6,520
Enterococcus faecium	2,627	2,264	2,152
Escherichia coli	4,363	3,657	3,515
Fonsecaea pedrosoi	12,509	9,768	6,589
Glycine max	55,798	23,349	12,422
Haemophilus influenzae	1,661	1,583	1,545
Helicobacter pylori	1,538	1,460	1,352

Continued on next page

Table E.1 – *Continued from previous page*

Organism name	**Volume of structures**	**Volume of non-redundant structures**	**Volume of high-quality models (Avg pLDDT $\geq$ 70%)**
Homo sapiens	23,327	13,755	8,489
Klebsiella pneumoniae	5,727	4,782	4,409
Leishmania infantum	7,914	7,336	3,941
Madurella mycetomatis	9,561	8,325	5,809
Methanocaldococcus jannaschii	1,773	1,633	1,545
Mus musculus	21,562	12,577	8,090
Mycobacterium leprae	1,602	1,520	1,344
Mycobacterium tuberculosis	3,988	3,336	3,043
Mycobacterium ulcerans	9,033	7,876	6,000
Neisseria gonorrhoeae	2,106	1,967	1,713
Nocardia brasiliensis	8,372	6,087	5,685
Onchocerca volvulus	1,459	1,391	926
Oryza sativa	41,894	30,164	10,980
Paracoccidioides lutzii	8,794	8,358	4,532
Plasmodium falciparum	5,115	4,597	2,053
Pseudomonas aeruginosa	5,556	4,366	4,207
Rattus norvegicus	19,258	12,220	7,874
Saccharomyces cerevisiae	6,038	5,116	3,584
Salmonella typhimurium	4,526	3,834	3,660
Schistosoma mansoni	13,856	10,505	5,401
Schizosaccharomyces pombe	5,112	4,539	3,429
Shigella dysenteriae	3,893	3,189	3,003
Sporothrix schenckii	8,652	7,691	4,801
Staphylococcus aureus	2,888	2,570	2,382

Continued on next page

Table E.1 – *Continued from previous page*

Organism name	**Volume of structures**	**Volume of non-redundant structures**	**Volume of high-quality models (Avg pLDDT $\geq$ 70%)**
Streptococcus pneumoniae	2,030	1,834	1,751
Strongyloides stercoralis	12,609	10,233	6,243
Trichuris trichiura	9,563	8,581	5,662
Trypanosoma brucei	8,491	7,592	4,354
Trypanosoma cruzi	19,036	10,280	6,004
Wuchereria bancrofti	12,721	12,315	7,245
Zea mays	38,914	21,881	10,206
Total	549,225	387,810	241,134

Table E.1: Table of the 48 different organisms, along with initial volume of protein structures downloaded from AlphaFold DB and then the volumes after using MMSeq2 to remove redundant proteins, and after removing low quality AF2 structural models.

Appendix F

Table of chemicals and materials

Name	Supplier	Use
Phusion High-Fidelity (HF) Mastermix	Thermo Fisher	PCR
Phusion High-Fidelity (HF) Polymerase	NEB	PCR
Phusion High-Fidelity (HF) Buffer	NEB	PCR
Deoxyribonucleotide triphosphate (dNTPs)	NEB	PCR
Dimethyl sulfoxide (DMSO)	NEB	PCR
UltraPure Agarose	Life Technologies	Gel electrophoresis
Tris-Borate-EDTA (TBE) buffer (10X)	Sigma-Aldrich	Gel electrophoresis
SYBR Safe DNA Gel Stain	Thermo Fisher	Gel electrophoresis
HyperLadder 1 Kb DNA Ladder	Thermo Fisher	Gel electrophoresis
GeneRuler 1 kb DNA Ladder	Thermo Fisher	Gel electrophoresis
DNA Gel Loading Dye (6X)	Thermo Fisher	Gel electrophoresis
Agar	Formedium	Making plates
2x Yeast Extract Tryptone (YT) medium	Sigma-Aldrich	Bacterial cell growth

Name	Supplier	Use
Potassium phosphate dibasic	Sigma-Aldrich	Bacterial cell growth
Potassium phosphate dibasic	Sigma-Aldrich	Bacterial cell growth
Ampicillin	Sigma-Aldrich	Bacterial cell growth
Kanamycin	Sigma-Aldrich	Bacterial cell growth
Chloramphenicol	Sigma-Aldrich	Bacterial cell growth
Glycerol	Sigma-Aldrich	Storing cells
DH5 alpha competent cells	NEB	Transformations
Dpn1 digestion kit (enzyme and buffer)	NEB	Gibson assembly
Gibson assembly master mix	NEB	Gibson assembly
HEPES	Sigma-Aldrich	Energy solution component
tRNA	Sigma-Aldrich/Roche	Energy solution component
Adenosine 5'-triphosphate dipotassium salt hydrate (ATP)	Sigma-Aldrich	Energy solution component
Guanosine 5'-triphosphate sodium salt hydrate (GTP)	Sigma-Aldrich	Energy solution component
Cytidine 5'-triphosphate disodium salt (CTP)	Sigma-Aldrich	Energy solution component
Uridine 5'-triphosphate trisodium salt dihydrate (UTP)	Sigma-Aldrich	Energy solution component
Coenzyme A (CoA)	Sigma-Aldrich	Energy solution component
Nicotinamide adenine dinucleotide (NAD)	Sigma-Aldrich	Energy solution component
Adenosine 3',5'-cyclic monophosphate sodium salt monohydrate (cAMP)	Sigma-Aldrich	Energy solution component
Folinic acid	Sigma-Aldrich	Energy solution component
Spermidine	Sigma-Aldrich	Energy solution component

Name	Supplier	Use
3PGA	Sigma-Aldrich	Energy solution component
Poly(ethylene glycol) (PEG 8k)	Sigma-Aldrich	Energy solution component / Cell-free reaction component
L-Glutamic acid hemimagnesium salt tetrahydrate (Mg glutamate)	Sigma-Aldrich	Energy solution component / Cell-free reaction component
L-Glutamic acid monopotassium salt monohydrate (K glutamate)	Sigma-Aldrich	Energy solution component / Cell-free reaction component
Dithiothreitol (DTT)	Sigma-Aldrich	Energy solution component / Cell-free reaction component
Sigma L- Amino acids	Scientific Laboratory Supplies	Energy solution component / Cell-free reaction component
Maltose monohydrate	EMD Millipore Corp.	Cell-free reaction component
384 Well Black/Clear Bottom Plate, Non-Treated Surface, No Lid, Non-Sterile (Nunc 384 well plate)	Scientific Laboratory Supplies	Cell-free reactions
Nunc 384 well plate plastic seal	Thermo-Fisher	Cell-free reactions
Chill-out Liquid Wax	Bio-Rad	Cell-free reactions
Slide-A-Lyze Dialysis Cassettes, 10K MWCO, 3mL	Life Technologies	Dialysis step of crude lysis preparation

Table F.1: List of chemicals and materials used for the experimental work.

Appendix G

Table of molecular biology kits used

Name	Supplier	Use
Zymo DNA Clean and Concentrator -25 kit	Zymo Research	PCR clean up
GeneJET plasmid miniprep kit	Thermo-Fisher	Minipreps
Dpn1 digestion kit	NEB	Digestion of template plasmids after Gibson assembly
Bradford Assay (Quick start Bradford 1x Dye Reagent)	Biorad	Measuring protein concentration

Table G.1: Table of molecular biology kits used in the experimental work.

Appendix H

Tables of sequences

Name	Origin	DNA Sequence (5' to 3')
p70a-deGFP	IDT gblock of promoter, CDS and terminator regions from Addgene (92224)	CCAGCCAGAAAACGACCTTTCTGTGGTGAAACCGGATGCTGCAATTCAGAGC GGCAGCAAGTGGGGGACAGCAGAAGACCTGACCGCCGCAGAGTGGATGTTT GACATGGTGAAGACTATCGCACCATCAGCCAGAAAACCGAATTTTGCTGGGT GGGCTAACGATATCCGCCTGATGCGTGAACGTGACGGACGTAACCACCGCGA CATGTGTGTGCTGTTCCGCTGGGCATGCTGAGCTAACACCGTGCGTGTTGACA ATTTTACCTCTGGCGGTGATAATGGTTGCAGCTAGCAATAATTTTGTTTAACT TTAAGAAGGAGATATACCATGGAGCTTTTCACTGGCGTTGTTCCCATCCTGGT CGAGCTGGACGGCGACGTAAACGGCCACAAGTTCAGCGTGTCCGGCGAGGGC GAGGGCGATGCCACCTACGGCAAGCTGACCCTGAAGTTCATCTGCACCACCG GCAAGCTGCCCGTGCCCTGGCCCACCCTCGTGACCACCCTGACCTACGGCGTG CAGTGCTTCAGCCGCTACCCCGACCACATGAAGCAGCACGACTTCTTCAAGTC CGCCATGCCCGAAGGCTACGTCCAGGAGCGCACCATCTTCTTCAAGGACGAC GGCAACTACAAGACCCGCGCCGAGGTGAAGTTCGAGGGCGACACCCTGGTGA ACCGCATCGAGCTGAAGGGCATCGACTTCAAGGAGGACGGCAACATCCTGGG GCACAAGCTGGAGTACAACTACAACAGCCACAACGTCTATATCATGGCCGAC AAGCAGAAGAACGGCATCAAGGTGAACTTCAAGATCCGCCACAACATCGAG GACGGCAGCGTGCAGCTCGCCGACCACTACCAGCAGAACACCCCCATCGGCG ACGGCCCCGTGCTGCTGCCCGACAACCACTACCTGAGCACCCAGTCCGCCCT GAGCAAAGACCCCAACGAGAAGCGCGATCACATGGTCCTGCTGGAGTTCGTG ACCGCCGCCGGGATCTAAACAATAACTGAATAGGGGATCCCGACTGGCGAGA GCCAGGTAACGAATGGATCCCCGAGCTCGAGCAAAGCCCGCCGAAAGGCGG GCTTTTCTGTGTCGACCGATGCCCTTGAGAGCCTTCAACCCAGTCAGCTCCTT CCGGTGGGCGCGGGGCATGACTATCGTCGCCGCACTTATGACTGTCTTCTTTA TCATGCAACTCGTAGGACAGGTGCCGGCAGCGCTCTTCCGCTTCCTCGCTCAC TGACTCGCTGCGCTCGGTCGTTCGGCTGCGGCGAGCGGTATCAGCTCACTCAA AGGCGGTAATACGGTTATCCACAGAATCAGGGGATAACGCAGGAAAGAACA TGTGAGCAAAAGG

Name	Origin	DNA Sequence (5' to 3')
T7p14-deGFP	Arbor Biosciences, #502111	TGCTGCAAGGCGATTAAGTTGGGTAACGCCAGGGTTTTCCCAGTCACGACGT TGTAAAACGACGGCCAGTGCCAAGCTTGCATGCAAGGAGATGGCGCCCAACA GTCCCCCGGCCACGGGGCCTGCCACCATACCCACGCCGAAACAAGCGCTCAT GAGCCCGAAGTGGCGAGCCCGATCTTCCCCATCGGTGATGTCGGCGATATAG GCGCCAGCAACCGCACCTGTGGCGCCGGTGATGCCGGCCACGATGCGTCCGG CGTAGAGGATCGAGATCTCGATCCCGCGAAATTAATACGACTCACTATAGGG AGACCACAACGGTTTCCCTCTAGAAATAATTTTGTTTAACTTTAAGAAGGAGA TATACCATGGAGCTTTTCACTGGCGTTGTTCCCATCCTGGTCGAGCTGGACGG CGACGTAAACGGCCACAAGTTCAGCGTGTCCGGCGAGGGCGAGGGCGATGCC ACCTACGGCAAGCTGACCCTGAAGTTCATCTGCACCACCGGCAAGCTGCCCG TGCCCTGGCCCACCCTCGTGACCACCCTGACCTACGGCGTGCAGTGCTTCAGC CGCTACCCCGACCACATGAAGCAGCACGACTTCTTCAAGTCCGCCATGCCCG AAGGCTACGTCCAGGAGCGCACCATCTTCTTCAAGGACGACGGCAACTACAA GACCCGCGCCGAGGTGAAGTTCGAGGGCGACACCCTGGTGAACCGCATCGAG CTGAAGGGCATCGACTTCAAGGAGGACGGCAACATCCTGGGGCACAAGCTGG AGTACAACTACAACAGCCACAACGTCTATATCATGGCCGACAAGCAGAAGAA CGGCATCAAGGTGAACTTCAAGATCCGCCACAACATCGAGGACGGCAGCGTG CAGCTCGCCGACCACTACCAGCAGAACACCCCCATCGGCGACGGCCCCGTGC TGCTGCCCGACAACCACTACCTGAGCACCCAGTCCGCCCTGAGCAAAGACCC CAACGAGAAGCGCGATCACATGGTCCTGCTGGAGTTCGTGACCGCCGCCGGG ATCTCTAGAGTGCACCACCACCACCATCACGTGTAAGATCCGGCTGCTAACA AAGCCCGAAAGGAAGCTGAGTTGGCTGCTGCCACCGCTGAGCAATAACTAGC ATAACCCCTTGGGGCCTCTAAACGGGTCTTGAGGGGTTTTTTGCTGAAAGGAG GAACTATATCCGGATATCCACAGGACGGGTGTGGTCGCCATGATCGCGTAGT CGATAGTGGCTCCAAGTAGCGAAGCGAGCAGGACTGGGCGGCGGCCAAAGC GGTCGGACAGTGCTCCGAGAACGGGTGCGCATAGAAATTGCATCAACGCATA TAGCGCTAGCAGCACGCCATAGTGACTGGCGATGCTGTCGGAATGGACGATA TCCCGCAAGAGGCCCGGCAGTACCGGCATAACCAAGCCTATGCCTACA

Name	Origin	DNA Sequence (5' to 3')
p70a-4m5.3-deGFP	IDT gblock of promoter and terminator regions from Addgene (92224). 4m5.3 amino acid sequence comes from Baran et al. [159], and DNA sequence obtained from IDTs codon optimisation tool	CCAGCCAGAAAACGACCTTTCTGTGGTGAAACCGGATGCTGCAATTCAGAGC GGCAGCAAGTGGGGGACAGCAGAAGACCTGACCGCCGCAGAGTGGATGTTT GACATGGTGAAGACTATCGCACCATCAGCCAGAAAACCGAATTTTGCTGGGT GGGCTAACGATATCCGCCTGATGCGTGAACGTGACGGACGTAACCACCGCGA CATGTGTGTGCTGTTCCGCTGGGCATGCTGAGCTAACACCGTGCGTGTTGACA ATTTTACCTCTGGCGGTGATAATGGTTGCAGCTAGCAATAATTTTGTTTAACT TTAAGAAGGAGATATACCATGTCCGAAGTAAAGCTCGATGAAACCGGTGGCG GCCTGGTGCAGCCAGGTGGTGCCATGAAGTTAAGTTGTGTCACTTCTGGGTTT ACATTCGGGCACTACTGGATGAACTGGGTGCGGCAGTCCCCAGAGAAAGGTT TGGAATGGGTCGCACAATTCCGTAATAAACCATATAATTATGAAACGTACTA TAGTGATTCCGTTAAAGGCCGCTTTACGATCTCTCGGGATGACTCTAAGTCGT CGGTTTACCTGCAGATGAATAACCTTCGGGTGGAGGACACGGGTATCTATTAT TGTACAGGGGCTTCTTACGGTATGGAATACTTGGGGCAAGGGACGTCCGTAA CCGTGTCGAGTGGTGGTGGGGGTTCGGGGGGTGGTGGCTCAGGCGGCGGGGG GAGCGATGTTGTAATGACTCAGACACCATTGTCATTGCCAGTTTCGTTAGGTG ACCAAGCAAGTATTAGCTGCCGCTCCAGCCAGAGCCTCGTGCACTCCAACGG TAATACTTATCTGCGTTGGTATCTGCAAAAGCCTGGTCAATCGCCAAAAGTTT TGATTTATAAGGTTTCAAACCGTGTCAGTGGGGTCCCGGATCGTTTTAGCGGG TCAGGCAGTGGCACTGATTTTACGCTTAAAATTAATCGGGTTGAAGCAGAGG ATCTCGGGGTGTACTTCTGTTCACAAAGTACGCACGTTCCGTGGACATTTGGT GGCGGCACAAAGTTGGAAATCAAAGGTGGGGGGGGTAGTCCGGCACCTGCA CCTCCGATGGAGCTTTTCACTGGCGTTGTTCCCATCCTGGTCGAGCTGGACGG CGACGTAAACGGCCACAAGTTCAGCGTGTCCGGCGAGGGCGAGGGCGATGCC ACCTACGGCAAGCTGACCCTGAAGTTCATCTGCACCACCGGCAAGCTGCCCG TGCCCTGGCCCACCCTCGTGACCACCCTGACCTACGGCGTGCAGTGCTTCAGC CGCTACCCCGACCACATGAAGCAGCACGACTTCTTCAAGTCCGCCATGCCCG AAGGCTACGTCCAGGAGCGCACCATCTTCTTCAAGGACGACGGCAACTACAA GACCCGCGCCGAGGTGAAGTTCGAGGGCGACACCCTGGTGAACCGCATCGAG CTGAAGGGCATCGACTTCAAGGAGGACGGCAACATCCTGGGGCACAAGCTGG AGTACAACTACAACAGCCACAACGTCTATATCATGGCCGACAAGCAGAAGAA CGGCATCAAGGTGAACTTCAAGATCCGCCACAACATCGAGGACGGCAGCGTG CAGCTCGCCGACCACTACCAGCAGAACACCCCCATCGGCGACGGCCCCGTGC TGCTGCCCGACAACCACTACCTGAGCACCCAGTCCGCCCTGAGCAAAGACCC CAACGAGAAGCGCGATCACATGGTCCTGCTGGAGTTCGTGACCGCCGCCGGG ATCTAAACAATAACTGAATAGGGGATCCCGACTGGCGAGAGCCAGGTAACGA ATGGATCCCCGAGCTCGAGCAAAGCCCGCCGAAAGGCGGGCTTTTCTGTGTC GACCGATGCCCTTGAGAGCCTTCAACCCAGTCAGCTCCTTCCGGTGGGCGCG GGGCATGACTATCGTCGCCGCACTTATGACTGTCTTCTTTATCATGCAACTCG TAGGACAGGTGCCGGCAGCGCTCTTCCGCTTCCTCGCTCACTGACTCGCTGCG CTCGGTCGTTCGGCTGCGGCGAGCGGTATCAGCTCACTCAAAGGCGGTAATA CGGTTATCCACAGAATCAGGGGATAACGCAGGAAAGAACATGTGAGCAAAA GG

Name	Origin	DNA Sequence (5' to 3')
pET24(+)-4m5.3-deGFP-lac	Clonal gene ordered from Twist Bioscience. 4m5.3 amino acid sequence comes from Baran et al. [159], and DNA sequence obtained from Twists codon optimisation tool. deGFP sequence comes from Addgene (92224).	CAGCCCAGTAGTAGGTTGAGGCCGTTGAGCACCGCCGCCGCAAGGAATGGTG CATGCAAGGAGATGGCGCCCAACAGTCCCCCGGCCACGGGGCCTGCCACCAT ACCCACGCCGAAACAAGCGCTCATGAGCCCGAAGTGGCGAGCCCGATCTTCC CCATCGGTGATGTCGGCGATATAGGCGCCAGCAACCGCACCTGTGGCGCCGG TGATGCCGGCCACGATGCGTCCGGCGTAGAGGATCGAGATCTCGATCCCGCG AAATTAATACGACTCACTATAGGGGAATTGTGAGCGGATAACAATTCCCCTC TAGAAATAATTTTGTTTAACTTTAAGAAGGAGATATACCATATGATGAGCGA GGTCAAATTGGATGAGACTGGCGGCGGCTTGGTACAACCGGGCGGCGCCATG AAGCTGAGCTGCGTTACCTCGGGCTTTACCTTTGGGCATTACTGGATGAATTG GGTCCGCCAGAGCCCGGAAAAGGGGCTGGAATGGGTGGCGCAGTTTCGTAAT AAACCGTACAACTATGAGACGTATTATAGCGACTCAGTGAAAGGGCGTTTCA CCATTTCTCGCGACGATTCGAAAAGCTCGGTGTATCTGCAAATGAATAATCTC CGCGTTGAAGATACGGGCATCTATTATTGTACTGGCGCGTCTTATGGCATGGA ATATCTGGGTCAAGGAACCTCAGTGACCGTGAGCTCGGGTGGTGGTGGCAGC GGCGGTGGTGGCTCGGGTGGTGGCGGAAGTGATGTTGTTATGACCCAGACTC CGCTGTCACTGCCCGTGTCTTTGGGCGATCAGGCATCCATTAGCTGCCGCTCC AGCCAATCGCTGGTCCACAGCAATGGCAATACCTATCTGCGCTGGTATCTGCA GAAACCGGGTCAGTCCCCCAAGGTCCTGATCTATAAGGTTTCGAATCGCGTGT CAGGGGTCCCGGATCGCTTCTCAGGGTCGGGCAGCGGCACCGATTTCACCCTT AAAATTAACCGTGTTGAAGCGGAAGACCTGGGCGTGTACTTCTGTTCGCAGA GCACACACGTTCCATGGACCTTTGGCGGCGGTACTAAGTTAGAGATTAAAGG TGGTGGCGGCTCGCCGGCCCCCGCCCCGCCTATGGAGCTTTTCACTGGCGTTG TTCCCATCCTGGTCGAGCTGGACGGCGACGTAAACGGCCACAAGTTCAGCGT GTCCGGCGAGGGCGAGGGCGATGCCACCTACGGCAAGCTGACCCTGAAGTTC ATCTGCACCACCGGCAAGCTGCCCGTGCCCTGGCCCACCCTCGTGACCACCCT GACCTACGGCGTGCAGTGCTTCAGCCGCTACCCCGACCACATGAAGCAGCAC GACTTCTTCAAGTCCGCCATGCCCGAAGGCTACGTCCAGGAGCGCACCATCTT CTTCAAGGACGACGGCAACTACAAGACCCGCGCCGAGGTGAAGTTCGAGGGC GACACCCTGGTGAACCGCATCGAGCTGAAGGGCATCGACTTCAAGGAGGACG GCAACATCCTGGGGCACAAGCTGGAGTACAACTACAACAGCCACAACGTCTA TATCATGGCCGACAAGCAGAAGAACGGCATCAAGGTGAACTTCAAGATCCGC CACAACATCGAGGACGGCAGCGTGCAGCTCGCCGACCACTACCAGCAGAACA CCCCCATCGGCGACGGCCCCGTGCTGCTGCCCGACAACCACTACCTGAGCAC CCAGTCCGCCCTGAGCAAAGACCCCAACGAGAAGCGCGATCACATGGTCCTG CTGGAGTTCGTGACCGCCGCCGGGATCCTCGAGCACCACCACCACCACCACT GAGATCCGGCTGCTAACAAAGCCCGAAAGGAAGCTGAGTTGGCTGCTGCCAC CGCTGAGCAATAACTAGCATAACCCCTTGGGGCCTCTAAACGGGTCTTGAGG GGTTTTTTGCTGAAAGGAGGAACTATATCCGGATTGGCGAATGGGACGCGCC CTGTAGCGGCGCATTAAGCGCGGCGGGTGTGGTGGTTACGCGCAGCGTGACC GCTACACTTGCCAGCGCCCTAGCGCCCGCTCCTTTCGCTTTCTTCCCTTCCTTT CTCGCCACGTTCGCCGGCTTTCCCCGTCAAGCTCTAAATCGGGGGCTCCCTTT AGGGTTCCGATTTAGTGCTTTACGGCACCTCGACCCCAAAAAACTTGATTAGG GTGATGGTTCACGTAGTGGGCCATCGCCCTGATAGACGGTTT

Name	Origin	DNA Sequence (5' to 3')
pET24(+)-4m5.3-deGFP-wo-lac	Clonal gene ordered from Twist Bioscience. 4m5.3 amino acid sequence comes from Baran et al. [159], and DNA sequence obtained from Twists codon optimisation tool. deGFP sequence comes from Addgene (92224).	CAGCCCAGTAGTAGGTTGAGGCCGTTGAGCACCGCCGCCGCAAGGAATGGTG CATGCAAGGAGATGGCGCCCAACAGTCCCCCGGCCACGGGGCCTGCCACCAT ACCCACGCCGAAACAAGCGCTCATGAGCCCGAAGTGGCGAGCCCGATCTTCC CCATCGGTGATGTCGGCGATATAGGCGCCAGCAACCGCACCTGTGGCGCCGG TGATGCCGGCCACGATGCGTCCGGCGTAGAGGATCGAGATCTCGATCCCGCG AAATTAATACGACTCACTATAGGCCTCTAGAAATAATTTTGTTTAACTTTAAG AAGGAGATATACCATATGATGAGCGAGGTCAAATTGGATGAGACTGGCGGCG GCTTGGTACAACCGGGCGGCGCCATGAAGCTGAGCTGCGTTACCTCGGGCTT TACCTTTGGGCATTACTGGATGAATTGGGTCCGCCAGAGCCCGGAAAAGGGG CTGGAATGGGTGGCGCAGTTTCGTAATAAACCGTACAACTATGAGACGTATT ATAGCGACTCAGTGAAAGGGCGTTTCACCATTTCTCGCGACGATTCGAAAAG CTCGGTGTATCTGCAAATGAATAATCTCCGCGTTGAAGATACGGGCATCTATT ATTGTACTGGCGCGTCTTATGGCATGGAATATCTGGGTCAAGGAACCTCAGTG ACCGTGAGCTCGGGTGGTGGTGGCAGCGGCGGTGGTGGCTCGGGTGGTGGCG GAAGTGATGTTGTTATGACCCAGACTCCGCTGTCACTGCCCGTGTCTTTGGGC GATCAGGCATCCATTAGCTGCCGCTCCAGCCAATCGCTGGTCCACAGCAATG GCAATACCTATCTGCGCTGGTATCTGCAGAAACCGGGTCAGTCCCCCAAGGT CCTGATCTATAAGGTTTCGAATCGCGTGTCAGGGGTCCCGGATCGCTTCTCAG GGTCGGGCAGCGGCACCGATTTCACCCTTAAAATTAACCGTGTTGAAGCGGA AGACCTGGGCGTGTACTTCTGTTCGCAGAGCACACACGTTCCATGGACCTTTG GCGGCGGTACTAAGTTAGAGATTAAAGGTGGTGGCGGCTCGCCGGCCCCCGC CCCGCCTATGGAGCTTTTCACTGGCGTTGTTCCCATCCTGGTCGAGCTGGACG GCGACGTAAACGGCCACAAGTTCAGCGTGTCCGGCGAGGGCGAGGGCGATG CCACCTACGGCAAGCTGACCCTGAAGTTCATCTGCACCACCGGCAAGCTGCC CGTGCCCTGGCCCACCCTCGTGACCACCCTGACCTACGGCGTGCAGTGCTTCA GCCGCTACCCCGACCACATGAAGCAGCACGACTTCTTCAAGTCCGCCATGCC CGAAGGCTACGTCCAGGAGCGCACCATCTTCTTCAAGGACGACGGCAACTAC AAGACCCGCGCCGAGGTGAAGTTCGAGGGCGACACCCTGGTGAACCGCATCG AGCTGAAGGGCATCGACTTCAAGGAGGACGGCAACATCCTGGGGCACAAGCT GGAGTACAACTACAACAGCCACAACGTCTATATCATGGCCGACAAGCAGAAG AACGGCATCAAGGTGAACTTCAAGATCCGCCACAACATCGAGGACGGCAGCG TGCAGCTCGCCGACCACTACCAGCAGAACACCCCCATCGGCGACGGCCCCGT GCTGCTGCCCGACAACCACTACCTGAGCACCCAGTCCGCCCTGAGCAAAGAC CCCAACGAGAAGCGCGATCACATGGTCCTGCTGGAGTTCGTGACCGCCGCCG GGATCCTCGAGCACCACCACCACCACCACTGAGATCCGGCTGCTAACAAAGC CCGAAAGGAAGCTGAGTTGGCTGCTGCCACCGCTGAGCAATAACTAGCATAA CCCCTTGGGGCCTCTAAACGGGTCTTGAGGGGTTTTTTGCTGAAAGGAGGAA CTATATCCGGATTGGCGAATGGGACGCGCCCTGTAGCGGCGCATTAAGCGCG GCGGGTGTGGTGGTTACGCGCAGCGTGACCGCTACACTTGCCAGCGCCCTAG CGCCCGCTCCTTTCGCTTTCTTCCCTTCCTTTCTCGCCACGTTCGCCGGCTTTCC CCGTCAAGCTCTAAATCGGGGGCTCCCTTTAGGGTTCCGATTTAGTGCTTTAC GGCACCTCGACCCCAAAAAACTTGATTAGGGTGATGGTTCACGTAGTGGGCC ATCGCCCTGATAGACGGTTT

Name	Origin	DNA Sequence (5' to 3')
T7p14-4m5.3-deGFP	Cloning 4m5.3 DNA sequence from pET24(+)-4m5.3-deGFP-wo-lac into T7p14-deGFP	TGCTGCAAGGCGATTAAGTTGGGTAACGCCAGGGTTTTCCCAGTCACGACGT TGTAAAACGACGGCCAGTGCCAAGCTTGCATGCAAGGAGATGGCGCCCAACA GTCCCCCGGCCACGGGGCCTGCCACCATACCCACGCCGAAACAAGCGCTCAT GAGCCCGAAGTGGCGAGCCCGATCTTCCCCATCGGTGATGTCGGCGATATAG GCGCCAGCAACCGCACCTGTGGCGCCGGTGATGCCGGCCACGATGCGTCCGG CGTAGAGGATCGAGATCTCGATCCCGCGAAATTAATACGACTCACTATAGGG AGACCACAACGGTTTCCCTCTAGAAATAATTTTGTTTAACTTTAAGAAGGAGA TATACCATGAGCGAGGTCAAATTGGATGAGACTGGCGGCGGCTTGGTACAAC CGGGCGGCGCCATGAAGCTGAGCTGCGTTACCTCGGGCTTTACCTTTGGGCAT TACTGGATGAATTGGGTCCGCCAGAGCCCGGAAAAGGGGCTGGAATGGGTGG CGCAGTTTCGTAATAAACCGTACAACTATGAGACGTATTATAGCGACTCAGT GAAAGGGCGTTTCACCATTTCTCGCGACGATTCGAAAAGCTCGGTGTATCTGC AAATGAATAATCTCCGCGTTGAAGATACGGGCATCTATTATTGTACTGGCGCG TCTTATGGCATGGAATATCTGGGTCAAGGAACCTCAGTGACCGTGAGCTCGG GTGGTGGTGGCAGCGGCGGTGGTGGCTCGGGTGGTGGCGGAAGTGATGTTGT TATGACCCAGACTCCGCTGTCACTGCCCGTGTCTTTGGGCGATCAGGCATCCA TTAGCTGCCGCTCCAGCCAATCGCTGGTCCACAGCAATGGCAATACCTATCTG CGCTGGTATCTGCAGAAACCGGGTCAGTCCCCCAAGGTCCTGATCTATAAGG TTTCGAATCGCGTGTCAGGGGTCCCGGATCGCTTCTCAGGGTCGGGCAGCGG CACCGATTTCACCCTTAAAATTAACCGTGTTGAAGCGGAAGACCTGGGCGTG TACTTCTGTTCGCAGAGCACACACGTTCCATGGACCTTTGGCGGCGGTACTAA GTTAGAGATTAAAGGTGGTGGCGGCTCGCCGGCCCCCGCCCCCGCCTATGGAG CTTTTCACTGGCGTTGTTCCCATCCTGGTCGAGCTGGACGGCGACGTAAACGG CCACAAGTTCAGCGTGTCCGGCGAGGGCGAGGGCGATGCCACCTACGGCAAG CTGACCCTGAAGTTCATCTGCACCACCGGCAAGCTGCCCGTGCCCTGGCCCAC CCTCGTGACCACCCTGACCTACGGCGTGCAGTGCTTCAGCCGCTACCCCGACC ACATGAAGCAGCACGACTTCTTCAAGTCCGCCATGCCCGAAGGCTACGTCCA GGAGCGCACCATCTTCTTCAAGGACGACGGCAACTACAAGACCCGCGCCGAG GTGAAGTTCGAGGGCGACACCCTGGTGAACCGCATCGAGCTGAAGGGCATCG ACTTCAAGGAGGACGGCAACATCCTGGGGCACAAGCTGGAGTACAACTACAA CAGCCACAACGTCTATATCATGGCCGACAAGCAGAAGAACGGCATCAAGGTG AACTTCAAGATCCGCCACAACATCGAGGACGGCAGCGTGCAGCTCGCCGACC ACTACCAGCAGAACACCCCCATCGGCGACGGCCCCGTGCTGCTGCCCGACAA CCACTACCTGAGCACCCAGTCCGCCCTGAGCAAAGACCCCAACGAGAAGCGC GATCACATGGTCCTGCTGGAGTTCGTGACCGCCGCCGGGATCCTCGAGCACC ACCACCACCACCACTGAGATCCGGCTGCTAACAAAGCCCGAAAGGAAGCTGA GTTGGCTGCTGCCACCGCTGAGCAATAACTAGCATAACCCCTTGGGGCCTCTA AACGGGTCTTGAGGGGTTTTTTGCTGAAAGGAGGAACTATATCCGGATATCC ACAGGACGGGTGTGGTCGCCATGATCGCGTAGTCGATAGTGGCTCCAAGTAG CGAAGCGAGCAGGACTGGGCGGCGGCCAAAGCGGTCGGACAGTGCTCCGAG AACGGGTGCGCATAGAAATTGCATCAACGCATATAGCGCTAGCAGCACGCCA TAGTGACTGGCGATGCTGTCGGAATGGACGATATCCCGCAAGAGGCCCGGCA GTACCGGCATAACCAAGCCTATGCCTACA

Name	Origin	DNA Sequence (5' to 3')
T7p14-4m5.3-mCherry		TGCTGCAAGGCGATTAAGTTGGGTAACGCCAGGGTTTTCCCAGTCACGACGT TGTAAAACGACGGCCAGTGCCAAGCTTGCATGCAAGGAGATGGCGCCCAACA GTCCCCCGGCCACGGGGCCTGCCACCATACCCACGCCGAAACAAGCGCTCAT GAGCCCGAAGTGGCGAGCCCGATCTTCCCCATCGGTGATGTCGGCGATATAG GCGCCAGCAACCGCACCTGTGGCGCCGGTGATGCCGGCCACGATGCGTCCGG CGTAGAGGATCGAGATCTCGATCCCGCGAAATTAATACGACTCACTATAGGG AGACCACAACGGTTTCCCTCTAGAAATAATTTTGTTTAACTTTAAGAAGGAGA TATACCATGAGCGAGGTCAAATTGGATGAGACTGGCGGCGGCTTGGTACAAC CGGGCGGCGCCATGAAGCTGAGCTGCGTTACCTCGGGCTTTACCTTTGGGCA TTACTGGATGAATTGGGTCCGCCAGAGCCCGGAAAAGGGGCTGGAATGGGTG GCGCAGTTTCGTAATAAACCGTACAACTATGAGACGTATTATAGCGACTCAG TGAAAGGGCGTTTCACCATTTCTCGCGACGATTCGAAAAGCTCGGTGTATCTG CAAATGAATAATCTCCGCGTTGAAGATACGGGCATCTATTATTGTACTGGCGC GTCTTATGGCATGGAATATCTGGGTCAAGGAACCTCAGTGACCGTGAGCTCG GGTGGTGGTGGCAGCGGCGGTGGTGGCTCGGGTGGTGGCGGAAGTGATGTTG TTATGACCCAGACTCCGCTGTCACTGCCCGTGTCTTTGGGCGATCAGGCATCC ATTAGCTGCCGCTCCAGCCAATCGCTGGTCCACAGCAATGGCAATACCTATCT GCGCTGGTATCTGCAGAAACCGGGTCAGTCCCCCAAGGTCCTGATCTATAAG GTTTCGAATCGCGTGTCAGGGGTCCCGGATCGCTTCTCAGGGTCGGGCAGCG GCACCGATTTCACCCTTAAAATTAACCGTGTTGAAGCGGAAGACCTGGGCGT GTACTTCTGTTCGCAGAGCACACACGTTCCATGGACCTTTGGCGGCGGTACTA AGTTAGAGATTAAAGGTGGTGGCGGCTCGCCGGCCCCCGCCCCGCCTATGAG CAAGGGCGAAGAAGATAACATGGCCATCATCAAGGAGTTCATGCGCTTCAAG GTGCACATGGAGGGCTCCGTGAACGGCCACGAGTTCGAGATCGAGGGCGAG GGCGAGGGCCGCCCCTACGAGGGCACCCAGACCGCCAAGCTGAAGGTGACC AAGGGTGGCCCCCTGCCCTTCGCCTGGGACATCCTGTCCCCTCAGTTCATGTA CGGCTCCAAGGCCTACGTGAAGCACCCCGCCGACATCCCCGACTACTTGAAG CTGTCCTTCCCCGAGGGCTTCAAGTGGGAGCGCGTGATGAACTTCGAGGACG GCGGCGTGGTGACCGTGACCCAGGACTCCTCCCTGCAGGACGGCGAGTTCAT CTACAAGGTGAAGCTGCGCGGCACCAACTTCCCCTCCGACGGCCCCGTAATG CAGAAGAAGACCATGGGCTGGGAGGCCTCCTCCGAGCGGATGTACCCCGAGG ACGGCGCCCTGAAGGGCGAGATCAAGCAGAGGCTGAAGCTGAAGGACGGCG GCCACTACGACGCTGAGGTCAAGACCACCTACAAGGCCAAGAAGCCCGTGCA GCTGCCCGGCGCCTACAACGTCAACATCAAGTTGGACATCACCTCCCACAAC GAGGACTACACCATCGTGGAACAGTACGAACGCGCCGAGGGCCGCCACTCCA CCGGCGGCATGGACGAGCTGTACAAGCTCGAGGTGCACCACCACCACCATCA CGTGTAAGATCCGGCTGCTAACAAAGCCCGAAAGGAAGCTGAGTTGGCTGCT GCCACCGCTGAGCAATAACTAGCATAACCCCTTGGGGCCTCTAAACGGGTCT TGAGGGGTTTTTTGCTGAAAGGAGGAACTATATCCGGATATCCACAGGACGG GTGTGGTCGCCATGATCGCGTAGTCGATAGTGGCTCCAAGTAGCGAAGCGAG CAGGACTGGGCGGCGGCCAAAGCGGTCGGACAGTGCTCCGAGAACGGGTGC GCATAGAAATTGCATCAACGCATATAGCGCTAGCAGCACGCCATAGTGACTG GCGATGCTGTCGGAATGGACGATATCCCGCAAGAGGCCCGGCAGTACCGGCA TAACCAAGCCTATGCCTACA

Table H.1: Linear DNA sequences for all the constructs used in this project, including promoter regions, coding domain sequence and terminator regions.

Name	DNA Sequence (5' to 3')
p70 promoter	TAACACCGTGCGTGTTGACAATTTTACCTCTGGCGGTGATAATGGTTGC
T7 promoter	TAATACGACTCACTATAG
t500 Terminator	CAAAGCCCGCCGAAAGGCGGGCTTTTCTGT
T7 terminator	AACCCCTTGGGGCCTCTAAACGGGTCTTGAGGGGTTTTTT
lac operator	GGGAATTGTGAGCGGATAACAATTCCCC

Table H.2: DNA sequences for the promoter, terminator and lac operator regions used in this project

Name	Amino Acid Sequence
4m5.3 scFv	MSEVKLDETGGGLVQPGGAMKLSCVTSGFTFGHYWMNWVRQSPEKGLEWVAQFRNKPYNYETYYSDSVKGRFTISRDDSKSSVYLQMNNLRVEDTGIYYCTGASYGMEYLGQGTSVTVSSGGGGSGGGGSGGGGSDVVMTQTPLSLPVSLGDQASISCRSSQSLVHSNGNTYLRWYLQKPGQSPKVLIYKVSNRVSGVPDRFSGSGSGTDFTLKINRVEAEDLGVYFCSQSTHVPWTFGGGTKLEIK
deGFP	MELFTGVVPILVELDGDVNGHKFSVSGEGEGDATYGKLTLKFICTTGKLPVPWPTLVTTLTYGVQCFSRYPDHMKQHDFFKSAMPEGYVQERTIFFKDDGNYKTRAEVKFEGDTLVNRIELKGIDFKEDGNILGHKLEYNYNSHNVYIMADKQKNGIKVNFKIRHNIEDGSVQLADHYQQNTPIGDGPVLLPDNHYLSTQSALSKDPNEKRDHMVLLEFVTAAGI
mCherry	MSKGEEDNMAIIKEFMRFKVHMEGSVNGHEFEIEGEGEGRPYEGTQTAKLKVTKGGPLPFAWDILSPQFMYGSKAYVKHPADIPDYLKLSFPEGFKWERVMNFEDGGVVTVTQDSSLQDGEFIYKVKLRGTNFPSDGPVMQKKTMGWEASSERMYPEDGALKGEIKQRLKLKDGGHYDAEVKTTYKAKKPVQLPGAYNVNIKLDITSHNEDYTIVEQYERAEGRHSTGGMDELYK
GSPA linker	GGGGSPAPAPP

Table H.3: Amino acid sequences for 4m5.3 scFv, deGFP. mCherry and the GSPA linker

Name	ID	Sequence (5' to 3')	Ta (°C)	Use
T7p14_deGFP_250_fwd	T7GFPFWD	TGCTGCAAGGCGATTAAGTT	61	PCR
T7p14_deGFP_250_rev	T7GFPREV	TGTAGGCATAGGCTTGGTTA	61	PCR
pBEST_deGFP_fwd_zf10	ZF10	CCAGCCAGAAAACGACCTTTCTGTG	67	PCR

Name	ID	Sequence (5' to 3')	Ta (°C)	Use
pBEST_deGFP_rev_nl32	NL32	CCTTTTGCTCACATGTTCTTTCCTGC	67	PCR
pBEST_deGFP_fwd_nl001	NL001FWD	CTTTCTGTGGTGAAACCGGATGCTGCAATTCAGA	72	PCR
pBEST_deGFP_rev_nl001	NL001REV	ATGTTCTTTCCTGCGTTATCCCCTGATTCTGTGGA	72	PCR
ms_pet24+_4m53_fwd1	MS001	CAGCCCAGTAGTAGGTTGAGGC	66	PCR
ms_pet24+_4m53_rev1	MS002	AAACCGTCTATCAGGGCGATGG	66	PCR
ms_T7p14_4m53_insert_fwd1	MS003	CTTTAAGAAGGAGATATACCATGAGCGAGGTCAAATTGG	61	Cloning
ms_T7p14_4m53_insert_rev1	MS004	GCTTTGTTAGCAGCCGGATCTCAGTGGTGGTGGTGG	61	Cloning
ms_T7p14_4m53_insert_fwd2	MS005	ATTTTGTTTAACTTTAAGAAGGAGATATACCATGAGCGAGGTCAAATTGG	61	Cloning
ms_T7p14_4m53_insert_rev2	MS006	CAGCTTCCTTTCGGGCTTTGTTAGCAGCCGGATCTCAGTGGTGGTGGTGG	61	Cloning
ms_T7p14_bb_fwd1	MS007	GATCCGGCTGCTAACAAAG	61	Cloning
ms_T7p14_bb_rev1	MS008	GGTATATCTCCTTCTTAAAGTTAAACAAAATTATTT	61/63	Cloning
ms_T7p14_4m53_insert_rev3	MS009	TCGGATCCACCACTTCCACCTTTAATCTCTAACTTAGTACCGCCG	61	Cloning
ms_T7p14_4m53_insert_rev4	MS010	TTATCTTCTTCGCCCTTGCTCATAGGCGGGGCGGGGGCCGGCGAGCCGCC	61	Cloning
ms_T7p14_mcherry_bb_fwd1	MS011	GGTGGAAGTGGTGGATCCG	61	Cloning
ms_T7p14_mcherry_bb_fwd2	MS012	ATGAGCAAGGGCGAAGAAG	63	Cloning
Chi6_Fwd	Chi6_Fwd	TCACTTCACTGCTGGTGGCCACTGCTGGTGGCCACTGCTGGTGGCCACTGCTGGGGCCACTGCTGGTGGCCACTGCTGGTGGCCA		Chi6 forward primer
Chi6_Rev	Chi6_Rev	TGGCCACCAGCAGTGGCCACCAGCAGTGGCCACCAGCAGTGGCCACCAGCAGTGGCCACCAGCAGTGAAGTGA		Chi6 reverse primer

Table H.4: Full list of primer sequences along with their annealing temperatures

Appendix I

Buffer and media preparation

Step	Description
1	Weigh out 31 g of 2x Yeast Extract Tryptone (YT) media and place into 1 L glass bottle.
2	Add 938 mL of milliQ water.
3	Autoclave and allow media to cool to room temperature.
4	Add 40 mL of sterile 1 M Potassium Dibasic Solution.
5	Add 22 mL of sterile 1 M Potassium Monobasic Solution.

Table I.1: Recipe for preparing 1 L of 2x YTP media.

Step	Description
1	Weigh out 10 g of formedium tryptone, 5 g of formedium yeast extract and 10 g of sodium chloride (NaCl) and add these to 1 L of distilled water.
2	Adjust pH to 7.2 with 10N of NaOH (approx 0.2 mL/L).
3	Autoclave at 121 °C for 15 minutes.

Table I.2: Recipe for preparing 1 L of Luria-Bertani (LB) broth. This was prepared by the Media Preparation and Wash-up team in the Roger Land building at the University of Edinburgh's Kings Buildings campus.

Step	Description
1	Weigh out 20 g of bacto tryptone, 5 g of oxoid yeast extract, 0.58 g of sodium chloride (NaCl), 0.186 g of potassium chloride (KCl), 2.47 g of magnesium sulphate (MgSO4), 2.03 g of magnesium chloride (MgCl2) and 3.60 g of glucose. Add all of these to 900 mL of distilled water.
2	Autoclave at 116 °C for 20 minutes.

Table I.3: Recipe for preparing 1 L of Super Optimal broth with Catabolite repression (SOC) media. This was prepared by the Media Preparation and Wash-up team in the Roger Land building at the University of Edinburgh's Kings Buildings campus.

Bibliography

[1] B. Alberts, "Molecular biology of the cell," *Garland science*, 2017.

[2] D. P. Clark, N. J. Pazdernik, and M. R. McGehee, "Chapter 3 - Nucleic Acids and Proteins," in *Molecular Biology (Third Edition)* (D. P. Clark, N. J. Pazdernik, and M. R. McGehee, eds.), pp. 63–94, Academic Cell, third edit ed., 2019.

[3] P. Y. Bruice, "Organic chemistry," *Pearson*, 2017.

[4] S. Wang and E. T. Kool, "Origins of the Large Differences in Stability of DNA and RNA Helixes: C-5 Methyl and 2-Hydroxyl Effects," *Biochemistry*, vol. 34, no. 12, pp. 4125–4132, 1995.

[5] D. Ulveling, C. Francastel, and F. Hubé, "When one is better than two: RNA with dual functions," *Biochimie*, vol. 93, pp. 633–644, 4 2011.

[6] M. McCarty and O. T. Avery, "Studies on the chemical nature of the substance inducing transformation of pneumococcal types: II. Effect of desoxyribonuclease on the biological activity of the transforming substance," *The Journal of experimental medicine*, vol. 83, no. 2, p. 89, 1946.

[7] F. Crick, "Central Dogma of Molecular Biology," *Nature 1970 227:5258*, vol. 227, no. 5258, pp. 561–563, 1970.

[8] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock, "Gene Ontology: tool for the unification of biology," *Nature Genetics 2000 25:1*, vol. 25, pp. 25–29, 5 2000.

[9] J. B. Olmsted and G. G. Borisy, "Microtubules," *Annual review of biochemistry*, vol. 42, no. 1, pp. 507–540, 1973.

[10] C. G. Dos Remedios, D. Chhabra, M. Kekic, I. V. Dedova, M. Tsubakihara, D. A. Berry, and N. J. Nosworthy, "Actin binding proteins: Regulation of cytoskeletal microfilaments," *Physiological Reviews*, vol. 83, no. 2, pp. 433–473, 2003.

[11] B. C. Powell and G. E. Rogers, "The role of keratin proteins and their genes in the growth, structure and properties of hair.," *EXS*, vol. 78, pp. 59–148, 1 1997.

[12] M. D. Shoulders and R. T. Raines, "Collagen Structure and Stability," *Annual Review of Biochemistry*, vol. 78, no. 1, pp. 929–958, 2009.

[13] H. Mauser, W. A. King, J. E. Gready, and T. J. Andrews, "CO2 Fixation by Rubisco: Computational Dissection of the Key Steps of Carboxylation, Hydration, and CC Bond Cleavage," *Journal of the American Chemical Society*, vol. 123, pp. 10821–10829, 11 2001.

[14] R. J. Ellis, "The most abundant protein in the world," *Trends in Biochemical Sciences*, vol. 4, pp. 241–244, 11 1979.

[15] R. R. Burgess, "RNA polymerase," *Annual review of biochemistry*, vol. 40, no. 1, pp. 711–740, 1971.

[16] J. M. Caruthers and D. B. McKay, "Helicase structure and mechanism," *Current Opinion in Structural Biology*, vol. 12, pp. 123–133, 2 2002.

[17] K. Brix and W. Stöcker, *Proteases: structure and function*. Springer, 2013.

[18] M. H. Meisler and C. N. Ting, "The remarkable evolutionary history of the human amylase genes," *Critical Reviews in Oral Biology and Medicine*, vol. 4, no. 3-4, pp. 503–509, 1993.

[19] R. Beckerman and C. Prives, "Transcriptional Regulation by P53," *Cold Spring Harbor Perspectives in Biology*, vol. 2, p. a000935, 8 2010.

[20] P. A. Shah, C. Huang, Q. Li, S. A. Kazi, L. A. Byers, J. Wang, F. M. Johnson, and M. J. Frederick, "NOTCH1 Signaling in Head and Neck Squamous Cell Carcinoma," *Cells 2020, Vol. 9, Page 2677*, vol. 9, p. 2677, 12 2020.

[21] S. Artavanis-Tsakonas, "The molecular biology of the Notch locus and the fine tuning of differentiation in Drosophila," *Trends in Genetics*, vol. 4, no. 4, pp. 95–100, 1988.

[22] L. Miele and B. Osborne, "Arbiter of Differentiation and Death: Notch Signaling Meets Apoptosis," *J. Cell. Physiol*, vol. 181, pp. 393–409, 1999.

[23] K. C. Murphy, "Lambda Gam protein inhibits the helicase and chi-stimulated recombination activities of Escherichia coli RecBCD enzyme," *Journal of Bacteriology*, vol. 173, no. 18, pp. 5808–5821, 1991.

[24] Y. Sakaki, A. E. Karu, S. Linn, and H. Echols, "Purification and properties of the $γ$-protein specified by bacteriophage $λ$: an inhibitor of the host recBC recombination enzyme," *Proceedings of the National Academy of Sciences*, vol. 70, no. 8, pp. 2215–2219, 1973.

[25] J. M. Baldwin, "Structure and function of haemoglobin," *Progress in Biophysics and Molecular Biology*, vol. 29, pp. 225–320, 1976.

[26] B. Thorens and M. Mueckler, "Glucose transporters in the 21st Century," *American Journal of Physiology-Endocrinology and Metabolism*, vol. 298, no. 2, pp. E141–E145, 2010.

[27] C. S. Yost, "Potassium Channels Basic Aspects, Functional Roles, and Medical Significance," *Anesthesiology*, vol. 90, pp. 1186–1203, 4 1999.

[28] S. Maloy, "Amino Acids," *Brenner's Encyclopedia of Genetics: Second Edition*, pp. 108–110, 2 2013.

[29] G. E. Schulz and R. H. Schirmer, "Amino Acids," in *Principles of Protein Structure*, pp. 1–16, New York, NY: Springer New York, 1979.

[30] S. H. Leung, "Amino Acids, Aromatic Compounds, and Carboxylic Acids: How Did They Get Their Common Names?," *In the Classroom 48 Journal of Chemical Education* •, vol. 77, no. 1, 2000.

[31] C. B. Anfinsen, "Principles that govern the folding of protein chains," *Science*, vol. 181, no. 4096, pp. 223–230, 1973.

[32] K. Asano, "Why is start codon selection so precise in eukaryotes?," *Translation*, vol. 2, p. e28387, 1 2014.

[33] T. M. SONNEBORN, "Degeneracy of the Genetic Code: Extent, Nature, and Genetic Implications," in *Evolving Genes and Proteins* (V. Bryson and H. J. Vogel, eds.), pp. 377–397, Academic Press, 1965.

[34] I. M. Walsh, M. A. Bowman, I. F. Soto Santarriaga, A. Rodriguez, and P. L. Clark, "Synonymous codon substitutions perturb cotranslational protein folding in vivo and impair cell fitness," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 117, pp. 3528–3534, 2 2020.

[35] C. B. Anfinsen, E. Haber, M. Sela, and F. H. White Jr, "The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain," *Proceedings of the National Academy of Sciences*, vol. 47, no. 9, pp. 1309–1314, 1961.

[36] P. W. Atkins, J. De Paula, and J. Keeler, "Atkins' physical chemistry," *Oxford university press*, 2023.

[37] R. J. Ellis and S. M. der Vies, "Molecular chaperones," *Annual review of biochemistry*, vol. 60, no. 1, pp. 321–347, 1991.

[38] L. Lins and R. Brasseur, "The hydrophobic effect in protein folding," *The FASEB journal*, vol. 9, no. 7, pp. 535–540, 1995.

[39] J. E. Lennard and I. Jones, "On the determination of molecular fields.—I. From the variation of the viscosity of a gas with temperature," *Proceedings of the Royal Society of London. Series A, containing papers of a mathematical and physical character*, vol. 106, no. 738, pp. 441–462, 1924.

[40] K. A. Dill, "Dominant forces in protein folding," *Biochemistry*, vol. 29, pp. 7133–7155, 8 1990.

[41] C. Nick Pace, J. M. Scholtz, and G. R. Grimsley, "Forces stabilizing proteins," *FEBS Letters*, vol. 588, no. 14, pp. 2177–2184, 2014.

[42] D. Schell, J. Tsai, J. M. Scholtz, and C. N. Pace, "Hydrogen bonding increases packing density in the protein interior," *Proteins: Structure, Function, and Bioinformatics*, vol. 63, pp. 278–282, 5 2006.

[43] R. E. Hubbard and M. K. Haider, "Hydrogen Bonds in Proteins: Role and Strength," *Encyclopedia of Life Sciences*, 2 2010.

[44] S. R. Trevino, K. Gokulan, S. Newsom, R. L. Thurlkill, K. L. Shaw, V. A. Mitkevich, A. A. Makarov, J. C. Sacchettini, J. M. Scholtz, and C. N. Pace, "Asp79 Makes a Large, Unfavorable Contribution to the Stability of RNase Sa," *Journal of Molecular Biology*, vol. 354, pp. 967–978, 12 2005.

[45] I. Bošnjak, V. Bojović, T. S. Šegvić-Bubić, and A. Bielen, "Occurrence of protein disulfide bonds in different domains of life: a comparison of proteins from the Protein Data Bank," *Protein Engineering, Design and Selection*, vol. 27, pp. 65–72, 3 2014.

[46] A. McAuley, J. Jacob, C. G. Kolvenbach, K. Westland, H. J. Lee, S. R. Brych, D. Rehder, G. R. Kleemann, D. N. Brems, and M. Matsumura, "Contributions of a disulfide bond to the structure, stability, and dimerization of human IgG1 antibody CH3 domain," *Protein Science : A Publication of the Protein Society*, vol. 17, p. 95, 1 2008.

[47] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, "The Protein Data Bank," 1 2000.

[48] M. Novotny and G. J. Kleywegt, "A Survey of Left-handed Helices in Protein Structures," *Journal of Molecular Biology*, vol. 347, pp. 231–241, 3 2005.

[49] G. N. t. Ramachandran and V. Sasisekharan, "Conformation of polypeptides and proteins," *Advances in protein chemistry*, vol. 23, pp. 283–437, 1968.

[50] R. R. Crichton, *Biological inorganic chemistry: a new introduction to molecular structure and function*. Elsevier, 2012.

[51] M. N. Fodje and S. Al-Karadaghi, "Occurrence, conformational features and amino acid propensities for the π-helix," *Protein Engineering, Design and Selection*, vol. 15, pp. 353–358, 5 2002.

[52] P. Tompa, "Intrinsically disordered proteins: a 10-year recap," *Trends in Biochemical Sciences*, vol. 37, pp. 509–516, 12 2012.

[53] R. Pancsa and P. Tompa, "Structural disorder in eukaryotes," *PloS one*, vol. 7, no. 4, p. e34687, 2012.

[54] M. Buljan and A. G. Bateman, "The evolution of protein domain families," *Biochemical Society Transactions*, vol. 37, pp. 751–755, 8 2009.

[55] I. Letunic, S. Khedkar, and P. Bork, "SMART: recent updates, new developments and status in 2020," *Nucleic Acids Research*, vol. 49, pp. D458–D460, 1 2021.

[56] N. Bordin, I. Sillitoe, V. Nallapareddy, C. Rauer, S. D. Lam, V. P. Waman, N. Sen, M. Heinzinger, M. Littmann, S. Kim, S. Velankar, M. Steinegger, B. Rost, and C. Orengo, "AlphaFold2 reveals commonalities and novelties in protein structure space for 21 model organisms," *Communications Biology 2023 6:1*, vol. 6, pp. 1–12, 2 2023.

[57] A. Andreeva, D. Howorth, C. Chothia, E. Kulesha, and A. G. Murzin, "SCOP2 prototype: a new approach to protein structure mining," *Nucleic Acids Research*, vol. 42, pp. D310–D314, 1 2014.

[58] A. Andreeva, E. Kulesha, J. Gough, and A. G. Murzin, "The SCOP database in 2020: expanded classification of representative family and superfamily domains of known protein structures," *Nucleic Acids Research*, vol. 48, pp. D376–D382, 1 2020.

[59] J. O. Nealon, L. S. Philomina, and L. J. McGuffin, "Predictive and Experimental Approaches for Elucidating Protein–Protein Interactions and Quaternary Structures," *International Journal of Molecular Sciences 2017, Vol. 18, Page 2623*, vol. 18, p. 2623, 12 2017.

[60] T. U. Consortium, "UniProt: the Universal Protein Knowledgebase in 2023," *Nucleic Acids Research*, vol. 51, pp. D523–D531, 3 2023.

[61] T. Hu, N. Chitnis, D. Monos, and A. Dinh, "Next-generation sequencing technologies: An overview," *Human Immunology*, vol. 82, pp. 801–811, 11 2021.

[62] M. S. Smyth and J. H. Martin, "x Ray crystallography," *Molecular Pathology*, vol. 53, no. 1, p. 8, 2000.

[63] L. E. Kay, "NMR studies of protein structure and dynamics," *Journal of Magnetic Resonance*, vol. 213, pp. 477–491, 12 2011.

[64] E. Callaway, "Revolutionary cryo-EM is taking over structural biology," *Nature*, vol. 578, p. 201, 2 2020.

[65] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A. W.

Senior, K. Kavukcuoglu, P. Kohli, and D. Hassabis, "Highly accurate protein structure prediction with AlphaFold," *Nature 2021*, pp. 1–11, 7 2021.

[66] Z. Lin, H. Akin, R. Rao, B. Hie, Z. Zhu, W. Lu, N. Smetanin, R. Verkuil, O. Kabeli, Y. Shmueli, A. dos Santos Costa, M. Fazel-Zarandi, T. Sercu, S. Candido, and A. Rives, "Evolutionary-scale prediction of atomic-level protein structure with a language model," *Science*, vol. 379, pp. 1123–1130, 3 2023.

[67] M. Varadi, S. Anyango, M. Deshpande, S. Nair, C. Natassia, G. Yordanova, D. Yuan, O. Stroe, G. Wood, A. Laydon, A. Zídek, T. Green, K. Tunyasuvunakool, S. Petersen, J. Jumper, E. Clancy, R. Green, A. Vora, M. Lutfi, M. Figurnov, A. Cowie, N. Hobbs, P. Kohli, G. Kleywegt, E. Birney, D. Hassabis, and S. Velankar, "AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models," *Nucleic Acids Research*, vol. 50, pp. D439–D444, 1 2022.

[68] J. Koehler Leman, P. Szczerbiak, P. D. Renfrew, V. Gligorijevic, D. Berenberg, T. Vatanen, B. C. Taylor, C. Chandler, S. Janssen, A. Pataki, N. Carriero, I. Fisk, R. J. Xavier, R. Knight, R. Bonneau, and T. Kosciolek, "Sequence-structure-function relationships in the microbial protein universe," *Nature Communications*, vol. 14, p. 2351, 9 2023.

[69] "Method of the Year 2021: Protein structure prediction," *Nature Methods 2022 19:1*, vol. 19, pp. 1–1, 1 2022.

[70] L. M. Bertoline, A. N. Lima, J. E. Krieger, and S. K. Teixeira, "Before and after AlphaFold2: An overview of protein structure prediction," *Frontiers in Bioinformatics*, vol. 3, no. February, pp. 1–8, 2023.

[71] A. Elofsson, "Progress at protein structure prediction, as seen in CASP15," *Current Opinion in Structural Biology*, vol. 80, p. 102594, 6 2023.

[72] R. Bonneau, J. Tsai, I. Ruczinski, D. Chivian, C. Rohl, C. E. Strauss, and D. Baker, "Rosetta in CASP4: Progress in ab initio protein structure prediction," *Proteins: Structure, Function, and Bioinformatics*, vol. 45, pp. 119–126, 1 2001.

[73] D. T. Jones, "Predicting novel protein folds by using FRAGFOLD," *Proteins: Structure, Function, and Bioinformatics*, vol. 45, pp. 127–132, 1 2001.

[74] J. Lee, P. L. Freddolino, and Y. Zhang, "Ab initio protein structure prediction," *From Protein Structure to Function with Bioinformatics: Second Edition*, pp. 3–35, 4 2017.

[75] J. Yang, W. Zhang, B. He, S. E. Walker, H. Zhang, B. Govindarajoo, J. Virtanen, Z. Xue, H. B. Shen, and Y. Zhang, "Template-based protein structure prediction in CASP11 and retrospect of I-TASSER in the last decade," *Proteins: Structure, Function, and Bioinformatics*, vol. 84, pp. 233–246, 9 2016.

[76] A. Šali and T. L. Blundell, "Comparative Protein Modelling by Satisfaction of Spatial Restraints," *Journal of Molecular Biology*, vol. 234, pp. 779–815, 12 1993.

[77] A. Fiser and A. Šali, "Modeller: Generation and Refinement of Homology-Based Protein Structure Models," *Methods in Enzymology*, vol. 374, pp. 461–491, 1 2003.

[78] A. W. Senior, R. Evans, J. Jumper, J. Kirkpatrick, L. Sifre, T. Green, C. Qin, A. Žídek, A. W. Nelson, A. Bridgland, H. Penedones, S. Petersen, K. Simonyan, S. Crossan, P. Kohli, D. T. Jones, D. Silver, K. Kavukcuoglu, and D. Hassabis, "Improved protein structure prediction using potentials from deep learning," *Nature*, vol. 577, no. 7792, pp. 706–710, 2020.

[79] M. Alquraishi, "AlphaFold at CASP13," *Bioinformatics (Oxford, England)*, vol. 35, pp. 4862–4865, 11 2019.

[80] A. Patil and M. Rane, "Convolutional Neural Networks: An Overview and Its Applications in Pattern Recognition," *Smart Innovation, Systems and Technologies*, vol. 195, pp. 21–30, 2021.

[81] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.

[82] Z. Li, F. Liu, W. Yang, S. Peng, and J. Zhou, "A Survey of Convolutional Neural Networks: Analysis, Applications, and Prospects," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, pp. 6999–7019, 12 2022.

[83] S. Pouyanfar, S. Sadiq, Y. Yan, H. Tian, Y. Tao, M. P. Reyes, M. L. Shyu, S. C. Chen, and S. S. Iyengar, "A survey on deep learning: Algorithms, techniques, and applications," *ACM Computing Surveys*, vol. 51, no. 5, 2018.

[84] A. D. Baxevanis, G. D. Bader, and D. S. Wishart, *Bioinformatics*. Wiley, 2020.

[85] A. Kryshtafovych, T. Schwede, M. Topf, K. Fidelis, and J. Moult, "Critical assessment of methods of protein structure prediction (CASP)—Round XIV," *Proteins: Structure, Function and Bioinformatics*, vol. 89, no. 12, pp. 1607–1617, 2021.

[86] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Kaiser, and I. Polosukhin, "Attention is All you Need," in *Advances in Neural Information Processing Systems* (I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds.), vol. 30, Curran Associates, Inc., 2017.

[87] D. V. Laurents, "AlphaFold 2 and NMR Spectroscopy: Partners to Understand Protein Structure, Dynamics and Function," *Frontiers in Molecular Biosciences*, vol. 9, p. 906437, 5 2022.

[88] R. Wu, F. Ding, R. Wang, R. Shen, X. Zhang, S. Luo, C. Su, Z. Wu, Q. Xie, B. Berger, J. Ma, and J. Peng, "High-resolution de novo structure prediction from primary sequence," *bioRxiv*, p. 2022.07.21.500999, 7 2022.

[89] M. Baek, F. DiMaio, I. Anishchenko, J. Dauparas, S. Ovchinnikov, G. R. Lee, J. Wang, Q. Cong, L. N. Kinch, R. Dustin Schaeffer, C. Millán, H. Park, C. Adams, C. R. Glassman, A. DeGiovanni, J. H. Pereira, A. V. Rodrigues, A. A. Van Dijk, A. C. Ebrecht, D. J. Opperman, T. Sagmeister, C. Buhlheller, T. Pavkov-Keller, M. K. Rathinaswamy, U. Dalwadi, C. K. Yip, J. E. Burke, K. Christopher Garcia, N. V. Grishin, P. D. Adams, R. J. Read, and D. Baker, "Accurate prediction of protein structures and interactions using a three-track neural network," *Science*, vol. 373, pp. 871–876, 8 2021.

[90] Z. Yang, X. Zeng, Y. Zhao, and R. Chen, "AlphaFold2 and its applications in the fields of biology and medicine," *Signal Transduction and Targeted Therapy 2023 8:1*, vol. 8, pp. 1–14, 3 2023.

[91] P.-S. Huang, S. E. Boyken, and D. Baker, "The coming of age of de novo protein design," *Nature*, vol. 537, pp. 320–327, 9 2016.

[92] E. V. Koonin, Y. I. Wolf, and G. P. Karev, "The structure of the protein universe and genome evolution," *Nature 2002 420:6912*, vol. 420, pp. 218–223, 11 2002.

[93] *ESA - How many stars are there in the Universe?* 2024-01-31.

[94] A. J. Link, M. L. Mock, and D. A. Tirrell, "Non-canonical amino acids in protein engineering," *Current Opinion in Biotechnology*, vol. 14, pp. 603–609, 12 2003.

[95] I. Drienovská and G. Roelfes, "Expanding the enzyme universe with genetically encoded unnatural amino acids," *Nature Catalysis 2020 3:3*, vol. 3, pp. 193–202, 1 2020.

[96] J. Maynard Smith, "Natural Selection and the Concept of a Protein Space," *Nature 1970 225:5232*, vol. 225, no. 5232, pp. 563–564, 1970.

[97] A. D. Keefe and J. W. Szostak, "Functional proteins from a random-sequence library," *Nature 2001 410:6829*, vol. 410, pp. 715–718, 4 2001.

[98] E. Bornberg-Bauer, "How are model protein structures distributed in sequence space?," *Biophysical Journal*, vol. 73, no. 5, pp. 2393–2403, 1997.

[99] C. A. Orengo and J. M. Thornton, "PROTEIN FAMILIES AND THEIR EVOLUTION—A STRUCTURAL PERSPECTIVE," *Annual Review of Biochemistry*, vol. 74, pp. 867–900, 6 2005.

[100] D. A. Drummond, J. D. Bloom, C. Adami, C. O. Wilke, and F. H. Arnold, "Why highly expressed proteins evolve slowly," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, pp. 14338–14343, 10 2005.

[101] J. D. Bloom, D. A. Drummond, F. H. Arnold, and C. O. Wilke, "Structural Determinants of the Rate of Protein Evolution in Yeast," *Molecular Biology and Evolution*, vol. 23, pp. 1751–1761, 9 2006.

[102] M. Kimura and T. Ota, "On Some Principles Governing Molecular Evolution*," *Proceedings of the National Academy of Sciences*, vol. 71, pp. 2848–2852, 7 1974.

[103] R. E. Cobb, R. Chao, and H. Zhao, "Directed evolution: Past, present, and future," *AIChE Journal*, vol. 59, pp. 1432–1440, 5 2013.

[104] P. A. Dalby, "Strategy and success for the directed evolution of enzymes," *Current Opinion in Structural Biology*, vol. 21, pp. 473–480, 8 2011.

[105] I. V. Korendovych and W. F. DeGrado, "De novo protein design, a retrospective," *Quarterly Reviews of Biophysics*, vol. 53, p. e3, 2020.

[106] L. Regan, D. Caballero, M. R. Hinrichsen, A. Virrueta, D. M. Williams, and C. S. O'Hern, "Protein design: Past, present, and future," *Biopolymers*, vol. 104, pp. 334–350, 7 2015.

[107] D. N. Woolfson, "A Brief History of De Novo Protein Design: Minimal, Rational, and Computational," *Journal of Molecular Biology*, vol. 433, p. 167160, 10 2021.

[108] I. V. Korendovych, "Rational and semirational protein design," *Methods in Molecular Biology*, vol. 1685, pp. 15–23, 2018.

[109] F. Cedrone, A. Ménez, and E. Quéméneur, "Tailoring new enzyme functions by rational redesign," *Current opinion in structural biology*, vol. 10, no. 4, pp. 405–410, 2000.

[110] T. Simonson, D. Mignon, K. Druart, E. Michael, V. Opuu, S. Polydorides, F. Villa, T. Gaillard, N. Panel, and G. Archontis, "Physics-based computational protein design: An update," *Journal of Physical Chemistry A*, vol. 124, pp. 10637–10648, 12 2020.

[111] N. Ferruz, M. Heinzinger, M. Akdel, A. Goncearenco, L. Naef, and C. Dallago, "From sequence to function through structure: Deep learning for protein design," *Computational and Structural Biotechnology Journal*, vol. 21, pp. 238–250, 1 2023.

[112] S. Ovchinnikov and P. S. Huang, "Structure-based protein design with deep learning," *Current Opinion in Chemical Biology*, vol. 65, pp. 136–144, 12 2021.

[113] R. Vos and L. G. Bellù, "Global Trends and Challenges to Food and Agriculture into the 21st Century," *Sustainable Food and Agriculture: An Integrated Approach*, pp. 11–30, 1 2019.

[114] M. Beauregard and M. A. Hefford, "Enhancement of essential amino acid contents in crops by genetic engineering and protein design," *Plant Biotechnology Journal*, vol. 4, pp. 561–574, 9 2006.

[115] D. J. Mandell, M. J. Lajoie, M. T. Mee, R. Takeuchi, G. Kuznetsov, J. E. Norville, C. J. Gregg, B. L. Stoddard, and G. M. Church, "Biocontainment of genetically modified organisms by synthetic protein design," *Nature*, vol. 518, pp. 55–60, 2 2015.

[116] D. Zilberman, T. G. Holland, and I. Trilnick, "Agricultural GMOs—What We Know and Where Scientists Disagree," *Sustainability 2018, Vol. 10, Page 1514*, vol. 10, p. 1514, 5 2018.

[117] D. R. Ort, S. S. Merchant, J. Alric, A. Barkan, R. E. Blankenship, R. Bock, R. Croce, M. R. Hanson, J. M. Hibberd, S. P. Long, T. A. Moore, J. Moroney, K. K. Niyogi, M. A. Parry, P. P. Peralta-Yahya, R. C. Prince, K. E. Redding, M. H. Spalding, K. J. Van Wijk, W. F. Vermaas, S. Von Caemmerer, A. P. Weber, T. O. Yeates, J. S. Yuan, and X. G. Zhu, "Redesigning photosynthesis to sustainably meet global food and bioenergy demand," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 112, pp. 8529–8536, 7 2015.

[118] T. J. Erb and J. Zarzycki, "A short history of RubisCO: The rise and fall (?) of Nature's predominant CO2 fixing enzyme," *Current opinion in biotechnology*, vol. 49, p. 100, 2 2018.

[119] A. Quijano-Rubio, H. W. Yeh, J. Park, H. Lee, R. A. Langan, S. E. Boyken, M. J. Lajoie, L. Cao, C. M. Chow, M. C. Miranda, J. Wi, H. J. Hong, L. Stewart, B. H. Oh, and D. Baker, "De novo design of modular and tunable protein biosensors," *Nature 2021 591:7850*, vol. 591, pp. 482–487, 1 2021.

[120] O. Herud-Sikimić, A. C. Stiel, M. Kolb, S. Shanmugaratnam, K. W. Berendzen, C. Feldhaus, B. Höcker, and G. Jürgens, "A biosensor for the direct visualization of auxin," *Nature 2021 592:7856*, vol. 592, pp. 768–772, 4 2021.

[121] R. Casanova-Sáez, E. Mateo-Bonmatí, and K. Ljung, "Auxin Metabolism in Plants," *Cold Spring Harbor Perspectives in Biology*, vol. 13, p. a039867, 3 2021.

[122] K. Shimizu, B. Mijiddorj, M. Usami, I. Mizoguchi, S. Yoshida, S. Akayama, Y. Hamada, A. Ohyama, K. Usui, I. Kawamura, and R. Kawano, "De novo design of a nanopore for single-molecule detection that incorporates a β-hairpin peptide," *Nature Nanotechnology 2021 17:1*, vol. 17, pp. 67–75, 11 2021.

[123] Y. L. Ying and Y. T. Long, "Nanopore-Based Single-Biomolecule Interfaces: From Information to Knowledge," *Journal of the American Chemical Society*, vol. 141, pp. 15720–15729, 10 2019.

[124] D. Branton, D. W. Deamer, A. Marziali, H. Bayley, S. A. Benner, T. Butler, M. Di Ventra, S. Garaj, A. Hibbs, X. Huang, S. B. Jovanovich, P. S. Krstic, S. Lindsay, X. S. Ling, C. H. Mastrangelo, A. Meller, J. S. Oliver, Y. V. Pershin, J. M. Ramsey, R. Riehn, G. V. Soni, V. Tabard-Cossa, M. Wanunu, M. Wiggin, and J. A. Schloss, "The potential and challenges of nanopore sequencing," *Nature Biotechnology 2008 26:10*, vol. 26, pp. 1146–1153, 10 2008.

[125] D. Deamer, M. Akeson, and D. Branton, "Three decades of nanopore sequencing," *Nature Biotechnology 2016 34:5*, vol. 34, pp. 518–524, 5 2016.

[126] B. Zakeri, J. O. Fierer, E. Celik, E. C. Chittock, U. Schwarz-Linek, V. T. Moy, and M. Howarth, "Peptide tag forming a rapid covalent bond to a protein, through engineering a bacterial adhesin," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 109, p. E690, 3 2012.

[127] R. Boni, E. A. Blackburn, D. J. Kleinjan, M. Jonaitis, F. Hewitt-Harris, M. Murdoch, S. Rosser, D. C. Hay, and L. Regan, "Chemically cross-linked hydrogels from repetitive protein arrays," *Journal of Structural Biology*, vol. 215, p. 107981, 9 2023.

[128] S. B. Liyanagedera, J. Williams, J. P. Wheatley, A. Y. Biketova, M. Hasan, A. P. Sagona, K. J. Purdy, R. J. Puxty, T. Feher, and V. Kulkarni, "SpyPhage: A Cell-Free TXTL Platform for Rapid Engineering of Targeted Phage Therapies," *ACS Synthetic Biology*, vol. 11, pp. 3330–3342, 10 2022.

[129] Y. Yang, Z. Xiao, K. Ye, X. He, B. Sun, Z. Qin, J. Yu, J. Yao, Q. Wu, Z. Bao, and W. Zhao, "SARS-CoV-2: characteristics and current advances in research," *Virology Journal*, vol. 17, pp. 1–17, 7 2020.

[130] L. Cao, I. Goreshnik, B. Coventry, J. B. Case, L. Miller, L. Kozodoy, R. E. Chen, L. Carter, A. C. Walls, Y. J. Park, E. M. Strauch, L. Stewart, M. S. Diamond, D. Veesler, and D. Baker, "De novo design of picomolar SARS-CoV-2 miniprotein inhibitors," *Science*, vol. 370, pp. 426–431, 10 2020.

[131] F. Sesterhenn, C. Yang, J. Bonet, J. T. Cramer, X. Wen, Y. Wang, C. I. Chiang, L. A. Abriata, I. Kucharska, G. Castoro, S. S. Vollers, M. Galloux, E. Dheilly, S. Rosset, P. Corthésy, S. Georgeon, M. Villard, C. A. Richard, D. Descamps, T. Delgado, E. Oricchio, M. A. Rameix-Welti, V. Más, S. Ervin, J. F. Eléouët, S. Riffault, J. T. Bates, J. P. Julien, Y. Li, T. Jardetzky, T. Krey, and B. E. Correia, "De novo protein design enables the precise induction of RSV-neutralizing antibodies," *Science*, vol. 368, 5 2020.

[132] D. Phillips, H.-C. Gasser, S. Kamp, A. Pałkowski, L. Rabalski, D. A. Oyarzún, O. Oyarzún, A. Rajan, and J. A. Alfaro, "Generating Immune-aware SARS-CoV-2 Spike Proteins for Universal Vaccine Design,"

[133] F. Jiang and J. A. Doudna, "CRISPR–Cas9 Structures and Mechanisms," *https://doi.org/10.1146/annurev-biophys-062215-010822*, vol. 46, pp. 505–529, 5 2017.

[134] J. Y. Wang and J. A. Doudna, "CRISPR technology: A decade of genome editing is only the beginning," *Science*, vol. 379, 1 2023.

[135] C. Wong, "UK first to approve CRISPR treatment for diseases: what you need to know," *Nature*, vol. 623, pp. 676–677, 11 2023.

[136] R. Gupta, D. Gupta, K. T. Ahmed, D. Dey, R. Singh, S. Swarnakar, V. Ravichandiran, S. Roy, and D. Ghosh, "Modification of Cas9, gRNA and PAM: Key to further regulate genome editing and its applications," *Progress in Molecular Biology and Translational Science*, vol. 178, pp. 85–98, 1 2021.

[137] P. S. Huang, S. E. Boyken, and D. Baker, "The coming of age of de novo protein design," *Nature 2016 537:7620*, vol. 537, pp. 320–327, 9 2016.

[138] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter, "From RNA to Protein," *Molecular Biology of the Cell*, 2002.

[139] J. T. MacDonald and P. S. Freemont, "Computational protein design with backbone plasticity," 10 2016.

[140] B. Kuhlman, G. Dantas, G. C. Ireton, G. Varani, B. L. Stoddard, and D. Baker, "Design of a Novel Globular Protein Fold with Atomic-Level Accuracy," *Science*, vol. 302, pp. 1364–1368, 11 2003.

[141] N. Ferruz, J. Noske, and B. Höcker, "Protlego: a Python package for the analysis and design of chimeric proteins," *Bioinformatics*, vol. 37, pp. 3182–3189, 10 2021.

[142] G. Grigoryan and W. F. DeGrado, "Probing Designability via a Generalized Model of Helical Bundle Geometry," *Journal of Molecular Biology*, vol. 405, pp. 1079–1100, 10 2011.

[143] C. W. Wood and D. N. Woolfson, "CCBuilder 2.0: Powerful and accessible coiled-coil modeling," *Protein Science : A Publication of the Protein Society*, vol. 27, p. 103, 1 2018.

[144] F. A. Croitoru, V. Hondru, R. T. Ionescu, and M. Shah, "Diffusion Models in Vision: A Survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, pp. 10850–10869, 9 2023.

[145] J. L. Watson, D. Juergens, N. R. Bennett, B. L. Trippe, J. Yim, H. E. Eisenach, W. Ahern, A. J. Borst, R. J. Ragotte, L. F. Milles, B. I. Wicky, N. Hanikel, S. J. Pellock, A. Courbet, W. Sheffler, J. Wang, P. Venkatesh, I. Sappington, S. V. Torres, A. Lauko, V. De Bortoli, E. Mathieu, S. Ovchinnikov, R. Barzilay, T. S. Jaakkola, F. DiMaio, M. Baek, and D. Baker, "De novo design of protein structure and function with RFdiffusion," *Nature 2023 620:7976*, vol. 620, pp. 1089–1100, 7 2023.

[146] X. Pan and T. Kortemme, "Recent advances in de novo protein design: Principles, methods, and applications," *Journal of Biological Chemistry*, vol. 296, 1 2021.

[147] C. Lee and S. Subbiah, "Prediction of protein side-chain conformation by packing optimization," *Journal of Molecular Biology*, vol. 217, pp. 373–388, 1 1991.

[148] B. Kuhlman and D. Baker, "Native protein sequences are close to optimal for their structures," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 97, pp. 10383–10388, 9 2000.

[149] I. Georgiev, D. Keedy, J. S. Richardson, D. C. Richardson, and B. R. Donald, "Algorithm for backrub motions in protein design," *Bioinformatics*, vol. 24, pp. i196–i204, 7 2008.

[150] L. V. Castorina, S. M. Ünal, K. Subr, and C. W. Wood, "TIMED-Design: Flexible and Accessible Protein Sequence Design with Convolutional Neural Networks," *Protein Engineering, Design and Selection*, p. gzae002, 2024.

[151] J. Dauparas, I. Anishchenko, N. Bennett, H. Bai, R. J. Ragotte, L. F. Milles, B. I. M. Wicky, A. Courbet, R. J. de Haas, N. Bethel, P. J. Y. Leung, T. F. Huddy, S. Pellock, D. Tischer, F. Chan, B. Koepnick, H. Nguyen, A. Kang, B. Sankaran, A. K. Bera, N. P. King, and D. Baker, "Robust deep learning based protein sequence design using ProteinMPNN," tech. rep., 9 2022.

[152] N. Ferruz, S. Schmidt, and B. Höcker, "ProtGPT2 is a deep unsupervised language model for protein design," *Nature Communications*, vol. 13, 12 2022.

[153] D. Akpinaroglu, K. Seki, A. Guo, E. Zhu, M. J. S. Kelly, and T. Kortemme, "Structure-conditioned masked language models for protein sequence design generalize beyond the native sequence space," *bioRxiv*, p. 2023.12.15.571823, 12 2023.

[154] C. B. Anfinsen, "Principles that Govern the Folding of Protein Chains," Tech. Rep. 4096, 1973.

[155] R. F. Alford, A. Leaver-Fay, J. R. Jeliazkov, M. J. O'Meara, F. P. DiMaio, H. Park, M. V. Shapovalov, P. D. Renfrew, V. K. Mulligan, K. Kappel, J. W. Labonte, M. S. Pacella, R. Bonneau, P. Bradley, R. L. Dunbrack, R. Das, D. Baker, B. Kuhlman, T. Kortemme, and J. J. Gray, "The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design," *Journal of Chemical Theory and Computation*, vol. 13, pp. 3031–3048, 6 2017.

[156] B. I. Dahiyat and S. L. Mayo, "Protein design automation," *Protein Science*, vol. 5, no. 5, pp. 895–903, 1996.

[157] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, "Basic local alignment search tool," *Journal of molecular biology*, vol. 215, no. 3, pp. 403–410, 1990.

[158] J. K. Leman, B. D. Weitzner, S. M. Lewis, J. Adolf-Bryfogle, N. Alam, R. F. Alford, M. Aprahamian, D. Baker, K. A. Barlow, P. Barth, B. Basanta, B. J. Bender, K. Blacklock, J. Bonet, S. E. Boyken, P. Bradley, C. Bystroff, P. Conway, S. Cooper, B. E. Correia, B. Coventry, R. Das, R. M. De Jong, F. DiMaio, L. Dsilva, R. Dunbrack, A. S. Ford, B. Frenz, D. Y. Fu, C. Geniesse,

L. Goldschmidt, R. Gowthaman, J. J. Gray, D. Gront, S. Guffy, S. Horowitz, P. S. Huang, T. Huber, T. M. Jacobs, J. R. Jeliazkov, D. K. Johnson, K. Kappel, J. Karanicolas, H. Khakzad, K. R. Khar, S. D. Khare, F. Khatib, A. Khramushin, I. C. King, R. Kleffner, B. Koepnick, T. Kortemme, G. Kuenze, B. Kuhlman, D. Kuroda, J. W. Labonte, J. K. Lai, G. Lapidoth, A. Leaver-Fay, S. Lindert, T. Linsky, N. London, J. H. Lubin, S. Lyskov, J. Maguire, L. Malmström, E. Marcos, O. Marcu, N. A. Marze, J. Meiler, R. Moretti, V. K. Mulligan, S. Nerli, C. Norn, S. Ó'Conchúir, N. Ollikainen, S. Ovchinnikov, M. S. Pacella, X. Pan, H. Park, R. E. Pavlovicz, M. Pethe, B. G. Pierce, K. B. Pilla, B. Raveh, P. D. Renfrew, S. S. Burman, A. Rubenstein, M. F. Sauer, A. Scheck, W. Schief, O. Schueler-Furman, Y. Sedan, A. M. Sevy, N. G. Sgourakis, L. Shi, J. B. Siegel, D. A. Silva, S. Smith, Y. Song, A. Stein, M. Szegedy, F. D. Teets, S. B. Thyme, R. Y. R. Wang, A. Watkins, L. Zimmerman, and R. Bonneau, "Macromolecular modeling and design in Rosetta: recent methods and frameworks," *Nature Methods 2020 17:7*, vol. 17, pp. 665–680, 6 2020.

[159] D. Baran, M. G. Pszolla, G. D. Lapidoth, C. Norn, O. Dym, T. Unger, S. Albeck, M. D. Tyka, and S. J. Fleishman, "Principles for computational design of binding antibodies," *Proceedings of the National Academy of Sciences*, vol. 114, pp. 10900–10905, 10 2017.

[160] D. Reichert, H. Schepers, J. Simke, H. Lechner, W. Dörner, B. Höcker, B. J. Ravoo, and A. Rentmeister, "Computational design and experimental characterization of a photo-controlled mRNA-cap guanine-N7 methyltransferase," *RSC Chemical Biology*, vol. 2, pp. 1484–1490, 10 2021.

[161] J. Huang, X. Xie, Z. Zheng, L. Ye, P. Wang, L. Xu, Y. Wu, J. Yan, M. Yang, and Y. Yan, "De Novo Computational Design of a Lipase with Hydrolysis Activity towards Middle-Chained Fatty Acid Esters," *International Journal of Molecular Sciences*, vol. 24, p. 8581, 5 2023.

[162] Y. Kipnis, A. O. Chaib, A. A. Vorobieva, G. Cai, G. Reggiano, B. Basanta, E. Kumar, P. R. Mittl, D. Hilvert, and D. Baker, "Design and optimization of enzymatic activity in a de novo β-barrel scaffold," *Protein Science*, vol. 31, p. e4405, 11 2022.

[163] X. Huang, R. Pearce, and Y. Zhang, "EvoEF2: Accurate and fast energy function

for computational protein design," *Bioinformatics*, vol. 36, pp. 1135–1142, 2 2020.

[164] E. Ong, X. Huang, R. Pearce, Y. Zhang, and Y. He, "Computational design of SARS-CoV-2 spike glycoproteins to increase immunogenicity by T cell epitope engineering," *Computational and Structural Biotechnology Journal*, vol. 19, pp. 518–529, 1 2021.

[165] L. Li, M. Gao, P. Jiao, S. Zu, Y. q. Deng, D. Wan, Y. Cao, J. Duan, S. R. Aliyari, J. Li, Y. Shi, Z. Rao, C. f. Qin, Y. Guo, G. Cheng, and H. Yang, "Antibody engineering improves neutralization activity against K417 spike mutant SARS-CoV-2 variants," *Cell and Bioscience*, vol. 12, pp. 1–15, 12 2022.

[166] S. McIntosh-Smith, T. Wilson, A. Ibarra, J. Crisp, and R. Sessions, "Benchmarking energy efficiency, power costs and carbon emissions on heterogeneous systems," *The Computer Journal*, vol. 55, pp. 192 – 205, 2 2012.

[167] S. McIntosh-Smith, J. Price, R. B. Sessions, and A. A. Ibarra, "High performance in silico virtual drug screening on many-core processors," *The International Journal of High Performance Computing Applications*, vol. 29, p. 119, 5 2015.

[168] D. W. Watkins, J. M. Jenkins, K. J. Grayson, N. Wood, J. W. Steventon, K. K. Le Vay, M. I. Goodwin, A. S. Mullen, H. J. Bailey, M. P. Crump, F. MacMillan, A. J. Mulholland, G. Cameron, R. B. Sessions, S. Mann, and J. L. Anderson, "Construction and in vivo assembly of a catalytically proficient and hyperthermostable de novo enzyme," *Nature Communications 2017 8:1*, vol. 8, pp. 1–9, 8 2017.

[169] A. R. Thomson, C. W. Wood, A. J. Burton, G. J. Bartlett, R. B. Sessions, R. L. Brady, and D. N. Woolfson, "Computational design of water-soluble α-helical barrels," *Science (New York, N.Y.)*, vol. 346, no. 6208, pp. 485–488, 2014.

[170] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, and others, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[171] T. Wu, S. He, J. Liu, S. Sun, K. Liu, Q. L. Han, and Y. Tang, "A Brief Overview of ChatGPT: The History, Status Quo and Potential Future Development," *IEEE/CAA Journal of Automatica Sinica*, vol. 10, no. 5, pp. 1122–1136, 2023.

[172] A. Madani, B. Krause, E. R. Greene, S. Subramanian, B. P. Mohr, J. M. Holton, J. L. Olmos, C. Xiong, Z. Z. Sun, R. Socher, J. S. Fraser, and N. Naik, "Large language models generate functional protein sequences across diverse families," *Nature Biotechnology 2023 41:8*, vol. 41, pp. 1099–1106, 1 2023.

[173] A. Rives, J. Meier, T. Sercu, S. Goyal, Z. Lin, J. Liu, D. Guo, M. Ott, C. L. Zitnick, J. Ma, and R. Fergus, "Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences," *Proceedings of the National Academy of Sciences*, vol. 118, 4 2021.

[174] R. Verkuil, O. Kabeli, Y. Du, B. I. M. Wicky, L. F. Milles, J. Dauparas, D. Baker, S. Ovchinnikov, T. Sercu, and A. Rives, "Language models generalize beyond natural proteins," *bioRxiv*, p. 2022.12.21.521521, 2022.

[175] S. Alamdari, N. Thakkar, R. v. d. Berg, A. X. Lu, N. Fusi, A. P. Amini, and K. K. Yang, "Protein generation with evolutionary diffusion: sequence is all you need," *bioRxiv*, p. 2023.09.11.556673, 2023.

[176] J. S. Lee, J. Kim, and P. M. Kim, "Score-based generative modeling for de novo protein design," *Nature Computational Science 2023 3:5*, vol. 3, pp. 382–392, 5 2023.

[177] J. B. Maguire, H. K. Haddox, D. Strickland, S. F. Halabiya, B. Coventry, J. R. Griffin, S. V. S. R. K. Pulavarti, M. Cummins, D. F. Thieker, E. Klavins, and others, "Perturbing the energy landscape for improved packing during computational protein design," *Proteins: Structure, Function, and Bioinformatics*, vol. 89, no. 4, pp. 436–449, 2021.

[178] J. B. Ingraham, M. Baranov, Z. Costello, K. W. Barber, W. Wang, A. Ismail, V. Frappier, D. M. Lord, C. Ng-Thow-Hing, E. R. Van Vlack, S. Tie, V. Xue, S. C. Cowles, A. Leung, J. V. Rodrigues, C. L. Morales-Perez, A. M. Ayoub, R. Green, K. Puentes, F. Oplinger, N. V. Panwar, F. Obermeyer, A. R. Root, A. L. Beam, F. J. Poelwijk, and G. Grigoryan, "Illuminating protein space with

a programmable generative model," *Nature 2023 623:7989*, vol. 623, pp. 1070–1078, 11 2023.

[179] W. R. Pearson, "Protein Function Prediction: Problems and Pitfalls," *Current Protocols in Bioinformatics*, vol. 51, pp. 1–4, 9 2015.

[180] I. Jarmoskaite, I. Alsadhan, P. P. Vaidyanathan, and D. Herschlag, "How to measure and evaluate binding affinities," *eLife*, vol. 9, pp. 1–34, 2020.

[181] V. L. Cruz, V. Souza-Egipsy, M. Gion, J. Pérez-García, J. Cortes, J. Ramos, and J. F. Vega, "Binding Affinity of Trastuzumab and Pertuzumab Monoclonal Antibodies to Extracellular HER2 Domain †," *International Journal of Molecular Sciences*, vol. 24, p. 12031, 8 2023.

[182] M. Stitt and Y. Gibon, "Why measure enzyme activities in the era of systems biology?," *Trends in Plant Science*, vol. 19, no. 4, pp. 256–265, 2014.

[183] I. L. H. Ong and K. L. Yang, "Recent developments in protease activity assays and sensors," *Analyst*, vol. 142, pp. 1867–1881, 5 2017.

[184] S. Martínez Cuesta, S. A. Rahman, N. Furnham, and J. M. Thornton, "The Classification and Evolution of Enzyme Function," *Biophysical Journal*, vol. 109, no. 6, pp. 1082–1086, 2015.

[185] "The Gene Ontology resource: enriching a GOld mine," *Nucleic acids research*, vol. 49, no. D1, pp. D325–D334, 2021.

[186] G. L. Rosano, E. S. Morales, and E. A. Ceccarelli, "New tools for recombinant protein production in Escherichia coli: A 5-year update," *Protein Science*, vol. 28, pp. 1412–1422, 8 2019.

[187] I. Anishchenko, S. J. Pellock, T. M. Chidyausiku, T. A. Ramelot, S. Ovchinnikov, J. Hao, K. Bafna, C. Norn, A. Kang, A. K. Bera, F. DiMaio, L. Carter, C. M. Chow, G. T. Montelione, and D. Baker, "De novo protein design by deep network hallucination," *Nature*, vol. 600, pp. 547–552, 9 2021.

[188] T. Di Mambro, T. Vanzolini, M. Bianchi, R. Crinelli, B. Canonico, F. Tasini, M. Menotta, and M. Magnani, "Development and in vitro characterization of a humanized scFv against fungal infections," *PLOS ONE*, vol. 17, p. e0276786, 10 2022.

[189] C. Yang, F. Sesterhenn, J. Bonet, E. A. van Aalen, L. Scheller, L. A. Abriata, J. T. Cramer, X. Wen, S. Rosset, S. Georgeon, T. Jardetzky, T. Krey, M. Fussenegger, M. Merkx, and B. E. Correia, "Bottom-up de novo design of functional proteins with complex structural features," *Nature Chemical Biology*, pp. 1–9, 1 2021.

[190] F. Katzen, G. Chang, and W. Kudlicki, "The past, present and future of cell-free protein synthesis," 3 2005.

[191] Y. C. Kwon and M. C. Jewett, "High-throughput preparation methods of crude extract for robust cell-free protein synthesis," *Scientific Reports*, vol. 5, pp. 1–8, 3 2015.

[192] Z. Swank, N. Laohakunakorn, and S. J. Maerkl, "Cell-free gene-regulatory network engineering with synthetic transcription factors," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 116, pp. 5892–5901, 3 2019.

[193] A. D. Silverman, A. S. Karim, and M. C. Jewett, "Cell-free gene expression: an expanded repertoire of applications," *Nature Reviews Genetics 2019 21:3*, vol. 21, pp. 151–170, 11 2019.

[194] R. Matsumoto, T. Niwa, Y. Shimane, Y. Kuruma, H. Taguchi, and T. Kanamori, "Regulated N-Terminal Modification of Proteins Synthesized Using a Reconstituted Cell-Free Protein Synthesis System," *ACS Synthetic Biology*, vol. 12, pp. 1935–1942, 7 2023.

[195] Y. Shimizu, A. Inoue, Y. Tomari, T. Suzuki, T. Yokogawa, K. Nishikawa, and T. Ueda, "Cell-free translation reconstituted with purified components," *Nature Biotechnology 2001 19:8*, vol. 19, pp. 751–755, 8 2001.

[196] Y. Shimizu, T. Kanamori, and T. Ueda, "Protein synthesis by pure translation systems," *Methods*, vol. 36, pp. 299–304, 7 2005.

[197] M. J. Stam and C. W. Wood, "DE-STRESS: a user-friendly web application for the evaluation of protein designs," *Protein Engineering, Design and Selection*, vol. 34, pp. 1–6, 2021.

[198] M. J. Stam, D. A. Oyarzún, N. Laohakunakorn, and C. W. Wood, "Large scale analysis of predicted protein structures links model features to in vivo behaviour," *bioRxiv*, 2024.

[199] M. A. Coleman, V. H. Lao, B. W. Segelke, , and P. T. Beernink*, "High-Throughput, Fluorescence-Based Screening for Soluble Protein Expression," *Journal of Proteome Research*, vol. 3, pp. 1024–1032, 9 2004.

[200] H. Liu and Q. Chen, "Computational protein design with data-driven approaches: Recent developments and perspectives," *Wiley Interdisciplinary Reviews: Computational Molecular Science*, vol. 13, p. e1646, 5 2023.

[201] R. F. Alford, A. Leaver-Fay, J. R. Jeliazkov, M. J. O'Meara, F. P. DiMaio, H. Park, M. V. Shapovalov, P. D. Renfrew, V. K. Mulligan, K. Kappel, J. W. Labonte, M. S. Pacella, R. Bonneau, P. Bradley, R. L. Dunbrack, R. Das, D. Baker, B. Kuhlman, T. Kortemme, and J. J. Gray, "The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design," *Journal of Chemical Theory and Computation*, vol. 13, pp. 3031–3048, 6 2017.

[202] M. Suárez, P. Tortosa, and A. Jaramillo, "PROTDES: CHARMM toolbox for computational protein design," *Systems and Synthetic Biology*, vol. 2, pp. 105–113, 7 2008.

[203] J. A. Maier, C. Martinez, K. Kasavajhala, L. Wickstrom, K. E. Hauser, and C. Simmerling, "ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB," *Journal of Chemical Theory and Computation*, vol. 11, pp. 3696–3713, 10 2015.

[204] M. ElHefnawi, M. ElGamacy, and M. Fares, "Multiple virtual screening approaches for finding new Hepatitis c virus RNA-dependent RNA polymerase inhibitors: Structure-based screens and molecular dynamics for the pursue of new poly pharmacological inhibitors," *BMC Bioinformatics*, vol. 13, p. S5, 11 2012.

[205] A. M. Poole and R. Ranganathan, "Knowledge-based potentials in protein design," *Current Opinion in Structural Biology*, vol. 16, pp. 508–513, 11 2006.

[206] K. T. Simons, C. Kooperberg, E. Huang, and D. Baker, "Assembly of protein tertiary structures from fragments with similar local sequences using simulated

annealing and bayesian scoring functions11Edited by F. E. Cohen," *Journal of Molecular Biology*, vol. 268, pp. 209–225, 11 1997.

[207] Y. Yang and Y. Zhou, "Ab initio folding of terminal segments with secondary structures reveals the fine difference between two closely related all-atom statistical energy functions," *Protein Science*, vol. 17, pp. 1212–1219, 10 2008.

[208] C. Negron and A. E. Keating, "A Set of Computationally Designed Orthogonal Antiparallel Homodimers that Expands the Synthetic Coiled-Coil Toolkit," 2014.

[209] H. Zhou and J. Skolnick, "GOAP: A generalized orientation-dependent, all-atom statistical potential for protein structure prediction," *Biophysical Journal*, vol. 101, pp. 2043–2052, 10 2011.

[210] V. Parthiban, M. M. Gromiha, and D. Schomburg, "CUPSAT: prediction of protein stability upon point mutations," *Nucleic Acids Research*, vol. 34, pp. W239–W242, 10 2006.

[211] J. Laimer, H. Hofer, M. Fritz, S. Wegenkittl, and P. Lackner, "MAESTRO - multi agent stability prediction upon point mutations," *BMC Bioinformatics*, vol. 16, p. 116, 10 2015.

[212] E. C. Alley, G. Khimulya, S. Biswas, M. AlQuraishi, and G. M. Church, "Unified rational protein engineering with sequence-based deep representation learning," *Nature Methods*, vol. 16, pp. 1315–1322, 10 2019.

[213] X. Gao, F. Ramezanghorbani, O. Isayev, J. S. Smith, and A. E. Roitberg, "TorchANI: A Free and Open Source PyTorch-Based Deep Learning Implementation of the ANI Neural Network Potentials," *Journal of Chemical Information and Modeling*, vol. 60, pp. 3408–3415, 10 2020.

[214] C. Pancotti, S. Benevenuta, V. Repetto, G. Birolo, E. Capriotti, T. Sanavia, and P. Fariselli, "A Deep-Learning Sequence-Based Method to Predict Protein Stability Changes Upon Genetic Variations," *Genes*, vol. 12, p. 911, 10 2021.

[215] H. Yang, Z. Xiong, and F. Zonta, "Construction of a Deep Neural Network Energy Function for Protein Physics," *Journal of Chemical Theory and Computation*, vol. 18, pp. 5649–5658, 10 2022.

[216] L. M. Blaabjerg, M. M. Kassem, L. L. Good, N. Jonsson, M. Cagiada, K. E. Johansson, W. Boomsma, A. Stein, and K. Lindorff-Larsen, "Rapid protein stability prediction using deep learning representations," tech. rep., 11 2022.

[217] A. M. Hermosilla, C. Berner, S. Ovchinnikov, and A. A. Vorobieva, "Validation of de novo designed water-soluble and transmembrane proteins by in silico folding and melting," *bioRxiv*, p. 2023.06.06.543955, 8 2023.

[218] O. Conchillo-Solé, N. S. de Groot, F. X. Avilés, J. Vendrell, X. Daura, and S. Ventura, "AGGRESCAN: a server for the prediction and evaluation of "hot spots" of aggregation in polypeptides," *BMC Bioinformatics*, vol. 8, p. 65, 10 2007.

[219] A. Kuriata, V. Iglesias, M. Kurcinski, S. Ventura, and S. Kmiecik, "Aggrescan3D standalone package for structure-based prediction of protein aggregation properties," *Bioinformatics*, vol. 35, pp. 3834–3835, 10 2019.

[220] A. Trovato, F. Chiti, A. Maritan, and F. Seno, "Insight into the Structure of Amyloid Fibrils from the Analysis of Globular Proteins," *PLOS Computational Biology*, vol. 2, p. e170, 10 2006.

[221] I. Walsh, F. Seno, S. C. E. Tosatto, and A. Trovato, "PASTA 2.0: an improved server for protein aggregation prediction," *Nucleic Acids Research*, vol. 42, pp. W301–W307, 10 2014.

[222] M. Hebditch, M. A. Carballo-Amador, S. Charonis, R. Curtis, and J. Warwicker, "Protein–Sol: a web tool for predicting protein solubility from sequence," *Bioinformatics*, vol. 33, pp. 3098–3100, 10 2017.

[223] J. Hon, M. Marusiak, T. Martinek, A. Kunka, J. Zendulka, D. Bednar, and J. Damborsky, "SoluProt: prediction of soluble protein expression in ¡i¿Escherichia coli¡/i¿," *Bioinformatics*, vol. 37, pp. 23–28, 10 2021.

[224] V. Thumuluri, H.-M. Martiny, J. J. Almagro Armenteros, J. Salomon, H. Nielsen, and A. R. Johansen, "NetSolP: predicting protein solubility in Escherichia coli using language models," *Bioinformatics*, vol. 38, pp. 941–946, 10 2022.

[225] N. S. de Groot, I. Pallarés, F. X. Avilés, J. Vendrell, and S. Ventura, "Prediction of "hot spots" of aggregation in disease-linked polypeptides," *BMC Structural Biology*, vol. 5, p. 18, 11 2005.

[226] A. Kuriata, V. Iglesias, J. Pujols, M. Kurcinski, S. Kmiecik, and S. Ventura, "Aggrescan3D (A3D) 2.0: prediction and engineering of protein solubility," *Nucleic Acids Research*, vol. 47, pp. W300–W307, 10 2019.

[227] R. Rao, J. Meier, T. Sercu, S. Ovchinnikov, and A. Rives, "Transformer protein language models are unsupervised structure learners," tech. rep., 11 2020.

[228] T. A. Binkowski, S. Naghibzadeh, and J. Liang, "CASTp: Computed Atlas of Surface Topography of proteins," *Nucleic Acids Research*, vol. 31, pp. 3352–3355, 7 2003.

[229] J. Dundas, Z. Ouyang, J. Tseng, A. Binkowski, Y. Turpaz, and J. Liang, "CASTp: computed atlas of surface topography of proteins with structural and topographical mapping of functionally annotated residues," *Nucleic Acids Research*, vol. 34, pp. W116–W118, 7 2006.

[230] W. Tian, C. Chen, X. Lei, J. Zhao, and J. Liang, "CASTp 3.0: computed atlas of surface topography of proteins," *Nucleic Acids Research*, vol. 46, pp. W363–W367, 10 2018.

[231] P. Garg, S. Sacher, P. Gautam, and A. Ray, "CICLOP: a robust and accurate computational framework for protein inner cavity detection," *Bioinformatics*, vol. 38, pp. 2153–2161, 10 2022.

[232] A. Jurcik, D. Bednar, J. Byska, S. M. Marques, K. Furmanova, L. Daniel, P. Kokkonen, J. Brezovsky, O. Strnad, J. Stourac, A. Pavelka, M. Manak, J. Damborsky, and B. Kozlikova, "CAVER Analyst 2.0: analysis and visualization of channels and tunnels in protein structures and molecular dynamics trajectories," *Bioinformatics*, vol. 34, pp. 3586–3588, 10 2018.

[233] A. Pavelka, E. Sebestova, B. Kozlikova, J. Brezovsky, J. Sochor, and J. Damborsky, "CAVER: Algorithms for Analyzing Dynamics of Tunnels in Macromolecules," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 13, pp. 505–517, 5 2015.

[234] P. B. O. S. J. B. B. K. A. G. V. M. K. P. M. L. B. J. S. Eva Chovancová Antonín Pavelka and J. Damborský, "CAVER 3.0: A Tool for the Analysis of Transport Pathways in Dynamic Protein Structures," *Pathways in Dynamic Protein Structures, PLoS Computational Biology 8: e1002708*, 2012.

[235] L. Schrödinger, "The PyMOL Molecular Graphics System, Version˜2.3." 11 2015.

[236] H. William, "VMD-visual molecular dynamics," *Journal of molecular graphics*, vol. 14, pp. 33–38, 1996.

[237] M. S. Weiss, "On the interrelationship between atomic displacement parameters (ADPs) and coordinates in protein structures," *Acta Crystallographica Section D: Biological Crystallography*, vol. 63, pp. 1235–1242, 10 2007.

[238] E. S. Huang, S. Subbiah, and M. Levitt, "Recognizing Native Folds by the Arrangement of Hydrophobic and Polar Residues," *Journal of Molecular Biology*, vol. 252, pp. 709–720, 10 1995.

[239] C. W. Wood, J. W. Heal, A. R. Thomson, G. J. Bartlett, A. Ibarra, R. L. Brady, R. B. Sessions, and D. N. Woolfson, "{ISAMBARD}: an open-source computational environment for biomolecular analysis, modelling and design," *Bioinformatics*, vol. 33, pp. 3043–3050, 10 2017.

[240] V. G, "WHAT IF: a molecular modeling and drug design program," *Journal of molecular graphics*, vol. 8, no. 1, pp. 52–56, 1990.

[241] W. Wang, Z. Li, J. Wang, D. Xu, and Y. Shang, "PSICA: a fast and accurate web service for protein model quality analysis," *Nucleic Acids Research*, vol. 47, p. W443, 7 2019.

[242] C. J. Williams, J. J. Headd, N. W. Moriarty, M. G. Prisant, L. L. Videau, L. N. Deis, V. Verma, D. A. Keedy, B. J. Hintze, V. B. Chen, S. Jain, S. M. Lewis, W. B. Arendall, J. Snoeyink, P. D. Adams, S. C. Lovell, J. S. Richardson, and D. C. Richardson, "MolProbity: More and better reference data for improved all-atom structure validation," *Protein Science*, vol. 27, pp. 293–315, 1 2018.

[243] R. A. Laskowski, M. W. MacArthur, D. S. Moss, and J. M. Thornton, "PROCHECK: a program to check the stereochemical quality of protein structures," *Journal of Applied Crystallography*, vol. 26, pp. 283–291, 4 1993.

[244] J. Ouyang, N. Huang, and Y. Jiang, "A single-model quality assessment method for poor quality protein structure," *BMC Bioinformatics 2020 21:1*, vol. 21, pp. 1–10, 4 2020.

[245] N. Ferruz, S. Schmidt, and B. Höcker, "ProteinTools: a toolkit to analyze protein structures," *Nucleic Acids Research*, vol. 49, pp. W559–W566, 7 2021.

[246] H. M. Berman, "The Protein Data Bank: a historical perspective," *Acta Crystallographica Section A*, vol. 64, pp. 88–95, 1 2008.

[247] C. W. Wood, J. W. Heal, A. R. Thomson, G. J. Bartlett, A. Ibarra, R. L. Brady, R. B. Sessions, and D. N. Woolfson, "ISAMBARD: an open-source computational environment for biomolecular analysis, modelling and design," *Bioinformatics*, vol. 33, pp. 3043–3050, 4 2017.

[248] W. MS, "On the interrelationship between atomic displacement parameters (ADPs) and coordinates in protein structures," *Acta crystallographica. Section D, Biological crystallography*, vol. 63, pp. 1235–1242, 11 2007.

[249] R. AS and H. PW, "NGL Viewer: a web application for molecular visualization," *Nucleic acids research*, vol. 43, no. W1, pp. W576–W579, 2015.

[250] A. S. Rose, A. R. Bradley, Y. Valasatava, J. M. Duarte, A. Prlić, and P. W. Rose, "Web-based molecular graphics for large complexes," *Proceedings of the 21st International Conference on Web3D Technology, Web3D 2016*, pp. 185–186, 7 2016.

[251] K. W and S. C, "Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features," *Biopolymers*, vol. 22, no. 12, pp. 2577–2637, 1983.

[252] W. G. Touw, C. Baakman, J. Black, T. A. te Beek, E. Krieger, R. P. Joosten, and G. Vriend, "A series of PDB-related databanks for everyday needs," *Nucleic Acids Research*, vol. 43, p. D364, 1 2015.

[253] H. Deng, Y. Jia, and Y. Zhang, "3DRobot: automated generation of diverse and well-packed protein structure decoys," *Bioinformatics*, vol. 32, pp. 378–387, 2 2016.

[254] K. Pearson, "LIII. On lines and planes of closest fit to systems of points in space," *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, vol. 2, pp. 559–572, 9 1901.

[255] L. Mcinnes, J. Healy, and J. Melville, "UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction," tech. rep., 2018.

[256] R. Phillips and R. Milo, "A feeling for the numbers in biology," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 106, pp. 21465–21471, 12 2009.

[257] R. J. Ellis, "Tackling unintelligent design," *Nature 2010 463:7278*, vol. 463, pp. 164–165, 1 2010.

[258] M. A. Parry, P. J. Andralojc, J. C. Scales, M. E. Salvucci, A. E. Carmo-Silva, H. Alonso, and S. M. Whitney, "Rubisco activity and regulation as targets for crop improvement," *Journal of Experimental Botany*, vol. 64, pp. 717–730, 1 2013.

[259] P. Horton, S. P. Long, P. Smith, S. A. Banwart, and D. J. Beerling, "Technologies to deliver food and climate security through agriculture," *Nature Plants 2021 7:3*, vol. 7, pp. 250–255, 3 2021.

[260] T. Hauser, L. Popilka, F. U. Hartl, and M. Hayer-Hartl, "Role of auxiliary proteins in Rubisco biogenesis and function," *Nature plants*, vol. 1, 6 2015.

[261] L. Schulz, Z. Guo, J. Zarzycki, W. Steinchen, J. M. Schuller, T. Heimerl, S. Prinz, O. Mueller-Cajar, T. J. Erb, and G. K. Hochberg, "Evolution of increased complexity and specificity at the dawn of form I Rubiscos," *Science*, vol. 378, 10 2022.

[262] H. Aigner, R. H. Wilson, A. Bracher, L. Calisse, J. Y. Bhat, F. U. Hartl, and M. Hayer-Hartl, "Plant RuBisCo assembly in E. coli with five chloroplast chaperones including BSD2," *Science*, vol. 358, pp. 1272–1278, 12 2017.

[263] F. Cesaratto, O. R. Burrone, and G. Petris, "Tobacco Etch Virus protease: A shortcut across biotechnologies," *Journal of Biotechnology*, vol. 231, pp. 239–249, 8 2016.

[264] R. B. Kapust, J. Toözseór, T. D. Copeland, and D. S. Waugh, "The P1 specificity of tobacco etch virus protease," *Biochemical and Biophysical Research Communications*, vol. 294, pp. 949–955, 6 2002.

[265] M. C. Wehr and M. J. Rossner, "Split protein biosensor assays in molecular pharmacological studies," *Drug Discovery Today*, vol. 21, pp. 415–429, 3 2016.

[266] G. Barnea, W. Strapps, G. Herrada, Y. Berman, J. Ong, B. Kloss, R. Axel, and K. J. Lee, "The genetic design of signaling cascades to record receptor activation," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 105, pp. 64–69, 1 2008.

[267] S. Enríquez-Flores, J. I. De la Mora-De la Mora, L. A. Flores-López, N. Cabrera, C. Fernández-Lainez, G. Hernández-Alcántara, C. E. Guerrero-Beltrán, G. López-Velázquez, and I. García-Torres, "Improved yield, stability, and cleavage reaction of a novel tobacco etch virus protease mutant," *Applied Microbiology and Biotechnology*, vol. 106, pp. 1475–1492, 2 2022.

[268] K. Valegård, D. Hasse, I. Andersson, and L. H. Gunn, "Structure of Rubisco from Arabidopsis thaliana in complex with 2-carboxyarabinitol-1,5-bisphosphate," *Acta Crystallographica Section D: Structural Biology*, vol. 74, pp. 1–9, 1 2018.

[269] C. W. Wood, A. A. Ibarra, G. J. Bartlett, A. J. Wilson, A. J. Wilson, D. N. Woolfson, D. N. Woolfson, D. N. Woolfson, R. B. Sessions, and R. B. Sessions, "BAlaS: fast, interactive and accessible computational alanine-scanning using BudeAlaScan," *Bioinformatics*, vol. 36, pp. 2917–2919, 5 2020.

[270] A. W. Senior, R. Evans, J. Jumper, J. Kirkpatrick, L. Sifre, T. Green, C. Qin, A. Žídek, A. W. R. Nelson, A. Bridgland, H. Penedones, S. Petersen, K. Simonyan, S. Crossan, P. Kohli, D. T. Jones, D. Silver, K. Kavukcuoglu, and D. Hassabis, "Improved protein structure prediction using potentials from deep learning," *Nature*, vol. 577, pp. 706–710, 1 2020.

[271] C. Hsu, R. Verkuil, J. Liu, Z. Lin, B. Hie, T. Sercu, A. Lerer, and A. Rives, "Learning inverse folding from millions of predicted structures," in *Proceedings of the 39th {International} {Conference} on {Machine} {Learning}*, pp. 8946–8970, PMLR, 6 2022.

[272] D. Akpinaroglu, J. A. Ruffolo, S. P. Mahajan, and J. J. Gray, "Improved antibody structure prediction by deep learning of side chain conformations," *bioRxiv*, p. 2021.09.22.461349, 9 2021.

[273] D. M. Mason, S. Friedensohn, C. R. Weber, C. Jordi, B. Wagner, S. M. Meng, R. A. Ehling, L. Bonati, J. Dahinden, P. Gainza, B. E. Correia, and S. T. Reddy, "Optimization of therapeutic antibodies by predicting antigen specificity from antibody sequence via deep learning," *Nature Biomedical Engineering 2021 5:6*, vol. 5, pp. 600–612, 4 2021.

[274] X. Zhou, W. Zheng, Y. Li, R. Pearce, C. Zhang, E. W. Bell, G. Zhang, and Y. Zhang, "I-TASSER-MTD: a deep-learning-based platform for multi-domain protein structure and function prediction," *Nature Protocols 2022 17:10*, vol. 17, pp. 2326–2353, 8 2022.

[275] T. Yu, H. Cui, J. C. Li, Y. Luo, G. Jiang, and H. Zhao, "Enzyme function prediction using contrastive learning," *Science*, vol. 379, pp. 1358–1363, 3 2023.

[276] M. Mirdita, K. Schütze, Y. Moriwaki, L. Heo, S. Ovchinnikov, and M. Steinegger, "ColabFold: making protein folding accessible to all," *Nature Methods 2022 19:6*, vol. 19, pp. 679–682, 5 2022.

[277] P. Eastman, J. Swails, J. D. Chodera, R. T. McGibbon, Y. Zhao, K. A. Beauchamp, L. P. Wang, A. C. Simmonett, M. P. Harrigan, C. D. Stern, R. P. Wiewiora, B. R. Brooks, and V. S. Pande, "OpenMM 7: Rapid development of high performance algorithms for molecular dynamics," *PLOS Computational Biology*, vol. 13, p. e1005659, 7 2017.

[278] Z. Gao, C. Tan, P. Chacón, and S. Z. Li, "PiFold: Toward effective and efficient protein inverse folding," *International Conference on Learning Representations*, 9 2022.

[279] T. M. Chidyausiku, S. R. Mendes, J. C. Klima, M. Nadal, U. Eckhard, J. Roel-Touris, S. Houliston, T. Guevara, H. K. Haddox, A. Moyer, C. H. Arrowsmith, F. X. Gomis-Rüth, D. Baker, and E. Marcos, "De novo design of immunoglobulin-like domains," *Nature Communications*, vol. 13, p. 5661, 10 2022.

[280] X. Huang, R. Pearce, and Y. Zhang, "De novo design of protein peptides to block association of the SARS-CoV-2 spike protein with human ACE2," *Aging*, vol. 12, pp. 11263–11276, 10 2020.

[281] J. Loughrey and P. Cunningham, "Overfitting in wrapper-based feature subset selection: The harder you try the worse it gets," in *International conference on innovative techniques and applications of artificial intelligence*, pp. 33–43, Springer, 2004.

[282] F. Urbina, F. Lentzos, C. Invernizzi, and S. Ekins, "Dual use of artificial-intelligence-powered drug discovery," *Nature Machine Intelligence 2022 4:3*, vol. 4, pp. 189–191, 3 2022.

[283] D. Baker and G. Church, "Protein design meets biosecurity," *Science*, vol. 383, p. 349, 1 2024.

[284] J. G. Greener, S. M. Kandathil, and D. T. Jones, "Deep learning extends de novo protein modelling coverage of genomes using iteratively predicted structural constraints," *Nature Communications 2019 10:1*, vol. 10, pp. 1–13, 9 2019.

[285] J. Durairaj, A. M. Waterhouse, T. Mets, T. Brodiazhenko, M. Abdullah, G. Studer, G. Tauriello, M. Akdel, A. Andreeva, A. Bateman, T. Tenson, V. Hauryliuk, T. Schwede, and J. Pereira, "Uncovering new families and folds in the natural protein universe," *Nature 2023 622:7983*, vol. 622, pp. 646–653, 9 2023.

[286] D. Ochoa, A. Hercules, M. Carmona, D. Suveges, J. Baker, C. Malangone, I. Lopez, A. Miranda, C. Cruz-Castillo, L. Fumis, M. Bernal-Llinares, K. Tsukanov, H. Cornu, K. Tsirigos, O. Razuvayevskaya, A. Buniello, J. Schwartzentruber, M. Karim, B. Ariano, R. Martinez Osorio, J. Ferrer, X. Ge, S. Machlitt-Northen, A. Gonzalez-Uriarte, S. Saha, S. Tirunagari, C. Mehta, J. Roldán-Romero, S. Horswell, S. Young, M. Ghoussaini, D. Hulcoop, I. Dunham, and E. McDonagh, "The next-generation Open Targets Platform: reimagined, redesigned, rebuilt," *Nucleic Acids Research*, vol. 51, pp. D1353–D1359, 9 2023.

[287] C. Hsu, R. Verkuil, J. Liu, Z. Lin, B. Hie, T. Sercu, A. Lerer, and A. Rives, "Learning inverse folding from millions of predicted structures," in *International Conference on Machine Learning*, pp. 8946–8970, PMLR, 9 2022.

[288] D. Moi, C. Bernard, M. Steinegger, Y. Nevers, M. Langleib, and C. Dessimoz, "Structural phylogenetics unravels the evolutionary diversification of communication systems in gram-positive bacteria and their viruses," *bioRxiv*, p. 2023.09.19.558401, 10 2023.

[289] E. Callaway, "'A Pandora's box': map of protein-structure families delights scientists," *Nature*, vol. 621, pp. 455–455, 9 2023.

[290] J. Dunbar, K. Krawczyk, J. Leem, T. Baker, A. Fuchs, G. Georges, J. Shi, and C. M. Deane, "SAbDab: the structural antibody database," *Nucleic Acids Research*, vol. 42, pp. D1140–D1146, 1 2014.

[291] C. Schneider, M. I. Raybould, and C. M. Deane, "SAbDab in the age of biotherapeutics: updates including SAbDab-nano, the nanobody structure tracker," *Nucleic Acids Research*, vol. 50, pp. D1368–D1372, 1 2022.

[292] C. Spearman, "The Proof and Measurement of Association between Two Things," *The American Journal of Psychology*, vol. 100, no. 3/4, pp. 441–471, 1987.

[293] D. J. C. Mackay, *Information Theory, Inference and Learning Algorithms*. Cambridge University Press, 2003.

[294] L. Breiman, "Random forests," *Machine Learning*, vol. 45, pp. 5–32, 10 2001.

[295] M. Steinegger and J. Söding, "MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets," *Nature Biotechnology*, vol. 35, pp. 1026–1028, 9 2017.

[296] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*, vol. 5.1, pp. 281–298, University of California Press, 9 1967.

[297] W. M. Rand, "Objective Criteria for the Evaluation of Clustering Methods," *Journal of the American Statistical Association*, vol. 66, pp. 846–850, 12 1971.

[298] S. C. Johnson, "Hierarchical clustering schemes," *Psychometrika*, vol. 32, pp. 241–254, 9 1967.

[299] C. L. Schoch, S. Ciufo, M. Domrachev, C. L. Hotton, S. Kannan, R. Khovanskaya, D. Leipe, R. Mcveigh, K. O'Neill, B. Robbertse, S. Sharma, V. Soussov, J. P. Sullivan, L. Sun, S. Turner, and I. Karsch-Mizrachi, "NCBI Taxonomy: a comprehensive update on curation, resources and tools," *Database*, vol. 2020, p. baaa062, 9 2020.

[300] M. R. Smith, "Information theoretic generalized Robinson–Foulds metrics for comparing phylogenetic trees," *Bioinformatics*, vol. 36, pp. 5007–5013, 9 2020.

[301] I. Letunic and P. Bork, "Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation," *Nucleic Acids Research*, vol. 49, pp. W293–W296, 9 2021.

[302] W. J. Jones, J. A. Leigh, F. Mayer, C. R. Woese, and R. S. Wolfe, "Methanococcus jannaschii sp. nov., an extremely thermophilic methanogen from a submarine hydrothermal vent," *Archives of Microbiology*, vol. 136, pp. 254–261, 9 1983.

[303] K. D. Doig, K. E. Holt, J. A. Fyfe, C. J. Lavender, M. Eddyani, F. Portaels, D. Yeboah-Manu, G. Pluschke, T. Seemann, and T. P. Stinear, "On the origin of Mycobacterium ulcerans, the causative agent of Buruli ulcer," *BMC Genomics*, vol. 13, pp. 1–19, 6 2012.

[304] P. Fey, A. S. Kowal, P. Gaudet, K. E. Pilcher, and R. L. Chisholm, "Protocols for growth and development of Dictyostelium discoideum," *Nature Protocols 2007 2:6*, vol. 2, pp. 1307–1316, 5 2007.

[305] J. M. T. Carneiro, K. Chacón-Madrid, B. C. M. Maciel, and M. A. Z. Arruda, "Arabidopsis thaliana and omics approaches: A review," 6 2015.

[306] B. Müller and U. Grossniklaus, "Model organisms — A historical perspective," *Journal of Proteomics*, vol. 73, pp. 2054–2063, 10 2010.

[307] W. Basile, M. Salvatore, C. Bassot, and A. Elofsson, "Why do eukaryotic proteins contain more intrinsically disordered regions?," *PLOS Computational Biology*, vol. 15, p. e1007186, 9 2019.

[308] L. Brocchieri and S. Karlin, "Protein length in eukaryotic and prokaryotic proteomes," *Nucleic Acids Research*, vol. 33, pp. 3390–3400, 9 2005.

[309] J. Kiraga, P. Mackiewicz, D. Mackiewicz, M. Kowalczuk, P. Biecek, N. Polak, K. Smolarczyk, M. R. Dudek, and S. Cebrat, "The relationships between the isoelectric point and: Length of proteins, taxonomy and ecology of organisms," *BMC Genomics*, vol. 8, pp. 1–16, 6 2007.

[310] A. K. Chamberlain and J. U. Bowie, "Analysis of side-chain rotamers in transmembrane proteins," *Biophysical Journal*, vol. 87, pp. 3460–3469, 11 2004.

[311] J. Zhang and J.-R. Yang, "Determinants of the rate of protein sequence evolution," *Nature Reviews Genetics*, vol. 16, pp. 409–420, 7 2015.

[312] B. Tihanyi and L. Nyitray, "Recent advances in CHO cell line development for recombinant protein production," *Drug Discovery Today: Technologies*, vol. 38, pp. 25–34, 9 2020.

[313] J. E. N. Müller, F. Meyer, B. Litsanov, P. Kiefer, E. Potthoff, S. Heux, W. J. Quax, V. F. Wendisch, T. Brautaset, J.-C. Portais, and J. A. Vorholt, "Engineering Escherichia coli for methanol conversion," *Metabolic Engineering*, vol. 28, pp. 190–201, 9 2015.

[314] The UniProt Consortium, "{UniProt}: the {Universal} {Protein} {Knowledgebase} in 2023," *Nucleic Acids Research*, vol. 51, pp. D523–D531, 1 2023.

[315] G. B. Kim, J. Y. Kim, J. A. Lee, C. J. Norsigian, B. O. Palsson, and S. Y. Lee, "Functional annotation of enzyme-encoding genes using deep learning with transformer layers," *Nature Communications 2023 14:1*, vol. 14, pp. 1–11, 11 2023.

[316] P. Li and S. Chen, "A review on Gaussian Process Latent Variable Models," *CAAI Transactions on Intelligence Technology*, vol. 1, pp. 366–376, 10 2016.

[317] D. Mattanovich, P. Branduardi, L. Dato, B. Gasser, M. Sauer, and D. Porro, "Recombinant protein production in yeasts," *Methods in molecular biology (Clifton, N.J.)*, vol. 824, pp. 329–358, 2012.

[318] J. Osz-Papai, L. Radu, W. Abdulrahman, I. Kolb-Cheynel, N. Troffer-Charlier, C. Birck, and A. Poterszman, "Insect cells-baculovirus system for the production of difficult to express proteins," *Methods in molecular biology (Clifton, N.J.)*, vol. 1258, pp. 181–205, 2015.

[319] M. J. Burnett and A. C. Burnett, "Therapeutic recombinant protein production in plants: Challenges and opportunities," *Plants, People, Planet*, vol. 2, pp. 121–132, 3 2020.

[320] A. L. Demain and P. Vaishnav, "Production of recombinant proteins by microbes and higher organisms," *Biotechnology Advances*, vol. 27, pp. 297–306, 5 2009.

[321] N. J. Claassens, S. Burgener, B. Vögeli, T. J. Erb, and A. Bar-Even, "A critical comparison of cellular and cell-free bioproduction systems," *Current Opinion in Biotechnology*, vol. 60, pp. 221–229, 12 2019.

[322] K. L. Garner, "Principles of synthetic biology," *Essays in Biochemistry*, vol. 65, pp. 791–811, 11 2021.

[323] C. G. Kurland, "Codon bias and gene expression," *FEBS Letters*, vol. 285, pp. 165–169, 7 1991.

[324] B. Kelley, "Developing therapeutic monoclonal antibodies at pandemic pace," *Nature Biotechnology 2020 38:5*, vol. 38, pp. 540–545, 4 2020.

[325] Z. Kis, R. Shattock, N. Shah, and C. Kontoravdi, "Emerging Technologies for Low-Cost, Rapid Vaccine Manufacture," *Biotechnology Journal*, vol. 14, p. 1800376, 1 2019.

[326] S. Kapoor, A. Rafiq, and S. Sharma, "Protein engineering and its applications in food industry," *Critical Reviews in Food Science and Nutrition*, vol. 57, pp. 2321–2329, 7 2017.

[327] J. K. Hong, M. Lakshmanan, C. Goudar, and D. Y. Lee, "Towards next generation CHO cell line development and engineering by systems approaches," *Current Opinion in Chemical Engineering*, vol. 22, pp. 1–10, 12 2018.

[328] L. A. Bui, S. Hurst, G. L. Finch, B. Ingram, I. A. Jacobs, C. F. Kirchhoff, C. K. Ng, and A. M. Ryan, "Key considerations in the preclinical development of biosimilars," *Drug Discovery Today*, vol. 20, pp. 3–15, 5 2015.

[329] N. McGillicuddy, P. Floris, S. Albrecht, and J. Bones, "Examining the sources of variability in cell culture media used for biopharmaceutical production," *Biotechnology Letters*, vol. 40, pp. 5–21, 1 2018.

[330] A. Pekarsky, M. Reininger, and O. Spadiut, "The impact of technical failures on recombinant production of soluble proteins in Escherichia coli: a case study on process and protein robustness," *Bioprocess and Biosystems Engineering*, vol. 44, pp. 1049–1061, 6 2021.

[331] M. Kim, P. M. O'Callaghan, K. A. Droms, and D. C. James, "A mechanistic understanding of production instability in CHO cell lines expressing recombinant monoclonal antibodies," *Biotechnology and Bioengineering*, vol. 108, pp. 2434–2446, 10 2011.

[332] T. Tharmalingam, H. Barkhordarian, N. Tejeda, K. Daris, S. Yaghmour, P. Yam, F. Lu, C. Goudar, T. Munro, and J. Stevens, "Characterization of phenotypic and genotypic diversity in subclones derived from a clonal cell line," *Biotechnology Progress*, vol. 34, pp. 613–623, 5 2018.

[333] S. Plotkin, J. M. Robinson, G. Cunningham, R. Iqbal, and S. Larsen, "The complexity and cost of vaccine manufacturing – An overview," *Vaccine*, vol. 35, pp. 4064–4071, 7 2017.

[334] L. Falzon, M. Suzuki, and M. Inouye, "Finding one of a kind: advances in single-protein production," *Current Opinion in Biotechnology*, vol. 17, pp. 347–352, 8 2006.

[335] Z. Z. Sun, C. A. Hayes, J. Shin, F. Caschera, R. M. Murray, and V. Noireaux, "Protocols for Implementing an Escherichia coli Based TX-TL Cell-Free Expression System for Synthetic Biology," *J. Vis. Exp*, no. 79, p. 50762, 2013.

[336] R. Marshall, C. S. Maxwell, S. P. Collins, C. L. Beisel, and V. Noireaux, "Short DNA containing χ sites enhances DNA stability and gene expression in E. coli cell-free transcription–translation systems," *Biotechnology and Bioengineering*, vol. 114, pp. 2137–2141, 9 2017.

[337] K. Sitaraman, D. Esposito, G. Klarmann, S. F. Le Grice, J. L. Hartley, and D. K. Chatterjee, "A novel cell-free protein synthesis system," *Journal of Biotechnology*, vol. 110, pp. 257–263, 6 2004.

[338] M. Chamberlin, J. Mcgrath, and L. Waskell, "New RNA Polymerase from Escherichia coli infected with Bacteriophage T7," *Nature 1970 228:5268*, vol. 228, no. 5268, pp. 227–231, 1970.

[339] J. L. Dopp, Y. R. Jo, and N. F. Reuel, "Methods to reduce variability in E. Coli-based cell-free protein expression experiments," *Synthetic and Systems Biotechnology*, vol. 4, pp. 204–211, 12 2019.

[340] S. Wang, S. Majumder, N. J. Emery, and A. P. Liu, "Simultaneous monitoring of transcription and translation in mammalian cell-free expression in bulk and in cell-sized droplets," *Synthetic Biology*, vol. 3, no. 1, pp. 1–9, 2018.

[341] M. Drost, J. B. Zonneveld, L. Van Dijk, H. Morreau, C. M. Tops, H. F. Vasen, J. T. Wijnen, and N. De Wind, "A cell-free assay for the functional analysis of variants of the mismatch repair protein MLH1," *Human Mutation*, vol. 31, pp. 247–253, 3 2010.

[342] L. E. Contreras-Llano and C. Tan, "High-throughput screening of biomolecules using cell-free gene expression systems," *Synthetic Biology*, vol. 3, 1 2018.

[343] H. Sun, N. Hu, and J. Wang, "Application of microfluidic technology in antibody screening," *Biotechnology Journal*, vol. 17, p. 2100623, 8 2022.

[344] J. B. McManus, C. B. Bernhards, C. E. Sharpes, D. C. Garcia, S. D. Cole, R. M. Murray, P. A. Emanuel, and M. W. Lux, "Rapid Characterization of Genetic Parts with Cell-Free Systems," *JoVE (Journal of Visualized Experiments)*, vol. 2021, p. e62816, 8 2021.

[345] X. Jin and S. H. Hong, "Cell-free protein synthesis for producing 'difficult-to-express' proteins," *Biochemical Engineering Journal*, vol. 138, pp. 156–164, 10 2018.

[346] L. Thoring, S. K. Dondapati, M. Stech, D. A. Wüstenhagen, and S. Kubick, "High-yield production of "difficult-to-express" proteins in a continuous exchange cell-free system based on CHO cell lysates," *Scientific Reports 2017 7:1*, vol. 7, pp. 1–15, 9 2017.

[347] X. Jin, W. Kightlinger, and S. H. Hong, "Optimizing Cell-Free Protein Synthesis for Increased Yield and Activity of Colicins," *Methods and Protocols 2019, Vol. 2, Page 28*, vol. 2, p. 28, 4 2019.

[348] C. Rodríguez-Nava, C. Ortuño-Pineda, B. Illades-Aguiar, E. Flores-Alfaro, M. A. Leyva-Vázquez, I. Parra-Rojas, O. del Moral-Hernández, A. Vences-Velázquez, K. Cortés-Sarabia, and L. d. C. Alarcón-Romero, "Mechanisms of

Action and Limitations of Monoclonal Antibodies and Single Chain Fragment Variable (scFv) in the Treatment of Cancer," *Biomedicines 2023, Vol. 11, Page 1610*, vol. 11, p. 1610, 6 2023.

[349] D. Zahavi and L. Weiner, "Monoclonal Antibodies in Cancer Therapy," *Antibodies 2020, Vol. 9, Page 34*, vol. 9, p. 34, 7 2020.

[350] S. Jin, Y. Sun, X. Liang, X. Gu, J. Ning, Y. Xu, S. Chen, and L. Pan, "Emerging new therapeutic antibody derivatives for cancer treatment," *Signal Transduction and Targeted Therapy 2022 7:1*, vol. 7, pp. 1–28, 2 2022.

[351] A. Mullard, "FDA approves 100th monoclonal antibody product," *Nature Reviews Drug Discovery*, vol. 20, pp. 491–495, 7 2021.

[352] R. B. de Aguiar, T. d. A. da Silva, B. A. Costa, M. F. M. Machado, R. Y. Yamada, C. Braggion, K. R. Perez, M. A. S. Mori, V. Oliveira, and J. Z. de Moraes, "Generation and functional characterization of a single-chain variable fragment (scFv) of the anti-FGF2 3F12E7 monoclonal antibody," *Scientific Reports 2021 11:1*, vol. 11, pp. 1–11, 1 2021.

[353] W. Ren, Y. Xu, Z. Huang, Y. Li, Z. Tu, L. Zou, Q. He, J. Fu, S. Liu, and B. D. Hammock, "Single-chain variable fragment antibody-based immunochromatographic strip for rapid detection of fumonisin B1 in maize samples," *Food Chemistry*, vol. 319, p. 126546, 7 2020.

[354] W. Chen, Y. Yuan, and X. Jiang, "Antibody and antibody fragments for cancer immunotherapy," *Journal of Controlled Release*, vol. 328, pp. 395–406, 12 2020.

[355] M. Stech, M. Hust, C. Schulze, S. Dübel, and S. Kubick, "Cell-free eukaryotic systems for the production, engineering, and modification of scFv antibody fragments," *Engineering in Life Sciences*, vol. 14, pp. 387–398, 7 2014.

[356] S. K. Krebs, N. Rakotoarinoro, M. Stech, A. Zemella, and S. Kubick, "A CHO-Based Cell-Free Dual Fluorescence Reporter System for the Straightforward Assessment of Amber Suppression and scFv Functionality," *Frontiers in Bioengineering and Biotechnology*, vol. 10, p. 873906, 4 2022.

[357] J. B. Eaglesham, A. Garcia, and M. Berkmen, "Production of antibodies in SHuffle Escherichia coli strains," *Methods in Enzymology*, vol. 659, pp. 105–144, 1 2021.

[358] J. Yang, G. Kanter, A. Voloshin, N. Michel-Reydellet, H. Velkeen, R. Levy, and J. R. Swartz, "Rapid expression of vaccine proteins for B-cell lymphoma in a cell-free system," *Biotechnology and bioengineering*, vol. 89, pp. 503–511, 3 2005.

[359] A. Kunamneni, E. C. Clarke, C. Ye, S. B. Bradfute, and R. Durvasula, "Generation and Selection of a Panel of Pan-Filovirus Single-Chain Antibodies using Cell-Free Ribosome Display," *The American Journal of Tropical Medicine and Hygiene*, vol. 101, no. 1, p. 198, 2019.

[360] B. Höger, C. Peifer, and E. Beitz, "Cell-free production of fluorescent proteins for the discovery of novel ribosome-targeting antibiotics," *Journal of Microbiological Methods*, vol. 213, p. 106814, 10 2023.

[361] J. Shin and V. Noireaux, "Efficient cell-free expression with the endogenous E. Coli RNA polymerase and sigma factor 70," *Journal of Biological Engineering*, vol. 4, pp. 1–9, 6 2010.

[362] N. C. Shaner, R. E. Campbell, P. A. Steinbach, B. N. Giepmans, A. E. Palmer, and R. Y. Tsien, "Improved monomeric red, orange and yellow fluorescent proteins derived from Discosoma sp. red fluorescent protein," *Nature Biotechnology 2004 22:12*, vol. 22, pp. 1567–1572, 11 2004.

[363] D. Garenne, S. Thompson, A. Brisson, A. Khakimzhan, and V. Noireaux, "The all-E. coliTXTL toolbox 3.0: new capabilities of a cell-free synthetic biology platform," *Synthetic Biology*, vol. 6, 2 2021.

[364] J. Garamella, R. Marshall, M. Rustad, and V. Noireaux, "The All E. coli TX-TL Toolbox 2.0: A Platform for Cell-Free Synthetic Biology," *ACS Synthetic Biology*, vol. 5, pp. 344–355, 4 2016.

[365] X. Chen, J. L. Zaro, and W. C. Shen, "Fusion protein linkers: Property, design and functionality," *Advanced Drug Delivery Reviews*, vol. 65, pp. 1357–1369, 10 2013.

[366] A. B. Al-Hawash, X. Zhang, and F. Ma, "Strategies of codon optimization for high-level heterologous protein expression in microbial expression systems," *Gene Reports*, vol. 9, pp. 46–53, 12 2017.

[367] M. A. Kercher, P. Lu, and M. Lewis, "Lac repressor—operator complex," *Current Opinion in Structural Biology*, vol. 7, no. 1, pp. 76–85, 1997.

[368] F. Sanger, S. Nicklen, and A. R. Coulson, "DNA sequencing with chain-terminating inhibitors," *Proceedings of the National Academy of Sciences*, vol. 74, pp. 5463–5467, 12 1977.

[369] F. Sievers and D. G. Higgins, "Clustal Omega for making accurate alignments of many protein sequences," *Protein Science : A Publication of the Protein Society*, vol. 27, p. 135, 1 2018.

[370] R. Young and U. Bliisi, "Holins: form and function in bacteriophage lysis," *FEMS Microbiology Reviews*, vol. 17, pp. 191–205, 8 1995.

[371] A. Didovyk, T. Tonooka, L. Tsimring, and J. Hasty, "Rapid and Scalable Preparation of Bacterial Lysates for Cell-Free Gene Expression," *ACS Synthetic Biology*, vol. 6, pp. 2198–2208, 12 2017.

[372] T. Mahmood and P. C. Yang, "Western Blot: Technique, Theory, and Trouble Shooting," *North American Journal of Medical Sciences*, vol. 4, p. 429, 9 2012.

[373] T. C. Terwilliger, D. Liebschner, T. I. Croll, C. J. Williams, A. J. McCoy, B. K. Poon, P. V. Afonine, R. D. Oeffner, J. S. Richardson, R. J. Read, and P. D. Adams, "AlphaFold predictions are valuable hypotheses and accelerate but do not replace experimental structure determination," *Nature Methods 2023*, pp. 1–7, 11 2023.

[374] N. Zarkar, M. A. Nasiri Khalili, S. Khodadadi, M. Zeinoddini, and F. Ahmadpour, "Expression and purification of soluble and functional fusion protein DAB389IL-2 into the E. coli strain Rosetta-gami (DE3)," *Biotechnology and Applied Biochemistry*, vol. 67, pp. 206–212, 3 2020.

[375] S. D. Cole, K. Beabout, K. B. Turner, Z. K. Smith, V. L. Funk, S. V. Harbaugh, A. T. Liem, P. A. Roth, B. A. Geier, P. A. Emanuel, S. A. Walper, J. L. Chávez, and M. W. Lux, "Quantification of Interlaboratory Cell-Free Protein Synthesis Variability," *ACS Synthetic Biology*, vol. 8, pp. 2080–2091, 9 2019.

[376] U. K. LAEMMLI, "Cleavage of Structural Proteins during the Assembly of the Head of Bacteriophage T4," *Nature 1970 227:5259*, vol. 227, no. 5259, pp. 680–685, 1970.

[377] A. Doerr, D. Foschepoth, A. C. Forster, and C. Danelon, "In vitro synthesis of 32 translation - factor proteins from a single template reveals impaired ribosomal processivity," *Scientific Reports*, pp. 1–12, 2021.

[378] W. Wang and C. J. Roberts, "Protein aggregation – Mechanisms, detection, and control," *International Journal of Pharmaceutics*, vol. 550, pp. 251–268, 10 2018.

[379] J. A. Schoborg, C. E. Hodgman, M. J. Anderson, and M. C. Jewett, "Substrate replenishment and byproduct removal improve yeast cell-free protein synthesis," *Biotechnology Journal*, vol. 9, pp. 630–640, 5 2014.