# Protein folding and structure prediction: biological and physical perspectives


Mark Chonofsky

Exeter College

University of Oxford

A thesis submitted for the degree of

*Doctor of Philosophy*

Hilary 2020

*In memoriam* Dr. Susan Offner

# Acknowledgements

The people for whose support I have been grateful over the last four years are too many to count and this list is incomplete. The Systems Biology Doctoral Training Centre funded this work, and the support of the staff of the Doctoral Training Centre and the Department of Statistics has been indispensable. I am particularly grateful to the Statistics IT staff, including Stuart McRobert, Susan Hutchinson, Mark Feasey, and Simon Patchett.

Charlotte Deane supervised this thesis, and it has been a tremendous privilege to learn from her talents in science and mentorship.

I grateful to a litany of friends, family, and colleagues, especially to my grandmother, Shirley, my mother, Helen, and my sister, Emma; and to Verity Algar, Andrew Allen, Fergus Boyles, Lyuba Bozhilova, Toby Cain, Ayanna Coleman-Potempa, Tim Ecott, Tanya Finkelstein, Eleanor Law, Claire Marks, Gabriela Minden, Miranda Miller, Shulamit Morris-Evans, Dan Nissley, Saulo Oliveira, Chloe Pinto, Sarah Robinson, Carlos Rubiera Outeiral, Dominik Schwarz, Zahra Seyyad, Karin Sigloch, Maya Silver, Miriam Stricker, Barnaby Taylor, Ayça Türkoğlu, Thomas Waters, and Clare West.

Thank you to the owners of every desktop computer on the second floor of the Department of Statistics for invaluable assistance with the calculations reported in Chapters 2 and 3.

Finally: the life of Susan Offner, my biology teacher, to which this thesis is dedicated, touched countless students and colleagues in Lexington and across the United States. To be unable to discuss this thesis with her leaves an enormous void. The memory of the righteous is truly a blessing.

# Abstract

Protein folding is poorly understood and distinct from protein structure prediction, encompassing the translation and folding process of proteins *in vivo* and *in vitro*. In this thesis, we examine the protein folding process from several perspectives. First, we implement models of the spatial constraints that the ribosome introduces in the protein folding process. We are unable to show that these constraints have a particular effect on the formation of protein structure. However, we identify and describe a method which improves protein structure prediction in SAINT2 by avoiding the disruptive effects of certain extension protocols. Then, we study the relationship between protein structure and the coevolutionary constraints under which the structure has evolved. We identify several relationships between the structure of the protein and the coevolutionary constraints that are found in the protein. We further compare contact prediction methods and show that different contact prediction methods identify different types of amino-acid contacts. Finally, we investigate obligate cotranslational protein folding in the *E. coli* proteome. We find a relationship between the behaviour of several folding analysis methods and the necessity of cotranslational folding for our proteins.

# Contents

# List of Figures

# List of Abbreviations

| | |
|---|---|
| aMIc | Adjusted Mutual Information, corrected, as described in Lee *et al.,* 2009. |
| ASTRAL | A database of protein domain information, as described in Brenner *et al.,* 2000. |
| BG | Background |
| CASP | Critical Assessment of Structure Prediction, a semiannual blind test of prediction methods (Moult *et al.,* 2018). |
| CATH | The CATH (Class, Architecture, Topology, Homologous superfamily) database of protein domains (Knudsen *et al.,* 2010; Dawson *et al.,* 2017). |
| CC-BY-SA | The Creative Commons Attribution-ShareAlike license. |
| CCMpred | A pseudolikelihood maximization program for contact inference due to Seemayer *et al.,* 2014. |
| DCA | Direct Coupling Analysis, which refers generally to the removal of transitive couplings from the set of informative couplings, usually by a pseudolikelihood maximization method. |
| DNA | Deoxyribonucleic acid |
| DSSP | A program, originally designed by Wolfgang Kabsch and Chris Sander (Kabsch *et al.,* 1983), that annotates protein structures with secondary structure types. |
| EVFold | A pioneering method for inferring protein contacts from sequence alignments (Marks *et al.,* 2011). |
| FM | The Free Modeling category in CASP. |
| gDCA | GaussDCA, a DCA method using a Gaussian modelling technique (Baldassi *et al.,* 2014). |
| GdmCl | Guanidinium chloride |
| GDT_TS | Global Distance Test (Total Score), a measure which describes the structural similarity of two protein structure models (Zemla *et al.,* 2003). |
| GREMLIN | Generative REgularized ModeLs of proteINs, a DCA-based method for contact inference (Balakrishnan *et al.,* 2011). |

| | |
|---|---|
| HHBlits | A protein multiple sequence alignment technique which uses a hidden Markov model approach (Remmert *et al.,* n.d.). |
| HIV | Human immunodeficiency virus |
| MD | Molecular dynamics |
| MetaPSICOV | A metaprediction algorithm for inferring protein contacts (Jones, Singh, *et al.,* 2015). |
| MI | Mutual information |
| MIc | Mutual information (corrected), as described in Lee *et al.,* 2009. |
| MIp | Mutual information without the influence of phylogeny, using the Average Product Correction (Dunn *et al.,* 2008). |
| mRNA | Messenger RNA |
| MSA | Multiple sequence alignment |
| NMR | Nuclear magnetic resonance |
| PconsC | A protein contact metaprediction program (Skwark *et al.,* 2013). |
| PDB | The Protein Data Bank, the international database of solved protein structures. |
| PDI | Protein disulfide isomerase |
| PPV | Positive predictive value |
| ProsPR | A fully-open implementation of the AlphaFold algorithm (Billings *et al.,* 2019). |
| PSI-BLAST | Position-Specific Iterative Basic Local Alignment Search Tool, a method for searching for similar protein sequences on the basis of a position-specific scoring matrix. |
| PSICOV | A DCA-based protein contact prediction method (Jones, Buchan, *et al.,* 2012). |
| PSIPRED | An algorithm for the prediction of protein secondary structure from sequence, implemented by Jones, Buchan, *et al.,* 2012. |

| | |
|---|---|
| PULCHRA | The PowerfUL CHain Restoration Algorithm, software which can be used to reconstruct chain or atom positions in protein models Rotkiewicz *et al.,* 2008. |
| RAPDF | The Residue-specific All-atom Probability Discriminatory Function (Samudrala *et al.,* 1998) |
| REMD | Replica-exchange molucular dynamics |
| RMSD | Root-mean-square deviation |
| RNA | Ribonucleic acid |
| ROC | Receiver operating characteristic |
| SAINT2 | A protein structure prediction program which uses fragment replacement under a Monte Carlo paradigm (de Oliveira, Law, *et al.,* 2018). |
| SCOPe | Structural Classification of Proteins-extended, a database of protein structural relationships that extends SCOP using automated methods (Fox *et al.,* 2014). |
| SCOP | Structural Classification of Proteins, a manually-curated database of protein structural relationships (Murzin, 1995). |
| SIFts | Structural Interaction Fingerprints |
| SS | Secondary structure |
| SSAP | The Sequential Structure Alignment Program, a method for aligning proteins by structure which has been used to construct the CATH database (Orengo *et al.,* 1996). |
| STRIDE | STRucture IDEntification, a knowledge-based algorithm for assigning protein secondary structures (Frishman *et al.,* 1995). |
| TBM | The Template-Based Modeling category in CASP. |
| TIM barrel | A transmembrane $\alpha/\beta$ structural motif, named after triosephosphate isomerase (TIM). |
| tRNA | Transfer RNA |

| | |
|---|---|
| VdW | Van der Waals |
| VemP | Vibrio protein export monitoring polypeptide, a regulatory protein involved in translation in *Vibrio alginolyticus*. |
| ZNMI | Z-scored Product Mutual Information, a pairwise mutual information-based scoring method that can be applied to multiple sequence alignments for contact inference(Brown *et al.,* 2010). |

# Chapter 1

# Introduction

Seventy-five percent of the eukaryotic cellular energy budget makes and destroys proteins (Lane *et al.,* 2010). These intricate molecular sculptures lie at the heart of the biochemistry of all life on Earth. Yet, since the discovery in 1957 that proteins have structure (Kendrew *et al.,* 1958), the mechanisms by which these structures are formed has remained poorly understood.

In this chapter, we review the science of protein structure formation and analysis, beginning with the chemical composition of proteins and the types of structures that proteins form. Then, we discuss what is known about how these structures are formed, and the methods that are used to predict protein structures and to study protein folding and structure formation.

## 1.1 The composition of protein structure

Proteins are chains of amino acids connected by peptide bonds (Fig. 1.1). Typically hundreds of amino acids long in natural organisms (Tiessen *et al.,* 2012), these molecules are formed through the ribosome-mediated condensation of amino acids according to an mRNA template (Berg *et al.,* 2019). As shown in Fig. 1.1, amino acids are composed of a backbone of two carbon atoms and one nitrogen atom, along with one of twenty side chains attached to the $\alpha$ carbon. (In the case of proline, this side chain is a five-membered ring that attaches to the terminal carbon as well.) The nitrogen forms part of an amino group, which bonds to other amino acids through a condensation reaction with the carbonyl group of another amino acid. After condensation, these amino acids are called residues.

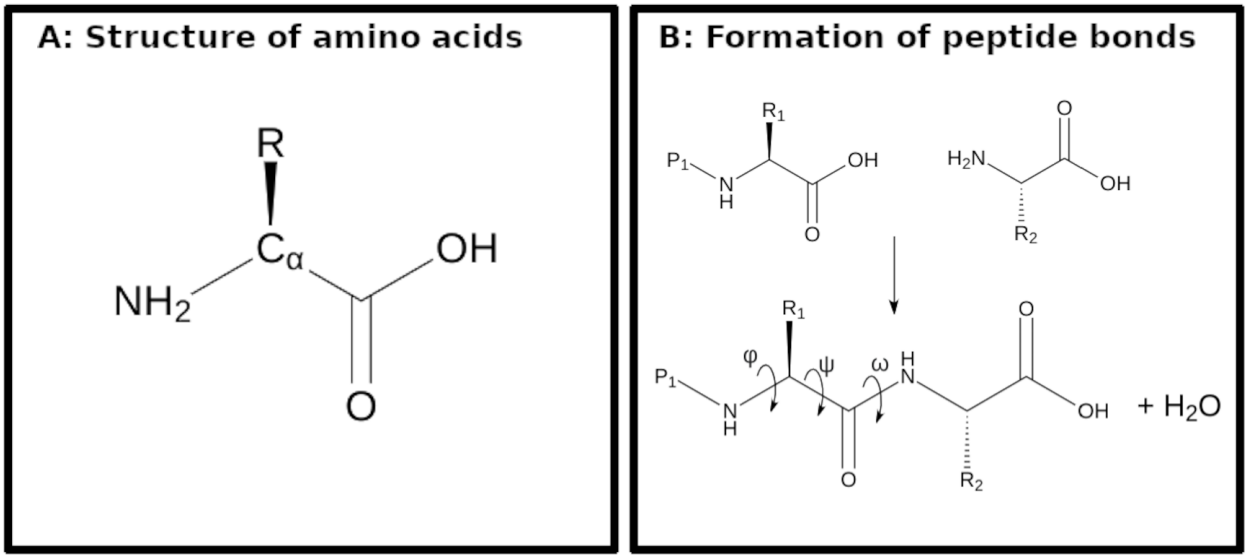


Figure 1.1: **Chemical structure of amino acids and the formation of peptide bonds through condensation. A:** The central $C_\alpha$ atom is bonded to an amino group ($NH_2$) and a carboxyl group (COOH), which together comprise the backbone atoms, along with a variable R group, which defines the unique chemical properties of the amino acid. **B:** The condensation of the amino and carboxyl groups from two amino acids results in the formation of the N-O peptide bond between amino acids. The backbone torsion angles are also shown.

### 1.1.1 Peptide structure at the molecular level

The 20 distinct side chains of the biological amino acids confer different chemical properties on each amino acid, and these side chains can be grouped by their physico-chemical properties (Fig. 1.2). Alanine, valine, leucine, and isoleucine have simple branched hydrocarbon side chains containing one to four carbons. Methionine and cysteine contain a sulfur atom, which, in the case of cysteine, is terminal and thus can form disulfide bonds with other cysteine residues.

Certain amino acids—phenylalanine, tyrosine, histidine, and tryptophan—contain aromatic rings, which can form aromatic stacking interactions with other aromatic residues. At physiological pH, arginine, histidine, and lysine are positively charged, while aspartic and glutamic acid are negatively charged. The positively-charged amino acids contain amino groups in their side chains, while the negatively-charged amino acids derive their charge from a carboxylic acid. By contrast, asparagine, glutamine, serine, and threonine contain amine and hydroxyl groups, which confer polarity without resulting in charge at physiological pH.

The remaining two amino acids, glycine and proline, have singular structural properties. In the case of glycine, its side-chain is simply a hydrogen atom, allowing a greater degree of flexibility than the remaining amino acids. By contrast, the five-membered ring which joins the carbonyl and $\alpha$ carbons in proline cause it to adopt a substantially lower degree of conformational flexibility than other amino acids (Richardson *et al.,* 2012).

These effects are particularly clear with reference to the internal bond angles that amino acids adopt in folded structures. The dihedral angles $\phi$ and $\psi$—the angles between the bonds before and after the N-$C_\alpha$ and $C_\alpha$-C bonds, respectively, when projected into the plane perpendicular to that connecting bond—are relatively unconstrained for glycine, while proline adopts a very narrow range of values unlike other amino acids (Fig. 1.3). The planar bond angles, *i.e.*, the angles between any three consecutive backbone atoms, also tend to exhibit little variability (Balasco *et al.,* 2017).

The configuration of the protein around the peptide bond, described by third torsion angle, $\omega$, tends to be planar, because the delocalization of carbonyl $\pi$ and nitrogen lone-pair electrons contributes to partial $\pi$ character of the peptide bond. Typically, the carbonyl oxygen and the amino hydrogen point in opposite directions (the *trans* conformation), minimizing steric

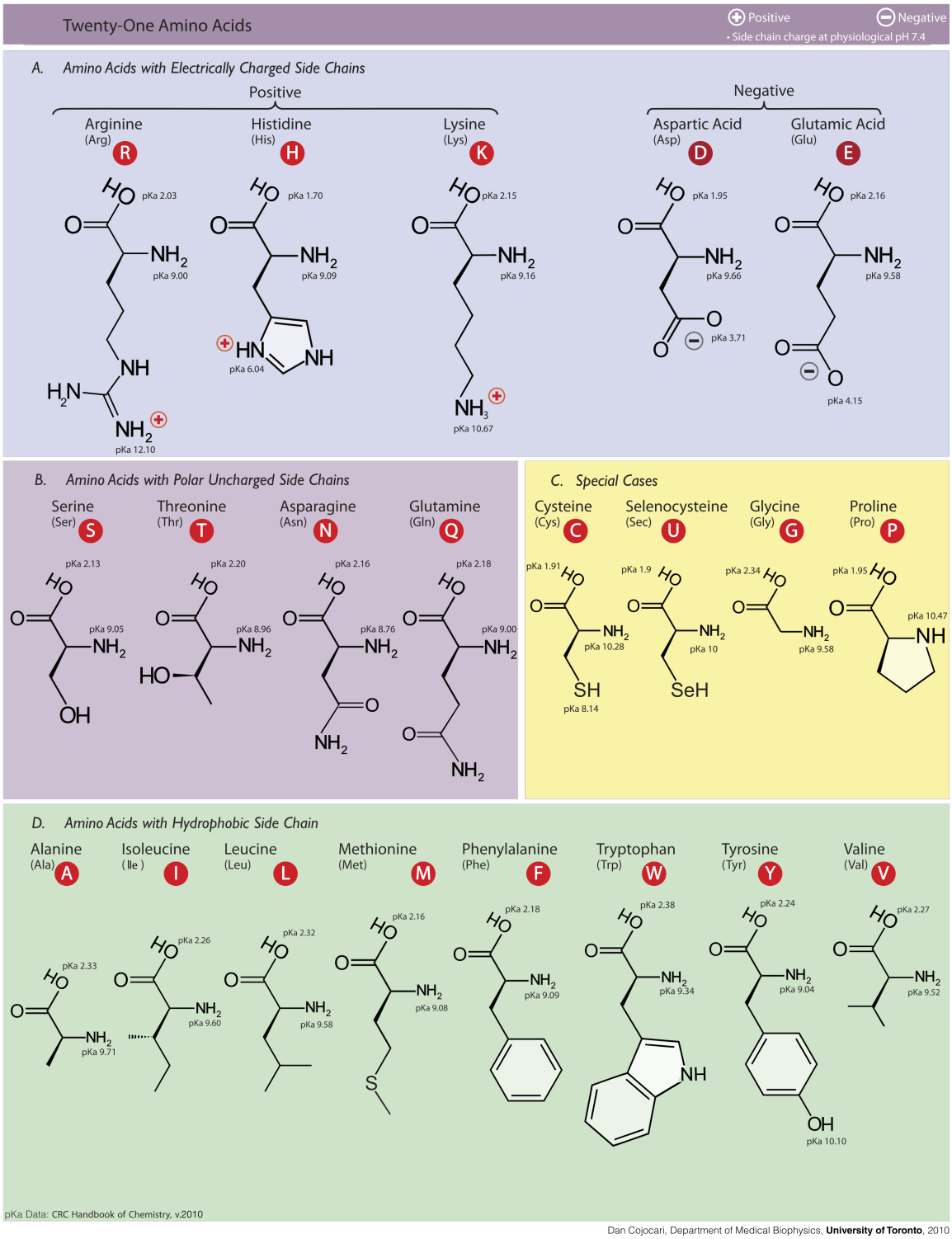


Figure 1.2: **Chemical structures of the 20 canonical amino acids.** Figure adapted from original by Dan Cojocari (Wikimedia Commons, CC-BY-SA 3.0).

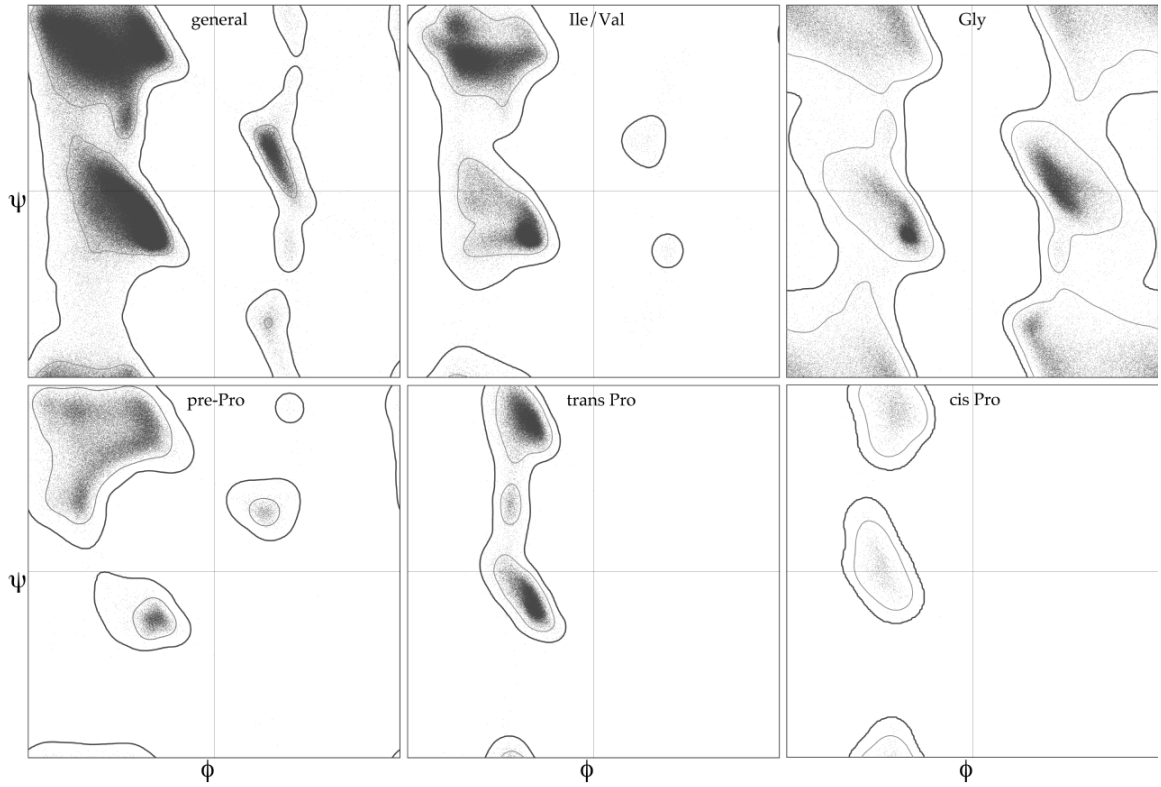


Figure 1.3: **Ramachandran plot for the six world-wide PDB amino acid validation categories.** Data from 8000 protein chains containing approximately 1.5 million amino acids, filtered for quality, as described in Richardson et al., 2012. Figure from Jane Shelby Richardson via Wikimedia Commons (CC-BY-SA 3.0).

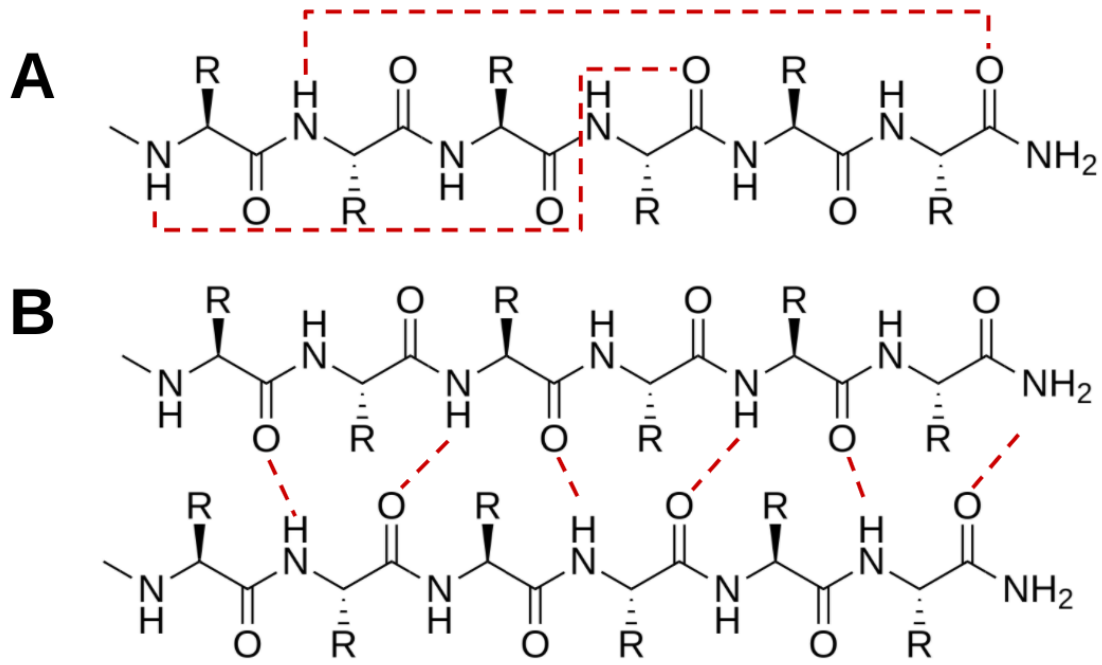


Figure 1.4: **Patterns of hydrogen bonding in common secondary structures** In $\alpha$-helices (**A**), the amino group of amino acid $i$ bond hydrogen bonds with the carbonyl group of amino acid $i + 3$ or $i + 4$. In parallel $\beta$-sheets (**B**), hydrogen bonds cross between the two sheets. Antiparallel $\beta$-sheets, in which the N-termini of adjacent strands are at opposite ends of the sheet, allow planar hydrogen bonds, which are energetically more favourable.

hindrance with nearby atoms. In a minority of cases (less than 1%), these atoms are found in *cis* instead. When followed by proline, this proportion rises to 5% in natural proteins and up to 30% in random coil peptides (Craveur *et al.,* 2013).

### 1.1.2 Chemical interactions within protein structures

As shown in Fig. 1.4, protein conformations frequently adopt patterns characterized by backbone hydrogen bonding, a property known as secondary structure. $\alpha$-helices are the most common of these structures (Haimov *et al.,* 2016). In $\alpha$-helices, the carbonyl oxygens in a series of amino acids form a hydrogen bonds with backbone nitrogens three or four amino acids away, creating a helical structure with side-chains oriented externally to the helix. Similarly, $\beta$-strands involve ladder-like hydrogen bonding between the same atoms in continuous sequences of residues, forming long hydrogen-bond-stabilized structures. By exploiting both the nitrogen and oxygen atoms

in each amino acid, an amino acid can form these kinds of interactions with two other amino acids, allowing the formation of arbitrarily-wide $\beta$-sheets. There are also other—less common—secondary structures, including the $3_{10}$ helix (a right-handed helix with three amino acids per turn) and the $\beta$-turn, which is a hairpin turn stabilized by a backbone hydrogen bond.

Hydrogen bonds are the principal stabilising interaction in protein secondary structures and are ubiquitous throughout protein structures. They are electrostatic interactions arising from the sharing of a hydrogen atom between a donor and an acceptor. Structural evidence suggests that these bonds have a geometry where the hydrogen lies approximately 2Å from the donor and the acceptor. The length of the bond principally accounts for its strength (Hubbard *et al.,* 2010). The strength of a single hydrogen bond is 10-40 kJ $mol^{-1}$. However, the energetic cost of a loss of a hydrogen bond is not significantly different from the energetic gain due to the resulting opportunity of the donor and acceptor to hydrogen bond with solvent molecules. Any energetic difference is dominated by the gain in entropy due to the change in solvation (Hubbard *et al.,* 2010).

Although hydrogen bonding drives backbone-backbone interactions through the formation of secondary structure, the principal interactions that drive the formation of high-level structure in proteins relate to side-chain interactions (Kayikci *et al.,* 2018). Of these, hydrophobicity is central. In a polar solvent, non-polar amino acids tend to collapse toward the centre of the protein, while polar side chains favour the exterior of the protein (Dyson *et al.,* 2006). Addition of non-polar constituents to the solvent, such as urea, is a common way to unfold proteins in the laboratory by solvating hydrophobic amino acids (Bennion *et al.,* 2003).

Hydrophobicity is a result of collective interactions between solvent molecules and amino acid side chains. In particular, there is a enthalpic benefit from the interactions between polar regions of side chains and the polar solvent molecules. These favourable interactions cause the coordination of water molecules around polar regions of the protein (Frank *et al.,* 1945), which decrease the number of thermodynamically-accessible conformations. As a result, the hydrophobic interaction is a balance of entropic and enthalpic effects (Lumry *et al.,* 1970). Much recent work uses MD to explore the nature of water interactions with proteins directly, modeling water molecules explicitly (Schauperl *et al.,* 2016; Zhu *et al.,* 2016).

Another fundamental interaction which causes the formation of protein structure are van der Waals forces. These forces arise due to the mutual interaction between dipoles, such as those due to the electrons in atoms. These forces cause an attraction at short distances between molecules, typically counterbalanced by the Pauli repulsion at very short distances. At large distances, van der Waals forces have no effect. The 12-6 Lennard-Jones potential $v(r) = 4\epsilon\left[\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6}\right]$ is the most common representation of this effect in simulation, though the true effect is much more complicated (Bordner *et al.,* 2003; X. Wang *et al.,* 2020). ($\sigma$ is a distance scale, $\epsilon$ is the energy scale, and $r$ is the separation between the two atoms.) This force generally promotes compactness in protein structures (Sung, 2015).

Some amino acids contain side-chain heavy atoms that are not carbon, and these atoms give rise to two further effects. First, the formation of ions in the side chains lead to electrostatic effects which stabilize protein structure. The primary contribution to the favourable interaction energy of two charged groups is the Coulombic interaction potential $V_{ij}(r) = k\frac{q_i q_j}{\epsilon r_{ij}}$, offset by the unfavourable desolvation potential and the entropic cost of ordering the structure (Bosshard *et al.,* 2004) for charges $q$ at distance $r$. ($k$ and the dielectric constant $\epsilon$ set the energy scale here.) Side-chain ions can also interact with structural dipoles, such as helices. Intrinsic dipoles originating with backbone atoms may have a particular role in the formation of secondary structures in folding proteins (Ganesan *et al.,* 2014).

Moreover, cysteine contains a sulfide group, which can form interchain covalent interactions. These bonds involve about half of cysteine residues and are present in a third of expressed proteins, particularly those proteins that are secreted (Bastolla *et al.,* 2005; Bosnjak *et al.,* 2014). Disulfide bonds are strongly stabilizing and appear to be under strongly positive selection over time (Wong *et al.,* 2011).

These interactions can be characterised geometrically from protein structures, using software such as ARPEGGIO (Jubb *et al.,* 2017) or GetContacts (Venkatakrishnan *et al.,* 2019). These tools use geometric and chemical heuristics to identify which of these interactions involve which atoms and the amino acid residues that they comprise. Yet, because of the fact that amino-acid interactions are inherently many-body, and the fact that free energies of binding for most interaction types are influenced by complex enthalpic and entropic factors, it is normally impossible

to ascribe structure formation to particular types of interactions.

### 1.1.3 Types of protein folds

Although proteins structures exhibit high levels of diversity, many workers have given identified commonalities between them. Attempts include the DALI (Holm *et al.,* 2006), ASTRAL (Brenner *et al.,* 2000; J.-M. Chandonia, 2004), SCOP (Murzin, 1995), SCOPe (Fox *et al.,* 2014), and CATH (Dawson *et al.,* 2017) databases, which categorise proteins by high-level structural composition. At their highest level, CATH, ASTRAL, and SCOP/SCOPe divide proteins into $\alpha$, $\beta$, and mixed $\alpha/\beta$ classes, and proteins with few secondary structures. (SCOP and SCOPe further divides mixed secondary structures into $\alpha + \beta$ and $\alpha/\beta$, depending on whether the secondary structures are segregated or interwoven.) Automated structural comparison methods, such as those used to build the CATH (Dawson *et al.,* 2017) and SCOPe databases (Fox *et al.,* 2014), enable these databases to scale as new structures are deposited in the PDB. SCOPe is an extension of the manually-curated SCOP database, aiming to maintain the same level of accuracy as SCOP through automatic curation and the correction of some SCOP errors.

These commonalities are reflective of highly-conserved folding motifs, of which the most common are the Rossman fold, TIM barrels, the alpha/beta plait, and immunoglobulin and oligonucleotide-binding folds (Mirny *et al.,* 1999). Within these folds, certain sites tend to be highly conserved, possibly pointing to convergent functional or folding requirements (Mirny *et al.,* 1999). On a larger scale, the number of protein 'families' of common structure observed in nature tends to lie in the low thousands, depending on estimation methodology (Anishchenko *et al.,* 2017; Ovchinnikov, Kamisetty, *et al.,* 2014; Dawson *et al.,* 2017). There are also efforts to classify the entire space of possible structures on the basis of geometrical criteria (Taylor, 2020).

## 1.2 Protein structure formation

Since Christian Anfinsen's seminal work five decades ago (Anfinsen, 1973), researchers have believed that in most cases, the sequence of amino acids in a protein fully determines the folded structure of the protein. In fact, Anfinsen observed that for a small protein—a ribonuclease—with eight cysteine residues, chemical unfolding and refolding caused all of the native disulfide

bonds to reform and the protein to regain native-like levels of activity. Yet the number of possible protein conformations is very large, and simple calculations establish that proteins *in vivo* cannot attempt all possible conformations (Levinthal, 1969).

Instead, it is likely that most proteins fold along one or more folding pathways in accordance with thermondynamic pressures (Karplus, 1997). The classical view of protein folding holds that proteins fold within a 'folding funnel' (Figure 1.5). The vertical position of a structural state encodes the energy of the fold, while the width of the funnel at a certain vertical position describes the number of the states available at that energy. At the top of the funnel, proteins transition from the fully-unfolded state to the so-called 'molten globule' state, in which the protein takes on some aspects of the final structure, but otherwise remains a part of a fluctuating ensemble. As the protein descends in energy toward the native state, the number of available states becomes restricted. Near to the bottom of the funnel, local minima may exist which trap the folding structure.

Though many protein sequences—even random sequences—fold to structures (Labean *et al.,* 2011; Tretyachenko *et al.,* 2017), protein folding is a many-body interaction that involves competing and contradictory energetic pressures, which is known as frustration (Ferreiro *et al.,* 2014). Since the native state is a Gibbs free energy minimum, proteins in their native state are minimally frustrated, yet substantial levels of frustration typically remain (Clementi *et al.,* 2003; Onuchic *et al.,* 1997; Panchenko *et al.,* 1996; Bryngelson *et al.,* 1987). In fact, most proteins are 'marginally stable': the free energy of the native state is not much lower than the typical energy of nearby states. It is likely that this fact is due to a combination of selective pressures to maintain functionality and unfoldability (Williams *et al.,* 2007; Loell *et al.,* 2018; Shah *et al.,* 2018; Taverna *et al.,* 2002) and the fact that mutations which increase stability are rare and diverse (Goldstein, 2011; Hart *et al.,* 2014).

For single-domain proteins, the protein-folding process has often been observed to occur through a 'two-state' mechanism. Two-state folding of a protein posits folding directly from the unfolded to the folded conformation via a transition state, without substantially populating metastable or intermediate states. Hydrogen exchange experiments have shown that many short proteins fold in this way (Yi *et al.,* 1996; Raschke *et al.,* 1998; Englander & Mayne, 2014;

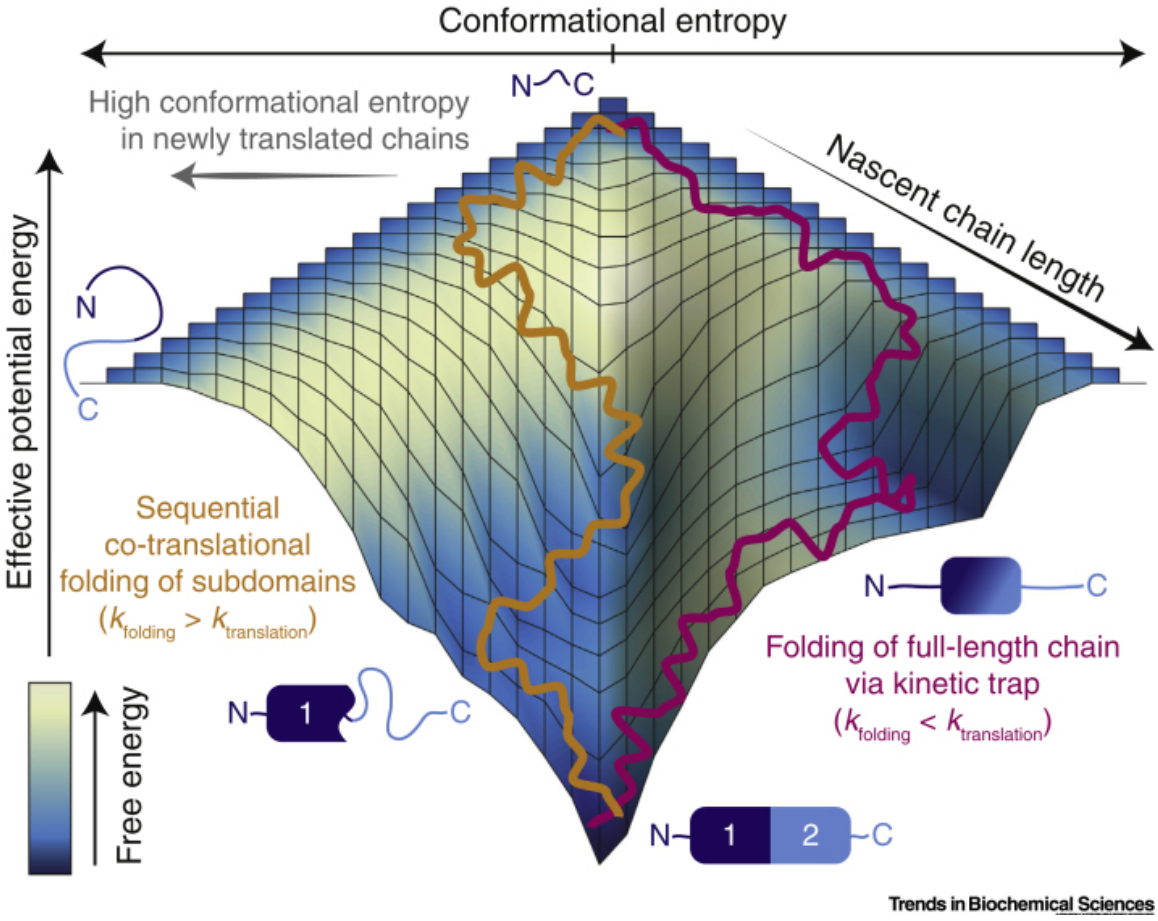


Figure 1.5: **The folding funnel in a cotranslational context.** The two trajectories, magenta and orange, represent different cotranslational folding paths for a hypothetical two-domain protein. As usual, the depth of the funnel indicates the chemical potential energy and the width of the funnel corresponds to the number of available conformations. These factors compete to produce the free energy (shading). Here, the orange trajectory folds cotranslationally and avoids the kinetic trap which the magenta trajectory finds. Both trajectories eventually terminate at the bottom of the funnel, a state in which the conformational entropy is very low and the potential energy is lower, resulting in a low free energy which maintains the protein in its native state. Figure from Waudby *et al.* (2019) with Creative Commons CC-BY license.

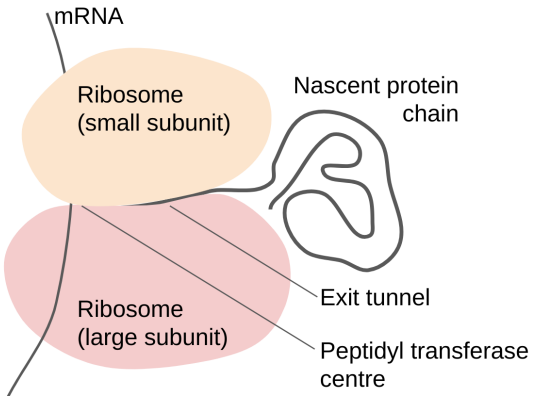


Figure 1.6: **Schematic view of cotranslational folding.** In cotranslational folding, the protein nascent chain folds as it leaves the ribosome exit tunnel. Amino acid addition takes place approximately 100 nm from the end of the ribosome exit tunnel. In some cases, it is likely that alpha helices fold inside the exit tunnel (Holtkamp *et al.,* 2015).

Englander, Mayne, *et al.,* 2016). Some larger proteins may fold through the stepwise formation of independently-folding sets of residues, or 'foldons' (Maity *et al.,* 2005; Bhardwaj *et al.,* 2008; Hu *et al.,* 2016), the folding of each of which is approximately two-state. The thermodynamic details of the folding process are described in more detail in Chapter 4.

It is clear that the cellular folding environment affects protein folding. For example, the presence of chaperones or the endoplasmic reticulum can be essential for correct protein folding for specific types of proteins (Y. E. Kim *et al.,* 2013). On a finer scale, the crowded protein environment is believed to hamper protein folding (Jefferys *et al.,* 2010). The presence of unfolded or misfolded products can also affect the production of correctly-folded products, as in the pathology of prion disease (Cobb *et al.,* 2009).

Perhaps the largest influence on folding trajectory is that of the folding machinery itself, through cotranslational protein folding. All proteins are synthesised on the ribosome—the cellular machine which catalyzes the mRNA-guided assembly of proteins from aminoacyl-tRNA—and it has a crucial role in protein structure formation. Simply: proteins fold orders of magnitude faster than they are translated, and signatures of cotranslational folding have been observed experimentally and theoretically.

Experimental and theoretical work has frequently found that proteins fold much faster than they are translated. In *E. coli*, the translation elongation rate is typically 18 amino acids per

second, falling to 8 amino acids per second in adverse conditions (Dai *et al.,* 2016), leading to the completion of translation of a protein of average length in 10-80 seconds in many bacteria (Wohlgemuth *et al.,* 2011). These rates are also typical in yeast (Riba *et al.,* 2019) and slightly higher than normal for many mouse tissues (Gerashchenko *et al.,* 2021). Yet proteins, at their fastest, can fold in microseconds (Yang *et al.,* 2003) and many proteins fold in less than a second (Naganathan *et al.,* 2005). Thus, it is not surprising that translation elongation rate, in some systems, is thought to affect folding products (O'Brien *et al.,* 2012; Jacobs *et al.,* 2017).

There is powerful experimental evidence that cotranslational folding is an important factor in the synthesis of real proteins with biological relevance. NMR studies suggest that native structures are reached by the first ten to twenty amino acids during translation of some proteins before elongation has concluded (Wright *et al.,* 1988). This observation is consistent with the fact that many single domains of multi-domain proteins fold successfully when translated independently (Batey *et al.,* 2008; Zhou *et al.,* 2019). Moreover, in many cases, proteins fold faster on the ribosome than they do in *in vitro* refolding experiments (sometimes by several orders of magnitude), or with a higher yield of correctly-folded protein. Others refold only at lower concentrations or temperatures than would normally be present *in vivo* (Kolb, 2001).

The formation of disulfide bonds provides strong evidence for the presence of cotranslational folding. Bergman and Kuehl showed that disulfide bond formation is necessarily cotranslation in murine immunoglobulin G. Disulfide bonds are formed in the endoplasmic reticulum, presenting an optimal chemical environment, and the presence of enzymes which assist disulfide formation, such as protein disulfide isomerase (Oka *et al.,* 2013). By disrupting the endoplasmic reticulum and alkylating free sulfhydryl groups, they showed that the first disulfide bond—between the first two cysteines in the protein—forms cotranslationally upon extrusion into the endoplasmic reticulum (Bergman *et al.,* 1979b). Moreover, murine immunoglobulin G does not form entirely correct disulfide bonds on *in vitro* refolding and refolds more slowly (Bergman *et al.,* 1979a). It also binds its dimeric partner cotranslationally (Bergman *et al.,* 1979a) and is glycosylated in a cotranslational fashion (Bergman *et al.,* 1979c). Rat serum albumin was shown, in a similar experiment, to close disulfide bonds in the amino-to-carbonyl direction during protein synthesis (Peters *et al.,* 1982). Other proteins, such as the low-density lipoprotein receptor (Kadokura

*et al.,* 2020) and the HIV *env* protein (Land *et al.,* 2003), exhibit PDI-mediated cotranslational disulfide formation as well.

These observations are borne out by experiments into protein function. In the case of firefly luciferase, *in vitro* refolding delivers poor yields and a propensity to aggregation. Yet proteinase experiments demonstrate that it folds progressively and cotranslationally from ribosomes (Frydman *et al.,* 1999). And, likewise, though it is not functional on ribosomes, it gains activity only seconds after release from the ribosome. *In vitro* experiments show that the refolding time of luciferase is approximately 14 minutes, so much of the structure must have been formed cotranslationally (Kolb *et al.,* 1994). When its stop codon is removed and a 26-residue extension is added, firefly luciferase folds into its active state while attached to the ribosome (Makeyev *et al.,* 1996). (The 12 C-terminal residues are crucial to the enzymatic activity of luciferase.) Bovine liver rhodanese also becomes active in a similar experimental set-up (Kudlicki *et al.,* 1995).

Several other proteins are known to fold cotranslationally. Certain proteins, such as chlorophyll *a* apoproteins CP43 and CP47 (Mullet *et al.,* 1990), and $\beta$-hemoglobin (Komar *et al.,* 1997), bind cofactors cotranslationally, while *E. coli* dihydrofolate reductase has higher specific activity when translated in the presence of dihydrofolate, suggesting that cotranslational presence of its ligand promotes correct folding (Mouat, 2000). Others oligomerize cotranslationally: both immunoglobulin (Bergman *et al.,* 1979a) and the collagen triple helix (Veis, Leibovich, *et al.,* 1985; Veis & Kirk, 1989) are examples of cases in which nascent chains bind cotranslationally to fully-synthesised binding partners or nascent chains.

Recent studies have enabled us to examine the process of cotranslational folding for a limited number of proteins at molecular or atomic detail (Liutkute *et al.,* 2020). Nilsson *et al.* (2015) found that a small (29-residue) zinc-finger domain folds inside the ribosme exit tunnel, and the energy released through folding is stored partially as increased tension on the translating nascent chain (Nilsson, Hedman, *et al.,* 2015). Likewise, by introducing a short section with a strong propensity to form $\alpha$-helices into dipeptidylaminopeptidase B, Bushan *et al.* (2010) were able to visualise cotranslational helix formation within the ribosome exit tunnel using cryo-electron microscopy. However, experiments with VemP suggest that the role of the inner part of the exit tunnel may be to inhibit the formation of certain tertiary structures in order to prevent ribosome

stalling (Eichmann *et al.,* 2010; Su *et al.,* 2017). The small globular protein HemK (Mercier *et al.,* 2018; Holtkamp *et al.,* 2015; Kemp *et al.,* 2019) and the spectrin domain (Nilsson, Nickson, *et al.,* 2017; Scott *et al.,* 2004), as well as a $\beta$-helix repeat protein (Notari *et al.,* 2018) have been used as test systems in which cotranslational folding can be followed at a molecular level. Although substantial engineering has been used to develop model systems which are stable and repeatable enough to allow force measurements or cryo-electron microscopy, it is likely that these proteins are representative of wider trends.

There is also evidence that codon-use patterns are reflective of regulation of protein folding, which would occur if the proteins fold cotranslationally (O'Brien *et al.,* 2012; Pechmann *et al.,* 2012; Nissley, Sharma, *et al.,* 2016). These patterns of codon use derive from the fact that some codons for the same amino acid lead to much faster translation than others (C.-H. Yu *et al.,* 2015), for example, by using amino acids that are cognate with tRNAs that are more or less common in the cell, leading to faster or slower translation, respectively. These patterns appear to be conserved (Jacobs *et al.,* 2017; Chaney, Steele, *et al.,* 2017; Nissley, Carbery, *et al.,* 2021) and frequently appear in clusters, even in highly-expressed sequences (Chaney & Clark, 2015). If proteins folded only after extrusion, regulation of extrusion speed could have no impact on the folding process. Likewise, it is observed experimenally that cross-species protein production is often inefficient when the expression host (often *E. coli*) has a different codon-usage pattern than the native species, i.e., the most common codons in the native species do not match those in the expression host. Aligning codon-usage patterns by making synonymous mutations to the DNA sequences often increases production efficiency, sometimes dramatically (Punde *et al.,* 2019). Codon harmonization is a well-established experimental tool (Mignon *et al.,* 2018; Asam *et al.,* 2018; Van Aalst *et al.,* 2020).

These pieces of evidence are complemented by simulation and observational studies of co-translational folding. Lattice models have provided evidence for the signatures of cotranslational folding and suggested that cotranslational folding may play a role in the stability of lattice models, although the evidence from these studies is mixed (Mann *et al.,* 2008; Lu *et al.,* 2007; Saunders *et al.,* 2011; Morrissey *et al.,* 2004). Analyses of protein structure data have suggested that there is statistical evidence of cotranslational folding in protein structures, especially $\alpha/\beta$

proteins (C. M. Deane *et al.,* 2007; Saunders *et al.,* 2011). Specifically, N-terminal regions are more buried on average, and many $\alpha/\beta$ proteins have an N-terminal set of core residues. Moreover, Srivastava et al. (2011) found that the the N-terminus tends to be more hydrophobic than the C-terminus of proteins. This observation seems sensible in a cotranslational model: more hydrophobic N-terminal regions would tend to be packed in the interior of globular proteins, where they would be more protected from water.

Cotranslational folding, thus, encompasses a large number of effects stemming from physical, biochemical, evolutionary, and other origins. It is likely that cotranslational folding occurs throughout bacterial and eukaryotic proteomes. That cotranslational folding is so widespread suggests it has a role in preventing accumulation of trapped intermediates and helping to select the nascent chain's path through configuration space. Taking account of these effects may assist physical approaches to protein structure prediction, particularly if protein native states are metastable, as has been demonstrated for some proteins (Thoden *et al.,* 1997; Sohl *et al.,* 1998).

## 1.3 Protein structure prediction

The last ten years have seen impressive advances in the accuracy of protein structure prediction, resulting in highly-sophisticated protein structure prediction algorithms. Fragment-based modelling—in which parts of known structures are concatenated to construct a three-dimensional model of the target protein—and increases in our ability to predict protein contacts (Abriata *et al.,* 2018), have each led to step changes in protein structure prediction accuracy. Assessment of these changes has been driven by the biennial Critical Assessment of Protein Structure (CASP), which brings the field together to benchmark performance on standardized blind structure-prediction problems (Moult *et al.,* 2018).

In 2020, AlphaFold2, a machine-learning program developed by DeepMind for protein structure prediction, achieved experimental accuracy for many *de novo* prediction targets, and claimed that their methodology had been used to correct at least one error in X-ray crystallographic protein-structure determination (DeepMind, 2020). This advance, for the first time, may permit reliable protein structure prediction as a tool rather than an object of scientific inquiry. However, the predictions from AlphaFold2 have yet to be validated systematically outside of the limited

CASP test set, and neither a large number of AlphaFold2's predictions nor the software itself are available for use by the structural biology community.

**Types of prediction** CASP defines two types of structure prediction (Moult *et al.,* 2018). Template-based modelling (TBM) consists of modelling where at least one template with known structure exists. This template is usually identified on the basis of sequence similarity and when more than one template exists, they are combined, extended, and refined to construct a model of the target molecule. Identification of distant homologs, i.e., sequences which are informative for the structure prediction task but share little sequence similarity with the target, is a central feature of cutting-edge prediction algorithms. (The sequences do not need to share a common evolutionary lineage, and such sequences are properly termed orthologs.) The second type of prediction is free modelling (FM), where no full length homologs with known structure can be identified. FM predictions are typically worse than TBM for the same targets.

**Assessment of predictions** The simplest metric for assessment of protein structure prediction is root-mean-square deviation (RMSD), which is the square root of a normalised sum of squared devations in atom positions between two structures. This methods suffers from a number of disadvantages, among which are the strong dependence of the measure on protein length, and the difficulty of interpreting the metric when part of the target structure has a large difference from the molecule with which it is compared.

To overcome these issues, several alternative scoring schemes have been developed. In CASP, protein structure predictions are scored with the GDT_TS metric (Zemla *et al.,* 2003). To construct GDT_TS, the target structure is aligned to a modelled structure and the number $n_\delta$ of modelled $C_\alpha$ atoms within $\delta$ Ångstrom is recorded for $\delta = 1\text{Å}, 2\text{Å}, 4\text{Å}$ and 8Å. GDT_TS is the average of these values:

$$\text{GDT\_TS} = \frac{n_1 + n_2 + n_4 + n_8}{4N}, \tag{1.1}$$

where $N$ is the number of atoms in the target structure. By binning accuracy thresholds, the metric limits the contribution that regions of very poor prediction can offer (Li *et al.,*

2016), while its basis as an average of proportions of atoms limits the effect of molecule size on the overall score.

Zhang and Skolnick (Zhang *et al.,* 2004) suggest TM-Score as an improvement on GDT_TS. In the case where both protein chains have the same length, properties, providing the basis of structural diversity and stability.

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

Here, $d_i$ is the distance between residue each of the corresponding residues $i$, while $d_0$ is a monotonically increasing normalisation of the length scale:

$$d_0(L) = 1.24 \sqrt[3]{L - 15} - 1.8$$

Like GDT_TS, it prevents large deviations from causing extremely large changes in the score, here by putting the deviations in atomic position into the denominator of the fraction in Equation 1.2. Since the fraction $\dfrac{1}{1 + \dfrac{d_i}{d_0(L)}}$ can be at most 1 (if $d_i/d_0$ is small) and no less than 0 (if $d_i/d_0$ is large), the TM-Score is bounded. The normalisation $d_0$ empirically minimises the length-dependence of the measure, enabling TM-Score to take into account all deviations between corresponding residues and enabling the use of a single measure for proteins of different lengths. A commonly-used criterion for the assessment of protein folding prediction accuracy is TM-Score $> 0.5$, where models with TM-Score above 0.5 are believed to have the same fold as the target (de Oliveira, 2015; Xu *et al.,* 2010).

As shown in Fig. 1.7, CASP results have been improving over time. In recent years, this effect has been due largely to improvements in contact prediction. A close relationship between contact prediction precision to GDT_TS for the best models in CASP12 is observed (Moult *et al.,* 2018), which is evidence that contact prediction is important for structure prediction. It has been suggested (D. E. Kim *et al.,* 2014) that one contact for every twelve residues is sufficient for structure prediction, and conversely, a lack of contact information is strongly detrimental for the quality of structure prediction (Moult *et al.,* 2018).

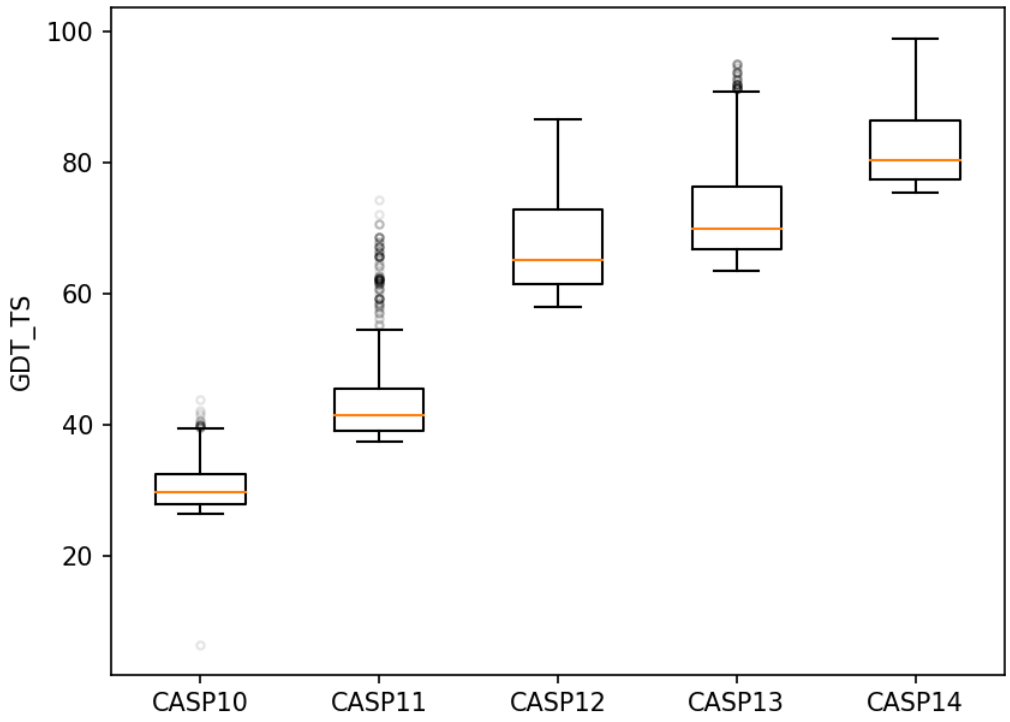


Figure 1.7: **CASP results in the free modelling category for the last five assessments (2012-2020).** GDT_TS is plotted for the top 1200 models for all FM targets. The jumps in 2016 (CASP12) and 2020 (CASP14) reflect the widespread adoption of contact prediction and machine learning, respectively. These plots are standard box-and-whisker plots: orange lines are medians, boxes indicate the range between the first and third quartile, and dots are outliers using the 1.5-interquartile range criterion. Upper and lower lines indicate the non-outlier range.

**Contact prediction** Contact prediction has been revolutionised in the last ten years by improvements in coevolution methods. These methods work on the principle that nearby residues tend to interact, and these interactions tend to be more preserved over evolution than the identities of individual residues. Thus, changes which are preserved co-occur in both residues of a contact, and this information can be interpreted as correlated changes between sites in large alignments of protein sequences.

The principle development that increased the accuracy of inter-residue contact prediction was the development of maximum-entropy models which enabled the deconvolution of transitive interactions from direct interactions (Stein *et al.,* 2015). The correlations between two pairs of amino acids $A, B$ and $B, C$ will result in an apparent correlation between $A$ and $C$, even though this correlation is only an artefact of the their mutual correlation with $B$. This inferred signal does not represent a contact and earlier methods, such as mutual information, or indeed any local pairwise measure of information (Marks *et al.,* 2011), were unable to differentiate between the correlation of $A$ and $B$ and $A$ and $C$.

There are a family of related methods that overcome these problems (Stein *et al.,* 2015). Known as direct information (Marks *et al.,* 2011) or direct coupling analysis (Morcos, Pagnani, *et al.,* 2011; Morcos, Schafer, *et al.,* 2014), these methods seek to fit a regularized version of the probability distribution

$$P(A_1, ..., A_L) = \frac{1}{Z} \exp\left( \sum_{1 \leq i \leq j \leq L} e_{ij}\left(A_i, A_j\right) + \sum_{1 \leq i \leq L} h_i\left(L_i\right) \right)$$

for particular amino acids $A_i, A_j$ at positions $i, j$ in the multiple sequence alignment. When summed over all possible amino acids, the couplings $e_{ij}$ and $h_i$ represent energies that indicate inter-residue coupling strengths. This model has been reimplemented several times, including mean-field implementations of the direct information model (Morcos, Pagnani, *et al.,* 2011) (which run faster, at the expense of accuracy) and efficient implementations of pseudo-likelihood methods (Marks *et al.,* 2011; Seemayer *et al.,* 2014).

The last decade has also seen the advent of contact meta-predictors and deep learning for contact prediction. MetaPSICOV achieved better predictions of contacts than CCMPred

(a DCA method) (Seemayer *et al.,* 2014) and PSICOV by combining predictions from each method using a two-stage feed-forward neural network (Jones, Singh, *et al.,* 2015). The highest levels of accuracy available at present are obtained by deep learning methods. Wang et al. (2017) have used an ultra-deep-learning method with two networks predicting pairwise and structural features simultaneously. MetaPSICOV has been recently expanded to include a more sophisticated deep-learning architecture (Buchan *et al.,* 2018). These methods take into account advances in image processing to determine contacts at higher levels of completeness and accuracy than before.

**Quality of contact prediction** Contact prediction accuracy can be assessed using a range of measures suitable for the assessment of prediction of binary outcomes, including ROC curves, the Matthews Correlation Coefficient, and others. The simplest and most widespread is top-$N$ positive predictive value (PPV):

$$\mathrm{PPV}_N = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}},$$

where the statistic takes into account the top-ranked $N$ predicted contacts. A common choice for $N$ is $L$, the length of the protein, which is commonly expressed as a percentage. In Chapter 3, we show typical levels of contact prediction accuracy for different methods and discuss the reasons for these differences.

**Fragment assembly** In free modelling, and, to a lesser extent, template-based modelling, an essential question is how to generate novel yet realistic protein structures, which has proven challenging for otherwise-promising methods (Alquraishi, 2019). One of the most common solutions to this problem is fragment assembly, in which a large collection of short peptide fragments from experimentally-determined structures, typically three to twenty amino acids long (Trevizani *et al.,* 2017), are substituted into a structural model. This process guarantees that each individual residue in the sequence will adopt a conformation that has been found in nature. The fragment, thus, is a set of torsion angles or Cartesian coordinates that defines the structure of a set of residues in the modelled structure (Trevizani *et al.,* 2017). Often, these fragments are chosen to match known or predicted properties of the

peptide, such as secondary or even tertiary structure.

In order to make a prediction, an initial structure is modified through an iterative process, in which a fragment is randomly chosen from the fragment library to replace the configuration of part of the initial structure. Then, the new structure, including the fragment substitution, may be assessed against the initial structure, typically *via* a Monte Carlo criterion, though some methods use simulated annealing. Implicit here is the use of a scoring function, which may be based on physical or knowledge-based potentials, or even machine learning (Senior *et al.,* 2020). A prediction will normally be based on running this procedure multiple times.

**Fragment libraries** The composition of fragment-based libraries is an area of active research because of the attractive properties of fragment-derived structures. Fragment libraries seek to balance fragment diversity with accessibility of native-like states, taking into account the fact that structure prediction methods will be able to make less effective use of larger fragment libraries because of the combinatorial size of the search space, which depends on the number of fragments.

Fragment library quality depends on the number of good fragments at a site. "Good" fragments are those which have less than 1.5 ÅRMSD with the native structure (de Oliveira, Shi, *et al.,* 2015). Fragment precision is the proportion which are good quality, and fragment coverage is the proportion of sites which have good-quality fragments. Fragments are often restricted to match the secondary structure of their targets, where available or predicted. Loop fragments are the most challenging fragments from a library-generation perspective, reflecting the fact that their structures are the least constrained and the hardest to predict (de Oliveira, Shi, *et al.,* 2015). Therefore, loop fragments are the least likely to reflect the native structure of the protein.

**Scoring functions** Fragment substitution methods generate large numbers of putative structures, both during fragment substitution and because the algorithm is typically run hundreds of times in order to sample the conformation space widely. In order to choose between these structures, a heuristic or approximate method of ranking structures is required: a full solution of the Schrödinger equation is computationally prohibitive using current methods.

Although a few cutting-edge methods involve deep learning (Senior *et al.,* 2020; Billings *et al.,* 2019; Gao *et al.,* 2020), more common is the use of a linear combination of physical or knowledge-based potentials. Knowledge-based potentials, such as the Residue-specific All-atom Conditional Probability Discriminatory Function (RAPDF) (Samudrala *et al.,* 1998), exploit the statistical properties of known protein structures. In the case of the RAPDF, the radial probability distribution function of each pair of amino acid types is computed for a set of known structures and used to score models. Physical potentials, by contrast, use properties which approximate physical effects, such as the Lennard-Jones 12-6 potential. In this work, we use the standard set of SAINT2 potentials, which are described in more detail in Chapter 2.

While recent developments have pushed back the boundaries of speed and especially accuracy, protein structure prediction remains a challenging problem of central importance to biological inquiry. For example, Ovchinnikov et al. (2015) identified 121 large protein families in prokaryotes for which no structures were available. They proposed 58 structures, corresponding to more than 400,000 individual proteins. There are likely to be significant differences in structure between the 58 exemplars and individual members of those families, irrespective of any modelling error, and a further 43 protein families still have no known structural exemplar. Membrane protein structure prediction suffers from a lack of benchmark 3D structures, despite progress in the field, which makes comparative modelling difficult (Hopf *et al.,* 2012). The effect of the availability of new protein structure prediction methods will become more clear as structure predictions for these and other proteins become widely available.

## 1.4 Protein folding

The description of protein folding—the process by which proteins transform from an unordered state into their final structure—is related to protein structure prediction but a different challenge. To understand protein folding, it is not sufficient to determine the tertiary structure of the protein, because protein folding requires knowledge of the history of conformational states. Conversely, accurate structure prediction does not demand knowledge of the folding process, and indeed AlphaFold2 is not known to provide a physical folding model. Software at the previous

cutting edge of structure prediction—the only software which is publicly-available (Moult *et al.,* 2018; Ovchinnikov, Kinch, *et al.,* 2015; Ovchinnikov, Park, D. E. Kim, *et al.,* 2016; Taylor *et al.,* 2012)—similarly does not assume any physical model of folding, but instead exploits statistical and physical potentials alongside information from multiple sequence alignments.

Moreover, folding is a special case of protein dynamics, and existing methods for understanding dynamics are incomplete. The most widely-used tool, molecular dynamics (MD), simulates protein structures through an explicit physical representation (which may be at the level of atoms, or which may 'coarse-grain' atoms into larger physical units) coupled to a representation of the forces which govern the dynamics of the protein. MD has become a useful tool for understanding motion in proteins, including for the development of small-molecule pharmaceuticals (Laurin *et al.,* 2020; Jennings *et al.,* 2018) and for relaxing proteins to physically-plausible structures (Park *et al.,* 2018), among many other applications.

Protein folding studies have made extensive use of molecular dynamics. Atomistic simulations have been able to reveal folding pathways for a collection of small proteins *in vitro* (Lindorff-Larsen *et al.,* 2011; Piana *et al.,* 2013). For larger proteins, and to understand the process of protein synthesis (which requires simulations over longer timescales) coarse-graining approaches have been used (Nilsson, Hedman, *et al.,* 2015; Trovato *et al.,* 2017), in which parts of the protein are abstracted to simpler structures. However, MD typically requires some knowledge of the native protein structure—though Shaw *et al.* (2010) is an exception, for a small protein domain—and developing an MD simulation which is physically realistic requires time and expertise. MD is also computationally intensive, especially when run to timescales relevant to folding or with large proteins at full atomic detail.

Understanding dynamics is especially important for certain classes of proteins, such as motor proteins, ion channels, and other proteins with large mobile components, as well as proteins for which allosteric changes are fundamental to their function. Knowing more about flexibility in protein loop regions, as well as the physical treatment of regions that exhibit large-scale motions, will require a more detailed physical understanding of the determinants of protein structures.

## 1.5 Structure of thesis

In the following chapters, we investigate the protein-folding problem by considering three complimentary lines of investigation. First, we implement a model of the ribosome in SAINT2 (de Oliveira, 2015) to assess whether the physical structure of the ribosome has an observable impact on the protein folding pathway of a collection of proteins. Then, in Chapter 3, we investigate the relationship of evolutionary constraints in protein structure to observable structural features, and use this information to evaluate the differences in certain methods used for structure prediction. This work is based on my paper "The evolution of contact prediction: evidence that contact selection in statistical contact prediction is changing" (Bioinformatics 36, March 2020) with Saulo de Oliveira, Konrad Krawczyk, and Charlotte Deane. Finally, we explore co-translational folding in the *E. coli* proteome, investigating the extent to which these effects can be predicted by SAINT2 and exploring other methods for this analysis.

I also include, as an appendix, "Ribosome occupancy profiles are conserved between structurally and evolutionarily related yeast domains" (Bioinformatics, January 2021) by Daniel Nissley, Anna Carbery, Mark Chonofsky, and Charlotte Deane, another piece of work which I was involved in, which demonstrates conservation of translation rate profiles between related genes in yeast cells, showing that on average, those sections of yeast genes that are translated particularly slowly or quickly tend to be preserved during evolution.

# Chapter 2

# Modelling the effects of the ribosome with SAINT2

Although a frequent metonym for protein structure prediction, protein folding properly refers to the process by which a protein attains its structure and not the task of determining the structure itself. Protein folding is difficult to study because it concerns a dynamic process on the nanometer scale, a process which is inherently stochastic and which is affected by thermodynamic noise.

As described in the Introduction, the process of natural protein folding takes place in a cellular environment, in which numerous competing physical effects influence the structure of the folded protein. However, since all proteins are extruded on the ribosome, it is likely that proteins have evolved in response to the ribosomal environment. We hypothesised that if the effect of the ribosome could be added to physical models of protein structure prediction, then these effects might lead modelled amino-acid chains to form native-like structures more consistently and to more closely match known native structures. In order to capture the true effect of the ribosome on protein structure formation in protein structure prediction pipelines, it is necessary to model protein folding at a level of detail which reflects the true effect of the ribosome. It is challenging to know which level of detail is required to capitalize on these effects in a computational context.

A recent approach for elucidating the relevance of cotranslational folding has been to assess its utility in protein structure prediction. Ellis et al. (2010) found that protein structure prediction is more effective when a protein is extruded from the N- to C-terminus, rather than in the reverse direction. In parallel, Chwastyk et al. (2015) have suggested a role for cotranslational folding

in knotted proteins, and a cotranslational approach has been used to predict structures in the UniCon3D pipeline (Bhattacharya *et al.,* 2016). Perhaps the most compelling evidence is due to de Oliveira et al. (2018, 2015) who showed that a cotranslational approach improves protein structure prediction.

SAINT2 is a template-free protein structure prediction method which models the structure of proteins through an explicit cotranslational approach. It was developed by the Deane group and used by de Oliveira et al. (2017) in their investigation of cotranslational protein structure prediction. SAINT2 uses a Markov chain Monte Carlo sampling strategy in order to move between protein structures. Specifically, it uses a fragment-based approach, in which the protein structure is varied through the substitution of fragments from known crystal structures. Fragment libraries, from which these fragments are drawn, are generated by the software Flib (de Oliveira, Shi, *et al.,* 2015). A full description of the method is given in de Oliveira, 2015.

Due to the success of simple cotranslational folding methods, it is reasonable to suspect that building other factors that are relevant to protein folding into protein structure prediction would also improve results. One such effect is that of the topological environment of the ribosome. Biophysical experiments show that some proteins part-fold inside the ribosome exit tunnel (Thommen *et al.,* 2017). Moreover, the effect of the ribosome is to occupy large regions of the space in which the protein could fold, so we would expect that it would have some effect on the folding process (Jefferys *et al.,* 2010). Similarly, proteins are known to fold at variable rates over the course of translation, and a large body of experimental evidence suggests that obtaining the correct rate of translation is important for physical protein production. This effect could also be assessed using computational methods.

In this chapter, we investigate that hypothesis by implementing three models of the ribosome in a cotranslational protein structure prediction pipeline, SAINT2. We introduced these models—analogues of the ribosome wall, the exit tunnel, and the exit tunnel with the ribosome wall—as spatial constraints within SAINT2's fragment replacement and extension protocol. We assessed their effect on protein structure prediction accuracy and found that none of these models led to higher levels of accuracy. In order to understand why no improvement was seen, we investigated the nature of SAINT2 protein folding pathways in terms of the SAINT2 energy

function. This analysis revealed that our new structural constraints we had added led to clashes between residues in predicted structures. Therefore, we implemented a new extension protocol which decreased the probability of these clashes and found that the use of this protocol improved protein structure predictions.

## 2.1 Methods

### 2.1.1 Protein dataset

We used a collection of 245 proteins which represent a diversity of structural, functional, and organismal origins (priv. comm., De Oliveira, S. H. P., 2017). As shown in Fig. 2.1, this collection of protein sequences ranges between 50 and 250 amino acids and spans the four most common SCOP fold categories (a-d) in approximately equal proportion. A diversity of folds is important because some types of folds are more difficult to predict than others (Crivelli *et al.,* 2002). We restrict the length of the protein chain because long proteins are particularly challenging for protein structure prediction.

We used Flib (de Oliveira, Shi, *et al.,* 2015) to generate fragment libraries for these proteins. Fig. 2.1 also demonstrates that contact prediction accuracy and secondary structure prediction accuracy are within acceptable ranges for protein structure prediction (de Oliveira, 2015). Contact predictions are generated by PSICOV (Jones, Buchan, *et al.,* 2012), which is a coevolution-based protein contact prediction method, and the median positive predictive value for these predictions is approximately 0.8. Secondary structure prediction, here using PSIPRED (McGuffin *et al.,* 2000), seeks to classify the sites in each protein in terms of the eight-state DSSP secondary structure classification (Kabsch *et al.,* 1983). With a median accuracy above 60% and nearly all cases above 40%, this accuracy is at the level needed to accurately predict protein structures using SAINT2. Taken together, these quality checks indicate that these proteins are reasonable structure prediction targets.

We assembled alignments and fragment libraries and we predicted contacts for these proteins using a previously-published protocol (Law *et al.,* 2017; de Oliveira, 2015). We used predicted contacts from the output of the MetaPSICOV stage 1 network, on the basis of previous results suggesting that protein structure predictions made on the basis of stage 1 predictions were more

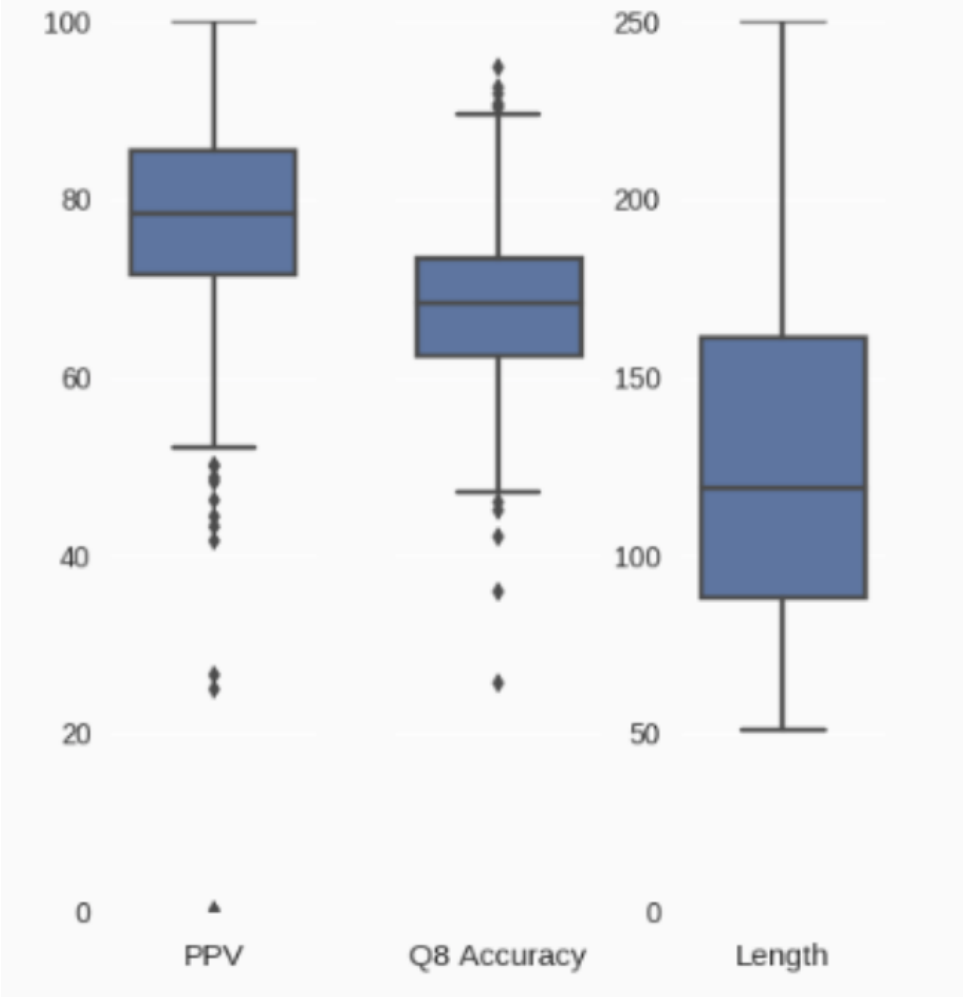


Figure 2.1: **Key characteristics of the inputs to SAINT2 protein structure prediction for the 245 test proteins.** From left, (1) contact prediction accuracy (positive predictive value); (2) Q8 accuracy, indicating the extent to which protein secondary structure prediction was successful; and (3) the distribution of protein lengths. Q8 accuracy is the proportion of residues with correct secondary structure predictions in an eight-state model of secondary structure.

accurate than those made on the basis of MetaPSICOV stage 2 predictions (priv. comm., de Oliveira, S. H. P., 2017).

### 2.1.2 Models

We implemented three models of the ribosome's influence in SAINT2 (Fig. 2.2) by identifying possible topologies might represent the excluded-volume effect of the ribosome on protein folding. First, we considered an infinite half-plane, which corresponds to the exclusion of the nascent chain from folding through the surface of the much-larger ribosome. (We refer to this model as the 'ribosome wall'.) We also implemented a model of the extrusion tunnel, which constrained the nascent chain to lie within 8 Å of the tunnel axis for 100 Å. We took these measurements from representative values for a variety of organisms (Schuwirth *et al.,* 2005; Yusupova *et al.,* 2014; Fedyukina *et al.,* 2011). Our model did not implement the bend in the tunnel that occurs approximately 40 Å from the end of the tunnel in many organisms (Fedyukina *et al.,* 2011). We assessed the wall individually and the two constraints together, in addition to comparing them to cotranslational folding without restraints. The third model was a control containing only the tunnel condition.

These models were implemented as simple geometrical constraints. When fragment replacement caused the resulting protein structure to intersect the ribosome wall, the move was assigned infinite energy and consequently rejected.

Initially, we observed that the modelled protein would not diffuse out of the extrusion tunnel, resulting in unsuccessful folding. Since we are not not able to move the nascent polypeptide chain out of the tunnel during or after extension, this effect was an artefact of our computational model. Therefore, during runs of 11,000 Monte Carlo steps, we uniformly shortened the length of the tunnel from 100 Å to 0 Å during steps 4,000 to 8,000. We additionally removed all constraints for the final 1,000 Monte Carlo steps because we expect some *in vivo* folding to occur after release from the ribosome.

### 2.1.3 SAINT2

The energy function from which the Markov chain samples is a linear combination of five different functions.

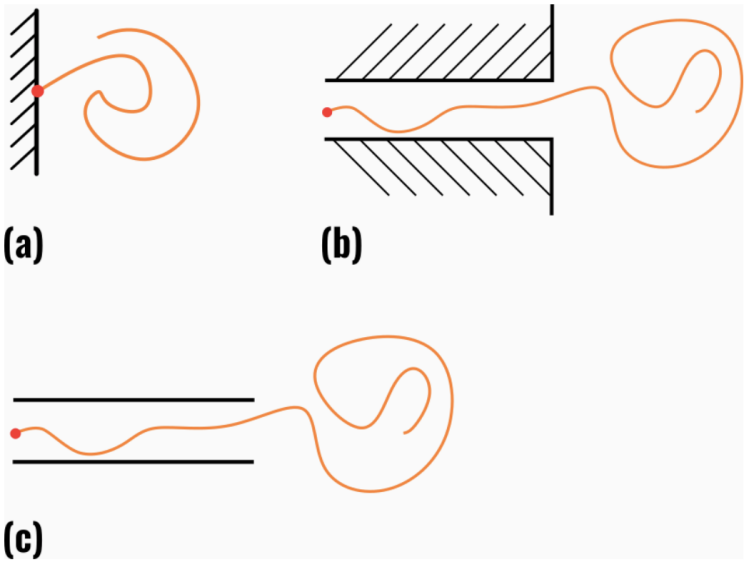


Figure 2.2: **Cross-sectional schematic diagrams of our ribosome models. Hatched areas indicate areas of exclusion.** The red dot indicates the point of extrusion. The orange line is a nascent chain. (a) The ribosome wall implemented as an infinite half-plane. (b) Protein extrusion through the ribosome tunnel in the presence of the ribosome wall (shown in cross-section), such that the protein nascent chain cannot intersect the tunnel or the wall. (c) Protein extrusion through the ribosome tunnel without volume exclusion representing the ribosome wall (shown again in cross-section), which was used as a control condition to assess the effect of the tunnel alone.

**Lennard-Jones** The Lennard-Jones potential describes interatomic interactions between backbone residues by means of a potential proportional to $(a/r)^{12}-(a/r)^{6}$. Here, $a$ is a constant that depends on the atoms involved in the interaction, and $r$ is the geometric distance between them. This potential favours backbone atoms which lie close to each other, but strongly penalises backbone atoms which clash with each other.

**Solvation** For each atom, we count the number of other atoms within a fixed distance of it. This distance is called the solvation radius. We normalise this atom count by the square root of the sequence length. The energetic contribution of the solvation score is linearly proportional to this count and favours more compact structures.

**RAPDF** The Residue-specific All-atom Probability Discriminatory Function (Samudrala *et al.,* 1998) is a radial potential which penalises structural configurations in which empirically-uncommon atom-residue configurations occur. In particular, the RAPDF uses the distribution of distances between pairs of atoms in a library of reference proteins to probabilistically score the distribution of distances between pairs of atoms in a structure model. The RAPDF uses individual probability distribuations for each pair of atom types. (An atom type is the combination of atom and residue label: "cysteine $C_\alpha$", for example, would be one atom type.)

**Orientation** The orientation score is an empirical score which compares the distribution of side chain angles of residue pairs to a reference distribution. The orientation score takes into account the distances between the residues by using a set of distance bins with different torsion angle distributions.

**Contact potential** For every pair of residues defined by the user of SAINT2, typically corresponding to the list of the top $L$ predicted contacts from MetaPSICOV Jones, Singh, *et al.,* 2015; de Oliveira, Shi, *et al.,* 2017, the SAINT2 contact potential adds a penalty proportional to the square root of the distance beyond 8 Å that those residues lie apart from each other.

In the default cotranslational mode, SAINT2 begins by sampling conformations for the first nine amino acids in the protein structure. It then elongates the structure by adding each

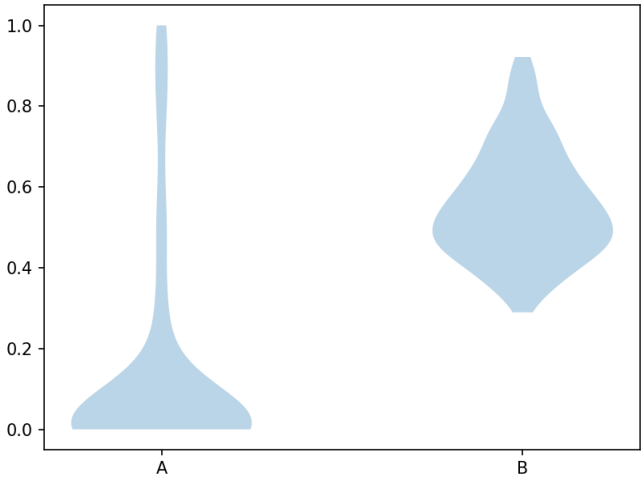


Figure 2.3: **Cotranslational protein structure prediction for 245 proteins.** As described in the text, 500 decoys were run for 11000 Monte Carlo steps, of which 10000 steps used the standard protein extension procedure and 1000 steps were completed after full protein extension. **A:** Proportion of decoys above $TM \geq 0.5$. **B:** Maximum TM-Score among the 500 decoys.

subsequent amino acid singly. We run SAINT2 for 10,000 steps, with the number of steps for each partial structure being proportional to the number of amino acids in that structure, and for 1,000 further steps once the chain is full extended. We run SAINT2 500 times for each target.

## 2.2 Results

For our test set of 245 proteins, Fig. 2.3 shows two measures of folding accuracy. Of the 245 proteins, for 145 we found one or more model out of the 500 generated that had $TM \geq 0.5$.. The median maximum TM-Score for these proteins was 0.54. Both of these statistics indicate that we are able to predict protein structures with a level of accuracy that is suitable for further investigations. The TM-Score $\geq 0.5$ threshold indicates a correct fold (Xu *et al.,* 2010). Although the maximum TM-Score is more reflective of our ability to succeed in a blind protein structure prediction setting than the proportion with TM-Score above 0.5, it has higher variability, especially at low sample sizes. However, the true protein structure would not be known in a real prediction campaign, so we would require a computational technique to pick out the best model from this larger collection of models.

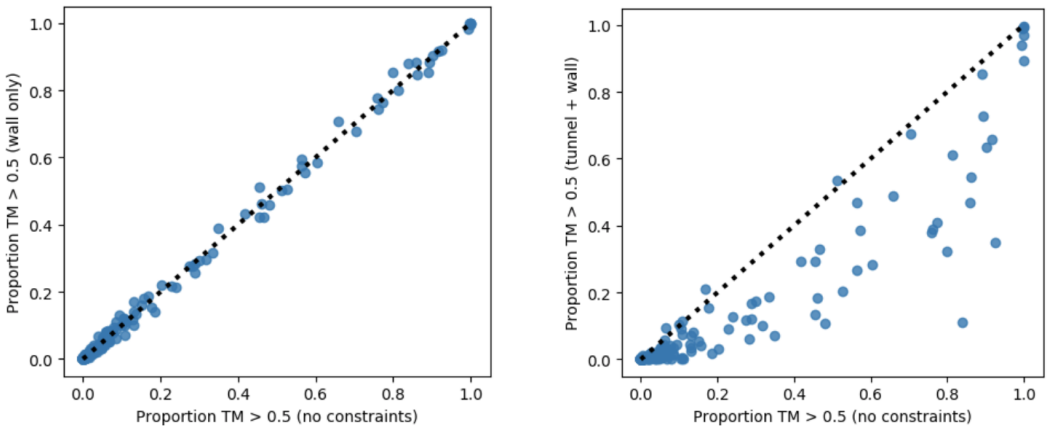


Figure 2.4: **Comparison of ribosome models in SAINT2 to control conditions.** Each point is one structure prediction target, and the value plotted is the proportion of models with TM-Score above 0.5. Left: wall folding accuracy as a function of unconstrained folding accuracy. Right: tunnel and wall folding accuracy as a function of unconstrained folding accuracy. Since the tunnel and wall condition did not deliver positive results, the tunnel only condition—which delivers slightly better results than tunnel with wall, and which was included as a control for it—is not shown or discussed here.

Fig. 2.4 displays an initial set of comparisons between folding simulations in three conditions: folding constrained by the ribosome wall only, folding constrained by the tunnel and wall, and folding in the absence of topological constraints. As above, we use 500 models for each of 245 targets. It shows that the wall condition makes apparently no difference to the quality of predicted models. Moreover, folding with the tunnel constraint leads to markedly worse structural predictions than standard cotranslational prediction. This effect may have multiple causes. First, the boundaries of the tunnel are only repulsive, in contrast to physical observations that suggest some role for adhesion onto the tunnel during folding (Thommen *et al.,* 2017). We also observed the nascent chain making physically-unrealistic hairpin turns in the tunnel. These were possible because of the presence of hairpin turn fragments in the fragment library. In a physical context, they would have required large lengths of protein chain to thread through the tunnel in a reverse direction. Threading of this kind is physically unrealistic (Chwastyk *et al.,* 2015). These effects are exacerbated by the rigidity of our fragment library, which excluded many fragments from use in the tunnel because they caused collisions with the tunnel wall. For these reasons, we chose not to progress the tunnel condition for further investigation.

## 2.3 SAINT2 energetics and folding pathways

In the course of our investigation into folding in SAINT2, we observed that some models displayed SAINT2 energies several orders of magnitude higher than the mean. This phenomenon occurred particularly when we retained structural constraints for all 11,000 Monte Carlo steps. Further investigation revealed that these high energies were due to the Lennard-Jones term in the SAINT2 energy.

We identified models for which this was a particular problem. For 1IXN_A, there are distinct maxima in the the Lennard-Jones energy (Fig. 2.5) which correspond to structural clashes (Fig. 2.6). In comparison to the solvation energy, which is of the same order of magnitude as the other contributions to the SAINT2 energy, the Lennard-Jones energy is extremely large. Since the Lennard-Jones energy depends on the inverse of $r$, the pairwise radius of separation between atoms, the Lennard-Jones energy is unbounded when $r$ becomes small. Small distances between atoms thus lead to the Lennard-Jones energy becoming much larger than the other

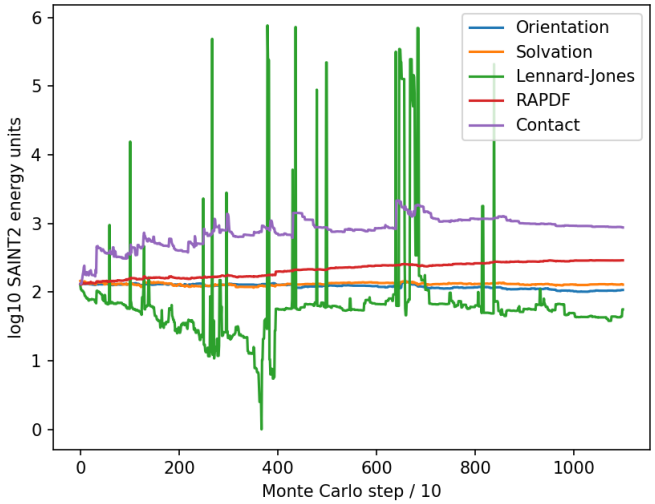


Figure 2.5: **Comparison of different components of the SAINT2 energy function for one standard SAINT2 folding run of 1IXN_A.** The logarithm of the Lennard-Jones energy (green) in SAINT2 energy units experiences large excursions several times, sometimes returning immediately to its previous value, and at other times returning to a different value, indicative of a new structure having been formed. These excursions are often accompanied by step increases in the contact potential, indicating that changes to the Lennard-Jones energy often cause less native-like structures to appear. The other energies are typically stable and remain within one order of magnitude of each other.

energy terms. Such structures are unphysical and represent poor predictions. Fig. 2.5 shows the variation of SAINT2 scores over a run of *standard* SAINT2, but these effects were present for longer periods of Monte Carlo 'time' in our constrained SAINT2 runs.

This effect originates from the default SAINT2 extension procedure. At every stage of the simulation, SAINT2 proposes a new structure by replacing a section of the structure with a fragment from the fragment library. When SAINT2 extends the structure by applying a fragment to a newly added residue, it accepts the move provided that the move satisfies any structural constraints. It does so for technical reasons: the energetic penalty of adding a new residue means that residue addition would be unlikely to be accepted under a Monte Carlo criterion. It is likely that the enforcement of structural constraints increased the likelihood that new structures will contain interatomic clashes, resulting in high Lennard-Jones energies, but we found that these

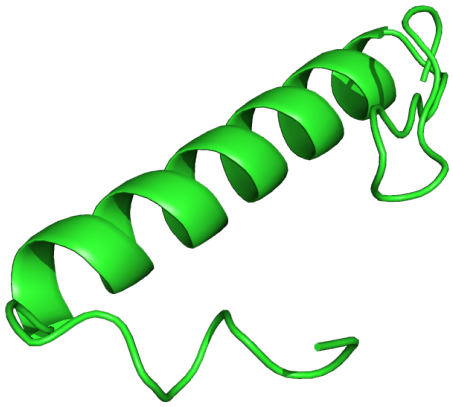

Figure 2.6: **Structural clash following extension in 1IXN_A.** In this run, the C-terminal residues of 1IXN_A, following extension of residue 195, form a strongly clashing interaction that raises the SAINT2 Lennard-Jones energy contribution to almost 700,000 SAINT2 energy units. Residues 1-150 have been removed for clarity.

effects were also present in the default version of SAINT2.

We then hypothesised that these moves would then relax into worse structures than would be accessible from the main folding pathway. This effect is partially visible in the Lennard-Jones energy trace of Fig. 2.5. Here, in several locations, the Lennard-Jones energy does not immediately decrease to the normal range: instead, it decreases in several steps, each of which represents an unfavourable, clashing structure. Since the difference in energy between these states is so large, highly unfavourable changes could have been introduced to the rest of the structure in circumvention of the Monte Carlo criterion.

To test this effect, we introduced an extension procedure that avoids clashes by adding amino acids in an unobstructed direction on the outside of the structure. We use a heuristic for this step: we add the new amino acid such that its $C_\alpha$ atom is collinear with the $C_\alpha$ of the previously-extruded residue and the $C_\alpha$ centre of mass of all of the extruded residues. This protocol is more physically analogous to cotranslational folding: on the ribosome, amino acids are always added in an unobstructed direction. Fig. 2.7 shows the positive effect of this extension procedure. A majority of proteins exhibit some improvement relative to the previous extension procedure, and

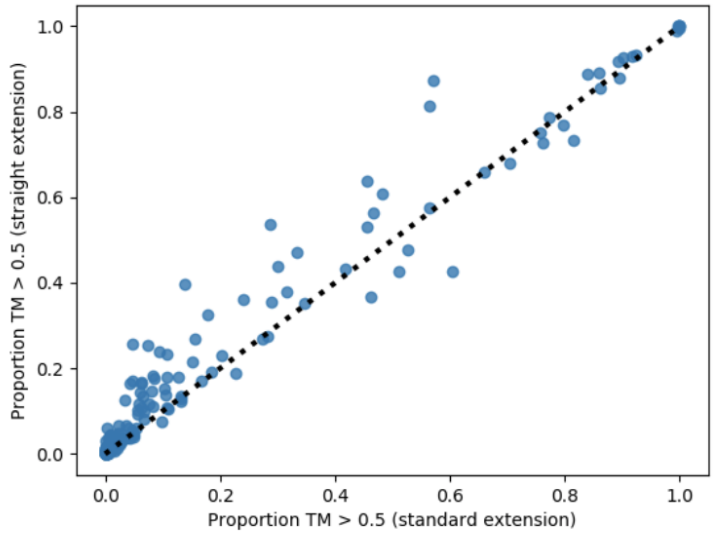


Figure 2.7: **Proportion of models with TM-Score $\geq 0.5$ for the modified extension protocol as a function of proportion with TM-Score $\geq 0.5$ for the unmodified extension protocol.** All points above the diagonal line indicate increased proportion of good TM scores for the modified protocol.

in some cases, this improvement is large.

Moreover, detailed inspection confirms that the connection between cotranslational folding and our modified computational procedure is physically meaningful. Fig. 2.8 shows two representative plots of the move acceptance probability over the course of the run (data shown for 1JLJ_C). The plot for the original extension algorithm contains vertical bands of high move acceptance probability when additional residues are introduced. This implies that moves throughout the entire structure were more likely when a new amino acid was added to the structure. By contrast, the new extension procedure shows fewer of these vertical bands. Addition of amino acids under the new procedure is less likely to unphysically disturb the interior of an already-folded structure. Thus, proteins in the standard extension protocol fold "less cotranslationally" than proteins folding under the new protocol, in addition to folding in a less physically-realistic way due to the introduction of clashes.

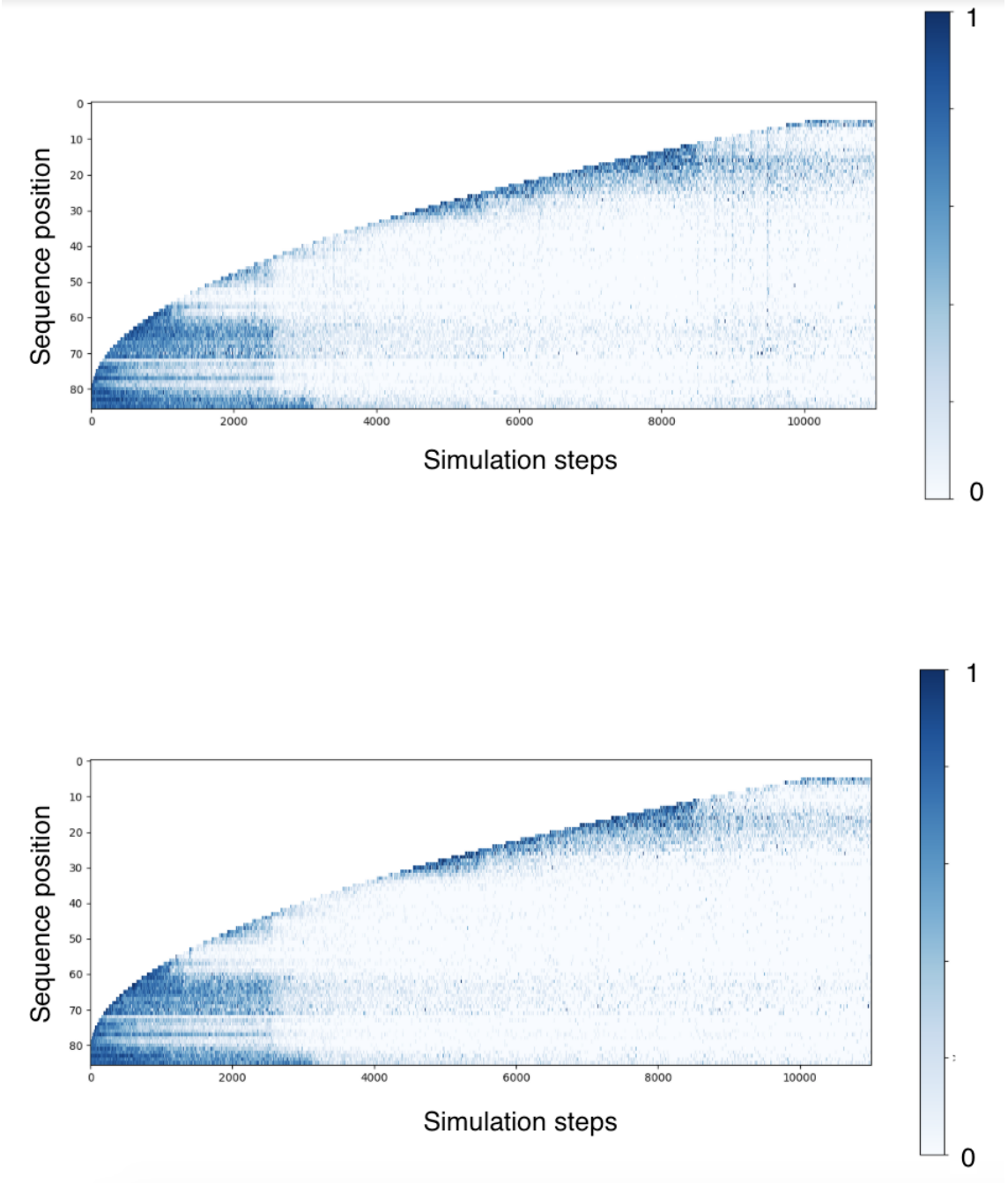


Figure 2.8: **Move acceptance probability plotted as a function of simulation step ($x$-axis) and sequence position ($y$-axis) for target 1JLJ_C.** The upper image shows results for the previous extension protocol, while the lower image shows results for the modified extension protocol. Colour weight indicates probability: darker is more probable. The key feature which distinguishes the two plots are the vertical bands between 8000 and 10000 steps in the upper figure, which do not appear under the modified extension protocol.

## 2.4 Discussion

This investigation resulted in two principal conclusions. First, the introduction of topological constraints to represent the ribosome does not lead to an improvement in predicted models in SAINT2. Tunnel models led to significantly worse outcomes, while wall models led to no observable changes. Therefore we were unable demonstrate a connection between the topological enviroment of the ribosome and physically-relevant protein folding effects.

The effect of the ribosome is likely to be more subtle than simple exclusion of volume. There are several complicating phenomena. Physically, some authors suspect that local structural chemistry may have important effects on the formation of different structural components near to the surface of the ribosome Joiret *et al.,* 2020. Moreover, our model of the ribosome may have been physically-unrealistic in at least two ways. We observed that hairpin turns inside the tunnel resulted in structures that deviated from a physical protein folding pathway, and in a larger sense, the process of fragment substitution may enable leaps in protein structure space which would be difficult to achieve *in vivo*, while excluding the possibility of smaller movements which would be likely to occur *in vivo*. Our model of the ribosome was also 'fully hard', meaning that the folding structure felt no effects in proximity to the modeled ribosome. It is possible that such a strategy could have led to less effective structural search because of the discontinuous nature of this constraint.

Additionally, we found that the SAINT2 extension protocol, even in the absence of structural constraints, can be improved by removing the opportunity for clashes and otherwise-unfavourable relaxations. The efficacy of the new extension protocol underlines the importance of a cotranslational folding, because it decreases the probability of disruptive moves occurring in parts of the protein that have already folded.

SAINT2 is effective for protein structure prediction because of the availability of accurate contact predictions derived from coevolutionary analysis of large alignments of protein sequences. These contact predictions are critical to protein structure prediction for many pipelines. Yet, contact predictions depend on the evolutionary history of the protein under study and the method used to analyse protein sequence alignments. These effects could result in differences in protein structure prediction and may also be independent sources of scientific information. We

analyse these effects in Chapter 3.

# Chapter 3

# Detecting contact-selection bias in contact prediction

This chapter is based on the publication

M. Chonofsky, S. H. P. de Oliveira, K. Krawczyk, and C. M. Deane. The evolution of contact prediction: evidence that contact selection in statistical contact prediction is changing. *Bioinformatics*, 36(6):1750-1756, 15 March 2020. doi: `https://doi.org/10.1093/bioinformatics/btz816`.

In the previous chapter I described my investigation into the ways that the physical extrusion enviroment could affect the formation of native protein protein structures. In this chapter, I examine one of the most important components of structural prediction algorithms—the contact predictions. As described in the introduction, contact prediction has enabled a step change in structural prediction accuracy.

This work was completed in 2019, before results of CASP14 demonstrated the power of AlphaFold2. Once AlphaFold2 or similar software is available, it would be interesting to assess what, if anything, its model has learnt about the physico-chemical properties of the structures it predicts.

## 3.1 Introduction

The development of advanced methods to detect correlation between sites in large multiple sequence alignments has increased the accuracy of protein contact prediction. The predicted contacts output by these methods have resulted in improvements in many areas of structural biology, including template-free protein structure prediction (Jones, Buchan, *et al.,* 2012; Kamisetty *et al.,* 2013). Machine learning-assisted contact prediction methods, such as AlphaFold, have recently demonstrated unprecedented ability to accurately predict protein structures at the level of topology or better (Moult *et al.,* 2018).

These contact prediction methods are based on the idea of coevolution between residues in the protein structure. If a protein is to keep its folded shape when a residue mutates, at least one of the residues with which it is in contact is likely to undergo a compensatory mutation. For example, a mutation which removes one cysteine in a disulfide bond might be compensated by a mutation of the remaining cysteine in order to preserve a bonding interaction between those two sites in the protein. Sites where such compensatory mutations occur frequently can be identified by statistical techniques from multiple sequence alignments. For these techniques to be successful, it is necessary that the multiple sequence alignments contain sufficient levels of sequence diversity to reveal these correlations.

Early contact prediction methods used mutual information between alignment columns to infer contacts. Even with a number of corrections, particularly including the average product correction (Dunn *et al.,* 2008) for phylogenetic and entropic noise, these methods (such as MIp (Dunn *et al.,* 2008), MIc and aMIc (Lee *et al.,* 2009), and ZNMI (Brown *et al.,* 2010)) were unable to accurately infer protein contacts (*i.e.,* residues that share spatial proximity, typically those with $C_\beta$ less than 8 Å apart). Gomes *et al.* (2012) found less than 30% precision at 20% recall for any of the available mutual information-based methods. The low precision of these methods was due in part to their inability to identify contacts within a larger number of transitive correlations.

Direct coupling analysis (DCA) (Morcos, Pagnani, *et al.,* 2011; Marks *et al.,* 2011; Jones, Buchan, *et al.,* 2012) overcame some of the weaknesses of MI methods by correcting for the effect of transitive couplings between residues. Methods such as CCMpred (Seemayer *et al.,* 2014),

Freecontact (Kaján *et al.,* 2014), EVFold (Sheridan *et al.,* 2015), GREMLIN (Balakrishnan *et al.,* 2011), and PSICOV (Jones, Buchan, *et al.,* 2012) all use variations of this methodology. DCA-based contact predictors reached accuracies approaching 50% for the top $L/5$ contacts where $L$ is the length of the protein (Jones, Buchan, *et al.,* 2012). Despite higher accuracy, these methods still obtains a low recall, and it remains unclear why certain contacts are not predicted. A recent paper by Hockenberry *et al.,* 2018 has suggested that DCA methods detect side-chain interactions, while most studies assess recall using an 8Å $C_\beta$ backbone distance cut-off.

In an effort to further increase accuracy and recall, the next development in protein contact prediction was the introduction of meta-predictors, which combined the output of different contact predictors to create aggregate predictions (*e.g.* MetaPSICOV (Jones, Singh, *et al.,* 2015) and PConsC (Skwark *et al.,* 2013)). MetaPSICOV outperforms its constituent predictors (CCMpred, DCA, and PSICOV) by 10% precision or more, as assessed on the top $L$ contacts (Jones, Singh, *et al.,* 2015). Although these methods increase the number of correctly predicted contacts, they also predict a set of contacts which is different from the sets that their constituent predictors predict, for example, by removing contacts that are predicted with low confidence or by only one constituent predictor, or by 'filling in' contacts from secondary structures (Jones, Singh, *et al.,* 2015).

Subsequent developments have centered on the application of deep learning approaches to contact prediction. DNCON2 (Adhikari *et al.,* 2018) and RaptorX (S. Wang *et al.,* 2017) are currently the only published examples of deep learning based contact predictors. (CASP13 and CASP14 featured numerous examples of this class of approach, but these programmes have not yet been released to the community.) Neither RaptorX nor DNCON2 operates directly on the multiple sequence alignment, instead using features derived from statistical coupling inference methods and sequence property predictions, such as predicted secondary structure and predicted solvation. DNCON2 outperforms MetaPSICOV and RaptorX on the CASP10, CASP11, and CASP12 datasets (Adhikari *et al.,* 2018), achieving a precision of 53.4% on the CASP12 dataset, compared with 42.9% and 46.3%, respectively, for MetaPSICOV and RaptorX, for the top $L/5$ predictions of long-range contacts. These methods treat contact prediction as a problem in computer vision, enabling the application of higher-order structures to the data, and resulting

in a set of correctly-predicted contacts that is again larger than those predicted by DCA or meta-prediction methods. This larger set must again contain different contacts from those identified by DCA or meta-prediction.

Contact prediction methods have been used to approach many bioinformatics problems, from protein structure prediction to inference of functional interactions, but little work has been done to understand the nature of the contacts that they predict. Given that these methods were all initially based on identifying co-evolving sites, it could be expected that the contacts that they predict relate to specific types of interactions. It is also likely that there are differences between contacts predicted by different methods. While more modern prediction methods may improve the accuracy of the predictions, as they move further from attempting to extract coevolutionary signal, the physico-chemical nature of the sets of predicted contacts may change. Direct coupling methods identify contacts that exhibit strong statistical coevolutionary signal, and may therefore identify contacts that have particular evolutionary significance. The effect of adding other information to these predictions through deep learning is not known. These differences might be key in understanding their utility for different problems.

Since the publication of DNCON2 and RaptorX, the field has taken a dramatic turn toward end-to-end protein structure prediction using deep neural networks (ref CASP14). In particular, DeepMind has made blind protein structure predictions of a standard which approximate and in some cases surpass experimental accuracy. This new development has not been examined in this chapter for two reasons. Most importantly, the DeepMind's AlphaFold2 has not been released for use by the community and may require highly-specialized hardware (ref Deepmind Blog). Our work in this chapter was carried out before the publication of DeepMind's advance. Indeed, differences between different methods of contact prediction originating in different methods of analysis may reveal patterns that are physically or biologically meaningful despite being leading to less accurate structures than AlphaFold2.

In this chapter, we investigate the nature of the predicted contacts from different contact prediction methods. We compare aMIc, CCMpred, MetaPSICOV, and DNCON2 as examples of the different types of contact predictors currently available, and we assess the differences between true contacts predicted by the methods and random true contacts in protein structures.

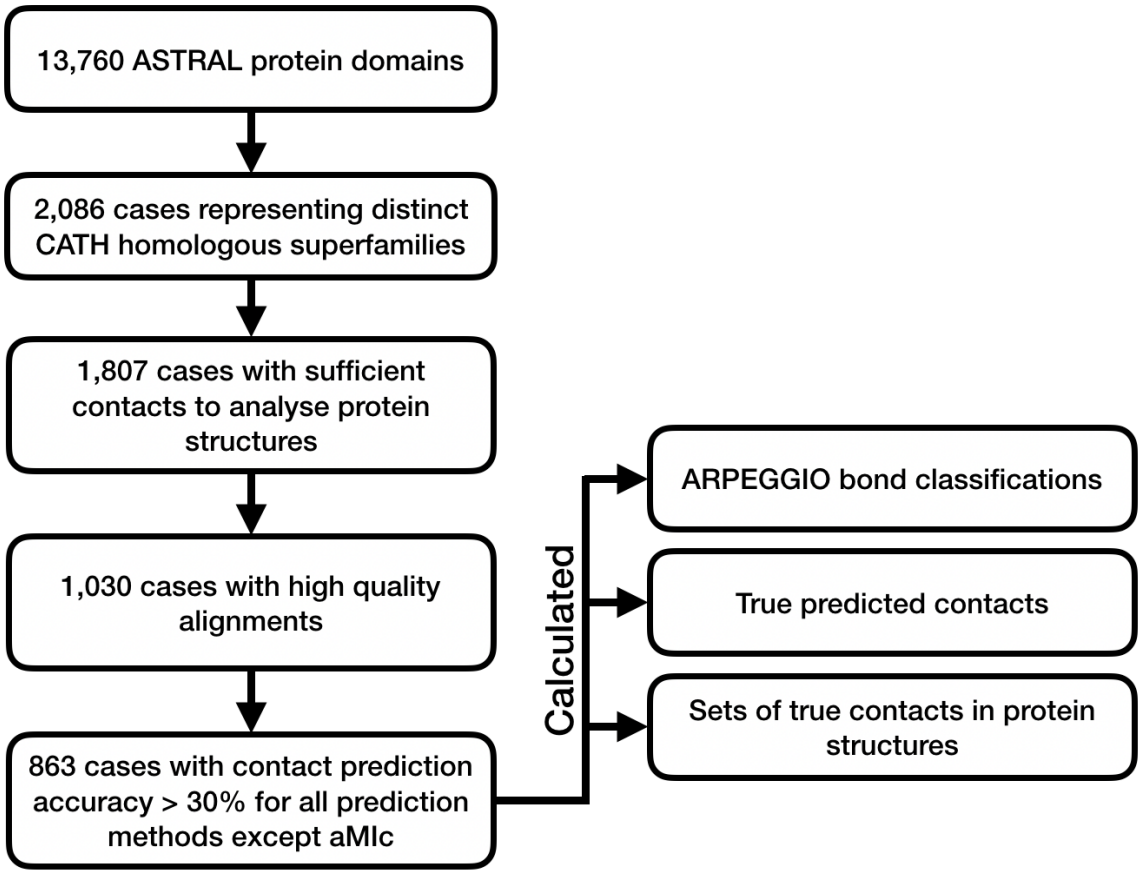


Figure 3.1: **A schematic of the data processing pipeline for our analysis.** As described in the main text, we filtered domains from ASTRAL to produce a set of domains with structural and functional diversity. This set of domains was used as the basis for contact prediction and categorisation of structural properties.

We classify the bonds which are formed between residues in our sets of contacts, and we show differences in the number and kind of physico-chemical bonding interactions between different methods, and between predicted contacts and random contacts. We show commonalities between machine-learning based methods (MetaPSICOV and DNCON2) and direct coupling analysis. Further, we find differences in the extent to which bonds are conserved between different sets of contact predictions and between contact predictions and the set of all contacts.

## 3.2 Methods

### Approach

In this chapter, we consider a set of protein domains from the ASTRAL database. A schematic of our pipeline is shown in Figure 3.1. We introduce the following terminology to describe contacts:

**Predicted set** Of the top $L$ predicted contacts for a given protein structure, the predicted set is the set of residue pairs which are in contact in that protein structure (true contacts), where $L$ is the length in residues of the structure. Therefore, the size of the predicted set is at most $L$.

**Background set** A randomly-selected set of residue pairs which are not in the predicted set but which are in contact in a given protein structure (false negatives) . For each protein structure, we select the same number of contacts for the background set as are in the predicted set. For most analyses, we use 20 randomly-selected background sets for each structure to improve statistical reliability.

### Structural domain set

From the 13,760 domains in ASTRAL (06.02.2016 build at 40% sequence identity cut-off) (Fox *et al.,* 2014; J.-M. Chandonia, 2004; J. M. Chandonia *et al.,* 2002; Brenner *et al.,* 2000), we selected a single exemplar domain for each CATH (Dawson *et al.,* 2017) homologous superfamily, giving 2,086 protein domains. For each protein domain, we assembled a Multiple Sequence Alignment (MSA) and predicted contacts for that alignment. (See below for more details.)

**Multiple sequence alignment generation** For each domain, we generated an MSA using HHBlits 3.0.0 (15-03-2015, default options except `-n 3`, `-maxfilt 500000`, `-id 99`, `-cov 0.90`) with the Uniprot20 database (2016.02) (Bateman *et al.,* 2017). In order to ensure alignments of sufficient quality for use in contact prediction, we removed MSAs which had $N_f < 32$ (Ovchinnikov, Park, Varghese, *et al.,* 2017).

### Protein contact properties

**Contact definition** Contacts are defined as residue pairs where the distance between $C_\beta$ atoms ($C_\alpha$ for glycine) is less than 8Å. While this cut-off is arbitrary, it is in accordance with

convention in the field, and in particular it is the cut-off with which DNCON2 and MetaPSICOV were trained (Adhikari *et al.,* 2018; Jones, Singh, *et al.,* 2015). We consider only those contacts which are separated by five or more residues.

**Contact prediction** We used our MSAs as input to four contact prediction methods: aMIc (Lee *et al.,* 2009), CCMpred (Seemayer *et al.,* 2014), gDCA (Baldassi *et al.,* 2014), MetaPSICOV version 1 (Jones, Singh, *et al.,* 2015), and DNCON2 (Adhikari *et al.,* 2018). For each of these prediction methods, we used default parameters except in the following ways. For aMIc, we used a pseudocount value of 0.05 in pairwise residue counts so that the marginal contributions of the pseudocounts for each residue was 1. We also modified the DNCON2 pipeline to use our HHBlits alignments so that all four methods had identical input. After contact prediction, we assessed contact prediction accuracy and removed cases in which any of CCMpred, MetaPSICOV, and DNCON2 had contact prediction accuracy over the top $L$ contacts below 30%, where $L$ is the length of the protein domain. We also removed structures where there were too few real contacts to populate the background set (see below). A full list of all 2,086 cases and their alignment and contact prediction statistics are given in SI.

**Physico-chemical interactions** We used ARPEGGIO (Jubb *et al.,* 2017) to identify the types of physico-chemical interactions between amino acids in the three-dimensional protein structures of our domains. ARPEGGIO uses molecular geometry to classify physico-chemical interactions into 13 Structural Interaction Fingerprints (SIFts) (Deng *et al.,* 2004). The most common interaction types by overall count were `hydrophobic`; `polar`, `hydrogen_bond`, and `weak_polar` and `weak_hydrogen_bond`; and `vdw` (van der Waals). We also observed `carbonyl`, `aromatic`, `ionic`, and `covalent` interactions. We did not count the `proximal` category because it is a $d \leq 5$Å distance bin, overlapping substantially with other interaction types without implying a specific physico-chemical interaction. A full list of physico-chemical interaction types is given in Table 3.1 and the geometric and chemical criteria used to identify and label these bonds are given and discussed in Jubb *et al.,* 2017 . We call these attractive physico-chemical interactions "bonds" because they represent attractive physical interactions between atoms. While some (i.e., disulfide bonds) are covalent, most are not.

| | |
|---|---|
| `clash` | Denotes if the covalent radii of the two atoms are clashing, i.e. steric clash |
| `covalent` | Denotes if the two atoms appear to be covalently bonded |
| `vdw_clash` | Denotes if the van der Waals radii of the two atoms are clashing |
| `vdw` | Denotes if the van der Waals radii of the two atoms are interacting |
| `proximal` | Denotes the two atoms being > the VdW interaction distance, but with in 5 Angstroms of each other |
| `hbond_like` | Denotes if the atoms form a hydrogen-like bond. ARPEGGIO has four related classifications, `weak_hydrogen_bond`, `hydrogen_bond`, `weak_polar`, and `polar`, which we combine. |
| `ionic` | Denotes if the atoms may interact via charges |
| `aromatic` | Denotes two aromatic ring atoms interacting |
| `hydrophobic` | Denotes two hydrophobic atoms interacting |
| `carbonyl` | Denotes a carbonyl-carbon:carbonyl-carbon interaction |

Table 3.1: **Structural Interaction Fingerprints** This table gives the SIFts identified in our collection of protein domains. SIFts are defined according to the definitions in Jubb et al. (2017), except that we we combine the `hydrogen_bond`, `weak_hydrogen_bond`, `polar_bond`, and `weak_polar_bond` categories into one category `hbond_like`. Identification of interacting pairs is on the basis of bond geometry and atom type. Further specification of the identification of interactions is available in Jubb et al. (2017).

## Structural analysis

**Structural alignment** Protein-protein structural alignments were carried out with CATH-SSAP (Dawson *et al.,* 2017), since we used CATH homologous superfamilies in structural classification.

**Secondary structure classification** STRIDE (Frishman *et al.,* 1995) was used to assign contacts to secondary structures. We classified contacts into four categories: Loop-Loop (contacts formed between residues in loops), SS-Loop (contacts formed between a residue in a loop and a residue in a secondary structure elements), within-SS (contacts formed between residues within one secondary structure element), and between-SS (contacts formed between residues within two different secondary structure elements). We classified contacts as within-SS by considering runs of consecutive $\alpha$ or $\beta$ residues. If two contacting residues $A$ and $B$ were situated in runs $R_A$ and $R_B$ of the same secondary structure type, we classified the contact $(A, B)$ as within-SS if there was a main-chain hydrogen bond between any of the residues in $R_A$ and $R_B$, or if $A$ and $B$ were situated in the same run. We also allowed transitive effects: if a third residue $C$ were located in a run $R_C$ that had a main-chain hydrogen bond with $R_B$, the contact $(A, C)$ would have been classified as within-SS.

**Effective isolated contacts** To assess the distribution of contacts, we sought the largest set of contacts which could be considered isolated. Specifically, we considered a contact $A : (A_1, A_2)$ between an amino acid with residue index $A_1$ and amino acid with residue index $A_2$ to be isolated if there was no predicted contact $B : (B_1, B_2)$ such that $|A_1 - B_1| \leq 1$ and $|A_2 - B_2| \leq 1$. We constructed an undirected graph on predicted contacts, with contacts corresponding to vertices and edges between contacts $A$ and $B$ iff $|A_1 - B_1| \leq 1$ or $|A_2 - B_2| \leq 1$. We then found a minimal vertex cover on this graph using a 2-approximation algorithm (Savage, 1982), *i.e.*, we identified the minimal set $C$ of contacts such that $C$ was adjacent to every contact not in $C$. The number of effective isolated contacts was the number of contacts not present in the vertex cover. We computed the vertex cover for all correct contacts inferred by any method.

**Adjusted probabilities** We computed the probability that a contact of a particular bond type was predicted by a each prediction method. In order to account for different sizes of contact sets from different prediction methods, we adjusted these probabilities by a factor equal

to the ratio of the length $L$ of the protein to the number of correct contacts in the set under consideration *i.e.*,

$$(N_{i,\text{set}}/N_{i,\text{all contacts}})\,(L/N_{\text{set}})\,,$$

for a bond type $i$ and the number $N_{\text{set}}$ of contacts in the predicted set for a contact prediction method. $N_{i,\text{set}}$ is the number of contacts displaying bond type $i$ which are in the predicted set of a particular prediction method. $N_{i,\text{allcontacts}}$ is the number of contacts displaying bond type $i$ in the set of all contacts in the protein domain. Thus, these probabilities are scaled to compensate for the effect of predicted sets of different sizes due to different contact prediction accuracies. These adjusted probabilities were averaged over the 863 cases.

## 3.3 Results and discussion

### Trends in contact prediction accuracy

We predicted contacts on 1,030 high-quality alignments of protein domains using four contact prediction methods (aMIc, CCMpred, MetaPSICOV, and DNCON2). We also considered gDCA, a Gaussian-based direct coupling method. Its prediction accuracy is similar to CCMpred (Potts model) and it represents the same generation of contact prediction as CCMpred. We have included results for gDCA alongside CCMPred where possible, but since these results tend to recapitulate the patterns seen in the output of CCMPred, we have omitted analyses involving gDCA later in the chapter.

Figures 3.2 shows the accuracy achieved over the top $L$ contacts, where $L$ is the length of the protein. As expected, aMIc (the mutual information method) performed worst (average accuracy of 15%). The best-performing method was DNCON2 (average accuracy of 77%) followed by MetaPSICOV (average accuracy of 64%) and CCMpred (average accuracy of 47%). CCMpred and gDCA, which are similar methods originating in the same generation of contact prediction algorithms, had similar average accuracy. We found that alignment quality was correlated with prediction accuracy for all prediction methods (Figure 3.3). We have used identical alignments for all methods with the aim of reducing the effect of this potentially confounding factor.

Since the purpose of this study is to investigate the physico-chemical properties of the true predicted contacts, we did not take aMIc contact predictions forward for further analysis, because

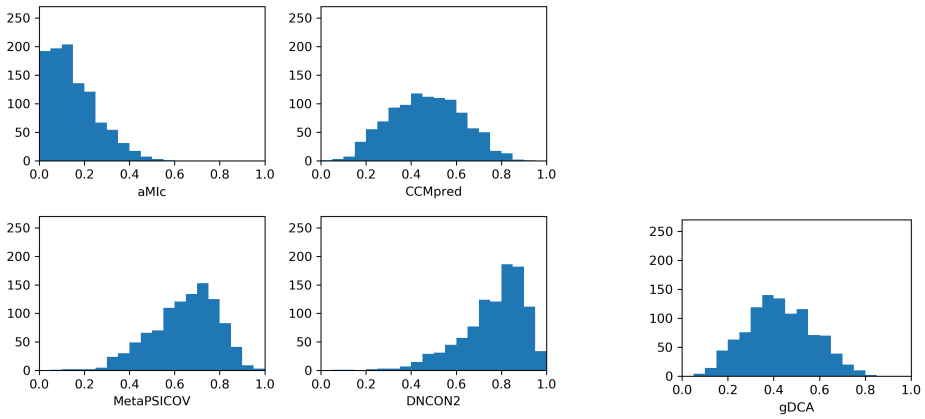


Figure 3.2: **Top-$L$ accuracy histograms of different contact prediction methods.** Accuracy was computed with respect to the top $L$ scoring predictions, where $L$ is the length of the protein domain, for five prediction methods – aMIc, CCMpred, gDCA, MetaPSICOV, and DNCON2 – over 1,030 protein domains. The $y$ axis is the number of protein domains, and the $x$ axis is the top-$L$ accuracy. This analysis excludes cases where effective sequences $N_f < 32$, which is known to result in poor predictions (Ovchinnikov, Park, Varghese, *et al.*, 2017).

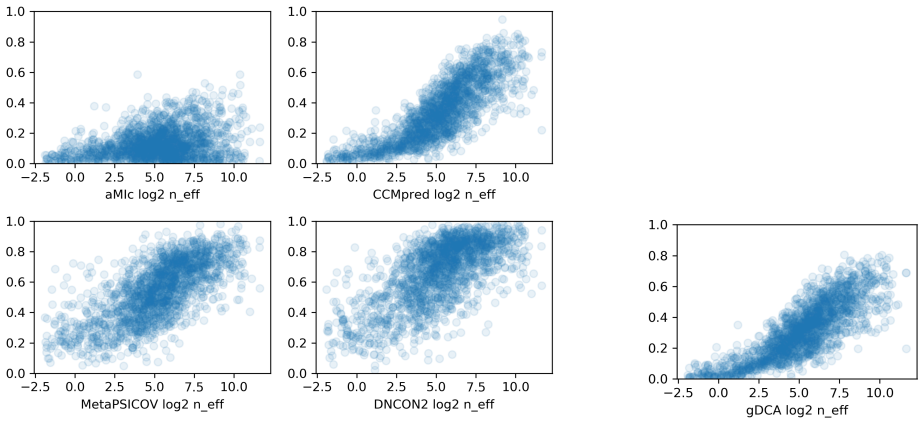


Figure 3.3: **Prediction accuracy as a function of alignment quality.** For each prediction method, top-L accuracy is plotted as a function of $\log_2 N_f$ (Ovchinnikov *et al.* (2017)).

only 102 cases had top-$L$ accuracy equal to 30% or higher. To fairly compare the three methods in terms of the physico-chemical properties of their predicted contacts, we used only the 863 cases for which all three methods had top-$L$ prediction accuracy above 30% and sufficient contacts available in the structure to form a predicted set and a background set for our analyses.

### Predicted contacts have more bonds than background contacts

Using this set of 863 cases, we compared the properties of the correct predicted contacts for each case (predicted set) to those of a randomly-selected set of residue pairs that are in contact in that protein structure and which were not in the predicted set (background set). The bonds between residue pairs in both the background and predicted sets were identified by ARPEGGIO (see Methods). Fig. 3.4 shows the number of bonds per contact averaged over the 863 prediction cases. For all three contact prediction methods, there are more bonds per contact for the predicted contact sets than the background contact sets. CCMpred exhibits the largest increase (58%), while MetaPSICOV has the smallest increase (47%). The bias toward selecting heavily-bonded contacts for all prediction methods suggests that physico-chemical bonds play a role in determining the coevolutionary signal in alignments. If the need to preserve existing chemical interactions drives the correlated mutations that give rise to the evolutionary signal in protein multiple sequence alignments, then it makes sense that those contacts which have the largest number of bonds are likely to be predicted, and that introducing other sources of contacts would result in fewer bonds per contact.

### MetaPSICOV and DNCON2 predict almost twice as many within-secondary-structure contacts as CCMpred

To further probe the nature of this difference, we separated the counts of contacts that occurred between loops and secondary structures. Although contacts in general are disproportionately found between secondary structure elements, MetaPSICOV and DNCON2 predict almost twice as many within-secondary-structure contacts as CCMpred, despite their background sets having similar compositions (Fig. 3.4 (b)). These general measures of the sets of all contacts mask sharper effects of individual contact predictors because all contact predictors predict some of the same contacts. In order to more precisely identify the properties of individual contact predictors, we considered those contacts which were predicted only by particular contact predictors.

For each of the 863 protein domain cases, and restricting ourselves to the top $L$ predictions, we considered separately those correct contacts that were predicted uniquely by CCMpred, gDCA, DNCON2, and MetaPSICOV. We also considered those contacts that were predicted

| Predictor | A | B | C | D | E | F | G | H | I | J |
|---|---|---|---|---|---|---|---|---|---|---|
| CDM | 0.26 | - | 2.33 | 0.37 | 0.38 | 0.234 | 0.364 | 0.147 | 0.210 | 0.279 |
| CM | 0.06 | - | 1.87 | 0.13 | 0.41 | 0.086 | 0.126 | 0.064 | 0.085 | 0.134 |
| DM | 0.27 | - | 1.81 | 0.55 | 0.17 | 0.441 | 0.569 | 0.299 | 0.357 | 0.349 |
| CD | 0.06 | - | 2.17 | 0.18 | 0.36 | 0.173 | 0.326 | 0.119 | 0.183 | 0.231 |
| M | 0.08 | 0.12 | 1.26 | 0.26 | 0.23 | 0.160 | 0.246 | 0.112 | 0.158 | 0.208 |
| C | 0.11 | 0.23 | 1.74 | 0.07 | 0.34 | 0.074 | 0.139 | 0.052 | 0.082 | 0.129 |
| D | 0.21 | 0.27 | 1.56 | 0.37 | 0.17 | 0.311 | 0.465 | 0.224 | 0.276 | 0.337 |

Table 3.2: **Statistics for unique predictions of CCMPred, MetaPSICOV, and DNCON2.** Contact prediction methods are abbreviated to their initial letters. CDM indicates those contacts predicted by CCMpred, DNCON2, and MetaPSICOV, while CM indicates those predicted only by CCMpred and MetaPSICOV, and so forth. All statistics are averages over 863 cases. **A:** Ratio of number of correct contacts to $L$, the length of the protein. **B:** Ratio of the number of contacts that are unique to each predictor to the number of the contacts predicted correctly by that contact predictor. **C:** Number of bonds per contact. **D:** Ratio of contacts within secondary structures to the number of contacts in each category, following the definition given in the main text. **E:** Proportion of correct predicted contacts with a bond between two side-chain atoms, where main-chain heavy atoms are N, O, C, $C_\alpha$, and $C_\beta$. **F-J:** Proportion of all contacts (**F**), within-secondary structure contacts (**G**), between-secondary structure contacts (**H**), secondary structure to loop contacts (**I**), or loop-loop contacts (**J**) which have main-chain/main-chain bonds.

by pairs of contact predictors, and those which were predicted by all three contact prediction methods (Table 3.2).

For each of the 863 protein domain cases, and restricting ourselves to the top $L$ predictions for each prediction method, we considered the union of the predicted sets for CCMpred, DNCON2, and MetaPSICOV. We then considered the ratio of the number of contacts in subsets of this group to $L$, the maximum number in the predicted set for any predictor. Specifically, we considered the three subsets that contained those contacts that were predicted uniquely by

| Predictor | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| DMG | 0.25 | - | 2.02 | 0.37 | 0.38 | - |
| DM | 0.29 | - | 1.45 | 0.55 | 0.18 | - |
| DG | 0.06 | - | 1.91 | 0.17 | 0.36 | - |
| MG | 0.06 | - | 1.67 | 0.13 | 0.41 | - |
| G | 0.11 | 0.22 | 1.57 | 0.07 | 0.34 | 16.69 |
| D | 0.21 | 0.25 | 1.24 | 0.37 | 0.17 | 11.78 |
| M | 0.09 | 0.12 | 1.13 | 0.25 | 0.24 | 7.94 |

Table 3.3: **Statistics for unique predictions of gDCA, MetaPSICOV, and DNCON2.** As in Table 3.2, contact prediction methods are abbreviated to their initial letters. GDM indicates those contacts predicted by gDCA, DNCON2, and MetaPSICOV, while GM indicates those predicted only by gDCA and MetaPSICOV, and so forth. All statistics are averages over 863 cases. **A:** Ratio of number of correct contacts to $L$, the length of the protein. **B:** Ratio of the number of contacts that are unique to each predictor to the number of the contacts predicted correctly by that contact predictor. **C:** Number of bonds per contact. **D:** Ratio of contacts within secondary structures to the number of contacts in each category, following the definition given in the main text. **E:** Proportion of correct predicted contacts with a bond between two side-chain atoms, where main-chain heavy atoms are N, O, C, $C_\alpha$, and $C_\beta$. **F:** Mean number of effective isolated contacts. We observe the same trends for CCMpred (Table 3.2) as for gDCA.

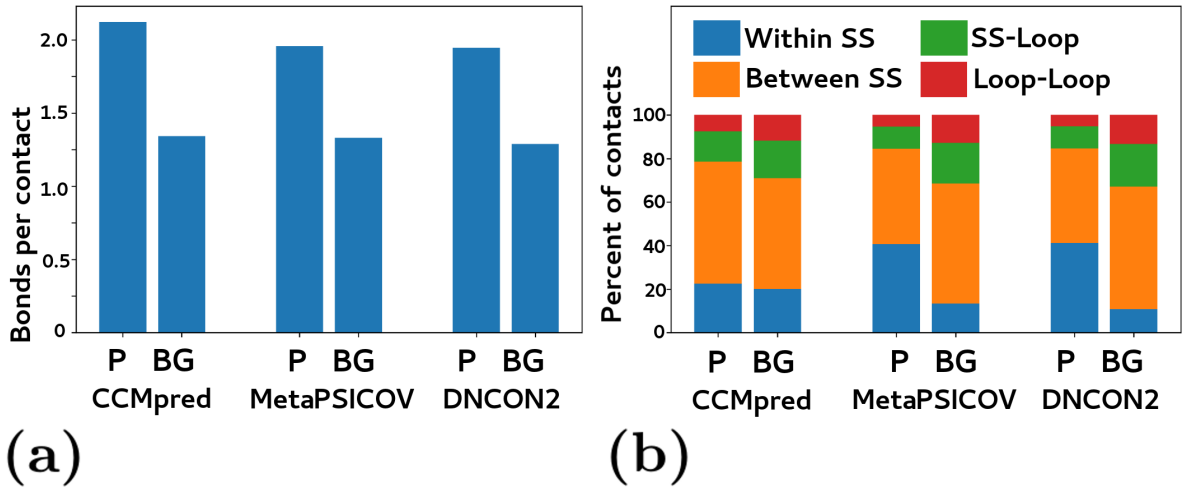


Figure 3.4: **A comparison of interactions between predicted set and background set contacts.** (a) shows the number of bonds per contact for the prediction methods in terms of the background and predicted sets of contacts. The figure shows the average value of bonds per contact 863 protein domains with top-$L$ prediction accuracy above 0.3 for all three methods. (b) shows the difference in secondary structure composition of contacts between the predicted and background sets for different prediction methods. The average count of contacts between secondary structures, within secondary structures, between loop regions (Loop-Loop), or between loops and secondary structure (SS-Loop), is plotted.

CCMpred, DNCON2, or MetaPSICOV, as well as the three subsets that contained those contacts predicted by only two of the three predictors, and the subset containing contacts predicted by all three predictors (Table 3.2). The largest group is the set of contacts that are predicted by all three methods ($0.27L$). DNCON2 and MetaPSICOV share an equivalently large number of contacts ($0.27L$) while CCMPred shares with MetaPSICOV and DNCON2 only $0.07L$ and $0.06L$, respectively. This points to a strong link between DNCON2 and MetaPSICOV predictions. Moreover, MetaPSICOV has the lowest proportion of unique predictions ($0.11L$ of its correct predictions), while DNCON2 and CCMPred have comparable proportions ($0.24L$ and $0.22L$, respectively), despite DNCON2's higher predictive accuracy. This analysis points to differences between raw DCA-based methods and methods which incorporate information from other sources. DNCON2 and MetaPSICOV predict similar sets of contacts, while the CCMpred predicted sets tend to contain different contacts than the other two predicted sets. In light of

the broader trend that CCMpred tends to predict fewer within-secondary structure contacts, and that there are similarities between the predictions of DNCON2 and MetaPSICOV that are not shared by CCMpred, we repeated earlier analyses to consider their distribution over those contacts that were predicted uniquely by one predictor, by pairs of predictors, and by all three predictors together. In all cases, the standard errors were less than $0.003L$.

First, considering the numbers of bonds per contact, we found that the contacts with the largest numbers of bonds on average were those that were predicted by all three methods (Table 3.2, column C). Those predicted by two or more methods also had more bonds per contact than those predicted by only one method. Of the contacts predicted by only one method, those contacts predicted only by MetaPSICOV had the lowest number of bonds per contact (1.26), while those predicted by CCMpred had the highest number of bonds per contact (1.74). Those contacts predicted by both CCMpred and DNCON2 had the highest number of bonds per contact (2.17), exceeding both sets of combinations which involved MetaPSICOV (1.87 and 1.81). As expected, in light of our findings related to secondary structures, contacts predicted by both DNCON2 and MetaPSICOV had the highest number of hydrogen bonds per contact (0.67, compared to 0.32 and 0.49 for those predicted by both CCMpred, and MetaPSICOV and DNCON2, respectively). These data confirm the idea that coevolutionary couplings are linked to the strength of the bonds between the residues that comprise them. Those contacts that are easiest to predict, in the sense that they are predicted by all three predictors, have the highest numbers of bonds per contact. This relationship is likely due to contacts with particularly strong and numerous bonds generating strong co-evolutionary signal which results in their prediction by all three methods. As noted below, there is not an unusually large proportion of within-secondary structure contacts in this group, suggesting that these predictions are not due to presence within secondary structures.

Those contacts predicted only by CCMpred have the largest number of bonds per contact of those sets from an individual contact prediction method. CCMpred uses raw co-evolutionary signal, and this signal appears to reflect the number of bonds in the contacts.

Further, CCMpred-predicted contacts have more side-chain contacts than those contacts predicted by other methods. We defined side-chain contacts as those contacts that had at least one

side-chain to side-chain bond. As shown in Table 3.2, column E, contacts predicted by CCMpred had a consistently higher proportion of these contacts than those predicted by MetaPSICOV and DNCON2 and consistently lower proportions of main-chain/main-chain contacts in all secondary structure contexts (defined as those contacts with at least one main-chain/main-chain bond, Table 3.2, columns F-J). This result is consistent with recent work, *e.g.* (Hockenberry *et al.,* 2018) but extends it by quantifying the extent of the difference in side-chain contacts. We find that only a minority of all contacts contained side-chain/side-chain bonds. It is plausible that machine-learning algorithms which are trained to maximise the proportion of $C_\beta$ contacts are more likely to omit contacts where there are significant side-chain interactions because those residues may be farther apart on average. Therefore, they may fail to detect important chemical interactions between side chains. By contrast, covariation-based methods use an unsupervised approach, and hence the types of contacts they recover depends on the biophysical mechanisms that create the covariation. These mechanisms may be more closely tied to the identity and position of side chains than to the backbone atoms.

We also assessed the secondary structure characteristics of the predicted contact sets (Table 3.2, column D). The set with the highest level of contacts within a secondary structure (55%) are between DNCON2 and MetaPSICOV. The lowest level of within-secondary-structure contacts were those predicted by CCMpred alone (7%), followed by those shared between CCMpred and one of the other predictors. These data suggest that the co-evolutionary signal within secondary structures is relatively weak, presumably because these structures are harder to disrupt than supersecondary interactions. Machine-learning methods may also capitalize on the ease with which it is possible to recognise and suggest contacts within secondary structures, increasing their proportion of these types of contacts in order to increase their total accuracy.

### CCMpred contacts are distributed more widely in protein structures

We also examined to consider the distribution of contacts within protein structures. As described in Methods, we considered a contact $(A_1, A_2)$ between amino acid $A_1$ and amino acid $A_2$ to be isolated if there was no predicted contact $(B_1, B_2)$ from the set of all predicted contacts such that $|A_1 - B_1| \leq 1$ and $|A_2 - B_2| \leq 1$. As a measure of the distribution of the contacts throughout the protein, we used an established algorithm to remove contacts from the contact sets until all

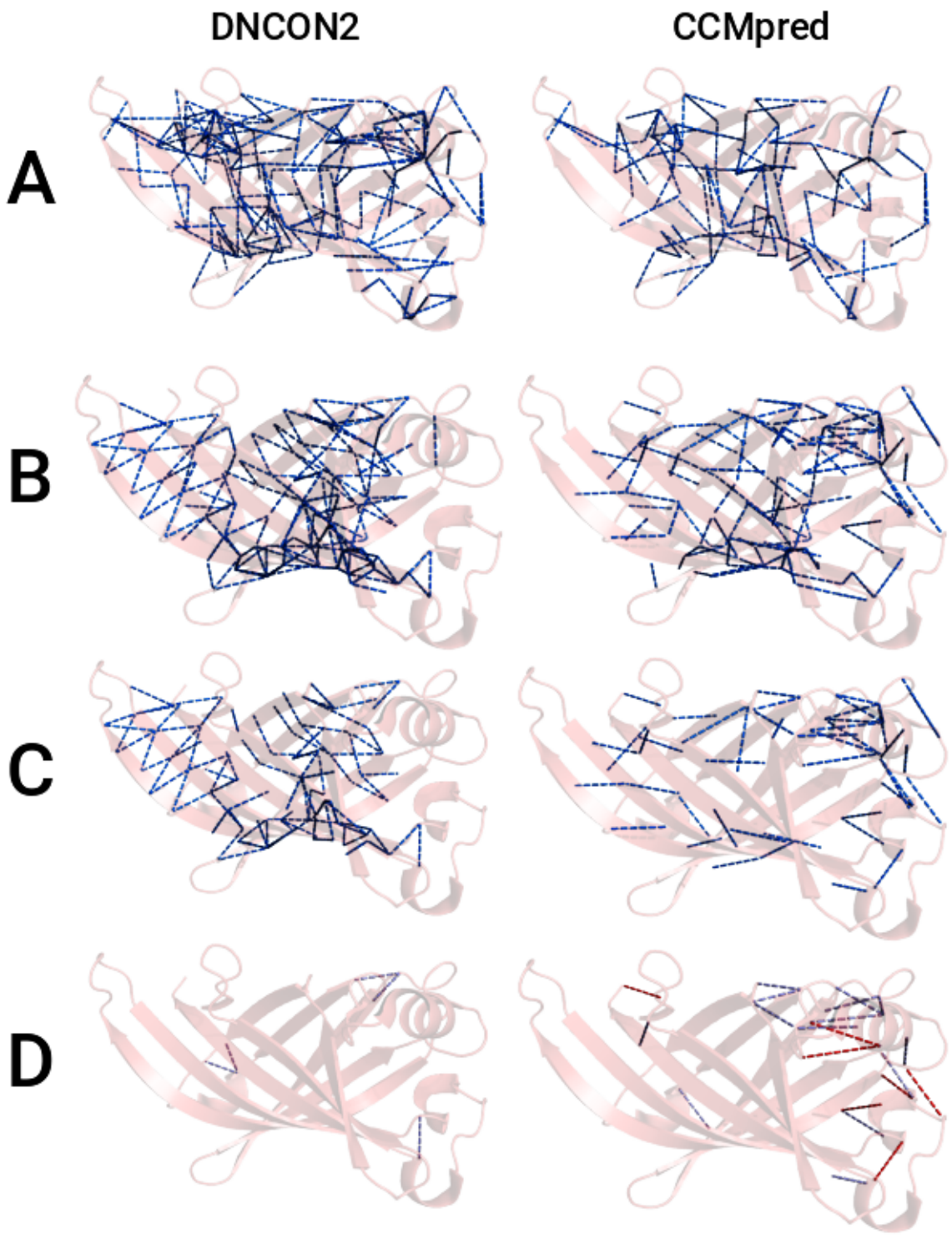


Figure 3.5: **A comparison of predicted contacts for PDB 1Y0G. A:** Background sets. **B:** Predicted sets. **C:** Contacts predicted by only one of the two predictors, *e.g.,* those predicted by CCMpred but not DNCON2. **D:** Contacts from **C** associated with bonds that are not within a single secondary structure. Contacts drawn in purple connect residues that have at least one hydrogen bond; contacts drawn in red have no hydrogen bonds associated with them.

remaining contacts were isolated (Savage, 1982). We refer to the number of remaining contacts as *effective* isolated contacts. CCMpred had more effective isolated contacts than DNCON2 ($0.090L$ and $0.052L$) and both had more effective isolated contacts than MetaPSICOV ($0.033L$). Only 6% of those contacts that were predicted by both DNCON2 and MetaPSICOV were isolated, the lowest proportion of any combination of predictors or individual predictor. These data suggest that CCMpred predicts contacts which have a broader distribution within protein structures than MetaPSICOV and DNCON2. Specifically, our evidence is that DNCON2 and MetaPSICOV tend to predict blocks of contacts corresponding to complete secondary structures. CCMpred, however, tends to make more isolated predictions. These results suggest that machine learning-based predictors are learning to 'fill in' secondary structure contacts. Additionally, isolated predictions are more likely to be incorrect, so predictors may learn to discard 'riskier' isolated contacts and promote 'safer' contacts which are connected to other blocks of contacts. Other work relating to machine learning for contact prediction have also noted that if a residue is in contact with another, then their neighboring residues are more likely to be in contact (Wozniak *et al.,* 2014) and it appears that this effect is incorporated into DNCON2 and MetaPSICOV.

As an example of these differences, we plotted the predicted contacts for PDB structure 1Y0G (Figure 3.5). Both CCMpred and DNCON2 exhibit noticeable ordering of their predicted contacts (4B) compared to background (4A). Although CCMpred predicts fewer contacts than DNCON2, its predictions include a greater proportion of SS-loop and between-SS contacts (4C). Excluding the within-SS contacts and those without bonds, DNCON2 predicts only five contacts, all of which are associated with hydrogen bonds, while CCMpred predicts 22, of which seven have hydrogen bonds (4D). This example demonstrates the possibility of divergence between contacts predicted by CCMpred and DNCON2 in terms of structural and chemical factors.

These differences between bond numbers and between kinds of contacts among the contact predictors led us to consider whether bond types differed in similar ways.

### Types of bonding interactions differ between contact predictors

Predicted contacts have more bonds, which suggests a link between coevolutionary signal and the physical effects which bonds mediate. We sought to investigate whether this difference also manifested in a change in physico-chemical properties of the bonds that mediate contact

predictions. We used the Cochran-Mantel-Haenszel procedure (Cochran, 1954; Mantel *et al.,* 1959) to test whether the distribution of bonding interactions in the background sets of proteins were different from the distribution of bonding interactions in the predicted set. In all cases, $p \ll 0.01$, so we considered the differences between the predicted and background sets in further detail.

We considered the probabilities that a contact with a particular type of bond would be found in the predicted set using the adjusted probability methodology described in Methods. These probabilities are given in Table 3.4. (Raw probabilities are available in Table 3.5.) For each contact type, cases in which no contacts of that type were found in the protein structure were excluded from the average.

| | | | | BG | BG | BG |
|---|---|---|---|---|---|---|
| | CCMpred | MetaPSICOV | DNCON2 | CCMpred | MetaPSICOV | DNCON2 |
| `covalent` | 0.97 | 0.49 | 0.43 | 0.21 | 0.31 | 0.33 |
| `ionic` | 0.84 | 0.43 | 0.4 | 0.23 | 0.35 | 0.36 |
| `hydrophobic` | 0.61 | 0.48 | 0.44 | 0.31 | 0.34 | 0.35 |
| `aromatic` | 0.58 | 0.35 | 0.34 | 0.31 | 0.38 | 0.39 |
| `vdw` | 0.47 | 0.47 | 0.49 | 0.36 | 0.34 | 0.32 |
| `vdw_clash` | 0.46 | 0.52 | 0.55 | 0.36 | 0.32 | 0.28 |
| `hbond_like` | 0.43 | 0.5 | 0.54 | 0.37 | 0.33 | 0.3 |
| `carbonyl` | 0.25 | 0.65 | 0.72 | 0.41 | 0.25 | 0.18 |

Table 3.4: **Adjusted probabilities of prediction of bond types for background and predicted sets.** Table 1 gives adjusted condititional probabilities for finding a bond in the predicted set, given that it is of a certain type. In this table we give the probability of finding a bond in the background set, given that it is of a certain type. We also repeat the probabilities from Table 1 for comparison.

A difference between contact prediction methods is evident from these data. The range of probabilities for CCMpred is larger than the range for DNCON2 or MetaPSICOV. Moreover, CCMpred has a different distribution of conditional probabilities than the other two contact prediction methods, where the figures are broadly similar. The contacts most likely to be selected

| | CCMpred | MetaPSICOV | DNCON2 | BG CCMpred | BG MetaPSICOV | BG DNCON2 |
|---|---|---|---|---|---|---|
| covalent | 0.45 | 0.31 | 0.33 | 0.1 | 0.21 | 0.26 |
| ionic | 0.43 | 0.29 | 0.32 | 0.11 | 0.24 | 0.29 |
| hydrophobic | 0.31 | 0.32 | 0.34 | 0.16 | 0.23 | 0.28 |
| aromatic | 0.3 | 0.23 | 0.26 | 0.15 | 0.26 | 0.31 |
| vdw | 0.24 | 0.32 | 0.39 | 0.18 | 0.23 | 0.25 |
| vdw_clash | 0.23 | 0.36 | 0.44 | 0.18 | 0.21 | 0.22 |
| hbond_like | 0.22 | 0.34 | 0.43 | 0.19 | 0.22 | 0.23 |
| carbonyl | 0.13 | 0.45 | 0.58 | 0.21 | 0.17 | 0.14 |

Table 3.5: **Raw probabilities of prediction of bond types for background and predicted sets.** Table 1 gives adjusted condititional probabilities for finding a bond in the predicted set, given that it is of a certain type. In this table we give the probability of finding a bond in the background set. These probabilities have not been adjusted for the different average size of the predicted sets, so we would expect the probabilities for predicting each type to be lower for less accurate methods.

in the top $L$ are those which display `covalent` or `ionic` interactions. `carbonyl` interactions are the least likely to be chosen by CCMpred. These results suggest that CCMpred preferentially predicts stronger bond types, once again pointing to CCMpred contacts being more closely related to evolutionary significance.

### Conservation of predicted contacts

In order to further test the role of evolutionary pressure in the formation of evolutionary signal which generates these correlations, we sought to investigate whether the predicted sets were particularly highly conserved in comparison to the background sets. In order to estimate this phenomenon, we compared the extent to which the predicted set of contacts for each case $P$ were present in other members of the same CATH homologous superfamily. For the CATH homologous superfamily in which $P$ occurred, we filtered the homolous superfamily at a 90% sequence identity threshold and then performed structural alignment between every protein

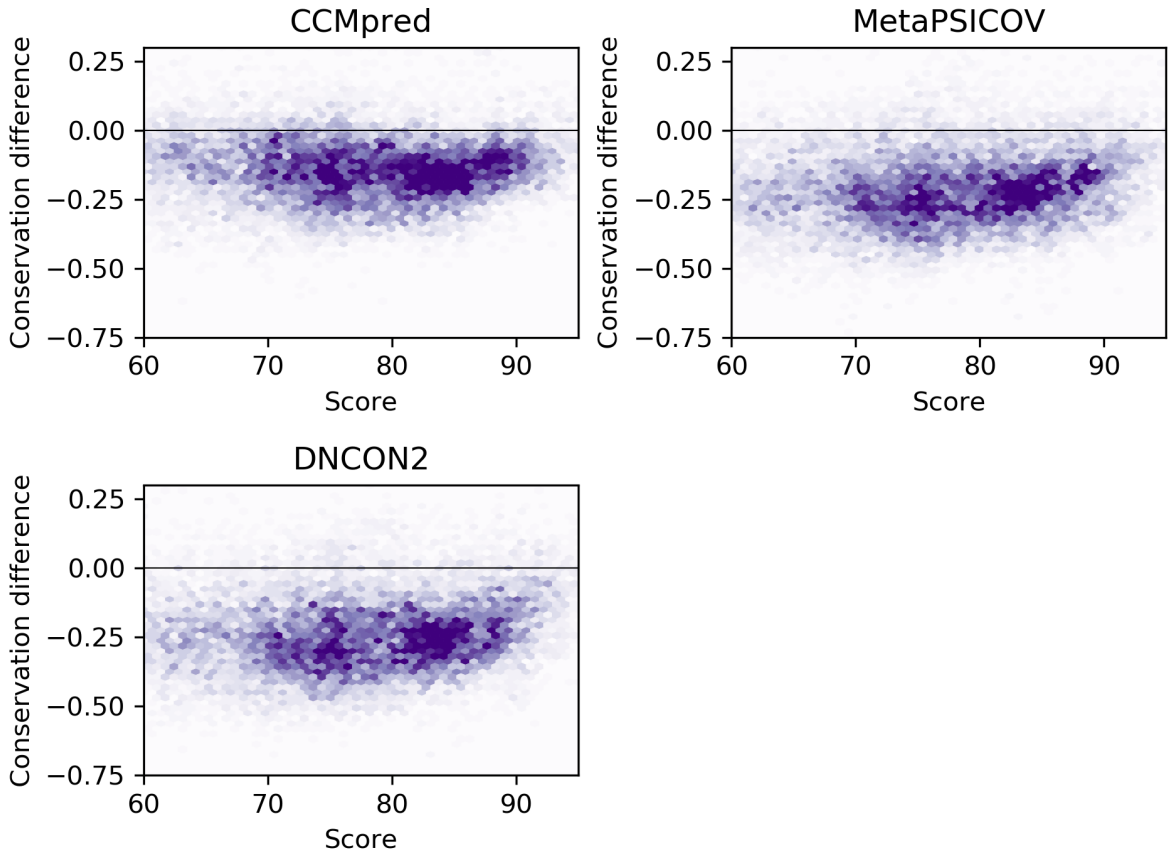


Figure 3.6: **Difference in conservation between predicted set of contacts and background set for different contact predictors as a function of structural dissimilarity.** SSAP structural alignment score is used as a measure of structural dissimilarity. The $y$ axis is backround conservation − predicted conservation.

remaining in the homologous superfamily and $P$. (There were 155 CATH superfamilies which had more than one family member after filtering at 90% sequence identity.) We then recorded the proportion of the contacts in the predicted set of $P$ that were also correct in the aligned family member. We performed the same process for the contacts in the background set. For all three contact prediction methods, the contacts in the predicted sets were more conserved than the background sets for more than 70% of protein-family member pairs (Table 3.6). This excess was present for a range of CATH-SSAP alignment scores and grew as family members became more distant from the exemplar. Fig. 3.6 demonstrates how, as structural relationships become more distant, the predicted set of contacts is more strongly conserved than the background set. This effect is stronger for DNCON2 and MetaPSICOV than for CCMpred.

| **Predictor** | $D \geq 0$ | $D < 0$ | Proportion $D < 0$ |
|---|---|---|---|
| CCMpred | 456 | 5906 | 92.8% |
| MetaPSICOV | 423 | 5939 | 93.3% |
| DNCON2 | 370 | 5992 | 94.2% |

Table 3.6: **Conservation differences between predicted and background sets for different contact predictors.** We compared every protein structure $i$ in the CATH homologous superfamily of each our 863 prediction cases after filtering at a 90% sequence identity threshold. For each protein structure $i$ in the family $J$ of prediction case $i$, we computed the conservation difference $D_{J[i]} \equiv \frac{1}{N_J}\left(\sum_{c\in\text{background set}_J} \epsilon_i^c - \sum_{c\in\text{predicted set}_J} \epsilon_i^c\right)$, the difference between the proportion of conserved contacts in the predicted set and in the background set, where $\epsilon_i^c$ is 1 if contact $c$ is a contact in structure $i$ and zero otherwise. $D_{J[i]} < 0$ indicates a greater level of conservation for the predicted contacts than the background contacts. $N_J$ is the number of homologues in family $J$.

This analysis confirms the centrality of coevolutionary constraints on our ability to predict contacts. Those contacts which are less evolutionarily important and therefore less evolutionarily conserved are more present in the background set than the predicted set. This effect is persistent over the full range of structural similarity scores within proteins. Moreover, CCMpred evinces a lower difference, which varies less as a function of alignment score than the other contact predictors. This difference may originate in CCMpred's comparative bias against secondary structure sites, causing the predicted set to appear to be less strongly conserved than for MetaPSICOV or DNCON2.

## 3.4 Conclusion

As described in the introduction, contact prediction has seen remarkable gains in the accuracy of its predictions and its utility for biological applications over the last ten years. The field of contact prediction has been able to identify larger numbers of contacts, and our results show that this improvement has resulted in changes to the kinds of contacts predicted by state-of-the-art methods. These differences complicate the recent drive to increase prediction accuracy

because not all predicted contacts may be of the same importance. In this chapter, we have placed the differences between predicted and non-predicted contacts in their structural and physico-chemical context.

We found that predicted contacts and background contacts have different properties. Predicted contacts have more bonds than background contacts. For MetaPSICOV and DNCON2, more predicted contacts are within secondary structures than are background contacts. Considering those sets that are uniquely predicted by one contact predictor, these effects are heightened: the unique predictions of CCMpred have more bonds than the unique predictions of MetaPSICOV or DNCON2 and fewer within-secondary structure contacts. CCMpred contacts were more often unique to CCMpred than were MetaPSICOV or DNCON2 unique to those contact predictors. Further, CCMpred contacts were more widely distributed within the protein structures. Contact prediction methods varied in terms of the kinds of bonds that they favoured. These effects throw into relief the relationship between contact prediction and chemical bonds.

Structural constraints that are relevant to the evolutionary history of proteins, and which can be detected in multiple sequence alignments, must be mediated by some kind of physical effect. Our evidence suggests that one component of this effect are physico-chemical bonding interactions, which can be inferred from three-dimensional protein structures. These effects manifest as changes in chemical properties of contact predictions.

If contact prediction is used in the inference of structural properties, such as in the prediction of functional properties, studies of protein mechanism, or simply in structure prediction, future work must take note of the implications for contact type that its choice of prediction method entails. Indeed, in some instances, it may be appropriate to train new approaches on a different definition of contact (*e.g.* physico-chemical interactions, rather than main-chain $C_\beta$ distance).

The accuracy and location of predicted contacts are known to have an important effect on protein structure prediction accuracy. For this reason, a great deal of effort has been dedicated to improving the accuracy of protein contact prediction. However, our data suggest that the raw evolutionary signal of less advanced and less accurate methods may be a source of independently interesting biological information. This approach may be complementary to ever-increasing levels of structure prediction accuracy in helping to unravel the relationship between physico-chemical

structural properties and the formation of protein structures *in vivo*.

# Chapter 4

# Analyzing protein folding pathways in *E. coli* using proteomics data

The work in this chapter was executed with the assistance of Carlos Rubiera Outeiral, who wrote SAINT3, and ran SAINT3 on the protein test set described in the text. He also provided some files to enable Flib2 to be run. Daniel A. Nissley provided parameters for molecular dynamics calibration.

## 4.1 Introduction

Protein folding is a difficult question because of the competing thermodynamic demands that define the protein-folding environment. Proteins must be stable enough to fold, yet unstable enough to be degraded; they must be rigid enough to maintain an active site or binding region, yet flexible enough to permit allosteric structural changes; mutations must maintain structural integrity and efficient function; and many native inter-residue interactions comprise favourable and unfavourable interatomic interactions. The balance between these and other constraints means that precise reckoning of the energy of the protein molecule is essential to reliable physical calculations.

On a physical level, protein folding is a simple expression of the thermodynamics of protein structures, and it is helpful to review the basic principles of this theory. For a particular sequence, three-dimensional configurations $\mathbf{x_i}$ of protein structures are distributed according

to the Boltzmann distribution $P(\mathbf{x_i}) = \exp(E(\mathbf{x_i}/k_BT))$, where $E(\mathbf{x_i})$ is the energy of each configuration.

In almost all cases, there are multiple atomic and molecular states that lead to a particular observed configuration of the protein molecule. For example, a model which incorporates the effect of solvent interactions could admit different solvent configurations that are equivalent in energy and which are labeled by the same protein structural configuration. Then, if two protein structural configurations have the same energy but different numbers of underlying states—each state of which is equally probable—the protein structual configuration with more underlying states will be more likely. Even though the chemical energy of the two structural states are the same, the system is more likely to attain the more-probable structural state, and therefore it behaves as if its chemical energy is lower. This effective chemical energy, produced by the balance of chemical energy (enthalpy) and the number of available states (entropy), is the free energy of the system. In the case of protein structures, for an enthalpy $H$ and entropy $S$, the Gibbs free energy $\Delta G = \Delta H - T\Delta S$ is the relevant quantity.

By considering structual configurations as a function of the Gibbs free energy, the protein-folding problem becomes equivalent to most problems in chemical kinetics. The protein folds from a high-energy, high-entropy configuration (due to a lack of interresidue attractive interactions and a large configuration space populated by unfolded structures) to a low-energy, low-entropy configuration, in which the structural configuration is almost fixed and many attractive interresidue interactions are present. For many such transitions, a free energy *maximum* lies on the reaction trajectory, and it is the crossing time of this barrier that determines the speed of the reaction.

Protein folding takes place in a rough energy landscape. The tendency for proteins to descend toward a global free energy minimum is counteracted by two phenomena. The first is that the free energy minimum generally does not lie far below the free energies of unfolded or misfolded states (Taverna *et al.,* 2002; Goldstein, 2011). The origin of marginal stability is controversial, with simulation studies suggesting that it originates from a paucity of stability-enhancing mutations during natural evolution (Williams *et al.,* 2007; Goldstein, 2011), while a line of experimental studies tend to see a tradeoff between stability and activity (Vogl *et al.,* 1997; Giver *et al.,* 1998).

It is not clear that these interpretations are mutually exclusive. Additionally, since thermal fluctuations necessarily cause random structural changes, these denaturation experiments are most insightful when the folding landscape is uncomplicated.

Additionally, the size of energy barriers between states is often large compared to the energy difference between folded and unfolded states (Goldstein *et al.,* 1992; Onuchic *et al.,* 1997; Hills *et al.,* 2009; Ferreiro *et al.,* 2014). These barriers are sometimes due to the need to backtrack from previously-formed native interactions (Tripathi *et al.,* 2013) but can also be due to the loss of side-chain entropy (Baxa, Haddadian, *et al.,* 2014). However, the fastest-folding proteins fold on a millisecond timescale with no observable barrier (W. Yu *et al.,* 2016).

There is a lack of direct experimental evidence about the trajectory of protein folding. Much experimental work relates to a few small protein domains (Clark *et al.,* 1998; Dill *et al.,* 2008; Baxa, Freed, *et al.,* 2008; Žoldák *et al.,* 2013), which likely exhibit different folding dynamics than larger, more complicated structures. Yet, some proteins are likely to be metastable in their native state (Sohl *et al.,* 1998; Tsutsui *et al.,* 2012), and some degree of fold-switching may be present throughout the genome (Lella *et al.,* 2017; Porter *et al.,* 2018). These considerations underline the need to understand the exact structure of the protein-folding landscape.

We have already reviewed the extent to which the protein folding landscape is influenced by the presence of the ribosome. Since the process of translation and folding are, at least for many proteins, closely linked, the consequences of perturbing cotranslational folding may reveal important features of the folding process. This chapter is driven by a dataset from Philip To and colleagues (To *et al.,* 2020) that assesses the propensity for proteins to fold cotranslationally. As described in Section 4.2 (Methods), this dataset identifies elements of the *E. coli* proteome which must fold cotranslationally, and others that refold readily *in vitro.* These data may provide direct access to the protein-folding trajectories of many proteins in the *E. coli* proteome.

We are interested in two questions. First, do these data provide evidence for the importance of the entropy of loop formation to the protein folding process? The contribution of loop formation entropy in the protein folding process has largely been restricted to molecular dynamics simulations (Baxa, Haddadian, *et al.,* 2014; Gavrilov *et al.,* 2015) because it is transition states that are most likely to be high in entropy (Jacobs *et al.,* 2017), and transition states are not

accessible in studies which focus on energies of stable states. Secondly, do these data provide evidence that the SAINT protein structure prediction methodology can provide insight into the folding trajectory of real proteins?

In this chapter, we first describe our efforts to implement a model which would make explicit the contribution of entropy to the thermodynamics of loop formation. We demonstrate exploratory and preliminary results and describe how they might be extended to use this dataset to study the contribution of loop formation entropy to protein folding trajectories. Then, we explore whether the predictions of SAINT2, and an extension that we call SAINT3, recapitulate the experimental results in this dataset.

## 4.2 Methods

### 4.2.1 Data

To *et al.* have used a mass spectrographic system to systematically identify proteins which do not fully refold after denaturation. *E. coli* lysates were unfolded in urea and GdmCl and refolded by dilution. After two hours, the lysate was probed with a short pulse of proteinase K, cleaving the proteins in the lysate at exposed sites in their structures. The lysate was then fully digested with trypsin. By comparing the fragments identified in the refolded sample with an equivalent mass spectrum identified for *E. coli* lysate which had not been subjected to the refolding procedure, fragments which were unusually abundant or absent in the refolded lysate were identified. These fragments were mapped to proteins in the *E. coli* proteome. Where fragments appear in significantly different quantities in the two samples, it is an indication that the protein is failing to refold to the structure that was present in the initial lysate. We refer to these peptides as 'significant peptides'.

Initial data from these experiments resulted in fragments that could be associated with 1370 genes in *E. coli* found in 1633 PDB structures. After removing genes that were associated with only one fragment, genes whose protein products were approximately larger than 250 amino acids, and proteins that had no identifiable PDB structure, we reduced the size of this set to 323 proteins. Due to the possible influence of complexed proteins on the protein's structure, we then removed proteins that were part of large complexes, by culling those proteins that comprised

less than one third of the total modeled weight of their associated PDB structure. We also removed structures that were obtained by cryo-electron microscopy because of their generally lower resolution and structures under 50 amino acids. After making these changes, our final set comprised 258 proteins for analysis.

As a figure of merit for *in vitro* folding $\mathcal{I}$, we consider the number of significant peptides $N_S$ as a proportion of the number of recovered peptides $N$ and add a pseudocount of 0.1: $\mathcal{I} = (N_S + 0.1)/(N + 0.1)$. When there are no significant peptides and a large number of peptides recovered, this statistic will be close to zero, corresponding to a protein which refolds *in vitro* without observed differences from its native state. When the statistic is far from zero, there are many differences between the *in vitro* and cotranslational fold products, suggesting poor *in vitro* folding.

Introducing a pseudocount confers a particular advantage. Proteins for which a large number of fragments are recovered provide stronger evidence about the nature of their folding than those for which fewer fragments are recovered. This effect is especially important for proteins that may fold entirely successfully in *in vitro*, in which our signal is the absence of significant peptides. Without the pseudocount, proteins where no significant peptides were found amongst few total peptides would be indistinguishable from those where many peptides were recovered and none were significant.

The properties of the 258 proteins used for analysis are shown in Fig. 4.1. The median value of $\mathcal{I}$ was 0.1, corresponding to the presence of at least one significant fragment in 155 of 258 cases. Of the proteins for which at least one significant fragment was found, the median number of fragments recovered was 12, and for those where no significant fragments found, the median number of fragments recovered was 5. Since the number of significant fragments depends on the number of fragments found, it is possible that experimental methodology that recovered more fragments for the less abundant proteins in the dataset would have led to the identification of a greater number of significant fragments. This is an additional reason to introduce statistical procedures that account for uncertainty in the number of fragments recovered, such as the pseudocount we use in $\mathcal{I}$. Fig. 4.1 demonstrates that this set has uniform structural characteristics over a range of values of $\mathcal{I}$.

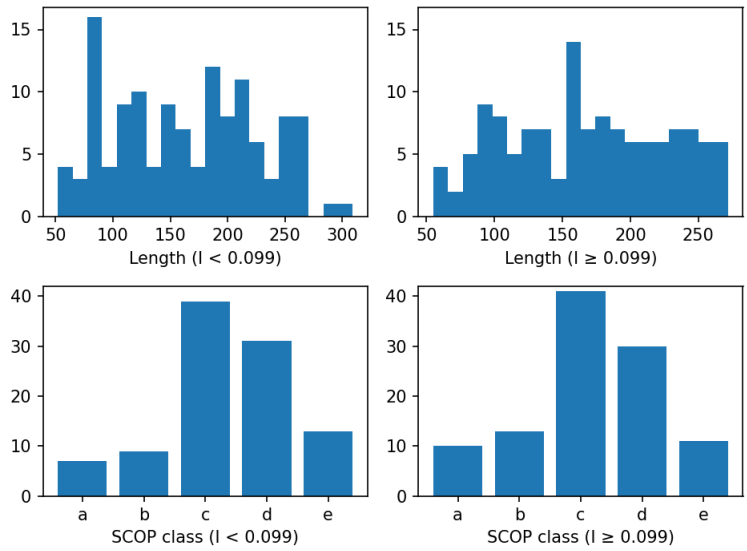


Figure 4.1: **Properties of the proteins identified for folding analysis.** In order to verify that fragment recovery did not depend on basic protein properties, we divided the analysis into high- and low-$\mathcal{I}$ groups at the median value of $\mathcal{I}$. The distribution of SCOP class and length for both groups are similar. The largest constituent SCOP classes (**a**: $\alpha$, **b**: $\beta$, **c**: $\alpha/\beta$, **d**: $\alpha+\beta$, **e**: multidomain) for both groups are **c** and **d** - the $\alpha/\beta$ and $\alpha+\beta$ classes. A significant fraction are multidomain (**e**). For the large-$\mathcal{I}$ group, eight proteins classified as membraine or coiled-coil have been included in the **e** column count. These counts do not sum to 258 because 50 proteins were not annotated in the SCOPe database.

### 4.2.2 Models

**Energetic folding model** We have began by implementing an analysis scheme that follows Jacobs and Shakhnovich (2016). The idea is to classify part-folded states by the presence or absence of secondary structures. The fully-folded state is built up from the ensemble of available secondary structures. Thus, if the secondary structures are *a*, *b*, and *c*, one possible folding sequence is *b*, *ab*, *abc*. Estimation of the height of the energy barriers between each of these states and between the other accessible states (*b*, *c*, *bc*, and *ac*) enables analysis of flow through the network of possible reaction paths. Transition state barriers are calculated using a mean-field theory and an energy derived from native-state contacts. The output of the model is a folding flux between each state and a more advanced folded state.

The central idea of this model is to quantify the conformational entropy loss due to the formation of secondary structures. The first contact to form in a secondary structure, particularly $\beta$ sheets, will typically reduce conformational entropy by a larger amount than further contacts, since loop conformational entropy is approximately proportional to the sum of the logarithms of loop length, and the formation of an isolated contact will result in the division of a larger loop into three smaller loops. However, the addition of a contact to an existing isolated native contact will, for the same reason, only slightly reduce total loop conformational entropy while conferring the energetic benefit associated with the formation of a native contact. Thus, folding would be energetically favourable after the formation of a native contact, but formation of the first native contact would be unfavoured and represent the principal transition barrier between states.

We model the enthalpy as the sum of enthalpies associated with interactions between amino acids in the protein chain. Our initial model uses the two-parameter energy function defined in Jacobs *et al.,* 2016, which incorporates the number $n_{uv}^{c}$ of heavy-atom contacts, whether or not there is a hydrogen bond ($1_{uv}^{\mathrm{hb}}$), and the whether or not the interaction is within an alpha helix:

$$\epsilon_{uv} = \begin{cases} 5/8\left(n_{uv}^{c} + 16 \times 1_{uv}^{\text{hb}}\right) & \text{if } uv \text{ is alpha-helical} \\ \left(n_{uv}^{c} + 16 \times 1_{uv}^{\text{hb}}\right) & \text{if } uv \text{ is not alpha-helical} \end{cases}.$$

Two heavy atoms are considered to be in contact if their interatomic distance is less than 4 Å. We set the chemical potential $\mu = 2$.

Again, following Shakhnovich and Jacobs, we define the loop contact entropy as follows:

$$\frac{\Delta S_L(g)}{k_B} = \sum_{\text{loops } l \in L(g)} \begin{cases} -\dfrac{|l|}{k_B T} & \text{if } |l| \leq b, \\ -\frac{b\mu}{k_B T} - \dfrac{3}{2}\left(\log\left(\dfrac{|l|}{b} + \dfrac{r(l)^2}{b^2|l|}\right)\right) & \text{if } |l| > b \end{cases}$$

$|l|$ is the number of residues in the loop, and $r(l)$ is the number of backbone or inter-residue bonds separating the first and last unbonded residues in the loop. (Here, any inter-residue edge in the contact map is counted as one bond.)

Shakhovich and Jacobs introduce a residue contact entropy $S_R = -\dfrac{\mu}{T}(N_c - 1)$. Thus, the free energy $F(\mathbf{x})$ of each configuration state $\mathbf{x}$ is

$$\frac{F(\mathbf{x})}{k_B T} = \frac{\sum_{u,v} 1_{\mathbf{x}}^{u,v} \epsilon_{uv}}{k_B T} - \frac{S_R}{k_B} - \frac{S_L}{k_B}.$$

**Sampling and analysis under this model** We use Wang-Landau sampling to estimate the free energy of the states. Wang-Landau sampling is standard Metropolis-Hastings Monte Carlo with the addition of a penalty $\xi^i$. If the sampler reaches state with order parameter $i$ at time $t$, the penalty $\xi^i$ is incremented by a value $f_\tau$. Over the course of the sampling, $f_\tau$ is reduced toward zero and the $F_i$ converge to the free energy of the states $i$. We use the $1/\tau$ method of Belardinelli (Belardinelli *et al.,* 2008) to set $f_\tau$, which, unlike other methods, does not require tracking and analysis of the histogram of visited states. Another theoretical advantage of this scheme over related methods is the superior performance of the $1/\tau$ method near to convergence, although we have found that, in practice, our simulations do not reach this point. We attempted a sampling scheme that sampled each state separately before uniting the free energies with a further round of Wang-Landau sampling, but we found that this scheme was inefficient and inaccurate.

At the conclusion of Wang-Landau sampling, the sequence of states should visit all values $i$ of the order parameter with approximately equal frequency. Therefore, standard Metropolis-Hastings sampling using $\xi^i$ as a biasing potential should also visit all order parameters $i$ with approximately equal frequency.

Algorithm 4.1: Wang-Landau sampling

***input***

$f_{\text{start}} \leftarrow 4$

$f_{\text{end}} \leftarrow 5\text{e}{-5}$

$t \leftarrow 0$

steps $\leftarrow$ 20000

$\xi^i \leftarrow 0$

N $\leftarrow$ length of the protein sequence

***output***

states $\mathbf{x}_i$

***begin***

initialize state $\mathbf{x}$

***while*** $f > f_{\text{end}}$ :

$\tau \leftarrow \tau + 1$

$f \leftarrow 3N/\tau$

$\mathbf{x}_{\text{trial}} \leftarrow p(\mathbf{x})$

$\text{A} \leftarrow \exp\left((F(\mathbf{x}) - F(\mathbf{x}_{\text{trial}})) + (\log\left(P(\mathbf{x}_{\text{trial}} \rightarrow \text{x})\right) - \log\left(P(\text{x} \rightarrow \mathbf{x}_{\text{state}})\right)) + \left(\xi^{\mathbf{x}_{\text{trial}}} - \xi^{\text{state}}\right)\right)$

***if*** r $<$ A:

$\mathbf{x}_\tau \leftarrow \mathbf{x}_{\text{trial}}$

***else***

$\mathbf{x}_\tau \leftarrow \mathbf{x}$

$\xi^{i(\mathbf{x})} = \xi^{i(\mathbf{x})} + f$

***end***

***end***

The Wang-Landau penalties $\xi^i$ are related to the free energies of each configuration $i$ : $(X, t)$ by $F_i/k_BT = \max_j \xi^j - \xi^i$. The average free energy $\overline{F_t}/k_BT$ of each topological configuration $\overline{F_t}/k_BT = -\log \sum_X \exp(-F_{(X,t)}/k_BT)$.

In order to calculate the folding flux from topology $s$ to topology $t$, we treat the problem using the theory of Markov reaction pathways and we proceed by calculating

- The proportion of sampled states of each value of order parameter $X$ and topology $s$ for for which residue $v$ is active: $\langle \mathbf{1}_v \rangle_{(s,X)}$.

- The equilibrium probabilities $\pi_t$ of each topological configuration $t$
- The energy barriers $\dfrac{\Delta F^{\dagger}_{s\to t}}{k_B T}$ between topological configurations, and thus the equilibrium constants $k_{st}$
- The 'forward committor' $P_{\text{fold}}(s)$, which is the probability that a reaction trajectory reaches $s$ and then the fully folded state before reaching the fully unfolded state.

The equilibrium probabilities $\pi_t$ are given by $\pi_t = \exp(-\overline{F_t}/k_B T)/\sum_{t'} \exp(-\overline{F_{t'}}/k_B T)$. The energy barriers $\dfrac{\Delta F^{\dagger}_{s\to t}}{k_B T}$ are the sum of the barriers from $s$ to $t$ at each order parameter $X$:

$$\frac{\Delta F^{\dagger}_{s\to t}}{k_B T} = -\log \sum_X \exp\left(\frac{-\Delta F^{\dagger}_{(s,X)\to(t,X+1)} - (F_{(s,X)} - \overline{F_s})}{k_B T}\right).$$

The values $\dfrac{\Delta F^{\dagger}_{(s,X)\to(t,X+1)}}{k_B T}$ are calculated according to the following algorithm.

Algorithm 4.2: Barrier calculations

***input***
    states $\mathbf{x}_i$
    order parameter $X$
    topology $s$
    topology $t$
    weight = 0
    contribution = 0
    $H_{u=1\ldots X} = 0$
***output***
    Free energy barrier $\Delta F^{\dagger}_{(s,X)\to(t,X+1)}/k_B T$
***begin***
    ***foreach*** residue $u$ ***if*** $u$ contributes to $s$:
        ***foreach*** state $\mathbf{x}_i$
            ***if*** $\mathbf{x}_i$ has configuration $s$ ***and*** order parameter $X$:
                $S_1 = S_L(\mathbf{x}_i)$
                $S_2 = S_L(\mathbf{x}_i + u)$
                contribution += $(S_2 - S_1)\exp(-F_{(X,s)})$
                weight += $\exp(-F_{(X,s)})$
        $\left\langle \dfrac{\Delta S_u}{k_B} \right\rangle = \dfrac{\text{contribution}}{\text{weight}}$

        ***foreach*** residue $v$ ***if*** $v$ contributes to $t$ but not $s$:
            $H_u \mathrel{+}= -\dfrac{\epsilon_{uv}}{k_B T} \times \langle \mathbf{1}_v \rangle_{(s,X)}$

23 $\frac{\Delta F^{\dagger}_{(s,X)\to(t,X+1)}}{k_B T} = -\log \sum_{u=1\ldots X} \exp\left\langle \frac{\Delta S_u}{k_B} \right\rangle \times (\exp H_u - 1)$

24 ***end***

---

Following Metzner *et al.,* 2009, the equilibrium constants

$$k_{st} = \begin{cases} \min\left(1, \exp\left(-\frac{\Delta F^{\dagger}_{s\to t}}{k_B T}\right)\right) & \text{if } s,t \text{ adjacent,} \\ -\sum_{t' \neq s} k_{st'} & \text{if } s = t, \\ 0 & \text{otherwise.} \end{cases}$$

The forward committor $P_{\text{fold}}(s)$ is calculated by solving the linear system

$$\sum_t k_{st} P_{\text{fold}}(s) = 0,$$

for all $s$ which are not the fully unfolded and fully folded states, which have $P_{\text{fold}}(s)$ respectively equal to 0 and 1. Then, the reactive flux for every $s \neq t$ is $f(s \to t) = \pi_s(1 - P_{\text{fold}}(s))k_{st}P_{\text{fold}}(t)$ and the net flux is $f^{+}_{st} = \max(f_{st} - f_{ts}, 0)$.

**Fragment libraries and SAINT** We also investigated whether the SAINT2 cotranslational folding model could provide insight into these phenomena. SAINT2 was configured according to its defaults. Decoys for each sequence were generated using 10,000 Monte Carlo extension steps and 1,000 further steps in the cotranslational mode, and for 11,000 Monte Carlo steps in *in vitro* mode. All cotranslational runs used the straight extension protocol described in Chapter 2. Fragment libraries were generated with Flib2, an unpublished reimplementation of the Flib protocol developed by Carlos Rubiera Outeiral. This program is faster than Flib, and incorporates several bug fixes, but otherwise follows the original protocol. We used PSI-BLAST with default settings to detect homologs of these proteins in the PDB. These sets were also used as an input to Flib fragment selection.

Initially, we configured Flib2 to output 20 fragments per site using the standard protocol. However, in order to improve fragment selection and guarantee high-quality fragments at each site, we then adopted the following protocol. After creating large fragment libraries (more than 200 fragments per site, including those from homologs and those not) we sampled twenty fragments at each site using Markov Chain Monte Carlo from a chi-squared

distribution of fragment RMSD to the native structure. After initial testing, we used a mean of 0.9 for this distribution, although at most sites it resulted in the twenty best fragments being selected (see Results).

We also used SAINT3, which is a similar cotranslational-extension/fragment-replacement software to SAINT2, but it introduces a short molecular dynamics simulation between each fragment replacement. Fragment replacement takes place in a full backbone representation, which is then coarse-grained to the $C_\alpha$ representation for molecular dynamics runs. In cotranslational mode, each extrusion is followed by 50,000 15-femtosecond time steps using Langevin dynamics within the Betancourt-Thirumalai potential (Betancourt *et al.,* 2008) and the backbone coordinates were re-generated using PULCHRA (Rotkiewicz *et al.,* 2008). Then, ten fragment replacement steps were executed under the following procedure. Fragments were drawn as in SAINT2, following which 10,000 molecular dynamics steps (0.15 nanoseconds total) were run. After these molecular dynamics steps, the new structure was compared to the previous structure and accepted or rejected using the standard SAINT2 Monte Carlo criterion. In the *in vitro* mode, 1000 fragment replacement/molecular dynamics steps are executed on the extended protein chain using the same protocol as described above. Values of the Betancourt-Thirumalai free parameter for each protein were obtained with a linear model trained on protein physical properties. We then verified that molecular dynamics starting from the native state remained within 6ÅRMSD in 99.8% of MD frames for 1 microsecond (Leininger *et al.,* 2019). The precise size of the 6Åcutoff is relatively unimportant: stable MD models normally stay well within this bound, while those that are not stable cross it quickly. This criterion ensures that the forces introduced through the MD model are unlikely to lead to artefactual results during the MD steps of the SAINT3 protocol.

In both cases and for each protein under consideration, we built hundreds of decoys in SAINT2 in cotranslational and in *in vitro* modes. The precise number of decoys varied between analyses, depending on predicted run-time and time available, but was greater than 50 in all cases and greater than 200 for SAINT2. SAINT2 runs were analyzed with respect to the proportion of decoys with $TM \geq 0.5$, as described in Chapter 2. As SAINT3

reports much more accurate decoys, we considered the median TM-Score for the sampled decoys.

## 4.3 Results and discussion

### 4.3.1 Energetic folding model

As the number of states is combinatoric in the number of substructures, the sampling time for the Wang-Landau sampler increases quickly for proteins with especially complex secondary structures. As a consequence, we were able to run the sampler only for the 86 smallest proteins. Of these, we experienced sampling difficulties for the majority of cases, resulting in predicted flux of 0 to the final folded state through the pathway that most closely followed the activation of residues cotranslationally. (In other words, if the residues of the protein were activated sequentially from residue 1 to the final residue, we are referring to the order in which the substructures would be activated.) This issue was likely due to the inaccessibility of one of the penultimate folded states by the sampler. Although we attempted to introduce an alternative state proposal scheme, these issues persisted and we reverted to the earlier proposal scheme because it was simpler.

Fig. 4.2 shows the results from this investigation. We compare the folding flux through the cotranslation pathway, as described above, and the most efficient pathway in the network. Each pathway is ranked according to its rate-limiting flux, *i.e.*, the lowest flux between folding states. It appears possible that at low $\mathcal{I}$, folding flux through the cotranslational pathway is much lower than flux through the unconstrained pathway, suggesting that the unconstrained pathway, through which the protein would be more likely to fold *in vitro*, is much more efficient than the cotranslational pathway in these cases. It would therefore be unsurprising that these proteins would be the ones that are the most efficient *in vitro* folders in the experimental data. However, there are only a few points and most are from the shortest proteins in the set. More simulation data is needed to confirm this trend.

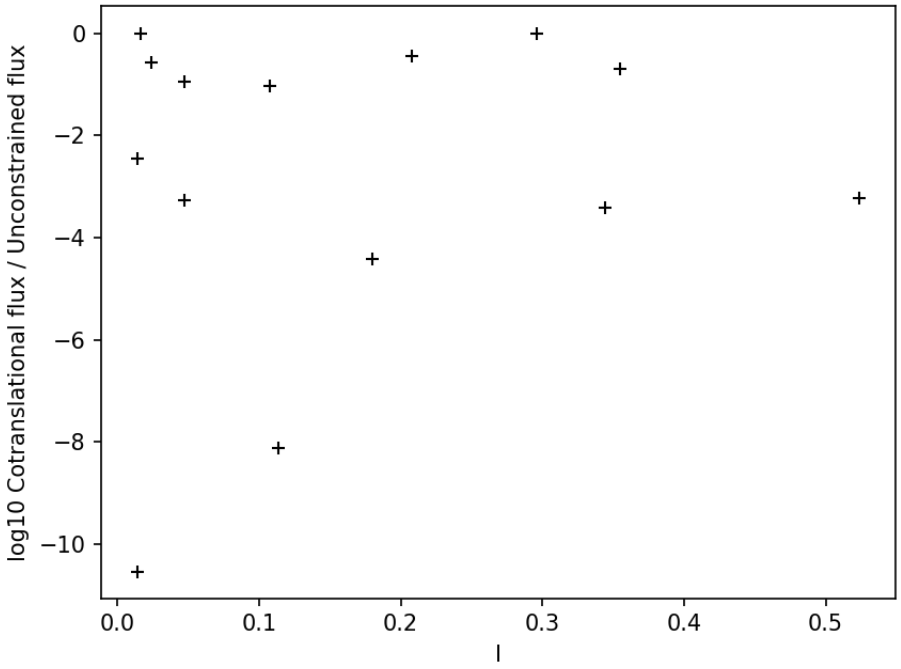


Figure 4.2: **Comparative efficiencies of the cotranslational and unconstrained pathways as a function of $\mathcal{I}$ as predicted by the energetic folding model.** The $x$-axis is the figure of merit $\mathcal{I}$, and the $y$-axis is the logarithm of the ratio of folding fluxes through the cotranslational pathway and the best pathway in the modelled folding network.

### 4.3.2 SAINT2

We then investigated the ability of SAINT2 to predict these effects. We began by running standard SAINT2 with default Flib settings and examined SAINT2's ability to predict our set of structures. Fig. 4.3 demonstrates that the cotranslational mode was consistently better than the *in vitro* mode. However, only a minority of prediction targets had TM-Score $\geq 0.5$, and since the set of informative results would be small, we investigated ways to increase prediction quality.

We suspected that improving fragment accuracy would cause prediction quality to improve. As described in Methods, we allowed fragments from proteins that were homologous to the protein targets to be included in the library. Using a Markov Chain Monte Carlo procedure to force the distribution of fragments at every alignment site to approximate the chi-squared distribution with mean 0.9 closely as possible, we sampled libraries of over 200 fragments per site into libraries of 20 fragments per site. However, the limiting factor was the availablity of accurate fragments. Fig. 4.4 demonstrates that although there was an effect of fragment accuracy on prediction results, most fragment libraries had very good fragment accuracy and the effect of median fragment RMSD was not very strong. Here, the 95th percentile of TM-Score is used because it is close to the maximum TM-Score but more statistically stable, allowing better estimates with fewer datapoints. The $x$-axis—the median of the median fragment RMSD—is the median value over the entire protein of the median fragment RMSD at each site. Without additional options to improve the fragment RMSD, we decided to investigate the improvement due to cotranslational folding using our existing libraries.

Fig. 4.5 demonstrates that there is no clear effect of $\mathcal{I}$ on cotranslational folding improvement. It is possible that the data indicate a slight decrease in the advantage of cotranslational folding for large values of $\mathcal{I}$. This effect might suggest that our cotranslational folding procedure is most advantageous when the protein targets are 'easier'—when they fold consistently in a natural setting. Fig. 4.6 demonstrates further that this result is not due to a length-related bias in our data, which we felt was the most likely source of SAINT2 bias. In order to gain more insight into these results, we were able to run these experiments again, integrating molecular dynamics, using SAINT3.

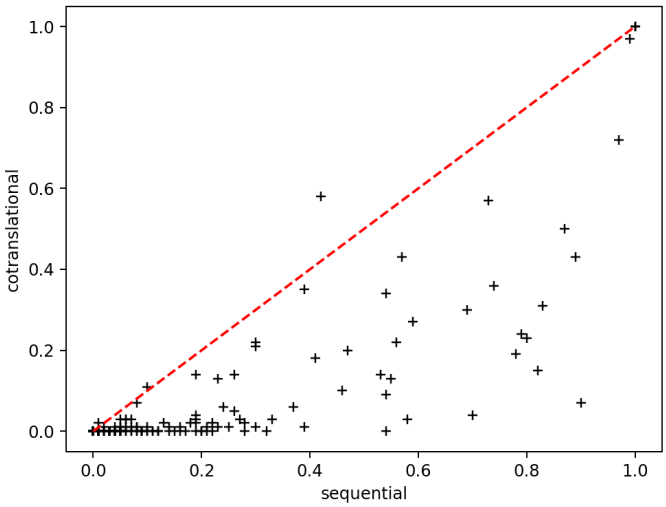


Figure 4.3: **Grishin plot of structure prediction improvement due to the SAINT2 cotranslational mode for the test set of 258 proteins.** The proportion of TM-Scores above 0.5 is plotted on both axes, and the red line indicates equivalence between the two methods.

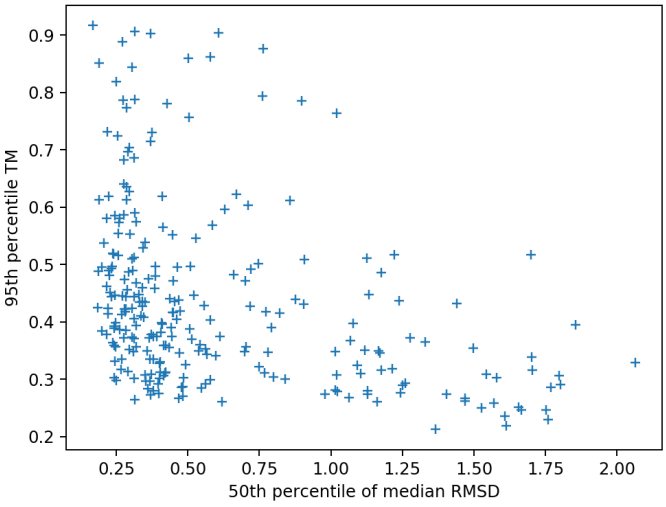


Figure 4.4: **TM-Score for cotranslational folding as a function of median fragment RMSD.** 250 decoys were generated for every protein in the test set. Most proteins have few decoys with TM-Score above 0.5, and these are concentrated at low fragment RMSD values.

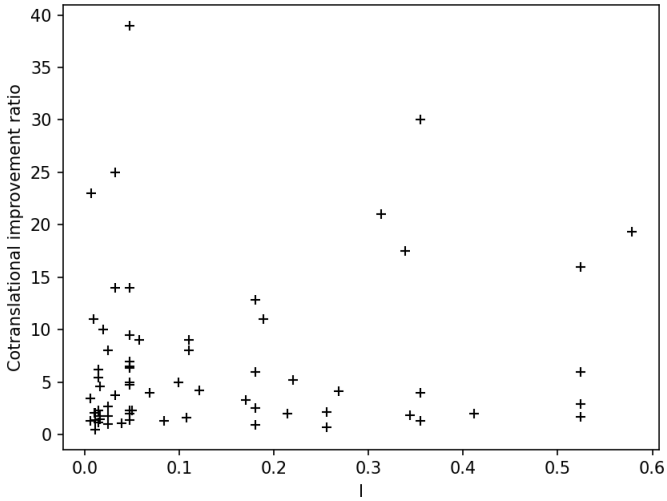


Figure 4.5: **Structure prediction improvement using cotranslational folding as a function of $\mathcal{I}$.** This figure uses a run of 100 decoys for each protein target, and does not show proteins where *in vitro* or cotranslational extension delivered either 0 or 100 decoys with $\mathrm{TM-Score} \geq 0.5$. These proteins were removed because they are less informative about folding difficulty in a relative sense because they are at the maximum or minimum value of our folding difficulty measurment.

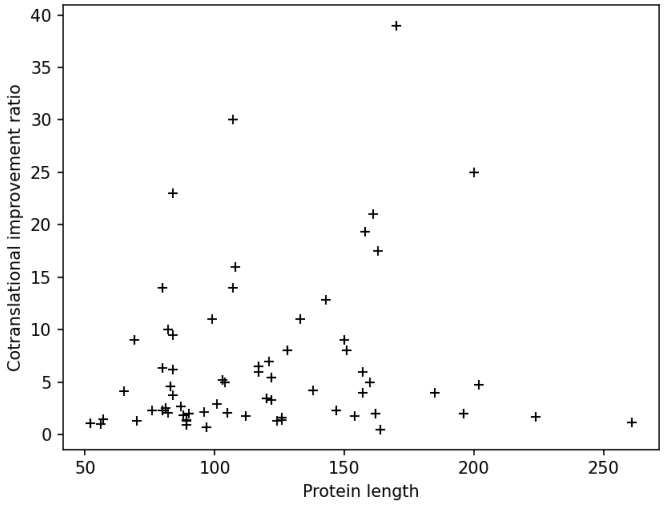


Figure 4.6: **Structure prediction improvement using cotranslational folding as a function of protein length.** Like the preceeding figure, this figure uses a run of 100 decoys for each protein target and does not show proteins where *in vitro* or cotranslational extension delivered either 0 or 100 decoys with $\mathrm{TM-Score} \geq 0.5$.

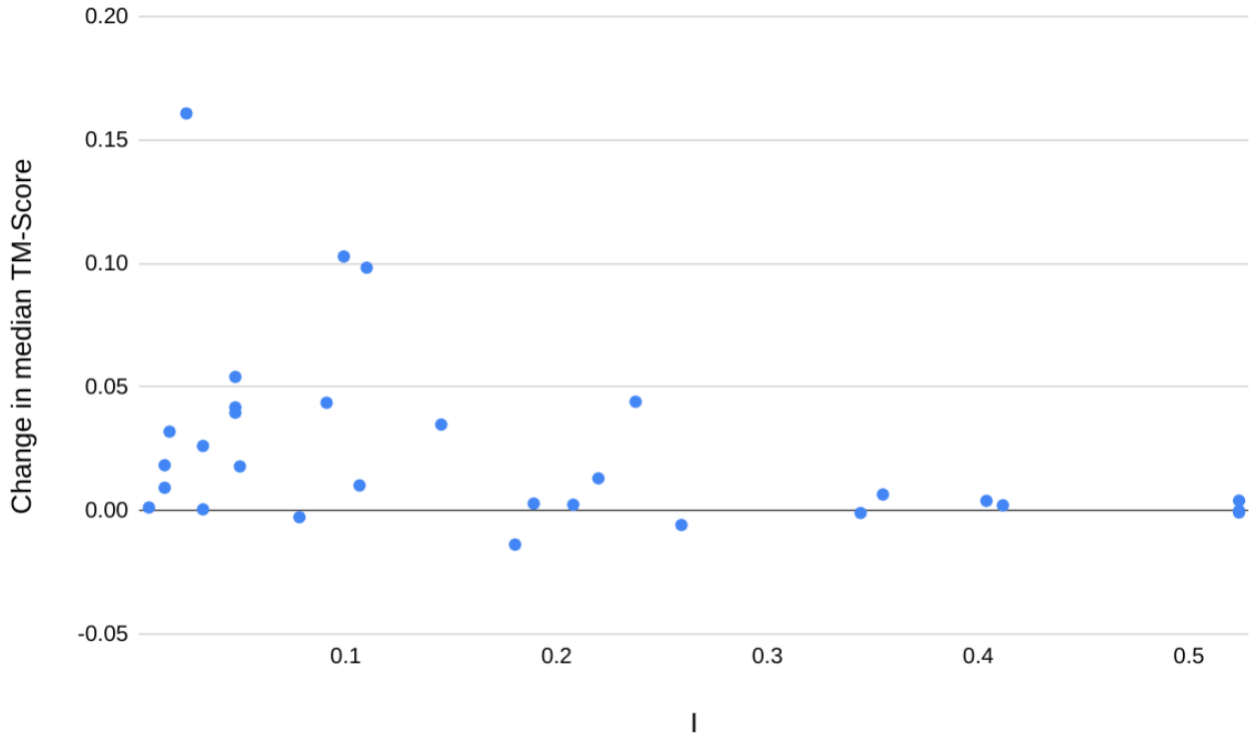


Figure 4.7: **Cotranslational folding improvement in SAINT3 as a function of the experimental level of inhibition of refolding.** The $x$-axis is the figure of merit $\mathcal{I}$, and the $y$-axis is difference between the median TM-Scores in cotranslational and *in vitro* folding using SAINT3. The Spearman's rank correlation coefficient is -0.4.

### 4.3.3 SAINT3

SAINT3 delivered results which were more clear. There is a weak but consistent trend in the data that suggests that the cotranslational folding mode is most helpful for those proteins which refold consistently *in vitro*. The reason for this is not clear, but it may suggest that the cotranslational folding mode is able to take advantage of unambiguous structural energy minima more effectively than the *in vitro* mode. In this sense, we would be able to identify those proteins which fold most consistently *in vitro* because they are the most stable proteins overall, and not necessarily because of a commonality between the simulated and real folding mechanism.

## 4.4 Conclusion

In this chapter, we have investigated ways of assessing the protein folding landscape in order to better understand the protein folding pathways associated with *in vitro* refolding and cotranslational folding, which has led to a collection of preliminary results which are in tension. There is a weak indication from the energy-landscape analysis that the cotranslational pathways are relatively the weakest for those proteins which readily refold *in vitro*. We speculate that these may also be the proteins which are the most stable, leading SAINT2 and SAINT3 to exploit this stability disproportionately in its cotranslational mode. It would not be surprising that a better search algorithm—cotranslation—would be more able to identify an energy minimum, and equally it would be no surprise, though not inherent, that many proteins which fold reliably are also those which fold to the deepest energy minima.

Next, it would be pertinent to investigate the internal energetic dynamics of the SAINT3 folding trajectories in order to understand better how the proteins which have been assisted disproportionately by the cotranslational mode are achieving those structures. This type of analysis might enable us to comment directly on the location of kinetic traps and the shape of the energy landscape. We anticipate more data becoming available for the SAINT3 and energy landscape analyses, which will help to confirm or refute these trends. Finally, it is possible that these methods could be adapted to make predictions on the basis of the patterns we have identified.

# Chapter 5

# Discussion and conclusion

"The infinite had two levels," writes Gilles Deleuze: "the coils of matter, and the folds of the soul" (Deleuze *et al.,* 1991). This thesis is concerned with the former. Although the last decade has revolutionized protein structure prediction, protein folding remains a complex and difficult problem. The problems in addressing protein folding remain several-fold: a paucity of time-course data, relating primarily to a narrow set of small proteins in artificial environments—or in simulations which may represent our existing biases—along with computational techniques that only poorly represent the underlying physics. In this environment, we have sought to expand our understanding of protein folding by improving models of interactions within proteins.

## 5.1 Three-dimensional constraints on protein folding pathways

The experimental evidence for cotranslational folding across protein families and the domains of life suggests that an account of protein folding which omits cotranslational folding will be incomplete. By definition, cotranslational folding takes place in the environment of the ribosome, and the structural restraints that the ribosome imposes have been shown to affect the protein-folding pathway in multiple model systems (Kolb, 2001; Mercier *et al.,* 2018; Liutkute *et al.,* 2020). In many model systems, $\alpha$-helices fold in the ribosome exit tunnel, with small folds occurring in the later stages of the exit tunnel, and larger-scale compaction occurring outside the exit tunnel.

In the first chapter, we investigated a model of protein folding that introduced an explicit

spatial model of the ribosome in order to evaluate and exploit these spatial constraints. In particular, we modified SAINT2 to incorporate explict spatial constraints and executed fragment-replacement protein structure prediction with extension in the presence of those constraints. These constraints tended to cause the resulting models to correspond less well to their known tertiary structures as determined by experiment, suggesting that introducing spatial constraints in protein structure prediction in this way is at best neutral (in the case of the infinite half-plane) and at worst deleterious (in the case of our model of the exit tunnel).

These conclusions are limited in scope. First, the models of the ribosome that we implemented involved hard step-like boundaries, which may have led to favourable moves being rejected because they slightly intersected the ribosome or tunnel wall. This would have led to effectively fewer net moves being executed on the extrusion models, which would have resulted in lower-quality models indepenent of any chemical or physical effects. Additionally, since our approach favours the formation of $\alpha$-helices where they are known to occur, and since the ribosome tunnel may be an environment which promotes the formation of $\alpha$-helices, we may have been unable to identify the true benefit of a ribosome tunnel environment by subsuming its functionality into our simulation methodology. But, by contrast, we also observed the introduction of hairpin turns into our structure prediction trajectories inside the exit tunnel, which were probably unphysical because they would have required large-scale translocations of extruded peptide into the tunnel. And, finally, our model took no account of the chemical environment within the tunnel or on the ribosome surface, which may affect the folding trajectory of the peptide, for example, by allowing the nascent chain to adhere to the surface of the ribosome or the interior of the exit tunnel. Therefore, this negative result ought to spur more sophisticated models to fully elucidate the extent to which modeling the ribosome might improve protein structure prediction and our understanding of protein folding.

This work resulted in the identification and improvement of one part of our structure-prediction pipeline. We noticed that the default SAINT2 extension algorithm would cause unphysical overlaps in the peptide structure, causing the energy of the protein to become temporarily very high. As a result, our Markov Chain Monte Carlo would accept almost any move which would resolve this conflict and divert the model from a genuine cotranslational trajec-

tory. Therefore, we introduced a sampling technique which ensured that the sampler would not introduce excursions in the protein energy. As a result, prediction accuracy improved. This effect is evidence for the utility of cotranslational folding in protein structure prediction because diversions from a cotranslation-like prediction trajectory resulted in worse predictions than the improved those that avoided those diversions.

## 5.2 Identification of coevolutionary constraints on protein structures

In order to better understand the coevolutionary analyses that underpinned most modern structure prediction methods, we undertook a large-scale survey of coevolutionary interactions in protein structures. We identified a set of 2,086 structures which represented different CATH superfamilies. After removing structures wth poor alignment quality, we predicted contacts for the remaining 1,030 cases and removed those with poor contact prediction accuracy, leaving 863 cases remaining. We adopted a comparative methodology, in which the top correct contacts were compared with the an equivalent number of random contacts, in order to control for the structural properties of each protein and illuminate the unique properties of those contacts that had been predicted.

These methods resulted in several insights. First, we were able to show that contacts which were predicted by the methods which relied on machine learning, both deep learning and metaprediction, were much more likely to predict contacts that were found in secondary structures than those methods which were designed only to quantify the strength of the coevolutionary pressure on pairs of sites in the dataset. Despite this bias toward secondary structures, the machine learning predictors identified slightly fewer bonding interactions than those that relied on raw coevolutionary information and exhibited a lower degree of structural diversity. However, the contacts identified by machine learning predictors appeared to be modestly more conserved than those identified due to coevolutionary signal alone, though all predicted contacts were more likely to be conserved than contacts which were not predicted. The difference between the two methods probably reflects the fact that secondary structures, in which machine learning predictors identify more contacts, are more likely to be conserved than loop regions.

Additionally, they are more likely to be able to be aligned using structural techniques than loop regions, which would result in artefactually greater levels of apparent conservation.

A straightforward explanation exists for these results: training machine learning models on $C_\beta$ contacts leads them to emphasise low-risk contacts implied by secondary structures and other recognizable patterns, and to remove or deprioritise 'risky' contacts which are isolated from these structures, supported only by cotranslational signal. Thus, the number of correct contacts increases, while contacts which are structurally or chemically important by virtue of their isolation or structural unpredictability are less likely to be reported. This may represent an opportunity for structure prediction algorithms that use contact predictions as input to extract greater amounts of information from their inputs by identifying these locations or otherwise combining different sources of contact predictions in order to extract the maximum amount of information that is relevant for structure prediction.

Some groups have begun incorporating techniques like these into their structure predictions. AlphaFold and its open-source copy ProsPR (Senior *et al.,* 2020; Billings *et al.,* 2019), among others, have used the 400 inter-amino acid coevolutionary channels as direct inputs to their force-field model, allowing correlations which arise between the channels, as well as between residues, to inform the folding process. And, more simply, a colleague reports (S. H. P. de Oliveira, private communication, 2017) that the contacts predicted by MetaPSICOV in its 'first stage'—before further refinement—was more effective for protein structure prediction than the 'second stage' predictions, possibly due to the inclusion of a greater diversity of contacts.

By focusing on the the utility of coevolutionary analyses for contact prediction, we may also ignore their utility as independent sources of information about protein structure. The evolutionary context that gives rise to coevolutionary couplings between protein sites is a reflection of the physical and biological constraints to which the protein is subject.

## 5.3 Protein-folding modeling with large-scale data

New approaches to the collection of experimental data, including the exploitation of proteomic techinques in mass spectrometry, will give rise to new insights about protein folding mechanisms. In Chapter 4, we used a dataset that measures the extent to which cotranslational folding is nec-

essary in all proteins in the *E. coli* proteome. Although these analyses are continuing, interesting preliminary results suggest that these data can be related to our SAINT structure prediction pipeline, as well as the predicted protein-folding fluxes from a native secondary structure-based energy-landscape analysis method.

In the case of SAINT, we found a negative correlation between refolding inhibition ($\mathcal{I}$) and the extent to which cotranslational folding was an improvement over the *in vitro* folding procedure. This effect was visible in SAINT2 and developed further in our SAINT3 analyses. If cotranslational folding is better able to exploit properties of the energy landscape, and the proteins which refold most easily are those that have the greatest level of energetic separation between folded and unfolded states, then it is plausible that these results reflect general properties of the SAINT analysis pipeline.

Our experiments with direct calculations on the energy landscape suggested a similar idea: those proteins that are most likely to refold were those that had the best-developed folding non-cotranslational folding pathway. However, these analyses suffered from sampling issues which we continue to work to resolve. It is our hope that the resolution of these issues will lead to the inclusion of more proteins in these analyses.

The data themselves are powerful because they reveal the widespread extent to which co-translational folding is obligate in the *E. coli* proteome. Thus, they indicate which proteins fold into kinetic traps outside of the ribosome. Eventually, tracing the location of these traps, even those that are reflective of non-native structure, will enable new understanding of the way that proteins fold *in vivo*.

## 5.4 Conclusion, future work, and outlook

This thesis examines the interface between the natural constraints on protein folding and the physical properties of the amino-acid sequence that lead to structure formation. In particular, we have considered the factors leading to cotranslational folding and the effect of evolutionary constraints on structure formation and the amino-acid sequences of proteins.

The data presented here lead to several paths of further investigation that may be fruitful. It is likely that our models of the ribosome could be improved, for example, by preventing

within-tunnel hairpins, introducing spatial constraints that are not fully 'hard', and accounting carefully for the number of effective Monte Carlo steps that elapse under different physical models. These simulations might be more fruitful in an MD context, such as that offered by SAINT3. Secondly, our investigation of contact properties could be made more sensitive by introducing better information about the importance of individual contacts, perhaps by relating them to allosteric interactions or MD simulations of protein structures. And, finally, we used a rich dataset relating to the cotranslational propensity of the *E. coli* proteome which merits further exploration. Our free-energy sampling implementation would be an excellent target because it directly limited our ability to draw conclusions about the proteins under investigation. There is also scope to study the trajectories of our SAINT3 simulations in order to see whether kinetic traps can be identified, perhaps relating the proteins which failed to refold cotranslationally to the structural characteristics of their folding trajectories.

As protein structure prediction continues to improve, it is likely that questions about protein folding, which suffer from worse raw data and which are less well-posed than the structure prediction problem, will take on new prominence. Understanding the formation of protein structure is essential to the utilization of protein products for industrial or pharmaceutical purposes, for example, by enabling engineers to disrupt aggregration and increase protein production. It will also help disentangle the molecular forces that drive protein structure formation from thermodynamic effects, perhaps enabling better simulation of protein dynamics for medical research. We hope that this thesis will in some way enable those discoveries.

# Chapter 6

# Appendix A: Ribosome occupancy profiles are conserved between structurally and evolutionarily related yeast domains

This paper, the analysis of which I helped to design and in the initial stages I supervised, is included here rather than in the main body of the text to reflect the lower level of contribution I made to it. It follows similar themes to the rest of the thesis, describing the role of cellular biology in the evolved translational environment *in vivo*.

*Bioinformatics*, 2021, 1–7
doi: 10.1093/bioinformatics/btab020
Advance Access Publication Date: 23 January 2021
Original Paper



Structural bioinformatics

# Ribosome occupancy profiles are conserved between structurally and evolutionarily related yeast domains

**Daniel A. Nissley , Anna Carbery , Mark Chonofsky and Charlotte M. Deane ***

Department of Statistics, University of Oxford, Oxford OX1 3LB, UK

*To whom correspondence should be addressed.
Associate Editor: Alfonso Valencia

Received on September 26, 2020; revised on December 11, 2020; editorial decision on January 6, 2021; accepted on January 12, 2021

## Abstract

**Motivation:** Protein synthesis is a non-equilibrium process, meaning that the speed of translation can influence the ability of proteins to fold and function. Assuming that structurally similar proteins fold by similar pathways, the profile of translation speed along an mRNA should be evolutionarily conserved between related proteins to direct correct folding and downstream function. The only evidence to date for such conservation of translation speed between homologous proteins has used codon rarity as a proxy for translation speed. There are, however, many other factors including mRNA structure and the chemistry of the amino acids in the A- and P-sites of the ribosome that influence the speed of amino acid addition.

**Results:** Ribosome profiling experiments provide a signal directly proportional to the underlying translation times at the level of individual codons. We compared ribosome occupancy profiles (extracted from five different large-scale yeast ribosome profiling studies) between related protein domains to more directly test if their translation schedule was conserved. Our analysis reveals that the ribosome occupancy profiles of paralogous domains tend to be significantly more similar to one another than to profiles of non-paralogous domains. This trend does not depend on domain length, structural classes, amino acid composition or sequence similarity. Our results indicate that entire ribosome occupancy profiles and not just rare codon locations are conserved between even distantly related domains in yeast, providing support for the hypothesis that translation schedule is conserved between structurally related domains to retain folding pathways and facilitate efficient folding.

**Availability and implementation:** Python3 code is available on GitHub at https://github.com/DanNissley/Compare-ribosome-occupancy.

**Contact:** deane@stats.ox.ac.uk

**Supplementary information:** Supplementary data are available at *Bioinformatics* online.

## 1 Introduction

Many protein domains acquire their native structure during synthesis by the ribosome through a process known as co-translational folding (Holtkamp *et al.*, 2015; Nicola et al., 1999; Nissley and O'Brien, 2014; Thommen et al., 2017). Folding during synthesis is intuitively beneficial in that it allows N-terminal sections of proteins to begin acquiring tertiary structure before synthesis of the full-length protein is complete (Frydman *et al.*, 1999). Vectoral folding of this nature helps avoid misfolded conformations (Frydman *et al.*, 1999) and thus leads to more efficient folding of the proteome. Changes to translation speed disrupt protein folding (Cortazzo *et al.*, 2002; Zhang et al., 2009) and function (Noriega et al., 2014; Walter and Johnson, 1994; Zhou et al., 2013) and are thought to be a causal factor in several human diseases including cystic fibrosis (Kim et al., 2015), certain cancers (Sauna and Kimchi-Sarfaty, 2011) and a type of haemophilia (Knobe et al., 2008). Co-translational folding is thus a key part of proteostasis, and its perturbation may lead to the accumulation of misfolded proteins, inducing proteotoxic stress (Nissley and O'Brien, 2016). Given that translation speed influences protein folding and function, it is natural to hypothesize that these speeds are evolutionarily conserved between proteins that share the same fold to aid correct and efficient folding. However, experimentally measuring codon-specific translation speeds was until recently challenging, leading researchers to seek convenient proxies.

The most common proxy for translation speed is rare codon usage. Rare codons tend to have correspondingly rare cognate aminoacyl-tRNA which, by simple chemical kinetic arguments, will increase the dwell time of the ribosome at rare codons relative to commonly occurring codons that are recognized by more common cognate aminoacyl-tRNAs (Fluitt *et al.*, 2007). Many experimental (Buhr *et al.*, 2016; Zhang *et al.*, 2009) and theoretical (Nissley et al., 2016) investigations have found that synonymous codon

substitutions can drastically alter the ability of proteins to fold. It makes logical sense, then, that studies of codon usage indicate that clusters of rare codons are conserved in a position-specific fashion (Chaney *et al.*, 2017; Chartier *et al.*, 2012) between homologous protein-coding sequences. One study also found evidence that rare codons are positioned to facilitate co-translational folding, with the odds of finding a rare codon cluster 20–60 codons downstream of a predicted folding intermediate roughly twice the odds of finding a rare codon cluster elsewhere in an mRNA (Jacobs and Shakhnovich, 2017). Rare codons have also been shown to be positioned to facilitate interactions between nascent proteins and the signal recognition particle (Pechmann et al., 2014) as well as with other proteins in the cell (Chartier *et al.*, 2012). However, rare codons are not the only factor that influences translation times: mRNA structure (Hershey *et al.*, 2012), the chemistry of the amino acid being added to the nascent protein (Artieri and Fraser, 2014; Pavlov et al., 2009), mechanical forces generated by nascent proteins (Fritch *et al.*, 2018; Fujiwara *et al.*, 2020; Leininger et al., 2019) and interactions between the nascent protein and the ribosome (Gumbart *et al.*, 2012) are all part of the picture.

Ribosome profiling is a next-generation sequencing technique that produces a signal relative to the number of ribosomes engaged in translation of specific codons across the ensemble of mRNA molecules in cells (Ingolia et al., 2009). The ribosome occupancy at a codon position, assuming experimental biases have been correctly accounted for, should be directly proportional to the mean time required by the ribosome to decode that codon. Ribosome profiling thus provides a more complete proxy for translation speed in living cells than metrics like rare codon usage. One potential downside to ribosome profiling is that no one dataset has sufficient read coverage to provide insight into translation kinetics over the entire translatome. Pooling reads from different experiments to increase read coverage is one way to overcome this shortcoming (Ahmed *et al.*, 2019).

In this article, we find evidence of translation speed conservation based on comparison of ribosome profiling data. Our results demonstrate in a more complete and robust way than all previous studies that translation speed is conserved between structurally and evolutionarily related protein domains.

We compare normalized ribosome occupancy profiles between yeast domains identified by SUPERFAM (Wilson et al., 2009) to be evolutionarily and structurally related. We find that the profiles of these paralogous domains tend to be much more similar to one another than to randomly selected unrelated domains across a Pooled dataset composed of five different ribosome profiling studies (Jan et al., 2014; Nissley *et al.*, 2016; Weinberg et al., 2016; Williams et al., 2014; Young et al., 2015). This trend is also present in the four highest-coverage individual datasets included in our Pooled dataset, with the signal increasing in strength as the number of mapped reads included in our analysis increases. This trend is statistically different from a random control, indicating that biases in the ribosome profiling data alone do not explain our results. Many of the paralogous domains that have highly similar normalized ribosome occupancy profiles also have low DNA sequence identity (<50%), suggesting that translation speed profiles can be conserved over long stretches of evolutionary time.

## 2 Materials and methods

### 2.1 Ribosome profiling datasets included in analysis

The citations and GEO accession numbers of the six individual ribosome profiling datasets from five different studies (Jan *et al.*, 2014; Nissley *et al.*, 2016; Weinberg *et al.*, 2016; Williams *et al.*, 2014; Young *et al.*, 2015) we analysed are provided in Supplementary Table S1. The individual datasets are referred to using the name of the first author of the original study. Results were computed using various different poolings of these six sets of data (Supplementary Table S2). The 'Pooled' dataset described below always refers to the dataset that includes reads from all six individual datasets.

### 2.2 Selection of paralogous domain pairs

Reads from each ribosome profiling experiment (Supplementary Table S1) were mapped to the sacCer3 reference transcriptome as described in Nissley *et al.* 2016 and the A-site position within each ribosome-protected fragment determined using an integer-programming method (Ahmed *et al.*, 2019). Only those reads mapped to frame 0 are considered for downstream analysis. The 5,404 domain assignments for yeast strain S288C were cross-referenced with the ribosome occupancy profiles generated from the Pooled dataset and all domains removed that had (i) non-contiguous primary sequence definitions, (ii) less than 100 residues or (iii) less than 70% read coverage.

Pairs of paralogous domains were identified as those domains in the same SUPERFAM family (Wilson *et al.*, 2009). The DNA sequences of each unique pair of related domains were aligned with MUSCLE (Edgar, 2004) and all pairs with less than 30 or greater than 80% DNA sequence identity removed to filter out pairs of distantly and closely related domains, respectively. All pairs of domains passing these criteria were considered for ribosome occupancy profile comparisons, though some are rejected due to the additional criteria described below related to processing raw ribosome profiling read profiles into normalized ribosome occupancy profiles. Ordered locus names, e.g. YEL066W, are used to refer to domains within specific open reading frames in the yeast genome.

### 2.3 Calculation and comparison of ribosome occupancy profiles

The raw ribosome occupancy profiles for pairs of domains were first aligned to the domain pair's MUSCLE amino acid alignment. Domains with more than ten individual gaps in their alignment or with at least one gap of five positions or more were excluded. Gaps at either end of alignments are not considered in this filtering step. These 'gappy' alignments are eliminated to ensure that processed profiles are predominantly composed of experimental data, as gaps in aligned profiles are filled in by univariate spline interpolation on the non-zero positions (see below). The first 40 and last 20 profile positions relative to the full-length gene sequence were then removed to control for biases related to the well-known increase in reads at the 5′ and 3′ ends of the mRNA, respectively (Weinberg *et al.*, 2016). Univariate spline interpolation was used to cover areas with zero read density or at alignment gaps while holding values at all other alignment positions fixed. The resulting profiles were then smoothed with a fifteen-codon moving average (Jacobs and Shakhnovich, 2017; Reuveni et al., 2011) and finally normalized to have an area under the curve of one. Any processed profiles less than 50 positions in length were discarded, leaving 664 pairs of paralogous profiles for the Pooled dataset (Supplementary Table S2).

All pairs of profiles were compared based on their $f_{\mathrm{smf}}$ value. To compute this metric for the similarity between two profiles, the fast, medium and slow positions in each profile are first identified as those positions in the bottom, middle and top thirds of normalized ribosome occupancy. The $f_{\mathrm{smf}}$ value is then computed as the fraction of positions between two normalized profiles with the same classification of fast, medium or slow. Paralogous domain profiles were aligned based on the MUSCLE alignment of their amino acid sequences before calculation of $f_{\mathrm{smf}}$. Visual representations of the processing of raw profiles into normalized ribosome occupancy profiles and the calculation of $f_{\mathrm{smf}}$ are provided in Supplementary Figures S1 and S2, respectively.

### 2.4 Selection of non-paralogous domains for comparisons

Nineteen non-paralogous domains were selected at random for each of the 664 paralogous domain pairs within the Pooled dataset. Non-paralogous read profiles were required to meet the same ≥70% A-site read coverage criterion as paralogous domains, to be ≥100 residues in length, and to be within 25 residues of the length of the paralogous domains to which they were compared. Non-paralogous domains were also required to be in a different SCOP class, superfamily and family (Andreeva *et al.*, 2014) from the paralogous

domains to which they were compared. Profiles were aligned based on the first common domain position counting from the 5′ end and then truncated at their 3′ end to exactly match the length of the aligned paralogous domain profiles to provide a fair comparison. In cases where the non-paralogous profile was too short, it was rejected and another selected at random. Twenty independent iterations of this selection process were carried out and the paralogous domain profiles ranked against each of these twenty sets of 19 non-paralogous profiles based on $f_{smf}$ (e.g. Fig. 2a and Supplementary Fig. S2). The results in Figure 2b represent the mean number of paralogous domain pairs in each rank over these twenty random trials.

### 2.5 Calculation of codon usage bias profiles

%MinMax profiles were generated for all domains within the Pooled dataset for which ribosome occupancy profiles were compared. To provide a metric for codon usage bias that has the same single-codon resolution as ribosome profiling data we used a sliding window size of $z = 1$ (Rodriguez et al., 2018). Codon usage frequencies for yeast were downloaded from the CoCoPUTs database (Alexaki *et al.*, 2019). %Min values are reported as negative numbers by convention, so a global additive shift was applied to each profile to set the minimum value within the profile to 1. Following these setup steps, %MinMax profiles were compared in precisely the same fashion as ribosome occupancy profiles.

## 3 Results

### 3.1 Two distantly related paralogous yeast domains have highly similar ribosome occupancy profiles

We constructed the Pooled dataset of ribosome profiling data by combining reads from six ribosome profiling experiments published in five different studies by four different laboratories (Supplementary Table S1) (Jan *et al.*, 2014; Nissley *et al.*, 2016; Weinberg *et al.*, 2016; Williams *et al.*, 2014; Young *et al.*, 2015). Reads were mapped to the sacCer3 reference transcriptome as previously described (Nissley *et al.*, 2016) and the A-site position within each ribosome-protected fragment determined using an integer-programming method (Ahmed *et al.*, 2019). The resulting A-site read counts in the canonical translation frame were then summed across all six experiments. Pairs of structurally related domains within *S. cerevisiae* strain S288C were then identified as those domains within the same SUPERFAM family (Wilson *et al.*, 2009) (i.e. paralogous domains that are structurally and evolutionarily related) and their normalized ribosome occupancy profiles computed as described in Section 2.

Figure 1 shows an example of two structurally related domains and their ribosome occupancy profiles. The two Bromodomains YDL070W residues 134–242 and YKR008W residues 51–152 (SUPERFAM family 47371) have highly similar translation speed profiles (Fig. 1a, left panel). The amino acid and DNA sequences of these two domains have just 17% and 44% sequence identity, respectively, indicating a significant amount of evolutionary time has elapsed since the gene duplication event that led to their emergence as paralogous domains. Despite their divergence in both amino acid and DNA sequence, their ribosome occupancy profiles are far more similar to one another than to a randomly selected non-paralogous domain of a similar size (Fig. 1a, right panel). The conservation of ribosome occupancy profiles between these related domains suggests that translation speed may be evolutionarily conserved despite divergence in sequence over evolutionary time.

### 3.2 Ribosome occupancy profiles are conserved between related domains across the yeast translatome

The high degree of similarity between the ribosome occupancy profiles of YDL070W residues 134–242 and YKR008W residues 51–152 raises the question of whether such conservation is a general phenomenon. That is—are ribosome occupancy profiles of related domains more similar to one another than to profiles of unrelated domains, despite divergence in sequence, across the yeast translatome? To answer this question, we generated and compared ribosome occupancy profiles between all pairs of related domains with reasonable read coverage within our Pooled dataset.

Comparisons were performed by first identifying pairs of related domains with sufficient read coverage in their ribosome occupancy profiles. Pairs of domains that are very closely or very distantly related to one another were filtered out by requiring that the DNA sequence identity of domains used in this analysis be between 30% and 80%. Nineteen unrelated domains of similar size were selected for each pair of related domains to serve as an objective comparison set (Fig. 2a). A total of 664 unique pairs of related domains passed all quality control criteria and are included in the analysis (see Section 2). Comparisons between pairs of occupancy profiles were then made by classifying each position in each profile as being in the top, middle or bottom thirds of ribosome occupancy for each individual profile and then computing the fraction of positions in the aligned profiles with the same classification, denoted $f_{smf}$ (Supplementary Fig. S2). This comparison procedure was carried out twenty times for each paralogous domain pair (once for the paralogous domain pair and nineteen times for the non-paralogous pairs), the results rank ordered and the position of the paralogous pair within the ranking determined. For example, if a paralogous domain pair displays the largest $f_{smf}$ (most similar profiles) it would be placed in rank 1 as in Figure 2a; if a paralogous domain pair displays the fifth-smallest $f_{smf}$ (fifth most dissimilar profiles) it is placed in rank 16. This procedure of selecting pairs of related domains along with 19 unrelated domains and comparing their profiles was performed twenty times with different random selections of 19 unrelated domains for each trial. Within the Pooled dataset, 11% of paralogous domain pairs rank in the first position, a 120% increase over the number expected by random chance (Fig. 2b). Only ranks 1, 2, 3, 4 and 5 contain more pairs of paralogous domain pairs than expected by random chance. Qualitatively similar results are obtained for the four highest-coverage individual datasets included in our Pooled dataset (Supplementary Tables S1 and S2, Supplementary Fig. S3). These results indicate that ribosome occupancy profiles are conserved between related yeast domains, suggesting that their translation schedules are conserved.

### 3.3 Accounting for biases in ribosome profiling data

Ribosome profiling experiments suffer from various biases that may cause occupancy profiles to be similar between related domains despite them having dissimilar translation speeds *in vivo*. For example, it is now well-known that the use of chemical agents such as cycloheximide to arrest translation leads to altered occupancy profiles that do not reflect real translation times (Hussmann *et al.*, 2015). Though we have specifically selected ribosome profiling datasets generated without the use of cycloheximide to arrest translation, other biases may be present for which we need to control. To account for such hidden biases, we also performed comparisons between sets of 20 randomly selected domains (random control in Fig. 2b). This selection procedure was performed precisely as for pairs of related domains with two exceptions: (i) selected pairs are not required to be in the same SUPERFAM family, though this may still occur by random chance, and (ii) no DNA sequence identity criterion is applied. A total of 664 random pairs were generated, allowing for fair comparisons. This random control is statistically differentiable from the Pooled dataset results for all ranks except 6, 7, 10, 11, 13 and 14 (permutation test, $\alpha = 0.05$, $1 \times 10^6$ samples). The random control trials find a mean of 38.15 pairs of domains in the first rank over twenty trials, somewhat higher than the 33.20 (=664/20) pairs expected if the result was completely random (Fig. 2b). This suggests that while biases and errors are likely present in the ribosome occupancy profiles, these biases alone cannot account for the observed similarity between profiles of related domains in yeast.

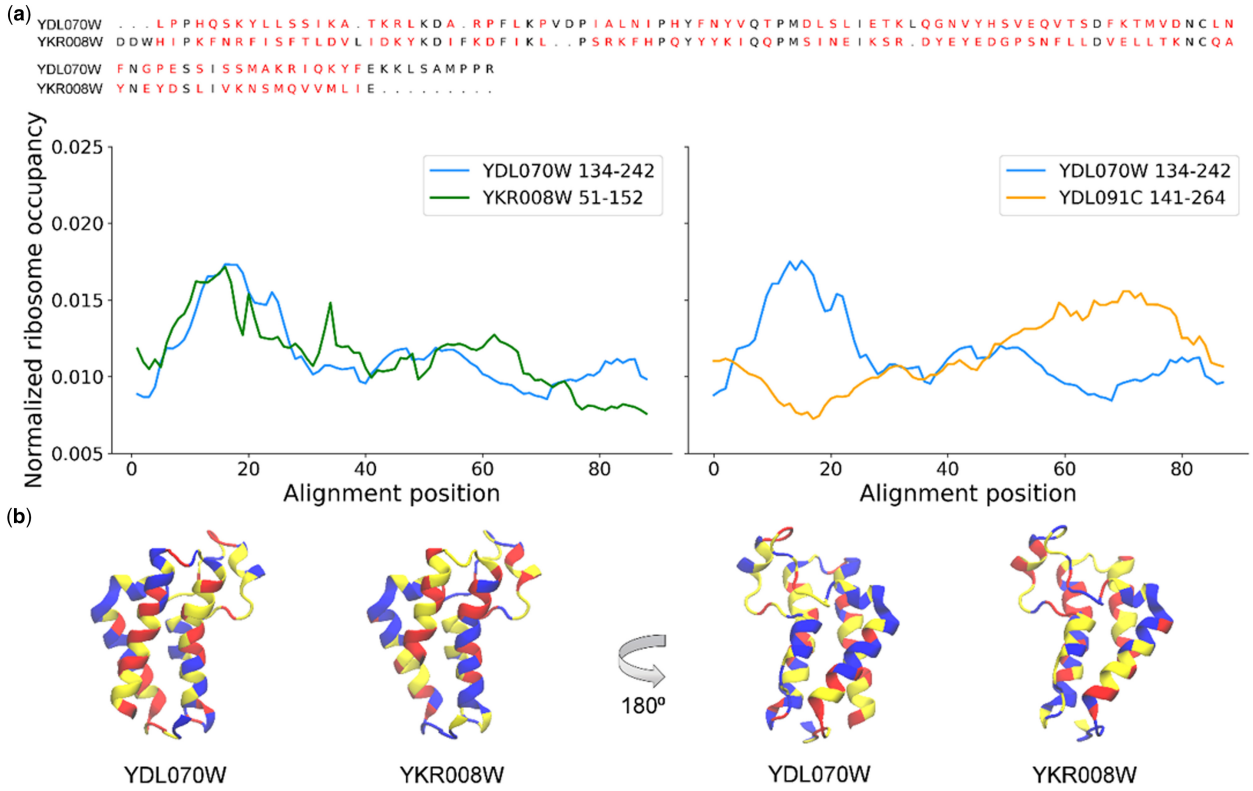


**Fig. 1.** Two yeast Bromodomains have highly similar ribosome occupancy profiles. (**a**) The MUSCLE sequence alignment between the amino acid sequences of the two Bromodomains YDL070W residues 134–242 and YKR008W residues 51–152 (SUPERFAM family 47371, top) has 17% identity. Positions that do not match between both sequences are colored red. The normalized ribosome occupancy profiles for YDL070W residues 134–242 (blue) and YKR008W residues 51–152 (green) were calculated based on the Pooled A-site read dataset as described in Section 2 and plotted as a function of position within the aligned and processed profiles (left panel). The right panel displays the processed ribosome occupancy profiles for YDL070W residues 134–242 (blue) and the randomly selected non-paralogous domain YDL091C residues 141–264 aligned from the first common profile position of their 5′ end (see Section 2). (**b**) PDB ID: 2R0V, which represents YKR008W, colored based on the ribosome occupancy profiles for YDL070W residues 134–242 and YKR008W residues 51–152. Sections of the structures colored red, yellow and blue correspond to the fastest, middle and slowest thirds of translation times within each profile. Note that no trimming, smoothing or normalization of ribosome occupancy profiles was performed in this instance to maintain a length similar to that of the domain itself for the sake of visualization

### 3.4 Highly similar ribosome occupancy profiles are found regardless of domain size, DNA sequence identity, structural class and amino acid composition

We next investigated the characteristics of paralogous domain pairs with highly similar ribosome occupancy profiles in comparison to those that rank poorly in Figure 2b. Importantly, many pairs of paralogous domains in the top rank have low DNA sequence identity, indicating that their high $f_{smf}$ values are not primarily due to favorable comparisons between domains with very recent common ancestors (Fig. 2c, left panel). Top-ranked paralogous domain pairs have a similar length distribution to pairs ranked in other positions (Fig. 2c, middle panel), though larger domains are slightly overrepresented in the top rank. There is also no clear dependence on SCOP class, with all four main structural classes (a, b, c and d) found in the top rank and all other ranks in similar proportions (Fig. 2c, right panel). Finally, we compared the amino acid composition of top-ranked domains versus domains with less similar ribosome occupancy profiles. Though overall very similar, top-ranked domain pairs are slightly enriched in His, Trp and Tyr and depleted in Asn (Supplementary Fig. S4). The fact that many top-ranked pairs of domains have low DNA sequence identity indicates that even when sequences have diverged significantly ribosome occupancy profiles remain highly similar. There is no obvious or general difference between pairs of related domains in the top-rank and pairs that rank lower in Figure 2b, suggesting that conservation of translation speed profiles between related domains is a general phenomenon.

## 4 Discussion

Our results indicate that ribosome occupancy profiles produced from ribosome profiling data are conserved between pairs of paralogous domains in yeast. The similarity of these profiles is apparent at the level of individual pairs of related domains (Fig. 1a) and across the set of all paralogous domains in yeast (Fig. 2b) with acceptable read coverage and sequence alignments.

Three hypotheses can explain in part or in whole why ribosome occupancy profiles are conserved between structurally similar domains. First, ribosome occupancy profiles may be conserved due to the influence of translation speed on co-translational folding. Structurally similar domains are likely to fold by similar pathways, meaning that perturbation of the translation speed profile may hinder their folding process and reduce the fitness of the protein. A second hypothesis, which is really a set of hypotheses, is that translation speed is not under selection at all, but factors like mRNA structure that influence translation speed are under selection. Evolutionary pressure on mRNA structure would lead to similar mRNA sequences between paralogous domains and, due to the relationship between codon usage and translation speed, similar ribosome occupancy profiles, despite the fact that the root cause is not related to translation speed. This second hypothesis is not cleanly separable from the co-translational folding hypothesis because mRNA structure, along with many other factors, influences translation speed. A third hypothesis is that we are considering domains that are too closely related, such that the paralogous domain sequences we compare have had too little evolutionary time to diverge and we are effectively comparing profiles to themselves. This third

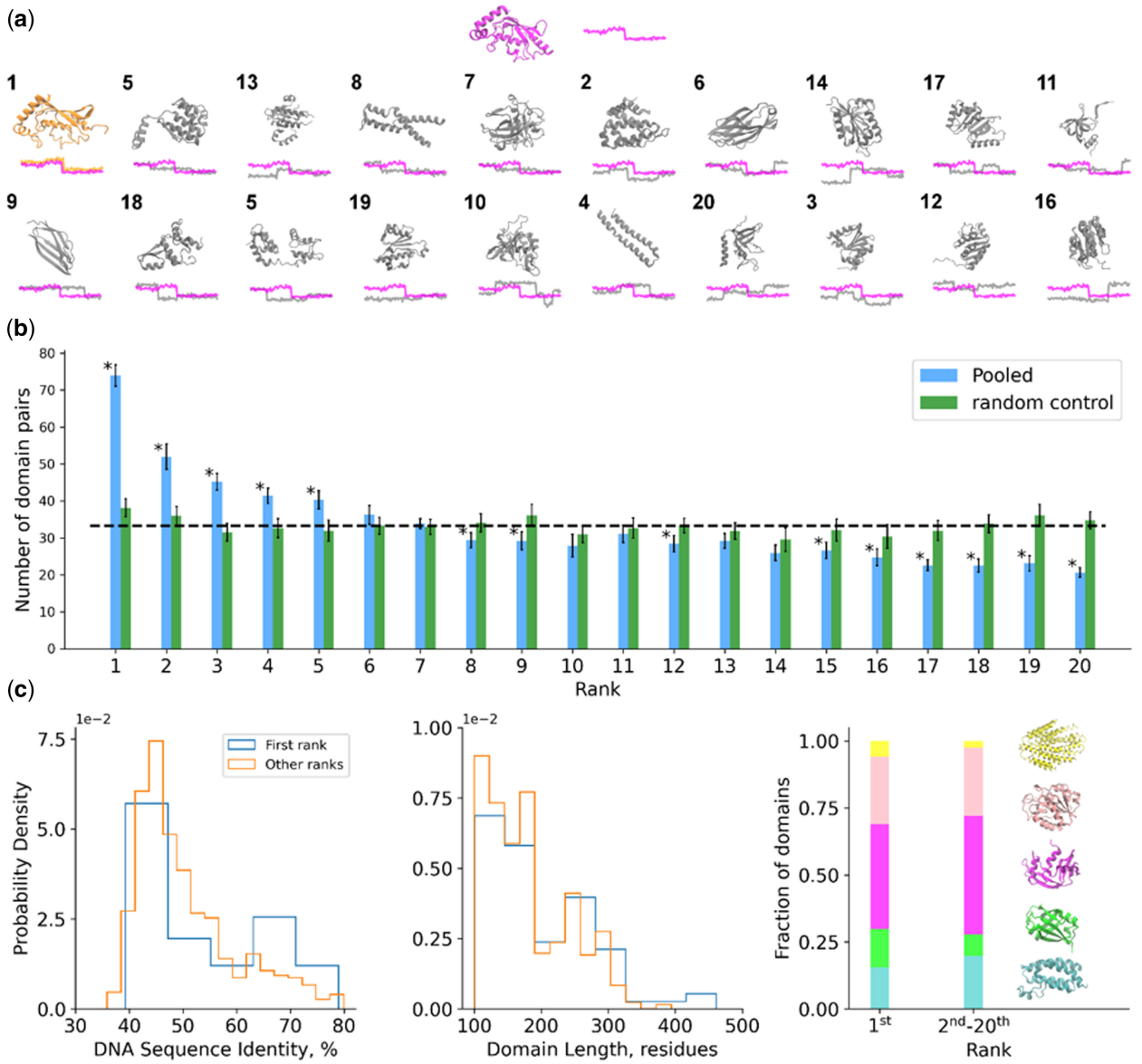


**Fig. 2.** Ribosome occupancy profiles are conserved between related pairs of domains across the yeast translatome. (**a**) Schematic of the comparison procedure for ribosome occupancy profiles of two related domains (magenta and orange). Nineteen unrelated domains of a similar size but in different SUPERFAM superfamilies, families and SCOP classes are also selected (grey structures). The processed ribosome occupancy profiles are then compared on the basis of the $f_{smf}$ metric. The resulting $f_{smf}$ scores are then ranked and the position of the pair of related domains in this ranking determined. Numbers represent the position of each domain in the final ranking. In this example, the pair of related domains have the highest $f_{smf}$ score, indicating they are the most similar out of all pairs of profiles, and they are therefore placed in the first rank. (**b**) The number of paralogous domain pairs from the set of 664 within the Pooled dataset that rank in positions 1st through 20th when compared against one another and against 19 non-paralogous domains as shown in (a) (blue). This analysis was also performed using randomly selected pairs of domains (green). The dotted line indicates the number of paralogous domain pairs expected in each rank if the results are completely random. Error bars are 95% confidence intervals calculated from the results of 20 independent trials. Asterisks indicate ranks for which there is a statistically significant difference between the Pooled results and random control (from permutation test, $\alpha = 0.05$, $1 \times 10^6$ samples). (**c**) (left) Histograms of DNA sequence identity for pairs of paralogous domains in the first rank (blue) and all other ranks (orange) in the first random trial. Subsequent random trials have similar results. (middle) Histograms of domain lengths for pairs of paralogous domains in the first rank (blue) and all other ranks (orange). (right) Stacked barplots indicating the fraction of domain pairs in the first rank and all other ranks that belong to SCOP classes a ($\alpha$, cyan), b ($\beta$, green), c ($\alpha+\beta$, magenta), d ($\alpha/\beta$, pink), e (multi-domain, yellow), f (membrane, yellow) and g (small proteins, yellow). For simplicity, the rarely occurring classes e, f and g are all colored yellow

hypothesis is unlikely given the range of sequence identity found for pairs of related domains in the first rank of results (Fig. 2c, left panel). Decoupling the first two hypotheses will provide more insight into the processes underlying conservation of ribosome occupancy profiles between evolutionarily related domains.

Co-translational folding and thus translation speed is often thought to be more important for larger domains with more complex folding landscapes and for domains with more $\beta$ character, as $\beta$-sheets often require forming hydrogen bond networks between portions of domains that are disparate in primary sequence. However, our results indicate no clear difference in domain length, amino acid composition, SCOP structural class or sequence identity between paralogous domains with highly similar profiles and those with dissimilar profiles (Fig. 2c). We have also found that many pairs of paralogous domains rank poorly, and some compare less favorably than all 19 randomly selected unrelated domains (Fig. 2b).

This raises a key question: if conservation of ribosome occupancy profiles between related domains is a general phenomenon, why do some pairs of even closely related domains result in poor comparisons?

One possible explanation is the high degree of noise inherent to ribosome profiling data and the low coverage found for many coding sequences. Comparing ribosome occupancy profiles generated for the same domain between the Williams and Weinberg ribosome profiling datasets reveals that some domains rank poorly even when compared to themselves between different datasets, though many more domains are found in the first rank than when pairs of related domains are compared within one dataset (Supplementary Fig. S5). This result suggests that ribosome occupancy profiles of individual domains are highly similar between different datasets but not identical. It may be the case that the signal for evolutionary conservation between related domains will become stronger as more high-quality ribosome profiling data becomes available; indeed, as we increase the total number of reads included in our analysis the signal becomes increasingly clear (Supplementary Fig. S6). Increasing sequence similarity also leads to an increasingly strong result (Fig. 2b, Supplementary Figs S5 and S7).

Data with high coverage over the entire translatome will result in more domain pairs with viable read coverage, allowing for extension of our method to more domains with more diverse sequence identity. When more domains are included in the analysis a trend may emerge between protein properties such as SCOP class or domain size and conservation of ribosome occupancy. We also computed codon usage bias profiles using the %MinMax algorithm (Rodriguez *et al.*, 2018) and compared them for the same sets of domains for which ribosome occupancy profiles were compared. We found related domain pairs are most likely to be in ranks 1 or 2 (Supplementary Fig. S8). This result is expected, as previous studies have reported various levels of conservation of codon usage (Chaney *et al.*, 2017; Jacobs and Shakhnovich, 2017). Our results suggest that, at least for the 664 pairs of domains for which we have compared both %MinMax and ribosome occupancy profiles, ribosome occupancy is more strongly conserved between related domains. Our results also indicate that only 1 in 3 pairs of domains with highly conserved %MinMax profiles also have highly conserved ribosome occupancy profile (Supplementary Fig. S8, inset). This suggests that comparison of codon positions alone may provide an incomplete picture of real translation kinetics. Our results show that entire ribosome occupancy profiles are conserved between structurally and evolutionarily related proteins. This result offers strong evidence that translation schedule is important for preserving folding pathways for proteins with similar structures.

One obvious extension of our methods is to compare ribosome occupancy profiles between orthologous proteins in different organisms. Unfortunately, while a general consensus has been reached about how to best process yeast ribosome profiling data, analysis for other organisms remains less clear, and even the best datasets remain low coverage in comparison to the best yeast datasets (Mohammad et al., 2019). We compared ribosome occupancy profiles between our Pooled yeast dataset and *E. coli* ribosome profiling data from Mohammed and co-workers (Mohammad *et al.*, 2019). We found that while some individual pairs of related domains have highly similar profiles (Supplementary Fig. S9) between the two organisms, too few pairs of related domains can be compared to provide confidence that this similarity is differentiable from random chance.

If translation schedule is critical to directing folding along optimal pathways, even evolutionarily unrelated proteins with similar folds (i.e. proteins that have undergone convergent evolution) will have similar translation speed profiles. It may be interesting to test this hypothesis in the future using ribosome profiling data.

A deeper understanding of the relationship between conservation of translation schedule and folding pathways may prove important in several areas of protein science. For example, as the quantity of high-quality ribosome profiling data increases it may be possible to extract characteristic translation schedule fingerprints for individual structural motifs [e.g. Greek key (Hutchinson and Thornton, 1993)]. This more detailed understanding of the relationship between translation schedule and structure could then be used for the rational design of proteins with robust co-translational folding characteristics for efficient folding *in vivo*. Conservation of translation schedule between related proteins also has implications for the recombinant expression of proteins. It is now common practice to harmonize the codon usage of the coding sequence to be expressed to match the codon usage of the expression organism (Angov *et al.*, 2008). Our results suggest that matching the translation schedule to preserve the endogenous co-translational folding pathway may result in an even higher fraction of correctly folded, functional protein.

In summary, our results indicate that ribosome occupancy profiles are conserved between structurally related yeast domains. We hypothesize that ribosome occupancy (and thus translation schedule) is conserved to preserve efficient co-translational folding pathways. As more high-quality ribosome profiling data become available more detailed translation schedule trends may be revealed.

## Acknowledgements

The authors acknowledge Ed O'Brien, Nabeel Ahmed and Nishant Soni for providing the ribosome profiling data and for helpful discussions.

*Financial Support*: A.C. and M.C. acknowledge funding from EPSRC grants EP/S024093/1 and EP/G03706X/1, respectively. M.C. was also supported by the Doctoral Training Centre at Oxford University.

*Conflict of Interest*: None declared.

## References

Ahmed,N. *et al.* (2019) Identifying A- and P-site locations on ribosome-protected mRNA fragments using Integer Programming. *Sci. Rep.*, **9**, 6256.

Alexaki,A. *et al.* (2019) Codon and Codon-Pair Usage Tables (CoCoPUTs): facilitating genetic variation analyses and recombinant gene design. *J. Mol. Biol.*, **431**, 2434–2441.

Andreeva,A. *et al.* (2014) SCOP2 prototype: a new approach to protein structure mining. *Nucleic Acids Res.*, **42**, 310–314.

Angov,E. *et al.* (2008) Heterologous protein expression is enhanced by harmonizing the codon usage frequencies of the target gene with those of the expression host. *PLoS One*, **3**, e2189.

Artieri,C.G. and Fraser,H.B. (2014) Accounting for biases in riboprofiling data indicates a major role for proline in stalling translation. *Genome Res.*, **24**, 2011–2021.

Buhr,F. *et al.* (2016) Synonymous codons direct cotranslational folding toward different protein conformations. *Mol. Cell*, **61**, 341–351.

Chaney,J.L. *et al.* (2017) Widespread position-specific conservation of synonymous rare codons within coding sequences. *PLoS Comput. Biol*, **e1005531**, 1–19.

Chartier,M. *et al.* (2012) Large-scale analysis of conserved rare codon clusters suggests an involvement in co-translational molecular recognition events. *Bioinformatics*, **28**, 1438–1445.

Cortazzo,P. *et al.* (2002) Silent mutations affect in vivo protein folding in *Escherichia coli*. *Biochem. Biophys. Res. Commun.*, **293**, 537–541.

Edgar,R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, **32**, 1792–1797.

Fluitt,A. *et al.* (2007) Ribosome kinetics and aa-tRNA competition determine rate and fidelity of peptide synthesis. *Comput. Biol. Chem.*, **31**, 335–346.

Fritch,B. *et al.* (2018) Origins of the mechanochemical coupling of peptide bond formation to protein synthesis. *J. Am. Chem. Soc.*, **140**, 5077–5087.

Frydman,J. *et al.* (1999) Co-translational domain folding as the structural basis for the rapid de novo folding of firefly luciferase. *Nat. Struct. Biol.*, **6**, 697–705.

Fujiwara,K. *et al.* (2020) Proteome-wide capture of co-translational protein dynamics in *Bacillus subtilis* using TnDR, a transposable protein-dynamics reporter. *Cell Rep.*, **33**, 108250.

Gumbart,J. *et al.* (2012) Mechanisms of SecM-mediated stalling in the ribosome. *Biophys. J.*, **103**, 331–341.

Hershey,J.W.B. *et al.* (2012) Principles of Translational Control: an Overview. *Cold Spring Harb. Perspect. Biol.*, **4**, a011528.

Holtkamp,W. *et al.* (2015) Cotranslational protein folding on the ribosome monitored in real time. *Science*, **350**, 1104–1107.

Hussmann,J.A. *et al.* (2015) Understanding biases in ribosome profiling experiments reveals signatures of translation dynamics in yeast. *PLoS Genet.*, **11**, e1005732.
Hutchinson,E.G. and Thornton,J.M. (1993) The Greek key motif: extraction, classification and analysis. *Protein Eng.*, **6**, 233–245.
Ingolia,N.T. *et al.* (2009) Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science*, **324**, 218–223.
Jacobs,W.M. and Shakhnovich,E.I. (2017) Evidence of evolutionary selection for cotranslational folding. *Proc. Natl. Acad. Sci. USA*, **114**, 11434–11439.
Jan,C.H. *et al.* (2014) Principles of ER cotranslational translocation revealed by proximity-specific ribosome profiling. *Science*, **346**, 1257521.
Kim,S.J. *et al.* (2015) Translational tuning optimizes nascent protein folding in cells. *Science*, **348**, 444–448.
Knobe,K.E. *et al.* (2008) Why does the mutation G17736A/Val107Val (silent) in the F9 gene cause mild haemophilia B in five Swedish families? *Haemophilia*, **14**, 723–728.
Leininger,S.E. *et al.* (2019) Domain topology, stability, and translation speed determine mechanical force generation on the ribosome. *Proc. Natl. Acad. Sci. USA*, **116**, 5523–5532.
Mohammad,F. *et al.* (2019) A systematically-revised ribosome profiling method for bacteria reveals pauses at single-codon resolution. *eLife*, **9**, 1–25.
Nicola,A.V. *et al.* (1999) Co-translational folding of an alphavirus capsid protein in the cytosol of living cells. *Nat. Cell Biol.*, **1**, 341–345.
Nissley,D.A. *et al.* (2016) Accurate prediction of cellular co-translational folding indicates proteins can switch from post- to co-translational folding. *Nat. Commun*, **7**, 10341.
Nissley,D.A. and O'Brien,E.P. (2016) Altered co-translational processing plays a role in Huntington's pathogenesis—a hypothesis. *Front. Mol. Neurosci.*, **9**, 54.
Nissley,D.A. and O'Brien,E.P. (2014) Timing is everything: unifying codon translation rates and nascent proteome behavior. *J. Am. Chem. Soc.*, **136**, 17892–17898.
Noriega,T.R. *et al.* (2014) Signal recognition particle-ribosome binding is sensitive to nascent chain length. *J. Biol. Chem.*, **289**, 19294–19305.
Pavlov,M.Y. *et al.* (2009) Slow peptide bond formation by proline and other N-alkylamino acids in translation. *Proc. Natl. Acad. Sci. USA*, **106**, 50–54.
Pechmann,S. *et al.* (2014) Local slowdown of translation by nonoptimal codons promotes nascent-chain recognition by SRP in vivo. *Nat. Struct. Mol. Biol.*, **21**, 1100–1105.
Reuveni,S. *et al.* (2011) Genome-scale analysis of translation elongation with a ribosome flow model. *PLoS Comput. Biol*, **7**, e1002127.
Rodriguez,A. *et al.* (2018) %MinMax: A versatile tool for calculating and comparing synonymous codon usage and its impact on protein folding. *Protein Sci.*, **27**, 356–362.
Sauna,Z.E. and Kimchi-Sarfaty,C. (2011) Understanding the contribution of synonymous mutations to human disease. *Nat. Rev. Genet.*, **12**, 683–691.
Thommen,M. *et al.* (2017) Co-translational protein folding: progress and methods. *Curr. Opin. Struct. Biol.*, **42**, 83–89.
Walter,P. and Johnson,A. (1994) Signal sequence recognition and protein targeting to the membrane. *Annu. Rev. Cell Biol.*, **10**, 87–119.
Weinberg,D.E. *et al.* (2016) Improved ribosome-footprint and mRNA measurements provide insights into dynamics and regulation of yeast translation. *Cell Rep.*, **14**, 1787–1799.
Williams,C.C. *et al.* (2014) Targeting and plasticity of mitochondrial proteins revealed by proximity-specific ribosome profiling. *Science*, **346**, 748–751.
Wilson,D. *et al.* (2009) SUPERFAMILY—sophisticated comparative genomics, data mining, visualization and phylogeny. *Nucleic Acids Res.*, **37**, 380–386.
Young,D.J. *et al.* (2015) Rli1/ABCE1 recycles terminating ribosomes and controls translation reinitiation in 3′UTRs in vivo. *Cell*, **162**, 872–884.
Zhang,G. *et al.* (2009) Transient ribosomal attenuation coordinates protein synthesis and co-translational folding. *Nat. Struct. Mol. Biol.*, **16**, 274–280.
Zhou,M. *et al.* (2013) Non-optimal codon usage affects expression, structure and function of clock protein FRQ. *Nature*, **495**, 111–115.

# Bibliography


1. Abriata, L. A., Tamò, G. E., Monastyrskyy, B., Kryshtafovych, A. & Dal Peraro, M. Assessment of hard target modeling in CASP12 reveals an emerging role of alignment-based contact prediction methods. *Proteins Struct. Funct. Bioinforma.* **86,** 97–112 (Mar. 2018).

2. Adhikari, B., Hou, J. & Cheng, J. DNCON2: improved protein contact prediction using two-level deep convolutional neural networks. *Bioinformatics* **34** (ed Valencia, A.) 1466–1472 (May 2018).

3. Alquraishi, M. End-to-End Differentiable Learning of Protein Structure. *Cell Syst.* **8,** 301 (2019).

4. Anfinsen, C. B. *Principles that govern the folding of protein chains* July 1973.

5. Anishchenko, I., Ovchinnikov, S., Kamisetty, H. & Baker, D. Origins of coevolution between residues distant in protein 3D structures. *Proc. Natl. Acad. Sci. U. S. A.* **114,** 9122–9127 (Aug. 2017).

6. Asam, C. *et al.* Harmonization of the Genetic Code Effectively Enhances the Recombinant Production of the Major Birch Pollen Allergen Bet v 1. *Int. Arch. Allergy Immunol.* **177,** 116–122 (Sept. 2018).

7. Balakrishnan, S., Kamisetty, H., Carbonell, J. G., Lee, S.-I. & Langmead, C. J. Learning generative models for protein fold families. *Proteins Struct. Funct. Bioinforma.* **79,** 1061–1078 (Apr. 2011).

8. Balasco, N., Esposito, L., Thind, A. S., Guarracino, M. R. & Vitagliano, L. Dissection of factors affecting the variability of the peptide bond geometry and planarity. *Biomed Res. Int.* **2017** (2017).

9. Baldassi, C. *et al.* Fast and Accurate Multivariate Gaussian Modeling of Protein Families: Predicting Residue Contacts and Protein-Interaction Partners. *PLoS One* **9** (ed Hamacher, K.) e92721 (Mar. 2014).

10. Bastolla, U. & Demetrius, L. Stability constraints and protein evolution: The role of chain length, composition and disulfide bonds. *Protein Eng. Des. Sel.* **18,** 405–415 (Sept. 2005).

11. Bateman, A. *et al.* UniProt: the universal protein knowledgebase. *Nucleic Acids Res.* **45,** D158–D169 (Jan. 2017).

12. Batey, S., Nickson, A. A. & Clarke, J. *Studying the folding of multidomain proteins* Dec. 2008.

13. Baxa, M. C., Freed, K. F. & Sosnick, T. R. Quantifying the structural requirements of the folding transition state of protein A and other systems. *J. Mol. Biol.* **381,** 1362–81 (Sept. 2008).

14. Baxa, M. C., Haddadian, E. J., Jumper, J. M., Freed, K. F. & Sosnick, T. R. Loss of conformational entropy in protein folding calculated using realistic ensembles and its implications for NMR-based calculations. *Proc. Natl. Acad. Sci. U. S. A.* **111,** 15396–15401 (Oct. 2014).

15. Belardinelli, R. E., Manzi, S. & Pereyra, V. D. Analysis of the convergence of the 1/t and Wang-Landau algorithms in the calculation of multidimensional integrals (June 2008).

16. Bennion, B. J. & Daggett, V. The molecular basis for the chemical denaturation of proteins by urea. *Proc. Natl. Acad. Sci. U. S. A.* **100,** 5142–7 (Apr. 2003).

17. Berg, J., Stryer, L. & Tymoczko, J. *Biochemistry* 9th ed. (Macmillan, New York, 2019).

18. Bergman, L. W. & Kuehl, W. M. Formation of intermolecular disulfide bonds on nascent immunoglobulin polypeptides. *J. Biol. Chem.* **254,** 5690–5694 (July 1979).

19. Bergman, L. W. & Kuehl, W. M. Co-translational modification of nascent immunoglobulin heavy and light chains. *J. Supramol. Struct.* **11,** 9–24 (Jan. 1979).

20. Bergman, L. W. & Kuehl, W. M. Co-translational modification of nascent immunoglobulin heavy and light chains. *J. Supramol. Struct.* **11,** 9–24 (Jan. 1979).

21. Betancourt, M. R. & Thirumalai, D. Pair potentials for protein folding: Choice of reference states and sensitivity of predicted native states to variations in the interaction schemes. *Protein Sci.* **8,** 361–369 (Dec. 2008).

22. Bhardwaj, A., Walker-Kopp, N., Wilkens, S. & Cingolani, G. Foldon-guided self-assembly of ultra-stable protein fibers. *Protein Sci.* **17,** 1475–1485 (Sept. 2008).

23. Bhattacharya, D., Cao, R. & Cheng, J. UniCon3D: de novo protein structure prediction using united-residue conformational search via stepwise, probabilistic sampling. *Bioinformatics* **32,** 2791–2799 (Sept. 2016).

24. Billings, W. M., Hedelius, B., Millecam, T., Wingate, D. & Corte, D. D. *ProSPr: Democratized implementation of alphafold protein distance prediction network* Nov. 2019.

25. Bordner, A. J., Cavasotto, C. N. & Abagyan, R. A. Direct derivation of van der Waals force field parameters from quantum mechanical interaction energies. *J. Phys. Chem. B* **107,** 9601–9609 (Sept. 2003).

26. Bosnjak, I., Bojovic, V., Segvic-Bubic, T. & Bielen, A. Occurrence of protein disulfide bonds in different domains of life: a comparison of proteins from the Protein Data Bank. *Protein Eng. Des. Sel.* **27,** 65–72 (Mar. 2014).

27. Bosshard, H. R., Marti, D. N. & Jelesarov, I. *Protein stabilization by salt bridges: Concepts, experimental approaches and clarification of some misunderstandings* Jan. 2004.

28. Brenner, S. E., Koehl, P. & Levitt, M. The ASTRAL compendium for protein structure and sequence analysis. *Nucleic Acids Res.* **28,** 254–256 (Jan. 2000).

29. Brown, C. A. & Brown, K. S. Validation of coevolving residue algorithms via pipeline sensitivity analysis: ELSC and OMES and ZNMI, oh my! *PLoS One* **5,** e10779 (June 2010).

30. Bryngelson, J. D. & Wolynes, P. G. Spin glasses and the statistical mechanics of protein folding. *Proc. Natl. Acad. Sci. U. S. A.* **84,** 7524–7528 (Nov. 1987).

31. Buchan, D. W. A. & Jones, D. T. Improved protein contact predictions with the MetaPSICOV2 server in CASP12. *Proteins Struct. Funct. Bioinforma.* **86,** 78–83 (Mar. 2018).

32. Chandonia, J.-M. The ASTRAL Compendium in 2004. *Nucleic Acids Res.* (2004).

33. Chandonia, J. M. *et al.* ASTRAL compendium enhancements. *Nucleic Acids Res.* **30,** 260–263 (Jan. 2002).

34. Chaney, J. L. & Clark, P. L. Roles for Synonymous Codon Usage in Protein Biogenesis. *Annu. Rev. Biophys.* **44,** 143–166 (June 2015).

35. Chaney, J. L., Steele, A., *et al.* Widespread position-specific conservation of synonymous rare codons within coding sequences. *PLoS Comput. Biol.* **13** (May 2017).

36. Chwastyk, M. & Cieplak, M. Cotranslational folding of deeply knotted proteins. *J. Phys. Condens. Matter* **27,** 354105 (Sept. 2015).

37. Clark, P. L., Weston, B. F. & Gierasch, L. M. Probing the folding pathway of a $\beta$-clam protein with single-tryptophan constructs. *Fold. Des.* **3,** 401–412 (Oct. 1998).

38. Clementi, C., García, A. E. & Onuchic, J. N. Interplay among tertiary contacts, secondary structure formation and side-chain packing in the protein folding mechanism: all-atom representation study of protein L. *J. Mol. Biol.* **326,** 933–54 (Feb. 2003).

39. Cobb, N. J. & Surewicz, W. K. Prion diseases and their biochemical mechanisms. *Biochemistry* **48,** 2574–2585 (Mar. 2009).

40. Cochran, W. G. Some Methods for Strengthening the Common $\chi$ 2 Tests. *Biometrics* **10,** 417 (Dec. 1954).

41. Craveur, P., Joseph, A. P., Poulain, P., De Brevern, A. G. & Rebehmed, J. *Cis-trans isomerization of omega dihedrals in proteins* Aug. 2013.

42. Crivelli, S. *et al.* A Physical Approach to Protein Structure Prediction. *Biophys. J.* **82,** 36–49 (Jan. 2002).

43. Dai, X. *et al.* Reduction of translating ribosomes enables Escherichia coli to maintain elongation rates during slow growth. *Nat. Microbiol.* **2** (Dec. 2016).

44. Dawson, N. L. *et al.* CATH: an expanded resource to predict protein function through structure and sequence. *Nucleic Acids Res.* **45,** D289–D295 (Jan. 2017).

45. Deane, C. M., Dong, M., Huard, F. P., Lance, B. K. & Wood, G. R. Cotranslational protein folding: fact or fiction? *Bioinformatics* **23,** i142–i148 (July 2007).

46. DeepMind. *AlphaFold: a solution to a 50-year-old grand challenge in biology* 2020.

47. Deleuze, G. & Strauss, J. The Fold. *Yale French Stud.,* 227 (1991).

48. Deng, Z., Chuaqui, C. & Singh, J. Structural Interaction Fingerprint (SIFt): A Novel Method for Analyzing Three-Dimensional Protein-Ligand Binding Interactions. *J. Med. Chem.* **47,** 337–344 (Jan. 2004).

49. De Oliveira, S. *Biologically inspired de novo protein structure prediction* PhD thesis (University of Oxford, 2015).

50. De Oliveira, S. & Deane, C. Co-evolution techniques are reshaping the way we do structural bioinformatics. *F1000Research* **6,** 1224 (2017).

51. De Oliveira, S., Law, E. C., Shi, J., Deane, C. M. & Valencia, A. Sequential search leads to faster, more efficient fragment-based de novo protein structure prediction. *Bioinformatics* **34** (ed Valencia, A.) 1132–1140 (Apr. 2018).

52. De Oliveira, S., Shi, J. & Deane, C. M. Comparing co-evolution methods and their application to template-free protein structure prediction. *Bioinformatics* **33,** 373–381 (2017).

53. De Oliveira, S., Shi, J. & Deane, C. M. Building a Better Fragment Library for De Novo Protein Structure Prediction. *PLoS One* **10** (ed Zhang, Y.) e0123998 (Apr. 2015).

54. Dill, K. A., Ozkan, S. B., Shell, M. S. & Weikl, T. R. *The protein folding problem* 2008.

55. Dunn, S., Wahl, L. & Gloor, G. Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction. *Bioinformatics* **24,** 333–340 (Feb. 2008).

56. Dyson, H. J., Wright, P. E. & Scheraga, H. A. The role of hydrophobic interactions in initiation and propagation of protein folding. *Proc. Natl. Acad. Sci. U. S. A.* **103,** 13057–13061 (Aug. 2006).

57. Eichmann, C., Preissler, S., Riek, R. & Deuerling, E. Cotranslational structure acquisition of nascent polypeptides monitored by NMR spectroscopy. *Proc. Natl. Acad. Sci. U. S. A.* **107,** 9111–9116 (May 2010).

58. Ellis, J. J., Huard, F. P., Deane, C. M., Srivastava, S. & Wood, G. R. Directionality in protein fold prediction. *BMC Bioinformatics* **11** (2010).

59. Englander, S. W. & Mayne, L. The nature of protein folding pathways. *Proc. Natl. Acad. Sci. U. S. A.* **111,** 15873–15880 (2014).

60. Englander, S. W., Mayne, L., Kan, Z.-Y. & Hu, W. Protein Folding—How and Why: By Hydrogen Exchange, Fragment Separation, and Mass Spectrometry. *Annu. Rev. Biophys.* **45,** 135–152 (July 2016).

61. Fedyukina, D. V. & Cavagnero, S. Protein folding at the exit tunnel. *Annu. Rev. Biophys.* **40,** 337–59 (2011).

62. Ferreiro, D. U., Komives, E. A. & Wolynes, P. G. Frustration in biomolecules. *Q. Rev. Biophys.* **47,** 285–363 (Nov. 2014).

63. Fox, N. K., Brenner, S. E. & Chandonia, J.-M. SCOPe: Structural Classification of Proteins—extended, integrating SCOP and ASTRAL data and classification of new structures. *Nucleic Acids Res.* **42,** D304–D309 (Jan. 2014).

64. Frank, H. S. & Evans, M. W. Free volume and entropy in condensed systems III. Entropy in binary liquid mixtures; Partial molal entropy in dilute solutions; Structure and thermodynamics in aqueous electrolytes. *J. Chem. Phys.* **13,** 507–532 (Nov. 1945).

65. Frishman, D. & Argos, P. Knowledge-based protein secondary structure assignment. *Proteins Struct. Funct. Genet.* **23,** 566–579 (Dec. 1995).

66. Frydman, J., Erdjument-Bromage, H., Tempst, P. & Ulrich Hartl, F. Co-translational domain folding as the structural basis for the rapid de novo folding of firefly luciferase. *Nat. Struct. Biol.* **6,** 697–705 (1999).

67. Ganesan, S. J. & Matysiak, S. Role of Backbone Dipole Interactions in the Formation of Secondary and Supersecondary Structures of Proteins. *J. Chem. Theory Comput.* **10,** 2569–2576 (June 2014).

68. Gao, W., Mahajan, S. P., Sulam, J. & Gray, J. J. *Deep Learning in Protein Structural Modeling and Design* Dec. 2020.

69. Gavrilov, Y., Dagan, S. & Levy, Y. Shortening a loop can increase protein native state entropy. *Proteins Struct. Funct. Bioinforma.* **83,** 2137–2146 (Dec. 2015).

70. Gerashchenko, M. V., Peterfi, Z., Yim, S. H. & Gladyshev, V. N. Translation elongation rate varies among organs and decreases with age. *Nucleic Acids Res.* **49,** e9 (Jan. 2021).

71. Giver, L., Gershenson, A., Freskgard, P. O. & Arnold, F. H. Directed evolution of a thermostable esterase. *Proc. Natl. Acad. Sci. U. S. A.* **95,** 12809–12813 (Oct. 1998).

72. Goldstein, R. A., Luthey-Schulten, Z. A. & Wolynes, P. G. Optimal protein-folding codes from spin-glass theory. *Biophysics (Oxf).* **89,** 4918–4922 (1992).

73. Goldstein, R. A. The evolution and evolutionary consequences of marginal thermostability in proteins. *Proteins Struct. Funct. Bioinforma.* **79,** 1396–1407 (May 2011).

74. Haimov, B. & Srebnik, S. A closer look into the $\alpha$-helix basin. *Sci. Rep.* **6,** 1–12 (Dec. 2016).

75. Hart, K. M. *et al.* Thermodynamic System Drift in Protein Evolution. *PLoS Biol.* **12** (ed Theobald, D. L.) e1001994 (Nov. 2014).

76. Hills, R. D., Brooks, C. L., Brooks, C. L. & III. Insights from coarse-grained Gō models for protein folding and dynamics. *Int. J. Mol. Sci.* **10,** 889–905 (Mar. 2009).

77. Hockenberry, A. J. & Wilke, C. O. Evolutionary couplings detect side-chain interactions. *bioRxiv,* 447409 (Dec. 2018).

78. Holm, L., Kääriäinen, S., Wilton, C. & Plewczynski, D. in *Curr. Protoc. Bioinforma.* (John Wiley & Sons, Inc., July 2006).

79. Holtkamp, W. *et al.* Cotranslational protein folding on the ribosome monitored in real time. *Science (80-. ).* **350,** 1104–1107 (Nov. 2015).

80. Hopf, T. A. *et al.* Three-Dimensional Structures of Membrane Proteins from Genomic Sequencing. *Cell* **149,** 1607–1621 (June 2012).

81. Hu, W., Kan, Z.-Y., Mayne, L. & Englander, S. W. Cytochrome c folds through foldon-dependent native-like intermediates in an ordered pathway. *Proc. Natl. Acad. Sci. U. S. A.* **113,** 3809–14 (Apr. 2016).

82. Hubbard, R. E. & Kamran Haider, M. in *Encycl. Life Sci.* (John Wiley & Sons, Ltd, Chichester, UK, Feb. 2010).

83. Jacobs, W. M. & Shakhnovich, E. I. Evidence of evolutionary selection for co-translational folding. *PNAS* **114,** 11434–11439 (Mar. 2017).

84. Jacobs, W. M. & Shakhnovich, E. I. Structure-Based Prediction of Protein-Folding Transition Paths. *Biophys. J.* **111,** 925–936 (Sept. 2016).

85. Jefferys, B. R., Kelley, L. A. & Sternberg, M. J. E. Protein Folding Requires Crowd Control in a Simulated Cell. *J. Mol. Biol.* **397,** 1329–1338 (Apr. 2010).

86. Jennings, L. E. *et al.* BET bromodomain ligands: Probing the WPF shelf to improve BRD4 bromodomain affinity and metabolic stability. *Bioorganic Med. Chem.* **26,** 2937–2957 (July 2018).

87. Joiret, M., Rapino, F., Close, P. & Geris, L. *Ribosome exit tunnel electrostatics* Oct. 2020.

88. Jones, D. T., Buchan, D. W. A., Cozzetto, D. & Pontil, M. PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. *Bioinformatics* **28,** 184–190 (Jan. 2012).

89. Jones, D. T., Singh, T., Kosciolek, T. & Tetchner, S. MetaPSICOV: combining coevolution methods for accurate prediction of contacts and long range hydrogen bonding in proteins. *Bioinformatics* **31,** 999–1006 (Apr. 2015).

90. Jubb, H. C. *et al.* Arpeggio: A Web Server for Calculating and Visualising Interatomic Interactions in Protein Structures. *J. Mol. Biol.* **429,** 365–371 (2017).

91. Kabsch, W. & Sander, C. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **22,** 2577–2637 (1983).

92. Kadokura, H. *et al.* Observing the nonvectorial yet cotranslational folding of a multidomain protein, LDL receptor, in the ER of mammalian cells. *Proc. Natl. Acad. Sci. U. S. A.* **117,** 16401–16408 (July 2020).

93. Kaján, L., Hopf, T. A., Kalaš, M., Marks, D. S. & Rost, B. FreeContact: fast and free software for protein contact prediction from residue co-evolution. *BMC Bioinformatics* **15,** 85 (Mar. 2014).

94. Kamisetty, H., Ovchinnikov, S. & Baker, D. Assessing the utility of coevolution-based residue-residue contact predictions in a sequence- and structure-rich era. *Proc. Natl. Acad. Sci.* **110,** 15674–15679 (Sept. 2013).

95. Karplus, M. The Levinthal paradox: Yesterday and today. *Fold. Des.* **2,** S69–S75 (June 1997).

96. Kayikci, M. *et al.* Visualization and analysis of non-covalent contacts using the Protein Contacts Atlas. *Nat. Struct. Mol. Biol.* **25,** 185–194 (Feb. 2018).

97. Kemp, G., Kudva, R., de la Rosa, A. & von Heijne, G. Force-Profile Analysis of the Cotranslational Folding of HemK and Filamin Domains: Comparison of Biochemical and Biophysical Folding Assays. *J. Mol. Biol.* **431,** 1308–1314 (Mar. 2019).

98. Kendrew, J. C. *et al.* A Three-Dimensional Model of the Myoglobin Molecule Obtained by X-Ray Analysis. *Nature* **181,** 662–666 (Mar. 1958).

99. Kim, D. E., DiMaio, F., Yu-Ruei Wang, R., Song, Y. & Baker, D. One contact for every twelve residues allows robust and accurate topology-level protein structure modeling. *Proteins Struct. Funct. Bioinforma.* **82,** 208–218 (Feb. 2014).

100. Kim, Y. E., Hipp, M. S., Bracher, A., Hayer-Hartl, M. & Ulrich Hartl, F. Molecular Chaperone Functions in Protein Folding and Proteostasis. *Annu. Rev. Biochem.* **82,** 323–355 (June 2013).

101. Knudsen, M. & Wiuf, C. The CATH database. *Hum. Genomics* **4,** 207 (Feb. 2010).

102. Kolb, V. A. Cotranslational Protein Folding. *Mol. Biol.* **35,** 584–590 (2001).

103. Kolb, V. A., Makeyev, E. V. & Spirin, A. S. Folding of firefly luciferase during translation in a cell-free system. *EMBO J.* **13,** 3631–3637 (Aug. 1994).

104. Komar, A. A., Kommer, A., Krasheninnikov, I. A. & Spirin, A. S. Cotranslational folding of globin. *J. Biol. Chem.* **272,** 10646–10651 (Apr. 1997).

105. Kudlicki, W., Chirgwin, J., Kramer, G. & Hardesty, B. Folding of an Enzyme into an Active Conformation While Bound as Peptidyl-tRNA to the Ribosome. *Biochemistry* **34,** 14284–14287 (1995).

106. Labean, T. H., Butt, T. R., Kauffman, S. A. & Schultes, E. A. Protein folding absent selection. *Genes (Basel).* **2,** 608–26 (Aug. 2011).

107. Land, A., Zonneveld, D. & Braakman, I. Folding of HIV-1 Envelope glycoprotein involves extensive isomerization of disulfide bonds and conformation-dependent leader peptide cleavage. *FASEB J.* **17,** 1058–1067 (June 2003).

108. Lane, N. & Martin, W. The energetics of genome complexity. *Nature* **467,** 929–934 (Oct. 2010).

109. Laurin, C. M. *et al.* Fragment-Based Identification of Ligands for Bromodomain-Containing Factor 3 of Trypanosoma cruzi. *ACS Infect. Dis.* (2020).

110. Law, E. C., de Oliveira, S., Kelm, S., Shi, J. & Deane, C. M. Investigating Cotranslational Folding in Membrane Proteins using Fragment-Based Structure Prediction. *Biophys. J.* **112,** 61a (Feb. 2017).

111. Lee, B.-C. & Kim, D. A new method for revealing correlated mutations under the structural and functional constraints in proteins. *Bioinformatics* **25,** 2506–2513 (Oct. 2009).

112. Leininger, S. E., Trovato, F., Nissley, D. A. & O'Brien, E. P. Domain topology, stability, and translation speed determine mechanical force generation on the ribosome. *Proc. Natl. Acad. Sci. U. S. A.* **116,** 5523–5532 (Mar. 2019).

113. Lella, M. & Mahalakshmi, R. Metamorphic Proteins: Emergence of Dual Protein Folds from One Primary Sequence. *Biochemistry* **56,** 2971–2984 (June 2017).

114. Levinthal, C. *How to Fold Graciously* in *Mössbauer Spectrosc. Biol. Syst.* (eds DeBrunner, J. T. P. & Munck, E.) (University of Illinois Press, Monticello, Illinois, USA, 1969), 22–24.

115. Li, G. & De Clercq, E. HIV Genome-Wide Protein Associations: a Review of 30 Years of Research. *Microbiol. Mol. Biol. Rev.* **80,** 679–731 (Sept. 2016).

116. Lindorff-Larsen, K., Piana, S., Dror, R. O. & Shaw, D. E. How fast-folding proteins fold. *Science (80-. ).* **334,** 517–520 (Oct. 2011).

117. Liutkute, M., Samatova, E. & Rodnina, M. V. *Cotranslational folding of proteins on the ribosome* Jan. 2020.

118. Loell, K. & Nanda, V. Marginal protein stability drives subcellular proteome isoelectric point. *Proc. Natl. Acad. Sci. U. S. A.* **115,** 11778–11783 (Nov. 2018).

119. Lu, H.-M. & Liang, J. A model study of protein nascent chain and cotranslational folding using hydrophobic-polar residues. *Proteins Struct. Funct. Bioinforma.* **70,** 442–449 (Aug. 2007).

120. Lumry, R. & Rajender, S. Enthalpy–entropy compensation phenomena in water solutions of proteins and small molecules: A ubiquitous properly of water. *Biopolymers* **9,** 1125–1227 (Oct. 1970).

121. Maity, H., Maity, M., Krishna, M. M., Mayne, L. & Englander, S. W. Protein folding: The stepwise assembly of foldon units. *Proc. Natl. Acad. Sci. U. S. A.* **102,** 4741–4746 (Mar. 2005).

122. Makeyev, E. V., Kolb, V. A. & Spirin, A. S. Enzymatic activity of the ribosome-bound nascent polypeptide. *FEBS Lett.* **378,** 166–170 (Jan. 1996).

123. Mann, M., Maticzka, D., Saunders, R. & Backofen, R. Classifying proteinlike sequences in arbitrary lattice protein models using LatPack. *HFSP J.* **2,** 396–404 (Dec. 2008).

124. Mantel, N. & Haenszel, W. Statistical Aspects of the Analysis of Data From Retrospective Studies of Disease. *JNCI J. Natl. Cancer Inst.* **22,** 719–748 (Apr. 1959).

125. Marks, D. S. *et al.* Protein 3D Structure Computed from Evolutionary Sequence Variation. *PLoS One* **6** (ed Sali, A.) e28766 (Dec. 2011).

126. McGuffin, L. J., Bryson, K. & Jones, D. T. The PSIPRED protein structure prediction server. *Bioinformatics* **16,** 404–405 (Apr. 2000).

127. Mercier, E. & Rodnina, M. V. Co-Translational Folding Trajectory of the HemK Helical Domain. *Biochemistry* **57,** 3460–3464 (June 2018).

128. Metzner, P., Schütte, C. & Vanden-Eijnden, E. Transition Path Theory for Markov Jump Processes. *Multiscale Model. Simul.* **7,** 1192–1219 (Jan. 2009).

129. Mignon, C. *et al.* Codon harmonization – going beyond the speed limit for protein expression. *FEBS Lett.* **592,** 1554–1564 (May 2018).

130. Mirny, L. A. & Shakhnovich, E. I. Universally conserved positions in protein folds: Reading evolutionary signals about stability, folding kinetics and function. *J. Mol. Biol.* **291,** 177–196 (Sept. 1999).

131. Morcos, F., Pagnani, A., *et al.* Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proc. Natl. Acad. Sci.* **108,** E1293–E1301 (2011).

132. Morcos, F., Schafer, N. P., Cheng, R. R., Onuchic, J. N. & Wolynes, P. G. Coevolutionary information, protein folding landscapes, and the thermodynamics of natural selection. *Proc. Natl. Acad. Sci. U. S. A.* **111,** 12408–13 (Aug. 2014).

133. Morrissey, M., Ahmed, Z. & Shakhnovich, E. The role of cotranslation in protein folding: a lattice model study. *Polymer (Guildf).* **45,** 557–571 (Jan. 2004).

134. Mouat, M. F. Dihydrofolate influences the activity of Escherichia coli dihydrofolate reductase synthesised de novo. *Int. J. Biochem. Cell Biol.* **32,** 327–337 (Mar. 2000).

135. Moult, J., Fidelis, K., Kryshtafovych, A., Schwede, T. & Tramontano, A. Critical assessment of methods of protein structure prediction (CASP)—Round XII. *Proteins Struct. Funct. Bioinforma.* **86,** 7–15 (Mar. 2018).

136. Mullet, J. E., Gamble Klein, P. & Klein, R. R. Chlorophyll regulates accumulation of the plastid-encoded chlorophyll apoproteins CP43 and D1 by increasing apoprotein stability. *Proc. Natl. Acad. Sci. U. S. A.* **87,** 4038–4042 (June 1990).

137. Murzin, A. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* **247,** 536–540 (Apr. 1995).

138. Naganathan, A. N. & Muñoz, V. Scaling of folding times with protein size. *J. Am. Chem. Soc.* **127,** 480–481 (Jan. 2005).

139. Nilsson, O. B., Hedman, R., *et al.* Cotranslational Protein Folding inside the Ribosome Exit Tunnel. *Cell Rep.* **12,** 1533–1540 (Sept. 2015).

140. Nilsson, O. B., Nickson, A. A., *et al.* Cotranslational folding of spectrin domains via partially structured states. *Nat. Struct. Mol. Biol.* **24,** 221–225 (Mar. 2017).

141. Nissley, D. A., Carbery, A., Chonofsky, M. & Deane, C. M. Ribosome occupancy profiles are conserved between structurally and evolutionarily related yeast domains. *Bioinformatics* (Jan. 2021).

142. Nissley, D. A., Sharma, A. K., *et al.* Accurate prediction of cellular co-translational folding indicates proteins can switch from post-to co-translational folding. *Nat. Commun.* **7** (2016).

143. Notari, L., Martínez-Carranza, M., Farías-Rico, J. A., Stenmark, P. & von Heijne, G. Cotranslational Folding of a Pentarepeat $\beta$-Helix Protein. *J. Mol. Biol.* **430,** 5196–5206 (Dec. 2018).

144. O'Brien, E. P., Vendruscolo, M. & Dobson, C. M. Prediction of variable translation rate effects on cotranslational protein folding. *Nat. Commun.* **3,** 1–9 (May 2012).

145. Oka, O. B. & Bulleid, N. J. *Forming disulfides in the endoplasmic reticulum* Nov. 2013.

146. Onuchic, J. N., Luthey-Schulten, Z. & Wolynes, P. G. Theory of Protein Folding: The Energy Landscape Perspective. *Annu. Rev. Phys. Chem* **48,** 545–600 (1997).

147. Orengo, C. A. & Taylor, W. R. *SSAP: sequential structure alignment program for protein structure comparison* Jan. 1996.

148. Ovchinnikov, S., Kamisetty, H. & Baker, D. Robust and accurate prediction of residue–residue interactions across protein interfaces using evolutionary information. *Elife* **3** (May 2014).

149. Ovchinnikov, S., Kinch, L., *et al.* Large-scale determination of previously unsolved protein structures using evolutionary information. *Elife* **4,** e09248 (Sept. 2015).

150. Ovchinnikov, S., Park, H., Kim, D. E., *et al.* Structure prediction using sparse simulated NOE restraints with Rosetta in CASP11. *Proteins* (2016).

151. Ovchinnikov, S., Park, H., Varghese, N., *et al.* Protein structure determination using metagenome sequence data. *Science (80-. ).* **355,** 294–298 (2017).

152. Panchenko, A. R., Luthey-Schulten, Z. & Wolynes, P. G. Foldons, protein structural modules, and exons (protein folding/folding domains/structural domains). *Biophysics (Oxf).* **93,** 2008–2013 (1996).

153. Park, H., Ovchinnikov, S., Kim, D. E., DiMaio, F. & Baker, D. Protein homology model refinement by large-scale energy optimization. *Proc. Natl. Acad. Sci. U. S. A.* **115,** 3054–3059 (Mar. 2018).

154. Pechmann, S. & Frydman, J. Evolutionary conservation of codon optimality reveals hidden signatures of cotranslational folding. *Nat. Struct. Mol. Biol.* **20** (2012).

155. Peters, T. & Davidson, L. K. The biosynthesis of rat serum albumin. In vivo studies on the formation of the disulfide bonds. *J. Biol. Chem.* **257,** 8847–8853 (Aug. 1982).

156. Piana, S., Lindorff-Larsen, K. & Shaw, D. E. Atomic-level description of ubiquitin folding. *Proc. Natl. Acad. Sci. U. S. A.* **110,** 5915–5920 (Apr. 2013).

157. Porter, L. L. & Looger, L. L. Extant fold-switching proteins are widespread. *Proc. Natl. Acad. Sci. U. S. A.* **115,** 5968–5973 (June 2018).

158. Punde, N., Kooken, J., Leary, D., Legler, P. M. & Angov, E. Codon harmonization reduces amino acid misincorporation in bacterially expressed P. falciparum proteins and improves their immunogenicity. *AMB Express* **9,** 167 (Dec. 2019).

159. Raschke, T. M. & Marqusee, S. Hydrogen exchange studies of protein structure. *Curr. Opin. Biotechnol.* **9,** 80–86 (Feb. 1998).

160. Remmert, M., Biegert, A., Hauser, A. & Söding, J. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nat. Methods* **9,** 173–175.

161. Riba, A. *et al.* Protein synthesis rates and ribosome occupancies reveal determinants of translation elongation rates. *Proc. Natl. Acad. Sci. U. S. A.* **116,** 15023–15032 (July 2019).

162. Richardson, J. S., Keedy, D. A. & Richardson, D. C. in *Biomol. Forms Funct. A Celebr. 50 Years Ramachandran Map* 46–61 (World Scientific Publishing Co., Jan. 2012).

163. Rotkiewicz, P. & Skolnick, J. Fast procedure for reconstruction of full-atom protein models from reduced representations. *J. Comput. Chem.* **29,** 1460–1465 (July 2008).

164. Samudrala, R. & Moult, J. An all-atom distance-dependent conditional probability discriminatory function for protein structure prediction. *J. Mol. Biol.* **275,** 895–916 (Feb. 1998).

165. Saunders, R., Mann, M. & Deane, C. M. Signatures of co-translational folding. *Biotechnol. J.* **6,** 742–751 (June 2011).

166. Savage, C. Depth-first search and the vertex cover problem. *Inf. Process. Lett.* **14,** 233–235 (1982).

167. Schauperl, M., Podewitz, M., Waldner, B. J. & Liedl, K. R. Enthalpic and Entropic Contributions to Hydrophobicity. *J. Chem. Theory Comput.* **12,** 4600–4610 (Sept. 2016).

168. Schuwirth, B. S. *et al.* Structures of the bacterial ribosome at 3.5 A resolution. *Science* **310,** 827–34 (Nov. 2005).

169. Scott, K. A., Batey, S., Hooton, K. A. & Clarke, J. The folding of spectrin domains I: Wild-type domains have the same stability but very different kinetic properties. *J. Mol. Biol.* **344,** 195–205 (Nov. 2004).

170. Seemayer, S., Gruber, M. & Söding, J. CCMpred—fast and precise prediction of protein residue–residue contacts from correlated mutations. *Bioinformatics* **30,** 3128–3130 (Nov. 2014).

171. Senior, A. W. *et al.* Improved protein structure prediction using potentials from deep learning. *Nature* **577,** 706–710 (Jan. 2020).

172. Shah, M. A., Mishra, S. & Chaudhuri, T. K. Marginal stability drives irreversible unfolding of large multi-domain family 3 glycosylhydrolases from thermo-tolerant yeast. *Int. J. Biol. Macromol.* **108,** 1322–1330 (Mar. 2018).

173. Shaw, D. E. *et al.* Atomic-Level Characterization of the Structural Dynamics of Proteins. *Science (80-. ).* **330** (2010).

174. Sheridan, R. *et al.* EVfold.org: Evolutionary Couplings and Protein 3D Structure Prediction. *bioRxiv* (2015).

175. Skwark, M. J., Abdel-Rehim, A. & Elofsson, A. PconsC: combination of direct information methods and alignments improves contact prediction. *Bioinformatics* **29,** 1815–1816 (July 2013).

176. Sohl, J. L., Jaswal, S. S. & Agard, D. A. Unfolded conformations of $\alpha$-lytic protease are more stable than its native state. *Nature* **395,** 817–819 (Oct. 1998).

177. Srivastava, S., Patton, Y., Fisher, D. W. & Wood, G. R. Cotranslational Protein Folding and Terminus Hydrophobicity. *Adv. Bioinformatics* **2011,** 1–8 (June 2011).

178. Stein, R. R., Marks, D. S. & Sander, C. Inferring Pairwise Interactions from Biological Data Using Maximum-Entropy Probability Models. *PLOS Comput. Biol.* **11** (ed Chen, S.-J.) e1004182 (July 2015).

179. Su, T. *et al.* The force-sensing peptide VemP employs extreme compaction and secondary structure formation to induce ribosomal stalling. *Elife* **6** (May 2017).

180. Sung, S. S. Peptide folding driven by Van der Waals interactions. *Protein Sci.* **24,** 1383–1388 (Sept. 2015).

181. Taverna, D. M. & Goldstein, R. A. Why are proteins marginally stable? *Proteins Struct. Funct. Genet.* **46,** 105–109 (Jan. 2002).

182. Taylor, W. R. Exploring Protein Fold Space. *Biomolecules* **10,** 193 (Jan. 2020).

183. Taylor, W. R., Jones, D. T. & Sadowski, M. I. Protein topology from predicted residue contacts. *Protein Sci.* **21,** 299–305 (Feb. 2012).

184. Thoden, J. B. *et al.* Structure of the $\beta$2 homodimer of bacterial luciferase from Vibrio harveyi: X-ray analysis of a kinetic protein folding trap. *Protein Sci.* **6,** 13–23 (1997).

185. Thommen, M., Holtkamp, W. & Rodnina, M. V. Co-translational protein folding: progress and methods. *Curr. Opin. Struct. Biol.* **42,** 83–89 (Feb. 2017).

186. Tiessen, A., Pérez-Rodríguez, P. & Delaye-Arredondo, L. Mathematical modeling and comparison of protein size distribution in different plant, animal, fungal and microbial species reveals a negative correlation between protein size and protein number, thus providing insight into the evolution of proteomes. *BMC Res. Notes* **5,** 85 (Dec. 2012).

187. To, P., Whitehead, B., Tarbox, H. E. & Fried, S. D. *Non-refoldability is pervasive across the E. coli proteome* Aug. 2020.

188. Tretyachenko, V. *et al.* Random protein sequences can form defined secondary structures and are well-tolerated in vivo. *Sci. Rep.* **7** (Dec. 2017).

189. Trevizani, R., Custódio, F. L., dos Santos, K. B. & Dardenne, L. E. Critical Features of Fragment Libraries for Protein Structure Prediction. *PLoS One* **12** (ed Zhang, Y.) e0170131 (Jan. 2017).

190. Tripathi, S., Makhatadze, G. I. & Garcia, A. E. Backtracking due to residual structure in the unfolded state changes the folding of the third fibronectin type III domain from tenascin-C. *J. Phys. Chem. B* **117,** 800–810 (Jan. 2013).

191. Trovato, F. & O'Brien, E. P. Fast Protein Translation Can Promote co- and post-translational Folding of Misfolding-Prone Proteins. *Biophys. J.* **112,** 1807–1819 (2017).

192. Tsutsui, Y., Dela Cruz, R. & Wintrode, P. L. Folding mechanism of the metastable serpin 1-antitrypsin. *Proc. Natl. Acad. Sci.* **109,** 4467–4472 (Mar. 2012).

193. Van Aalst, E., Yekefallah, M., Mehta, A. K., Eason, I. & Wylie, B. Codon Harmonization of a Kir3.1-KirBac1.3 Chimera for Structural Study Optimization. *Biomolecules* **10,** 430 (Mar. 2020).

194. Veis, A. & Kirk, T. Z. The coordinate synthesis and cotranslational assembly of type I procollagen. *J. Biol. Chem.* **264,** 3884–3889 (Mar. 1989).

195. Veis, A., Leibovich, S. J., Evans, J. & Kirk, T. Z. Supramolecular assemblies of mRNA direct the coordinated synthesis of type I procollagen chain. *Proc. Natl. Acad. Sci. U. S. A.* **82,** 3693–3697 (June 1985).

196. Venkatakrishnan, A. J. *et al. Uncovering patterns of atomic interactions in static and dynamic structures of proteins* Nov. 2019.

197. Vogl, T., Jatzke, C., Hinz, H. J., Benz, J. & Huber, R. Thermodynamic stability of annexin V E17G: Equilibrium parameters from an irreversible unfolding reaction. *Biochemistry* **36,** 1657–1668 (Feb. 1997).

198. Wang, S., Sun, S., Li, Z., Zhang, R. & Xu, J. Accurate De Novo Prediction of Protein Contact Map by Ultra-Deep Learning Model. *PLOS Comput. Biol.* **13** (ed Schlessinger, A.) e1005324 (Jan. 2017).

199. Wang, X., Ramírez-Hinestrosa, S., Dobnikar, J. & Frenkel, D. The Lennard-Jones potential: When (not) to use it. *Phys. Chem. Chem. Phys.* **22,** 10624–10633 (May 2020).

200. Williams, P. D., Pollock, D. D. & Goldstein, R. A. Functionality and the evolution of marginal stability in proteins: inferences from lattice simulations. *Evol. Bioinform. Online* **2,** 91–101 (Jan. 2007).

201. Wohlgemuth, I., Pohl, C., Mittelstaet, J., Konevega, A. L. & Rodnina, M. V. *Evolutionary optimization of speed and accuracy of decoding on the ribosome* 2011.

202. Wong, J. W., Ho, S. Y. & Hogg, P. J. Disulfide bond acquisition through eukaryotic protein evolution. *Mol. Biol. Evol.* **28,** 327–334 (2011).

203. Wozniak, P. P. & Kotulska, M. Characteristics of protein residue-residue contacts and their application in contact prediction. *J. Mol. Model.* **20,** 2497 (Nov. 2014).

204. Wright, P. E., Dyson, H. J. & Lerner, R. A. Conformation of peptide fragments of proteins in aqueous solution: implications for initiation of protein folding. *Biochemistry* **27,** 7167–7175 (Sept. 1988).

205. Xu, J. & Zhang, Y. How significant is a protein structure similarity with TM-score = 0.5? *Bioinformatics* **26,** 889–95 (Apr. 2010).

206. Yang, W. Y. & Gruebele, M. Folding at the speed limit. *Nature* **423,** 193–197 (May 2003).

207. Yi, Q. & Baker, D. Direct evidence for a two-state protein unfolding transition from hydrogen-deuterium exchange, mass spectrometry, and NMR. *Protein Sci.* **5,** 1060–1066 (1996).

208. Yu, C.-H. *et al.* Codon Usage Influences the Local Rate of Translation Elongation to Regulate Co-translational Protein Folding. *Mol. Cell* **59,** 744–54 (Sept. 2015).

209. Yu, W., Baxa, M. C., Gagnon, I., Freed, K. F. & Sosnick, T. R. Cooperative folding near the downhill limit determined with amino acid resolution by hydrogen exchange. *Proc. Natl. Acad. Sci. U. S. A.* **113,** 4747–52 (Apr. 2016).

210. Yusupova, G. & Yusupov, M. High-Resolution Structure of the Eukaryotic 80S Ribosome. *Annu. Rev. Biochem.* **83,** 467–486 (June 2014).

211. Zemla, A. *et al.* LGA: a method for finding 3D similarities in protein structures. *Nucleic Acids Res.* **31,** 3370–3374 (July 2003).

212. Zhang, Y. & Skolnick, J. Scoring Function for Automated Assessment of Protein Structure Template Quality. *Proteins* **57,** 702–710 (2004).

213. Zhou, X., Hu, J., Zhang, C., Zhang, G. & Zhang, Y. Assembling multidomain protein structures through analogous global structural alignments. *Proc. Natl. Acad. Sci. U. S. A.* **116,** 15930–15938 (Aug. 2019).

214. Zhu, C. *et al.* Characterizing hydrophobicity of amino acid side chains in a protein environment via measuring contact angle of a water nanodroplet on planarpeptide network. *Proc. Natl. Acad. Sci. U. S. A.* **113,** 12946–12951 (Nov. 2016).

215. Žoldák, G., Stigler, J., Pelz, B., Li, H. & Rief, M. Ultrafast folding kinetics and cooperativity of villin headpiece in single-molecule force spectroscopy. *Proc. Natl. Acad. Sci. U. S. A.* **110,** 18156–18161 (Nov. 2013).