Exploring protein folding mechanisms to enable the protein structure prediction of previously intractable targets

Clare West

Lincoln College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Michaelmas 2019

Acknowledgements

First of all, I have Charlotte Deane to thank for giving me this opportunity, and for all the amazing things that have happened in the four years since. Charlotte is an exceptional supervisor — both a brilliant scientist and an endlessly generous and dedicated mentor. What I've learnt from Charlotte would fill several more theses, and I'm extremely proud to have been her student.

I am also grateful to Saulo de Oliveira, whose patience and enthusiasm has always been abundant even across oceans and time zones. Thank you to all my other mentors, official and unofficial: Sebastian Kelm, Jiye Shi, Helen Byrne, Eleanor Law, Jinwoo Leem, Ruth Mitchell, Mischka Byworth, Lydia Harriss, and Brian Nicholls-Lee.

This research — and two formative trips to San Francisco and Basel — were supported by the Engineering and Physical Sciences Research Council (EPSRC), Medical Research Council (MRC), UCB Pharma, the International Society for Computational Biology (ISCB) and Lincoln College.

I am incredibly grateful for the friendships I have made along the way that have made me a better scientist and a better person. Dan Wilson was my first collaborator and made me feel like there was a place for me in this scary interdisciplinary space; thanks for educating me a little about maths and a lot about life. I owe a particular debt of gratitude to Lincoln College for introducing me to my incomparable friend Jen Laws (and the gardens!). I was very lucky to share this journey with Laura Depner, who reassured me that she was confused as I was (I'm proud to say our 3D-printed dimer is now immortalised in these pages). I am also grateful to Eleanor Law and Dominik Schwarz for making the office so much fun, even on days when it was the last place I wanted to be. It hasn't always been easy; my friendship with Isabel Wilkinson has empowered and revived me many times. And I wouldn't have made it over the finish line without Susan Leung.

Thank you to all the others who have shared the highs and lows (and more than a few dinners and drinks) with me on this journey, in particular: Jared Field, Joe Bluck, Darren Valentine, Alex Skates, Lyuba Bozhilova, Qinrui Wang and Hannah Patel.

Thank you to all the members of the ever-expanding Oxford Protein Informatics Group (OPIG) (see Appendix D) for countless coffee breaks, cryptic crosswords, pub trips, explorations in Basel and invaluable discussions, scientific or otherwise. I am also grateful to my colleagues in the department, especially Beverley Lane

and Susan Hutchinson, not only for their essential IT and administrative support, but for giving me a sense of belonging in the Department of Statistics, which once seemed so unlikely.

The past four years would have been much more difficult without my real-world friends — Helena Cerski, Kat Yearley and Hannah Minns in particular — who always had unwavering faith in me, and reminded me that (in the figurative sense, at least) there is more to life than proteins.

Thank you to my family — Julie, Martin, Emma, Ritchie and Dan — for always welcoming me home whenever I needed it, and for bringing home to Oxford when I couldn't make it to Surrey. A special thank you to Emma for bringing positivity, for keeping me warm with knitted gifts, and for housing me during my internship. This thesis would not have been possible without my Mum and Dad, who instilled in me the knowledge that I can do anything, and reminded me of this whenever my conviction began to waver.

Finally, thank you to my co-author in life, Leo Speidel — aside from this thesis, the greatest thing to come out of my time at Oxford. Leo has improved every aspect of my life, thesis included. Thank you for being there for me always and for making my world so much bigger. DPhil 頑張りましたね。

Abstract

The prediction of a protein structure from its amino acid sequence is one of the grand challenges of computational biology. In this thesis, we describe extensions to our biologically-inspired fragment-based protein structure prediction pipeline, SAINT2, to improve the prediction of previously intractable targets.

While template-free protein structure prediction protocols can now produce good quality models for many targets, modelling failure remains common; users need to be able to identify if a good model exists among the many models produced for a given target. We describe Random Forest Quality Assessment (RFQAmodel), which assesses whether models produced by a protein structure prediction pipeline have the correct fold. By iteratively generating models and running RFQAmodel until a model is produced that is predicted to be correct with high confidence, we demonstrate how such a protocol can be used to focus computational efforts on difficult modelling targets.

It is common for a target of interest to have an incomplete structure or a partial homologous template. We describe Flib-Flex, a method for incorporating known structural information into the fragment library that is used during prediction, and its combination with SAINT2-ScafFold to complete partial protein structures. We demonstrate that the missing regions can be modelled accurately in the presence of the known structure, and correct models can be identified using a modified version of RFQAmodel.

Long proteins remain a challenge for template-free protein structure prediction. There is experimental evidence that proteins can adopt the native conformation via the sequential folding of small units known as foldons. Inspired by this folding pathway hypothesis, we have divided long protein structures into smaller, semi-stable segments, and predict these individually in succession. This protocol, ScafFoldOn, enables the prediction of previously intractable targets.

Declaration

I declare that no parts of this thesis or its research herein have been reproduced or accepted for another award or degree or diploma at any other university or learning institution. This thesis contains no other person's work except where stated in the text.

Clare West

20th December 2019

Contents

List of Figures

List of Tables

1 Introduction

Contents

Proteins perform a myriad of functions that are essential to life. Their sequences are encoded by genes and have been honed over millions of years of evolution. The varied roles of proteins include forming cellular structures, transmitting signals, transportation of small molecules or macromolecules, and catalysis of chemical

reactions. These functions are mediated by the dynamic three-dimensional structures that proteins are able to adopt. An understanding of these structures can elucidate functional understanding (Lee et al., 2007), explain disease mechanisms (J. Wang et al., 2016) and inform drug design (Nero et al., 2018).

Template-free protein structure prediction methods attempt to predict the structure of target proteins from the sequence. Most methods are founded on the principle that all the information required for protein folding is contained within the sequence. However, proteins fold more efficiently *in vivo* compared to refolding *in vitro*, and the extent of the influence of the cellular environment on protein folding is becoming increasingly clear (Clark, 2004; Komar, 2009; Waudby et al., 2019). Biologically-inspired approaches to computational protein structure prediction have already resulted in improvements in accuracy and efficiency (de Oliveira, Law, et al., 2018; Ellis et al., 2010). In this thesis, we explore whether we can apply such approaches for the prediction of challenging protein structures.

In this chapter, we give an overview of protein structure and its prediction, as well as how this relates to what is currently understood about how proteins fold in the cell. Finally, we outline the work presented in this thesis, describing extensions to the SAINT2 structure prediction pipeline to improve the prediction of previously intractable targets.

1.1 Protein structure

Protein structure is commonly considered at four levels: primary, secondary, tertiary and quaternary. These levels are illustrated in Figure 1.1.

1.1.1 Primary structure

With rare exceptions, proteins consist of the 20 canonical amino acids. The general structure of an amino acid is a central carbon atom (denoted C_α), a hydrogen

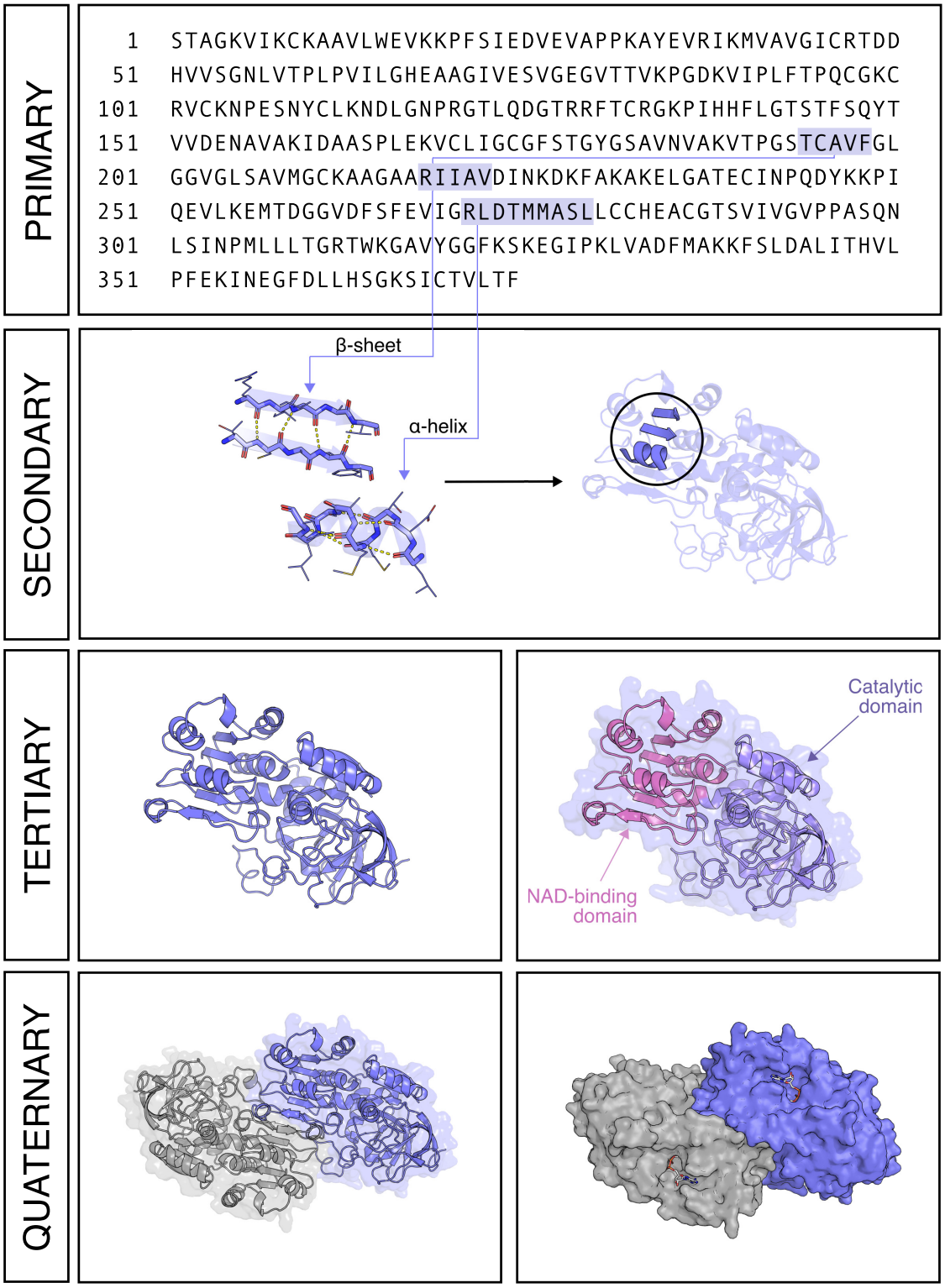

Figure 1.1: The four levels of protein structure, demonstrated with human alcohol dehydrogenase (PDB code 1HTB) as an example case. The primary structure of a protein is its sequence of amino acids. The secondary structure refers to local structure, mediated by backbone hydrogen bonding. The tertiary structure is the three-dimensional arrangement of the whole protein chain; this is often divided into domains. Quaternary structure is the arrangement of multiple chains. The surface representation demonstrates the overall shape of the alcohol dehydrogenase dimer; NADH is shown in the binding pocket, in which the reaction with substrate ethanol is catalysed.

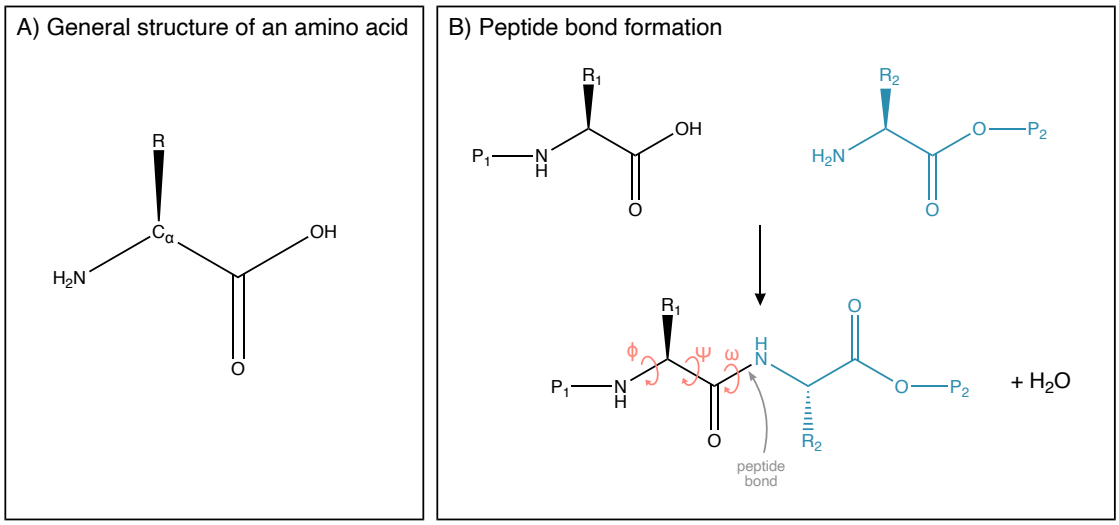

Figure 1.2: General structure of an amino acid and peptide bond formation. **A)** Amino acids consist of a central carbon atom (C_α), carboxyl (COOH) and amine (NH_2) functional groups, and a characteristic side chain (R), see Figure 1.3. **B)** Amino acids polymerise via a condensation reaction between the carboxyl group of one amino acid (black) and the amide group of another (blue), forming a peptide bond and a water molecule. The three backbone torsion angles are shown in red.

atom, carboxyl and amine functional groups, and a characteristic sidechain (R-group) (Figure 1.2A). Virtually all amino acids in proteins exist in the L arrangement around the chiral C_α atom. Amino acids polymerise via peptide bonds that are formed between the carboxyl and amine functional groups (Figure 1.2B).

Protein structure and function are mediated by the physiochemical properties of these amino acids, which can be broadly grouped into categories: hydrophobic, hydrophilic, and those that are charged at physiological pH (Figure 1.3). Cysteine, glycine and proline are of particular structural interest. Pairs of cysteine residues are capable of forming covalent disulphide bonds — known as disulphide bridges — under oxidising conditions, which typically occurs in the endoplasmic reticulum, periplasm or extracellular space (Sevier et al., 2002). Glycine has the smallest side chain, consisting only of a hydrogen atom. Proline has a unique cyclic sidechain with two chemical bonds to the backbone; as a result, proline is unable to act as a hydrogen bond donor, and has a tightly constrained conformation that is often found in tight turns and introduces kinks into helical structures.

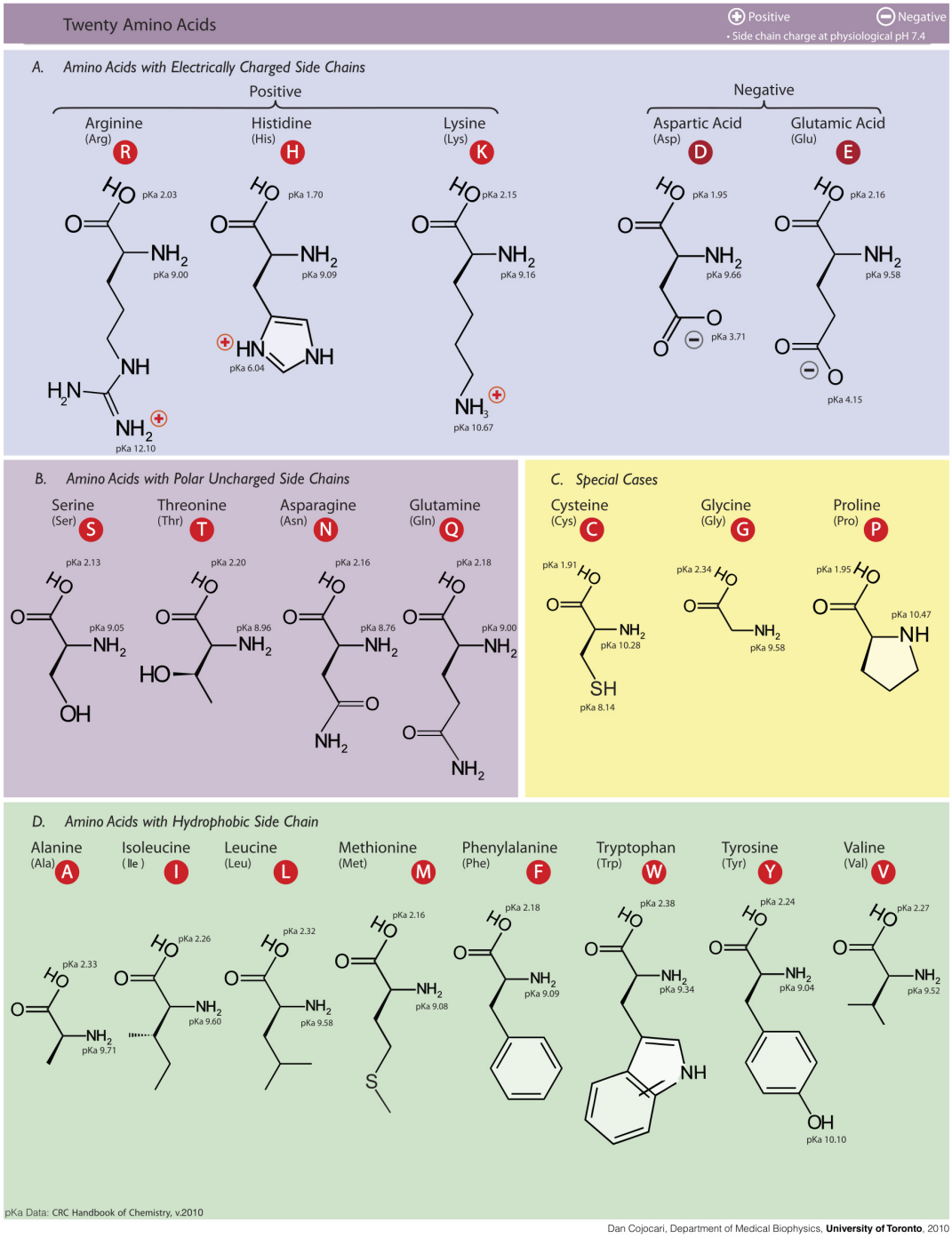

Figure 1.3: The molecular structures and nomenclature of the 20 canonical amino acids, grouped according to their properties. Figure adapted from original by Dan Cojocari under Creative Commons license (CC BY-SA 3.0; `https://creativecommons.org/licenses/by-sa/3.0/`), via Wikipedia Commons.

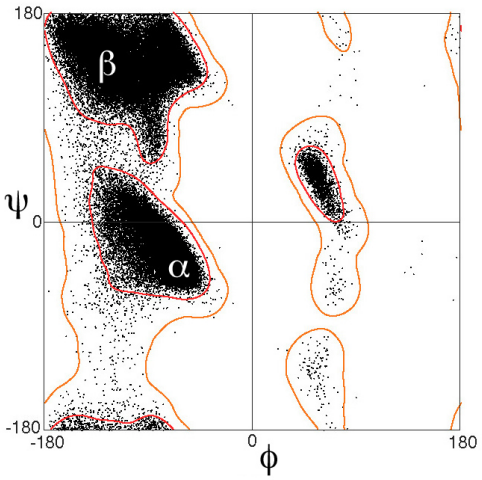

Figure 1.4: Ramachandran plot for the general case (excluding glycine, proline and residues preceding proline). ϕ and ψ angles for around 100,000 residues from high-resolution crystal structures, with data from Lovell et al., 2003. Regions corresponding to α-helical and β-sheet secondary structures are indicated. Figure reproduced from Dcrjsr under Creative Commons license (CC BY-SA 3.0; `https://creativecommons.org/licenses/by-sa/3.0/`), via Wikipedia Commons.

1.1.2 Secondary structure

Each residue in a chain of amino acids has three torsion angles (also known as dihedral angles) around the three backbone bonds: ϕ, ψ and ω (Figure 1.2B). The peptide bond is held in a planar shape by the π-bond character that restricts rotation around the peptide bond. This means that the ω angle can vary little from 180° in the more common *trans* isomer, or 0° in the less favourable *cis* isomer; the latter is rare for all amino acids with the exception of proline. The majority of the conformation of the backbone can therefore be described by the two rotatable bonds, ϕ and ψ. Some conformations occur more frequently than others, with many values being disfavoured due to steric clashes; the variation of these angles is often visualised using a Ramachandran plot (Ramachandran et al., 1963) (Figure 1.4).

Secondary structure refers to the local arrangement of protein chains that is mediated by hydrogen bonds formed between the polar backbone groups. The most highly-represented regions of a Ramachandran plot correspond to α-helices and β-sheets, which are the most common repeated secondary structure elements. α-helices are coiled structures that are stabilised by regular hydrogen bonds between

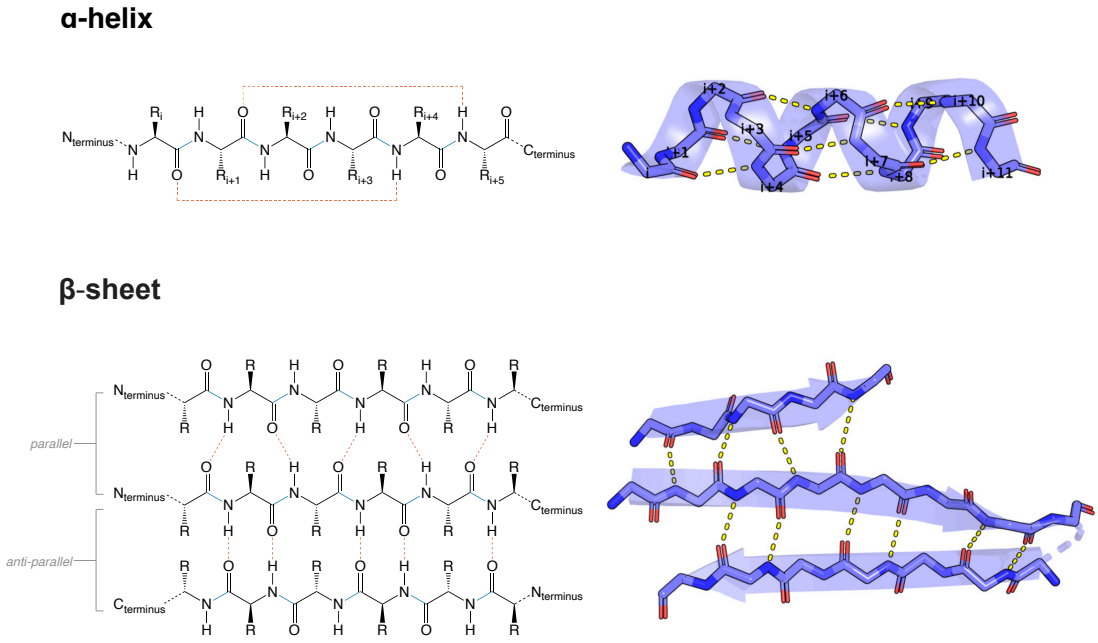

Figure 1.5: Structure and hydrogen bonding schemes of an α-helix and β-sheet. Left) General molecular structures, with hydrogen bonds indicated by orange dashed lines. Right) 3D example showing only the backbone atoms, overlaying a cartoon representation, with hydrogen bonds indicated by yellow dashed lines (PDB code 1I3A). N and O atoms are coloured in dark blue and red, respectively. C_α atoms are labelled with the relative residue number for the α-helix.

the C=O group of residue i with the N-H group of residue $i+4$ (Figure 1.5). More rarely, helical structures are formed by hydrogen bonds between residues i and $i+3$, or i and $i+5$, which are referred to as 3_{10}-helices and π-helices, respectively. The regular pattern of an α-helix has characteristic dimensions: each residue is separated by 1.5Å and a rotation of 100°, with 3.6 residues and 5.4Å per full turn of the helix.

β-sheets are formed by hydrogen bonds between two or more adjacent polypetide chain regions known as β-strands. The β-strands are almost fully extended, with approximately 3.5Å between consecutive residues. The strands in a β-sheet can be oriented in the same direction (parallel), the opposite direction (anti-parallel), or a mixture.

Each amino acid residue has a different propensity for adopting α-helices or β-sheets as a result of the compatibility of the sidechains with the geometry of the backbone conformations. In a protein structure, residues that do not adopt a secondary structure are commonly described as loops or coil.

Standard definitions of these secondary structure elements enable the automated annotation of experimentally derived structures. One of the most widely used of these is DSSP, which identifies characteristic hydrogen bonding patterns inferred from estimated electrostatic interaction energies (Kabsch et al., 1983).

1.1.3 Tertiary structure

The overall three-dimensional arrangement of the protein chain is known as the tertiary structure (Figure 1.1). Soluble proteins typically adopt compact structures in which hydrophobic sidechains are buried and hydrophilic sidechains are exposed. Proteins that interact with or are embedded in cell membranes — which are lipid environments — are instead arranged so that hydrophobic sidechains come into contact with the membrane.

Protein chains are often divided into distinct modular structures known as domains. There are two ways to define a domain. The first is an evolutionarily conserved unit of sequence that is found in multiple protein contexts (Sonnhammer et al., 1997). The second is an autonomously folding unit of protein structure (Holland et al., 2006). Although in practice the definitions often overlap, autonomous folding is more relevant to this thesis.

Domains are often genetically conserved and associated with particular functions, and sequences corresponding to entire domains are frequently duplicated and recombined within genomes (Chothia et al., 2009). The annotation and classification of protein structure typically occurs at a domain level: domains are grouped by structural similarity and evolutionary relationships and organised using hierarchical clustering.

The assignment of a domain is typically based on structural compactness, as well as sequence or structural similarity to known domains. Recently, semi- or fully-automated methods have been established to cope with the increasing number of solved protein structures (Fox et al., 2014; Sillitoe et al., 2015). This includes the methods employed by two comprehensive and widely-used resources: SCOP-extended (SCOPe) (Fox et al., 2014) and Class, Architecture, Topology, Homology

(CATH) (Sillitoe et al., 2015). Domain boundary assignment in SCOPe uses sequence similarity to previously classified domains (Chandonia et al., 2017), initially based on the manually-curated predecessor database, SCOP (Murzin et al., 1995). CATH uses a combination of methods that assess sequence identity and structure similarity to previously classified domains, as well as structural compactness (Greene et al., 2007; Sillitoe et al., 2015). Both methods employ manual inspection for ambiguous cases. There is a large proportion of agreement between these two databases, but subjectivity of domain classification is to be expected, and variation in methods and hierarchical structure sometimes result in inconsistencies (Holland et al., 2006). For example, SCOP tends to partition protein chains into fewer, larger domains than CATH (Csaba et al., 2009). Nevertheless, the concept of domains has been crucial for the understanding of protein function, structure, and evolution (Fox et al., 2015).

1.1.4 Quaternary structure

While some proteins are monomers, consisting of a single polypeptide chain, many are oligomeric assemblies consisting of more than one chain. The quaternary structure describes the arrangement of the constituent subunits; this may consist of multiple identical subunits (homomeric) or different chains heteromeric).

1.1.5 Experimental determination of protein structure

There are 146,831 experimentally determined protein structures deposited in the Protein Data Bank (PDB) (Berman, 2000) as of December 2019 (Figure 1.6). When clustered by 100% sequence identity, 88,163 unique protein sequences are represented, or 34,443 at 30% sequence identity.

The majority of these were determined using X-ray crystallography, the method that was used to solve the first protein structure, myoglobin (Kendrew et al., 1958). A purified protein in a crystallised form is irradiated with X-rays and the deflected X-rays are observed as a diffraction pattern, from which an electron density can

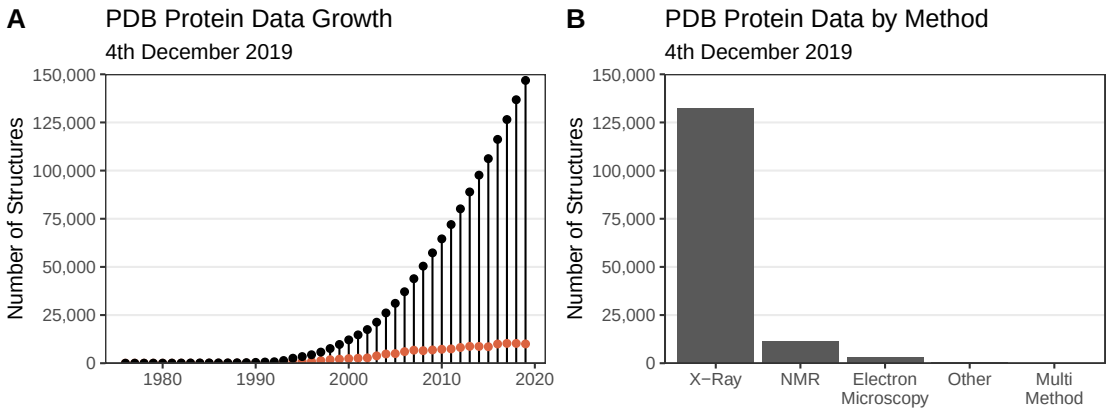

Figure 1.6: Number of protein structures in the PDB. A) The total number of structures by year (black) and the number of new structures added each year (orange). B) The number of structures by experimental method. Data from PDB statistics (`www.rcsb.org/stats/growth/overall`)

be reconstructed given phase information. This electron density is then used to model the underlying protein structure.

The quality of an X-ray crystallography model is reflected in two measures: the resolution and the R-factor or R-free (Brünger, 1997). The positions of sidechains are visible at around 3.5Å resolution, and atomic resolution, at which individual atoms can be distinguished, corresponds to around 1.2Å (Wlodawer et al., 2017). The R-factor describes how closely a simulated diffraction pattern based on the model fits the experimentally-observed data. R-free is a related measure that aims to reduce bias of model refinement to improve the R-factor: some of the experimental data is held out from model refinement, and R-free is calculated based on how well the resulting model predicts this remaining fraction (Brünger, 1992). Technical innovations have increased the resolution, throughput, and range of structures accessible by X-ray crystallography methods, but the crystallisation process is a major bottleneck (Holcomb et al., 2017).

The second most prevalent method for solving protein structures, nuclear magnetic resonance (NMR) spectroscopy, does not require crystallisation. This method exploits the magnetic spin inherent to the atomic nuclei of specific isotopes that can be introduced to protein structures. When the protein is irradiated with a radio frequency electromagnetic pulse under a magnetic field, these nuclei

resonate at characteristic frequencies, which reveal information about their chemical environment. This information is used to determine the relative location of atoms close together in space, and from these spatial constraints a model of the three-dimensional structure can be fitted. NMR structures are calculated as an ensemble of structures. Unlike X-ray crystallography, NMR is carried out in solution and can capture dynamic information such as conformational changes and disorder. However, it is generally limited to smaller proteins (<25kDa for routine NMR, Sugiki et al., 2017).

Electron microscopy (EM) is a method of structural determination suitable for larger macromolecules, including large proteins and complexes. A beam of electrons is fired at the protein and the electrons that penetrate the sample are detected, producing a two-dimensional image. Many thousands of images captured at different orientations can then be reconstructed into a three-dimensional structure. Historically, such methods have been limited to low-resolution models. However, recent advances in cryogenic EM (cryo-EM) and electron detection technology have produced structures with estimated resolutions comparable to X-ray crystallography, although measures of model quality are not as well-established (Murata et al., 2018).

Structural determination of proteins in a crystal or in solution does not necessarily reflect the conformation that may exist *in vivo*. Structures can be observed in a physiological environment using in-cell NMR (Serber et al., 2001; Sakakibara et al., 2009; Pan et al., 2016; Ikeya, Hanashima, et al., 2016), but solving detailed structures without prior knowledge remains difficult and is not yet routine (Ikeya, Güntert, et al., 2019).

In this thesis, experimentally-determined models of protein structures (usually derived using X-ray crystallography) are referred to as the native structure, as these models represent the current best observable "ground truth" for the conformation adopted by a target protein.

1.1.6 Protein formation in the biological system

In living cells, proteins are synthesised sequentially from the N-terminus (amide) to the C-terminus (carboxyl) by the process of translation. The protein sequence is encoded by messenger RNA (mRNA) transcripts, which are processed by the ribosome, a large molecular machine formed of proteins and RNA. As the ribosome moves along the transcript, transfer RNA (tRNA) molecules bring the next amino acid to the site by recognising cognate codon triplets in the mRNA code. The ribosome then catalyses the addition of the amino acid to the C-terminus of the nascent chain.

1.2 Protein folding

The process by which proteins adopt their native state is central to biological function, and underlies many pathologies caused by protein misfolding and aggregation (Chiti et al., 2017). Understanding of the protein folding problem has made great progress in the 50 years since Levinthal's observation that, considering the speed of protein folding relative to the vast number of possible conformations to be explored, protein folding cannot occur by random search of all possible conformations (Levinthal, 1968). The energetics of the conformational space to be explored can be visualised as a free energy landscape (Bryngelson et al., 1995; Dinner et al., 2000). To explain how proteins are able to fold on experimentally-observed timescales, it was proposed that the shape of this landscape reduces the conformational space explored by the protein by guiding towards the native structure (Dill et al., 1997). Rather than a flat landscape in which all non-native conformations are equally favourable, proteins are thought to fold into the thermodynamic global energy minimum via favourable interactions that bias towards this native state (Dill et al., 1997). The trajectory through the landscape may involve a defined pathway via reproducible folding intermediates (Dill et al., 1997), or a more

diffusive exploration of a rugged landscape with many possible pathways (Harrison et al., 1985; Onuchic et al., 1997; Mallamace et al., 2016); the extent to which each model is correct or dominant remains the subject of debate (Dinner et al., 2000; Englander et al., 2017; Baldwin, 2017; Eaton et al., 2017).

Much of the evidence relating to protein folding is derived from experiments and simulations on relatively short, full-length proteins refolding reversibly from a denatured state *in vitro* (Braselmann et al., 2013). However, for many proteins this kind of *in vitro* refolding is inefficient or impossible (Eiberle et al., 2010; Cabrita and Bottomley, 2004). A group of proteins known as chaperones are known to promote the successful folding of newly synthesised proteins by stabilising unfolded regions and preventing aggregation (Hartl, 2002); however, as less than 20% of cytoplasmic proteins require chaperones to fold, the absence of chaperones alone does not explain this discrepancy in folding efficiency (Hartl, 2002). It is therefore also important to consider the features of cellular folding pathways.

1.2.1 Foldons

One type of intermediate suggested to be important for protein folding are small semi-stable subunits known as "foldons" (Englander et al., 2017). Foldons are thought to be small enough to feasibly fold via a random search, with a sufficiently large energy benefit to overcome the entropy cost of folding and drive a bias towards the native conformation (Panchenko et al., 1996).

A number of proteins have been shown experimentally to refold into the native conformation via the cooperative folding of these constituent foldons (Y. Xu et al., 1998; Englander et al., 2014; Hu et al., 2016). Refolding pathways of proteins can be probed by hydrogen/deuterium exchange (HDX) experiments, which identify regions that quickly become solvent-inaccessible, and regions that remain unfolded and therefore undergo exchange for longer (Bai et al., 1995). Start2Fold is a database of curated HDX experimental data, in which 57 protein entries are annotated with early, intermediate and late folding residues as of November 2019 (bio2byte.be/start2fold/) (Pancsa et al., 2016).

1.2.2 Cotranslational folding

It is well-established that proteins can begin the folding process as soon as the N-terminus emerges from the ribosome tunnel (Kolb, 2001; Komar, 2009; Waudby et al., 2019). Proteins fold faster than they are synthesised and, except where mechanisms exist to actively hold the nascent peptide in an unfolded state, peptides will spontaneously adopt energetically-favourable conformations (Fedorov et al., 1997). Cotranslational folding is therefore likely to be the physiological pathway for many proteins, particularly for large and multidomain proteins (Ciryam et al., 2013).

The energy landscape of a cotranslationally folding protein differs from that of a fully extruded protein that is refolding (Waudby et al., 2019). The conformational space available to the nascent protein increases with each additional residue in the growing chain; the folding landscape can therefore be conceptualised as a nested series of length-dependent landscapes (Figure 1.7) (Clark, 2004).

Computational evidence for cotranslational folding

If a protein folds as it emerges from the ribosome, the N-terminus may tend to be more compact and buried compared to the C-terminus. Computational analysis revealed a bias towards these hypothetical structural features among known structures, particularly for domains with alternating α and β secondary structures (Deane et al., 2007; Saunders, Mann, et al., 2011). These features were also replicated in simulations of cotranslational folding based on the simplified hydrophobic-polar model of protein folding (Saunders, Mann, et al., 2011). An analysis of tertiary structure prediction methods found that predictions were more accurate at the N-terminal region than the C-terminal region; one possible explanation is that the N-terminus may be closer to the global minimum as a result of cotranslational folding (Saunders and Deane, 2010a).

Coarse-grained molecular dynamics simulations, which are able to recapitulate experimental folding pathways (Nissley and O'Brien, 2018), have also been used to explore cotranslational folding pathways (Nissley, Sharma, et al., 2016; O'Brien et al., 2010). In simulations, the nascent chain can form stable, native-like intermediates;

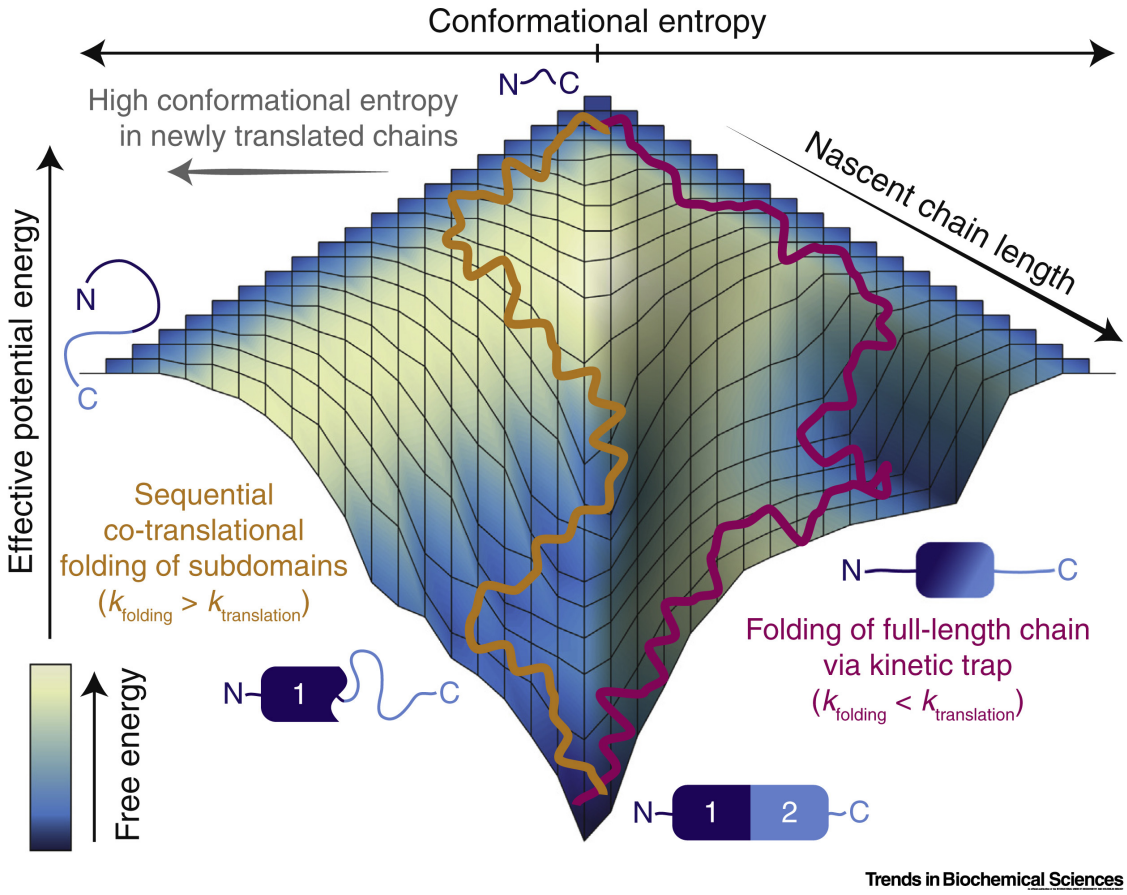

Figure 1.7: Schematic illustration of the free energy landscape of a hypothetical protein with two subdomains, demonstrating two possible cotranslational folding trajectories. The shading of the landscape indicates the total free energy at each point. This represents the competition between favourable conformational energy (width of the surface) or effective potential energy (depth of the surface). On the right, a magenta trajectory represents slow folding relative to translation: folding proceeds via a kinetically trapped, misfolded intermediate corresponding to a free energy minimum that emerges at a longer chain length. This is represented by interactions between the unfolded N-terminal subdomain and partially translated C-terminal subdomain. On the left, an orange trajectory represents rapid folding relative to translation: the N-terminal subdomain folds before translation of the C-terminus, as a result of an intermediate free energy minimum. Figure reproduced from Waudby et al. (2019) under Creative Commons CC-BY license.

these intermediates can correspond to complete or partial domains, and may differ from those observed in refolding simulations, suggesting a distinct cotranslational folding pathway (Jacobs et al., 2017).

Experimental evidence for cotranslational folding

A subset of soluble proteins are known to be able to correctly refold from a denatured state *in vitro* (Anfinsen, 1973), but this is often difficult, inefficient or impossible (Eiberle et al., 2010; Cabrita and Bottomley, 2004). A number of proteins that are slow or aggregation-prone when folding *in vitro* follow a cotranslational folding pathway *in vivo* that guides towards the native structure and avoids misfolding (Evans et al., 2008; Frydman et al., 1999; Fedorov et al., 1999; Ugrinov et al., 2010).

The conformations of partially extruded proteins on the ribosome can be explored by arresting translation at different nascent chain lengths and probing the resulting structures. Such experiments have revealed that some partially extruded proteins are capable of conformation-dependent interactions (Komar et al., 1997), characteristic protease resistance (G. Zhang et al., 2009), or enzymatic activity (Frydman et al., 1999); this demonstrates that native-like conformations can be adopted during translation.

Cotranslational protein folding has also been observed in real time; experiments use methods such as fluorescently-labelled residues (Holtkamp et al., 2015) or single-molecule force spectroscopy (Kaiser, Goldman, et al., 2011) to explore the compact structures formed by the extruding nascent chain.

Instances of non-native intermediates have also been seen (Holtkamp et al., 2015); in some cases these differ from the intermediates seen in the *in vitro* refolding pathway of the same protein, and represent more efficient pathways (Evans et al., 2008; Nilsson et al., 2015).

Features of the cotranslational folding landscape

In addition to directional elongation, a number of other features of cotranslational folding are thought to modulate the conformational search space (Waudby et al., 2019; Mahlab et al., 2014; Komar, 2009).

Non-uniform translation speeds

The shape of the cotranslational folding landscape is related to the relative rate of translation compared to folding (the diffusive search for a low energy state) (Figure 1.7). If translation is rapid relative to folding, the chain will remain disordered at longer lengths, increasing the likelihood of kinetic traps such as transient or aggregation-prone misfolded intermediates (Waudby et al., 2019). Conversely, rapid folding relative to translation results in sequential folding from the N- to the C-terminus (Waudby et al., 2019).

Non-uniform translation rates are thought to be one mechanism by which cotranslational folding pathways are mediated (Jacobson et al., 2016). Varying expression levels of multiple cognate tRNA molecules for a given amino acid, and therefore availability during translation, can influence the speed of translation. Rare, or non-optimal, codons can introduce pauses during translation (Mohammad et al., 2019; C. C. C. Wu et al., 2019).

Site-specific variations in translation rate have been observed using single-molecule measurements of ribosome activity (Uemura et al., 2010), as well as ribosome profiling, in which translation rates are inferred by systematically identifying the position of ribosomes on many mRNA molecules (Mohammad et al., 2019; Ingolia et al., 2019). It has been suggested that translational pauses are conserved, under evolutionary selection, to facilitate the folding of preceding domains or cotranslational folding intermediates (Pechmann et al., 2013; G. Zhang et al., 2009; Sander et al., 2014; Jacobs et al., 2017; Saunders and Deane, 2010b).

There is experimental evidence that non-uniform translation rates, mediated by rare codons, are important for protein folding. Translational pauses can be disrupted by altering tRNA concentrations or by inducing synonymous substitutions;

this has been found to perturb the adoption of native structure and function for some proteins (G. Zhang et al., 2009). The routine replacement of rare codons in cDNA sequences — with the aim of optimising translation speed in recombinant *Escherchia coli* expression systems — has been found to significantly increase the proportion of the yield that is insoluble; selective reintroduction of non-optimal codons improved the recovery of soluble protein (Konczal et al., 2019).

Interactions with the ribosome

Interactions between the nascent chain and the ribosome are increasingly understood to modulate the cotranslational folding pathway (Kaiser and Liu, 2018; Waudby et al., 2019). The ribosome tunnel can accommodate around 30 extended residues between the site of translation and the tunnel exit (Voss et al., 2006). While folding has primarily been seen to occur outside the ribosome tunnel, α-helices and very small domains have been observed to form within the tunnel (Nilsson et al., 2015; Marino et al., 2016; Kramer et al., 2009).

Computational simulations and experimental probing of the behaviour of the nascent chain on the ribosome suggest that interactions with the ribosome influence protein folding by stabilising the unfolded state, although the underlying processes are not yet fully understood (Hsu et al., 2007; Kaiser and Liu, 2018; Kaiser, Goldman, et al., 2011; Samelson et al., 2016; Cassaignau et al., 2016; Liu et al., 2017; Cabrita, Cassaignau, et al., 2016). Furthermore, experiments using single-molecule optical tweezers have demonstrated that the force exerted by the nascent chain folding at the tunnel exit can rescue translational stalling (Goldman et al., 2015).

The vast amount of theoretical, computational and experimental evidence described here demonstrates how the features of the cellular environment can have a significant influence on the protein folding pathway.

1.2.3 Protein folding and protein structure prediction

The majority of template-free protein structure prediction software is founded on the principle that all the information required for a protein to adopt its native structure

is contained within the primary sequence. However, given the high efficiency of protein folding in living cells compared to refolding *in vitro*, it is important to consider how these proteins fold when attempting to predict their structure.

Compared to conventional search strategies, a sequential approach generates predictions more efficiently and accurately (Ellis et al., 2010). Our prediction software, SAINT2, emulates the cotranslational folding pathway by predicting protein structures from the N- to the C-terminal (de Oliveira, Law, et al., 2018). This sequential approach results in more accurate models with shorter computation times compared to a non-sequential implementation of the same method sampling the full-length target (de Oliveira, Law, et al., 2018). In this thesis we explore how further understanding and inclusion of efficient *in vivo* search mechanisms may allow the computational prediction of previously intractable targets.

1.3 Protein structure prediction

Given the high cost of structure determination compared to sequencing, the prediction of protein structure from sequence has been a major goal of computational biology. Following decades of computational and methodological advances, as well as an ever increasing amount of data, much information can now be learnt about protein structure from sequence alone (Kryshtafovych, Schwede, et al., 2019).

1.3.1 Secondary structure

One of the most successful of these efforts is the prediction of secondary structure. Until recently, most methods focussed on three-state prediction: categorising residues as helical, sheet, or coil. The output is commonly an estimated probability of each state for every residue in the sequence. The theoretical limit of secondary structure prediction is suggested to be around 88-90% due to the inconsistent and sometimes arbitrary distinction between the three states, although this applies mostly to the

boundaries of secondary structure elements (Rost et al., 1994; Rost, 2001; Y. Yang et al., 2018). The accuracy of the top-performing methods now approaches this theoretical limit (Heffernan et al., 2017).

Early techniques were based on the statistical propensities of amino acid residues towards each secondary structure element. As the number of available protein structures and sequences increased, it became possible to incorporate additional information first from neighbouring residues using a sliding window of 10-30 residues, and later from sequence evolution profiles derived from multiple sequence alignments. The highest reported accuracies are achieved using neural networks trained on large datasets (Y. Yang et al., 2018; Heffernan et al., 2017). One limitation of the sliding window approach is that it only captures local information; this limits performance on medium to large proteins with many non-local contacts (Kihara, 2005), even when considering deep neural network methods (Y. Yang et al., 2018). A recent method, SPIDER3, replaced the window technique with a Long Short-Term Memory Bidirectional Recurrent Neural Network, which overcame this limitation and achieved a reported accuracy of up to 84% (Heffernan et al., 2017).

Several patterns in secondary structure prediction errors have been identified (Y. Yang et al., 2018). Misclassification is more common in boundary regions than internal regions; sheet residues are predicted less accurately than helical residues; and helices and sheets that are more solvent accessible are more likely to be misclassified as coil residues (Y. Yang et al., 2018).

A number of recent methods attempt to predict all eight states classified by the most widely used assignment method, DSSP (Kabsch et al., 1983). These states are: α-helix, 3_{10}-helix, π-helix, isolated β-bridge, extended β-sheet, helix turn (a hydrogen-bonded turn), bend (presenting high curvature), or coil (no secondary structure) (Joosten et al., 2011). The highest-performing methods predicting all eight states now achieve accuracies of over 70% of residues correctly classified (S. Wang et al., 2016; Y. Yang et al., 2018). Such methods capture a broader range of backbone conformation, although the issue of inconsistent and arbitrary distinction remains.

1.3.2 Torsion angles

Torsion angles define backbone conformations continuously and precisely. In addition to local steric restrictions imposed on backbone conformations, torsion angles are influenced by the side chain of each residue and those of its neighbours in sequence (Betancourt et al., 2004). Prediction of the ϕ and ψ values from sequence alone was first attempted in 2008 with the method Real-SPINE (Xue et al., 2008). Additional prediction of the Cα_{i-1}–Cα_i–Cα_{i+1} (θ) angle and the dihedral angle rotation about the Cα_i–C–N–Cα_{i+1} bond (τ), which capture the local structure of neighbouring residues, was later introduced by SPIDER (Lyons et al., 2014).

Most torsion angle prediction methods employ supervised machine learning methods trained on sets of known protein structures. Inputs are typically sequence profiles derived from multiple sequence alignments, as well as secondary structure and, later, solvent accessibility predictions (S. Wu et al., 2008). A more recent version of Real-SPINE, SPINE-X, adopted an iterative process that included predicted ϕ and ψ values as input, enabling correlations between the two values to be learnt (Faraggi, Y. Yang, et al., 2009; Faraggi, T. Zhang, et al., 2012). The current state-of-the-art methods employ deep learning algorithms. SPIDER2 achieves the highest reported accuracy of 19.2° and 29.9° mean absolute error (MAE) for ϕ and ψ predictions, respectively (Bai et al., 1995).

In protein structures, the ϕ angle adopts a narrower distribution of values than ψ due to the larger steric restriction of the oxygen atom compared to the hydrogen atom; prediction performance is worse for ψ angles, as well as for the small glycine residues (S. Wu et al., 2008). While the entire backbone structure can be constructed using only the predicted values of these angles, small errors in their values can result in large deviations in the overall structure.

1.3.3 Predicted contacts

The prediction of inter-residue contacts from coevolutionary information extracts non-local information from the sequence, and has revolutionised template-free

protein structure prediction.

Coevolution methods

One way to predict contacts is to identify covarying residues, which is indicative of coevolution. This suggests that compensatory mutations are important to maintain the protein structure, and the residues are therefore potentially in close physical proximity in three-dimensional space (Göbel et al., 1994). Such correlated mutations can be revealed by applying statistical techniques to large multiple sequence alignments (MSAs) and used to predict contacts. The required depth of the MSA for such inference is on the order of $3L$-$5L$, where L is the number of non-redundant residues in the target (Kamisetty et al., 2013; Seemayer et al., 2014).

In the protein structure prediction and contact prediction literature, contacts are typically defined as pairs of residues, at least four residues apart in sequence, with less than 8Å distance between the C_β residues (or C_α in the case of proline). The accuracy of the top $L/5$ long-range contacts (where L is the length of the target) for the top-performing method increased from 26.7% in CASP11 (Monastyrskyy et al., 2016), to 47.1% at CASP12 (Schaarschmidt et al., 2018), and most recently to 70% at CASP13 (Shrestha et al., 2019).

The improvements in accuracy seen in CASP12 were achieved using direct coupling analysis (DCA) methods that distinguish directly coevolving residues from indirect coupling effects, which previously lead to many false positives. Many of the most successful methods used variations of DCA, including Gremlin (Kamisetty et al., 2013), PSICOV (D. T. Jones, Buchan, et al., 2012), CCMpred (Seemayer et al., 2014), FreeContact (Kaján et al., 2014) and EVFold (D. S. Marks, Colwell, et al., 2011).

Meta-predictors further improved the accuracy of contact prediction. MetaPSICOV is a consensus contact prediction method combining predictions from PSICOV, CCMpred and FreeContact via a machine learning approach (D. T. Jones, Singh, et al., 2015). Predictions are additionally supplemented with predicted secondary structure and solvent accessibility information. The procedure involves two stages of prediction output, in which the second is a refined version of the first. The

predictions produced in the first stage have been shown to be less precise but lead to improved protein structure prediction (D. T. Jones, Singh, et al., 2015; de Oliveira, J. Shi, et al., 2017).

Recently, the application of deep learning methods to contact prediction have led to further accuracy gains while decreasing the size of the MSA required (Shrestha et al., 2019). A number of these deep learning methods are trained to predict bins of inter-atomic distances, rather than binary presence or absence of a contact (Kandathil et al., 2019a; J. Xu and S. Wang, 2019; Ji et al., 2019). While this has a clear performance benefit when incorporated into protein structure prediction methods, it has been noted that the physical principles underlying these predictions are not clear, and may not be possible to analyse (Kryshtafovych, Schwede, et al., 2019; Chonofsky et al., 2019). It is known that different methods predict different sets of contacts (Shrestha et al., 2019; D. T. Jones, Singh, et al., 2015) and result in different performance gains when used for protein structure prediction (de Oliveira, J. Shi, et al., 2017). A recent analysis suggested that the contacts captured by the less accurate DCA-based methods are more likely to mediate physico-chemical bonding interactions than deep learning methods (Coucke et al., 2016; Chonofsky et al., 2019).

Coevolution applications

Coevolution information has been utilised in a range of structural biology applications (Simkovic et al., 2017; de Oliveira and Deane, 2017). Patterns of predicted contacts have been used to estimate domain boundaries (Rigden, 2002; Sadowski, 2013), protein folding rates (Censoni et al., 2018), and protein-protein interactions (Anishchenko et al., 2017; Jana et al., 2014; Simkovic et al., 2017; Croce et al., 2019). Predicted contacts have been used as constraints to guide predicted protein-protein docking (Yu et al., 2017), coarse-grained molecular simulations (Sutto et al., 2015), and the solution of X-ray crystallography and cryo-EM structures (Simkovic et al., 2017; Thomas et al., 2017).

Predicted contacts have had a dramatic impact on the accuracy of template-free protein structure prediction (D. S. Marks, Hopf, et al., 2012). This information

can be incorporated into protein structure prediction, either by constraining a distance geometry algorithm (e.g. D. S. Marks, Colwell, et al., 2011; Adhikari et al., 2015), or by inclusion as a potential in the scoring function for a conformational search algorithm (e.g. Ovchinnikov, Kim, et al., 2016). The extent to which predicted contacts are satisfied can also be used to estimate the quality of a model (e.g. Michel, Skwark, et al., 2017; Maghrabi et al., 2017).

Number of effective sequences

The performance gain in protein structure prediction as a result of coevolutionary information depends on the number and accuracy of contacts that are predicted. For many methods improved performance is correlated with a larger effective number of sequences in the MSA, (Michel, Skwark, et al., 2017) which is defined as:

$$B_{\text{eff}} = \sum_{b=1}^{B} \frac{1}{m_b} \tag{1.1}$$

where m_b is the number of sequences in the multiple sequence alignment with at least 90% sequence identity to the b-th sequence in the alignment, and B is the total number of sequences.

In large-scale structure prediction studies, good predicted contacts and good resulting model structures have been generated for targets with at least 100 effective sequences (Michel, Menéndez Hurtado, et al., 2017). While the latest methods have enabled better performance for targets with fewer effective sequences, the quality of models and the number of effective sequences relative to the target length remain correlated (Abriata et al., 2019).

1.3.4 Evaluation of protein structure prediction

Critical Assessment of Structure Prediction (CASP)

The prediction of tertiary protein structure from sequence has long been a goal of computational biology, and for the last 25 years progress in the field has been benchmarked by Critical Assessment of Structure Prediction (CASP). CASP is a biennial community experiment in which protein modelling methods are evaluated by

independent assessors (Kryshtafovych, Schwede, et al., 2019). The target structures are unknown to the participants and, crucially, are absent from data sets used to develop or train methods. This emulates the prediction of a targets without known structures, and is increasingly important as more methods rely on machine learning from large datasets, and the validation sets used to report performance are not always non-overlapping (Kandathil et al., 2019b).

Measures of model accuracy

A number of evaluation methods are commonly used to assess the accuracy of a model compared to the experimentally-derived structure. Such assessment is not trivial — an ideal metric should have a fixed range, capture relevant structural similarity and variance, distinguish between related and unrelated pairs, be robust to small deviations and experimental errors, and intuitive to understand (Kufareva et al., 2011). The balance of local and global similarity is a particular challenge. Regions that are difficult to model — such as unstructured, flexible or disordered regions — or the incorrect global orientation of domains may obscure otherwise structurally similar models.

Many metrics depend on the optimal superimposition of the model and the native structure, which is itself often a complex task without a single solution (Hasegawa et al., 2009). Sequence-dependent structural alignment ensures that each residue in the model is aligned to the corresponding residue in the native structure, and requires identical sequences. Sequence-independent structural alignment searches for the optimal structural alignment without considering the sequence relationship. Except for distantly related structures, the results using sequence-dependent or sequence-independent methods are highly correlated (Kufareva et al., 2011).

The simplest method of evaluation is to calculate the root-mean-square deviation (RMSD) of atoms in the model compared to the native structure. This may include all atoms or a subset of atoms, such as the backbone atoms, or only those that can be structurally aligned to the native structure. While this method is widely-used and useful when structures are close to native, it has a length dependency

and is disproportionally sensitive to local deviations, particularly when structures are globally more dissimilar (Betancourt et al., 2001). For small targets (< 120 residues), models that are < 3Å RMSD to the native structure are considered very accurate models; at higher values, the RMSD becomes less informative (Y. Zhang, 2009). RMSD is sensitive even to the small deviations in structure that arise due to inherent protein flexibility and experimental resolution limitations (Kufareva et al., 2011). These limitations make it difficult to compare RMSD values across targets

Template-modelling (TM) score and Global Distance Score (GDT) are two methods of evaluation that aim to overcome this problem. TM-score weights the residues at closer distances more strongly relative to those that are more distant. TM-score has range (0,1] and is defined as:

$$\text{TM-score} = \max \left[\frac{1}{L_\text{T}} \sum_{i}^{L_A} \frac{1}{1 + \left(\frac{d_i}{d_0(L_\text{T})} \right)^2} \right] \tag{1.2}$$

where L_T is the number of residues in the native structure, L_A is the number of residues aligned to the model, and d_i is the distance between the ith pair of aligned residues, when the score is maximised by the optimal superposition of the model and native structure. The distance parameter, $d_0(L_T) = 1.24\sqrt[3]{L_T - 15} - 1.8$, is an approximation of the average distance between corresponding residue pairs in two unrelated structures, and is used to remove the dependency on protein size. TM-align is a sequence-independent implementation of TM-align, that allows structures of different sequences to be compared (Y. Zhang, 2005). Using TM-align, pairs of structures scoring below 0.17 have no more similarity than random pairs, while scores greater than 0.5 correspond to structures in the same overall fold (J. Xu and Y. Zhang, 2010).

Global distance test (GDT) is an alternative metric that similarly attempts to avoid the sensitivity to outliers that affects RMSD (Zemla, 2003). For a range of thresholds, the largest proportion of residues falling within the threshold distance of the superimposed native structure is calculated. The total score (GDT-TS) is taken as the average of these scores, ranging between 0 and 100 (sometimes normalised

to the range 0-1). GDT-TS is routinely used for evaluation in CASP, using 1, 2, 4 and 8Å thresholds, and has a high level of agreement with TM-score.

Contact-based methods avoid the ambiguity associated with superposition-dependency. Local distance difference test (lDDT) is an alternative method of evaluation that, unlike RMSD, TM-score and GDT-TS, is superposition-independent (Mariani et al., 2013). It considers, for all pairs of atoms closer than a predefined distance (15Å, by default), the proportion of inter-atom distances that are correctly recapitulated in the model. The final score is calculated as an average over a range of thresholds: 0.5Å, 1Å, 2Å and 4Å, and ranges from 0 to 1. As lDDT is independent of the global superposition and relative domain arrangements it is well-suited to evaluating local similarity; however, is not as widely used as TM-score and GDT-TS for evaluating fold-level accuracy.

1.3.5 Template-based tertiary structure prediction

For many sequences, the structure of a related protein already exists among the growing number of experimentally-determined structures. In this case, the related structure can be used as a template for template-based modelling, also known as homology or comparative modelling. Once one or more suitable templates has been identified, equivalent residues between the template and target structures are defined by alignment; the backbone atom coordinates of the target structure are then generated based on the positions of the template structure residues (Lam et al., 2017; Muhammed et al., 2019).

A major challenge for template-based methods is modelling regions that do not align to any template, particularly loops (Fiser, 2010). Approaches for modelling these regions can be divided into categories: conformational search methods that sample possible conformations while optimising an objective function (e.g. Park et al., 2014; Fiser et al., 2000; Wong et al., 2017), and database search methods that identify compatible loops or fragments of loops from known structures and incorporate these into the model (e.g. (Choi et al., 2009; Deane, 2001)), or a combination of the two (e.g. (C. Marks et al., 2017)).

It is often possible to produce a model with the correct overall fold from a template with a global sequence identity of at least 20%, below which the quality deteriorates (Kryshtafovych, Monastyrskyy, Fidelis, Moult, et al., 2018). The alignment of the template and target is crucial, particularly for low-identity templates (Kryshtafovych, Monastyrskyy, Fidelis, Moult, et al., 2018). The quality of the template structure is also important; more accurate models are produced using high-resolution ($\leq$2.5Å) template structures (Croll et al., 2019).

1.3.6 Template-free tertiary structure prediction

When a suitable template is not available, template-free modelling, also known as *ab initio*, *de novo*, or free-modelling, is required. Methods of this type have improved dramatically in recent years, but challenges still remain.

1.3.7 Fragment assembly

The most well-established approach for template-free protein structure prediction, adopted by many of the most successful methods, is fragment assembly (Bowie et al., 1994; Ovchinnikov, Kim, et al., 2016; J. Yang et al., 2015; D. Xu and Y. Zhang, 2012). Many potential structures of the target are generated, each pieced together from a series of fragments derived from known structures. A fragment is a set of torsion angles representing the backbone conformation of a short sequence of residues. Conformational space is explored by random fragment replacements, which are accepted or rejected based on the score of a conformation compared to the previous conformation. Typically, the score is an objective function consisting of physics- and knowledge-based potentials (see Section 1.3.8).

Most methods use a Markov Chain Monte Carlo sampling strategy, in which less favourable conformations are accepted with a probability related to the size of the energy difference. Some methods also employ Simulated Annealing, in which the probability of unfavourable conformations being accepted is related to the "temperature", which is reduced as sampling proceeds.

Fragment libraries

Fragment-based methods rely on fragment libraries, which are generated specifically for each target sequence (Trevizani et al., 2017). The number of fragments in the library varies between methods, ranging from tens to hundreds of fragments per position (de Oliveira, J. Shi, et al., 2015). This is a trade-off between allowing sufficient exploration of conformational space to capture native-like fragments, and the combinatorial feasibility of exploring this space during sampling.

Two measures of quality are crucial for successful fragment assembly: the proportion of fragments that are of good quality (precision) and the proportion of positions with at least one good fragment (coverage) (de Oliveira, J. Shi, et al., 2015). A cutoff of ≤ 1.5Å backbone RMSD to the native structure is often used to classify fragments as good quality (de Oliveira, J. Shi, et al., 2015).

Fragments of predominantly α-helical secondary structure typically have the highest precision, followed by β-strand fragments, with predominantly loop fragments representing the greatest challenge for fragment generation (de Oliveira, J. Shi, et al., 2015).

1.3.8 Scoring functions

Many protein structure prediction methods, including fragment-based approaches, rely on a scoring function to evaluate candidate protein conformations. Scoring functions, also known as energy functions, are typically a weighted sum of potentials that are constructed to reward favourable conformations. Ideally, for a given target, conformations with lower scores should be closer to the native structure. Scoring functions vary between methods in terms of the implementation, the choice, and the weighting of component potentials.

Each potential typically represents either a knowledge-based (or statistical) term or a physical force. Knowledge-based potentials describe the likelihood of a given property, such as pairwise distances or solvent accessibility, based on propensities derived from known protein structures. Physics-based potentials represent physical forces known to coordinate protein folding; an example is the Lennard-Jones

potential, which is a mathematical model describing the attractive and repulsive forces between two atoms (J. E. Jones, 1924).

1.3.9 Deep-learning based methods

Using geometric constraint methods to fit a model into predicted contacts is an alternative to fragment-assembly approaches that avoids explicit conformational sampling, requiring less computation and generating far fewer models (Senior et al., 2019; Greener et al., 2019; Adhikari et al., 2015; J. Xu, 2019). Such methods rely heavily on large numbers of accurate predicted contacts, usually derived from coevolution methods. With the emergence of deep learning approaches for contact prediction, including distance-based predictions, these methods have dramatically increased in accuracy, but models for targets with very little evolutionary information remain poor (Kandathil et al., 2019b; AlQuraishi, 2019a; Kryshtafovych, Schwede, et al., 2019).

One recent method attempted end-to-end prediction without explicit coevolutionary information using a recurrent geometric network (RGM) that implicitly learnt protein structure features from a large training set of known structures (AlQuraishi, 2019b). A limitation of models produced by deep learning methods, identified during CASP13, is that though models may have correct overall topology, the atomic level structure is often poorly optimised (Won et al., 2019).

1.3.10 Remaining challenges

There are still many remaining challenges in the area of template-free protein structure prediction. Foremost are the performance on long proteins and multidomain proteins.

In CASP12, template-free methods produced models in the correct fold for only half of targets longer than 100 residues (Moult et al., 2018). This is a limitation for protein structure prediction as the median protein lengths for bacteria and eukaryotic organisms are around 267 and 361 residues, respectively (Brocchieri,

2005). In CASP13, at least one correct model was produced for all target domains (which ranged in length from 41 to 431 residues, with an average of 155), but accuracy is still higher for targets less than 150 residues (Won et al., 2019).

A related challenge is that of multidomain proteins. Targets are routinely divided into constituent domains for fragment-based structure prediction, requiring assembly using rigid docking (Inbar et al., 2005) or by sampling the conformational space of the linkers and interfaces between the otherwise rigid domains (Wollacott et al., 2007; D. Xu, Jaroszewski, et al., 2015). The division of the target sequence into appropriate domains is not trivial, and incorrectly assigned boundaries is one of the biggest contributors to modelling failure (Moult et al., 2018). Evaluation of CASP13 entries found that the interactions between domains tend to be modelled inaccurately, even if the individual domains are well-predicted (Won et al., 2019).

1.3.11 Model quality assessment

For the prediction of unknown structure, it is essential that the best of many models produced can be selected, and that the quality of this predicted model can be estimated. Reliable local and global estimations of model accuracy are therefore crucial and are evaluated as an independent category in CASP.

There are two main classes of model quality assessment methods. Consensus (or multi-model) methods compare many models produced for the same target, typically using the similarity between the models to predict the accuracy (Wallner, 2006). Such methods perform well in CASP assessments, particularly at distinguishing good and bad models, but are dependent on the composition of the model ensemble (Kryshtafovych, Monastyrskyy, Fidelis, Schwede, et al., 2018). Single-model methods consider each model individually, and output a score that is independent of other models (Uziela et al., 2017; Benkert et al., 2008). Recent methods of this type can now outperform consensus methods, particularly for selecting a single top-ranked model (Elofsson et al., 2018). Some methods compare each model individually to a set of internally-generated models; these are sometimes referred to as quasi-single model methods (Maghrabi et al., 2017). Recently, deep learning

methods have improved the performance of model quality assessment methods, but little progress was made in the field between CASP12 and CASP13 (Won et al., 2019; Cheng, Choe, et al., 2019).

1.4 SAINT2: Sequential protein structure prediction

SAINT2 is a biologically inspired, fragment-based template-free protein structure predictor that uses a sequential approach (de Oliveira, J. Shi, et al., 2017). The scoring function used in SAINT2 is a combined physics and knowledge-based potential containing RAPDF (Samudrala et al., 1998), Lennard-Jones (J. E. Jones, 1924), solvation, orientation, and predicted contacts (D. T. Jones, Singh, et al., 2015) (Table 1.1). A diagram outlining SAINT2 is shown in Figure 1.8.

To emulate the cotranslational folding of proteins *in vivo*, prediction proceeds by the sequential addition of residues starting from the N-terminus. In the Reverse mode, the direction is the opposite of the biological direction, with prediction starting at the C-terminus. Fragment replacements resulting in a better score are always accepted; those resulting in a worse score are accepted according to the probability: $P = e^{-\Delta/1000}$, where Δ is the difference between the scores of the previous and proposed conformations. Random fragment replacements are performed between extrusion steps, in which fragment replacements extend the peptide length by one residue and are always accepted. The number of replacements carried out between extrusion steps increases linearly as the protein length increases. By default, decoys are generated with a total of 11,000 moves: 10,000 replacement steps are performed during extension and a final 1,000 moves post-extrusion. In the Non-sequential mode all 11,000 replacement steps are performed on the complete chain, which is initialised in a fully extended conformation, with idealised bond lengths and angles.

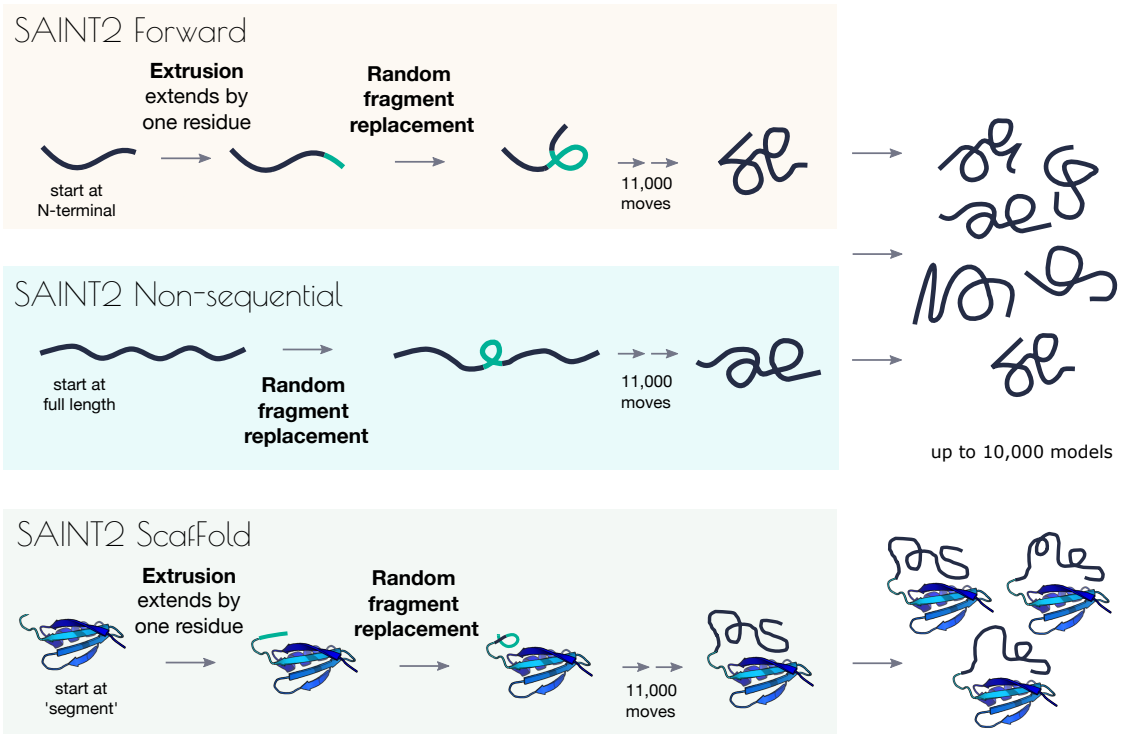

Figure 1.8: A diagram outlining protein structure prediction by SAINT2 Forward, SAINT2 Non-sequential and SAINT2-ScafFold. SAINT2 Forward begins at the N-terminus, performing extrusion moves that extend the chain length by one residue, as well as random fragment replacement moves. SAINT2 Reverse follows the same procedure but from the C- to N-terminus. In the Non-sequential mode, the target begins as a fully extended chain and only random fragment replacements are made. SAINT2-ScafFold begins with a fixed 'segment' of the target structure; the remainder of the structure is completed using SAINT2 Forward or Reverse, without sampling of the segment region.

Optimally, 10,000 predicted structures are generated for each target. Generating more than 10,000 models does not significantly increase the TM-score of the best model for proteins of less than 250 residues (de Oliveira, Law, et al., 2018). Generation of a single model takes approximately 11 seconds for the shortest (81 residues), and 65 seconds for the longest protein in our sets (498 residues), calculated using an Intel(R) Xeon(R) CPU E5-2699 v4 at 2.20GHz with 44 virtual cores and 490GB of RAM.

SAINT2 is motivated by previous work suggesting that a sequential approach generates predictions more efficiently and accurately than conventional search strategies (Ellis et al., 2010) (see Section 1.2.3).

SAINT2 predicts protein structures more successfully in the biologically-relevant

Forward mode than the Non-sequential mode, and to a lesser extent the Reverse mode, in both accuracy and computational efficiency (de Oliveira, J. Shi, et al., 2017). An individual model is produced 1.5-2.5 times faster using a sequential approach compared to non-sequential prediction, and resulted in better quality models for 31 of 41 soluble protein targets and 18 of 42 membrane protein targets. These results highlight the potential to improve model generation by adapting protein structure prediction methods in light of the efficient folding pathways utilised biologically. This may be due to a more efficient exploration of conformational space relevant to the *in vivo* folding pathway.

1.4.1 SAINT2-ScafFold

SAINT2-ScafFold is an implementation of SAINT2 in which part of the target structure is provided (Law, 2017), (Figure 1.8). This region, referred to as the segment, is not sampled during prediction. The remaining region is completed by sequential prediction that takes into account the interactions with the segment during scoring, including predicted contact information.

This method was developed to complete homology models of transmembrane proteins for which at least one terminal helix is not covered by the best available template, a scenario that is common (Law, 2017). SAINT2-ScafFold was able to complete one terminal helix with an accuracy comparable to that of a homology model (<5Å backbone atom RMSD) for the majority of 48 simulated example cases (Law, 2017).

1.4.2 Flib: Fragment library generation

The fragment libraries used by SAINT2 are generated using Flib (de Oliveira, J. Shi, et al., 2015) or its successor Flib-Coevo (de Oliveira and Deane, 2018).

Predicted secondary structure and torsion angles are generated from the target sequence (Figure 1.9, green). For each position in the target sequence, a total of around 3,000 fragments are extracted from the template library, 1,000 from an

Table 1.1: Default SAINT2 scoring function weights and parameters. For some components, the weights differ for short structures (< 150 residues) and long structures (≥ 150 residues); in these cases, the weight for long structures is indicated in italics. The components and weights were trained using a set of 5,425 high-quality crystal structures derived from PISCES (G. Wang et al., 2003), see de Oliveira, J. Shi, et al. (2017) for details.

Component	Value	Description
Scoring function weights		
RAPDF	0.156 *0.303*	Residue-specific All-atom conditional Probability Discriminatory Function (RAPDF) (Samudrala et al., 1998), a knowledge-based score for the frequencies of contacts between atom types at a range of distances (18 bins).
Solvation Potential	0.262 *0.282*	Knowledge-based score for the propensity of a given residue to be exposed. As implemented in Tosatto (2005).
Lennard-Jones Potential	0.505 *0.304*	Mathematical model that approximates the favorability of inter-atomic interaction distances (J. E. Jones, 1924).
Predicted Contact Potential	1	A penalty for each a pair of residues predicted to be in contact that are more than 8Å apart. The penalty is $\delta - 8$ where δ is the distance between the residues in Å.
Core Potential	1	Combined RAPDF and Lennard-Jones, which in the SAINT2 implementation are calculated together for computational purposes.
Orientation Potential	0.077 *0.111*	Knowledge-based score for the frequencies of contacts between atom types at a range of distances (18 bins) and side chain orientations (10 bins). As implemented in Tosatto (2005).
Parameters		
Temperature	2.5	
Extrusion moves	10000	Number of fragment sampling moves on the partial structure during extrusion.
Final moves	1000	Number of fragment sampling moves on the full-length, fully extruded structure.
Move distribution	Linear	The number of fragment sampling moves increases linearly with the length of the structure during extrusion.
Initial length	9	Length of fragment used to initialise sequential prediction.
Mode	Forward	Direction of extrusion. Prediction is carried out sequentially from the N- to C-terminus.

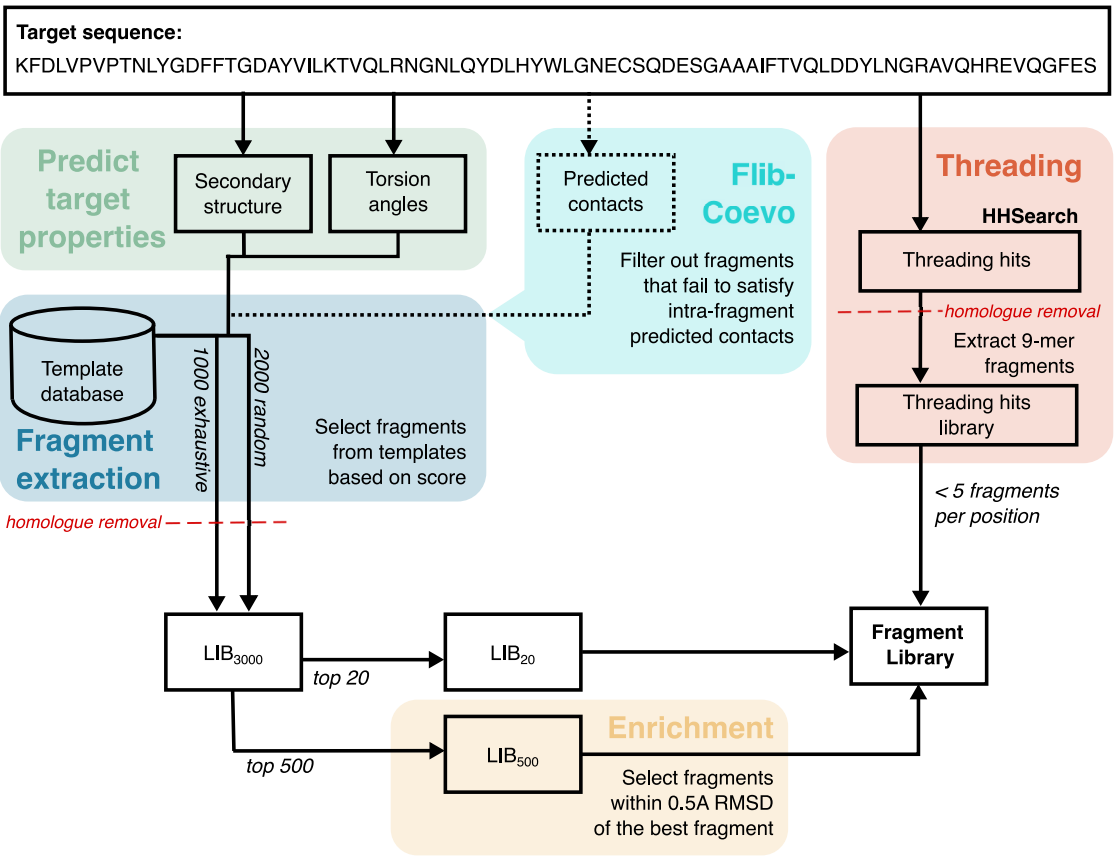

Figure 1.9: Diagram of fragment library generation using Flib and Flib-Coevo. The secondary structure and torsion angles are predicted from the target sequence (green). These are used to score fragments extracted from the template database (dark blue). An average of 3,000 fragments per position are extracted (LIB_{3000}). The 20 fragments with the highest predicted torsion angle score are selected (LIB_{20}), and enriched with fragments among the top 500 (LIB_{500}) that are structurally similar to the top-scoring fragment (yellow), and fragments derived from protein threading hits (orange), to comprise the final fragment library. For benchmarking, to emulate a real template-free prediction scenario, fragments derived from structures with sequence homology to the target are removed (red dashed lines). For the Flib-Coevo protocol, predicted contacts are predicted from the target sequence, and fragments that do not satisfy intra-fragment predicted contacts are filtered out (light blue, dotted lines).

exhaustive search approach and 2,000 from a random search approach (Figure 1.9, dark blue). The exhaustive search scores the secondary structure and torsion angles of every 6–20 residue fragment in the template database against the predicted values for all positions in the target, and the top 1,000 fragments per position are selected. For the random approach, 5,000 randomly selected fragments are scored for each target position, and all fragments that satisfy a minimum score are selected, resulting in an average of 2,000 fragments per position.

From the resulting library of 3,000 fragments per position, the top 20 according to the predicted torsion angle score are included in the final fragment library. This is enriched with fragments from the top 500 that have structural similarity ($<$ 0.5Å RMSD) to the best fragment per position (Figure 1.9, yellow). This enrichment applies to fragments with no predominant secondary structure. Finally, 9-residue fragments extracted from potential template structures — identified using the protein threading software HHSearch — are added to the fragment library (Figure 1.9, orange).

A recent comparison of five fragment library generation programmes found that despite having by far the fewest fragments per position, Flib produced fragment libraries with higher precision for the difficult β-strand and coil regions of target structure for more than 400 targets from three different datasets (de Oliveira and Deane, 2018). When used for structure prediction, these Flib fragment libraries resulted in better models compared to the second-best software, reaffirming the relationship between fragment library precision and model quality.

An extension of the software, Flib-Coevo, exploits coevolutionary information to improve the quality of the fragment libraries (de Oliveira and Deane, 2018). Where contacts are predicted to occur within a fragment, fragments in which these contacts are satisfied were found to have a lower RMSD to the native structure, while fragments that do not satisfy the contacts were more likely to be of poor quality. This is the case despite the fact that not all predicted contacts are correct. In Flib-Coevo, fragments that do not satisfy these contacts are removed and the fragment library is enriched with fragments in which at least one predicted contact is satisfied (Figure 1.9, light blue). This was found to improve the overall precision of the fragment libraries and the quality of the resulting models, with no loss of coverage.

The predicted contacts used for SAINT2 and Flib-Coevo are generated using MetaPSICOV. MetaPSICOV outputs an estimated Positive Predictive Value (PPV) for each predicted contact, which estimates the likelihood that it is a true positive; only predicted contacts with an estimated PPV of at least 0.5 are considered.

1.5 Outline of thesis

In this thesis, we describe extensions to the SAINT2 structure prediction pipeline to improve the prediction of previously intractable targets. These extensions are summarised in Figure 1.10.

In Chapter 2, we describe a new model quality assessment method, RFQAmodel, to identify where modelling has succeeded. This method is published in PLOS ONE and is available to download (West et al., 2019). We demonstrate how RFQAmodel can be used to focus computational effort more efficiently and improve the number of targets for which we have correct models.

In Chapter 3, we describe the incorporation of regions of known structure into the fragment library generation. In combination with RFQAmodel, we explore a new application of SAINT2-ScafFold to complete missing terminal regions of partial protein structures.

Chapter 4 describes a protocol, ScafFoldOn, for improving prediction of long targets, which are currently difficult to predict using a fragment-based approach.

Finally, in Chapter 5 we summarise the results and implications of these extensions. We discuss possible future directions for building on this work, and that of others, towards more successful protein structure prediction.

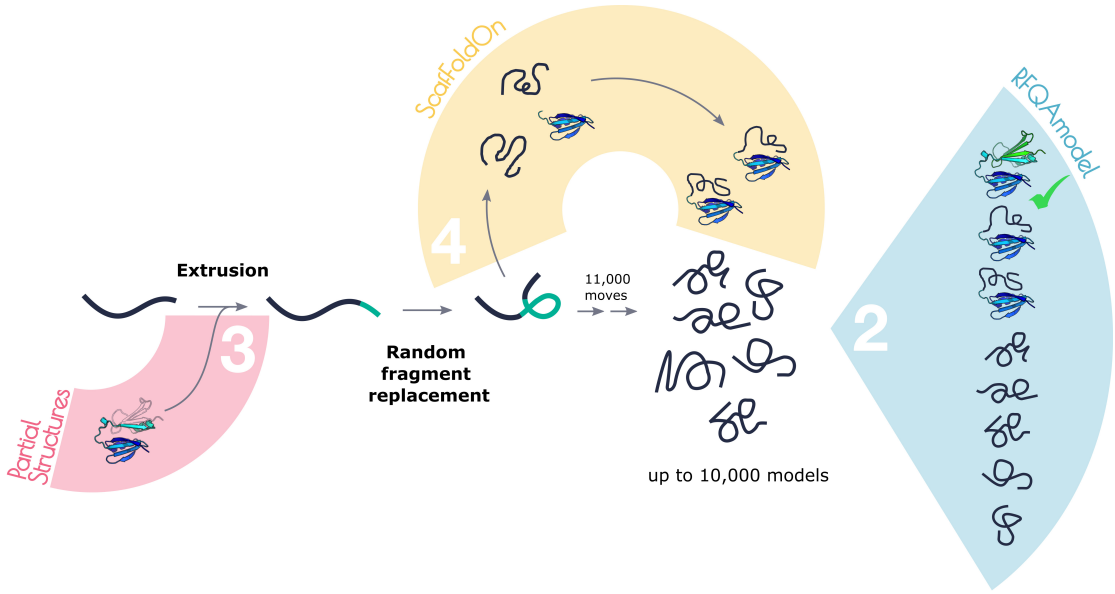

Figure 1.10: An overview of the thesis. The extensions to SAINT2 explored in each chapter of this thesis are illustrated in blue, yellow and red.

2

RFQAmodel: Random Forest Quality Assessment to identify a predicted protein structure in the correct fold

Contents

This chapter is based on a manuscript published in PLOS ONE (West et al., 2019), of which I am the first author. I carried out all the work except constructing the Training and Validation sets and generating the initial 500 models for these targets, which was done by Saulo de Oliveira. Additional analyses carried out by Saulo de Oliveira are included in the appendices.

2.1 Background

Template-free protein structure prediction protocols routinely produce hundreds to thousands of models for a given target (de Oliveira, Law, et al., 2018). Users need to be able to identify whether a good model exists in this ensemble. The final step in a typical structure prediction pipeline is therefore to select a representative subset of five or fewer models as output (Kryshtafovych, Monastyrskyy, Fidelis, Schwede, et al., 2018). This model selection step is critical, and the community's ability to select good models is assessed as part of the Critical Assessment of protein Structure Prediction (CASP) experiments (Moult et al., 2018).

2.1.1 Model Quality Assessment

Protocols for model quality assessment can be divided into three classes: single-model methods, quasi-single model methods, and consensus methods (Kryshtafovych, Monastyrskyy, Fidelis, Schwede, et al., 2018). Single-model methods calculate a score for each model independently, and this score does not take into account any of the other models generated for a particular target. The objective function optimised during protein structure prediction can usually be used as a single-model quality estimator, but better results have been reported if different scores are used for modelling and ranking (Kryshtafovych, Monastyrskyy, Fidelis, Schwede, et al., 2018). Examples of single-model scores include ProQ3D (Uziela et al., 2017) and the ROSETTA energy terms (Leaver-Fay et al., 2011). For quasi-single model methods, the score of a given model is calculated based on its relative score compared to a subset of all models (reference set) produced for the target, for example MQAPsingle (Pawlowski et al., 2016). Consensus methods, such as Pcons (Michel, Skwark, et al., 2017), perform pairwise comparison of the predicted structures to identify clusters of similar models or regions, and assume that structures with high consensus are more likely to be correct.

Predicted contacts derived from co-evolution analysis of multiple sequence alignments have been used by single or quasi-single model methods to improve model quality assessment (e.g. de Oliveira, J. Shi, et al., 2017; Michel, Skwark, et al., 2017). Existing contact-based methods for quality assessment often consider the proportion of predicted contacts that are satisfied in each model (i.e. how many of the pairs of residues predicted to be in contact are within a certain threshold distance) (Michel, Skwark, et al., 2017). ModFOLD6, a quasi-single model quality assessment method, includes a term describing the local agreement with predicted contacts for each residue in the model (Maghrabi et al., 2017). An alternative way to use predicted contact information is to align predicted contact maps for a particular target to the observed contacts maps of models. Contact map alignment has been used to select regions of models to be hybridised (Ovchinnikov, Park, et al., 2017) or to perform protein threading (Buchan and D. T. Jones, 2017). Prior to this work, contact map alignment has not been used for model quality assessment, but the principles that govern these techniques should also be applicable for quality assessment tasks.

In combination with recent advances in model quality due to better contact prediction techniques, improvements in model quality assessment have made template-free protein structure prediction more reliable (e.g. de Oliveira, J. Shi, et al., 2017; Michel, Skwark, et al., 2017). The most recent CASP competition demonstrated remarkable progress in the field: the highest-performing method produced a model in the correct fold (TM-score $\geq$0.5) in the top five models for 23 of 32 free-modelling target domains, although performance decreases when considering only the top model. This level of predictive ability has driven efforts to perform large-scale modelling of significant numbers of protein families without a member of known structure (Ovchinnikov, Park, et al., 2017; Michel, Menéndez Hurtado, et al., 2017). While these studies offer reliable topologies for many protein families, the recall of their quality assessment protocol remains low enough that some predictions with the correct topology may not be identified. Furthermore, such studies were limited by the computational expense of model generation, opting either

to produce models for a subset of these families of unknown structure (Ovchinnikov, Park, et al., 2017) or to produce a reduced number of models per target (Michel, Menéndez Hurtado, et al., 2017).

2.1.2 Random Forests

Random forests are a widely used ensemble method of supervised machine learning used for classification and regression problems (Breiman, 2001). The overall prediction is the mean or mode of the votes made by a large ensemble of decision trees, where each tree is generated using a subset of the training data, and each tree split considers a random subset of features. This method is robust to overfitting and can handle different types of variables as input features without scaling. Furthermore, estimates of feature importance can be computed internally, offering additional insight.

Random forests have performed well for a wide range of structural biology problems (Boyles et al., 2019; Basu et al., 2017; Hou et al., 2017; Luttrell et al., 2019). For model quality assessment, the method RFMQA uses a random forest model to estimate the TM-score of predicted protein structures from potential energy terms and the agreement with secondary structure and solvent accessibility predictions (Manavalan et al., 2014). While RFMQA performed comparably to the state-of-the-art at the time, the performance of model quality assessment methods have since improved.

2.1.3 Overview

In this chapter, we introduce RFQAmodel, a random forest quality assessment classifier developed to evaluate models produced by template-free protein structure prediction pipelines. The classifier combines existing quality assessment scores with predicted contact map alignment scores. Unlike most established quality assessment methods, RFQAmodel is trained to evaluate whether models are in the correct fold (TM-score ≥ 0.5) rather than estimating the absolute model quality. For each

model, RFQAmodel outputs an estimated probability that the model is correct. This probability can be used to estimate whether the model is correct with high, medium, or low confidence, or if modelling is predicted to have failed.

We compiled Training and Validation sets each comprising 244 structurally diverse protein domains. We ensured that these sets were well-balanced in terms of protein length, number of effective sequences (Michel, Skwark, et al., 2017), SCOP class (Fox et al., 2014), and other properties that are known to have an effect on modelling success (Section 1.3.10). We used our sequential protein structure prediction protocol SAINT2 (de Oliveira, Law, et al., 2018) to generate 500 models for each of the 488 protein domains. Using the Training set, we show that predicted contact map alignment scores are as effective for ranking models as existing state-of-the-art quality assessment scores. Furthermore, the models ranked highly by these contact map alignment scores are different from those ranked highly by conventional scores. We incorporate several state-of-the-art quality assessment scores alongside contact map alignment scores into a random forest classifier, RFQAmodel, which classifies models as correct (i.e. in the correct topology) or incorrect, and outperforms the component quality assessment scores. Of the 244 targets in the Validation set, RFQAmodel predicts that the highest-ranking model may be correct for 185 targets, of which 86 are correct (out of a possible 142 for which at least one correct model was generated by SAINT2). The 185 are further split by RFQAmodel into those where the highest-ranking model is predicted to be correct with high confidence, 67 targets, of which 52 are correct. Of the 59 targets predicted to be modelling failures, 5 had at least one correct model, and none had a correct highest-ranking model. We demonstrate that similar results are achieved when applied to the server models submitted to CASP12 and CASP13. Finally, we demonstrate how RFQAmodel can be used to estimate when sufficient models have been generated for a particular target, enabling more efficient use of computational power.

2.2 Methods

2.2.1 Training and Validation Sets

To construct the Training and Validation data sets, we used the mapping between Pfam (Punta et al., 2012) domains and PDB (Berman, 2000) structures as available on the EBI repository in February 2017. To represent each of these families, we selected the first protein chain listed for that family (Appendix A.1).

We annotated each of the protein chains according to the 2.06 stable build of SCOPe (Fox et al., 2014). If the protein chain selected to represent a Pfam family was not annotated in SCOPe, we tested all the remaining members of the family sequentially (as ordered on the mapping) to maximise the number of Pfam families with SCOPe annotations (Table A.1 and Figure A.1).

We excluded all families longer than 250 residues (see Section 1.3.10), and performed a culling and cleaning process (Appendix A.1.2) that resulted in a data set of 488 structurally diverse protein domains (Table A.3). The average length and number of effective sequences, B_{eff} (see Section 1.3.3), of these domains were similar to those of the original PDB-mapped and SCOPe-annotated Pfam domain sets.

The 488 protein domains were divided into Training and Validation sets of equal size. For each SCOP class, we selected two domains at a time in order of increasing B_{eff} and randomly assigned one to the Training and the other to the Validation set. We used the B_{eff} of the multiple sequence alignments used for contact prediction. While this ensured that the sets have similar B_{eff} medians and have roughly the same number of protein domains for each SCOP class, the overall length and resolution distributions differed between sets (Figure A.2). In particular, proteins in the Validation set with B_{eff} <100 tended to be longer than proteins on the Training set with B_{eff} <100, which suggests that the Validation set may be more challenging for protein structure prediction.

2.2.2 Protein Structure Prediction

To produce models for all targets in our Training and Validation sets, we used our fragment-assembly protocol SAINT2 (de Oliveira, Law, et al., 2018) (see Section 1.4) with the following input and parameters.

Prediction of sequence-based descriptors

We used SPIDER3 (Heffernan et al., 2017) with standard parameters to perform torsion angle prediction.

Contact prediction was carried out using metaPSICOV (D. T. Jones, Singh, et al., 2015) with standard parameters. MetaPSICOV uses a two-stage neural network that outputs two sets of predictions: stage1 and stage2. Although metaPSICOV stage2 predictions are reported to be more precise, we use the stage1 predictions, which have been shown to lead to better modelling results (D. T. Jones, Singh, et al., 2015; de Oliveira, J. Shi, et al., 2017).

We used DeepCNF Q8 secondary structure predictions (S. Wang et al., 2016), which had a slightly higher precision for targets with large B_{eff} values compared to the previously used method, PSIPRED (D. T. Jones, 1999), and result in marginal improvements in fragments with predominantly loop secondary structure, according to benchmarking performed by Saulo de Oliveira (see Appendix A.2.1).

Further analysis by Saulo de Oliveira found that the precision of the predicted torsion angles, secondary structure and contacts are comparable between the Training and Validation sets (Appendix A.2.2).

Fragment library generation

Flib-Coevo was used to produce fragment libraries for all proteins in our Training and Validation sets. Fragments from homologues were discarded to represent a more realistic template-free protein structure prediction scenario. Homologues were identified using a protein-protein BLAST (Camacho et al., 2009) search against the target protein with an E-value cutoff of 0.05 as in de Oliveira and Deane (2018).

Model generation

In order for SAINT2 to produce the best possible model, the optimal number of models to generate is 10,000 (de Oliveira, J. Shi, et al., 2017). However, for the purpose of developing a quality assessment protocol, Saulo de Oliveira estimated that only 500 models were required to produce correct models for a sizeable number of targets (see Appendix A.3).

We used SAINT2 with standard parameters to produce 500 models for each target in our Training and Validation sets. We assessed the number of modelling successes - targets for which at least one correct model (TM-score ≥ 0.5, Y. Zhang and Skolnick, 2004) was produced - as well as the TM-score of the best model produced for each target.

2.2.3 CASP12 and CASP13 Test Sets

To test our classifier on models produced by methods other than SAINT2, and to compare its performance to other quality assessment methods, we used the stage2 server models used in the blind test of model quality assessment methods at CASP12 and CASP13. These consist of the 150 top-ranking server models submitted for 60 targets each for CASP12 and CASP13 targets. The models, model quality predictions, and model quality evaluations were accessed from the CASP website (http://www.predictioncenter.org/download_area/). This resulted in a total of 17,976 models for 120 targets. The lengths of the target structures range from 41 to 863 residues, with an average length of 289 residues.

2.2.4 Model Validation

To assess the quality of the SAINT2 models, we used TM-align to calculate TM-score (Y. Zhang and Skolnick, 2004). For the models in the CASP12 and CASP13 sets, we used the TM-score calculated by the CASP assessors. We consider all models with a TM-score ≥ 0.5 to be in the correct topology (J. Xu and Y. Zhang, 2010).

2.2.5 Classification Features

We used the randomForest package in R (Liaw et al., 2002) to build a classifier, RFQAmodel, that classifies models as correct or incorrect. All training was carried out on our Training set. Our random forest classifier was initially trained using 5-times cross-validation to test the sensitivity of the classifier to its hyperparameters. Cross-validation was performed by splitting the data set on a per-target basis, meaning that all models for a target were either used to train or to cross-validate results. We used 500 trees and 7 maximum features per tree for classification. Features were not normalised as tree-based classification does not require scaling. For reproducibility, the results presented here were attained with a fixed seed of 1011.

For model classification, we used a set of 58 features, which can be divided into three groups: target-specific (3), model-specific (12), and ensemble-specific (43). The target-specific features are calculated from the target's sequence, and are common to all models produced for that target. The model-specific features are calculated for each model, and include five existing single-model quality assessment scores, a consensus method quality assessment score, two scores based on the predicted contacts, and three predicted contact map alignment scores. The ensemble-specific features are summary statistics (maximum, median, minimum, and spread) of our model-specific features calculated across all models produced for each target. For all methods we used SAINT2 models and the predicted contacts generated by metaPSICOV. We note that many of the assessment scores used were not originally trained using these inputs, so their performance may be worse than expected.

Target-specific features (3): The domain length, the B_{eff}, and the total number of predicted contacts output by metaPSICOV with a score greater than 0.5.

Single-model quality assessment scores (5): The final modelling score output by SAINT2, and the global score output by ProQ3D and component scores ProQ2D, ProQRosFAD and ProQRosCenD (Uziela et al., 2017). ProQ3D uses a deep neural net model to estimate the quality of each model from a large number of input features. ProQRosCenD and ProQRosFAD are based on the Rosetta centroid and full atom (Leaver-Fay et al., 2011) energy functions, respectively, which were

calculated on relaxed models with repacked side chains. Relaxation was carried out using the *ab initio* relax protocol of Rosetta 3.7 as described in Uziela et al. (2017). For CASP12 and CASP13 models, the SAINT2 score was calculated for each model without performing any minimisation. For ranking models, we have additionally considered the SAINT2 score without its contact component (SAINT2 Raw); this was not included as a feature in the random forest classifier.

Consensus quality assessment score (2): We used the global score output by Pcons (Wallner, 2006) with standard parameters. Pcons compares each model to all others for a given target, and scores well-conserved residue positions more highly. We also include PcombC (Michel, Menéndez Hurtado, et al., 2017), a weighted sum of three features: the ProQ3D global score, the Pcons consensus score, and the proportion of predicted contacts present in the model (positive predictive value, PPV).

Contact-based features (2): The contact component of the SAINT2 score and the proportion of satisfied predicted contacts (positive predictive value, PPV). Here, we considered a predicted contact to be a satisfied if the C-β atoms (C-α in the case of glycine) of the two residues predicted to be in contact were less than 8Å apart in the model output by SAINT2.

Predicted contact map alignment scores (3): We used BioPython (Cock et al., 2009) to calculate an observed contact map for each model, with an 8Å distance cut-off between residue C-β atoms (C-α in case of glycine). We aligned the observed contact maps to the predicted contact maps produced from the output of metaPSICOV stage1. Two methods of contact map alignment were tested: map_align (Ovchinnikov, Park, et al., 2017), and EigenTHREADER (Buchan and D. T. Jones, 2017). Map_align uses a dynamic programming algorithm to perform local contact map alignment and identify consensus regions. We used as features the best hit score and the best hit length produced by map_align. EigenTHREADER uses eigenvector decomposition and dynamic programming to align the principal eigenvectors of the two maps. For an ensemble of structures, EigenTHREADER assesses which of the models is most likely to be in the same fold as the one described

by the reference predicted contact map, assigning a relative score per model. We used the score output by EigenTHREADER for each model as a feature.

Ensemble-specific features (43): The maximum, minimum, median, and spread (the difference between the maximum and the median) of 10 of our 12 model-specific features, excluding map_align's hit length and the proportion and absolute number of satisfied predicted contacts, for which only the maximum value for each target is included. These features were calculated per target across all models.

2.3 Results

2.3.1 Modelling Results

Correct models were produced for 151 out of 244 protein domains in our Training set, and 145 out of 244 protein domains in our Validation set. This corresponds to around 60% of the targets in each set, in line with numbers reported previously (de Oliveira, Law, et al., 2018).

Our modelling results vary by B_{eff}, SCOP class and domain length. When considering the modelling results according to three B_{eff} bins (Figure 2.1A), our results corroborate previous findings that modelling is more likely to succeed when more effective sequences are available (de Oliveira, J. Shi, et al., 2017). For our Training set, we observe a modelling success rate of 46% at B_{eff} values below 100, and a success rate of 69% for $B_{\text{eff}} \geq 1000$ (Figure 2.1A).

We also find that modelling success rates vary by SCOP class (Figure 2.1B) and domain length (Figure 2.1C). SAINT2 produced a correct model for 85% of all-α targets, 65% of α/β targets, 61% of $\alpha+\beta$ targets, and 30% of all-β targets (Figure 2.1B). As expected, modelling success rate decreases as targets increase in length (Figure 2.1C). For our Training set, SAINT2 produced a correct model for 83% of the targets that were 50 to 99 residues-long, for 65% of targets that were 100 to 149 residues-long, for 41% of targets that were 150 to 199 residues-long, and 39%

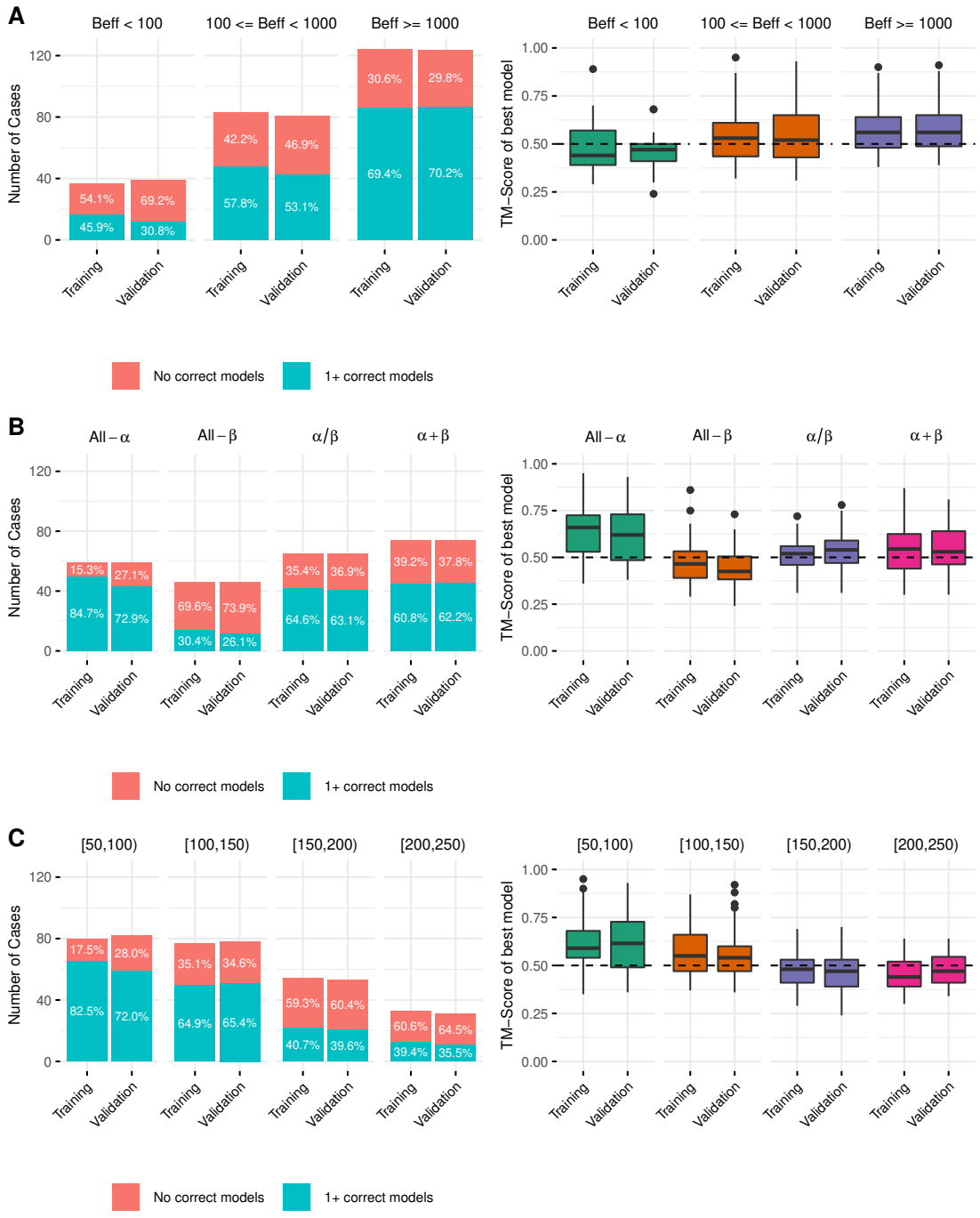

Figure 2.1: Modelling success rate by B_{eff}, SCOP class, and domain length. Number of targets for which correct models were (blue) or were not (red) produced (left) and distributions of the TM-score of best model (right) for the 244 protein domains in each of our Training and Validation sets. Results are shown for different **A**) B_{eff} bins **B**) SCOP class and **C**) domain length. A model with a TM-score ≥ 0.5 is considered correct.

of targets longer than 200 residues. When considering the combined effect of B_{eff} and domain length, SAINT2 failed to produce a correct model for all targets longer than 200 residues with a $B_{\text{eff}} < 100$ (Figure A.7). Comparable modelling success rates and distributions of TM-score of the best models were obtained for Training and Validation sets across all B_{eff} bins, SCOP classes and domain length, with marginally worse performance for Validation set targets with B_{eff} values below 100.

Given the effect of these three features on modelling success, it is important to ensure that Training and Validation sets have similar distributions of domain length, effective sequences, and SCOP classes. A validation set that is comprised of shorter targets, or that contains more targets with a high B_{eff}, or a disproportionate number of α-helical targets may lead to overestimation of classification performance.

2.3.2 Comparing Quality Assessment methods

To assess the usefulness of including predicted contact map alignment scores as features for model quality assessment, we compared these scores with ten other model quality estimators: three SAINT2 scores and seven existing quality assessment scores. We ranked the 500 models produced by SAINT2 for each of the 244 targets in our Training set according to each of these model quality scores. For each score, we assessed the number of targets for which the highest-ranking model was correct (TM-score ≥ 0.5). Given that the quality of models is dependent on the availability of a sufficient number of effective sequences (B_{eff}), we stratified this comparison across three B_{eff} bins (Figure 2.2).

We consider modelling to be a success if at least one correct model is produced for a target. For $B_{\text{eff}} \geq 1{,}000$, our set contained correct models for 86 out of 124 targets ("Total Successes" in Figure 2.2). The two best methods for selecting correct models in this B_{eff} bin were the SAINT2 score and EigenTHREADER's predicted contact map score; the highest-ranking models of these methods were correct (TM-score ≥ 0.5) for 58 and 57 targets, respectively. The predicted contact potential of the SAINT2 score, SAINT2_Contact, also identified correct models for 57 targets, while only 38 were identified when this potential is excluded (SAINT2_Raw). Within

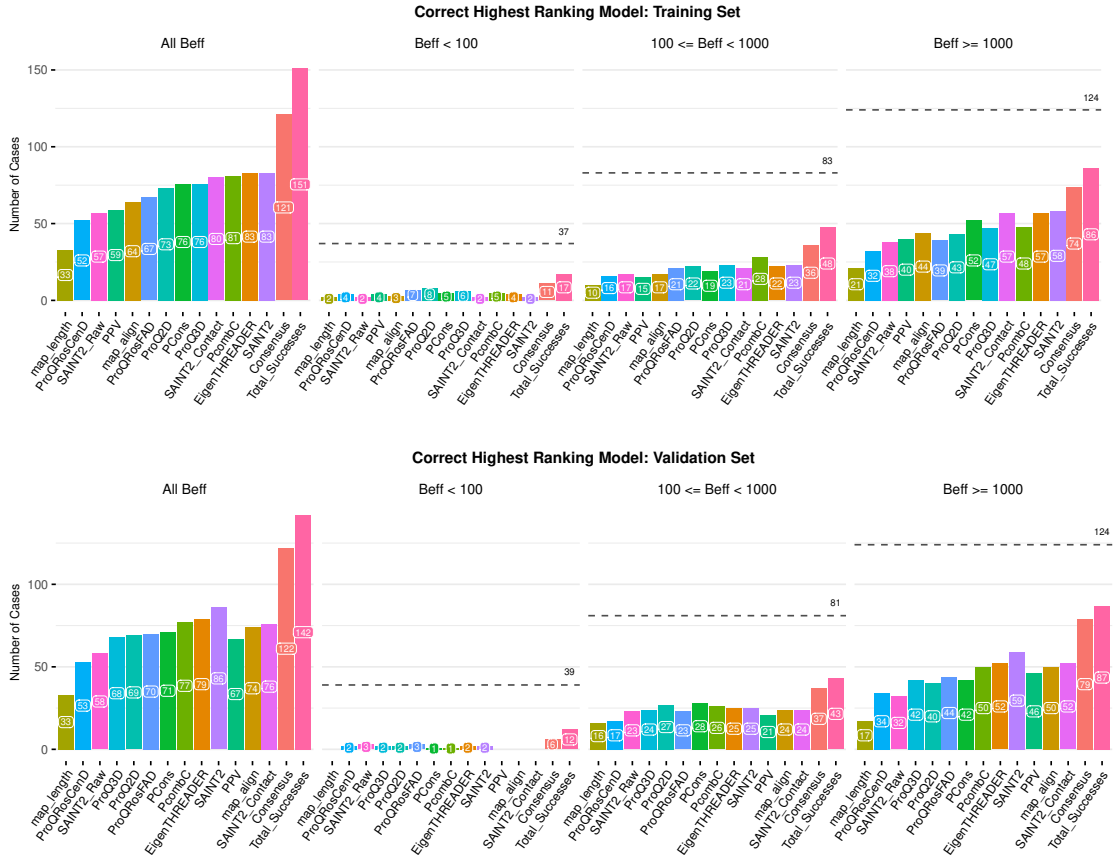

Figure 2.2: Number of targets out of the 244 targets in our Training set (top) and Validation set (bottom) for which a correct model was produced and selected as the highest-ranked model according to 13 methods. Three SAINT2 scores (SAINT2, SAINT2_Contact and SAINT2_Raw), seven existing quality assessment scores (ProQ3D, ProQRosCenD, ProQRosFAD, Pcons, PcombC, ProQ2D and PPV), and three predicted contact map alignment scores (EigenTHREADER, Map_align and map_length) are shown, as well as all methods combined ("Consensus") and the total number of targets with a correct model ("Total Successes"), for three B_{eff} bins and across all bins. The total number of targets in each B_{eff} bin is indicated with a dashed line.

this B_{eff} bin, the length of the map_align predicted contact map alignment selected correct models for the smallest number of targets, followed by PPV, the proportion of predicted contacts satisfied in the model. ProQRosCenD, a score based on the centroid knowledge-based energy potential Rosetta Centroid, also identified fewer correct models than the other scores, with a similar performance to SAINT2_Raw.

When considering $100 \leq B_{\text{eff}} < 1{,}000$, our set contained correct models for 48 out of 83 targets ("Total Successes" in Figure 2.2). PcombC performed the best at identifying correct models for this B_{eff} bin, with correct highest-ranking models

for 28 targets, followed by the SAINT2 score and ProQ3D, each with 23 correct highest-ranking models. For $B_{\text{eff}} < 100$, our set contained correct models for 17 out of 37 targets. For these targets ProQ2D was the most successful, selecting a correct model for eight targets. Similar results were observed when considering the models for targets in the Validation set (Figure 2.2).

As expected, these results demonstrate that methods using predicted contact information perform well on targets with more sequence data available, while knowledge-based scores are more informative for targets with less of this data. Overall, the SAINT2 score and EigenTHREADER identified correct highest-ranking models for 83 targets each, more targets than any other method (Figure 2.2). The best three methods, SAINT2, EigenTHREADER and PcombC identify correct models for different targets. Of the targets correctly identified by EigenTHREADER, 12 are not identified using SAINT2, and 17 are not identified by PcombC (Figure 2.3A). Incorporating EigenTHREADER scores when ranking the models produced by SAINT2 may therefore improve our ability to identify correct models. While these three methods are the major contributors (Figure 2.3B), we included all 12 methods into our random forest classifier as all methods had some predictive power.

2.3.3 RFQAmodel: model quality assessment

Among the Validation set of 244 targets, 142 have a correct model within the 500 models produced by SAINT2. Selecting the highest-ranking model according to the SAINT2 score results in a correct model for 86 targets in this set. However, as the SAINT2 score cannot easily be compared between targets, it is difficult to infer for which targets the highest-ranked models are correct. We have trained a classifier, RFQAmodel, that assesses each model produced for a target and outputs a score, between 0 and 1, that the model has the correct fold.

We assessed the performance of RFQAmodel on our Validation set. Using a Receiver Operating Characteristic (ROC) curve, RFQAmodel achieved an area under the curve (AUC) of 0.95 for classifying all models for all targets as correct or incorrect, higher than all the individual component scores, including the best

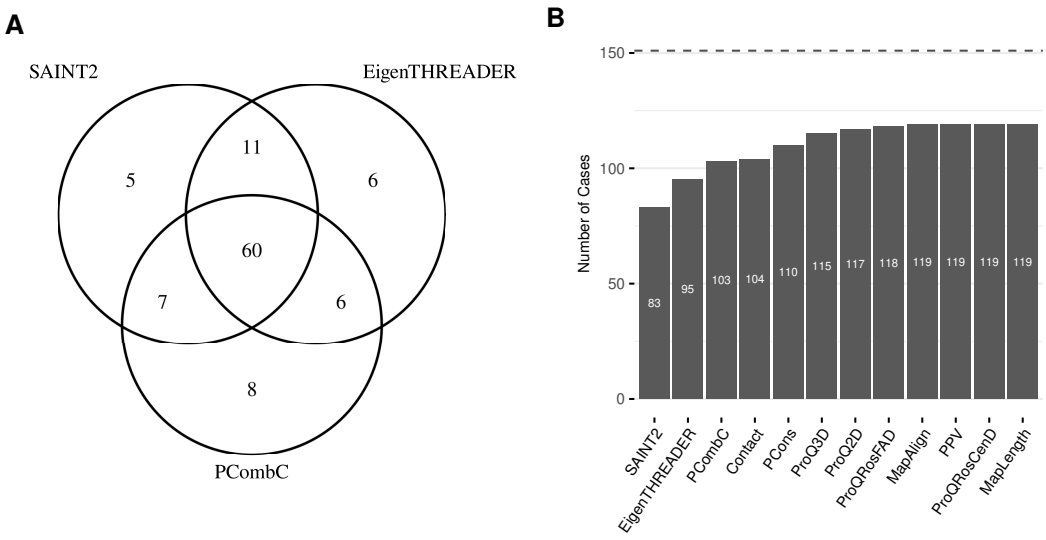

Figure 2.3: **A**) The number of targets in our Training Set for which the highest-ranking model is correct, when ranked according to the three overall best methods: the SAINT2 score, quality assessment score PcombC, and predicted contact alignment score EigenTHREADER. **B**) The number of targets for which at least one method identified a correct model as the highest-ranking, when methods are sequentially added in order of ranking ability as shown in Figure 2.2 (from left to right). The total number of targets with at least one correct model is indicated with a dashed line.

individual quality assessment score, Pcons (0.91), EigenTHREADER (0.84), and the SAINT2 score (0.77), as well as the other quality assessment scores ProQ2D (0.90), ProQ3D (0.89), ProQRosFAD (0.88), and PcombC (0.79) (Figure 2.4A). In practice, we are interested in the classification of the highest-ranked model per target as correct or incorrect; for this task, RFQAmodel also outperforms the component methods (Figure 2.4B).

The estimated relative importance of each feature included in the model is shown in Appendix Figure A.8. The inclusion of the ensemble features improves the number of targets for which the highest-ranking model is in the correct fold and is classified correctly, when compared to the performance when the ensemble features are excluded (Figure 2.5).

We divided the score output by RFQAmodel into four broad categories based on the Training set data: correct with high (>0.5), medium (between 0.3 and 0.5), or low (between 0.1 and 0.3) confidence, or predicted modelling failures ($\leq$0.1) (Figure 2.6).

The models for a given target were ranked according to the RFQAmodel score,

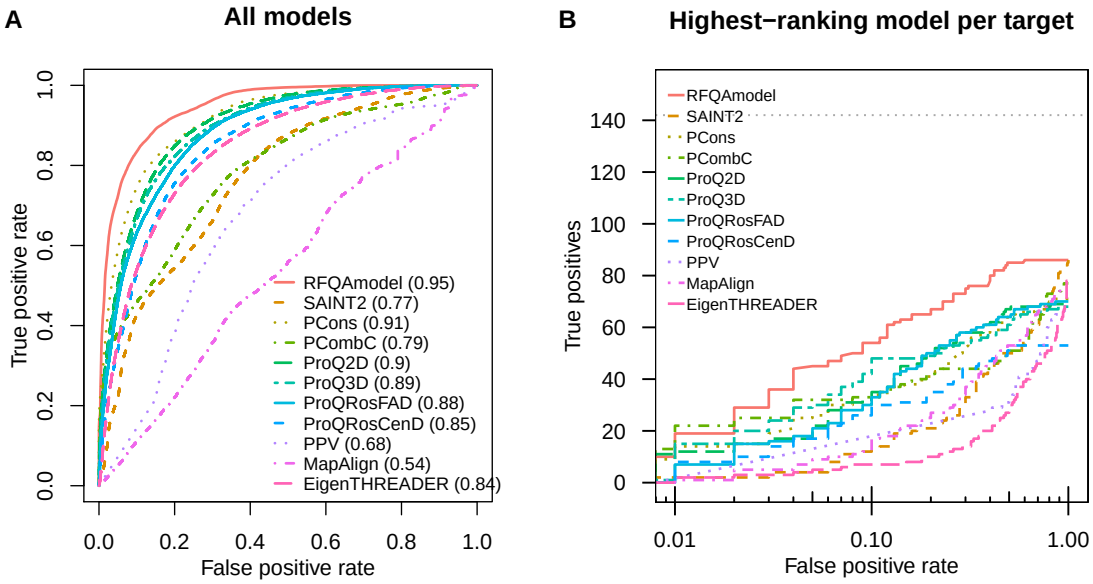

Figure 2.4: RFQAmodel classification of Validation Set targets.
A) Receiver Operating Characteristic (ROC) Curves for the classification of all models into whether they were correct (TM-score ≥ 0.5) or incorrect according to RFQAmodel and the 10 component methods for the 244 targets in our Validation set. The area under the ROC curve (AUC) for each method is shown in brackets. The EigenTHREADER score was normalised by the maximum score for each target. B) The number of targets with a correct highest-ranking model (true positives) plotted against the false positive rate on a logarithmic scale. The grey dotted line indicates the total number of targets that had at least one correct model.

Table 2.1: The performance of our classification protocol for the 244 modelling targets in our Validation set. Modelling is considered successful for a given target if at least one model is correct (TM-score ≥ 0.5). The number of targets (Total) as well as the number of targets with at least one correct model (Max) is reported for each confidence category. The number of targets for which the highest-ranked model (Top1) and the best of the top five highest-ranked models (Top5) are shown, with the corresponding precision in brackets.

Confidence	Total	Max	Top1		Top5	
High	67	63	52	(77.6%)	60	(89.6%)
Medium	50	38	21	(42.0%)	30	(60.0%)
Low	68	36	13	(19.1%)	21	(30.9%)
Failed	59	5	0	(0.0%)	1	(1.7%)

and targets were categorised based on the RFQAmodel score of the highest-ranking model. For each level of confidence, we assess whether the highest-ranking model (Top1) or the best of the top five highest-ranking models (Top5) is correct (Table 2.1).

When the models for each target in the Validation set are ranked according to the RFQAmodel score, the highest-ranking (Top1) model is correct for 86 of 244

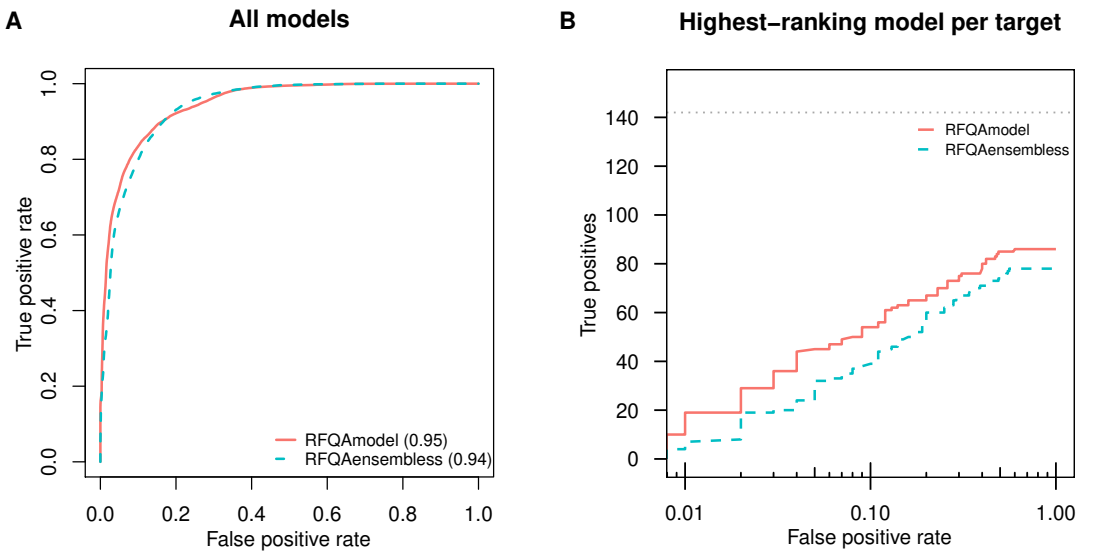

Figure 2.5: RFQAmodel classification of Validation Set targets when ensemble features are excluded.
A) Receiver Operating Characteristic (ROC) Curves for the classification of all models into whether they were correct (TM-score ≥ 0.5) or incorrect according to RFQAmodel and a version of the model trained without the ensemble features (RFQAensembless). The area under the ROC curve (AUC) for each method is shown in brackets. **B)** The number of targets with a correct highest-ranking model (true positives) plotted against the false positive rate on a logarithmic scale. The grey dotted line indicates the total number of targets for which at least one correct model was produced by SAINT2.

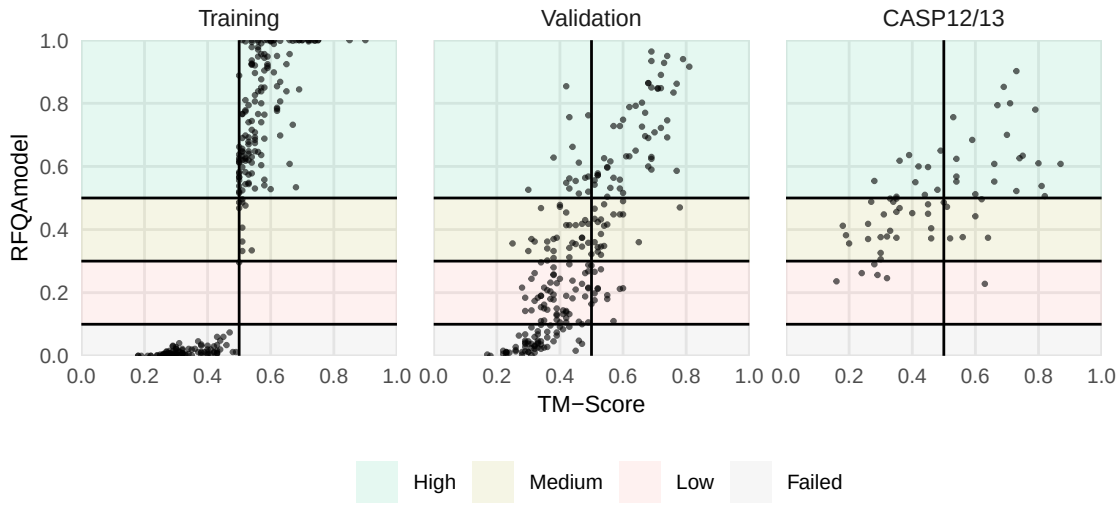

Figure 2.6: The RFQAmodel score and TM-score of the highest-ranking model per target in the Training and Validation sets, as well as the CASP12 and CASP13 targets. Predictions are categorised into high (>0.5, green), medium (between 0.3 and 0.5, yellow), or low (between 0.1 and 0.3, red) confidence, or are predicted to have failed ($\leq$0.1, grey).

targets. This is exactly the same as the number of correct highest-ranking models when ranked according to SAINT2; the difference is that RFQAmodel assigns a likelihood that each model is correct. RFQAmodel predicts that modelling has failed ($\leq$0.1) for all models for 59 targets. For 5 of these targets there was at least one correct model in the 500, but the highest-ranked model was not correct for any. Excluding these 59 targets reduces our Validation set from 244 to 185 targets, of which 137 have a correct model.

The highest-ranking (Top1) model was predicted to be correct with low confidence for 68 targets. This model was correct for 13 of these targets (19% precision), and 21 targets had a correct model in the top five (Top5) highest-ranking models (31% precision).

The highest-ranking model was predicted to be correct with medium confidence for 50 targets. The highest-ranked model was correct for 21 of these targets (42% precision), and the best out of the top five highest-ranking models was correct for 30 targets (60% precision).

The highest-ranking model was predicted to be correct with high confidence for 67 targets. This model was correct for 52 out of these 67 high-confidence targets (78% precision), and the best out of the top five highest-ranking models was correct for 60 of these targets (90% precision).

When considering the combined results for the 117 targets with highest-ranking models predicted to be correct with high or medium confidence, this model was correct for 73 targets (62% precision), and the best out of the top five highest-ranking models was correct for 90 of these targets (77% precision).

2.3.4 Comparison to methods used in large-scale studies

We compared RFQAmodel to two methods that have been used to evaluate the success of large-scale predictions of unknown protein structures by Michel et al. (Michel, Menéndez Hurtado, et al., 2017) and Ovchinnikov et al. (Ovchinnikov, Park, et al., 2017). In the study by Michel et al., the authors used the PcombC score cut-off that achieved a false positive rate (FPR) of 0.01 and 0.1 on the benchmarking

set to predict whether models were correct (TM-score ≥ 0.5) (Michel, Menéndez Hurtado, et al., 2017). PcombC is one of the scores used in RFQAmodel, so it is unsurprising that RFQAmodel is able to achieve better performance (Figure 2.4). Compared to PcombC, RFQAmodel performs similarly at an FPR of 0.01, but identifies a correct model for more targets at 0.1 (Figure 2.4B).

To compare RFQAmodel with the method used in Ovchinnikov et al., we calculated the mean pairwise TM-score of the 10 models with the highest ProQ3RosCenD score out of the 500 models generated for each target, and classified targets above 0.65 as correct (Ovchinnikov, Park, et al., 2017). This method classified 21 targets as correct, of which 19 had a correct highest-ranking model. A similarly high precision was achieved using ProQ3RosFAD instead of ProQ3RosCenD (19 out of 22). Using RFQAmodel, a similar precision with higher recall can be achieved with a cut-off of 0.7, with 26 of 29 targets having a correct highest-ranking model (Figure 2.7, solid lines). Using the high confidence cut-off for RFQAmodel we achieve 78% precision and 37% recall. At this level of recall, the ProQRosCenD method achieves a precision of 36% (Figure 2.7, dashed lines). The difference between the methods appears to be the ability of RFQAmodel to identify correctly modelled targets with fewer correct models (Figure 2.7), which is known to be particularly challenging (Kryshtafovych, Monastyrskyy, Fidelis, Schwede, et al., 2018).

2.3.5 Comparison to a regression model

Another way to evaluate the success of protein structure prediction is to estimate the actual TM-score of models. We trained a version of our random forest classifier, RFQAregression, that attempts to predict the TM-score of each model. The methods perform similarly when used to classify models as correct or incorrect (Figure 2.8A and B). We evaluated the ability of each method to identify the best model among the ensemble for each target. At this task, the performance of RFQAregression is slightly better than RFQAmodel when considering the top1 model, while both methods outperform the best individual component method, PcombC (Figure 2.8C and D). However, when a predicted TM-score ≥ 0.5 is used

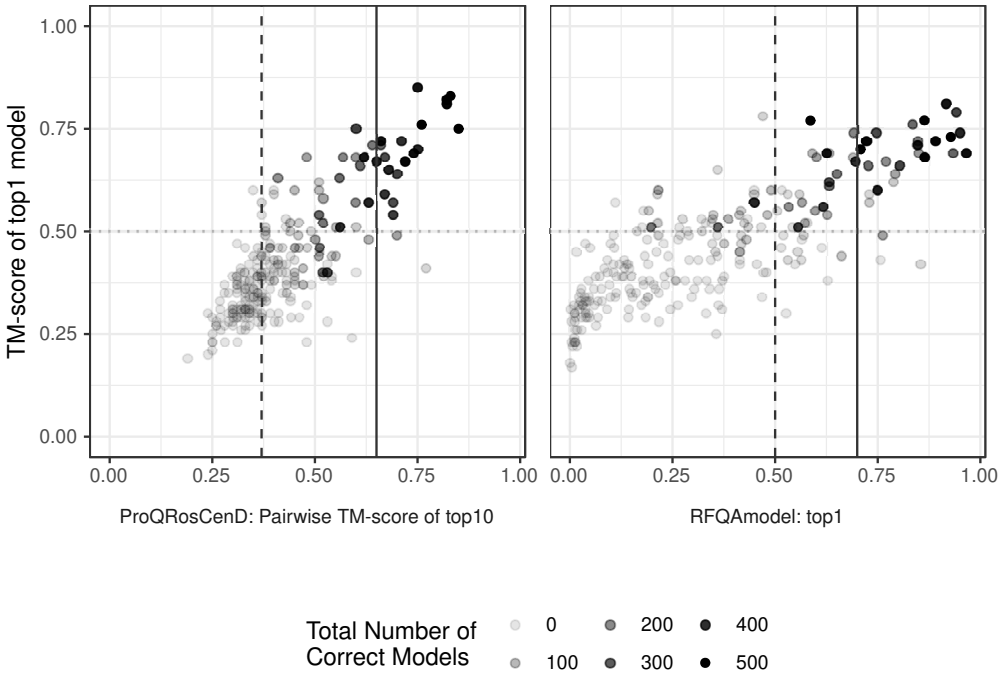

Figure 2.7: Using convergence or RFQAmodel to identify correct models. The TM-score of the highest-ranking model for each of the 244 targets in the Validation set according to ProQRosCenD and RFQAmodel, against the mean pairwise TM-score of the 10 highest-ranking models (ProQRosCenD, left) or the score of the highest-ranking model (RFQAmodel, right). Targets with a mean pairwise TM-score greater than 0.65 are predicted to be correct (solid line, left); a similar precision is achieved with an RFQAmodel cut-off of 0.7 (solid line, right). A pairwise TM-score cut-off of 0.37 (dashed line, left) achieves a similar recall to the high confidence cut-off of RFQAmodel (dashed line, right). Targets for which fewer correct models were generated among the 500 models are shown with lighter circles.

as a cutoff to predict modelling success, RFQAregression has a similar precision but recovers correct models for fewer targets than the equivalent High confidence category of RFQAmodel (Table 2.2). This measure demonstrates that RFQAmodel is better able to identify whether the estimated best model is in the correct fold or not, which is the purpose of this method, and is crucial for real-world applications of protein structure prediction methods.

2.3.6 CASP12 and CASP13 Quality Assessment

RFQAmodel was trained and validated on models generated using SAINT2. In order to test its performance on models generated by other methods, we used RFQAmodel

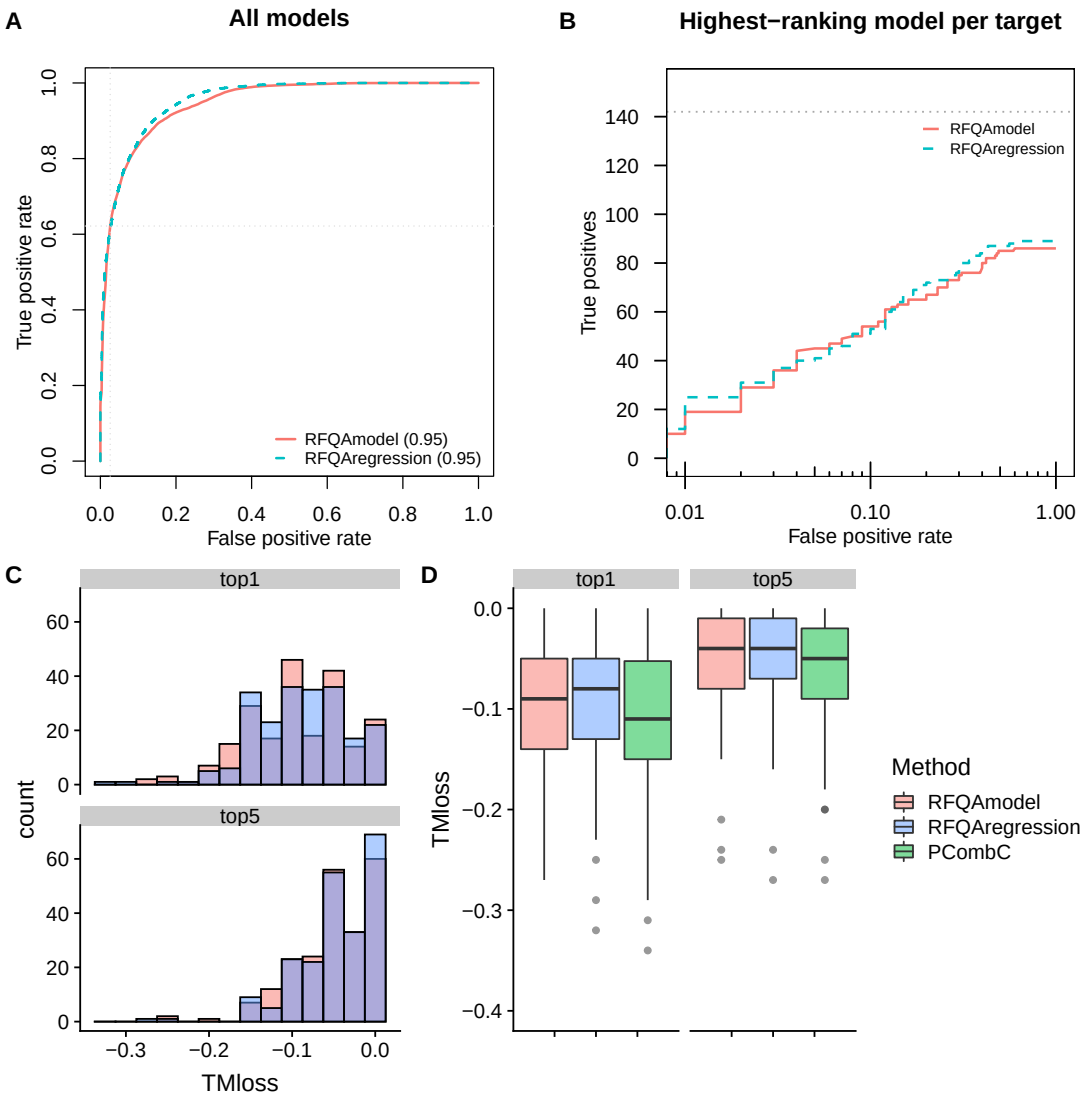

Figure 2.8: Performance of RFQAmodel compared to a version of the model trained to predict the actual TM-score of each model (RFQAregression) on the Validation Set

A) Receiver Operating Characteristic (ROC) Curves for the classification of all models into whether they were correct (TM-score ≥ 0.5) or incorrect according to RFQAmodel and RFQAregression. The area under the ROC curve (AUC) for each method is shown in brackets. The grey dotted lines indicate the performance when a cutoff of 0.5 is used. **B)** The number of targets with a correct highest-ranking model (true positives) plotted against the false positive rate on a logarithmic scale. The grey dotted line indicates the total number of targets for which at least one correct model was produced by SAINT2. **C)** The distribution of TMloss - the difference between the TM-score of the highest-ranking model (top1) or the best of the top five highest ranking model (top5) and the best model in the ensemble - for RFQAmodel and RFQAregression. Results are shown for all targets with at least one model with a TM-score above 0.4. **D)** The distributions of TMloss for RFQAmodel and RFQAregression compared to the component method PcombC.

Table 2.2: Performance of RFQAmodel compared to a version of the model trained to predict the actual TM-score of each model (RFQAregression) on the Validation Set. The results are divided into confidence categories for RFQAmodel, or predicted TMscore range for RFQAregression, as well as for all targets overall (All). The total number of targets (Total) and the number of targets for which there is at least one correct model in the ensemble (Max) are shown for each category. The number of targets for which the highest-ranking model (Top1) or the best of the top five highest-ranking models (Top5) is correct is shown, with the corresponding precision in brackets.

	Confidence	Total	Max	Top1		Top5	
RFQAmodel	High	67	63	52	(77.6%)	60	(89.6%)
	Medium	50	38	21	(42.0%)	30	(60.0%)
	Low	68	36	13	(19.1%)	21	(30.9%)
	Failed	59	5	0	(0.0%)	1	(1.7%)
	All	244	142	86	(35.2%)	112	(45.9%)
RFQAregression	≥ 0.5	57	54	46	(80.7%)	52	(91.2%)
	(0.3, 0.5)	171	88	43	(25.1%)	65	(38.0%)
	(0.1, 0.3)	16	0	0	(0.0%)	0	(0.0%)
	All	244	142	89	(36.5%)	117	(48.0%)

to classify models for the 57 CASP12 and 72 CASP13 Quality Assessment targets (see Methods). We used the stage2 set: the 150 highest-ranking models per target selected from the server predictions, with up to five models contributed by 93 different methods. The targets are not divided into constituent domains for the evaluation of quality assessment methods in CASP. As RFQAmodel is designed to assess the output of template-free protein structure prediction protocols as correct or incorrect, here, we only evaluate its performance on the 33 CASP12 and 34 CASP13 targets containing domains classified as free-modelling targets. RFQAmodel performs well on models of the easier template-based modelling targets, which tend to be globally more accurate (Table 2.3).

We used RFQAmodel, trained on the SAINT2 Training set, to classify models in the CASP12 and CASP13 sets as either correct or incorrect. Of the 67 free-modelling targets, 47 targets had at least one correct model. When classified using RFQAmodel, 31 targets had a high confidence highest-ranking model, of which 21 were correct (68% precision, 31% recall).

To assess the performance against other quality assessment techniques, we compared RFQAmodel to the predictions submitted to CASP13 for free-modelling

Table 2.3: RFQAmodel performance for all CASP12 and CASP13 free-modelling and template-based modelling targets.
Targets are grouped into free-modelling (FM) and template-based modelling (TBM) targets. For each RFQAmodel confidence category, we report: the total number of targets (Total); the number of targets for which there was a correct model among the 500 models (Max); and the number of targets for which the highest-ranking model (Top1) or best of the top five highest-ranking models (Top5) is correct with the corresponding precision in brackets.

	Type	Confidence	Total	Max	Top1		Top5	
CASP12	FM	High	9	8	5	(55.6%)	7	(77.8%)
		Medium	20	11	6	(30.0%)	6	(30.0%)
		Low	4	3	1	(25.0%)	1	(25.0%)
		All	33	22	12	(36.4%)	14	(42.4%)
	TBM	High	23	23	23	(100.0%)	23	(100.0%)
		Medium	1	1	1	(100.0%)	1	(100.0%)
		All	24	24	24	(100.0%)	24	(100.0%)
CASP13	FM	High	22	21	16	(72.7%)	20	(90.9%)
		Medium	10	4	1	(10.0%)	3	(30.0%)
		Low	2	0	0	(0.0%)	0	(0.0%)
		All	34	25	17	(50.0%)	23	(67.6%)
	TBM	High	36	35	31	(86.1%)	33	(91.7%)
		Medium	1	1	1	(100.0%)	1	(100.0%)
		All	37	36	32	(86.5%)	34	(91.9%)

targets. These blind predictions were submitted between May and July 2018, and made publicly available in December 2018. We find that RFQAmodel performs similarly to the top performing methods at classifying individual models and the highest-ranking model as correct or incorrect (Figure 2.9).

2.3.7 Iterative model generation and quality assessment

The optimal number of models to generate using SAINT2 is 10,000, but RFQAmodel may enable us to focus our computational efforts more efficiently by identifying the targets for which fewer models are sufficient to generate good models. It may be possible to improve modelling results by iteratively generating more models for the predicted modelling failures and applying RFQAmodel until modelling is predicted to have succeeded with the required confidence.

In order to assess this application, we chose six targets for which RFQAmodel

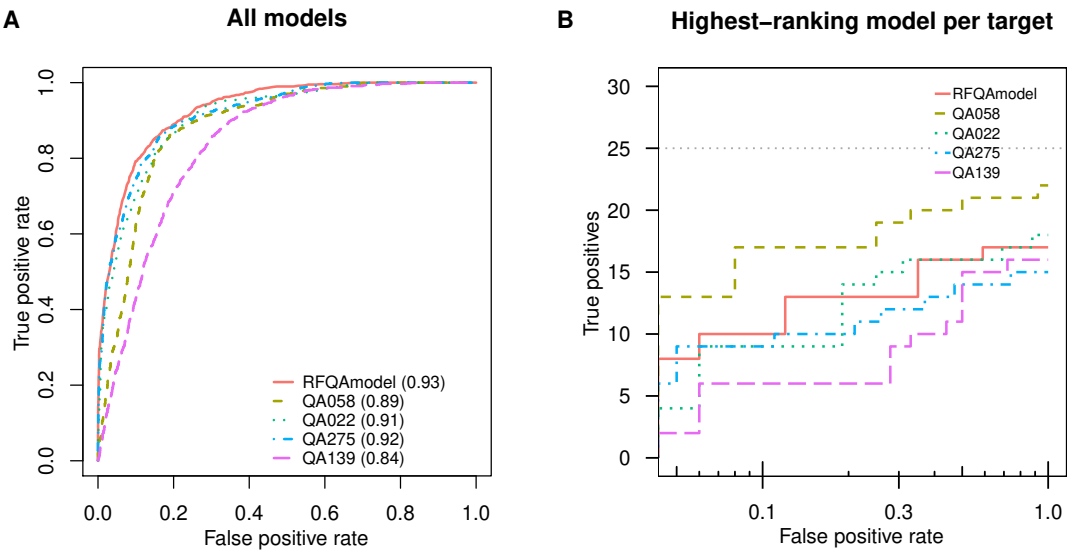

Figure 2.9: RFQAmodel classification of CASP13 free-modelling targets. **A)** Receiver Operating Characteristic (ROC) Curves for the classification of all models into whether they were correct (TM-score ≥ 0.5) or incorrect according to RFQAmodel and four quality assessment scores submitted for the 34 free-modelling targets in the CASP13 set. The area under the ROC curve (AUC) for each method is shown in brackets. **B)** The number of targets with a correct highest-ranking model (true positives) plotted against the false positive rate on a logarithmic scale. The grey dotted line indicates the total number of targets that had at least one correct model.

predicted the highest-ranking model to be correct with low confidence or modelling failures based on the initial 500 models. We then iteratively generated 10,000 models in intervals of 500 models; at each interval we reassessed the model ensemble and compared the TM-score of the best of the top5 highest-ranking models (Figure 2.10). As generating and assessing 10,000 models is computationally expensive, carrying out this analysis on all 244 targets in the Validation set is infeasible.

For one target, 2FZPA, no correct models were generated, and RFQAmodel classified the highest-ranking model as failed or low confidence for all ensemble sizes. For another target, 2CAYB, the confidence increased from low to medium confidence, but a correct model was never identified. For 1IN0A, a high-confidence model was identified once the ensemble size reached 4,000, and this model was correct. Interestingly, if model generation continues, the quality of the highest-ranking models decreases after 6,000 models. For the remaining three targets, RFQAmodel selected better models with higher confidence as the ensemble size

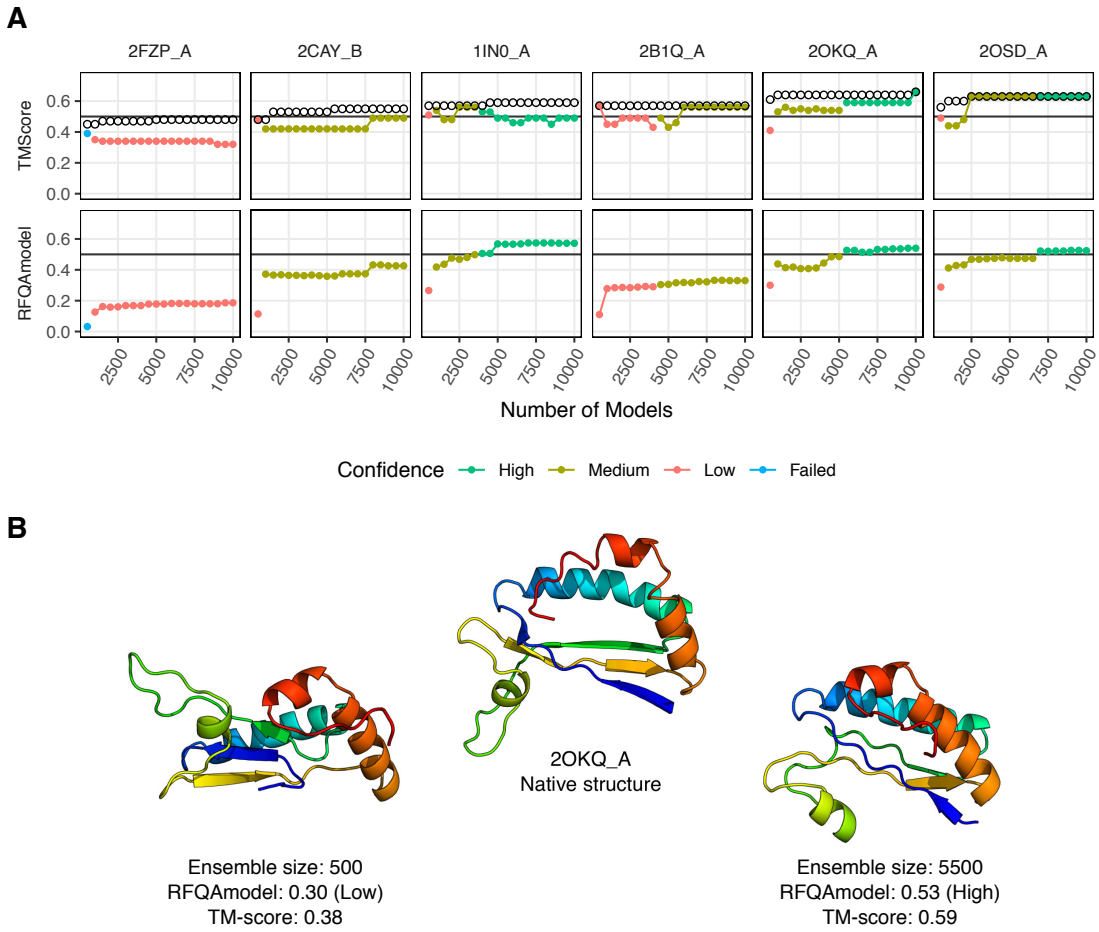

Figure 2.10: RFQAmodel classification improves with ensemble size. **A)** Six targets that were initially classified by RFQAmodel as low confidence or failed were chosen. The number of models generated for each target was increased in increments of 500 from 500 to 10,000. The best model (highest TM-score) is highlighted with a black circle. The TM-score of the best of the top5 highest-ranking models according to RFQAmodel (above) and the highest RFQAmodel score (below) for each ensemble size is indicated with a filled circle, coloured according to the Confidence. **B)** The native structure of 2OKQA (centre) compared to the highest-ranked model according to RFQAmodel after 500 models were generated (left) and after 5,500 models were generated (right), at which point a high confidence RFQAmodel score was achieved.

increased (2B1QA, 2OKQA and 2OSDA). For example, for 2OKQA the highest-ranked model of the initial 500 models had a low-confidence RFQAmodel score of 0.3 (TM-score 0.38). After 1,000 models were generated, the highest-ranked model had a medium-confidence score of 0.44 (TM-score 0.53). Once the ensemble size reaches 5,500 the highest-ranked model had a high-confidence RFQAmodel score of 0.53, and a TM-score of 0.59 (Figure 2.10B). These results demonstrate how RFQAmodel could be used to guide computational efforts and thus and increase

the number of targets for which we have a good predicted structure.

As an example of such a protocol, we iteratively generated and assessed up to 10,000 models for the 50 targets that RFQAmodel predicted to be correct with medium confidence. For the ensembles of 500 models, 38 of these targets had at least one correct model, and 30 had a correct model in the top five highest-ranked models according to RFQAmodel (60% precision). When model generation is continued for each target until at least one high-confidence model is identified, or 10,000 models have been generated, the number of targets with at least one correct model increases to 45 out of 50 targets (Table 2.4). On these ensembles, a further 36 targets are classified by RFQAmodel as correct with high confidence. Of these 36 targets, 28 have a correct model in the five highest-ranked models (78% precision). Using this protocol, the median number of models generated for each target is 3,250. When 10,000 models are generated for all 50 targets, correct models are generated for 46 targets. Of these, 34 are categorised as high confidence, of which correct models are identified in the top five highest-ranked models for 26 (77% precision). Iterative RFQAmodel therefore achieves similar performance to producing 10,000 models for all targets with much less computational expense. This demonstrates that RFQAmodel can be used to improve the efficiency of protein structure prediction, either by identifying a subset of targets for which many models should be generated, or by guiding the termination of model generation.

2.4 Discussion

We find, as have others, that both modelling and quality assessment are more likely to succeed for targets that are shorter, mostly alpha-helical, or have higher B_{eff} values (e.g. Figure 2.3) (de Oliveira, Law, et al., 2018; D. T. Jones, Singh, et al., 2015; de Oliveira, J. Shi, et al., 2017). Previous attempts at estimating quality assessment success have used training and test sets that were not balanced in length

Table 2.4: Performance of iterative RFQAmodel on the 50 targets categorised as medium confidence. Results are shown for the original ensembles of 500 models, the ensembles when models are generated until either a model is categorised as high confidence or 10,000 models have been generated (Iterative RFQAmodel), and for ensembles of 10,000 models for all targets. The total number of targets (Total) as well as the number of targets with at least one correct model is reported for each confidence category (Confidence) and for all targets overall (All). The number of targets for which the highest-ranking model (Top1) or the best of the top five highest-ranking models (Top5) is correct is shown, with the corresponding precision in brackets.

	Confidence	Total	Max	Top1		Top5	
500 models	Medium	50	38	21	(42.0%)	30	(60.0%)
Iterative RFQAmodel	High	36	34	19	(52.8%)	28	(77.8%)
	Medium	14	11	3	(21.4%)	5	(35.7%)
	All	50	45	22	(44.0%)	33	(66.0%)
10000 models	High	34	34	22	(64.7%)	25	(73.5%)
	Medium	16	13	4	(25.0%)	7	(43.8%)
	All	50	47	26	(52.0%)	32	(64.%)0

and number of effective sequences (e.g. Michel, Menéndez Hurtado, et al., 2017), which may result in inconsistent performance when applied to other sets. In order to ensure as accurate an estimate of performance as possible, we designed our Training and Validation sets to be well-balanced in terms of these features.

Using our Training set we built RFQAmodel, which uses the contact map alignment scores EigenTHREADER and map_align in addition to existing quality assessment scores to estimate model quality. For targets with sufficient sequence information, we found that EigenTHREADER identifies correct models for more targets than a number of existing single-model, consensus, and hybrid model quality scores (Figure 2.2). Eight of these targets were not captured by the two other top performing methods, SAINT2 and PcombC. This indicates that predicted contact map alignment scores are, at least to some extent, orthogonal to existing model quality assessment scores.

Unlike many existing quality assessment scores, RFQAmodel was designed to output a score that indicates the likelihood that a model is correct. On our Validation set it identifies, with high confidence, a single correct model for 67 of 244 targets with 78% precision. RFQAmodel outperformed the component quality assessment methods, in agreement with previous studies where combining methods improves

performance (Uziela et al., 2017; Michel, Menéndez Hurtado, et al., 2017; Manavalan et al., 2014). When compared to methods used to identify successfully modelled targets in large-scale protein structure prediction studies (Michel, Menéndez Hurtado, et al., 2017; Ovchinnikov, Park, et al., 2017), RFQAmodel achieved a higher recall and was able to identify successfully modelled targets with fewer correct models in their ensemble. This suggests that by using RFQAmodel it may be possible to identify more modelling successes in large-scale studies.

While RFQAmodel was developed and trained using our template-free protein structure prediction protocol, SAINT2, we assessed its suitability for use with other protocols. We tested RFQAmodel on ensembles of models from a large number of different protocols for 56 CASP12 and CASP13 free-modelling targets. For these models we used the TM-scores calculated by the CASP assessors. RFQAmodel classified the highest-ranking model as correct with high confidence for 38% of targets with 81% precision and 85% recall. While this demonstrates that RFQAmodel can be used to classify models generated by methods other than SAINT2, the performance of RFQAmodel may be improved by training on models from a variety of other protocols.

RFQAmodel was not trained for other quality assessment tasks, such as predicting the absolute quality of models. Furthermore, unlike some methods (including ProQ3D and PCons), RFQAmodel does not estimate the local (per-residue) quality of models. However, at the task of selecting a correct model for each target, we found that it performed comparably to the top-performing methods in CASP13 and outperformed RFQAregression, a version trained to predict the absolute TM-score of each model.

Finally, our protocol is able to reduce the computational cost of protein structure prediction, which is a common limitation for large-scale studies. The assignment of confidence enables us to identify the targets for which 500 models are sufficient to generate good models with high confidence. We can then iteratively generate more models for the medium, low confidence, or failed targets and apply RFQAmodel

until modelling is predicted to have succeeded with high confidence, focussing computational efforts more efficiently.

3

Flib-Flex and predicting missing terminal regions of protein structures

Contents

3.1 Introduction

It is common for a protein target of interest to have only a partial structure due to experimental limitations and/or the lack of an available full-length homologue structure to use as a template for prediction. In particular, terminal regions are routinely removed from expression constructs to improve crystallisation yield in

structural determination by X-ray crystallography (Savitsky et al., 2010). This chapter describes the development of Flib-Flex, a method for incorporating known structural information into the fragment library, and its combination with SAINT2-ScafFold to complete partial protein structures.

While template-free structure prediction has improved dramatically in recent years, template-based modelling (also known as homology modelling) remains the most reliable approach where a homologous structure is available (Croll et al., 2019). However, loop regions are commonly absent from the template structure and may vary more between homologues, presenting a challenge for template-based modelling. A number of techniques have been developed for modelling and refinement of these regions alongside template-based modelling, including the knowledge-based method FREAD (Choi et al., 2009; Deane, 2001), the *ab initio* loop refinement tool of MODELLER (Fiser et al., 2000), and the combination method Sphinx (C. Marks et al., 2017).

Template-based modelling can therefore predict regions without a template structure, but such techniques are generally optimised for relatively short, unstructured loop regions (Li, 2013; Park et al., 2014; Wong et al., 2017; Fiser et al., 2000; C. Marks et al., 2017). Where the missing regions are longer and include secondary structure elements, template-free modelling may be more appropriate.

An alternative is to use a template-free approach to predict the entire structure. Template-free prediction methods may produce higher quality models where a homologous structure is available, for example due to the availability of homologous fragments or a larger number of homologous sequences. However, the challenges of template-free modelling — such as the large conformational space and dependence on co-evolutionary predicted contacts (Section 1.3.10) — still limit the performance compared to template-based modelling where a homologue is available, even for the top-performing deep learning methods (Croll et al., 2019; Senior et al., 2019).

To our knowledge there is no method to perform template-free structure prediction to complete a partially known structure. Such a method would be useful where an incomplete structure has been obtained experimentally or modelled

from a partial-length template. Using existing methods, the unknown region could be modelled separately using template-free methods and hybridised (docked) with the known structure. However, predicting the missing region in the environment of the known structure enables the orientation and interactions between the regions to be considered directly during model generation. This could potentially enable more efficient and effective exploration of conformational space.

SAINT2-ScafFold is an existing implementation of SAINT2 that is able to build models from a known segment of structure (Law, 2017). During prediction, only the missing region and a small section of the known structure – referred to as the 'Flex' region – is sampled during prediction. This method has previously been applied to the completion of partial models of transmembrane targets; models of $\leq$5Å RMSD to the native structure were successfully produced for 29 of 35 single helix predictions (Law, 2017).

In this chapter, we combine RFQAmodel (Chapter 2) and an updated version of SAINT2-ScafFold to predict large missing regions of target structures.

3.1.1 Overview

We first collected a Test set of protein targets for which two similar structures have been solved, with one at least 15 residues shorter than the other. These represent target structures for which a missing region can be predicted against the known shorter structure, emulating the completion of an incomplete X-ray crystallography structure. We then built larger Training and Validation sets based on these Test case examples.

We generated fragment libraries using Flib-Flex, which is able to incorporate structural information from the shorter structure, and improves the quality of the fragment libraries for the regions corresponding to the known structure. These Flib-Flex libraries can be used for prediction of the full-length target using SAINT2, or for completing only the missing region using SAINT2-ScafFold.

We used these fragment libraries to generate 500 models for each target structure, completing the missing region using SAINT2-ScafFold. Globally and locally correct

models are produced for 76% and 77% of Validation set targets, and 47% globally and 37% locally correct models are produced for the Test set targets. In addition, we find that predicting the missing region in the presence of the known region results in better models than when this region is predicted in isolation.

Finally, we attempt to identify locally and globally correct models using a modified version of RFQAmodel (Chapter 2) that is trained on the Training set targets. For the Validation set, the highest-ranking model is predicted to be globally correct for 77 of 171 targets, with 94% precision. When applied to the real cases in the Test set, the highest-ranking model was predicted to be globally correct for 23 of 80 targets, with 65% precision.

3.2 Methods

3.2.1 Datasets

The datasets were set up to emulate a scenario in which the target has a solved crystal structure, but a terminal region is unresolved or was excluded from the construct.

Training and Validation sets

In order to conserve all the real cases (see Section 3.2.1: Test sets) for testing, we created a set of targets with "missing regions" to use as Training and Validation sets based on the Training and Validation sets described in Section 2.2.1. All targets with fewer than 85 residues were excluded; this ensured a segment length of at least 50 residues when considering a minimum of 15 residues of missing residues and 15 residues of flexible region (see Figure 3.1). This resulted in 198 Training and 171 Validation set targets.

To simulate missing regions for the Training and Validation sets, we attempted to emulate the properties of the real cases in the Test set, taking into account both the length and proportion of the missing regions (see Section 3.2.1). The

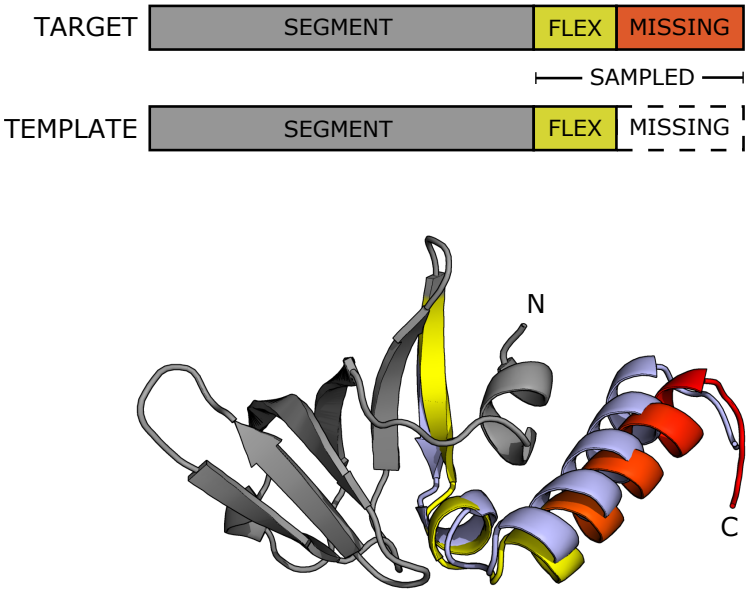

Figure 3.1: A demonstration of a pair of structures and the terminology used for each region.
The missing region (orange) is defined by the terminal region that is absent in the shorter template structure (indicated with dashed lines). The template structure consists of either a truncated version of the target crystal structure (Crystal Structure Test Set, Section 3.2.1) or the shorter homologous structure (Homology Model Test Set, Section 3.2.1). The template structure is used to build the corresponding segment and flex regions of the fragment library for the target structure. During prediction, the flex and missing region of the target structure are sampled. An example model is shown (PDB code: 1TS9A). The sampled flex (yellow) and missing (red) regions have a global RMSD of 2.7Å and local RMSD of 1.7Å to the equivalent region of the native structure (blue), once the segment (grey) region is superimposed.

missing regions of the Test set targets are never more than 25% of the full-length domain (Figure 3.2D), and there is some correlation between the lengths of the domain and the missing region (Figure 3.2C). The Test set targets are typically longer overall than the Training and Validation set targets, which have a maximum length of 250 residues (Figure 3.2C).

For each target in the Training and Validation of up to 200 residues in length, a simulated missing region length was sampled from the distribution of real missing region lengths for the Test set targets of 250 residues or fewer. For Training and Validation set targets longer than 200 residues, the simulated missing region lengths were sampled from the distribution for the Test set targets with more than 250

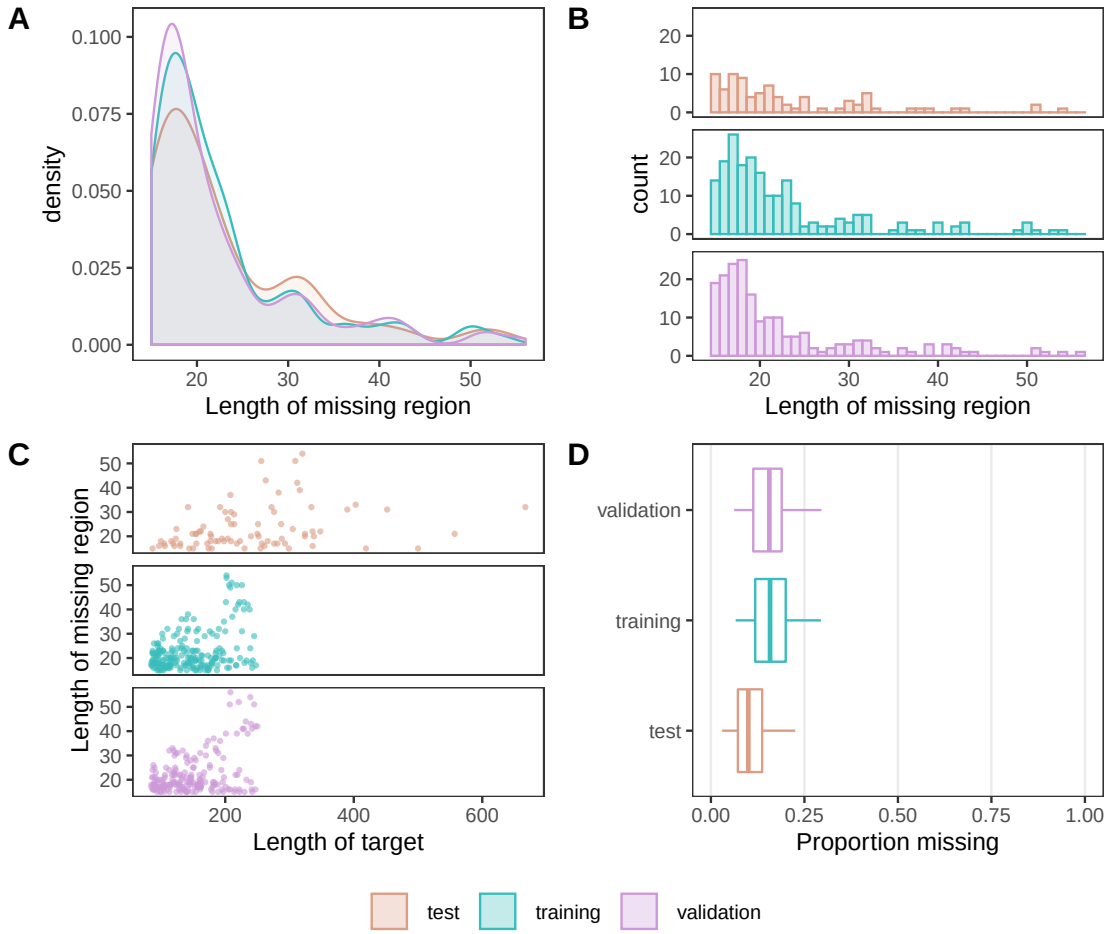

Figure 3.2: Properties of the missing regions for targets in the data sets. **A)** Density and **B)** histogram distributions for the lengths of the missing regions for targets in the Training, Validation and Test sets. **C)** The overall target length and missing region length for each target in these data sets. **D)** Boxplots showing the proportions of the missing region compared to the overall length of the targets in the data sets.

residues, while ensuring no missing region is more than 30% of the total length. The resulting sets have similar distributions of missing region lengths, but the missing regions make up a larger proportion of the entire structure for the Training and Validation set targets when compared to the Test set (Figure 3.2).

Test sets

We collected 80 pairs of homologous structures that differ in length by at least 15 residues at one terminus, and used these to build two test sets. These cases are examples of terminal regions that can be absent from a crystal structure but are still capable of forming a stable conformation. This extra terminal region

- the "missing region" - can therefore be predicted based on one structure and validated against the other (Figure 3.1).

The Test set pairs were selected from the CATH domain database full release (Sillitoe et al., 2015) (v4.2.0 , based on the 17-05-2017 release of the PDB, Cock et al., 2009). The sequence information for all 434,873 CATH domains was downloaded from the CATH website. All domains with a resolution worse than 2.5Å were excluded, resulting in 262,061 domains. The domains were grouped into superfamilies based on their S60 cluster (in which domains are clustered using a 60% sequence identity cutoff), resulting in 33,025 clusters. For each S60 superfamily with more than one domain (25,859), a multiple sequence alignment was performed using MUSCLE (Edgar, 2004) to identify alignments containing terminal gaps of at least 15 residues (1,064). Pairwise alignments using EMBOSS needle (Rice et al., 2000) were performed on all domains within these S60 superfamilies to identify pairs of domains with terminal gaps of 15-150 residues. The resulting alignments consisted of 14,123 unique CATH domains, the structures of which were downloaded and the coordinates and sequences of the corresponding regions were extracted. His-tags were removed (360); these have been found to have little effect on the structure (Carson et al., 2007). Confirmation by needle alignment of the extracted sequences resulted in 13,224 C-terminal and 11,405 N-terminal pairs, representing 270 and 239 S60 superfamilies, respectively. From these, we selected pairs in which both domains were: single-domain chains (domain identifier is "00"), monomeric (according to the PDB stoichiometry annotation) and had no gap in the chain exceeding nine residues. The resulting 1,274 C-terminal and 798 N-terminal pairs represented 41 and 43 superfamilies, respectively. Finally, for each terminus, for each S60 superfamily, we selected at random a pair with the shortest maximum gap in the crystal structure. Four pairs were discarded due to duplication of a target in both the N- and C-terminal sets or due to problems with the structures.

In the CATH database, modified residues are marked as "X" in the sequence files and removed from the coordinate files. These were identified from the original PDB

file and modelled as the corresponding standard amino acid using the homology modelling software MODELLER v9.19 (Šali et al., 1993; Fiser et al., 2000).

The missing regions of the resulting 80 pairs are of a maximum length of 87 for N-terminal and 53 for C-terminal cases. The sequence identity between each pair ranged from 59% to 100%, with a median of 98%. The backbone RMSD between the pairs ranged from 0.18Å to 3.02Å, with a median of 0.95Å.

Crystal Structure Test Set

We created two sets based on this set of 80 pairs of structures. For the Crystal Structure Test Set, the longer structure in each pair was truncated to the length of the shorter structure, and used as the template. This emulates a scenario in which an incomplete crystal structure is available.

Homology Model Test Set

We also evaluated whether the solved crystal structure of a shorter homologue can be used as the starting point to predict a longer target structure. For this case, the template is a homology model generated using the shorter structure of each homologue pair. This emulates a scenario in which the best available template does not cover the entire target length. This set was not used for protein structure prediction, but was used to evaluate the quality of the Flib-Flex fragment library when a homology model is used rather than an incomplete crystal structure.

All homology models were generated using MODELLER v9.19 (Šali et al., 1993; Fiser et al., 2000). The template and target sequences were aligned using needle (Rice et al., 2000), and this alignment was used to generate 10 homology models of the target based on the template structure. The model with the highest TM-score against the template structure was selected as the model structure.

3.2.2 Fragment library generation with Flib-Flex

Flib-Flex is a modified version of the fragment library generation software, Flib-Coevo (de Oliveira, J. Shi, et al., 2015; de Oliveira and Deane, 2018), that enables

information about the known region of the target structure to be incorporated into a fragment library.

Flib-Coevo

As described in Section 1.4.2, during fragment library generation using Flib-Coevo, fragments are extracted from the template database according to their score in comparison to predicted secondary structure and torsion angles (Fragment extraction, Figure 3.3). Fragments that do not satisfy intra-fragment predicted contacts are filtered out (Flib-Coevo, Figure 3.3). Candidate fragments are then ranked for inclusion in the fragment library according to their predicted torsion angle score. The library is enriched with fragments with low RMSD to high-scoring fragments (Enrichment, Figure 3.3), and with 9-residue fragments generated exhaustively from protein threading hits (Threading, Figure 3.3). For a more realistic approximation of template-free protein structure prediction, fragments from homologous structure are excluded from the fragment library.

Flib-Flex

Flib-Flex additionally takes into account the known part of the target structure (Template, segment and flex, Figure 3.1). The modifications to the Flib-Coevo protocol are illustrated in pink in Figure 3.3. Predicted contacts within the template that are incorrect are removed; all predicted contacts involving the missing region are kept. During fragment library generation, the true secondary structure and torsion angles from the template structure (derived using DSSP) are used instead of the predicted values for the corresponding regions of the target, with the predicted values used for the remaining regions. Fragments from homologous structures are only excluded from the missing region of the fragment library. Candidate fragments for the known regions are ranked for inclusion according to their RMSD to the template structure, rather than the predicted torsion angle score. Where there are at least the minimum of 20 fragments for a given position, fragments with an RMSD exceeding 1.5Å are excluded.

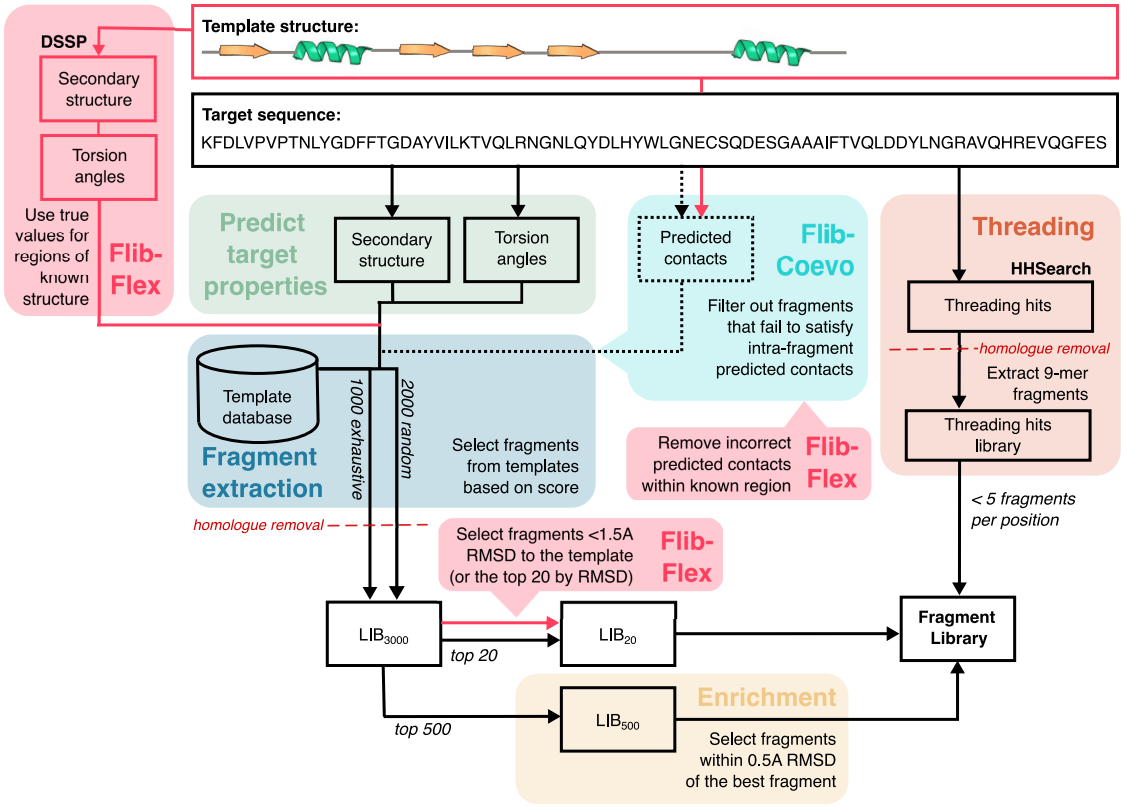

Figure 3.3: Diagram of fragment library generation using Flib-Flex. The secondary structure and torsion angles are predicted from the target sequence (green). These are used to score fragments extracted from the template database (dark blue). For the Flib-Coevo protocol, predicted contacts are generated from the target sequence, and fragments that do not satisfy intra-fragment predicted contacts are filtered out (light blue, dotted lines). An average of 3,000 fragments per position are extracted (LIB_{3000}). The 20 fragments with the highest predicted torsion angle score are selected (LIB_{20}), and enriched with fragments among the top 500 (LIB_{500}) that are structurally similar to the top-scoring fragment (yellow), and fragments derived from protein threading hits (orange), to comprise the final fragment library. To emulate a real template-free prediction scenario, fragments derived from structures with sequence homology to the target are removed (red dashed lines). For the Flib-Flex protocol (pink), information is incorporated from regions with known structure (Template structure). Incorrect predicted contacts within the known region are removed, and the true secondary structure and torsion angles (derived using DSSP, Kabsch et al., 1983) are used to score fragments during extraction. Fragments are ranked for inclusion by the RMSD to the known structure; fragments with > 1.5Å RMSD are excluded, unless required to ensure a minimum of 20 fragments per position. Fragments from homologous structures are not removed from the fragments in the known region.

The incorporation of a range of similar fragments, rather than only fragments derived from the template structure, allows some flexibility during sampling. This may be important to accommodate the regions of unknown structure to be predicted, and because the template may be a homologous structure.

For the missing regions, as well as any positions for which fragments overlap with the missing region, the standard behaviour of Flib-Coevo applies. While in this work the missing regions are always terminal, Flib-Flex can handle missing regions anywhere in the target sequence. Although a full fragment library is generated for the target, when it is used with SAINT2-ScafFold the only regions of the fragment library that are used for sampling are the missing region and a 15-residue "Flex" region of the known structure (see Figure 3.1 and Section 3.2.3, below).

We generated the predicted secondary structure using DeepCNF (Q8 predictions) (S. Wang et al., 2016), torsion angles using SPIDER3 (Heffernan et al., 2017), and contacts using metaPSICOV (stage1 predictions) (D. T. Jones, Singh, et al., 2015). These inputs were used to generate fragment libraries using Flib-Flex, with an RMSD cut-off of 1.5Å. For the Training and Validation sets, we compared these fragment libraries to the fragment libraries described in Section 2.2.2, which were generated with Flib-Coevo rather than Flib-Flex.

3.2.3 Missing region prediction with SAINT2-ScafFold

SAINT2-ScafFold is an implementation of SAINT2 in which a segment of a target protein is provided as a starting point from which the remainder of the structure is grown (see Section 1.4.1). We ran SAINT2-ScafFold on the missing regions of the target structures in the Training set, Validation set, and the Crystal Structure Test set. Conformational sampling is restricted to the missing region, as well as a 15-residue "Flex" region immediately preceding it, which allows some flexibility in the known segment and may help accommodate the missing region (Figure 3.1).

For the Crystal Structure Test set targets, the template structure was created by removing the missing region from the native structure of the longer homologue in each pair, which was then used as the segment for SAINT2-ScafFold. To ensure

there were no gaps in the segment structure, complete models were generated with Modeller, using the crystal structure coordinates as a template. For the Training and Validation set, we truncated the simulated missing region and used SAINT2-ScafFold to predict this region with the remaining crystal structure as the segment.

3.2.4 Model evaluation

For full-length structures, as in the previous chapter, we evaluate models using TM-score. TM-score was trained on protein structures with a minimum length of 50 residues, and therefore is not applicable to the short sampled regions. Instead, we use another widely used measure of model accuracy, root-mean-square deviation (RMSD) of the backbone atoms.

To assess the local accuracy of the sampled region - which includes both the missing region and the flex region - we calculated Local RMSD, the minimised RMSD against the corresponding region of the native structure. To take into account the orientation relative to the rest of the structure, we additionally calculated the RMSD of the sampled region when the segment (the region where SAINT2 has made no changes) has been superimposed with the known region of the native structure, which we refer to as Global RMSD. We used 2.5Å and 5Å RMSD cutoffs to label models as locally and globally correct, respectively.

3.2.5 Model quality assessment with RFQAscaffold

We modified our random forest model quality assessment framework, RFQAmodel (Chapter 2), to identify successfully predicted models. We trained two classifiers to predict local and global quality separately.

To ensure reproducibility, the random seed was set to 1219 before the training of each model.

Features

We included the features used in RFQAmodel (see Section 2.2.5), excluding mapalign. For the local classifier, the global ProQ3D, ProQRosCenD, ProQRosFAD, ProQ2D and Pcons scores were excluded. In addition, we included:

Flex RMSD: The RMSD of the flexible region against the crystal structure or template structure used as the scaffold.

Local quality scores: the average local Pcons and ProQ3D scores for the sampled residues.

3.3 Results

3.3.1 Properties of the missing regions

We collected examples of targets with a terminal region that was present in one structure and absent in another, which we refer to as the missing region. In collating the Test set, we do not distinguish between cases for which this region was unresolved or was absent from the crystal. These regions may have particular properties, such as high solvent-accessibility, low stability, little secondary structure, or high B-factors. Some of these properties might be common among terminal regions, but they may also be particularly pronounced in these cases.

We compared these regions to the equivalent length region at the opposite end of each target structure ("Opposite"), as well as the remaining region of the structure ("Internal") (Figure 3.4). In addition, we compared to the "missing" regions in the simulated cases of the Training and Validation sets.

Solvent accessibility

We compared the average solvent accessibility of the residues in each region for the Test, Training and Validation sets (Figure 3.4A). Solvent accessibility was calculated using DSSP (Kabsch et al., 1983). As expected, the missing region is

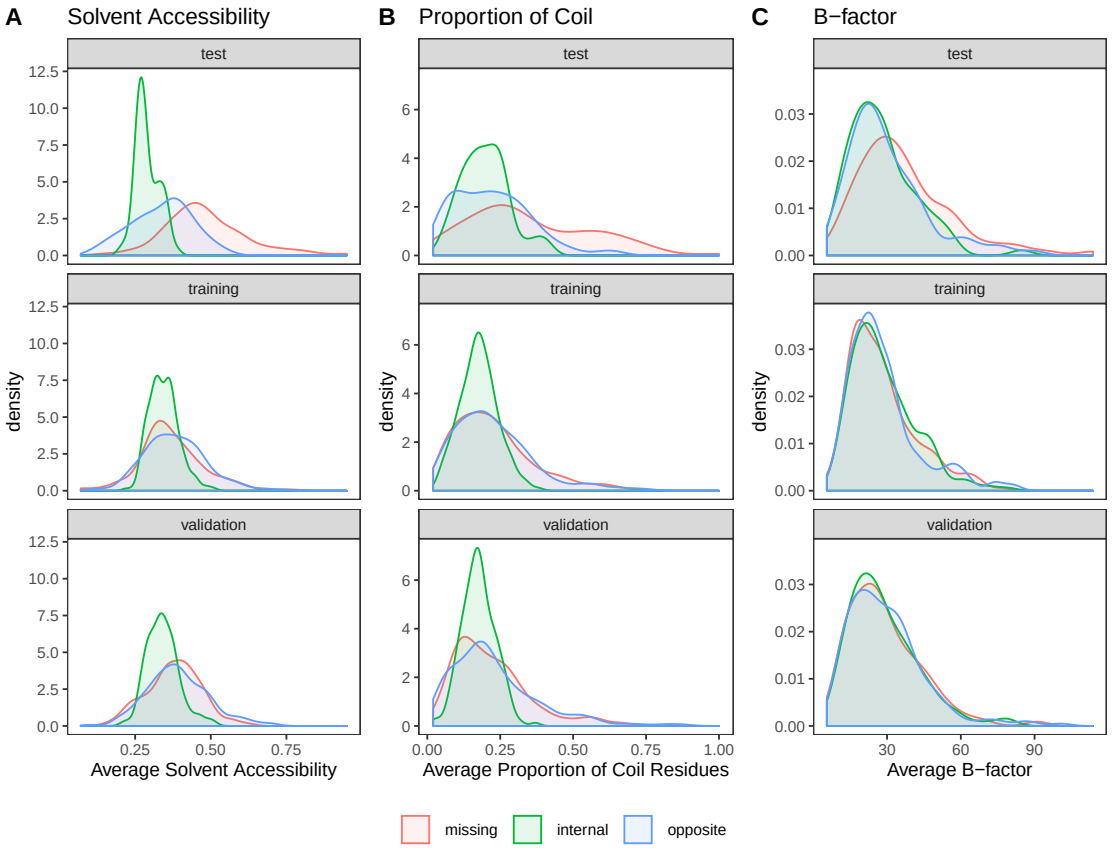

Figure 3.4: Properties of the Missing, Opposite and Internal regions.
The distribution of **A)** average solvent accessibility, **B)** proportion of coil residues, and **C)** average B-factor for residues in the Missing region (red), the equivalent length region at the opposite terminus ("Opposite", blue) and the remaining residues ("Internal", green) for targets in the Test (top), Training (middle) and Validation (bottom) sets.

more solvent-accessible than the rest of the structure for Test set targets. For the Training and Validation sets, both termini show similar increased average solvent accessibility compared to the internal region, but are less accessible than the missing regions of the Test set targets.

It has previously been suggested that the N-terminus may adopt a more compact structure and be more buried than the C-terminus as a result of cotranslational protein folding (Alexandrov, 1993; Deane et al., 2007; Saunders, Mann, et al., 2011); however, we find very little difference between N- and C-terminal regions for the Training and Validation set (Appendix B.1).

Prevalence of coil

We calculated the proportion of coil residues according to the DSSP assignment for the Missing, Opposite, and Internal regions (Figure 3.4B). For the Test set, the missing region is more likely to have little or no secondary structure than the equivalent region at the opposite end. Both termini have a higher proportion of coil than the internal region, which is also the case for the Training and Validation sets. However, at least half the residues of the Missing regions form secondary structure for 72% of the Test set targets.

B-factor

We compared the average B-factor for the residues in the Missing, Opposite and Internal regions (Figure 3.4C). For the Training and Validation sets, all regions have very similar distributions. In the case of the Test set, missing regions have higher B-factor values on average.

3.3.2 Improving fragment library quality for known regions using Flib-Flex

Flib-Flex is a modified version of our fragment library generator, Flib-Coevo, that incorporates information from a known template structure by using the true secondary structure and torsion angles to help select fragments. We used Flib-Flex to generate fragment libraries for all targets in our Training, Validation and Crystal Structure Test sets by using the template region of the target structure as the template (Figure 3.1).

Training and Validation Flib-Flex quality

We compared the template and missing regions of the Flib-Flex fragment libraries to the equivalent regions of standard Flib-Coevo libraries for the Training and Validation set targets (Figure 3.6). The missing regions of the fragment libraries should be very similar, as the standard fragment generation process applies to this region, which includes some stochasticity. An example of the precision of fragment

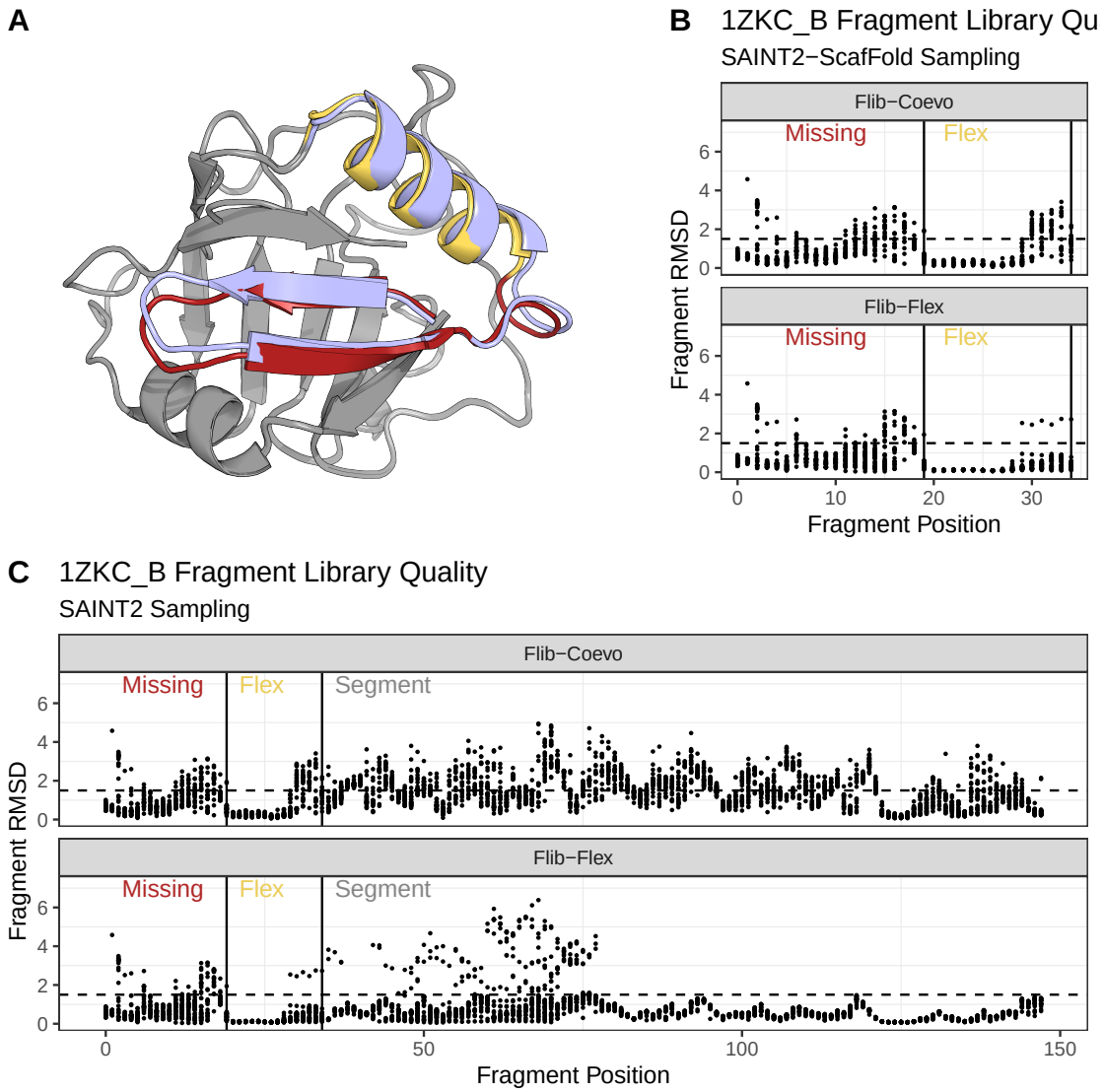

Figure 3.5: Flib-Flex and Flib-Coevo fragment library quality for example target 1ZKCB.
A) The native structure of Training set target 1ZKCB superimposed with the best model generated by SAINT2-ScafFold with the Flib-Flex fragment library. The segment region (grey) is not sampled during prediction. The sampled Flex region (yellow) and Missing region (red) together have a global RMSD of 1.0Å compared to the equivalent region of the native structure (light blue). **B)** The RMSD of each fragment to the native structure for each position in the fragment library, for the library generated using standard Flib-Coevo (top) and Flib-Flex (bottom) for the Training set target 1ZKCB. Fragments of $\leq$1.5 Å RMSD are considered good fragments, indicated with a dashed line. The Missing and Flex regions are indicated with solid vertical lines. **C)** The full Flib-Flex fragment library, including for the Segment region, which is sampled during prediction using SAINT2 but not SAINT2-ScafFold.

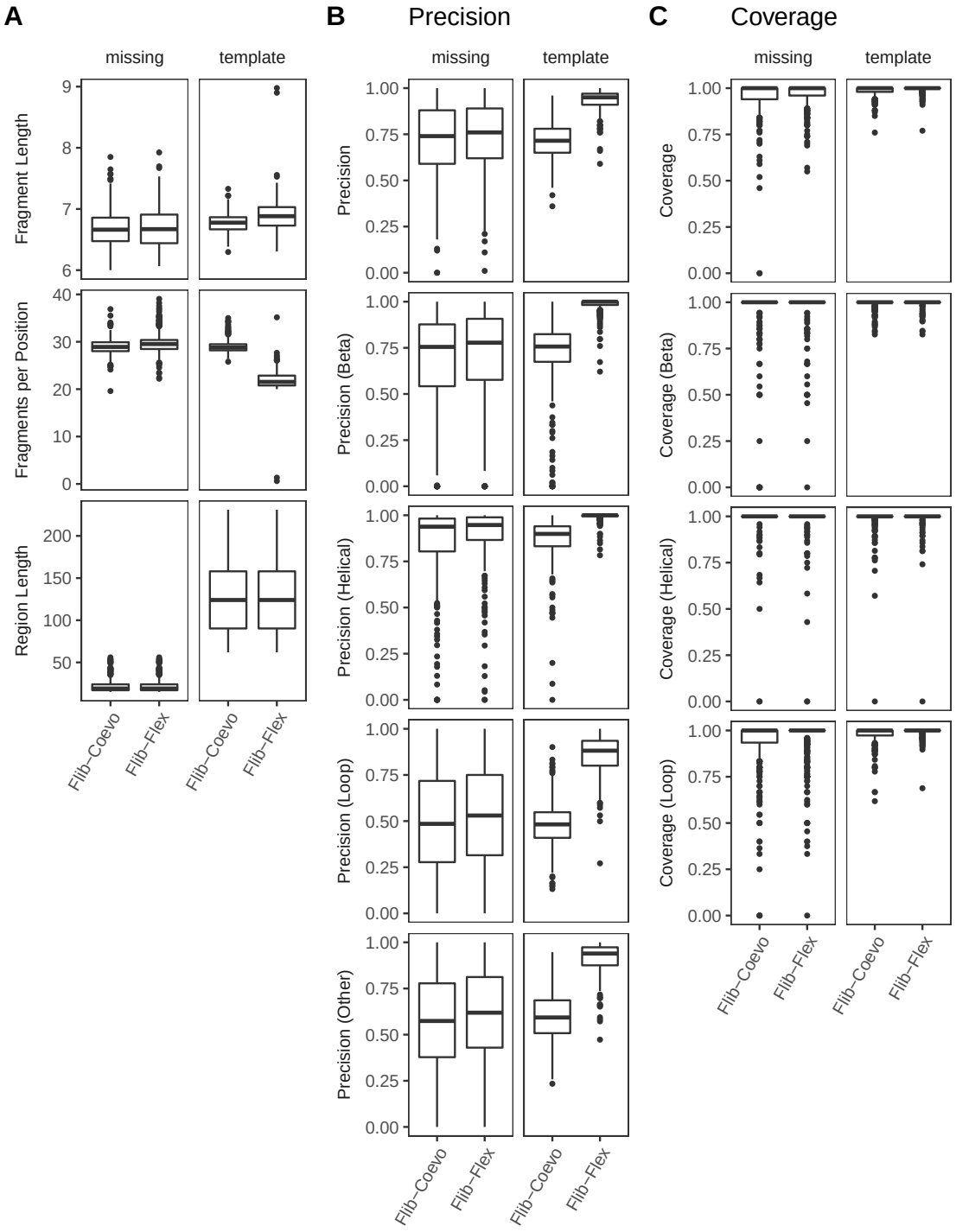

Figure 3.6: Comparison of fragment library quality for the "missing" region (left) and the region corresponding to the template structure (right), for Training and Validation set fragments libraries generated using Flib-Flex or Flib-Coevo.

A) The average fragment length and number of fragments per position, as well as the total lengths of each region. **B)** The overall proportion of fragments of $\leq$ 1.5Å RMSD to the native structure ("Precision") as well as for fragments categorised as predominantly beta, helical, loop, or other. **C)** The overall proportion of positions in the target structure with at least one fragment of $\leq$ 1.5Å to the native structure ("Coverage"), as well as for positions corresponding to beta, helical, or loop secondary structure.

libraries at each position is shown in Figure 3.5. Fragments of ≤1.5Å RMSD to the target structure are considered good fragments. We consider both the proportion of fragments that are good quality (precision), and the proportion of positions that have at least one good fragment (coverage).

For the template region of the fragment library, candidate fragments with >1.5Å RMSD to the template structure are excluded from the Flib-Flex libraries, as long as there are a minimum of 20 fragments per position. Furthermore, unlike the missing region, homologues are not excluded during fragment generation. As a result, the Flib-Flex libraries have fewer fragments per position on average for the template region (Figure 3.6A and B), as well as higher overall precision (Figure 3.6B). The improvement in coverage is less noticeable, as coverage is already very high for the template region of the Flib-Coevo libraries (Figure 3.6B). On average, the template region fragments are slightly longer in the Flib-Flex libraries (Figure 3.6A). These results demonstrate that Flib-Flex has the expected effect of enriching the fragment library with good quality fragments where structural information is known.

Crystal Structure Test Set Flib-Flex quality

The Crystal Structure Test set consists of targets for which a shorter version of the structure has also been deposited in the PDB. We compared the quality of the fragment libraries generated using Flib-Flex for the Crystal Structure Test set and the Training set (Figure 3.7). The missing regions of the Crystal Structure Test set fragment libraries have lower overall precision and coverage compared to the Training set targets. This suggests that SAINT2 may produce better quality models for the missing regions of the simulated Training and Validation set targets, compared to the Test set targets.

Homology Model Test Set Flib-Flex quality

We also generated Flib-Flex fragment libraries for the Homology Model Test set, for which a homology model is used for the template, rather than the native structure. We compared how these different template structures affect the quality of the

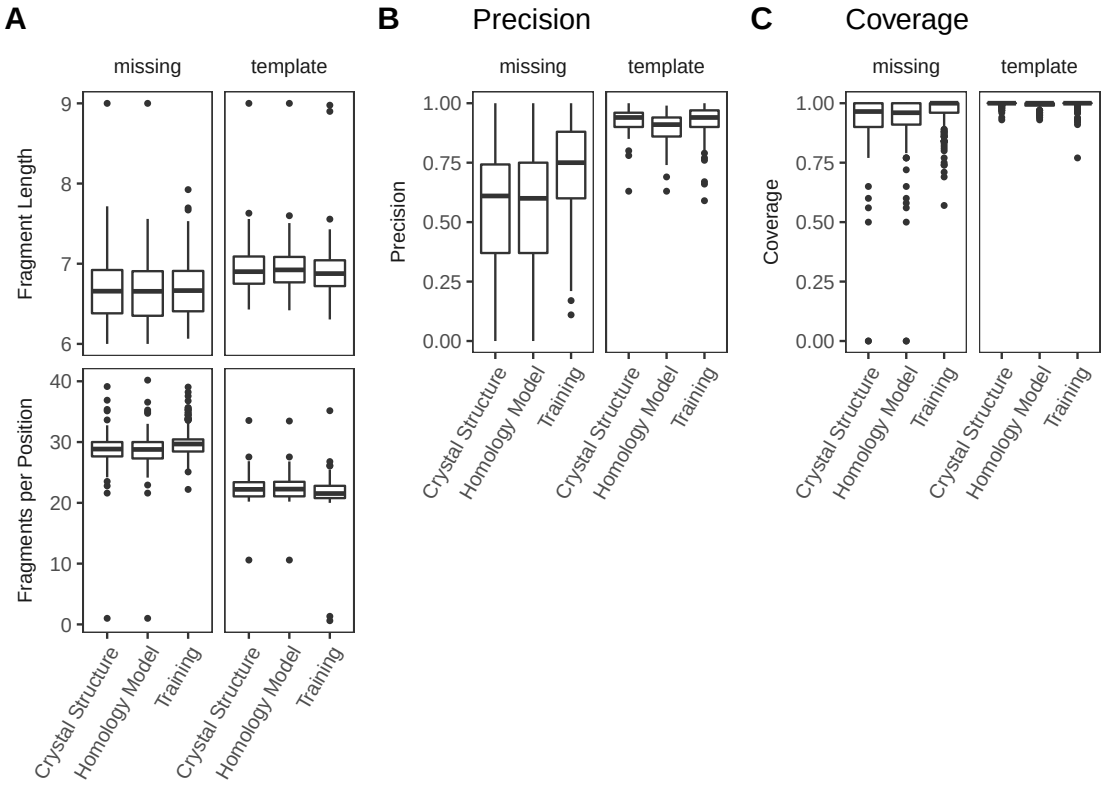

Figure 3.7: Fragment library quality for the missing regions for Crystal Structure Test set, Homology Model Test and Training set targets.
Boxplots showing the distribution of **A)** the average fragment length and number of fragments per position, **B)** the number of targets that are ≤1.5Å RMSD (Precision), and **C)** the proportion of positions with at least one fragment with ≤ 1.5Å RMSD (Coverage), for the portions of the fragment libraries corresponding to the missing regions (left) and template (right). Results are shown for the Crystal Structure Test set and the Homology Model Test set, as well as for the Training set.

resulting fragment libraries (Figure 3.7). While using the homology model has a small effect on the precision of the template region, the effect on the quality of the missing region is minimal. This suggests that while the homology model may not accurately reflect the native structure for all cases, it may have only a small effect when predicting the missing structure using SAINT2-ScafFold, in which only the missing and flex regions are sampled.

3.3.3 Protein structure prediction of missing regions using SAINT2 or SAINT2-ScafFold

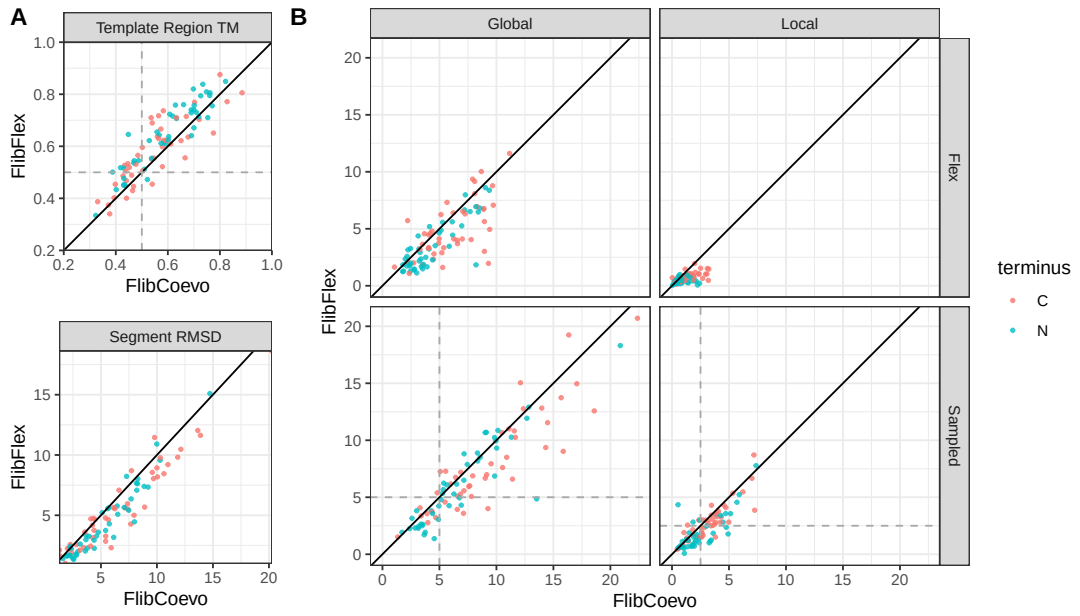

Figure 3.8: Performance of SAINT2 on the full-length target structure using Flib-Flex fragment libraries for 91 Training set targets.
A) The TM-score of the template region of the target structure (top) and the RMSD of the segment region (bottom) for the best model produced by SAINT2 Forward on the full-length target, when either Flib-Flex or Flib-Coevo fragment libraries are used. Models with a TM-score ≥ 0.5 are considered to be in the correct overall fold, indicated with dashed grey lines. **B)** The best global and local RMSD scores for the Flex region (top) and the sampled region (bottom) of the models produced using SAINT2 Forward on the full length structure with either a Flib-Flex or Flib-Coevo fragment library. The 2.5Å local RMSD and 5Å global RMSD cutoffs are indicated with dashed grey lines. Targets are coloured according to the terminus of the missing region.

SAINT2 on the full-length target

Using Flib-Flex, a fragment library is created for the full-length of the target, in which the fragments up to the start of the missing region are closer to the template structure than when using standard Flib-Coevo (Figure 3.8C). Although only the missing and flex regions are sampled using SAINT2-ScafFold, an alternative method is to predict the full-length target using SAINT2 with this Flib-Flex fragment library. To evaluate the improvement gained from using the Flib-Flex library compared to the Flib-Coevo library, we generated 500 models for 91 targets randomly selected from the Training set using SAINT2 in the Forward direction.

While the best models produced for each target using Flib-Flex generally had large improvements in TM-scores compared to those achieved using the Flib-Coevo fragment libraries, the improvement is not sufficiently high to compete with a

homology model or experimentally derived structure (Figure 3.8A). Nevertheless, producing higher quality models of the known region (Figure 3.8A) appears to result in better structures for the missing region (Figure 3.8B, Sampled region).

Predicting the missing region with SAINT2-ScafFold

SAINT2 does not produce sufficiently high-quality models of the region with known structure; it would therefore be useful to directly incorporate the template structure when predicting the missing region. We used SAINT2-ScafFold to generate 500 models completing the missing region for each target structure in the Training, Validation and Crystal Structure Test sets. We consider models with ≤ 2.5Å local and ≤ 5Å global RMSD to the sampled region of the native structure to be correct. While many globally and locally correct models were produced for the Training set targets, fewer correct models were produced for the Crystal Structure Test set targets, with models more likely to be locally than globally correct (Figure 3.9A). At least one globally or locally correct model was produced 76% or 64% of Training set targets, respectively, compared to 47% and 37% of Crystal Structure Test set targets (Figure 3.9B). Targets with longer sampled regions tend to have higher best local and global RMSD values (Figure 3.9C). While longer targets are usually more difficult to predict, this may also reflect that RMSD is sensitive to protein length. The effect is more noticeable for global RMSD, and is most pronounced for the Crystal Structure Test set, as longer missing regions take up a greater proportion of this set (Figure 3.2). These results indicate that the missing region can be accurately modelled for many targets using SAINT2-ScafFold, although performance is lower for the more difficult Crystal Structure Test set targets.

Predicting the missing region without SAINT2-ScafFold

SAINT2-ScafFold enables us to predict the structure directly onto the known structure. Without such a protocol, the missing region would have to be predicted separately. The flexible region could then be used to hybridise the predicted region with the known structure.

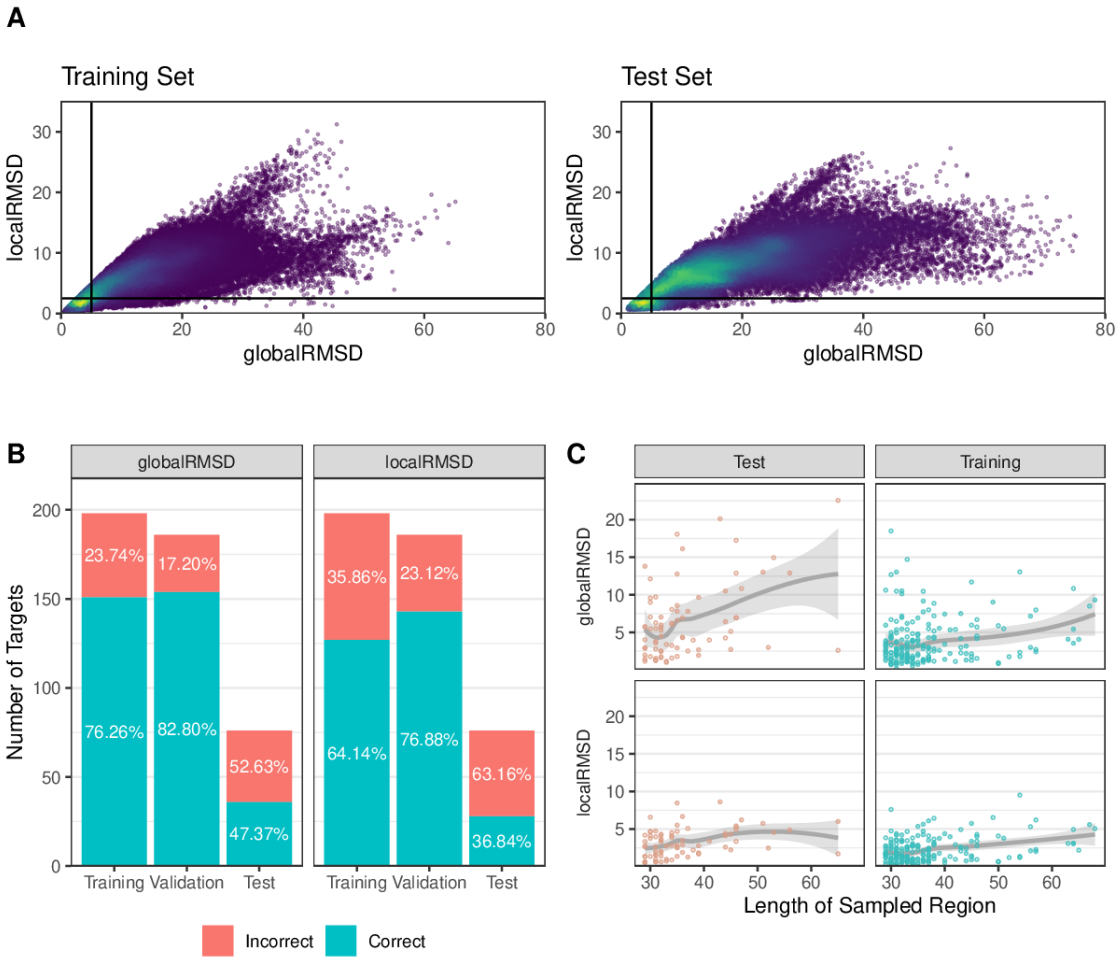

Figure 3.9: Performance of SAINT2-ScafFold for on the Training, Validation and Test sets.
A) The global and local RMSD compared to the native structure for all models produced for the Training (left, 98,973 models) and Test set (right, 38,475 models) targets. Regions with higher density are shown in lighter colour. The 2.5Å and 5Å RMSD cutoffs for a locally and globally correct model are indicated with black solid lines. **B)** The number of targets for which at least one correct model was produced for each set, when considering global RMSD (left) or local RMSD (right). **C)** The best global RMSD (top) and local RMSD (bottom) model for each target in the Test (left) and Training sets (right) compared to the length of the sampled region. Loess (locally estimated scatterplot smoothing) curves are shown in grey, with shaded 95% confidence intervals.

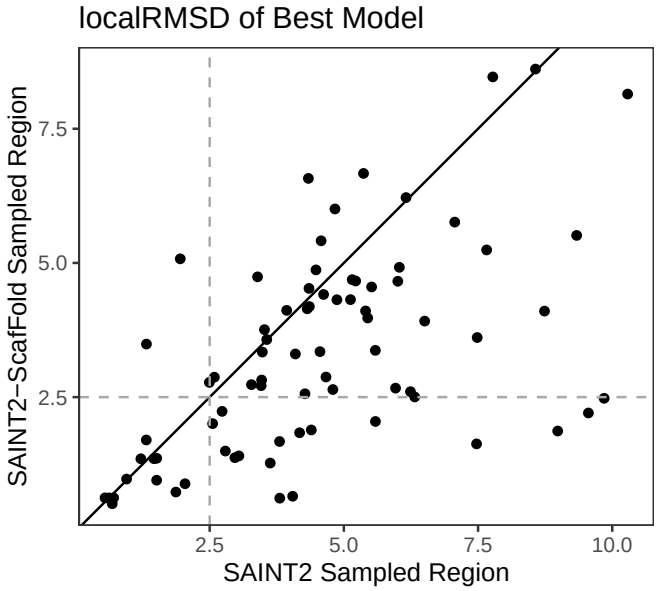

Figure 3.10: Performance of SAINT2 compared to SAINT2-ScafFold for the prediction of the missing regions of the Test set targets
The local RMSD of the sampled region of the best model produced by SAINT2 (on the sampled region alone) or using SAINT2-ScafFold (in the presence of the known structure). Targets below the solid black line had a more accurate best model using SAINT-ScafFold than using SAINT2 of the sampled region alone. The 2.5Å RMSD cutoffs for a locally correct model are indicated with grey dashed lines.

We generated 500 models of only the sampled regions of the Test set targets using SAINT2. We found that the local RMSD of the best models generated using SAINT2-ScafFold were generally lower than the best local RMSD achieved by SAINT2 without the known structure present (Figure 3.10). SAINT2 achieved a local RMSD of ≤ 2.5 for only three targets that SAINT2-ScafFold was unable to predict. Conversely, SAINT2-ScafFold generated correct models for 16 targets that were not successfully predicted using SAINT2 alone. This indicates that the presence of the known region of the structure improves prediction of the missing region.

3.3.4 Model quality assessment of predicted missing regions using RFQAlocal and RFQAglobal

When 500 models were generated using SAINT2-ScafFold for each of the 171 targets in the Validation set, at least one locally or globally correct model was produced

for 132 (78%) and 140 (82%) targets, respectively. To identify correct models for these partially sampled structures, we trained two modified versions of our random forest classifier on the 198 structures in our Training set. Like RFQAmodel (Chapter 2), these classifiers assess each model and output a score, between 0 and 1, that the model is correct.

Two versions of the model were trained: RFQAlocal assesses the local quality of the model, using a local RMSD of $\leq$ 2.5Å to label models as correct, while RFQAglobal assesses the global quality, using a global RMSD of $\leq$ 5Å to label correct models, and includes additional global quality assessment features (see Section 3.2.5). The estimated relative importance of each feature is shown in Appendix Figure B.2.

We assessed the performance of RFQAlocal and RFQAglobal on the Validation set targets using a Receiver Operating Characteristic (ROC) curve. While these classifiers perform similarly, each slightly outperforms the other at the task for which each was trained (Figure 3.11). For classifying all individual models as locally correct or incorrect, RFQAlocal and RFQAglobal achieved an area under the curve (AUC) of 0.9 and 0.85, respectively (Figure 3.11A and C). The performance was more similar for classifying models as globally correct or incorrect, with an AUC of 0.95 for RFQAglobal compared to 0.94 for RFQAlocal (Figure 3.11A and C). Both models outperform the best component method - the average local ProQ3D score of the sampled residues - which achieved an AUC of 0.79 for local assessment and 0.87 for global assessment (Figure 3.11A and C). For the task of classifying the single highest-ranking model per target as correct or incorrect, RFQAlocal and RFQAglobal were also able to correctly classify models for more targets than sProQ3D, when considering false positive rates above 0.03 (Figure 3.11B and D).

We divided the scores output by RFQAlocal and RFQAglobal into the same four broad categories used by RFQAmodel: correct with high ($\geq$ 0.5), medium (between 0.3 and 0.5), or low (between 0.1 and 0.3) confidence, or predicted modelling failures ($\leq$ 0.1) (Figure 3.12A). For each classifier, the models for a given target were ranked according to the output score, and targets were categorised based on the score of

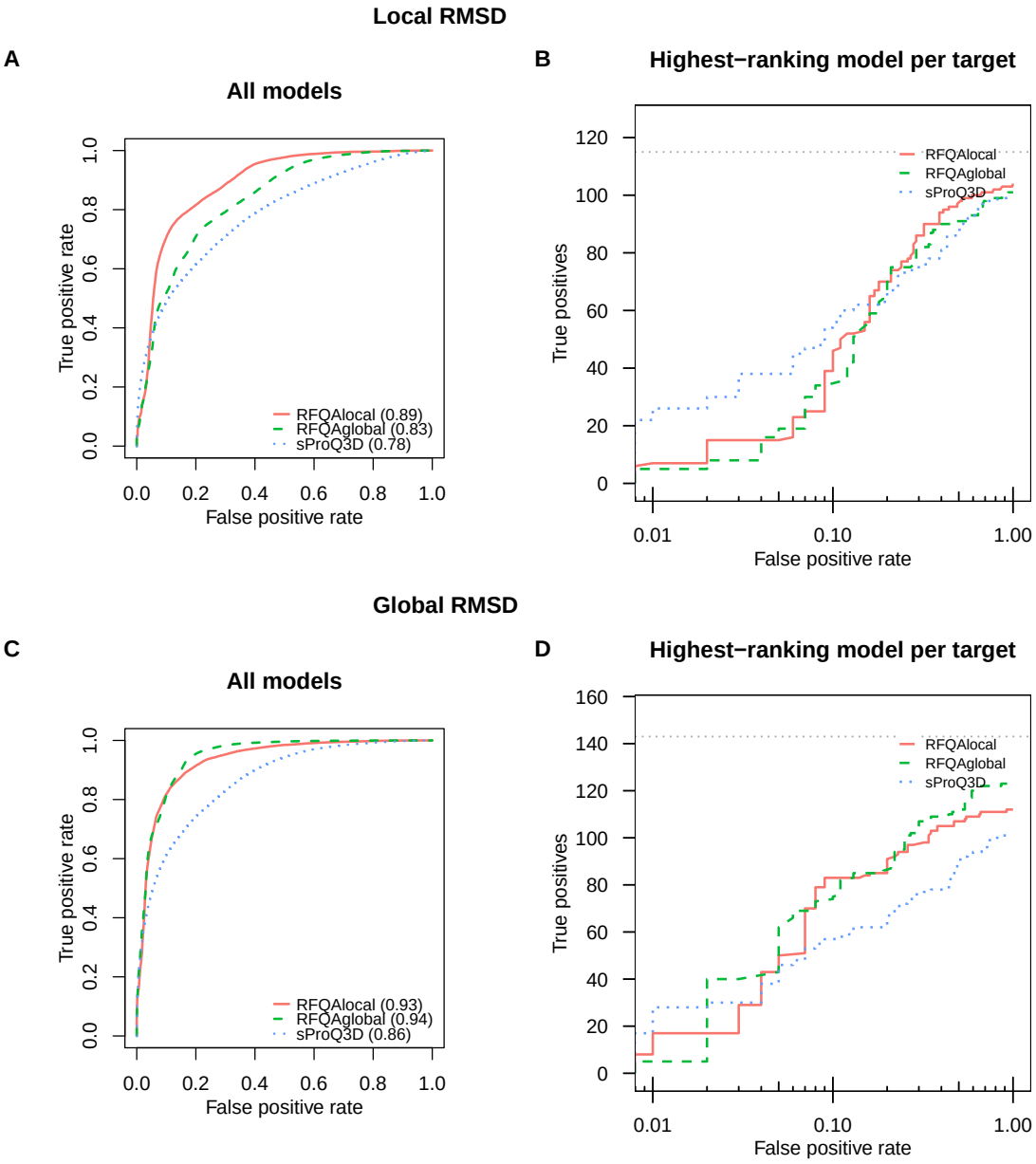

Figure 3.11: Classification of Validation set targets.
A,C) Receiver Operating Characteristic (ROC) Curves for the classification of all models into whether they were **A)** locally correct (local RMSD $\leq$ 2.5Å) or incorrect, or **C)** globally correct (global RMSD $\leq$ 5Å) or incorrect, according to RFQAlocal, RFQAglobal, and the average local ProQ3D score of the sampled residues (sProQ3D). The area under the ROC curve (AUC) for each method is shown in brackets. **B,D)** The number of targets with a **B)** locally or **D)** globally correct highest-ranking model (true positives) plotted against the false positive rate on a logarithmic scale, for the 171 targets in our Validation set. The grey dotted line indicates the total number of targets that had at least one correct model.

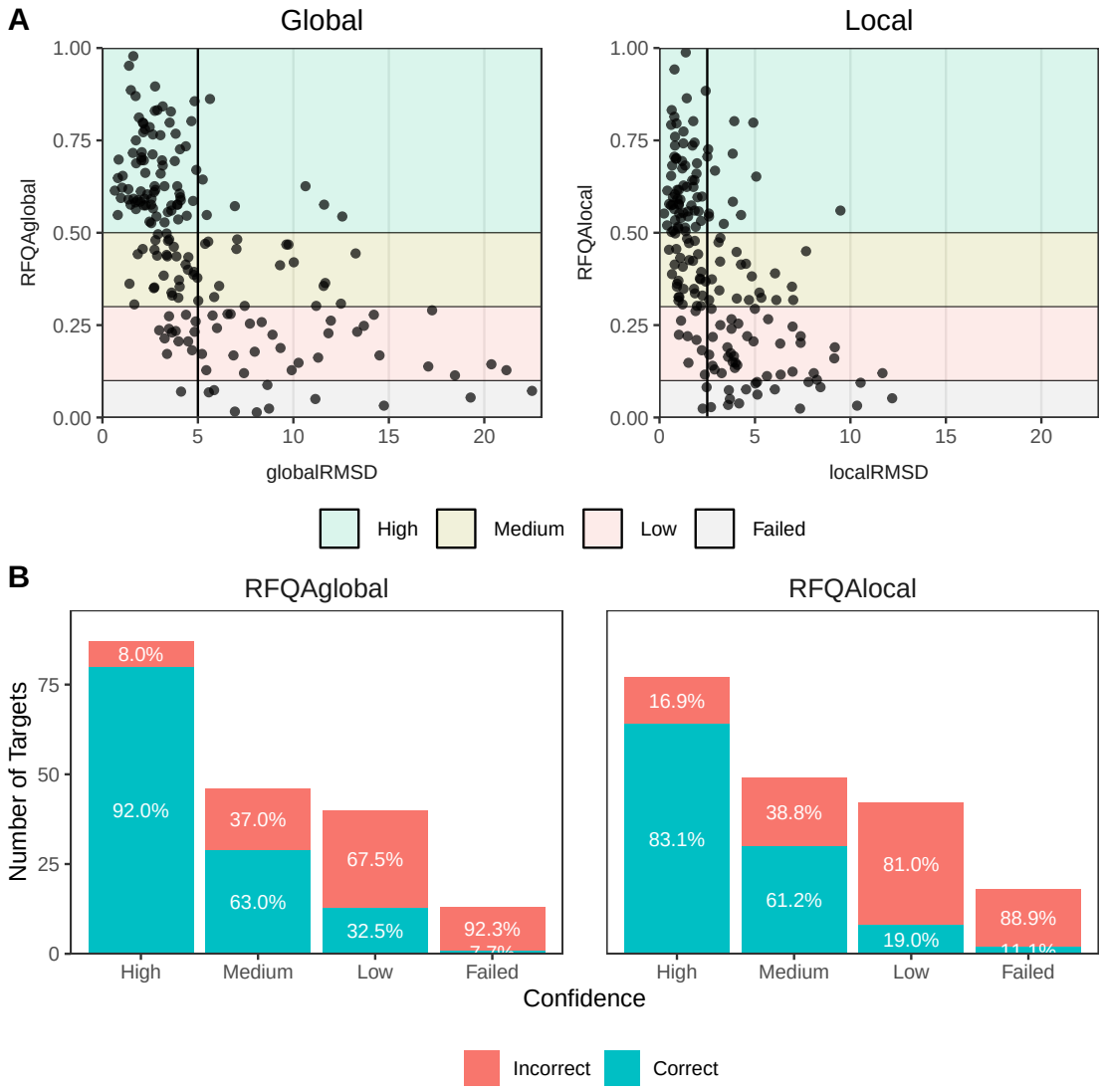

Figure 3.12: Performance of RFQAlocal and RFQAglobal on the Validation Set.
A) The Global (left) and Local (right) RFQA score and RMSD of the highest-ranking model per target in the Validation set. Predictions are categorised into high (>0.5, green), medium (between 0.3 and 0.5, yellow), or low (between 0.1 and 0.3, red) confidence, or are predicted to have failed (≤0.1, grey). The 2.5Å local RMSD and 5Å global RMSD cutoffs are indicated with vertical lines. B) The number of targets in each confidence category after classification with RFQAglobal (left) or RFQAlocal (right), and the percentage for which the highest-ranking model was correct (blue) or incorrect (red).

the highest-ranking model. For each level of confidence, we assess whether the highest-ranking model (Top1) is correct (Figure 3.12B).

RFQAlocal predicted that modelling had failed for 18 targets, of which 2 targets had a correct highest-ranked model. RFQAglobal predicted that modelling had

failed for 13 targets, of which only a single target had a correct highest-ranked model.

The highest-ranking (Top1) model was predicted to be locally correct with high confidence according to RFQAlocal for 70 targets. This model was correct for 62 of these targets (87% precision).

The highest-ranking (Top1) model was predicted to be globally correct with high confidence according to RFQAglobal for 77 targets. This model was correct for 72 targets (94% precision).

Performance on the Crystal Structure Test Set

RFQAmodel and RFQAlocal were trained on targets for which a missing region had been simulated. The Crystal Structure Test set consists of targets for which a shorter version of a similar structure has also been deposited in the PDB. We used RFQAmodel and RFQAlocal to classify the 500 models produced for each target in the Crystal Structure Test set as correct or incorrect.

For this set, SAINT2-ScafFold produced lower quality models overall, and correct models were generated for fewer targets (Figure 3.9). Of the 76 Crystal Structure Test set targets for which models were successfully produced and evaluated, 28 targets had at least one locally correct model (36%) and 36 had at least one globally correct model (47%). We found that RFQAlocal and RFQAmodel also showed a worse performance, with a large proportion of false positives for local quality; nevertheless, the majority of High-confidence predictions by RFQAglobal were correct (Figure 3.13). Modelling was predicted to have failed for 12 and 10 targets according to RFQAlocal or RFQAglobal, respectively, none of which had a correct highest-ranking model. When classified using RFQAlocal, 11 targets had a high confidence highest-ranking model, of which 5 were correct (45% precision). When classified using RFQAglobal, 23 targets had a high confidence highest-ranking model, of which 15 were correct (65% precision).

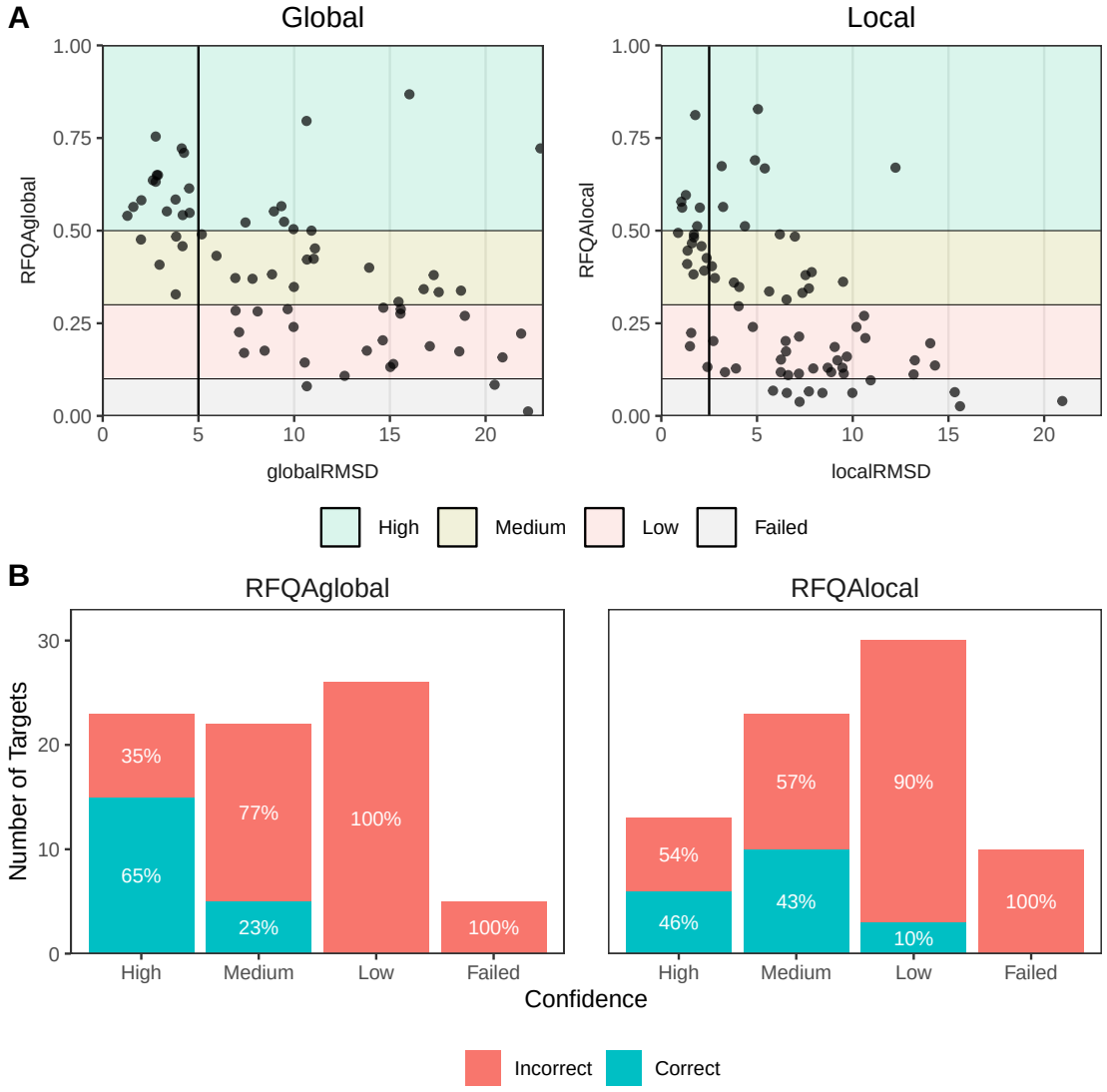

Figure 3.13: Performance of RFQAlocal and RFQAglobal on the Test Set. **A)** The Global (left) and Local (right) RFQA score and RMSD of the highest-ranking model per target in the Validation set. Predictions are categorised into high (>0.5, green), medium (between 0.3 and 0.5, yellow), or low (between 0.1 and 0.3, red) confidence, or are predicted to have failed (≤0.1, grey). The 2.5Å local RMSD and 5Å global RMSD cutoffs are indicated with vertical lines. **B)** The number of targets in each confidence category after classification with RFQAglobal (left) or RFQAlocal (right), and the percentage for which the highest-ranking model was correct (blue) or incorrect (red).

3.4 Discussion

In this chapter we describe a protocol to predict the missing terminal regions of partially-known structures. Such a method would be useful where structures are partially solved or have no full-length template structure available.

We collected a Test set of targets for which a shorter version of a similar structure has also been solved, and simulated larger Training and Validation sets with similar distributions of missing regions. These regions generally had increased solvent accessibility, larger proportion of coil, and higher B-factor values than the equivalent control regions. These properties likely reflect increased uncertainty about the conformation, and are likely to make structure prediction and validation more difficult. Nevertheless, the majority of the Test set missing regions had significant secondary structure, which indicates that these regions may be of structural and functional interest.

We developed Flib-Flex, which incorporates information about known structures into the fragment library generation process. Flib-Flex improves the coverage and accuracy for any region of the fragment library where structural information is known, both when the native structure or a homologue is used as the template structure. The current implementation does not guarantee that the template structure is included in the fragment library; this is due to the large conformational space that is searched during fragment library generation. This could be addressed by specifically enriching the library with template fragments, which is likely to improve the precision and coverage even further.

While the Flib-Flex fragment library improved the performance of SAINT2 when the entire structure was predicted, the quality of the models produced currently do not compete with that of a homology model. As SAINT2 performs prediction sequentially, screening partially extruded models against the template structure and abandoning poor intermediates could improve performance and efficiency.

Where a reliable template structure is available, it is therefore useful to incorporate this partial structure directly. We have demonstrated that SAINT2-ScafFold is

able to complete the missing regions of partial structures. SAINT2-ScafFold was able to produce locally and globally accurate models for the majority of Training and Validation set targets, although performance was lower on the more difficult Crystal Structure Test set targets. Furthermore, predicting the missing region in the presence of the known structure resulted in better models than when the missing region is predicted independently.

It is also important to be able to identify the cases for which modelling has succeeded. RFQAlocal and RFQAglobal include local model quality assessment scores and the RMSD of the flexible region as features and were trained to classify SAINT2-ScafFold models as locally or globally correct. These classifiers performed well on the Training and Validation set targets, and showed promising performance on the Crystal Structure Test set targets, particularly for assessing global quality. Out of 80 Crystal Structure Test set targets, 23 targets had a highest-ranking model predicted to be globally correct with High confidence, of which 15 were correct. Furthermore, RFQAlocal and RFQAglobal were able to accurately identify many of the cases for which modelling failed; this is particularly important given that the modelling success rate was low for the difficult Crystal Structure Test set targets.

The difference in performance on the Crystal Structure Test set for both SAINT2-ScafFold and model quality assessment indicates that the simulated Training and Validation cases are not adequately representative of the Test set cases. In particular, a larger proportion of the Training and Validation set have shorter missing regions, which is likely to contribute to the higher proportion of targets with a globally correct model. Furthermore, the Training and Validation sets had higher quality fragment libraries, which probably improved the performance of SAINT2-ScafFold.

In conclusion, SAINT2-ScafFold combines a template structure with template-free modelling and results in better quality models than homology modelling or template-free modelling alone. The combination of Flib-Flex, SAINT2-ScafFold and RFQAlocal/RFQAglobal can successfully produce and identify correct models for missing terminal regions for the majority of simulated cases, and a smaller number of example Crystal Structure Test cases, providing a promising foundation for

future work. Performance would be improved by training on a more representative set of example cases, provided a larger number could be identified. This could be achieved by relaxing some of the conditions used when deriving the Test set from the CATH database.

4

Stepwise protein structure prediction of long proteins using SAINT2-ScafFoldOn

Contents

4.1 Introduction

4.1.1 Long proteins

In this chapter we describe a method to improve the prediction of large protein structures, which are biologically prevalent but experimentally difficult to determine. Template-free protein structure prediction performs poorly on long proteins (Kinch et al., 2016; KC, 2016) (Section 1.3.10). While exploration of the vast conformational space and complex topologies available to long protein structures currently requires a prohibitive amount of computational time *in silico*, proteins are able to adopt their native structure efficiently *in vivo* (Section 1.2). Our protein structure prediction software, SAINT2, enables more efficient structure prediction by emulating the cotranslational folding pathway seen *in vivo* (de Oliveira, Law, et al., 2018). In this chapter, we incorporate another mechanism used by proteins to fold efficiently *in vivo* to similarly improve the structure prediction of long proteins.

4.1.2 Foldons and stepwise prediction

There is experimental evidence that protein folding can occur via an ordered process of foldon-determined steps, in which small units (foldons) fold sequentially and guide the folding of subsequent units (Englander et al., 2014; Hu et al., 2016). These foldon units are small enough to have a tractable conformational search, but large enough that adoption of their compact structure provides sufficient energy bias to drive folding. Under this model, protein folding consists of local folding that results in small folded substructures, which then seed and stabilise global folding.

The refolding pathways of proteins can be probed using Hydrogen/deuterium exchange (HDX) experiments, which identify regions that quickly fold and become solvent-inaccessible, and which regions remain unfolded and undergo exchange for longer (Bai et al., 1995). There are 57 protein entries annotated with early, intermediate and late folding residues based on this experimental data in the database Start2Fold as of November 2019 (bio2byte.be/start2fold/) (Pancsa et al., 2016).

As described in the Introduction (Section 1.2), folding intermediates have also been observed in a cotranslational setting. A nascent chain may adopt a native-like structure, identified by conformation-dependent interactions (Komar et al., 1997) or enzymatic activity (Frydman et al., 1999). Protein folding on the ribosome has been observed biochemically (Komar et al., 1997), kinetically (Kelkar et al., 2012), and even in real time with NMR spectroscopy (Cassaignau et al., 2016; Holtkamp et al., 2015). Secondary structures and small domains have been observed to fold within the ribosome tunnel, which can accommodate about 30 extended residues, and is able to entropically stabilise α-helices (Nilsson et al., 2015; Marino et al., 2016; Kramer et al., 2009).

These two sets of evidence demonstrate that substructures within proteins are capable of folding semi-independently. We have previously shown that structure prediction with SAINT2 tends to perform well on small domains (Chapter 2 and de Oliveira, J. Shi, et al., 2017). If foldons and subdomains – semi-stable structures that are not complete domains – can be identified within protein chains from the sequence, initial local prediction of these regions may improve prediction of large and multidomain proteins, which otherwise have intractably large conformational space.

In this chapter we discuss the incorporation of foldon-based prediction into SAINT2 by identifying semi-stable substructures within proteins and folding them individually in succession. Like sequential prediction, which was inspired by cotranslational folding, we hope that emulating this mechanism for efficient exploration of conformational space - used by proteins *in vivo* - will improve the prediction of previously intractable targets.

4.1.3 Identification of domains and foldons

In order to computationally sample the conformations of these foldons separately, the identification of foldons is a necessary step. To identify foldons, it may be possible to employ similar methods as those used to identify protein domains.

Domains are typically considered the smallest unit of protein structure (Kolodny et al., 2013). They are generally understood as independently folding sets of amino

acids that interact more strongly with each other than the rest of the protein, and are frequently associated with particular functions. While very small or large domains exist, they are typically between 50 to 200 residues (Chothia et al., 2009). Domains are often genetically conserved and leave modular signatures in protein evolution; sequences corresponding to entire domains are frequently duplicated and recombined within genomes (Chothia et al., 2009). They can therefore be considered both a structural and evolutionary unit.

Methods that assign or predict domain boundaries do so by identifying compact, stable regions with self-contained interactions. These methods could be extended to detect smaller, semi-stable subdomains, such as foldons.

The division of known protein structures into constituent domains is non-trivial. Precise definitions of domains vary; consequently, a range of domain assignment methods exist and the results are sometimes inconsistent (Schaeffer et al., 2011; Csaba et al., 2009; Holland et al., 2006).

Manual assignment, as used by Structural Classification of Proteins (SCOP) (Murzin et al., 1995), is regarded as the most reliable method, but the increasing number of solved protein structures has made it necessary to develop semi- or fully-automated methods (Fox et al., 2014). Automated domain recognition is informed by sequence and structure similarity to known domains, as well as protein topology. The semi-automated domain boundary assignment in Class, Architecture, Topology, Homology (CATH) (Sillitoe et al., 2015) uses a range of algorithms to probe for structural features such as compactness, self-contained interactions and independent hydrophobic cores to define domains without sequence or structural similarity to known domains.

We aim to divide proteins into foldons, which are sub-domain structures. This has previously been attempted using a knowledge-based energy function that approximates 'foldability' (Panchenko et al., 1996). For this technique, the average energy of the segments either side of each residue is determined from the N- to C-terminal. A foldon boundary is assigned at the first 'foldability' maximum, allowing a minimum foldon size of 15 residues. This process is then repeated - with each

foldon beginning where the previous foldon ends - until all residues are assigned to a foldon. The resulting foldons had an average size of 38 residues, and correlated relatively well with the limited experimental foldon data available at the time.

The scoring function used in SAINT2 is a physics and knowledge-based score that evaluates structural compactness and similarity of features to known protein structures (Section 1.4). So following the approach of Panchenko et al., we first examined the use of this scoring function to identify semi-stable units within protein structures. However, this approach uses the native structure of a protein, which would not be available in a structure prediction setting.

Other methods have been developed to predict domain boundaries from sequence alone. Domains are defined as compact units which have more interactions within the domain than with other parts of the protein. Such units can be identified in structures by evaluating patterns of non-trivial residue-residue contacts, defined where the C-β atoms (C-α in the case of Glycine) of two residues are less than 8Å apart and separated by more than four residues in sequence.

With a sufficiently large multiple sequence alignment, contacts can be predicted from sequence by inferring coevolution from correlated mutations (see Section 1.3.3). These methods are dependent on evolutionary information and performance is therefore correlated with the number and diversity of homologues and quality of the multiple sequence alignments available (Kamisetty et al., 2013). Predicted contacts have previously been used to predict domain boundaries (Rigden, 2002; Sadowski, 2013); the boundaries were chosen by maximising the number of contacts between residues within each domain. In a recent study (Sadowski, 2013), this method was shown to outperform other sequence-based methods of domain boundary prediction, including those based on domain linker sequence signatures (Suyama et al., 2003; Ebina et al., 2009), and predicted solvent accessibility and secondary structure (Cheng, Sweredoski, et al., 2006).

The amount of sequence data available and the accuracy of contact prediction has improved over recent years (Kamisetty et al., 2013; Ovchinnikov, Park, et al., 2017). We next investigated whether predicted contacts can be used to identify

small, self-contained regions within a protein sequence to facilitate a foldon-based approach to protein structure prediction.

4.1.4 Overview

Inspired by the foldon-based folding pathway seen *in vivo*, we describe the improvement of structure prediction of long proteins by dividing their structures into smaller, semi-independently folding segments. To do this, we used the physics and knowledge-based potential of SAINT2 to identify stable or quasi-stable foldon-like substructures within long proteins, reproducing early attempts at foldon identification (Panchenko et al., 1996). These smaller substructures should have a more tractable conformational space and score well during protein structure prediction.

We then predicted the protein structures with identified foldons using different implementations of SAINT2. We first predicted the structure of the constituent foldons separately, and found that individual foldons are capable of folding without the presence of the rest of the protein. Furthermore, independent prediction of the N-terminal foldon results in more accurate models with better SAINT2 scores. We then implemented a protocol, SAINT2-ScafFoldOn, in which the second foldon is built from a model of the first. When given a correct model of the first foldon, SAINT2-ScafFoldOn outperforms prediction of the whole structure; full-length model populations are enriched with correct answers and more targets are predicted correctly.

However, in a real template-free prediction scenario it would be necessary to identify these substructures from sequence alone. We therefore investigated whether a sequence-based method is able to identify similar foldons using contact or secondary structure information predicted from the sequence, and by dividing the chain into sections of around 150 residues.

We then tested the SAINT2-ScafFoldOn procotol on a larger Validation set. We found that our model quality assessment method, RFQAmodel, is able to identify correct models of the N-terminal foldon (foldon-1) as well as correct full-length

models produced using SAINT2-ScafFoldOn, built from the highest-ranking models of foldon-1, even when a sequence-based method of foldon determination is used. We find that by using SAINT2-ScafFoldOn to model these 87 long targets, with no structural information, 23 targets are predicted to be correct with High confidence, compared to 8 using SAINT2 alone.

4.2 Methods

4.2.1 Protein Data Sets

Short Two-domain and Single-domain CATH Set

We selected from the CATH database (Sillitoe et al., 2015) unbroken chains with a minimum resolution of 2.5Å, a maximum of 150 residues, monomeric stoichiometry, and consisting of two continuous domains to which all residues are assigned. We performed a pairwise BLAST (Camacho et al., 2009) against these 191 proteins and clustered them using an E-value cut-off of 0.05. A single protein was selected from each cluster, yielding a non-redundant set of 15 short, two-domain proteins.

For the single-domain set, we selected 583 monomeric, single-domain proteins with a maximum of 150 residues from CATH's set of non-redundant domains constructed using all-against-all BLAST with a 40% sequence identity cutoff.

Long Single-domain Set

SAINT2 performance has previously been validated on a PDB-Representative dataset of 41 structurally diverse, single chain, soluble proteins (de Oliveira, J. Shi, et al., 2017). These are single-domain proteins according to the SCOP database annotations, though some are considered to be two-domain by CATH. Of the 26 that are longer than 150 residues, 20 are divisible into two foldons using our method described in Section 4.2.2. The 20 proteins in this set were used to explore the application of foldons to long protein structure prediction. The PDB code,

length, domain assignments and foldon boundaries of each protein are shown in Appendix Table C.1.

Long Training and Validation Sets

The Training and Validation sets described in Chapter 2 each consist of 244 structurally diverse targets of length 50–250 that are single-domain according to SCOPe classification (see Section 2.2.1). In this chapter, we use all of those targets with 150 or more residues, resulting in 91 and 87 Training and Validation set targets, respectively. These sets were used for the training and validation of RFQAmodel (Chapter 2), but were not used for training in this chapter.

4.2.2 Identification of Foldons

Assignment of foldons using Folding-Unit Predictor (FoldUP)

We implemented a structure-based method to define foldons. FoldUP identifies semi-stable substructures within known protein structures, using an adaptation of the method described by Panchenko et al. (1996). The structure is cleaved after residue i and the average SAINT2 score (without the contact score) of the resulting two segments is calculated:

$$E_i = (E_{1,i} + E_{i,L})/2 \tag{4.1}$$

for $10 \leq i \leq L - 10$ where L is the length of the protein. The scoring components within the SAINT2 score are weighted and normalised as if all the foldons were "short" (less than 150 residues) (see Section 1.4).

The minimum of the resulting profile is the optimal point of cleavage, where the average stability of the two halves is maximised under our score. If the minimum is at least 15 residues away from both termini (the minimum foldon size), then this position is rounded to the nearest C-terminal end of a secondary structure to define the boundary between foldon-1 and foldon-2. Pseudocode describing the procedure is given in Appendix C.1.

Prediction of foldons using predicted contacts (conFoldUP)

We explore a sequence-based version of FoldUP that uses predicted contacts (conFoldUP). conFoldUP cleaves the protein after each residue and calculates the number of residue pairs predicted to be in contact within (intra-foldon contacts) or between (inter-foldon contacts) the resulting segments. The predicted contacts used for conFoldUP are those generated by MetaPSICOV (D. T. Jones, Singh, et al., 2015) with an estimated precision cutoff (PPV) of 0.5.

Prediction of foldons using predicted secondary structure (ssFoldUP)

We implemented a sequence-based version of FoldUP that uses predicted secondary structure (ssFoldUP). For each target, candidate breakpoints are all residues within regions of at least five consecutive residues with no predicted secondary structure, that are not between two beta strands. From these candidate breakpoints, the closest to half the length of the target is chosen as the foldon boundary. Pseudocode outlining how ssFoldUP determines foldon boundaries is included in Appendix C.1.

4.2.3 Prediction of protein structures using SAINT2

The fragment libraries were generated using Flib (de Oliveira, J. Shi, et al., 2015) for the Long Single-domain set. Fragments were selected based on predicted secondary structure and torsion angles, generated using PSIPRED (D. T. Jones, 1999; Buchan, Minneci, et al., 2013) and SPINE-X (Faraggi, Y. Yang, et al., 2009; Faraggi, T. Zhang, et al., 2012), respectively. Fragments from homologous structures were removed from the libraries. For the Long Training and Validation sets, the same fragment libraries described in Section 2.2.2 are used. Contacts were predicted using MetaPSICOV (D. T. Jones, Singh, et al., 2015); only those with an estimated PPV of at least 0.5 are used. Target structures were predicted using SAINT2 (Section 1.4) or SAINT2-ScafFoldOn (Section 4.2.3).

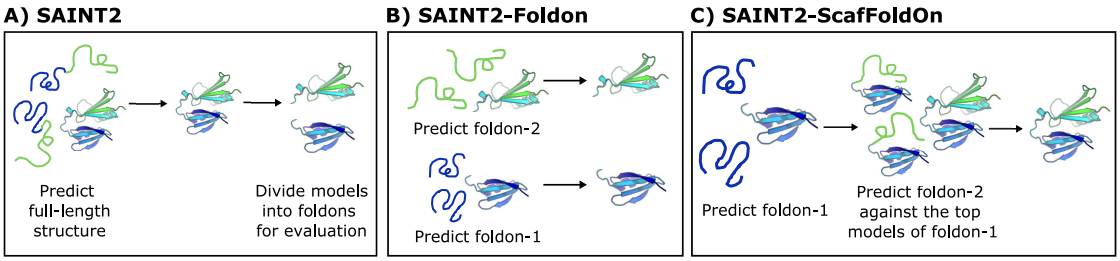

Figure 4.1: Diagram of the SAINT2, SAINT2-Foldon and SAINT2-ScafFoldOn protocols.
A) SAINT2: 500 models of the entire structure are generated using SAINT2 on the full-length target. For assessing the accuracy of individual foldons, the resulting models are divided into the constituent foldons to be scored individually. **B)** SAINT2-Foldon: 500 models each of foldon-1 (blue) and foldon-2 (green) are generated separately using SAINT2. **C)** SAINT2-ScafFoldOn: 500 models of the first foldon (blue) are generated using SAINT2. SAINT2-ScafFold is used to build 500 models of the remainder of the target from each of the best (using TM-score) or highest-ranking (using RFQAmodel) models of foldon-1, with sampling restricted to foldon-2 (green). A short C-terminal region of foldon-1 is also included in the sampling, consisting of either the final secondary structure or the final 15 residues.

Evaluation of predictions

As in Chapter 2, models were evaluated using the TM-score compared to the native structure, calculated with TM-align (Y. Zhang and Skolnick, 2004; Y. Zhang, 2005). A TM-score of 0.5 or above, which indicates they share the same overall fold, is used to identify correct models (Y. Zhang, 2005). To compare the success of different implementations of SAINT2, we calculate the best TM-score across all models, the number of correct models and the distribution of TM-scores.

SAINT2

Models of the full-length protein are produced using standard SAINT2 in the Forward direction. For comparison to SAINT2-Foldon, models are divided into constituent foldons and scored individually (Figure 4.1A).

SAINT2-Foldon

For this implementation, we predict the foldons of each target protein individually. The fragment library and predicted contacts generated for the target protein are divided at the foldon boundary and input separately into SAINT2 Forward, with

11,000 moves for each foldon (Figure 4.1B).

SAINT2-ScafFoldOn

SAINT2-ScafFold is an implementation of SAINT2 developed by Eleanor Law (Section 1.4.1). A segment of the target protein is provided as a starting point from which the remainder of the protein is grown. For our extension of this method, SAINT2-ScafFoldOn, we select the top-scoring model from the prediction of the N-terminal foldon (foldon-1) for use as the starting segment. As prediction proceeds, fragment replacement steps are restricted to foldon-2, as well as a short region of the preceding foldon. This region of foldon-1 is included in conformational sampling to ensure the linker is flexible and accessible, as the adoption of a restrictive, compact structure may have been preferred in isolation. Unless otherwise specified, a length of 15 residues was used for this flexible region. A diagram outlining the SAINT2-ScafFoldOn protocol is shown in Figure 4.1C.

4.2.4 Model Quality Prediction

The quality of models produced was predicted using RFQAmodel, which was trained on all 244 full-length Validation set targets (Chapter 2).

4.3 Results

4.3.1 Identifying potential foldon boundaries

Structure-based foldon identification using FoldUP

SAINT2 is able to accurately predict the topology of small domains (Chapter 2 and de Oliveira, J. Shi, et al., 2017). We aim to assess whether comparable results can be obtained when predicting the topology of foldons. We developed the FoldUP method to identify small, semi-stable substructures from protein structures. As

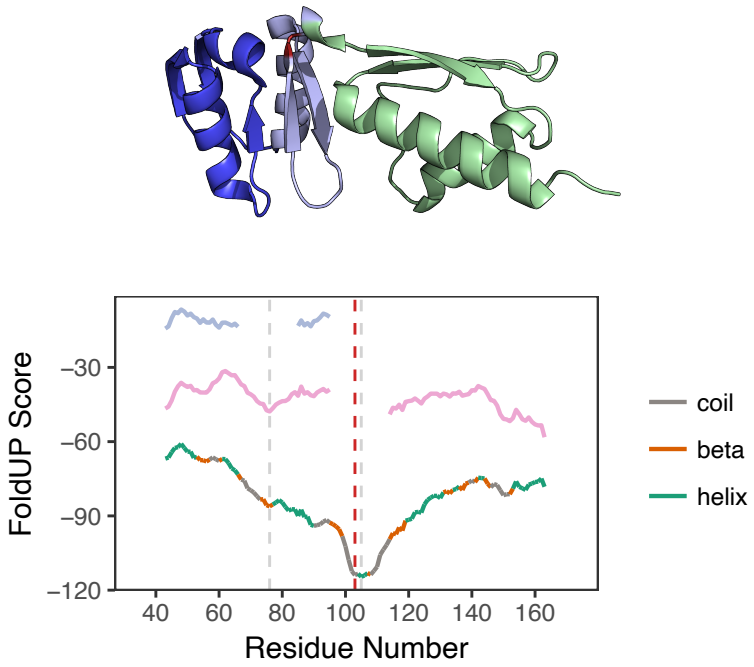

Figure 4.2: Example FoldUP profile for 4G08A.
The domain boundary assigned by CATH is indicated with a dashed red line. The DSSP (Kabsch et al., 1983) secondary structure assignments are shown: helix in turquoise, beta in orange or no secondary structure, 'coil', in grey. Dashed grey lines indicate where the minimum average score falls at least 15 residues from either end; this is rounded to the nearest C-terminal end of a secondary structure element and defined as the foldon boundary. The resulting foldons are scanned for further potential foldon boundaries, shown in pink and light blue. Above the plot is the structure of 4G08A, coloured according to the foldons defined by the FoldUP profile. The CATH domain boundary is highlighted in red.

a proof-of-principle, we tested whether the method is capable of dividing a set of small proteins into the two domains defined by CATH.

FoldUP profiles were calculated for two sets, 15 Short Two-domain proteins and a further 583 Short Single-domain proteins, using the method described in Section 4.2.2. If a foldon boundary was found, FoldUP was run recursively on the resulting protein sections. An example of a FoldUP profile is shown in Figure 4.2.

For the Short Two-domain set, the first FoldUP minimum was within 15 residues of the domain boundary assigned by CATH for 14 of the 15 proteins (Figure 4.3A). For the remaining case, the second domain was just 13 residues long, which is smaller than the minimum foldon size that FoldUP can assign. The domains in this

dataset are unusually small, being between 13 and 117 residues, with an average of 59, whereas domains in general tend to be 50–200 residues long (Chothia et al., 2009). The proteins in the Short Single-domain set tend to be shorter than those in the Short Two-domain set; the distribution of lengths for each set are shown in Figure 4.3B. Allowing a minimum foldon size of 15 residues, at least one potential FoldUP breakpoint was identified for 379 (65%) of these Short Single-domain proteins. FoldUP was therefore able to accurately divide the Short Two-domain set targets into their constituent small domains, and many of the Short Single-domain set may also contain similarly domain-like sub-structures.

The purpose of FoldUP is to identify where the protein can optimally be divided into compact and relatively stable foldons, which may have enough self-contained interactions to be predicted individually. It is therefore important to determine whether the minimum of a given FoldUP profile would be a useful foldon boundary. The SAINT2 scoring function gives low scores to compact, low-energy, protein-like conformations. The more stable the resulting segments are relative to random divisions, the more extreme the FoldUP minimum will be. Conversely, segments resulting from a shallow minimum might not be more stable than a random cut. To computationally distinguish between these possibilities, we calculated (for each potential FoldUP breakpoint) the difference between the scores at the minimum point and the lowest of either end of the profile; this value is referred to as M_d of the FoldUP minimum. This is an imperfect approximation, as does not capture local information about the minimum, and places undue importance on the ends of the profile, which correspond to the FoldUP score of residues 10 and $L - 10$. However, it is sufficient for exploratory analysis.

If the division of a single domain into two foldons has a stability benefit similar to the division of a protein into its two small constituent domains, then it may be a useful substructure for protein structure prediction. We therefore compared the M_d scores of FoldUP minima identified for the two-domain and single-domain sets, in order to determine what should be considered a meaningful FoldUP minimum. Figure 4.3 shows that the Short Two-domain proteins tend to have larger M_d scores

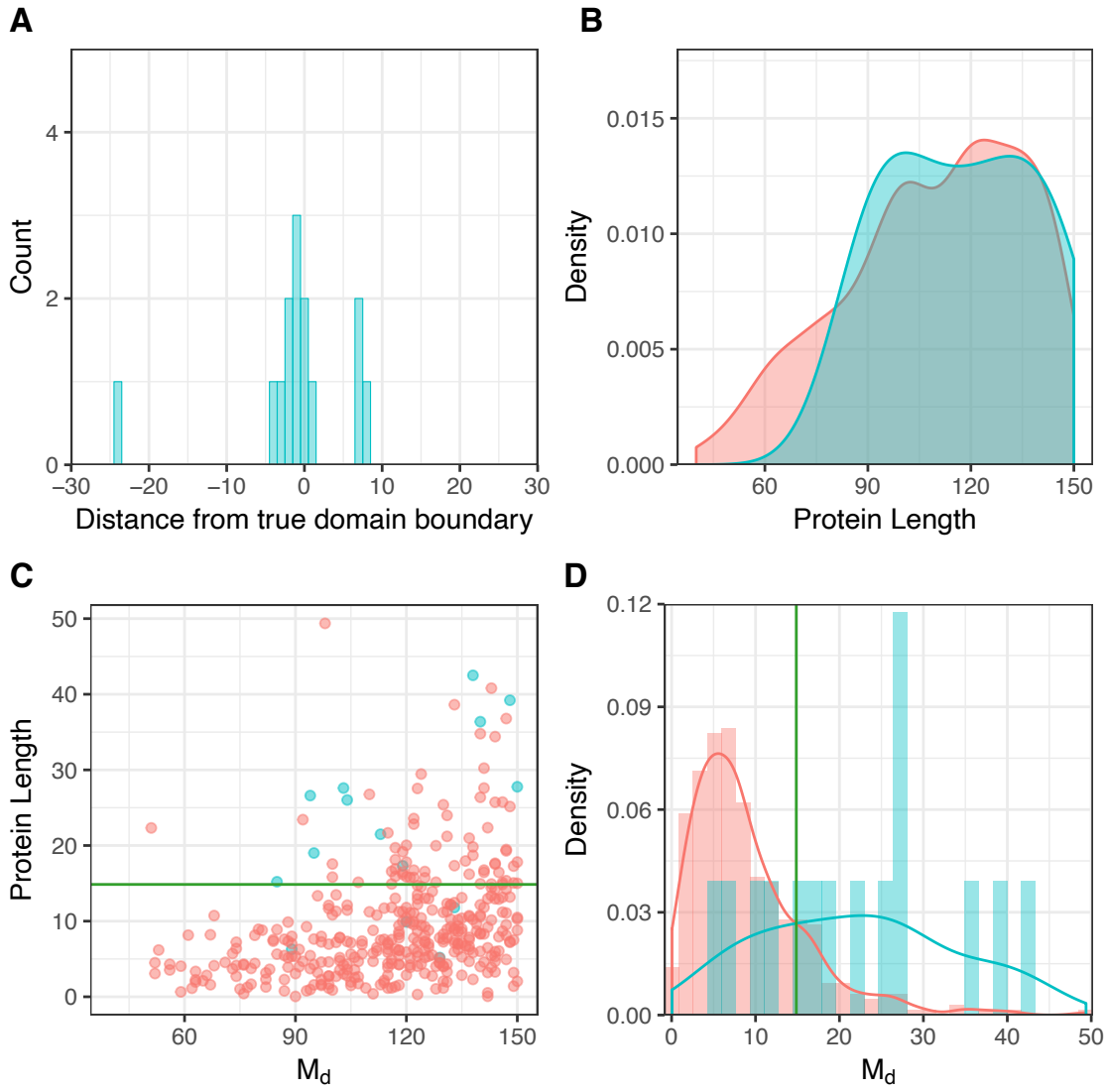

Figure 4.3: Comparison of FoldUP results for 15 two-domain (blue) and 583 single-domain (red) short proteins.
A) Histogram showing the number of residues between the first FoldUP minimum and the domain boundary assigned by CATH for the two-domain proteins. **B)** Distributions of lengths for proteins in each set. **C)** M_d value against protein length. **D)** distributions of M_d values for the proteins with identified FoldUP minima (15 two-domain and 379 single-domain proteins). The distributions overlap at a M_d value of 14.9, marked with a green line.

than the single-domain proteins, but there is also a correlation with protein length. The two M_d distributions overlap at 14.9, which we have used as a minimum cut-off for defining a FoldUP breakpoint (Figure 4.3D). This is a conservative cut-off that includes 11 of the 15 Short Two domain proteins (73%), as well as 67 of the 538 Short Single-domain proteins (11%) (18% of the Short Single-domain proteins that contained a potential FoldUP break).

FoldUP foldons were then determined for the proteins in the Long Single-domain set (see Section 4.2.1). Using the minimum M_d cut-off of 14.9, 20 of the 26 proteins had suggested foldons using FoldUP. Where a foldon boundary is found, FoldUP runs recursively on the resulting protein sections; six targets were divided twice, and the longest was divided four times. However, for simplicity, we only consider the first FoldUP boundary, resulting in two foldons per protein. The shortest foldon was 63 residues, and the longest was 283. Foldon-1 tended to be slightly longer than foldon-2, ranging from 44% to 75% of the protein sequence, with an average of 53%. Therefore, using an approach capable of identifying domains in small two-domain proteins, the majority of proteins in our Long Single-domain set can be divided into substructures. For these targets, it may be possible to predict these substructures sequentially and improve the results of protein structure prediction.

Sequence-based foldon prediction from predicted contacts using conFoldUP

The foldon boundaries described above were defined using FoldUP on the native structure of each target protein. This makes them unsuitable for protein structure prediction purposes. We next explored conFoldUP, an alternative method of predicting semi-stable substructures, which is based on predicted contact information inferred only from the target sequence.

One way to identify compact substructures is to maximise the number of residue-residue contacts within each resulting foldon (or, equivalently, minimising the number of contacts between foldons). We generated predicted contacts for the 20 proteins in the Long Single-domain set and used these to calculate conFoldUP profiles, as described in Section 4.2.2. The conFoldUP profile for 1VL1 is shown

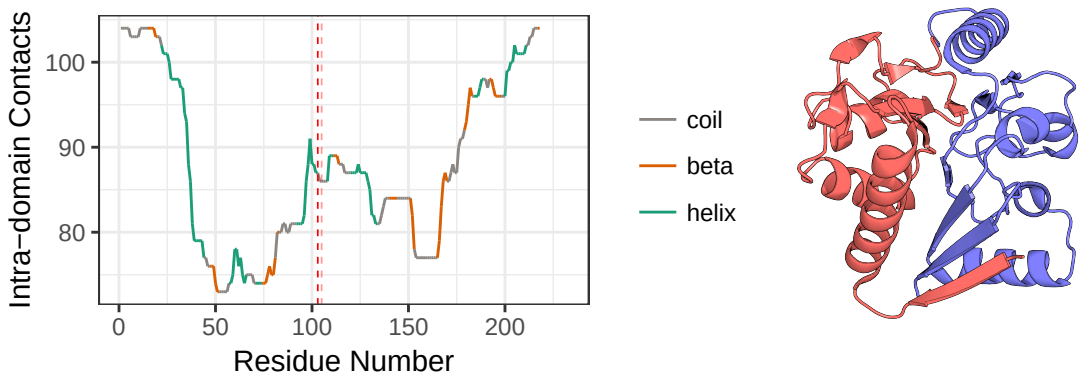

Figure 4.4: Example conFoldUP profile for 1VL1.
The foldon boundary assigned by structure-based FoldUP is indicated by the dashed red line. The actual FoldUP minimum, prior to rounding to the nearest secondary structure element, is shown in light red. The DSSP secondary structure assignments for each residue are shown: 'helix' in green, 'beta' in orange, or no secondary structure, 'coil', in grey. The structure of 1VL1 is shown on the right; Foldon-1 and Foldon-2, identified by FoldUP, are coloured red and blue, respectively.

in Figure 4.4. Profiles for the complete protein set, as well as full contact maps for the proteins, are included in Appendix Figures C.1 and C.2.

We have not yet implemented a way to select a foldon boundary from the conFoldUP profile. We conducted a preliminary comparison by ranking the local maxima in the conFoldUP profile by the number of intrafoldon contacts. For eight of the proteins in this set, the foldon boundary designated by FoldUP falls within five residues of at least one of the four best local maxima.

conFoldUP therefore shows promise as a method for the identification of foldons from sequence alone. It is possible that it could be improved by increasing the number of contacts included, for example by using a different estimated PPV cutoff, or by enriching with contacts that can be inferred from predicted secondary structure. This will be discussed further in future work.

Relevance of foldons to the prediction of Long Single-domain protein structures

FoldUP foldons were identified in 20 of the proteins in the Long Single-domain set. To explore whether the concept of foldons might be useful, we looked at the

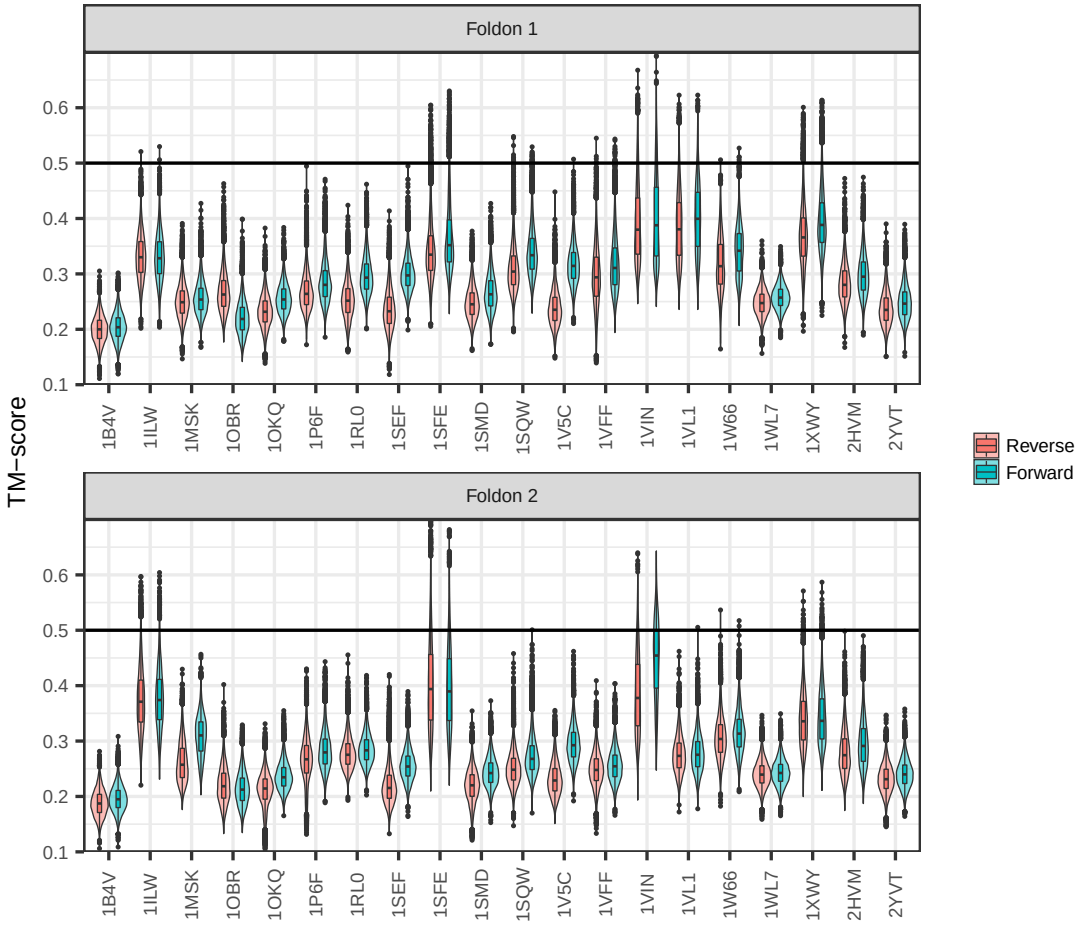

Figure 4.5: Comparison of SAINT2 Forward and Reverse for foldons in the Long Single-domain set.
The distribution of TM-scores for the 10,000 models generated in SAINT2 Forward (blue) and Reverse (red) modes for each protein in the Long Single-domain set, with incorrect predicted contacts included. Full length models were generated and the foldons scored separately. A model with a TM-Score of 0.5 or above is considered correct.

behaviour of these targets under sequential protein structure prediction. For each protein, 10,000 models had previously been generated using SAINT2 in the Forward and Reverse modes by Saulo de Oliveira (de Oliveira, J. Shi, et al., 2015). At least one correct full-length model was produced for five targets in both Forward and Reverse mode. The models were split by the foldon boundary assigned by FoldUP as described in Section 4.2.2. Foldon-1 and foldon-2 were scored separately against their native equivalents; the distributions of these scores are shown in Figure 4.5.

Foldon-1 tended to be predicted better than foldon-2, with average top TM-scores of 0.48 and 0.46, respectively. This is despite the fact that foldon-1 tends to

be longer than foldon-2, consisting of on average 53% of the target sequence. The Forward direction produced at least one model containing a correct foldon-1 for nine targets, compared to eight in Reverse mode. A greater number of correct models were generated in the Forward than the Reverse mode in all these cases, with an average increase of 138. In comparison, models containing a correct foldon-2 were produced for seven targets in Forward Mode, compared to five in Reverse mode. The Forward mode generated more correct models for three of these five. The distributions of models produced by SAINT2 Forward tends to show a shift towards better TM-scores for the population of models compared to SAINT2 Reverse.

The improvement in predictions using the Forward mode compared to Reverse adds to the evidence that the biological direction of translation is relevant to structure prediction *in silico* (Ellis et al., 2010; de Oliveira, J. Shi, et al., 2017). These results show that when predicting the entire structure in the biological direction, the first foldon is more likely to be correct than the second. This suggests it may be adopting the native conformation before the second foldon is extruded, and it may therefore be possible to predict it in isolation.

4.3.2 Stepwise prediction of Long Single-domain proteins using ScafFoldOn

We next investigated whether the identified foldons were stable enough for prediction individually. For the 20 proteins in the Long Single-domain dataset, we generated 10,000 models using SAINT2 in the Forward mode on the entire target (SAINT2), and on the foldons individually (Foldon). Using SAINT2-ScafFold, 10,000 full-length models were then generated by predicting foldon-2 against the best model of foldon-1 for each target (ScafFoldOn). For these tests, identical fragment libraries were used as the previous full-length predictions by Saulo de Oliveira (Section 4.3.1), but only the predicted contacts that were correct were used. This assumes 100% precision of contact prediction; in reality, the precision ranges from 70-100% for this set, with an average of 80.6%. A table showing the top scoring model and the number of correct models produced for each target in each mode is shown in

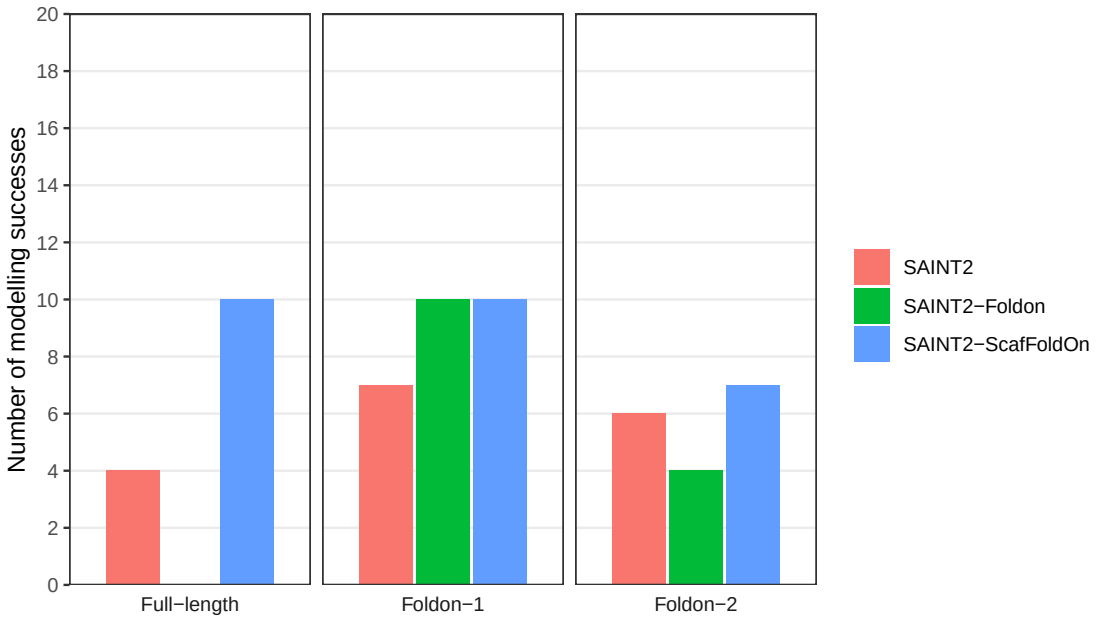

Figure 4.6: Comparison of modelling success using different protocols for the Long Single-domain set.
The number of targets, out of the 20 targets in the Long Single-domain set, for which at least one correct model was produced by each protocol. The entire structure (full-length) and individual foldons are evaluated. Ten thousand models were produced for each protein using SAINT2 Forward on the whole structure (SAINT2 Full-Length, green), the each foldon individually (SAINT2-Foldon, red) and by building foldon-2 on the top-scoring model of foldon-1 (SAINT2-ScafFoldOn, blue).

Appendix Table C.1. The number of targets for which at least one correct model was produced, for each implementation of SAINT2, is shown in Figure 4.6, and results of each protocol are discussed separately below.

Prediction of foldons individually

Predicting foldon-1 alone produced at least one correct model for 10 targets, compared to 7 when the entire protein was predicted (Figure 4.7A, "Foldon-1"). The number of correct models increased in all cases. The top scoring model was better in all but three cases. The largest decrease in TM-score score was 0.02.

Predicting the entire protein produced correct models of foldon-2 in six cases. Only four of these could be predicted correctly without the presence of foldon-1, and fewer correct models were produced for three of these (Figure 4.7A, "Foldon-2"). One exception, foldon-2 of 1SQW, performed much better when predicted

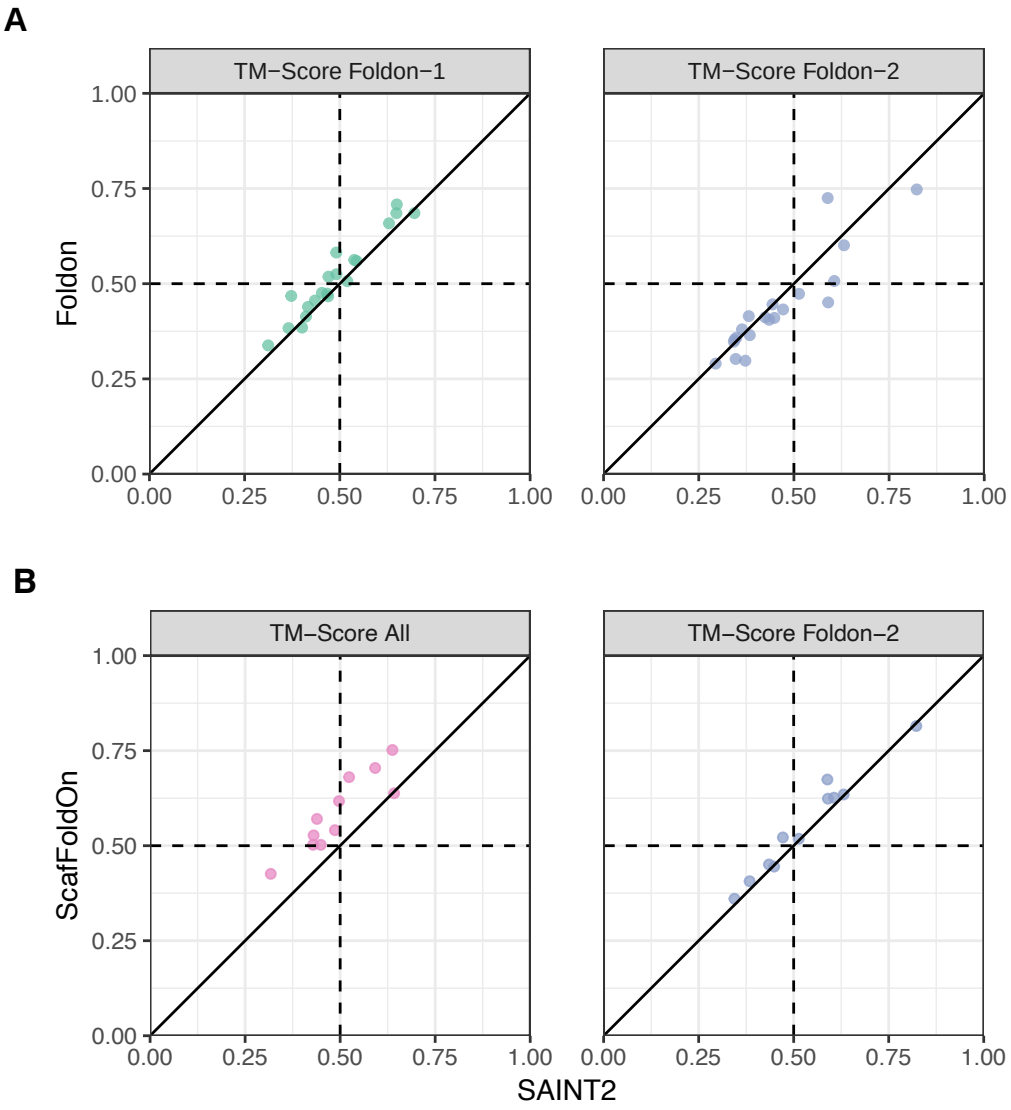

Figure 4.7: Performance of SAINT2-Foldon and SAINT2-ScafFoldOn on the Long Single-domain set targets.
Comparison of the TM-score of the best models achieved using SAINT2, against **A)** predicting the foldon individually (Foldon) or **B)** the SAINT2-ScafFoldOn protocol. TM-scores shown are either for the full-length protein (all) or for individual foldons, as specified. Points above the diagonal line indicate a better performance by the SAINT2-Foldon or SAINT2-ScafFoldOn protocol.

alone, and is also the only protein to have performed better using SAINT2 Reverse compared to SAINT2 Forward.

It is therefore possible to predict foldon-1 in isolation, with better performance than when it is predicted as part of the full-length structure. Conversely, prediction of foldon-2 is improved by the presence of foldon-1. These results add further evidence to the relevance of directional elongation to protein structure prediction, and suggest that stepwise prediction of foldon-1 and then foldon-2 may improve the modelling of these long targets.

Prediction using ScafFoldOn

We tested ScafFoldOn on the 10 targets where foldon-1 was predicted correctly individually. An extra protein, 2HVM, was included although the best model was incorrect (TM 0.47), in order to test whether a nearly correct foldon-1 can lead to a correct structure overall. 10,000 full-length models of each protein were generated, with foldon-2 building from the best model of foldon-1. To compare performance, the resulting models were scored as a whole against the native structure, as well as foldon-1 and 2 separately against the native equivalents. The last secondary structure and preceding coil regions of foldon-1 were included in the conformational sampling during prediction; there was therefore a small distribution of scores for foldon-1, but the top TM-score did not decrease in any case, and increased by an average of 0.02.

The presence of a "correct" foldon-1 improved the accuracy of foldon-2 in eight cases (Figure 4.7B, "Foldon-2"). One correct model for foldon-2 of 2HVM was produced, which had not been correctly predicted previously.

As a correct part of the structure was already provided, the technique resulted in improved overall top TM-score in all cases (Figure 4.7B, "All"). Correct models were produced for 10 of 11 cases compared to 4 previously, including a single correct model for 2HVM, and the model sets were enriched with correct answers. Figure 4.8 shows the best models produced for the 260 residue protein 1XWY.

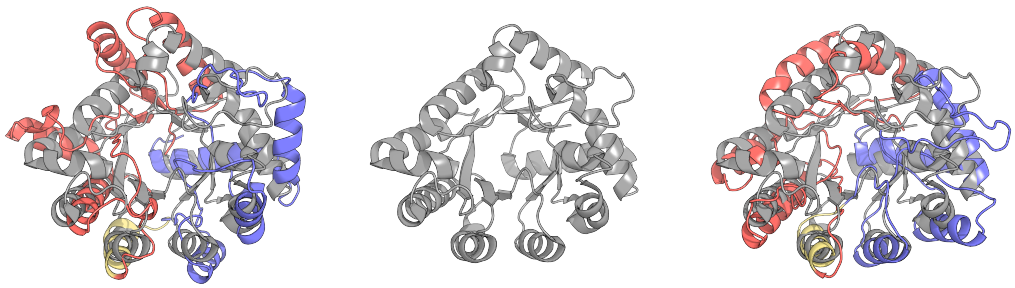

Figure 4.8: The best models of the 260 residue protein 1XWY.
The models with the highest TM-scores produced using SAINT2 Forward on the whole protein (left, TM-score 0.59) or ScafFoldOn (right, 0.70), superimposed with the native structure (grey). Foldon-1 and foldon-2 are shown in red and blue, respectively, with the final flexible region of foldon-1 included in foldon-2 sampling is shown in yellow.

Prediction using more moves

In a standard SAINT2 run, each model is produced using a total of 11,000 fragment replacement moves. Using SAINT2-Foldon and SAINT2-ScafFoldOn, each foldon benefits from 11,000 moves, which may account for the improvement in accuracy. In order to test this, we ran SAINT2 with 10,000 moves and 1,000 post-extrusion moves per 150 residues for five targets that had produced correct models using SAINT2-ScafFoldOn but not SAINT2. The TM-score of the top models increased for four of the five targets, by a maximum of 0.05 and an average of 0.02 (see Appendix Figure C.3 and Appendix Table C.2). In one case this resulted in a correct prediction: 8 correct models were produced using SAINT2 with extra moves for 1VL1, compared to 1,279 using ScafFoldOn. Therefore, this preliminary test suggests that with approximately the same computational power, SAINT2-ScafFoldOn achieves a greater improvement in models.

Prediction of 150 residue N-terminal segments individually

It is feasible that simply dividing the target structure into smaller segments is beneficial, without specifically identifying foldon-like structures. To test this, we predicted the first 150 residues of each target in the Long Single-domain set individually. This boundary was rounded to the nearest secondary structure element,

so the resulting segments actually ranged in size from 145 to 159 residues. In two cases, these segments were identical to the foldons.

The improvement over prediction of the entire target was similar for both foldon-1 and the 150 residue segments; prediction individually only reduced the top TM-score in three cases for both (Figure 4.9A and Appendix Table C.2). There is a correlation between length and TM-score: foldon-1 tended to perform better than the 150 segment when it was shorter, and similarly or slightly worse when it was longer (Figure 4.9B).

This suggests that where foldon boundaries cannot be inferred from sequence, a similar benefit may be gained by simply breaking into smaller segments, although the optimal length for these segments is yet to be determined.

Model ranking using foldons

For use in structure prediction, it is important that correct structures of foldon-1 can be identified. One benefit of a foldon-based approach to dividing long proteins into more manageable segments is that they are compact structures that should be favoured by the SAINT2 scoring function. From the 10,000 models produced by each protocol, we selected the 10 models with the best SAINT2 score for foldon-1 and compared their TM-scores (Figure 4.10A).

When the whole target is predicted at once, no correct models were selected in the top 10 by SAINT2 score, despite the fact that correct answers were generated in six cases. The best TM-score of the 10 models was on average 0.16 lower than the best TM-score of any model, with the decrease ranging from 0.07 to 0.30.

In contrast, when foldon-1 was predicted alone, the top 10 models by SAINT2 score included correct models for six of the ten cases for which a correct answer was generated. In one case, the best model produced was included in the top 10. The decrease between the best TM-score of the top 10 and the best TM-score of any model was on average 0.06, ranging from 0 to 0.14. In addition, the average SAINT2 score was much lower for the models of foldon-1 generated individually (Figure 4.10).

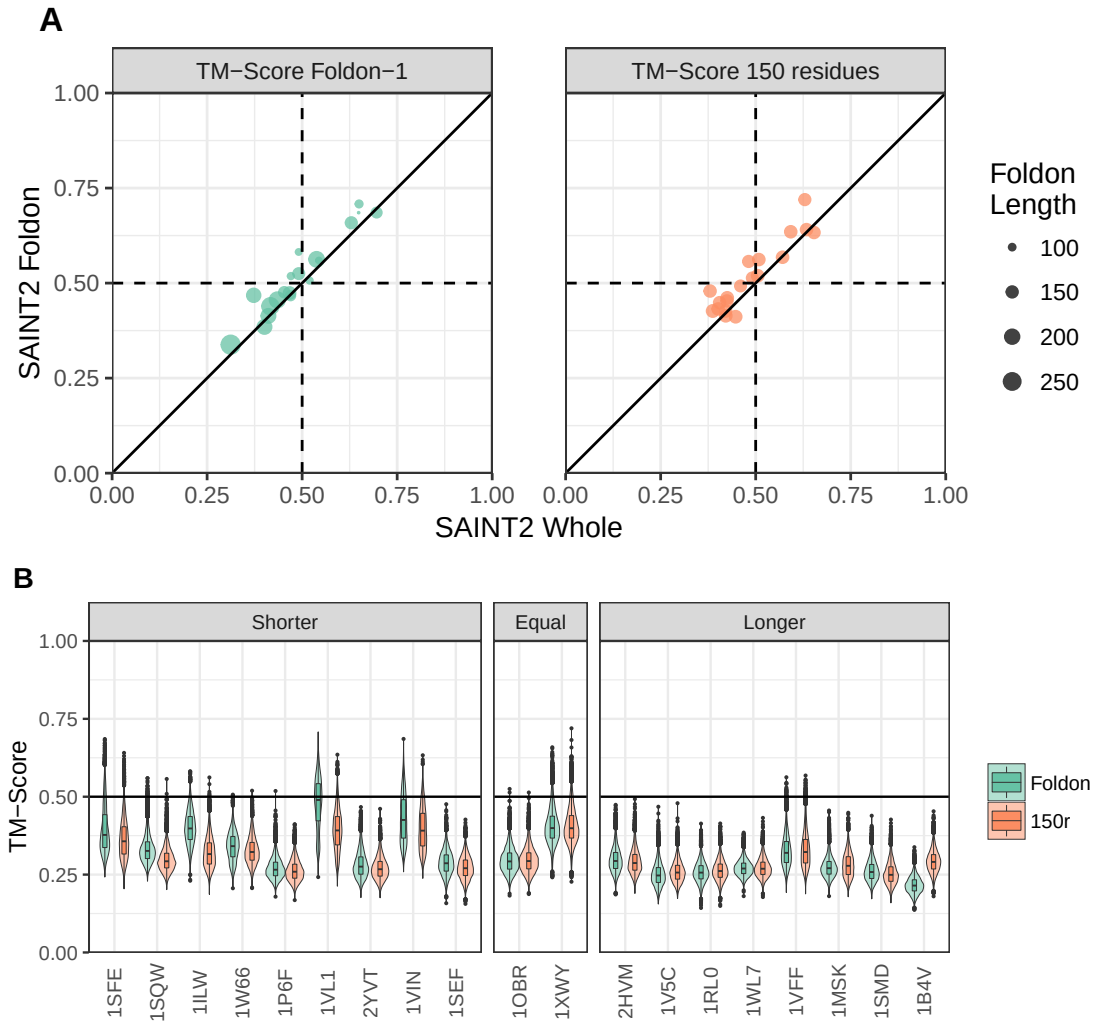

Figure 4.9: Modelling performance when using N-terminal 150 residues compared to FoldUP foldon-1.
A) Comparison of the TM-score of the top scoring models achieved using SAINT2 Forward on the full-length target, against predicting foldon-1 individually (Foldon-1) or the first 150 residues of the target, rounded to the nearest C-terminal end of a secondary structure (150 residues). The size of the point is proportional to the length of the foldon or segment. Points above the diagonal line indicate a better performance by predicting foldon-1 or the 150 residue segment individually. **B)** The distribution of TM-scores of the models produced by predicting foldon-1 (turquoise) or the 150 residue segment (orange) individually. Targets are arranged in increasing size of foldon-1; whether it is shorter, equal to or longer than the 150 residue segment is indicated.

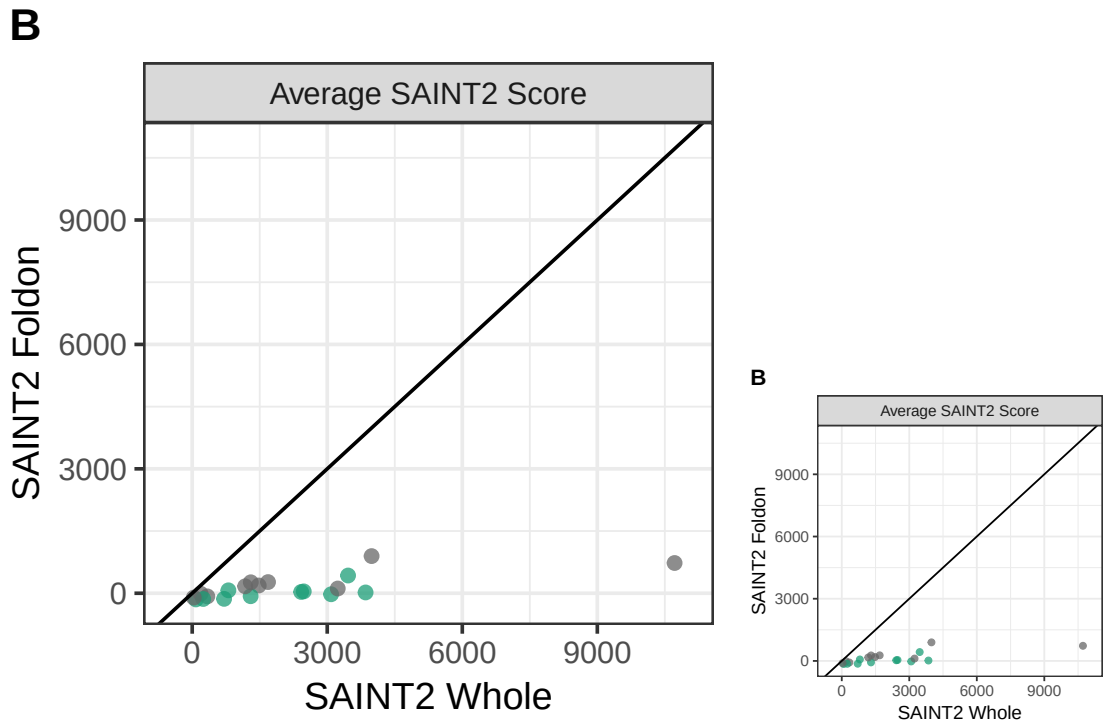

Figure 4.10: Improved ranking of the foldon-1 region of models using the SAINT2 score.
Comparison of **A)** the TM-score and **B)** the SAINT2 score of the 10 models with the best SAINT2 score for foldon-1 of each target in the Long Single-domain set, generated by prediction of the entire target at once (SAINT2 Whole) or individually (SAINT2 Foldon). Targets for which at least one correct structure was generated are coloured turquoise. "Best" shows the highest score out of the 10 models. Points above the diagonal line for TM-score, or below the diagonal line for SAINT2 score, indicate a better performance by SAINT2 Foldon.

Prediction of foldon-1 individually not only produces more accurate models, but correct models are favoured by the SAINT2 score and are easier to identify than when foldon-1 is predicted as part of the full-length target. This suggests that such substructures are likely to be favoured by our model quality assessment method, RFQAmodel, which includes the SAINT2 score as a feature.

4.3.3 Application of ScafFoldOn to the Long Training and Validation Sets

We have explored the concept of foldons in protein structure prediction using the 26 Long Single-domain Set targets (Section 4.2.1), derived from the original set

Table 4.1: Comparison of SAINT2 and RFQAmodel for the 87 long and 157 short modelling targets in our Validation set.
The total number of targets, the number of targets that were successfully modelled using SAINT2 ("SAINT2") and the number of targets in each Confidence category according to RFQAmodel ("RFQAmodel"). Targets with at least 150 residues ("Long") or fewer than 150 resides ("Short") are shown separately. Modelling is considered successful for a given target if at least one model is correct (TM-score ≥ 0.5).

	Total	SAINT2		RFQAmodel		
Long targets	87	34	39.1%	High	8	9.2%
				Medium	15	17.2%
				Low	27	31.0%
				Failed	37	42.5%
Short targets	157	108	68.8%	High	59	37.6%
				Medium	35	22.3%
				Low	41	26.1%
				Failed	22	14.0%

of 40 targets used to validate SAINT2, and found that the stepwise prediction of foldon substructures improved the prediction of these targets. To test the full ScafFoldOn protocol, including model quality assessment of the foldon-1 and full-length models, we applied the ScafFoldOn to our larger Training and Validation sets of single-domain targets (Section 4.2.1), in combination with our model quality assessment protocol, RFQAmodel.

We have previously shown that targets over 150 residues long were less likely to be successfully modelled using SAINT2 (Chapter 2 Figure 2.2). When 500 models were generated for each of the 87 targets of at least 150 residues in the Validation set, correct models were produced for 34 targets (39%), compared to 108 (69%) of the 157 targets under 150 residues (Table 4.1). Of the long targets, RFQAmodel predicted that 8 had been successfully modelled with High confidence, of which 6 had a correct highest-ranking model (Table 4.2).

We applied our SAINT2-ScafFoldOn protocol to these targets to determine whether the performance for these targets can be improved.

Structure-based SAINT2-ScafFoldOn on the Long Validation Set

Of the 91 and 87 targets in our Long Training and Long Validation sets, 89 and 84 were divided into at least two foldons by FoldUP, respectively. The FoldUP profiles

Table 4.2: Comparison of structure prediction and model quality assessment using RFQAmodel for the 87 Long Validation set targets using SAINT2, SAINT2-ScafFoldOn and SAINT2-ssScafFoldOn.
Results are shown for SAINT2-ScafFoldOn using structure-based foldons (ScafFoldOn, top), sequence-based foldons (ssScafFoldOn, middle), or using SAINT2 on the full length target (bottom). The number of targets (Total) as well as the number of targets with at least one correct model (Max) is reported for each confidence category, and for all targets overall (All). The number of targets for which the highest-ranked model (Top1) and the best of the top five highest-ranked models (Top5) is correct is shown, with the corresponding precision. Modelling is considered successful for a given target if at least one model is correct (TM-score ≥ 0.5).

	Confidence	Total	Max	Top1		Top5	
foldon-1	All	84	60	43	51.2%	48	57.1%
	High	28	27	23	82.1%	25	89.3%
	Medium	26	24	18	69.2%	20	76.9%
	Low	18	7	2	11.1%	3	16.7%
	Failed	12	2	0	0.0%	0	0.0%
foldon-2	All	84	45	11	13.1%	19	22.6%
	High	2	2	2	100.0%	2	100.0%
	Medium	6	5	1	16.7%	1	16.7%
	Low	52	33	7	13.5%	15	28.8%
	Failed	24	5	1	4.2%	1	4.2%
ScafFoldOn	All	46	34	23	50.0%	29	63.0%
	High	23	21	17	73.9%	21	91.3%
	Medium	14	10	5	35.7%	7	50.0%
	Low	6	3	1	16.7%	1	16.7%
	Failed	3	0	0	0.0%	0	0.0%
ssFoldon-1	All	84	56	26	31.0%	39	46.4%
	High	23	22	17	73.9%	21	91.3%
	Medium	28	23	8	28.6%	13	46.4%
	Low	26	11	1	3.8%	5	19.2%
	Failed	6	0	0	0.0%	0	0.0%
ssScafFoldOn	All	51	31	21	41.2%	25	49.0%
	High	27	24	19	70.4%	20	74.1%
	Medium	15	6	2	13.3%	5	33.3%
	Low	6	1	0	0.0%	0	0.0%
	Failed	3	0	0	0.0%	0	0.0%
SAINT2	All	87	34	17	19.5%	23	26.4%
	High	8	8	6	75.0%	7	87.5%
	Medium	15	11	5	33.3%	9	60.0%
	Low	27	14	6	22.2%	6	22.2%
	Failed	37	1	0	0.0%	1	2.7%

of the Long Validation set targets are shown in Appendix Figure C.4. For simplicity, only the first FoldUP break is considered here, resulting in two foldons per target.

When 500 models were generated for each foldon individually for these 84 Long Validation set targets, modelling was more successful for foldon-1 than for foldon-2 (Figure 4.11A). Correct models of foldon-1 were generated for 60 targets (71%), compared to 45 targets (54%) for foldon-2 (Table 4.2). Similar results were achieved for the Training set targets (Appendix Table C.3). This high modelling success rate for foldon-1 is consistent with the results seen on the smaller Long Single-domain set targets (Section 4.3.2), reinforcing the relevance of directionality to protein structure.

In order for SAINT2-ScafFoldOn to be useful, it is crucial to be able to identify cases for which modelling of foldon-1 has succeeded or failed. To estimate whether correct models had been produced for foldon-1 and foldon-2 we used RFQAmodel, which was trained on the full-length models of the Training Set targets, as described in Chapter 2. The results of modelling and RFQAmodel are summarised in Table 4.2.

Of the 84 targets in the Long Validation set, modelling of foldon-1 was predicted to have failed for 12 targets, none of which had a correct model in the top five highest-ranking models. Foldon-1 was predicted to have been successfully modelled with High confidence for 28 targets (Table 4.2). Of these, 23 had a correct highest-ranking model (82% precision), and at least one of the top five highest-ranking models was correct for 25 targets (90% precision).

While RFQAmodel predicted that foldon-1 was modelled successfully for 54 targets with High or Medium confidence, for foldon-2 this was the case for only 8 targets — the vast majority were categorised as Low confidence or modelling failures (Table 4.2). Like SAINT2, RFQAmodel therefore achieves a better performance on foldon-1 compared to foldon-2 (Figure 4.11A); this adds further evidence to the relevance of sequential folding to protein structure prediction, and confirms that SAINT2-ScafFoldOn should be applied in the biologically-relevant N to C direction.

For the 54 Long Validation set targets for which foldon-1 was predicted to have been modelled successfully with High or Medium confidence, we selected the five highest-ranking models to use as segments for the prediction of foldon-2 (the

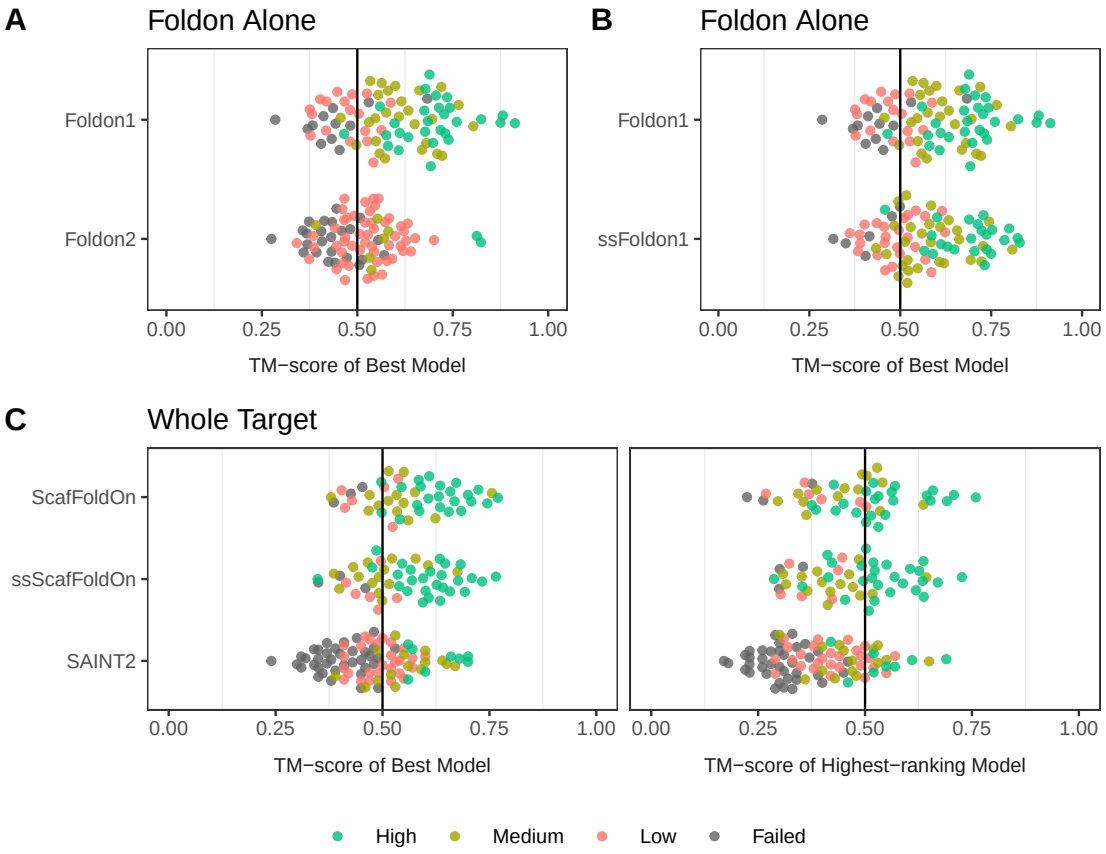

Figure 4.11: Performance of SAINT2 and SAINT2-ScafFoldOn on the Long Validation Set targets.
A) Comparison of the TM-score of the best model produced for foldon-1 and foldon-2, assigned using sequence-based FoldUP Comparison. **B)** Comparison of the TM-score of the best model produced for foldon-1 when identified using structure-based FoldUP ("Foldon1") or by sequence-based sFoldUP ("ssFoldon1"). **C)** Comparison of the TM-score of the best model (left) or the highest-ranked model according to RFQAmodel (right), when models are generated using SAINT2-ScafFoldOn with structure-based foldons ("ScafFoldOn"), SAINT2-ssScafFoldOn with sequence-based foldons ("ssScafFoldOn") or SAINT2 on the entire structure at once ("SAINT2"). Each point represents a single target, and colours indicate the confidence category assigned by RFQAmodel.

remainder of the structure) using SAINT2-ScafFold. For each of these foldon-1 models, 500 full-length models were generated, resulting in a total of 2,500 models per target. At least one of these full-length models were correct for 46 of the 84 targets, compared to 37 when using SAINT2 on the full-length target (Table 4.2).

For the prediction of unknown protein structures, it is crucial not only that correct models are generated, but that these can be identified without knowledge of the native structure. When RFQAmodel is applied to the 2,500 models produced

by SAINT2-ScafFoldOn for each of the 46 targets, 23 targets are predicted to have been successfully modelled with High confidence, of which 17 have a correct highest-ranking model and 21 have a correct model in the top five models (Table 4.2). This is compared to just 8 High confidence targets for the 500 models produced using SAINT2, with comparable precision.

Sequence-based SAINT2-ScafFoldOn on the Long Validation Set

The above section describes how the SAINT2-ScafFoldOn protocol using structure-based foldons improves the prediction and model quality assessment of targets of 150 residues or more. However, a sequence-based method of defining foldon boundaries is required when predicting unknown protein structures. On our small Long single-domain set, we found that using predicted contacts was promising but limited to targets with many predicted contacts, while simply dividing structures into 150 residue foldons led to some improvement. While the availability and accuracy of predicted contacts varies between targets, secondary structure prediction is reliably accurate (Section 1.3.1). For our Long Validation set targets, we used ssFoldUP, a foldon identification method based on predicted secondary structure that attempts to divide the structure within a region of coil close to the middle of the structure, while avoiding coil regions between beta strands (see Section 4.2.2). While this method does not ensure that the foldons are self-contained, it results in foldons of approximately equal size, with minimal disruption to secondary structure elements.

Of the 87 targets in the Long Validation set, 84 were divided into foldons using this method. For two targets, the resulting foldon boundary was identical to the structure-based FoldUP foldon boundary. A comparison of the FoldUP and ssFoldUP boundaries for all 87 targets is shown in Appendix Figure C.4.

Using this method of foldon identification — which is based only on the sequence — we evaluated the performance of SAINT2-ScafFold compared to SAINT2 for the Validation set targets. When 500 models are generated for foldon-1 of each of the 84 Validation set targets, correct models are produced for 56 (64%) (Table 4.2). This is only a slightly lower success rate than when using the structure-based

FoldUP (71%), and the distribution of the TM-scores of the best model produced for each model is similar (Figure 4.11B).

We then produced full-length models for the 51 of these targets for which foldon-1 was predicted to have been successfully modelled with High or Medium confidence. For each of these 51 targets, a further 500 models were generated using SAINT2-ScafFold from each of the five highest-ranking models of foldon-1. Among the resulting 2,500 full-length models for each of the 51 targets, 31 had at least one correct model; this is fewer than the 34 targets correctly modelled when using SAINT2 to predict the whole structure at once. However, 27 are predicted to be correct with High confidence using RFQAmodel, with a precision of 70%, compared to the 8 targets predicted to be correct with High confidence using SAINT2 alone.

These results demonstrate that SAINT2-ScafFoldOn achieves a large improvement over standard SAINT2 on the full-length target; this is the case even when foldons are determined from sequence rather than using a structure-based method (Figure 4.11C). Using only sequence information, SAINT2-ScafFoldOn results in a High confidence prediction for 19 targets that were previously intractable.

4.3.4 SAINT2-ScafFoldOn on a very long target structure

Our SAINT2-ScafFoldOn protocol, in which the target is divided into two foldons that are predicted sequentially, enabled the prediction of more of the 150–250 residue Validation set targets than using SAINT2 alone. It may be possible to improve the prediction of even longer targets by dividing them into more than two foldons. To explore this, we chose a very long target from the Long Single-domain set as a case study.

1VFF is a 423 residue target with a TIM beta/alpha-barrel fold that is classified as single-domain by both SCOP and CATH. No correct models were produced when using standard SAINT2 with 10,000 models; the best TM-score achieved was 0.43 (Figure 4.12, SAINT2). When considering the native structure, FoldUP divided 1VFF into two foldons of 208 and 215 residues. Using the SAINT2-ScafFoldOn

protocol with these foldons, as described in Section 4.2.3, two models with a TM-score of 0.5 were produced out of 10,000 models.

To divide 1VFF into three foldons without the use of the native structure, we used sequence-based ssFoldUP: rather than choosing a single foldon boundary close to $L/2$, two foldon boundaries were chosen, close to $L/3$ and $2L/3$. The resulting foldons were 144, 136 and 143 residues long.

After 500 models of foldon-1 were produced, the best TM-score achieved was 0.58 (Figure 4.12, foldon-1). RFQAmodel predicted modelling success with Low confidence; nevertheless, the five highest-ranked model included two models in the correct fold, with TM-scores of 0.53 and 0.58.

From each of these five highest-ranked models, 500 models of foldon-2 were generated using SAINT2-ScafFold (Figure 4.12, foldon1+2). RFQAmodel predicted modelling success with Medium confidence, and the top five highest-ranking models contained two correct models, including the best of the 2,500 models (TM-score 0.55).

Finally, the 500 models of the third foldon were generated against each of these top five highest-ranking models (Figure 4.12, full-length). Two of the resulting 2,500 models had a TM-score of exactly 0.5; RFQAmodel predicted modelling success with Medium confidence and no correct models were identified in the top five highest-ranking models. Nevertheless, the distribution of TM-scores was higher than that of the 10,000 models produced using SAINT2 on the full-length target (Figure 4.12), and for many models and the first two foldons of structure were in the correct fold. The improvement from using SAINT2-ScafFoldOn was greater than simply using additional fragment replacement moves, which increased the TM-score of the best model to just 0.47 (Appendix Table C.2). This case study demonstrates that SAINT2-ScafFoldOn could potentially be applied to very long targets with multiple foldons.

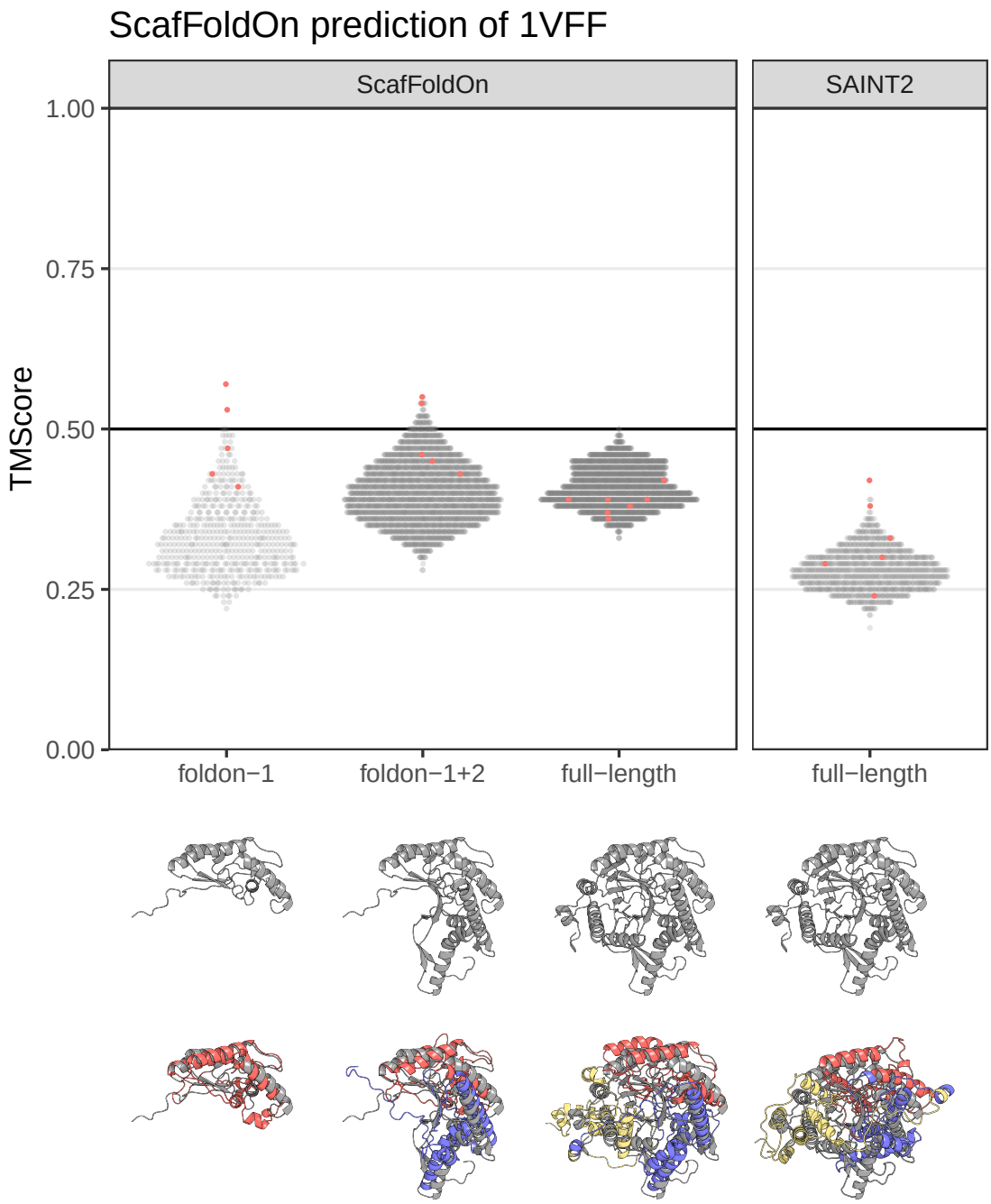

Figure 4.12: Performance of SAINT2-ScafFoldOn with three ssFoldUP foldons for the 423 residue protein 1VFF.
The TM-score of all models produced for foldon-1, foldon-1+2, and the full-length structure. Each model is shown as an individual grey point. The five highest-ranking models according to RFQAmodel are highlighted in red. Underneath, the corresponding native structure is shown in grey, and the best model (by TM-score) is shown superimposed with the native structure. Foldon-1, 2 and 3 are coloured red, blue and yellow.

4.4 Discussion and Future Work

Like many protein structure prediction methods, our software, SAINT2, successfully predicts small domains but performs poorly on proteins over 150 residues long. Inspired by the foldon hypothesis for how proteins navigate conformational space *in vivo*, we investigated whether long proteins can be divided into smaller, compact substructures that would be more tractable for protein structure prediction. We found that suitable substructures, referred to as foldons, can be identified from the protein structure using a physical and knowledge-based score. This score evaluates how well the features and topology of a structure resemble those of known proteins, and is the same score used in SAINT2. These foldons do not necessarily correspond to the small, cooperatively-folding structures seen *in vivo*, but can be predicted semi-independently *in silico* to reduce the search space.

Recently, a foldon-inspired approach to template-free protein structure prediction, UniCon3D, was developed (Bhattacharya et al., 2016). This method uses a united-residue probabilistic model, in which the bond angles and lengths separating C-α atoms, side chains and peptide groups, are sampled from continuous distributions. The target protein is divided into foldon units, assigned as one or more secondary structure elements of at least 20 residues. Prediction then proceeds by extruding and minimising one foldon at a time from the N- to C-terminus. For comparison, an implementation with conformational resampling of random 1-15 residue stretches of the full-length protein was used to emulate the conventional, non-sequential approach. The sequential, foldon-based method was found to produce more accurate models with lower energy conformations compared to the non-sequential implementation, for a test set of six small proteins.

In this chapter, we described Structure prediction of the N-terminal foldon-1 in isolation - which means generating partial protein structures using incomplete sequence information - increased the quality of the models in all cases. This is not the case for the prediction of the C-terminal foldon-2, which instead is improved by the presence of the rest of the protein chain. Extrusion against a constant foldon-1

that is in the correct overall fold also marginally improves the quality of foldon-2 models. This adds to evidence that protein folding has an element of directionality, which is relevant to computational structure prediction.

Extrusion of foldon-2 against a good model of foldon-1 would be a more efficient way of exploring the search space than predicting the full-length protein all at once. We explored this possibility using SAINT2-ScafFold to build models of the foldon-2 against the five best models produced for foldon-1. On a small set of ten long single-domain targets, we found that this method enabled the correct prediction of six previously intractable targets, and greatly enriched the model populations with correct structures.

The structure-based FoldUP method, which uses the SAINT2 score, identifies foldons that are compact, protein-like, and favoured during protein structure prediction. However, for use in template-free structure prediction where there is no prior structural knowledge, it is essential that foldons can be identified from sequence alone.

Maximising intra-foldon contacts shows some correlation with the foldon prediction by FoldUP (Appendix Figure C.1, suggesting that foldons could be detected in this fashion. Existing methods have been developed to identify the optimal partitioning of protein chains into domains based on predicted contact information (Rigden, 2002; Sadowski, 2013). As discussed in Section 4.1.3, a range of other information has also previously been used for sequence-based domain boundary prediction, including hydrophobicity and the coverage of multiple sequence alignments, which could supplement predicted contact information. These descriptors, along with predicted secondary structure information, could also be used to inform the selection of suitable foldon boundaries; a recent method of domain boundary prediction combined such information with a deep neural network approach(Q. Shi et al., 2019). Further work should investigate whether using FoldUP on a small preliminary set of clustered models can also give information on possible foldons. This would be a computationally expensive approach, but

would enable the incorporation of our structure-based technique, FoldUP, to detect protein-like substructures in a prediction scenario.

Accurate prediction of foldon boundaries may also increase the number of predicted contacts available. The number of predicted contacts is correlated with improved prediction quality (Ovchinnikov, Park, et al., 2017), although the range and distribution of these contacts (Kim et al., 2014), as well as the structure of the target protein (Adhikari et al., 2015), may also have an impact on performance. During contact prediction, columns with low coverage in the multiple sequence alignment are discarded in order to reduce noise. Carrying out contact prediction on foldon sequences individually may yield additional intra-foldon contact information by reducing gaps where sequences only align to one foldon.

As a preliminary investigation, we generated predicted contacts for the 40 individual foldon sequences of the 20 Long Single-domain proteins, and compared them to the intra-domain predicted contacts generated using the full-length sequences. Two examples of the resulting contact maps are shown in Figure 4.13. We found that additional intrafoldon predicted contacts were obtained for all targets, with an average increase of 31%. However, predictions were of poorer quality; a larger proportion of these contacts were incorrect and the estimated PPV of these contacts was lower on average than for those produced using the full-length sequence (Figure 4.14). We found that contact prediction using partial sequences is inherently more noisy; accuracy could be improved by selecting predictions using a more stringent estimated PPV cutoff.

Where foldon boundaries cannot be inferred, the results of predicting N-terminal 150 segments in isolation suggest that simply breaking the protein into smaller segments might be enough to improve on the full-length prediction. When additionally guided by predicted secondary structure, this method of foldon division resulted in similar foldons to those identified using structurally-derived approach on a set of 87 long single-domain targets.

A caveat of this protocol is that building from only five models of foldon-1 significantly reduces the search space; it is crucial that we select good models

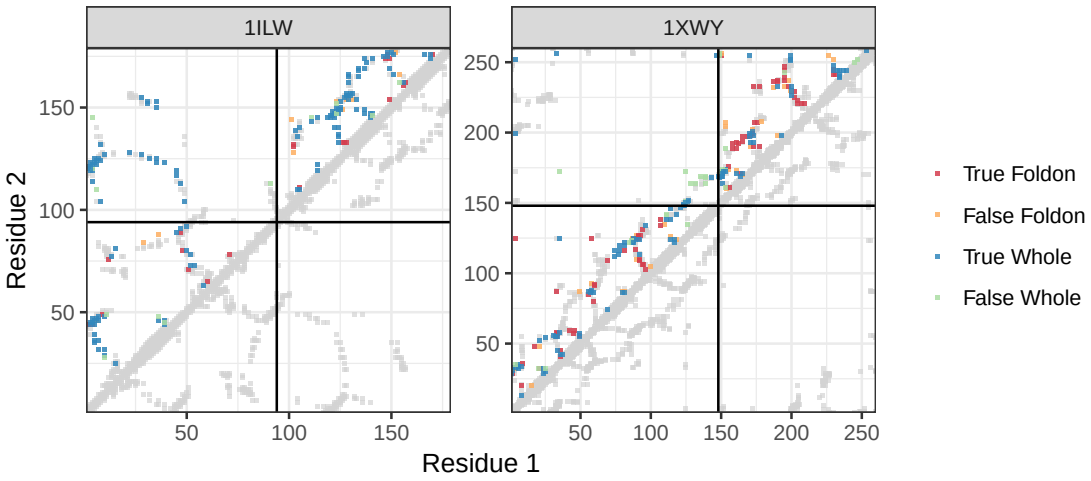

Figure 4.13: Additional intra-foldon predicted contact information from using foldon sequences individually.
Example contact maps for 1ILW and 1XWY, which have 179 and 260 residues, respectively, and are considered single-domain by both SCOP and CATH. The true residue-residue contacts are shown in light grey. True- and false-positive predicted contacts generated using the full-length protein sequences are shown in blue and green, respectively. Additional intra-foldon predicted confacts from prediction using the foldon sequences individually are shown in red for true positives and orange for false positives. The foldon boundaries assigned by FoldUP are indicated with black lines.

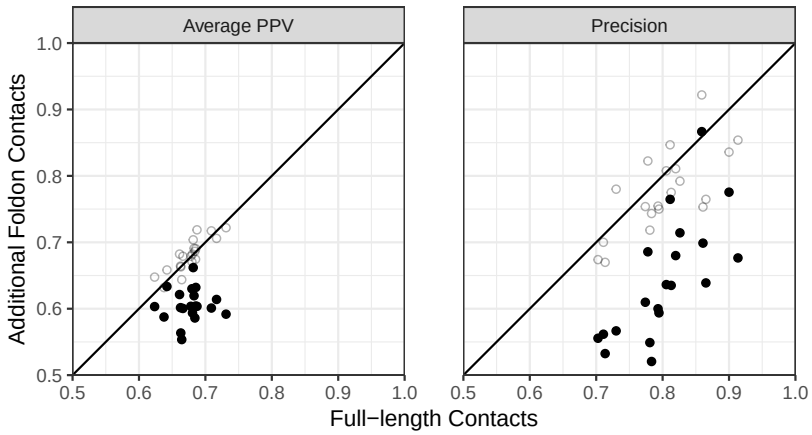

Figure 4.14: Comparison of the quality of predicted contacts generated using the full-length sequence (horizontal axis) and the additional intra-foldon contacts generated using the foldon sequences individually (vertical axis).
The equivalent comparison using all the contacts produced using foldon sequences individually, rather than just the additional contacts, is shown with light grey circles. The Average PPV value is the average of the estimated PPV values output by MetaPSICOV for each predicted contact. Precision refers to the number of true positives divided by the total number of predicted contacts.

with which to continue prediction. Furthermore, it would be more efficient not to expend further computation on targets for which correct models of foldon-1 have not been produced. In order to use the SAINT2-ScafFoldOn protocol in structure prediction, correct structures of foldon-1 must be identified without the use of TM-score. Quality assessment of predicted protein structures is difficult (Kryshtafovych, Barbato, et al., 2014), but we have shown that prediction of foldon-1 individually improves model ranking using the SAINT2 score. The 10 models with the lowest SAINT2 scores have better average SAINT2 scores, are more likely to have a TM-score over 0.5 and are closer to the best model when the foldon is predicted alone. Likewise, RFQAmodel was able to identify correctly modelled N-terminal foldons (foldon-1) with High or Medium confidence with high precision, as well as discarding modelling failures, for foldons identified by sequence alone.

Using the standard SAINT2 protocol, only 8 of our 87 Long Validation targets were predicted to be successfully modelled with High confidence according to RFQAmodel, of which 6 had a correct highest-ranking model. We assessed the performance of our SAINT2-ScafFoldOn protocol based on sequence alone, with foldon boundaries predicted using ssFoldUP and foldon-1 models selected using RFQAmodel. Using this protocol, the number of High confidence targets increased from 8 to 27, of which 19 had a correct highest-ranking model. The proportion of targets classified as High confidence, 31%, is close to the proportion attained for targets under 150 residues (37.6%). SAINT2-ScafFoldOn therefore successfully increased the modelling success rate for long proteins.

We have found that foldon-1 tends to be more tractable than foldon-2, in terms of both structure prediction and model quality assessment. We can extend this concept to assess partially correct structures. It may be possible to determine whether foldon-1 is likely to be correct, even if the whole structure is not. For large proteins, where experimental structure determination is difficult and structural knowledge is often scarce, partial structure prediction could prove useful.

It is likely that some protein structures should be divided into more than two foldons. In our Long Single-domain set, 15 foldons were longer than 150 residues,

and correct models were generated for only two of these. Further division might improve results, particularly for very long proteins and those for which no correct models for foldon-1 were generated. Determining the number of divisions is a complication for both structure- and sequence-based domain assignment, with some methods having a tendency to over-cut while others under-cut when compared to manual assignments (Kolodny et al., 2013; Sadowski, 2013). Using our sequence-based ssFoldUp approach for foldon identification avoids this problem, as as the chain is divided based only on secondary structure elements and a specified length that is feasible for protein structure prediction.

As a case study, we divided a 423 residue target into three foldons using the sequence-based ssFoldUP. When the whole structure is predicted using SAINT2, the mean TM-score produced is 0.28 and no correct models are produced. SAINT2-ScafFoldOn produced higher quality models with a mean TM-score of 0.41, many of which had the correct fold for the first and second foldon. Further work should investigate whether there is an optimal foldon length for structure prediction, and which method of making multiple divisions is appropriate. RFQAmodel can be used to determine whether correct models have been generated; the RFQAmodel confidence could additionally be explored as a basis to decide when to cut foldons further.

5 Conclusions

Contents

5.1 Protein structure prediction of previously intractable targets

In this thesis, we have investigated extensions to SAINT2 to enable the prediction of targets that were previously intractable, due to computational limitations or large conformational space.

5.1.1 RFQAmodel improves computational efficiency

The high computational cost of protein structure prediction, particularly fragment-based methods, has been a bottleneck in large-scale prediction studies (Ovchinnikov, Park, et al., 2017; Michel, Menéndez Hurtado, et al., 2017). Many methods generate up to hundreds of thousands of models, for example, the optimal number of models to generate per target using SAINT2 was estimated at 10,000 (de Oliveira, Law, et al., 2018). Furthermore, it was difficult to identify targets for which correct models had been generated among these large ensembles.

Using our model quality assessment method RFQAmodel, we are able to identify targets for which 500 models are sufficient (Chapter 2). Guided by RFQAmodel's confidence categories, this allows us to focus computational effort on more difficult targets and enables modelling success for a greater number of targets with an equivalent amount of computation. This reduction of up to 95% of the computation and storage space is particularly critical as the number of targets predicted in large-scale protein structure prediction studies grows, and researchers become increasingly conscious of the financial and environmental sustainability of computation.

5.1.2 SAINT2-ScafFold can accurately model missing terminal regions of structure

Template-free protein structure prediction is essential where there is no homologous structure available for a given target. However, it is often the case that there is an incomplete known structure or an available homologous structure that does not cover the entire target sequence (Kryshtafovych, Schwede, et al., 2019; Law, 2017). The unique sequential nature of SAINT2 is well-suited to this application, as it is possible to perform template-free prediction by building from a given structure (Law, 2017).

We have described a protocol in which the terminal missing regions of crystal structures are predicted against the known structure (Chapter 3). Prediction in this context results in better models than when the missing region is predicted independently.

Furthermore, while globally correct models were produced for only 43% of the Crystal Structure Test set targets, when the models were assessed using a modified version of RFQAmodel, all of the correct highest-ranking models were identified with high or medium confidence. The ability to accurately distinguish incorrect models makes the protocol more useful for real applications, in which it is not known whether a particular terminal region can be accurately modelled. The local quality was more challenging both for structure prediction and model quality assessment. It may be possible to improve on this promising start with a Training set of more representative example cases.

5.1.3 SAINT2-ScafFoldOn improves prediction of long proteins

Long proteins are another challenge that has hindered protein structure prediction. SAINT2 produced a correct model for only 39% of Validation set targets between 150 and 250 residues (Chapter 2), and in a previous study SAINT2 successfully modelled just 2 of 15 targets over 250 residues (de Oliveira, Law, et al., 2018). The most recent CASP found a large improvement in the modelling of domains over 150 residues long, but the best models are not as good as those for smaller domains (Abriata et al., 2019).

We have described SAINT2-ScafFoldOn, where long proteins are divided into smaller foldons based on length and predicted secondary structure (Chapter 4). These foldons are predicted sequentially from the N to the C terminus: the second foldon is predicted against the five highest-ranked models of the first. This process of foldon prediction was made possible due to the ability of RFQAmodel to identify good models among the top five highest-ranking models. Using this protocol, correct models were generated and identified for more targets than when using SAINT2 on the full-length structure. Furthermore, using a 423 residue target as a case study, we demonstrated that SAINT2-ScafFoldOn can improve the tractability of very long targets by dividing the structure into more than two foldons.

In the implementation of SAINT2-ScafFoldOn, several challenges remain. One such challenge is the best way to identify foldons from protein sequence. Integrating predicted contact information into foldon boundary determination is likely to become more feasible as the number and accuracy of predicted contacts improves, particularly as methods become less reliant on the depth of the multiple sequence alignment (Shrestha et al., 2019). Future work should determine the optimal procedure for determining foldon boundaries from sequence, which may include considering predicted contacts and solvent accessibility in addition to the foldon length and secondary structure.

5.2 Future perspectives

Over the past three years, the field of protein structure prediction has seen considerable advances. The application of deep learning methods has improved the accuracy of contact prediction (Shrestha et al., 2019) and structure prediction (Kryshtafovych, Schwede, et al., 2019; Kandathil et al., 2019b; Greener et al., 2019) so dramatically that the CASP13 assessors recently described the problem of predicting the topology of monomeric proteins as essentially solved (Kryshtafovych, Schwede, et al., 2019).

In this context, it is important to consider the future role of fragment-based methods. In particular, the advantages of such methods, and how they might be extended to different applications and contribute to new avenues of discovery.

5.2.1 Ensembles of models

Deep-learning methods have demonstrated high performance while only producing a single model or several very similar models (Greener et al., 2019). While this can be preferable for a user, there are potential advantages to fragment-based approaches that generate ensembles of models.

Firstly, for some targets, fragment-based methods have been found to produce a better quality model than the deep learning methods, but this model cannot be identified among the many possible models (Greener et al., 2019). This underscores the importance of model quality assessment methods for fully utilising the predictive power of such methods. In addition, having a range of candidate models may be preferable for some applications, such as solving crystallographic structures using molecular replacement (Thomas et al., 2017).

Secondly, ensembles of models produced by fragment-based methods have been found to capture some information about the multiple conformations that may exist for a target sequence (Kosciolek et al., 2017; Palopoli et al., 2016). This issue of multiple conformations was highlighted in the two most recent CASP competitions: experimental data from small angle X-ray scattering (SAXS), a high-throughput method that captures distances between electron pairs of a protein in solution, revealed that the conformations adopted in solution were inconsistent with the crystallographically determined structure for seven of the eleven SAXS-assisted targets in CASP13 (Hura et al., 2019; Ogorzalek et al., 2018). This reflects the wider issue in protein structure prediction that proteins are inherently flexible, and many proteins exist in multiple conformations or contain disordered regions that are often crucial to their function (e.g. Monzon et al., 2016; Garton et al., 2018; Dunker et al., 2002). Furthermore, machine learning methods that are trained on crystal structure data may bias methods away from solution-state conformations (Hura et al., 2019). The ability of fragment-assembly methods to capture multiple conformations is currently not rewarded under the CASP-style evaluation system, where models are compared to a single crystal structure. Such behaviour may become valuable as the accuracy and reliability of protein structure prediction improves, and consideration of the dynamic nature of protein structures may become increasingly feasible.

5.2.2 Improvements to SAINT2

Much of the improvement seen in CASP13 was the result of deep learning contact prediction methods that predict inter-residue distance probabilities, rather than

binary contact prediction (Kryshtafovych, Schwede, et al., 2019). As others have discussed, the biological basis for these predictions is not clear, and may not be possible to analyse; however, the additional information led to improvements in structure prediction methods (Kryshtafovych, Schwede, et al., 2019). The contact component of the SAINT2 score could easily be adapted to handle inter-residue distances, which may result in improvements in its performance, as it has for other fragment-assembly approaches (Senior et al., 2019).

5.2.3 Extensions of protein structure prediction tools

The flexible frameworks of SAINT2, SAINT2-ScafFold and RFQAmodel enable us to develop protocols for structure prediction scenarios that are not possible with other tools.

We investigated one example of this in Chapter 3, in which we attempted to predict the terminal missing regions of structures. While we achieved a good performance on simulated cases, further work is required to improve the performance on real cases.

5.2.4 Combined template-based and template-free modelling

We have established that structure prediction is improved by the presence of a partial crystal structure or a correctly modelled portion of the target chain (Chapter 3 and Chapter 4). Where a crystal structure or homologous template is available for a part of a target protein, a combination of template-based and template-free modelling in a single fragment-based framework would be convenient, and could result in improved modelling of the missing region.

The top-performing method in CASP13, AlphaFold, used their free-modelling approach for template-based targets — for which a template structure can be identified by sequence search (Senior et al., 2019). A template is therefore not explicitly used, although the template structure may have been included in the training set. The performance was often worse on template-based targets than

free-modelling targets with equivalent available sequence data (B_{eff}), and AlphaFold was outperformed by comparative modelling methods for a number of template-based targets. This suggests that deep learning methods are currently unable to compete with homology modelling when a template is available. A fragment-based approach, which can explicitly incorporate known regions of structure into the modelling process, could therefore be useful.

In Chapter 3, we generated Flib-Flex fragment libraries that incorporate information from a known region of structure. We found that sampling both the known and missing regions of structure using this fragment library led to improved performance compared to standard SAINT2, but template regions of the model were not sufficiently accurate to compete with homology models. Explicit inclusion of template structure fragments is likely to result in immediate improvement. Beyond this, larger fragment lengths could be explored for the regions of known structure, as well as more complex sampling strategies. This framework could then be extended to predict missing regions anywhere in the target structure.

5.2.5 Structural models for annotation

As the amount of available sequence information increases, a related area of interest is the prediction of protein function. Protein structures are more informative for the prediction of function and functionally important residues than sequences alone (Gligorijevic et al., 2019; K. Wang et al., 2008; Fetrow et al., 2001). Where solved structures are not available, using even low-resolution structural models in the overall correct fold can improve performance, facilitating structural and functional annotation on a genome-wide or microbiome-wide scale (C. Zhang et al., 2017; Y. Wang et al., 2019). Unlike most existing model quality assessment methods, RFQAmodel predicts whether a model has been produced that is in the correct overall fold. This classification approach was better able to identify successfully modelled targets than a regression model trained to predict the actual TM-score, and therefore could increase the number of targets that can be annotated.

Also of interest are expressed sequence tags (ESTs); these DNA fragments often contain only part of the coding region for a protein (Parkinson et al., 2009). One study that investigated whether structures could be predicted for such partial sequences found that it was possible for sequences that were at least half the full length of the coding region, but that fragment-based methods performed poorly compared to template-based (Laurenzi et al., 2013). We found that partial-length sequences could be predicted accurately using SAINT2, with better performance for N-terminal substructures (Chapter 4). Future work could investigate whether SAINT2 could therefore be used to improve the annotation of EST data, which is a large source of sequence and gene information.

5.2.6 Biologically-inspired protein structure prediction

Fragment-based methods involve the physical exploration of conformational space, which enables the inclusion of biologically-inspired features. SAINT2 uses a sequential procedure that is inspired by cotranslational folding and improves efficiency and accuracy of structure prediction (de Oliveira, J. Shi, et al., 2017). We found that using a foldon-inspired approach, long structures can be divided into more tractable sub-structures (Chapter 4). We also found that prediction of the N-terminal foldon alone increases tractability, while the C-terminal foldon achieves a better performance when predicted against the N-terminal foldon. This reiterates the importance of biologically-relevant directional prediction.

Further understanding and inclusion of efficient *in vivo* search mechanisms may allow the computational prediction of previously intractable targets. Possible mechanisms include non-uniform translation speeds, domain-wise folding, and biological spatial restrictions. In addition to varying the fragment library and search strategy, it may be possible to incorporate coarse-grained molecular dynamics into the prediction process; such methods have previously been used to predict protein structure (Cheung et al., 2018) and recapitulate folding pathways (Nissley, Sharma, et al., 2016; Nissley and O'Brien, 2018).

Multidomain proteins

Protein structure prediction methods perform more poorly on multidomain proteins (Abriata et al., 2019). Prediction software has historically approached multidomain proteins by modelling and evaluating domains separately, followed by assembly using rigid docking (Inbar et al., 2005) or by sampling the conformational space of the linkers and interfaces between the otherwise rigid domains (Wollacott et al., 2007; D. Xu, Jaroszewski, et al., 2015). While these techniques have had some success using crystal structures, they tend to fail when models are used due to incorrect domain parsing (Ovchinnikov, Kim, et al., 2016), long unstructured linkers (Wollacott et al., 2007), and modelling errors (D. Xu, Jaroszewski, et al., 2015). Compared to the complete multidomain structure, crystal structures of separately-solved constituent domains have slightly different conformations; these small variations can result in incorrect domain assembly (D. Xu, Jaroszewski, et al., 2015). Homology and template-free models are inevitably non-native, which makes evaluating docked positions especially difficult (Inbar et al., 2005).

While more recent deep learning methods show promise in the prediction of multidomain structures without dividing structures, there remain limitations (Abriata et al., 2019). Though the majority of eukaryotic proteins are multidomain (Zmasek et al., 2012), these structures are underrepresented in the Protein Data Bank (PDB) (Berman, 2000), which affects the availability of training data. Furthermore, the packing of domains in crystal structures has been found to differ from solution-state orientations; this may confound deep learning methods, which are known to be sensitive to dataset biases (Hura et al., 2019).

The top-performing method in CASP13, AlphaFold, used a fragment-assembly approach (employing a distance potential and simulated annealing) for some targets and gradient descent for others (Senior et al., 2019). Gradient descent slightly outperformed the fragment-assembly approach: the average accuracy of models produced using gradient descent was 64.4 compared to 63.4 GDT-TS, which the authors suggest may have been due to the need to predict estimated

domains separately using the fragment-based approach, which is not required for gradient descent.

Experimental evidence has demonstrated that domain-wise, cotranslational folding of multidomain proteins occurs *in vivo*, with functional activity detectable before protein synthesis has been completed (Hamlin et al., 1972; Komar et al., 1997; Frydman et al., 1999; Sakahira et al., 2002; Sánchez et al., 2004; Hsu et al., 2007; Evans et al., 2008)

Our SAINT2-ScafFoldOn protocol potentially emulates the *in vivo* folding pathway of multidomain proteins. Domains can be folded individually while retaining interdomain predicted contact information during prediction, and eliminating the need for explicit domain docking or loop closure. We tested SAINT2-ScafFoldOn with a set of long proteins that were annotated as one-domain by SCOP. However, 10 were divided into two domains by CATH. Of these, four were predicted correctly, two of which were only successfully predicted by the SAINT2-ScafFoldOn method. Future work could investigate whether this protocol can be applied to multidomain proteins.

Multichain proteins

Some proteins require the presence of a partner protein in order to adopt their native state. An example of this is Caspase-activated DNase (CAD), which folds cotranslationally but requires the presence of its inhibitor protein, ICAD, for both *in vivo* and *in vitro* folding. Preliminary investigations (not included in this thesis) reveal that the N-terminal ICAD binding domain of this protein was predicted slightly better using the non-sequential mode of SAINT2. If the relative performance of the prediction modes is indeed related to the biological folding pathway, then this may be an example of a protein whose biological folding pathway requires extra information not currently incorporated into SAINT2. The non-sequential mode may then outperform the sequential approaches as a result of wider exploration of conformational space. The functionality of folding against a previously predicted foldon or domain may be extended to allow folding against

a binding partner, which could be beneficial to proteins with folding mechanisms involving partner interactions.

Membrane proteins

Membrane proteins represent a challenge to structure prediction due to their typically large size and relatively limited structural understanding. Integral membrane proteins constitute only 4% of the protein structures in the PDB, despite representing around 30% of human genes (Almén et al., 2009) and over half of pharmaceutical targets (Bakheet et al., 2009).

Membrane proteins are inserted into the membrane cotranslationally and have unique environmental constraints inflicted by being embedded in lipid bilayers. Structural motifs within transmembrane regions are limited to α-helices or β-barrels (Tan et al., 2008). The distinctive folding mechanism, unique environmental constraints and limited transmembrane structural motifs of membrane proteins makes them an interesting case study for SAINT2.

In a recent investigation, structure prediction by SAINT2 Forward was found to outperform the Non-sequential mode for 18 out of a set of 24 membrane proteins (de Oliveira, J. Shi, et al., 2017). Future work could investigate whether the prediction of membrane structures can be improved by SAINT2-ScafFoldOn prediction and the imposition of additional biological constraints.

Biological understanding of protein folding

Finally, if emulating biological folding pathways results in improved structure prediction, the extent of improvement seen under such mechanisms may vary between protein targets that rely on different folding mechanisms *in vivo*. We may be able to infer information about the folding pathways of proteins, which are of vital importance but are experimentally challenging to study. In the same way biology has improved our computational methods for protein structure prediction, computational techniques can also inform our understanding of biology.

Appendices

A

Chapter 2 Appendices

A.1 Data Sets

A.1.1 Pfam PDB mapping

The mapping between Pfam domains and the PDB structures available on the EBI repository in Feb 2017 encompassed 453,717 protein chains from 113,618 PDB files, mapped to 8,005 unique Pfam families, from which we selected the first protein chain listed for each family as representative.

Table A.1: Properties of the 8,005 protein chains representing each of the Pfam domains mapped to PDB structures

B_{eff}	Number of protein chains	% of Total	Average length (± s.d.)	Average B_{eff} (± s.d.)
Less than 100	1,948	24.33	124 ± 121	30 ± 28
100 to 1,000	3,138	39.2	138 ± 112	429 ± 244
More than 1,000	2,919	36.46	164 ± 118	5,470 ± 8,094

Table A.2: Properties of the 4,728 protein chains with SCOPe annotations chosen to represent unique Pfam families mapped to PDB structures

B_{eff}	Number of protein chains	% of Total %	Average length (± s.d.)	Average B_{eff} (± s.d.)
Less than 100	918	19.41	124 ± 124	29 ± 28
100 to 1,000	1,498	31.68	129 ± 95	452 ± 254
More than 1,000	2,312	48.90	159 ± 112	6,030 ± 8,643

Table A.3: Properties of the 488 protein domains chosen to comprise our Training and Validation data sets.

B_{eff}	Number of protein chains	% of total	Average length (± s.d.)	Average B_{eff} (± s.d.)
Less than 100	76	25	124± 48	41± 31
100 to 1,000	164	55	134±53	455± 262
More than 1,000	248	83	136± 51	6924± 8867

A.1.2 Culling Process

We removed protein chains from the set of 4,728 proteins in Table 2 according to the following criteria:

- Remove all protein chains shorter than 50 residues (4,060 protein chains remaining).
- Remove all protein chains that are classified in SCOP as multi-domain, membrane, or small protein (3,688 protein chains remaining).
- Remove all non-crystallographic structures and those with a resolution of 2.5Åor worse (2,266 protein chains remaining).
- Remove all protein chains that contain more than one domain as annotated on the PDB mapping to Pfam (1,376 single-domain protein chains remaining).
- Remove all proteins that could not be parsed using the non-permissive parser of BioPython (Cock et al., 2009) (1,327 single-domain protein chains remaining). Extract the corresponding domain as annotated on the mapping.
- Remove all protein domains with chain breaks, incomplete sequences, and unresolved residues (606 protein domains remaining).

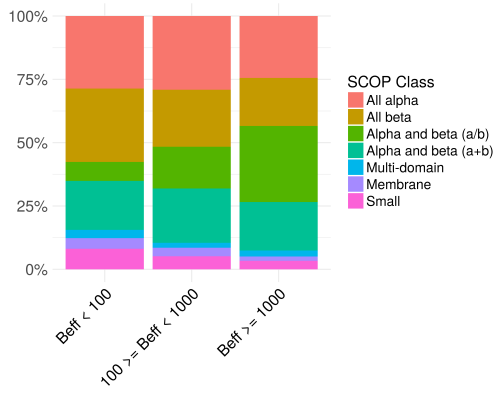

Figure A.1: SCOP classes of representative chains
Proportion of protein chains per SCOP class at different B_{eff} levels. Data is shown for the 4,728 protein chains with SCOP annotations chosen to represent unique families described on the PDB mapping to Pfam.

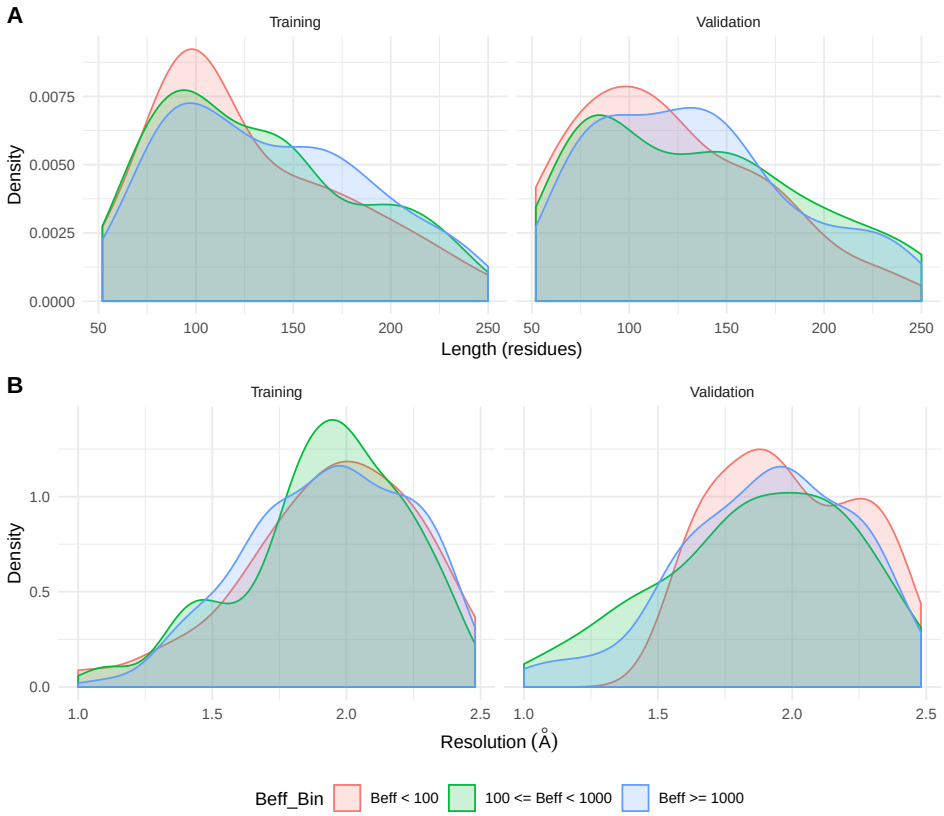

Figure A.2: Domain lengths and resolutions of Training and Validation sets.
Distribution of the A) lengths and B) resolution of protein domains in our Training (left) and Validation (right) sets according to three B_{eff} bins.

A.2 Benchmarking sequence-based descriptors

Saulo de Oliveira performed the following benchmarking of sequence-based descriptors.

A.2.1 Benchmarking secondary structure predictors

Secondary structure prediction is an essential step in most template-free protein structure prediction protocols. Errors in secondary structure prediction can have a drastic effect on modelling success, particularly if a large stretch of residues are assigned incorrectly. Recently, two methods, SPIDER3 (Heffernan et al., 2017), and DeepCNF (S. Wang et al., 2016), have used deep learning to improve the precision of secondary structure prediction. Given the importance of secondary structure prediction to accurate modelling, we compared the Q3 and Q8 precisions of three secondary structure predictors, PSIPRED (D. T. Jones, 1999), SPIDER3 (Heffernan et al., 2017), and DeepCNF (S. Wang et al., 2016) (Figure A.3). To perform this assessment, we considered the methods' ability to classify residues into three (helical, strand, or coil) or eight secondary structure types (refer to (Kabsch et al., 1983) for more details), respectively.

For B_{eff} <1000, no significant difference was observed when assessing the Q3 precision obtained by each of the methods. However, DeepCNF produced better Q3 predictions when sufficient sequence information was available (B_{eff} ≥1,000). DeepCNF produced better Q8 predictions than the other methods for all B_{eff} bins. Our subsequent analyses were carried out using the output of DeepCNF Q8.

A.2.2 Comparing prediction of sequence-based descriptors for the Training and Validation sets

We compared the Q3 and Q8 precisions of secondary structure prediction on our Training and Validation sets (Figure A.3) and found comparable Q3 and Q8 precisions between the two sets.

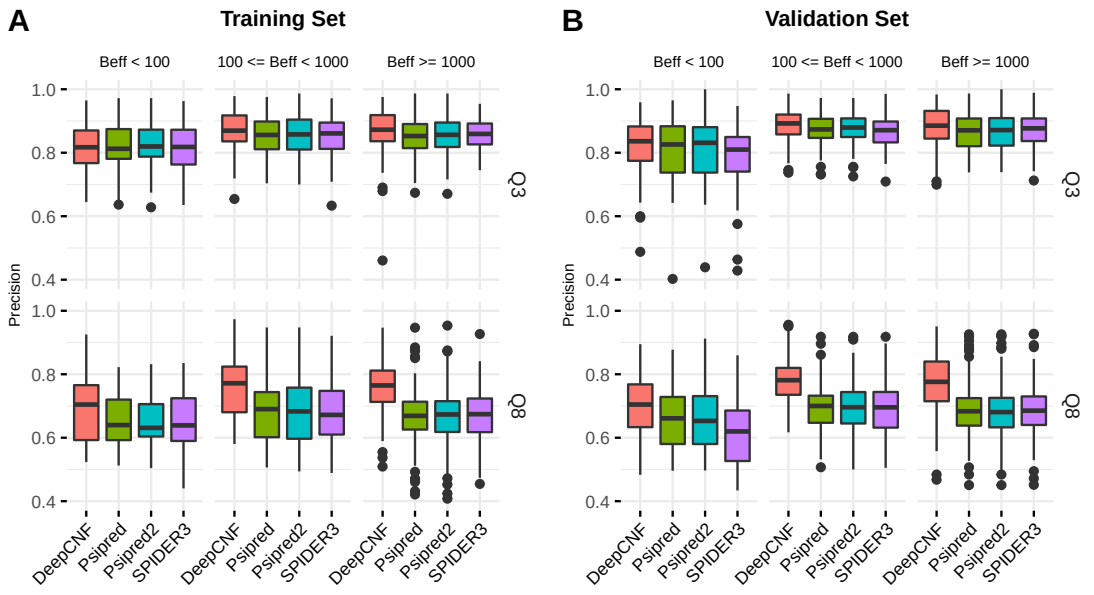

Figure A.3: Secondary structure prediction.
Precision of Q3 (top) and Q8 (bottom) secondary structure prediction according to B_{eff} for the 244 targets in each of the **A**) Training set and **B**) Validation set.

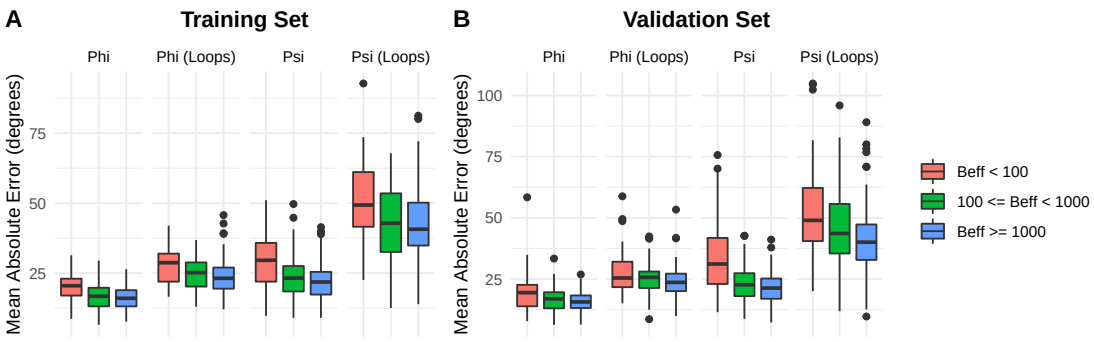

Figure A.4: Torsion angle prediction.
Mean Absolute Error (MAE) of Φ and Ψ torsion angle prediction for the 244 proteins in each of the A) Training set and B) Validation set (right). MAE was calculated across all residues or across loop residues exclusively.

We then compared the precision of Torsion Angle prediction between our Training and Validation sets. In our current modelling protocol, Torsion Angle prediction is used for building a fragment library. Previous findings suggest that poor torsion angle prediction leads to modelling failure (de Oliveira, Law, et al., 2018). To perform this comparison, we calculated the Mean Absolute Error (MAE) in degrees between predicted and observed angles for both Φ and Ψ dihedrals, across all domains in our Training and Validation sets (Figure A.4). Considering the challenges associated with predicting loop conformations during modelling (de Oliveira, J. Shi, et al., 2015), we have also assessed the MAE exclusively for loop residues of these domains. For this, loop residues were determined using the DSSP secondary structure assignments calculated previously.

Our results reveal that the precision of torsion angle prediction is slightly lower for lower B_{eff} values. Interestingly, the MAE for Ψ angles was significantly higher than for Φ angles. The difference in Φ against Ψ MAEs becomes significant when we considered the MAE for loop residues, exclusively. Comparable MAEs were observed between Training and Validation sets for both Φ and Ψ angles, considering all residues and considering loops exclusively.

Another property that has been shown to be crucial for modelling success (de Oliveira, J. Shi, et al., 2017) is the precision of contacts predicted from multiple sequence alignments. To ensure that our split between Training and Validation was balanced in this regard, we compared the precision of short-range ($<$23 residues separation), and long-range ($\geq$ 23 residues separation) predicted contacts output by metaPSICOV stage1 for all proteins in our Training and Validation sets (Figure A.5). The choice of stratifying at 23 residues separation is corroborated by previous findings suggesting that precision of long-range contacts is more important than short-range for modelling success (D. T. Jones, Singh, et al., 2015; de Oliveira, J. Shi, et al., 2017). Our analysis shows that the two data sets are comparable both in terms of precision of short-range and long-range predicted contacts across each of three B_{eff} bins.

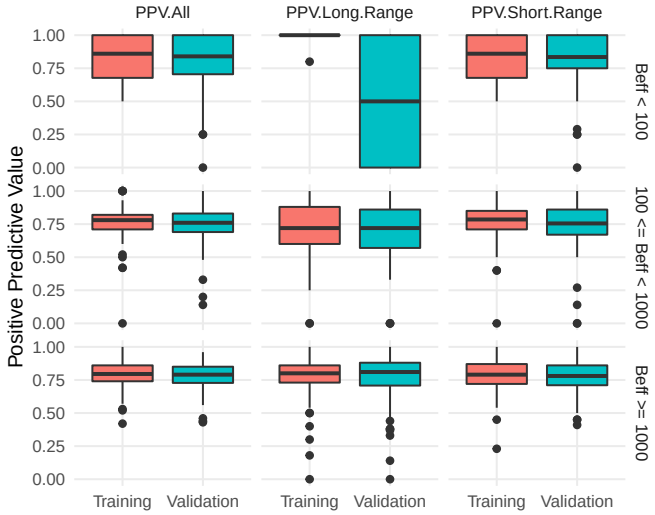

Figure A.5: Contact prediction.
Precision (Positive Predictive Value, PPV) of short-range (<23 residues separation) and long-range (≥23 residue-separation) predicted contacts as output by metaPSICOV stage1 for the 244 protein domains in our Training set and the 244 protein domains in our Validation set. Results are shown for three different B_{eff} bins.

A.3 Estimating the number of models required

First, we used SAINT2 to produce 1,000 models per target for 245 targets randomly selected from our Training and Validation sets. For each target, we sampled n models without replacement and counted the number of targets for which at least one correct model (TM-score ≥ 0.5) was included in the sample for varying values of n ($1 \leq n \leq 1,000$) (Figure A.6). We observed little improvement in the number of correctly predicted targets when sampling more than 500 models. To balance the number of modelling successes and computational feasibility, we therefore opted to produce 500 models using SAINT2 for our subsequent analyses.

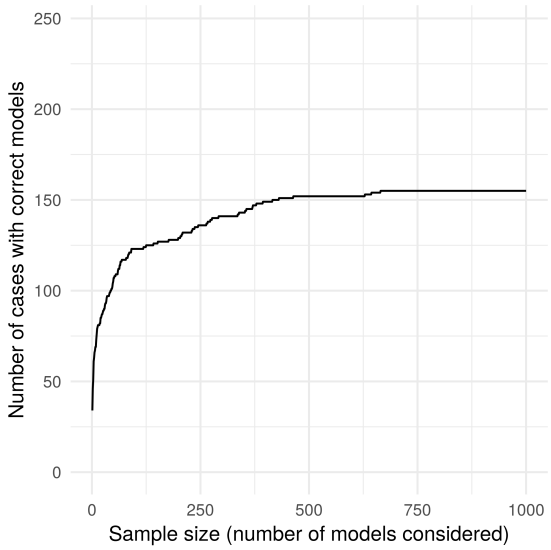

Figure A.6: The number of targets with at least one correct model (TM-Score ≥ 0.5) in a sample of all models produced by SAINT2. The sample size ranges from 1 to 1,000 models considered per target. Results are shown for 245 protein domains randomly selected from our Training and Validation sets.

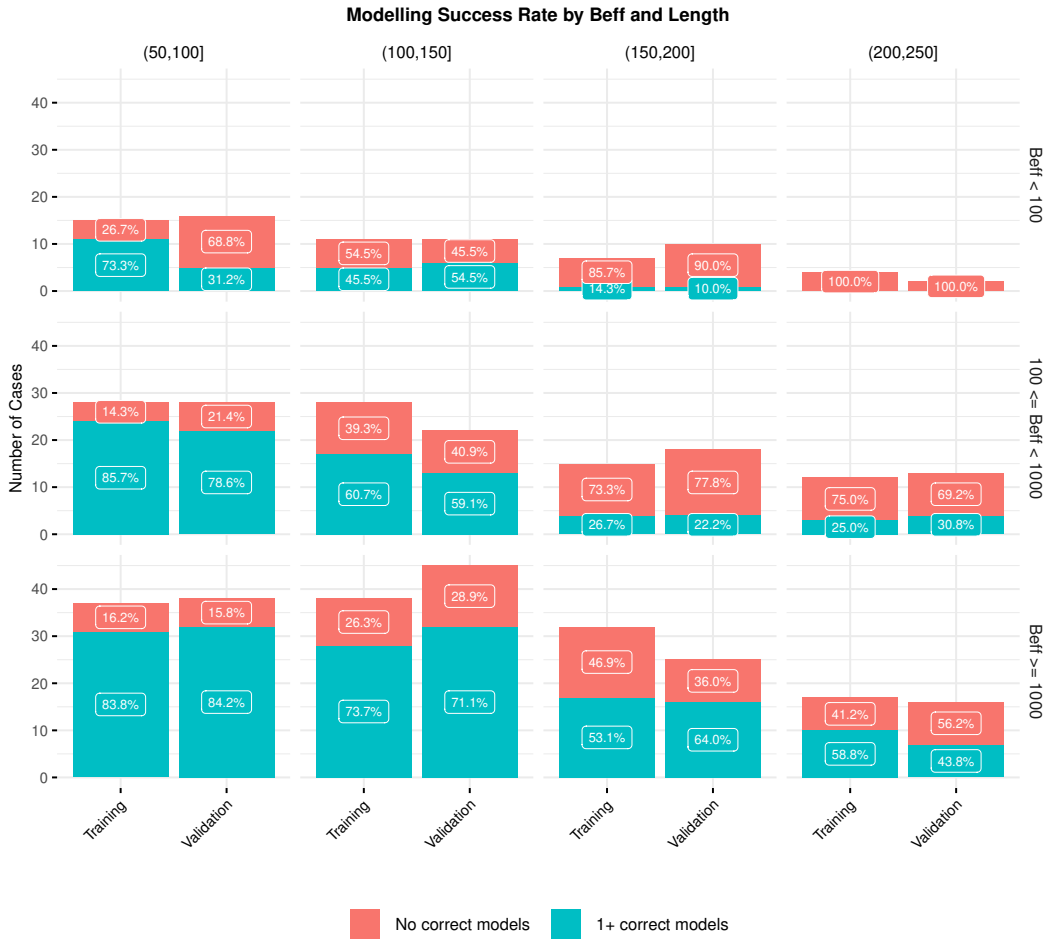

Figure A.7: Modelling success rate by both B_{eff} and length.
Modelling success rate for the 244 protein domains in each of our Training and Validation sets according to both length and B_{eff} bins. Any target for which at least one correct model (TM-Score ≥ 0.5) was produced is considered a modelling success.

A.4 RFQAmodel classifier feature importance

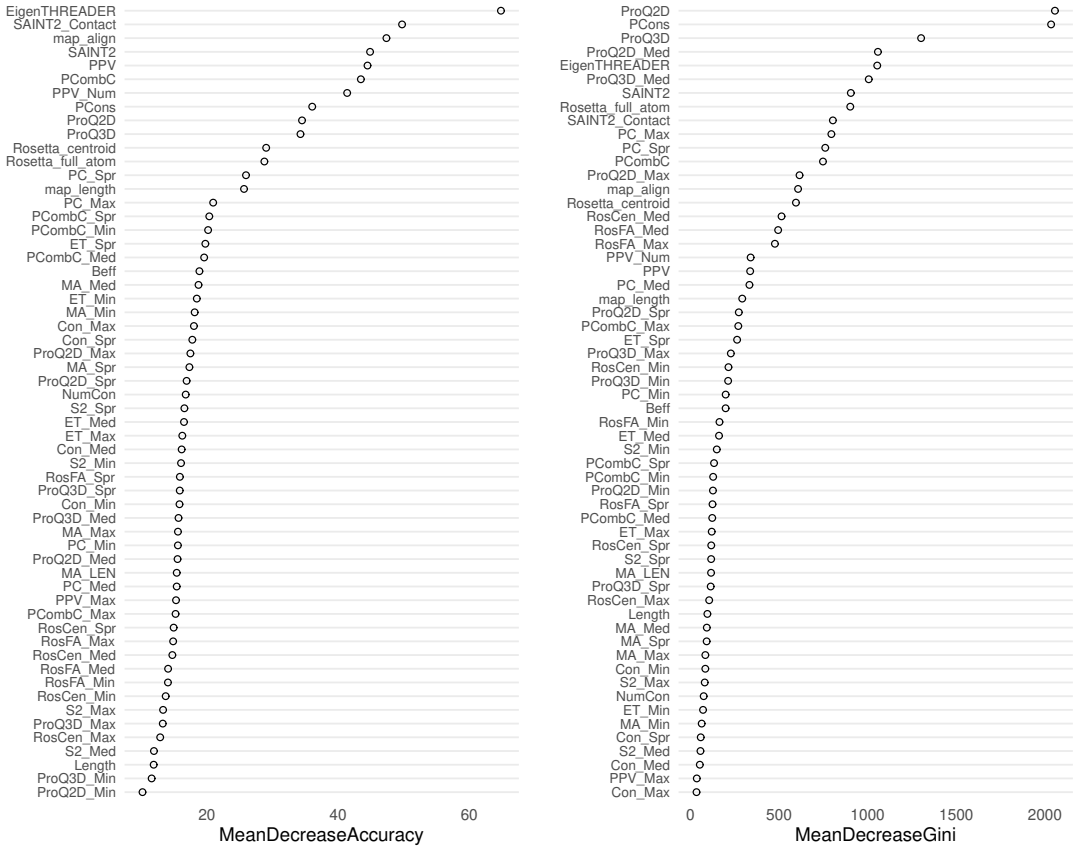

Figure A.8: Relative importance of the component features of RFQAmodel reported by randomForest.

B

Chapter 3 Appendices

B.1 Average solvent accessibility for the Missing and Opposite regions.

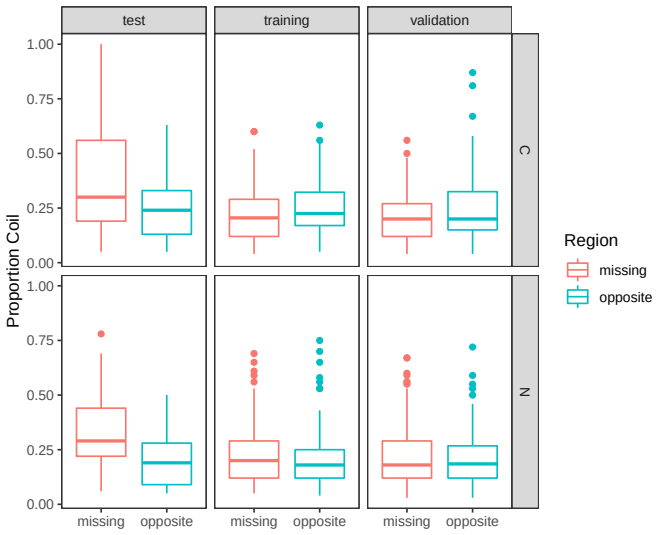

Figure B.1: Solvent accessibility for the Missing and Opposite regions. .
The distribution of average solvent accessibility for residues in the Missing region (red) and the equivalent length region at the opposite terminus ("Opposite", blue) for the Test, Training and Validation, where the Missing region is at the C- (top) or N-terminus (bottom).

B.2 Estimated relative feature importance

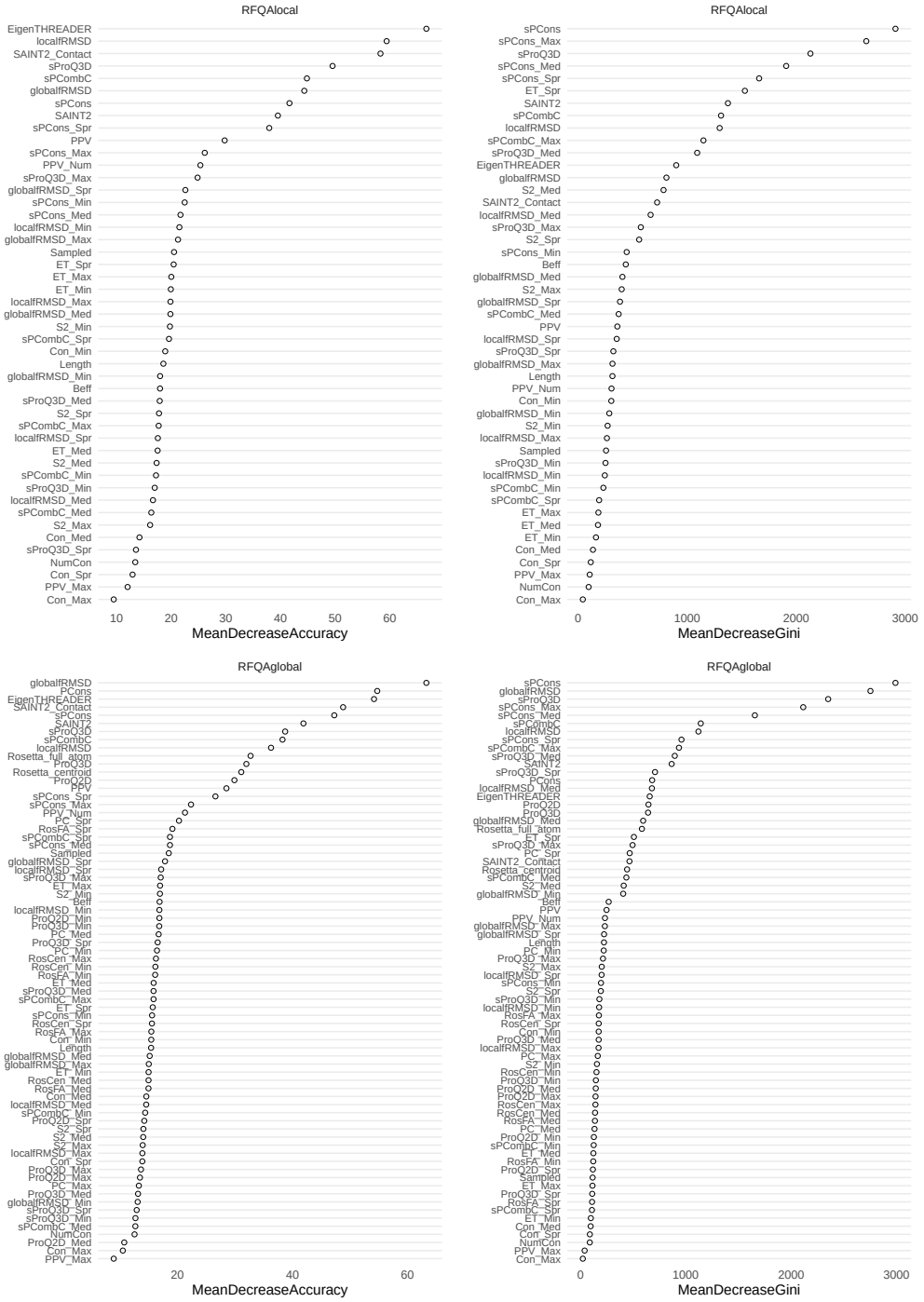

Figure B.2: Relative importance of the component features of RFQAlocal (top) and RFQAglobal (bottom) reported by randomForest.

C Chapter 4 Appendices

C.1 FoldUP and ssFoldUP pseudocode

Algorithm 1 Pseudocode demonstrating how FoldUP determines ideal foldon boundaries from the crystal structure of a protein.

min_foldon_length $\leftarrow 15$
min_$M_d \leftarrow 17.5$
$L \leftarrow$ protein length
for i in (10 **to** $L - 10$) **do**
    FoldUPscores[i] $\leftarrow \frac{\text{SAINT2score(Residues[1:}i\text{])+SAINT2score(Residues[}i\text{:}L\text{])}}{2}$
end for
$R_{min} \leftarrow$ index of minimum score
$M_d \leftarrow$ minimum of (FoldUPscores[10] $-$ FoldUPscores[R_{min}]) and (FoldUPscores[$L - 10$] $-$ FoldUPscore[R_{min}])
if $R_{min} >$ min_foldon_length **and** $(L - R_{min}) >$ min_foldon_length **then**

    if $M_d >$ min_M_d **then**
        foldon_boundary $\leftarrow R_{min}$
    end if
end if

Algorithm 2 Pseudocode demonstrating how ssFoldUP determines the foldon boundaries from the predicted secondary structure of a protein.

```
min_foldon_length ← 15
N_foldons ← number of foldons
min_coil ← 5
L ← protein length
i_min ← min_foldon_length
i_max ← L − min_foldon_length
for residue in sequence[i_min : i_max] do
    if residue is the start of a coil region then

        if length(coil) ≥ min_coil AND coil is not between beta strands then
            add coil to candidate_breakpoints
        end if
    end if
end for
for N_i in (1 to N_foldons) do
    N_i_foldon_boundary ← the candidate_breakpoint that is closest to (N_i / N_foldons) L
end for
```

Table C.1: Details and SAINT2 modelling results for the Long Single-domain set. The PDB code, full and FoldUP foldon lengths (L), and the modelling results using different protocols for each protein in the Long Single-domain set, when evaluating the whole structure (Whole) or individual foldons (foldon-1 and foldon-2). We tested three SAINT2 protocols: SAINT2 Forward on the full-length of the protein (SAINT2), SAINT2 Forward on each foldon individually (Foldon), and ScafFoldOn, building foldon-2 from the best prediction of foldon-1 (S-FoldOn). Out of 10,000 decoys generated for each protein and protocol, the number of correct (TM-score ≥ 0.5) decoys (left) and the top TM-score (right) are indicated. Note that these are single-domain according to SCOP, but some are divided into two domains by CATH. Where the domains are continuous, the final residue of the first domain is indicated (CATH). For non-continuous domains, the final residue of each domain segment is indicated, with domain one and two separated by a semicolon.

		Whole					foldon-1					foldon-2						
PDB	**CATH**	**L**	**SAINT2**		**S-FoldOn**		**L**	**SAINT2**		**Foldon**		**L**	**SAINT2**		**S-FoldOn**		**Foldon**	
1SFE	81	165	239	0.64	3128	0.75	81	301	0.65	1190	0.69	84	1274	0.82	1402	0.81	619	0.75
1SQW	93	176	0	0.49	9	0.54	93	9	0.54	26	0.56	83	27	0.59	33	0.67	505	0.72
1ILW		179	4	0.52	4508	0.68	94	0	0.49	344	0.58	85	187	0.61	761	0.63	1	0.51
1P6F	95	188	0	0.43	3	0.53	97	0	0.47	1	0.52	91	0	0.45	0	0.44	0	0.41
1VL1		218	0	0.50	1279	0.62	103	874	0.65	4494	0.71	115	0	0.44	0	0.45	0	0.41
1W66		218	0	0.44	25	0.57	96	2	0.52	3	0.51	122	2	0.51	6	0.52	0	0.47
1SEF	24,250;143,	250	0	0.37			138	0	0.45	0	0.48	112	0	0.38			0	0.41
1VIN	194,432;307,	252	138	0.64	242	0.64	132	1754	0.70	2150	0.69	120	864	0.63	1063	0.63	331	0.60
1RL0	181	255	0	0.36			192	0	0.41	0	0.41	63	0	0.44			0	0.45
2YVT		257	0	0.39			124	0	0.47	0	0.47	133	0	0.35			0	0.36
1XWY		260	49	0.59	6213	0.70	148	303	0.63	640	0.66	112	53	0.59	202	0.62	0	0.45
2HVM		273	0	0.45	1	0.50	155	0	0.47	0	0.47	118	0	0.47	1	0.52	0	0.43
1WL7		312	0	0.36			192	0	0.40	0	0.38	120	0	0.34			0	0.35
1OBR		323	0	0.32	0	0.43	145	0	0.49	2	0.52	178	0	0.34	0	0.36	0	0.35
1MSK	*	327	0	0.41			215	0	0.43	0	0.46	112	0	0.42			0	0.41
1OKQ	12,374;187	374	0	0.34			188	0	0.37	0	0.38	186	0	0.37			NA	NA
1V5C		386	0	0.32			183	0	0.37	0	0.47	203	0	0.36		0	0.38	
1VFF		423	0	0.43	2	0.50	208	8	0.54	26	0.56	215	0	0.38	0	0.41	0	0.36
1SMD	403	496	0	0.35			242	0	0.42	0	0.44	254	0	0.35			0	0.30
1B4V	**	434	0	0.28			283	0	0.31	0	0.34	215	0	0.29			0	0.29

* Full CATH domain boundaries for 1MSK: 35,164,327;95,192

** Full CATH domain boundaries for 1B4V: 1-147,210-314,377-393,435-497;148-205,315-376,398-434

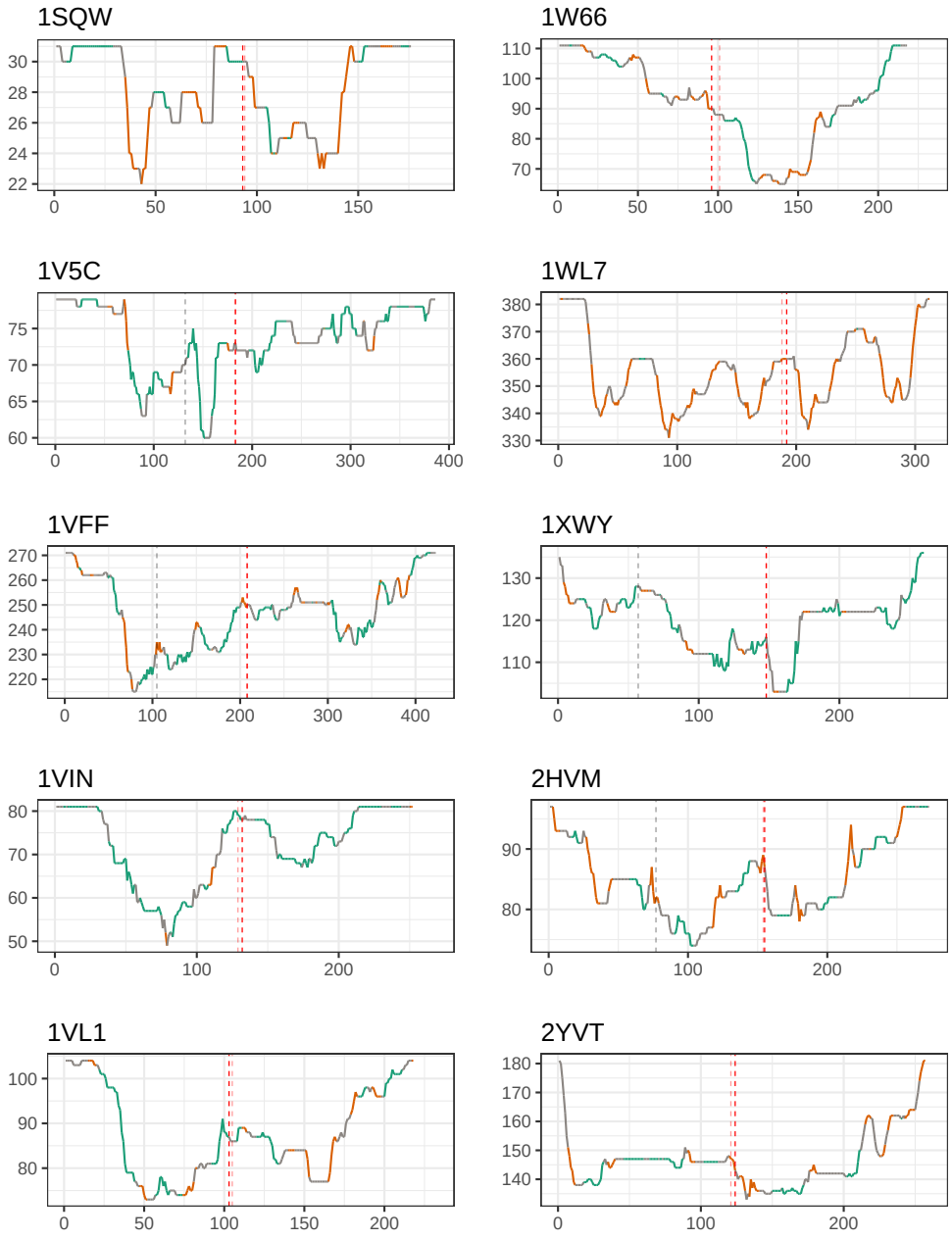

Figure C.1: conFoldUP profiles for all targets in the Long Single-domain set for which foldons were suggested by FoldUP, showing the number of intra-foldon contacts on either side of each residue. The dashed red line indicates the foldon boundary assigned by structure-based FoldUP. The actual FoldUP minimum, prior to rounding to the nearest secondary structure element, is shown in light red. Where FoldUP broke the protein twice, the second break is marked with a grey dashed line. The DSSP secondary structure assignments for each residue are designated as either helical ('helix', turquoise), beta-sheet ('beta', orange) or no secondary structure ('coil', grey).

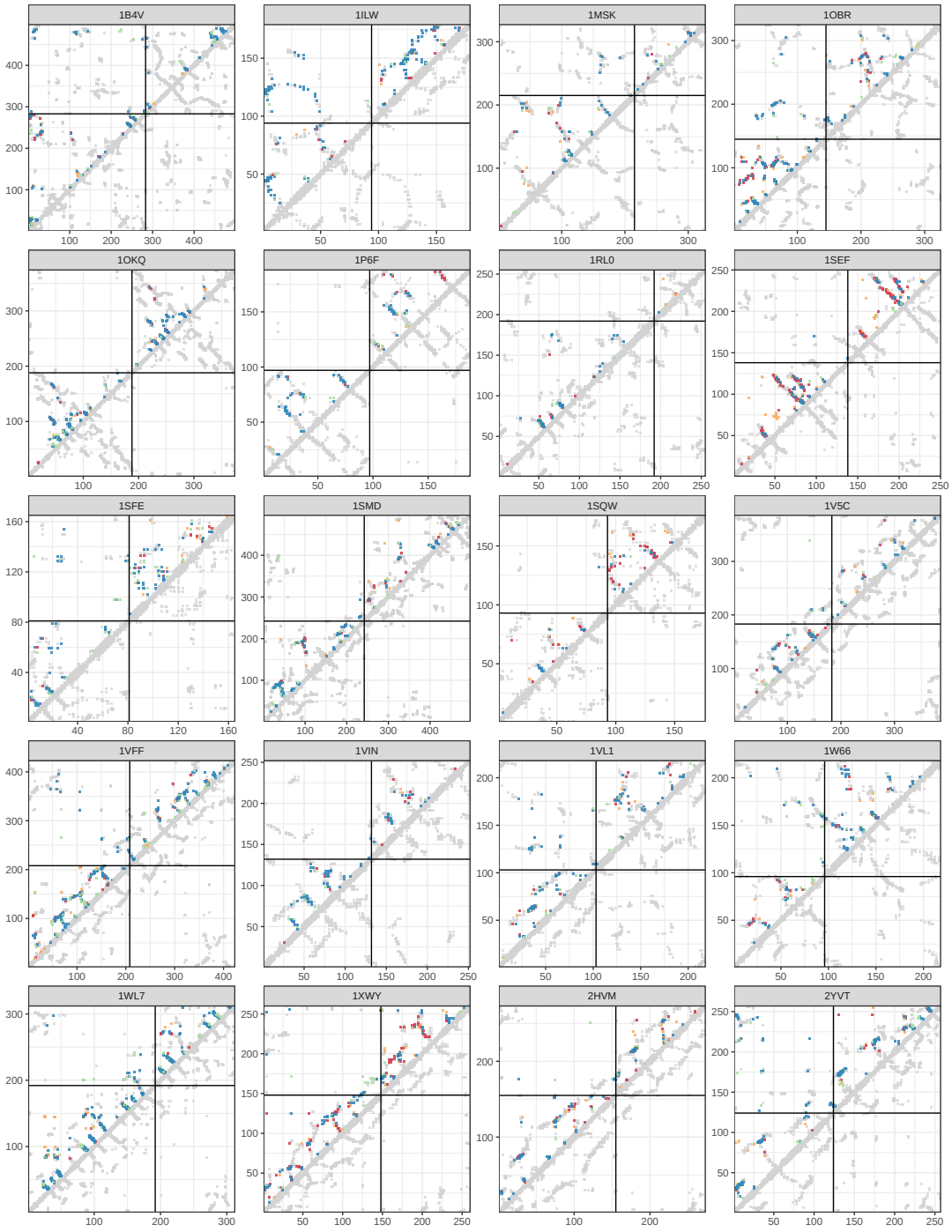

Figure C.2: Contact maps for all the targets in the Long Single-domain set. The true residue-residue contacts are shown in light grey. True and false positive predicted contacts generated using the full-length protein sequences are shown in blue and green, respectively. Additional intra-foldon predicted contacts from prediction using the foldon sequences individually are shown in red for true positives and orange for false positives. The foldon boundaries assigned by conFoldUP are indicated with black lines.

Table C.2: Results for N-terminal 150 residue segment individually compared (top, 150 residue) compared to the equivalent region when using SAINT2 on the entire structure (SAINT2), as well as for the the full-length structure using extra moves (bottom, SAINT2 Long) compared to standard SAINT2 with 10,000 moves (SAINT2). The number of correct models (TM-score $\geq$ 0.5) out of 10,000 produced (No. correct) and the TM-score of the best model (Top TM-score) are shown. The best result for each target is highlighted in bold.

PDB ID	Length	No. Correct		Top TM-score	
		SAINT2	**150 Residue**	**SAINT2**	**150 Residue**
1SFE	152	**298**	253	0.63	**0.64**
1SQW	148	0	**2**	0.48	**0.56**
1ILW	153	2	**7**	0.51	**0.56**
1P6F	156	0	0	**0.45**	0.41
1VL1	153	108	**551**	0.59	**0.64**
1W66	153	1	1	0.51	**0.52**
1SEF	153	0	0	0.42	**0.43**
1VIN	152	646	**776**	**0.65**	0.63
1RL0	156	0	0	**0.42**	0.41
2YVT	156	0	0	0.43	**0.46**
1XWY	148	306	**685**	0.63	**0.72**
2HVM	145	0	0	0.46	**0.49**
1WL7	152	0	0	0.40	**0.43**
1OBR	145	0	**1**	0.49	**0.51**
1MSK	152	0	0	0.41	**0.45**
1OKQ	NA	NA	NA	NA	NA
1V5C	152	0	0	0.38	**0.48**
1VFF	153	22	**50**	0.57	0.57
1SMD	159	0	0	0.39	**0.43**
1B4V	150	0	0	0.42	**0.45**
		SAINT2	**SAINT2 Long**	**SAINT2**	**SAINT2 Long**
1SQW	176	0	0	**0.49**	0.46
1VL1	218	0	**8**	0.50	**0.54**
1W66	218	0	0	0.54	**0.47**
2HVM	273	0	0	0.45	**0.46**
1VFF	423	0	0	0.42	**0.47**

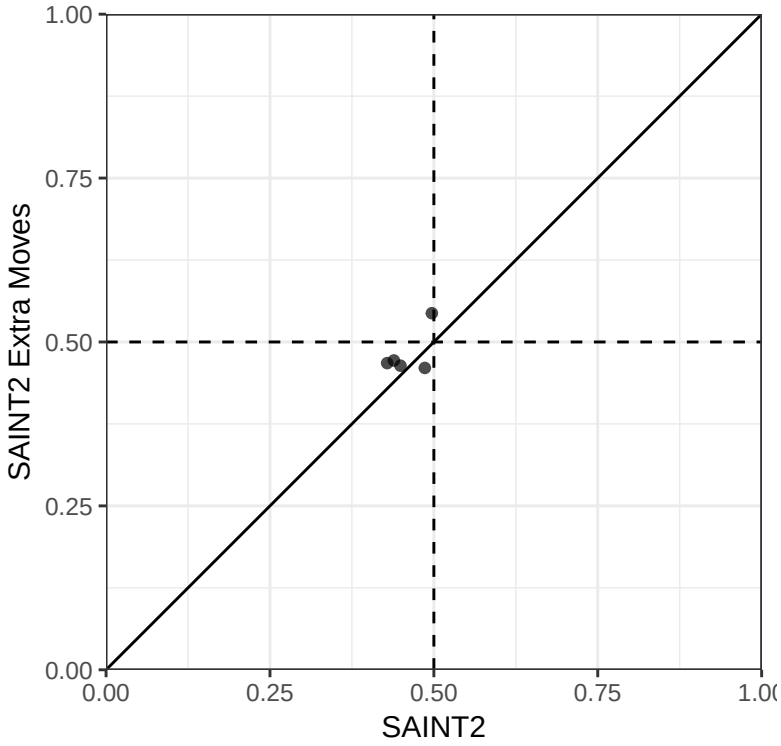

Figure C.3: Comparison of SAINT2 and SAINT2 with extra moves for five Long Single-domain targets. The TM-score of the top-scoring model produced for each target using SAINT2 with the standard 11,000 moves (SAINT2) against SAINT2 with 11,000 moves per 150 residues (SAINT2 Extra Moves). Points above the diagonal line indicate a better performance by SAINT2 with extra moves.

Table C.3: Performance of SAINT2 for the long and short modelling targets in our Training and Validation sets. The total number of targets, the number of targets that were successfully modelled using SAINT2 ("SAINT2"). Targets with at least 150 residues ("Long") or fewer than 150 resides ("Short") are shown separately. Modelling is considered successful for a given target if at least one model is correct (TM-score ≥ 0.5).

Set		Total	SAINT2	
Training	Long	91	37	40.7%
	Short	153	114	74.5%
Validation	Long	87	34	39.1%
	Short	157	108	68.8%

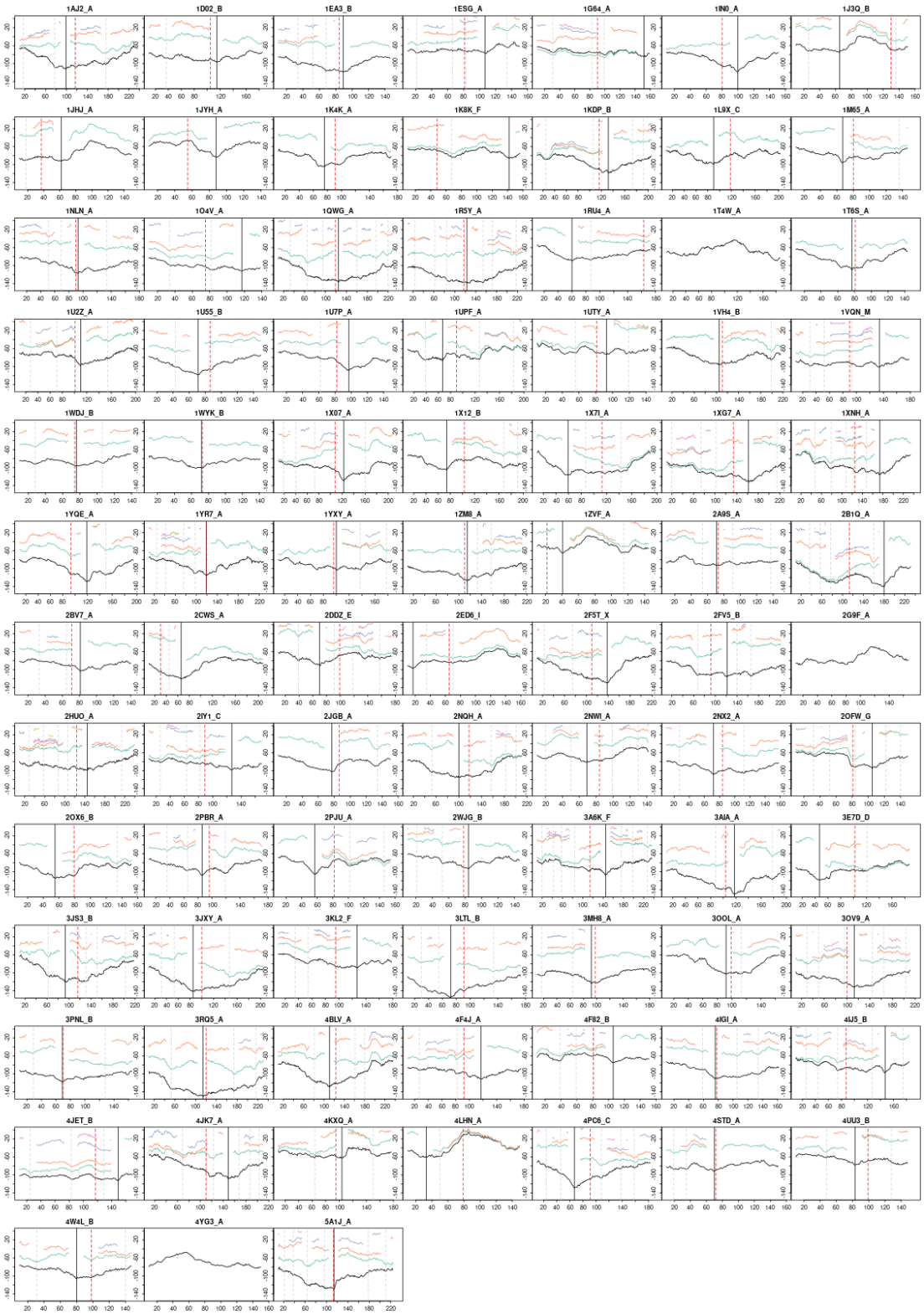

Figure C.4: FoldUP profiles and ssFoldUP boundaries for the Long Validation Set targets. The residue number is on the X-axis and the FoldUP score is on the Y-axis. Dashed grey lines indicate where the minimum average score falls at least 15 residues from either end; this is rounded to the nearest C-terminal end of a secondary structure element and defined as the foldon boundary. The resulting foldons are scanned for further potential foldon boundaries, shown in different colours. The foldon boundary assigned using sequence-based ssFoldUP is shown with a dashed ref line.

D

Acknowledgements Appendices

List of acknowledged Oxford Protein Informatics Group (OPIG) members

- Charlotte Deane (Professor, 2000–)
- Garrett Morris (Associate Professor, 2015–)
- Eoin Malins (Visitor, 2011–)
- Dan Nissley (Postdoc, 2019–)
- Claire Marks (Research Software Engineer, 2018–; Postdoc, 2016-2018; DPhil, 2012-2016)
- Clare West (DPhil, 2015–)
- Mark Chonofsky (DPhil, 2017–)
- Dominik Schwarz (DPhil, 2018–)
- Aleksandr Kovaltsuk (DPhil, 2016–)
- Matthew Raybould (DPhil, 2017–)
- Wing Ki (Catherine) Wong (DPhil, 2017–)
- Constantin Schneider (DPhil, 2018–)
- Eve Richardson (DPhil, 2018–)
- Florian Klimm (Postdoc, 2018-2019; DPhil, 2014-2018)
- Lyuba Bozhilova (DPhil, 2015–)
- Javier Pardo Diaz (DPhil, 2017–)
- James Wilsenach (DPhil, 2017–)
- Joe Bluck (DPhil, 2015-2019)
- Fergus Boyles (DPhil, 2015–)
- Susan Leung (DPhil, 2015–)
- Elliot Nelson (DPhil, 2015-2019)
- Lucian Chan (DPhil, 2017–)
- Fergus Imrie (DPhil, 2017–)
- Anne Nierobisch (DPhil, 2017–)
- Carlos Outeiral (DPhil, 2018–)
- Jack Scantlebury (DPhil, 2018–)
- Marc Moesser (DPhil, 2019–)
- Conor Wild (DPhil, 2018–)
- Angela Hellyer (Part II, 2019–)
- Mark Chin (Part II, 2019–)
- An Goto (DPhil, 2019–)
- Tom Hadfield (DPhil, 2019–)
- Sarah Robinson (DPhil, 2019–)
- Anna Carbery (Part II, 2018-2019)
- Hannah Patel (DPhil, 2014-2018)
- Laura Depner (MRes, 2015-2018)
- Mihai Cucuringu (Postdoc, 2017-2018)

Figure D.1: Members of OPIG. Group photos from (clockwise from top left) OPIGtreat 2017, OPIGtreat 2018, OPIGtreat 2019, OPIGmas 2018.

- Konrad Krawczyk (Postdoc, 2013-2018)
- Jinwoo Leem (DPhil, 2012-2016; Postdoc, 2017-2018)
- Saulo de Oliveira (Postdoc, 2015-2018)
- Jaroslaw Nowak (DPhil, 2013-2017)
- Cristian Regep (DPhil, 2013-2017)
- Eleanor Law (DPhil, 2013-2017)
- Luis Ospina (DPhil, 2013-2017)
- Nicholas Pearce (DPhil, 2012-2016; Postdoc, 2016-2017)
- James Dunbar (Postdoc, 2014-2017)
- Samuel Demharter (DPhil, 2012-2016)
- Malte Luecken (DPhil, 2012-2016)
- Alistair Martin (DPhil, 2012-2016)
- Anatol Wegner (Postdoc, 2015-2016)
- Bernhard Knapp (Postdoc, 2013-2016)

Bibliography

Abriata, L. A., G. E. Tamò, and M. Dal Peraro (2019). "A further leap of improvement in tertiary structure prediction in CASP13 prompts new routes for future assessments". In: *Proteins: Structure, Function, and Bioinformatics* 87.12, pp. 1100–1112 (cit. on pp. 24, 145, 151).

Adhikari, B., D. Bhattacharya, R. Cao, and J. Cheng (2015). "CONFOLD: Residue-residue contact-guided ab initio protein folding". In: *Proteins: Structure, Function and Bioinformatics* 83.8, pp. 1436–1449 (cit. on pp. 24, 30, 138).

Alexandrov, N. (1993). "Structural argument for N-terminal initiation of protein folding". In: *Protein Science* 2.11, pp. 1989–1991 (cit. on p. 84).

Almén, M., K. J. Nordström, R. Fredriksson, and H. B. Schiöth (2009). "Mapping the human membrane proteome: a majority of the human membrane proteins can be classified according to function and evolutionary origin". In: *BMC Biology* 7.1, p. 50 (cit. on p. 153).

AlQuraishi, M. (2019a). "AlphaFold at CASP13". In: *Bioinformatics* 35.22. Ed. by A. Valencia, pp. 4862–4865 (cit. on p. 30).

AlQuraishi, M. (2019b). "End-to-End differentiable learning of protein structure". In: *Cell Systems* 8.4, 292–301.e3 (cit. on p. 30).

Anfinsen, C. B. (1973). "Principles that govern the folding of protein chains". In: *Science* 181.4096, pp. 223–230 (cit. on p. 16).

Anishchenko, I., S. Ovchinnikov, H. Kamisetty, and D. Baker (2017). "Origins of coevolution between residues distant in protein 3D structures". In: *Proceedings of the National Academy of Sciences of the United States of America* 114.34, pp. 9122–9127 (cit. on p. 23).

Bai, Y., T. Sosnick, L. Mayne, and S. Englander (1995). "Protein folding intermediates: native-state hydrogen exchange". In: *Science* 269.5221, pp. 192–197 (cit. on pp. 13, 21, 104).

Bakheet, T. M. and A. J. Doig (2009). "Properties and identification of human protein drug targets". In: *Bioinformatics* 25.4, pp. 451–457 (cit. on p. 153).

Baldwin, R. L. (2017). "Clash between energy landscape theory and foldon-dependent protein folding". In: *Proceedings of the National Academy of Sciences of the United States of America* 114.32, pp. 8442–8443 (cit. on p. 13).

Basu, S., F. Söderquist, and B. Wallner (2017). "Proteus: a random forest classifier to predict disorder-to-order transitioning binding regions in intrinsically disordered proteins". In: *Journal of Computer-Aided Molecular Design* 31.5, pp. 453–466 (cit. on p. 44).

Benkert, P., S. C. E. Tosatto, and D. Schomburg (2008). "QMEAN: A comprehensive scoring function for model quality assessment". In: *Proteins: Structure, Function, and Bioinformatics* 71.1, pp. 261–277 (cit. on p. 31).

Berman, H. M. (2000). "The Protein Data Bank". In: *Nucleic Acids Research* 28.1, pp. 235–242 (cit. on pp. 9, 46, 151).

Betancourt, M. R. and J. Skolnick (2001). "Universal similarity measure for comparing protein structures". In: *Biopolymers* 59.5, pp. 305–309 (cit. on p. 26).

Betancourt, M. R. and J. Skolnick (2004). "Local propensities and statistical potentials of backbone dihedral angles in proteins". In: *Journal of Molecular Biology* 342.2, pp. 635–649 (cit. on p. 21).

Bhattacharya, D., R. Cao, and J. Cheng (2016). "UniCon3D: de novo protein structure prediction using united-residue conformational search via stepwise, probabilistic sampling". In: *Bioinformatics* 32.18, pp. 2791–2799 (cit. on p. 136).

Bowie, J. U. and D. Eisenberg (1994). "An evolutionary approach to folding small alpha-helical proteins that uses sequence information and an empirical guiding fitness function." In: *Proceedings of the National Academy of Sciences of the United States of America* 91.10, pp. 4436–4440 (cit. on p. 28).

Boyles, F., C. M. Deane, and G. M. Morris (2019). "Learning from the ligand: using ligand-based features to improve binding affinity prediction". In: *Bioinformatics*. Ed. by A. Elofsson (cit. on p. 44).

Braselmann, E., J. L. Chaney, and P. L. Clark (2013). "Folding the proteome". In: *Trends in Biochemical Sciences* 38.7, pp. 337–344 (cit. on p. 13).

Breiman, L. (2001). "Random Forests". In: *Machine Learning* 45.1, pp. 5–32 (cit. on p. 44).

Brocchieri, L. (2005). "Protein length in eukaryotic and prokaryotic proteomes". In: *Nucleic Acids Research* 33.10, pp. 3390–3400 (cit. on p. 30).

Brünger, A. T. (1992). "Free R value: a novel statistical quantity for assessing the accuracy of crystal structures". In: *Nature* 355.6359, pp. 472–475 (cit. on p. 10).

Brünger, A. T. (1997). "Free R value: Cross-validation in crystallography". In: pp. 366–396 (cit. on p. 10).

Bryngelson, J. D., J. N. Onuchic, N. D. Socci, and P. G. Wolynes (1995). "Funnels, pathways, and the energy landscape of protein folding: A synthesis". In: *Proteins: Structure, Function, and Genetics* 21.3, pp. 167–195 (cit. on p. 12).

Buchan, D. W. A. and D. T. Jones (2017). "EigenTHREADER: analogous protein fold recognition by efficient contact map threading". In: *Bioinformatics* 33.17. Ed. by A. Valencia, pp. 2684–2690 (cit. on pp. 43, 50).

Buchan, D. W. A., F. Minneci, T. C. O. Nugent, K. Bryson, and D. T. Jones (2013). "Scalable web services for the PSIPRED Protein Analysis Workbench." In: *Nucleic Acids Research* 41.Web Server issue, W349–57 (cit. on p. 111).

Cabrita, L. D. and S. P. Bottomley (2004). "Protein expression and refolding – A practical guide to getting the most out of inclusion bodies". In: *Biotechnology Annual Review*. Vol. 10, pp. 31–50 (cit. on pp. 13, 16).

Cabrita, L. D., A. M. E. Cassaignau, et al. (2016). "A structural ensemble of a ribosome–nascent chain complex during cotranslational protein folding". In: *Nature Structural & Molecular Biology* 23.4, pp. 278–285 (cit. on p. 18).

Camacho, C. et al. (2009). "BLAST+: architecture and applications". In: *BMC Bioinformatics* 10.1, p. 421 (cit. on pp. 47, 109).

Carson, M., D. H. Johnson, H. McDonald, C. Brouillette, and L. J. DeLucas (2007). "His-tag impact on structure". In: *Acta Crystallographica Section D: Biological Crystallography* 63.3, pp. 295–301 (cit. on p. 77).

Cassaignau, A. M. E. et al. (2016). "A strategy for co-translational folding studies of ribosome-bound nascent chain complexes using NMR spectroscopy". In: *Nature Protocols* 11.8, pp. 1492–1507 (cit. on pp. 18, 105).

Censoni, L. and L. Martínez (2018). "Prediction of kinetics of protein folding with non-redundant contact information". In: *Bioinformatics* 34.23. Ed. by A. Valencia, pp. 4034–4038 (cit. on p. 23).

Chandonia, J.-M., N. K. Fox, and S. E. Brenner (2017). "SCOPe: Manual Curation and Artifact Removal in the Structural Classification of Proteins – extended Database". In: *Journal of Molecular Biology* 429.3, pp. 348–355 (cit. on p. 9).

Cheng, J., M. H. Choe, et al. (2019). "Estimation of model accuracy in CASP13". In: *Proteins: Structure, Function and Bioinformatics* 87.12, pp. 1361–1377 (cit. on p. 32).

Cheng, J., M. J. Sweredoski, and P. Baldi (2006). "DOMpro: Protein domain prediction using profiles, secondary structure, relative solvent accessibility, and recursive neural networks". In: *Data Mining and Knowledge Discovery* 13.1, pp. 1–10 (cit. on p. 107).

Cheung, N. J. and W. Yu (2018). "De novo protein structure prediction using ultra-fast molecular dynamics simulation". In: *PLOS ONE* 13.11. Ed. by Y. Zhang, e0205819 (cit. on p. 150).

Chiti, F. and C. M. Dobson (2017). "Protein Misfolding, Amyloid Formation, and Human Disease: A Summary of Progress Over the Last Decade". In: *Annual Review of Biochemistry* 86.1, pp. 27–68 (cit. on p. 12).

Choi, Y. and C. M. Deane (2009). "FREAD revisited: Accurate loop structure prediction using a database search algorithm". In: *Proteins: Structure, Function, and Bioinformatics* 78.6, pp. 1431–1440 (cit. on pp. 27, 72).

Chonofsky, M., S. H. P. de Oliveira, K. Krawczyk, and C. M. Deane (2019). "The evolution of contact prediction: Evidence that contact selection in statistical contact prediction is changing". In: *Bioinformatics*. Ed. by Y. Ponty (cit. on p. 23).

Chothia, C. and J. Gough (2009). "Genomic and structural aspects of protein evolution". In: *Biochemical Journal* 419.1, pp. 15–28 (cit. on pp. 8, 106, 115).

Ciryam, P., R. I. Morimoto, M. Vendruscolo, C. M. Dobson, and E. P. O'Brien (2013). "In vivo translation rates can substantially delay the cotranslational folding of the Escherichia coli cytosolic proteome". In: *Proceedings of the National Academy of Sciences of the United States of America* 110.2, E132–E140 (cit. on p. 14).

Clark, P. (2004). "Protein folding in the cell: reshaping the folding funnel". In: *Trends in Biochemical Sciences* 29.10, pp. 527–534 (cit. on pp. 2, 14).

Cock, P. J. A. et al. (2009). "Biopython: freely available Python tools for Computational Molecular Biology and Bioinformatics". In: *Bioinformatics* 25.11, pp. 1422–1423 (cit. on pp. 50, 77, 159).

Coucke, A. et al. (2016). "Direct coevolutionary couplings reflect biophysical residue interactions in proteins". In: *The Journal of Chemical Physics* 145.17, p. 174102 (cit. on p. 23).

Croce, G. et al. (2019). "A multi-scale coevolutionary approach to predict interactions between protein domains". In: *PLOS Computational Biology* 15.10. Ed. by S. Maslov, e1006891 (cit. on p. 23).

Croll, T. I., M. D. Sammito, A. Kryshtafovych, and R. J. Read (2019). "Evaluation of template-based modeling in CASP13". In: *Proteins: Structure, Function, and Bioinformatics* 87.12, pp. 1113–1127 (cit. on pp. 28, 72).

Csaba, G., F. Birzele, and R. Zimmer (2009). "Systematic comparison of SCOP and CATH: a new gold standard for protein structure analysis." In: *BMC Structural Biology* 9.23, pp. 1–11 (cit. on pp. 9, 106).

De Oliveira, S. H. P. and C. M. Deane (2017). "Co-evolution techniques are reshaping the way we do structural bioinformatics". In: *F1000Research* 6, p. 1224 (cit. on p. 23).

De Oliveira, S. H. P. and C. M. Deane (2018). "Combining co-evolution and secondary structure prediction to improve fragment library generation". In: *Bioinformatics* 34.13. Ed. by A. Valencia (cit. on pp. 34, 37, 47, 78).

De Oliveira, S. H. P., E. C. Law, J. Shi, and C. M. Deane (2018). "Sequential search leads to faster, more efficient fragment-based de novo protein structure prediction". In: *Bioinformatics* 34.7. Ed. by A. Valencia, pp. 1132–1140 (cit. on pp. 2, 19, 33, 42, 45, 47, 51, 67, 104, 144, 145, 163).

De Oliveira, S. H. P., J. Shi, and C. M. Deane (2015). "Building a better fragment library for de novo protein structure prediction". In: *PLOS ONE* 10.4. Ed. by Y. Zhang, e0123998 (cit. on pp. 29, 34, 78, 111, 119, 163).

De Oliveira, S. H. P., J. Shi, and C. M. Deane (2017). "Comparing co-evolution methods and their application to template-free protein structure prediction". In: *Bioinformatics*, btw618 (cit. on pp. 23, 32, 34, 35, 43, 47, 48, 51, 67, 105, 109, 113, 120, 150, 153, 163).

Deane, C. M. (2001). "CODA: A combined algorithm for predicting the structurally variable regions of protein models". In: *Protein Science* 10.3, pp. 599–612 (cit. on pp. 27, 72).

Deane, C. M., M. Dong, F. P. Huard, B. K. Lance, and G. R. Wood (2007). "Cotranslational protein folding fact or fiction?" In: *Bioinformatics* 23.13, pp. i142–i148 (cit. on pp. 14, 84).

Dill, K. A. and H. S. Chan (1997). "From Levinthal to pathways to funnels". In: *Nature Structural & Molecular Biology* 4.1, pp. 10–19 (cit. on p. 12).

Dinner, A. R., A. Šali, L. J. Smith, C. M. Dobson, and M. Karplus (2000). "Understanding protein folding via free-energy surfaces from theory and experiment". In: *Trends in Biochemical Sciences* 25.7, pp. 331–339 (cit. on pp. 12, 13).

Dunker, A. K., C. J. Brown, J. D. Lawson, L. M. Iakoucheva, and Z. Obradović (2002). "Intrinsic Disorder and Protein Function". In: 41.21 (cit. on p. 147).

Eaton, W. A. and P. G. Wolynes (2017). "Theory, simulations, and experiments show that proteins fold by multiple pathways". In: *Proceedings of the National Academy of Sciences of the United States of America* 114.46, E9759–E9760 (cit. on p. 13).

Ebina, T., H. Toh, and Y. Kuroda (2009). "Loop-length-dependent SVM prediction of domain linkers for high-throughput structural proteomics". In: *Biopolymers* 92.1, pp. 1–8 (cit. on p. 107).

Edgar, R. C. (2004). "MUSCLE: multiple sequence alignment with high accuracy and high throughput". In: *Nucleic Acids Research* 32.5, pp. 1792–1797 (cit. on p. 77).

Eiberle, M. K. and A. Jungbauer (2010). "Technical refolding of proteins: Do we have freedom to operate?" In: *Biotechnology Journal* 5.6, pp. 547–559 (cit. on pp. 13, 16).

Ellis, J. J., F. P. Huard, C. M. Deane, S. Srivastava, and G. R. Wood (2010). "Directionality in protein fold prediction". In: *BMC Bioinformatics* 11.1, p. 172 (cit. on pp. 2, 19, 33, 120).

Elofsson, A. et al. (2018). "Methods for estimation of model accuracy in CASP12". In: *Proteins: Structure, Function, and Bioinformatics* 86, pp. 361–373 (cit. on p. 31).

Englander, S. W. and L. Mayne (2014). "The nature of protein folding pathways." In: *Proceedings of the National Academy of Sciences of the United States of America* 111.45, pp. 15873–80 (cit. on pp. 13, 104).

Englander, S. W. and L. Mayne (2017). "The case for defined protein folding pathways." In: *Proceedings of the National Academy of Sciences of the United States of America* 114.31, pp. 8253–8258 (cit. on p. 13).

Evans, M. S., I. M. Sander, and P. L. Clark (2008). "Cotranslational Folding Promotes β-Helix Formation and Avoids Aggregation In Vivo". In: *Journal of Molecular Biology* 383.3, pp. 683–692 (cit. on pp. 16, 152).

Faraggi, E., Y. Yang, S. Zhang, and Y. Zhou (2009). "Predicting continuous local structure and the effect of Iis substitution for secondary structure in fragment-free protein structure prediction". In: *Structure* 17.11, pp. 1515–1527 (cit. on pp. 21, 111).

Faraggi, E., T. Zhang, Y. Yang, L. Kurgan, and Y. Zhou (2012). "SPINE X: Improving protein secondary structure prediction by multistep learning coupled with prediction of solvent accessible surface area and backbone torsion angles". In: *Journal of Computational Chemistry* 33.3, pp. 259–267 (cit. on pp. 21, 111).

Fedorov, A. N. and T. O. Baldwin (1997). "Cotranslational Protein Folding". In: *Journal of Biological Chemistry* 272.52, pp. 32715–32718 (cit. on p. 14).

Fedorov, A. N. and T. O. Baldwin (1999). "Process of biosynthetic protein folding determines the rapid formation of native structure". In: *Journal of Molecular Biology* 294.2, pp. 579–586 (cit. on p. 16).

Fetrow, J. S. et al. (2001). "Genomic-scale comparison of sequence- and structure-based methods of function prediction: Does structure provide additional insight?" In: *Protein Science* 10.5, pp. 1005–1014 (cit. on p. 149).

Fiser, A. (2010). "Template-based protein structure modeling". In: *Methods in molecular biology (Clifton, N.J.)* Vol. 673, pp. 73–94 (cit. on p. 27).

Fiser, A., R. K. G. Do, and A. Šali (2000). "Modeling of loops in protein structures". In: *Protein Science* 9.9, pp. 1753–1773 (cit. on pp. 27, 72, 78).

Fox, N. K., S. E. Brenner, and J.-M. Chandonia (2014). "SCOPe: Structural Classification of Proteins—extended, integrating SCOP and ASTRAL data and

classification of new structures". In: *Nucleic Acids Research* 42.D1, pp. D304–D309 (cit. on pp. 8, 45, 46, 106).

Fox, N. K., S. E. Brenner, and J.-M. Chandonia (2015). "The value of protein structure classification information - surveying the scientific literature". In: *Proteins: Structure, Function, and Bioinformatics* 83.11, pp. 2025–2038 (cit. on p. 9).

Frydman, J., H. Erdjument-Bromage, P. Tempst, and F. U. Hartl (1999). "Co-translational domain folding as the structural basis for the rapid de novo folding of firefly luciferase". In: *Nature Structural Biology* 6.7, pp. 697–705 (cit. on pp. 16, 105, 152).

Garton, M., S. S. MacKinnon, A. Malevanets, and S. J. Wodak (2018). "Interplay of self-association and conformational flexibility in regulating protein function". In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 373.1749, p. 20170190 (cit. on p. 147).

Gligorijevic, V. et al. (2019). "Structure-based function prediction using graph convolutional networks". In: *bioRxiv*, p. 786236 (cit. on p. 149).

Göbel, U., C. Sander, R. Schneider, and A. Valencia (1994). "Correlated mutations and residue contacts in proteins". In: *Proteins: Structure, Function, and Genetics* 18.4, pp. 309–317 (cit. on p. 22).

Goldman, D. H. et al. (2015). "Mechanical force releases nascent chain-mediated ribosome arrest in vitro and in vivo". In: *Science* 348.6233, pp. 457–460 (cit. on p. 18).

Greene, L. H. et al. (2007). "The CATH domain structure database: new protocols and classification levels give a more comprehensive resource for exploring evolution". In: *Nucleic Acids Research* 35.Database, pp. D291–D297 (cit. on p. 9).

Greener, J. G., S. M. Kandathil, and D. T. Jones (2019). "Deep learning extends de novo protein modelling coverage of genomes using iteratively predicted structural constraints". In: *Nature Communications* 10.1, p. 3977 (cit. on pp. 30, 146, 147).

Hamlin, J. and I. Zabin (1972). "Beta-Galactosidase: immunological activity of ribosome-bound, growing polypeptide chains." In: *Proceedings of the National Academy of Sciences of the United States of America* 69.2, pp. 412–6 (cit. on p. 152).

Harrison, S. C. and R. Durbin (1985). "Is there a single pathway for the folding of a polypeptide chain?" In: *Proceedings of the National Academy of Sciences* 82.12, pp. 4028–4030 (cit. on p. 13).

Hartl, F. U. (2002). "Molecular chaperones in the cytosol: from nascent chain to folded protein". In: *Science* 295.5561, pp. 1852–1858 (cit. on p. 13).

Hasegawa, H. and L. Holm (2009). "Advances and pitfalls of protein structural alignment". In: *Current Opinion in Structural Biology* 19.3, pp. 341–348 (cit. on p. 25).

Heffernan, R., Y. Yang, K. Paliwal, and Y. Zhou (2017). "Capturing non-local interactions by long short-term memory bidirectional recurrent neural networks for improving prediction of protein secondary structure, backbone angles, contact numbers and solvent accessibility". In: *Bioinformatics* 33.18. Ed. by A. Valencia, pp. 2842–2849 (cit. on pp. 20, 47, 81, 161).

Holcomb, J. et al. (2017). "Protein crystallization: Eluding the bottleneck of X-ray crystallography". In: *AIMS Biophysics* 4.4, pp. 557–575 (cit. on p. 10).

Holland, T. A., S. Veretnik, I. N. Shindyalov, and P. E. Bourne (2006). "Partitioning protein structures into domains: Why is it so difficult?" In: *Journal of Molecular Biology* 361.3, pp. 562–590 (cit. on pp. 8, 9, 106).

Holtkamp, W. et al. (2015). "Cotranslational protein folding on the ribosome monitored in real time". In: *Science* 350.6264, pp. 1104–1107 (cit. on pp. 16, 105).

Hou, Q., P. De Geest, W. F. Vranken, J. Heringa, and K. A. Feenstra (2017). "Seeing the trees through the forest: Sequence-based homo- and heteromeric protein-protein interaction sites prediction using Random Forest". In: *Bioinformatics* 33.10, pp. 1479–1487 (cit. on p. 44).

Hsu, S.-T. D. et al. (2007). "Structure and dynamics of a ribosome-bound nascent chain by NMR spectroscopy." In: *Proceedings of the National Academy of Sciences of the United States of America* 104.42, pp. 16516–16521 (cit. on pp. 18, 152).

Hu, W., Z.-Y. Kan, L. Mayne, and S. W. Englander (2016). "Cytochrome c folds through foldon-dependent native-like intermediates in an ordered pathway." In: *Proceedings of the National Academy of Sciences of the United States of America* 113.14, pp. 3809–14 (cit. on pp. 13, 104).

Hura, G. L. et al. (2019). "Small angle X-ray scattering-assisted protein structure prediction in CASP13 and emergence of solution structure differences". In: *Proteins: Structure, Function, and Bioinformatics* 87.12, pp. 1298–1314 (cit. on pp. 147, 151).

Ikeya, T., P. Güntert, and Y. Ito (2019). "Protein structure determination in living cells". In: *International Journal of Molecular Sciences* 20.10, p. 2442 (cit. on p. 11).

Ikeya, T., T. Hanashima, et al. (2016). "Improved in-cell structure determination of proteins at near-physiological concentration". In: *Scientific Reports* 6.1, p. 38312 (cit. on p. 11).

Inbar, Y., H. Benyamini, R. Nussinov, and H. J. Wolfson (2005). "Combinatorial docking approach for structure prediction of large proteins and multi-molecular assemblies." In: *Physical Biology* 2.4, S156–65 (cit. on pp. 31, 151).

Ingolia, N. T., J. A. Hussmann, and J. S. Weissman (2019). "Ribosome Profiling: Global views of translation". In: *Cold Spring Harbor Perspectives in Biology* 11.5, a032698 (cit. on p. 17).

Jacobs, W. M. and E. I. Shakhnovich (2017). "Evidence of evolutionary selection for cotranslational folding". In: *Proceedings of the National Academy of Sciences of the United States of America* 114.43, pp. 11434–11439 (cit. on pp. 16, 17).

Jacobson, G. N. and P. L. Clark (2016). "Quality over quantity: optimizing co-translational protein folding with non-'optimal' synonymous codons". In: *Current Opinion in Structural Biology* 38, pp. 102–110 (cit. on p. 17).

Jana, B., F. Morcos, and J. N. Onuchic (2014). "From structure to function: the convergence of structure based models and co-evolutionary information". In: *Physical Chemistry Chemical Physics* 16.14, pp. 6496–6507 (cit. on p. 23).

Ji, S. et al. (2019). "DeepCDpred: Inter-residue distance and contact prediction for improved prediction of protein structure". In: *PLOS ONE* 14.1. Ed. by Y. Zhang, e0205214 (cit. on p. 23).

Jones, D. T. (1999). "Protein secondary structure prediction based on position-specific scoring matrices". In: *Journal of Molecular Biology* 292.2, pp. 195–202 (cit. on pp. 47, 111, 161).

Jones, D. T., D. W. A. Buchan, D. Cozzetto, and M. Pontil (2012). "PSICOV: Precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments". In: *Bioinformatics* 28.2, pp. 184–190 (cit. on p. 22).

Jones, D. T., T. Singh, T. Kosciolek, and S. Tetchner (2015). "MetaPSICOV: combining coevolution methods for accurate prediction of contacts and long range hydrogen bonding in proteins". In: *Bioinformatics* 31.7, pp. 999–1006 (cit. on pp. 22, 23, 32, 47, 67, 81, 111, 163).

Jones, J. E. (1924). "On the determination of molecular fields. II. From the equation of state of a gas". In: *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences* 106.738, pp. 463–477 (cit. on pp. 30, 32, 35).

Joosten, R. P. et al. (2011). "A series of PDB related databases for everyday needs". In: *Nucleic Acids Research* 39.Database, pp. D411–D419 (cit. on p. 20).

Kabsch, W. and C. Sander (1983). "Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features". In: *Biopolymers* 22.12, pp. 2577–2637 (cit. on pp. 8, 20, 80, 83, 114, 161).

Kaiser, C. M., D. H. Goldman, J. D. Chodera, I. Tinoco, and C. Bustamante (2011). "The ribosome modulates nascent protein folding". In: *Science* 334.6063, pp. 1723–1727 (cit. on pp. 16, 18).

Kaiser, C. M. and K. Liu (2018). "Folding up and moving on—nascent protein folding on the ribosome". In: *Journal of Molecular Biology* 430.22, pp. 4580–4591 (cit. on p. 18).

Kaján, L., T. A. Hopf, M. Kalaš, D. S. Marks, and B. Rost (2014). "FreeContact: fast and free software for protein contact prediction from residue co-evolution". In: *BMC Bioinformatics* 15.1, p. 85 (cit. on p. 22).

Kamisetty, H., S. Ovchinnikov, and D. Baker (2013). "Assessing the utility of coevolution-based residue-residue contact predictions in a sequence- and structure-rich era." In: *Proceedings of the National Academy of Sciences of the United States of America* 110.39, pp. 15674–9 (cit. on pp. 22, 107).

Kandathil, S. M., J. G. Greener, and D. T. Jones (2019a). "Prediction of interresidue contacts with DeepMetaPSICOV in CASP13". In: *Proteins: Structure, Function, and Bioinformatics* 87.12, pp. 1092–1099 (cit. on p. 23).

Kandathil, S. M., J. G. Greener, and D. T. Jones (2019b). "Recent developments in deep learning applied to protein structure prediction". In: *Proteins: Structure, Function, and Bioinformatics* 87.12, pp. 1179–1189 (cit. on pp. 25, 30, 146).

KC, D. B. (2016). "Recent advances in sequence-based protein structure prediction". In: *Briefings in Bioinformatics* 31, pp. 1–12 (cit. on p. 104).

Kelkar, D. A., A. Khushoo, Z. Yang, and W. R. Skach (2012). "Kinetic analysis of ribosome-bound fluorescent proteins reveals an early, stable, cotranslational

folding intermediate". In: *Journal of Biological Chemistry* 287.4, pp. 2568–2578 (cit. on p. 105).

Kendrew, J. C. et al. (1958). "A three-dimensional model of the myoglobin molecule obtained by X-Ray analysis". In: *Nature* 181.4610, pp. 662–666 (cit. on p. 9).

Kihara, D. (2005). "The effect of long-range interactions on the secondary structure formation of proteins". In: *Protein Science* 14.8, pp. 1955–1963 (cit. on p. 20).

Kim, D. E., F. Dimaio, R. Yu-Ruei Wang, Y. Song, and D. Baker (2014). "One contact for every twelve residues allows robust and accurate topology-level protein structure modeling". In: *Proteins: Structure, Function and Bioinformatics* 82.SUPPL.2, pp. 208–218 (cit. on p. 138).

Kinch, L. N., W. Li, B. Monastyrskyy, A. Kryshtafovych, and N. V. Grishin (2016). *Evaluation of free modeling targets in CASP11 and ROLL* (cit. on p. 104).

Kolb, V. A. (2001). *Cotranslational protein folding* (cit. on p. 14).

Kolodny, R., L. Pereyaslavets, A. O. Samson, and M. Levitt (2013). "On the Universe of Protein Folds". In: *Annual Review of Biophysics* 42.1, pp. 559–582 (cit. on pp. 105, 141).

Komar, A. A. (2009). "A pause for thought along the co-translational folding pathway". In: *Trends in Biochemical Sciences* 34.1, pp. 16–24 (cit. on pp. 2, 14, 17).

Komar, A. A., A. Kommer, I. A. Krasheninnikov, and A. S. Spirin (1997). "Co-translational folding of globin". In: *Journal of Biological Chemistry* 272.16, pp. 10646–10651 (cit. on pp. 16, 105, 152).

Konczal, J., J. Bower, and C. H. Gray (2019). "Re-introducing non-optimal synonymous codons into codon-optimized constructs enhances soluble recovery of recombinant proteins from Escherichia coli". In: *PLOS ONE* 14.4 (cit. on p. 18).

Kosciolek, T., D. W. A. Buchan, and D. T. Jones (2017). "Predictions of backbone dynamics in intrinsically disordered proteins using de novo fragment-based protein structure predictions". In: *Scientific Reports* 7.1, p. 6999 (cit. on p. 147).

Kramer, G., D. Boehringer, N. Ban, and B. Bukau (2009). "The ribosome as a platform for co-translational processing, folding and targeting of newly synthesized proteins". In: *Nature Structural & Molecular Biology* 16.6, pp. 589–597 (cit. on pp. 18, 105).

Kryshtafovych, A., A. Barbato, et al. (2014). "Assessment of the assessment: Evaluation of the model quality estimates in CASP10". In: *Proteins: Structure, Function, and Bioinformatics* 82, pp. 112–126 (cit. on p. 140).

Kryshtafovych, A., B. Monastyrskyy, K. Fidelis, J. Moult, et al. (2018). "Evaluation of the template-based modeling in CASP12". In: *Proteins: Structure, Function, and Bioinformatics* 86, pp. 321–334 (cit. on p. 28).

Kryshtafovych, A., B. Monastyrskyy, K. Fidelis, T. Schwede, and A. Tramontano (2018). "Assessment of model accuracy estimations in CASP12". In: *Proteins: Structure, Function, and Bioinformatics* 86, pp. 345–360 (cit. on pp. 31, 42, 60).

Kryshtafovych, A., T. Schwede, M. Topf, K. Fidelis, and J. Moult (2019). "Critical assessment of methods of protein structure prediction (CASP)—Round XIII". In: *Proteins: Structure, Function, and Bioinformatics* 87.12, pp. 1011–1020 (cit. on pp. 19, 23, 25, 30, 144, 146, 148).

Kufareva, I. and R. Abagyan (2011). "Methods of Protein Structure Comparison". In: *Methods in Molecular Biology*. Vol. 857, pp. 231–257 (cit. on pp. 25, 26).

Lam, S. D., S. Das, I. Sillitoe, and C. Orengo (2017). "An overview of comparative modelling and resources dedicated to large-scale modelling of genome sequences". In: *Acta Crystallographica Section D Structural Biology* 73.8, pp. 628–640 (cit. on p. 27).

Laurenzi, A., L.-H. Hung, and R. Samudrala (2013). "Structure Prediction of Partial-Length Protein Sequences". In: *International Journal of Molecular Sciences* 14.7, pp. 14892–14907 (cit. on p. 150).

Law, E. C. (2017). "Computational studies of structural motifs and cotranslational folding mechanisms in membrane and soluble proteins". PhD thesis. University of Oxford (cit. on pp. 34, 73, 144).

Leaver-Fay, A. et al. (2011). "Rosetta3". In: *Methods in Enzymology*, pp. 545–574 (cit. on pp. 42, 49).

Lee, D., O. Redfern, and C. Orengo (2007). "Predicting protein function from sequence and structure". In: *Nature Reviews Molecular Cell Biology* 8.12, pp. 995–1005 (cit. on p. 2).

Levinthal, C. (1968). "Are there pathways for protein folding?" In: *Journal de Chimie Physique* 65, pp. 44–45 (cit. on p. 12).

Li, Y. (2013). "Conformational sampling in template-free protein loop structure modeling: an overview". In: *Computational and Structural Biotechnology Journal* 5.6, e201302003 (cit. on p. 72).

Liaw, A. and M. Wiener (2002). "Classification and Regression by randomForest". In: *R news* (cit. on p. 49).

Liu, K., J. E. Rehfus, E. Mattson, and C. M. Kaiser (2017). "The ribosome destabilizes native and non-native structures in a nascent multidomain protein". In: *Protein Science* 26.7, pp. 1439–1451 (cit. on p. 18).

Lovell, S. C. et al. (2003). "Structure validation by Cα geometry: phi, psi and Cβ deviation". In: *Proteins: Structure, Function, and Bioinformatics* 50.3, pp. 437–450 (cit. on p. 6).

Luttrell, J., T. Liu, C. Zhang, and Z. Wang (2019). "Predicting protein residue-residue contacts using random forests and deep networks". In: *BMC Bioinformatics* 20.S2, p. 100 (cit. on p. 44).

Lyons, J. et al. (2014). "Predicting backbone Cα angles and dihedrals from protein sequences by stacked sparse auto-encoder deep neural network". In: *Journal of Computational Chemistry* 35.28, pp. 2040–2046 (cit. on p. 21).

Maghrabi, A. H. A. and L. J. McGuffin (2017). "ModFOLD6: an accurate web server for the global and local quality estimation of 3D protein models". In: *Nucleic Acids Research* 45.W1, W416–W421 (cit. on pp. 24, 31, 43).

Mahlab, S. and M. Linial (2014). "Speed controls in translating secretory proteins in eukaryotes - an evolutionary perspective". In: *PLOS Computational Biology* 10.1. Ed. by Y. Pilpel, e1003294 (cit. on p. 17).

Mallamace, F. et al. (2016). "Energy landscape in protein folding and unfolding". In: *Proceedings of the National Academy of Sciences of the United States of America* 113.12, pp. 3159–3163 (cit. on p. 13).

Manavalan, B., J. Lee, and J. Lee (2014). "Random Forest-based protein Model Quality Assessment (RFMQA) using structural features and potential energy terms". In: *PLOS ONE* 9.9. Ed. by G. P. S. Raghava, e106542 (cit. on pp. 44, 69).

Mariani, V., M. Biasini, A. Barbato, and T. Schwede (2013). "lDDT: a local superposition-free score for comparing protein structures and models using distance difference tests". In: *Bioinformatics* 29.21, pp. 2722–2728 (cit. on p. 27).

Marino, J., G. von Heijne, and R. Beckmann (2016). "Small protein domains fold inside the ribosome exit tunnel". In: *FEBS Letters* 590.5, pp. 655–660 (cit. on pp. 18, 105).

Marks, C. et al. (2017). "Sphinx: merging knowledge-based and ab initio approaches to improve protein loop prediction". In: *Bioinformatics* 33.9, pp. 1346–1353 (cit. on pp. 27, 72).

Marks, D. S., L. J. Colwell, et al. (2011). "Protein 3D Structure Computed from Evolutionary Sequence Variation". In: *PLOS ONE* 6.12. Ed. by A. Sali, e28766 (cit. on pp. 22, 24).

Marks, D. S., T. A. Hopf, and C. Sander (2012). "Protein structure prediction from sequence variation". In: *Nature Biotechnology* 30.11, pp. 1072–1080 (cit. on p. 23).

Michel, M., D. Menéndez Hurtado, K. Uziela, and A. Elofsson (2017). "Large-scale structure prediction by improved contact predictions and model quality assessment". In: *Bioinformatics* 33.14, pp. i23–i29 (cit. on pp. 24, 43, 44, 50, 59, 60, 68, 69, 144).

Michel, M., M. J. Skwark, D. Menéndez Hurtado, M. Ekeberg, and A. Elofsson (2017). "Predicting accurate contacts in thousands of Pfam domain families using PconsC3". In: *Bioinformatics* 33.18. Ed. by A. Valencia, pp. 2859–2866 (cit. on pp. 24, 42, 43, 45).

Mohammad, F., R. Green, and A. R. Buskirk (2019). "A systematically-revised ribosome profiling method for bacteria reveals pauses at single-codon resolution". In: *eLife* 8 (cit. on p. 17).

Monastyrskyy, B., D. D'Andrea, K. Fidelis, A. Tramontano, and A. Kryshtafovych (2016). "New encouraging developments in contact prediction: Assessment of the CASP11 results". In: *Proteins: Structure, Function, and Bioinformatics* 84, pp. 131–144 (cit. on p. 22).

Monzon, A. M., C. O. Rohr, M. S. Fornasari, and G. Parisi (2016). "CoDNaS 2.0: a comprehensive database of protein conformational diversity in the native state". In: *Database* 2016, baw038 (cit. on p. 147).

Moult, J., K. Fidelis, A. Kryshtafovych, T. Schwede, and A. Tramontano (2018). "Critical assessment of methods of protein structure prediction (CASP)-Round XII". In: *Proteins: Structure, Function, and Bioinformatics* 86, pp. 7–15 (cit. on pp. 30, 31, 42).

Muhammed, M. T. and E. Aki-Yalcin (2019). "Homology modeling in drug discovery: Overview, current applications, and future perspectives". In: *Chemical Biology & Drug Design* 93.1, pp. 12–20 (cit. on p. 27).

Murata, K. and M. Wolf (2018). "Cryo-electron microscopy for structural analysis of dynamic biological macromolecules". In: *Biochimica et Biophysica Acta (BBA) - General Subjects* 1862.2, pp. 324–334 (cit. on p. 11).

Murzin, A. G., S. E. Brenner, T. Hubbard, and C. Chothia (1995). "SCOP: A structural classification of proteins database for the investigation of sequences and structures". In: *Journal of Molecular Biology* 247.4, pp. 536–540 (cit. on pp. 9, 106).

Nero, T. L., M. W. Parker, and C. J. Morton (2018). "Protein structure and computational drug discovery". In: *Biochemical Society Transactions* 46.5, pp. 1367–1379 (cit. on p. 2).

Nilsson, O. B. et al. (2015). "Cotranslational Protein Folding inside the Ribosome Exit Tunnel". In: *Cell Reports* 12.10, pp. 1533–1540 (cit. on pp. 16, 18, 105).

Nissley, D. A. and E. P. O'Brien (2018). "Accurate prediction of forster resonance energy transfer during co-translational folding with coarse-grained molecular dynamics simulations". In: *Biophysical Journal* 114.3, 47a (cit. on pp. 14, 150).

Nissley, D. A., A. K. Sharma, et al. (2016). "Accurate prediction of cellular co-translational folding indicates proteins can switch from post- to co-translational folding". In: *Nature Communications* 7.1, p. 10341 (cit. on pp. 14, 150).

O'Brien, E. P., S.-T. D. Hsu, J. Christodoulou, M. Vendruscolo, and C. M. Dobson (2010). "Transient tertiary structure formation within the ribosome exit port". In: *Journal of the American Chemical Society* 132.47, pp. 16928–16937 (cit. on p. 14).

Ogorzalek, T. L. et al. (2018). "Small angle X-ray scattering and cross-linking for data assisted protein structure prediction in CASP 12 with prospects for improved accuracy". In: *Proteins: Structure, Function, and Bioinformatics* 86, pp. 202–214 (cit. on p. 147).

Onuchic, J. N., Z. Luthey-Schulten, and P. G. Wolynes (1997). "Theory of protein folding: The rnergy landscape perspective". In: *Annual Review of Physical Chemistry* 48.1, pp. 545–600 (cit. on p. 13).

Ovchinnikov, S., D. E. Kim, et al. (2016). "Improved de novo structure prediction in CASP11 by incorporating coevolution information into Rosetta". In: *Proteins: Structure, Function, and Bioinformatics* 84.S1, pp. 67–75 (cit. on pp. 24, 28, 151).

Ovchinnikov, S., H. Park, et al. (2017). "Protein structure determination using metagenome sequence data". In: *Science* 355.6322, pp. 294–298 (cit. on pp. 43, 44, 50, 59, 60, 69, 107, 138, 144).

Palopoli, N., A. M. Monzon, G. Parisi, and M. S. Fornasari (2016). "Addressing the role of conformational diversity in protein structure prediction". In: *PLOS ONE* 11.5. Ed. by S. C. E. Tosatto, e0154923 (cit. on p. 147).

Pan, B.-B. et al. (2016). "3D structure determination of a protein in living cells using paramagnetic NMR spectroscopy". In: *Chemical Communications* 52.67, pp. 10237–10240 (cit. on p. 11).

Panchenko, A. R., Z. Luthey-Schulten, and P. G. Wolynes (1996). "Foldons, protein structural modules, and exons." In: *Proceedings of the National Academy of Sciences of the United States of America* 93.5, pp. 2008–2013 (cit. on pp. 13, 106, 108, 110).

Pancsa, R., M. Varadi, P. Tompa, and W. F. Vranken (2016). "Start2Fold: a database of hydrogen/deuterium exchange data on protein folding and stability." In: *Nucleic Acids Research* 44.D1, pp. D429–34 (cit. on pp. 13, 104).

Park, H., G. R. Lee, L. Heo, and C. Seok (2014). "Protein loop modeling using a new hybrid energy function and its application to modeling in inaccurate structural environments". In: *PLOS ONE* 9.11. Ed. by Y. Zhang, e113811 (cit. on pp. 27, 72).

Parkinson, J. and M. Blaxter (2009). "Expressed Sequence Tags: An overview". In: *Methods in Molecular Biology* 533.1, pp. 1–12 (cit. on p. 150).

Pawlowski, M., L. Kozlowski, and A. Kloczkowski (2016). "MQAPsingle: A quasi single-model approach for estimation of the quality of individual protein structure models". In: *Proteins: Structure, Function, and Bioinformatics* 84.8, pp. 1021–1028 (cit. on p. 42).

Pechmann, S. and J. Frydman (2013). "Evolutionary conservation of codon optimality reveals hidden signatures of cotranslational folding". In: *Nature Structural & Molecular Biology* 20.2, pp. 237–243 (cit. on p. 17).

Punta, M. et al. (2012). "The Pfam protein families database". In: *Nucleic Acids Research* 40.D1, pp. D290–D301 (cit. on p. 46).

Ramachandran, G., C. Ramakrishnan, and V. Sasisekharan (1963). "Stereochemistry of polypeptide chain configurations". In: *Journal of Molecular Biology* 7.1, pp. 95–99 (cit. on p. 6).

Rice, P., L. Longden, and A. Bleasby (2000). *EMBOSS: The European Molecular Biology Open Software Suite* (cit. on pp. 77, 78).

Rigden, D. J. (2002). "Use of covariance analysis for the prediction of structural domain boundaries from multiple protein sequence alignments". In: *Protein Engineering, Design and Selection* 15.2, pp. 65–77 (cit. on pp. 23, 107, 137).

Rost, B. (2001). "Review: Protein Secondary Structure Prediction Continues to Rise". In: *Journal of Structural Biology* 134.2-3, pp. 204–218 (cit. on p. 20).

Rost, B., C. Sander, and R. Schneider (1994). "Redefining the goals of protein secondary structure prediction". In: *Journal of Molecular Biology* 235.1, pp. 13–26 (cit. on p. 20).

Sadowski, M. I. (2013). "Prediction of protein domain boundaries from inverse covariances". In: *Proteins: Structure, Function, and Bioinformatics* 81.2, pp. 253–260 (cit. on pp. 23, 107, 137, 141).

Sakahira, H. and S. Nagata (2002). "Co-translational folding of caspase-activated DNase with Hsp70, Hsp40, and inhibitor of caspase-activated DNase". In: *Journal of Biological Chemistry* 277.5, pp. 3364–3370 (cit. on p. 152).

Sakakibara, D. et al. (2009). "Protein structure determination in living cells by in-cell NMR spectroscopy". In: *Nature* 458.7234, pp. 102–105 (cit. on p. 11).

Šali, A. and T. L. Blundell (1993). "Comparative Protein Modelling by Satisfaction of Spatial Restraints". In: *Journal of Molecular Biology* 234.3, pp. 779–815 (cit. on p. 78).

Samelson, A. J., M. K. Jensen, R. A. Soto, J. H. D. Cate, and S. Marqusee (2016). "Quantitative determination of ribosome nascent chain stability". In: *Proceedings of the National Academy of Sciences of the United States of America* 113.47, pp. 13402–13407 (cit. on p. 18).

Samudrala, R. and J. Moult (1998). "An all-atom distance-dependent conditional probability discriminatory function for protein structure prediction". In: *Journal of Molecular Biology* 275.5, pp. 895–916 (cit. on pp. 32, 35).

Sánchez, I. E., M. Morillas, E. Zobeley, T. Kiefhaber, and R. Glockshuber (2004). "Fast folding of the two-domain Semliki Forest virus capsid protein explains co-translational proteolytic activity". In: *Journal of Molecular Biology* 338.1, pp. 159–167 (cit. on p. 152).

Sander, I. M., J. L. Chaney, and P. L. Clark (2014). "Expanding Anfinsen's Principle: Contributions of synonymous codon selection to rational protein design". In: *Journal of the American Chemical Society* 136.3, pp. 858–861 (cit. on p. 17).

Saunders, R. and C. M. Deane (2010a). "Protein structure prediction begins well but ends badly". In: *Proteins* 78.5, pp. 1282–90 (cit. on p. 14).

Saunders, R. and C. M. Deane (2010b). "Synonymous codon usage influences the local protein structure observed". In: *Nucleic Acids Research* 38.19, pp. 6719–6728 (cit. on p. 17).

Saunders, R., M. Mann, and C. M. Deane (2011). "Signatures of co-translational folding". In: *Biotechnology Journal* 6.6, pp. 742–751 (cit. on pp. 14, 84).

Savitsky, P. et al. (2010). "High-throughput production of human proteins for crystallization: The SGC experience". In: *Journal of Structural Biology* 172.1, pp. 3–13 (cit. on p. 72).

Schaarschmidt, J., B. Monastyrskyy, A. Kryshtafovych, and A. M. Bonvin (2018). "Assessment of contact predictions in CASP12: Co-evolution and deep learning coming of age". In: *Proteins: Structure, Function, and Bioinformatics* 86, pp. 51–66 (cit. on p. 22).

Schaeffer, R. D., A. L. Jonsson, A. M. Simms, and V. Daggett (2011). "Generation of a consensus protein domain dictionary". In: *Bioinformatics* 27.1, pp. 46–54 (cit. on p. 106).

Seemayer, S., M. Gruber, and J. Söding (2014). "CCMpred—fast and precise prediction of protein residue–residue contacts from correlated mutations". In: *Bioinformatics* 30.21, pp. 3128–3130 (cit. on p. 22).

Senior, A. W. et al. (2019). "Protein structure prediction using multiple deep neural networks in the 13th Critical Assessment of Protein Structure Prediction (CASP13)". In: *Proteins: Structure, Function, and Bioinformatics* 87.12, pp. 1141–1148 (cit. on pp. 30, 72, 148, 151).

Serber, Z. et al. (2001). "High-resolution macromolecular NMR spectroscopy inside living cells". In: *Journal of the American Chemical Society* 123.10, pp. 2446–2447 (cit. on p. 11).

Sevier, C. S. and C. A. Kaiser (2002). "Formation and transfer of disulphide bonds in living cells". In: *Nature Reviews Molecular Cell Biology* 3.11, pp. 836–847 (cit. on p. 4).

Shi, Q. et al. (2019). "DNN-Dom: predicting protein domain boundary from sequence alone by deep neural network". In: *Bioinformatics* 35.24. Ed. by A. Elofsson, pp. 5128–5136 (cit. on p. 137).

Shrestha, R. et al. (2019). "Assessing the accuracy of contact predictions in CASP13". In: *Proteins: Structure, Function, and Bioinformatics* 87.12, pp. 1058–1068 (cit. on pp. 22, 23, 146).

Sillitoe, I. et al. (2015). "CATH: comprehensive structural and functional annotations for genome sequences". In: *Nucleic Acids Research* 43.D1, pp. D376–D381 (cit. on pp. 8, 9, 77, 106, 109).

Simkovic, F., S. Ovchinnikov, D. Baker, and D. J. Rigden (2017). "Applications of contact predictions to structural biology". In: *IUCrJ* 4.3, pp. 291–300 (cit. on p. 23).

Sonnhammer, E. L., S. R. Eddy, and R. Durbin (1997). "Pfam: A comprehensive database of protein domain families based on seed alignments". In: *Proteins: Structure, Function, and Genetics* 28.3, pp. 405–420 (cit. on p. 8).

Sugiki, T., N. Kobayashi, and T. Fujiwara (2017). *Modern Technologies of Solution Nuclear Magnetic Resonance Spectroscopy for Three-dimensional Structure Determination of Proteins Open Avenues for Life Scientists* (cit. on p. 11).

Sutto, L., S. Marsili, A. Valencia, and F. L. Gervasio (2015). "From residue coevolution to protein conformational ensembles and functional dynamics". In: *Proceedings of the National Academy of Sciences of the United States of America* 112.44, pp. 13567–13572 (cit. on p. 23).

Suyama, M. and O. Ohara (2003). "DomCut: prediction of inter-domain linker regions in amino acid sequences". In: *Bioinformatics* 19.5, pp. 673–674 (cit. on p. 107).

Tan, S., H. T. Tan, and M. C. M. Chung (2008). "Membrane proteins and membrane proteomics". In: *Proteomics* 8.19, pp. 3924–3932 (cit. on p. 153).

Thomas, J. M. H. et al. (2017). "Approaches to ab initio molecular replacement of α-helical transmembrane proteins". In: *Acta Crystallographica Section D Structural Biology* 73.12, pp. 985–996 (cit. on pp. 23, 147).

Tosatto, S. C. (2005). "The Victor/FRST Function for Model Quality Estimation". In: *Journal of Computational Biology* 12.10, pp. 1316–1327 (cit. on p. 35).

Trevizani, R., F. L. Custódio, K. B. dos Santos, and L. E. Dardenne (2017). "Critical features of fragment libraries for protein structure prediction". In: *PLOS ONE* 12.1. Ed. by Y. Zhang, e0170131 (cit. on p. 29).

Uemura, S. et al. (2010). "Real-time tRNA transit on single translating ribosomes at codon resolution". In: *Nature* 464.7291, pp. 1012–1017 (cit. on p. 17).

Ugrinov, K. G. and P. L. Clark (2010). "Cotranslational folding increases GFP folding yield". In: *Biophysical Journal* 98.7, pp. 1312–1320 (cit. on p. 16).

Uziela, K., D. Menéndez Hurtado, N. Shu, B. Wallner, and A. Elofsson (2017). "ProQ3D: improved model quality assessments using deep learning". In: *Bioinformatics* 33.10, btw819 (cit. on pp. 31, 42, 49, 50, 69).

Voss, N., M. Gerstein, T. Steitz, and P. Moore (2006). "The geometry of the ribosomal polypeptide exit tunnel". In: *Journal of Molecular Biology* 360.4, pp. 893–906 (cit. on p. 18).

Wallner, B. (2006). "Identification of correct regions in protein models using structural, alignment, and consensus information". In: *Protein Science* 15.4, pp. 900–913 (cit. on pp. 31, 50).

Wang, G. and R. L. Dunbrack (2003). "PISCES: a protein sequence culling server". In: *Bioinformatics* 19.12, pp. 1589–1591 (cit. on p. 35).

Wang, J. et al. (2016). "Exploring Human Diseases and Biological Mechanisms by Protein Structure Prediction and Modeling". In: pp. 39–61 (cit. on p. 2).

Wang, K., J. A. Horst, G. Cheng, D. C. Nickle, and R. Samudrala (2008). "Protein meta-functional signatures from combining sequence, structure, evolution, and amino acid property information". In: *PLOS Computational Biology* 4.9. Ed. by R. B. Altman, e1000181 (cit. on p. 149).

Wang, S., J. Peng, J. Ma, and J. Xu (2016). "Protein secondary structure prediction using deep convolutional neural fields". In: *Scientific Reports* 6.1, p. 18962 (cit. on pp. 20, 47, 81, 161).

Wang, Y. et al. (2019). "Fueling ab initio folding with marine metagenomics enables structure and function predictions of new protein families". In: *Genome Biology* 20.1, p. 229 (cit. on p. 149).

Waudby, C. A., C. M. Dobson, and J. Christodoulou (2019). "Nature and regulation of protein folding on the ribosome". In: *Trends in Biochemical Sciences* 44.11, pp. 914–926 (cit. on pp. 2, 14, 15, 17, 18).

West, C. E., S. H. P. de Oliveira, and C. M. Deane (2019). "RFQAmodel: Random Forest Quality Assessment to identify a predicted protein structure in the correct fold". In: *PLOS ONE* 14.10. Ed. by Y. Zhang, e0218149 (cit. on pp. 38, 41).

Wlodawer, A. and Z. Dauter (2017). "'Atomic resolution': a badly abused term in structural biology". In: *Acta Crystallographica Section D Structural Biology* 73.4, pp. 379–380 (cit. on p. 10).

Wollacott, A. M., A. Zanghellini, P. Murphy, and D. Baker (2007). "Prediction of structures of multidomain proteins from structures of the individual domains." In: *Protein Science* 16.2, pp. 165–75 (cit. on pp. 31, 151).

Won, J., M. Baek, B. Monastyrskyy, A. Kryshtafovych, and C. Seok (2019). "Assessment of protein model structure accuracy estimation in CASP13: Challenges in the era of deep learning". In: *Proteins: Structure, Function, and Bioinformatics* 87.12, pp. 1351–1360 (cit. on pp. 30–32).

Wong, S. W. K., J. S. Liu, and S. C. Kou (2017). "Fast de novo discovery of low-energy protein loop conformations". In: *Proteins: Structure, Function, and Bioinformatics* 85.8, pp. 1402–1412 (cit. on pp. 27, 72).

Wu, C. C. C., B. Zinshteyn, K. A. Wehner, and R. Green (2019). "High-resolution ribosome profiling defines discrete ribosome elongation states and translational regulation during cellular stress". In: *Molecular Cell* 73.5, 959–970.e5 (cit. on p. 17).

Wu, S. and Y. Zhang (2008). "ANGLOR: A composite machine-learning algorithm for protein backbone torsion angle prediction". In: *PLoS ONE* 3.10. Ed. by D. Jones, e3400 (cit. on p. 21).

Xu, D., L. Jaroszewski, Z. Li, and A. Godzik (2015). "AIDA: Ab initio domain assembly for automated multi-domain protein structure prediction and domain-domain interaction prediction." In: *Bioinformatics* 31.13, pp. 2098–105 (cit. on pp. 31, 151).

Xu, D. and Y. Zhang (2012). "Ab initio protein structure assembly using continuous structure fragments and optimized knowledge-based force field". In: *Proteins: Structure, Function, and Bioinformatics* 80.7, pp. 1715–1735 (cit. on p. 28).

Xu, J. (2019). "Distance-based protein folding powered by deep learning". In: *Proceedings of the National Academy of Sciences of the United States of America* 116.34, pp. 16856–16865 (cit. on p. 30).

Xu, J. and S. Wang (2019). "Analysis of distance-based protein structure prediction by deep learning in CASP13". In: *Proteins: Structure, Function, and Bioinformatics* 87.12, pp. 1069–1081 (cit. on p. 23).

Xu, J. and Y. Zhang (2010). "How significant is a protein structure similarity with TM-score = 0.5?" In: *Bioinformatics* 26.7, pp. 889–895 (cit. on pp. 26, 48).

Xu, Y., L. Mayne, and S. W. Englander (1998). "Evidence for an unfolding and refolding pathway in cytochrome c". In: *Nature Structural Biology* 5.9, pp. 774–778 (cit. on p. 13).

Xue, B., O. Dor, E. Faraggi, and Y. Zhou (2008). "Real-value prediction of backbone torsion angles". In: *Proteins: Structure, Function, and Bioinformatics* 72.1, pp. 427–433 (cit. on p. 21).

Yang, J. and Y. Zhang (2015). "Protein Structure and Function Prediction Using I-TASSER". In: *Current Protocols in Bioinformatics* 52.1 (cit. on p. 28).

Yang, Y. et al. (2018). "Sixty-five years of the long march in protein secondary structure prediction: the final stretch?" In: *Briefings in Bioinformatics*, bbw129 (cit. on p. 20).

Yu, J., J. Andreani, F. Ochsenbein, and R. Guerois (2017). "Lessons from (co-)evolution in the docking of proteins and peptides for CAPRI Rounds 28-35". In: *Proteins: Structure, Function, and Bioinformatics* 85.3, pp. 378–390 (cit. on p. 23).

Zemla, A. (2003). "LGA: a method for finding 3D similarities in protein structures". In: *Nucleic Acids Research* 31.13, pp. 3370–3374 (cit. on p. 26).

Zhang, C., P. L. Freddolino, and Y. Zhang (2017). "COFACTOR: Improved protein function prediction by combining structure, sequence and protein–protein interaction information". In: *Nucleic Acids Research* 45.W1, W291–W299 (cit. on p. 149).

Zhang, G., M. Hubalewska, and Z. Ignatova (2009). "Transient ribosomal attenuation coordinates protein synthesis and co-translational folding". In: *Nature Structural & Molecular Biology* 16.3, pp. 274–280 (cit. on pp. 16–18).

Zhang, Y. (2005). "TM-align: a protein structure alignment algorithm based on the TM-score". In: *Nucleic Acids Research* 33.7, pp. 2302–2309 (cit. on pp. 26, 112).

Zhang, Y. (2009). "Protein structure prediction: when is it useful?" In: *Current Opinion in Structural Biology* 19.2, pp. 145–155 (cit. on p. 26).

Zhang, Y. and J. Skolnick (2004). "Scoring function for automated assessment of protein structure template quality". In: *Proteins: Structure, Function, and Bioinformatics* 57.4, pp. 702–710 (cit. on pp. 48, 112).

Zmasek, C. M. and A. Godzik (2012). "This deja vu feeling? Analysis of multidomain protein evolution in eukaryotic genomes". In: *PLOS Computational Biology* 8.11. Ed. by C. A. Orengo, e1002701 (cit. on p. 151).