Covariation-derived residue contacts in *ab initio* modelling and Molecular Replacement

Felix Šimkovic

Thesis submitted in accordance with the requirements of the
University of Liverpool
for the degree of
Doctor in Philosophy

September 2018

Institute of Integrative Biology
University of Liverpool
United Kingdom

Covariation-derived residue contacts in *ab initio* modelling and Molecular Replacement

Felix Šimkovic

This thesis is concerned with the application of predicted residue contacts in *ab initio* protein structure prediction and Molecular Replacement (MR).

Initially, in Chapter 3, research explored the use of predicted residue contacts to improve *ab initio* protein structure predictions, which were used to generate AMPLE ensemble search models for MR. The results proved highly encouraging: four additional targets were tractable where previous AMPLE attempts were unable to achieve structure solution. Furthermore, a novel approach to enhance β-rich decoy quality proved critical for an additional structure solution.

Leading on from the work in Chapter 3, it was essential to investigate different contact prediction algorithms and ROSETTA distance-restraints energy functions to optimise decoy quality. Results presented in Chapter 4 supported previous findings, which claimed that METAPSICOV produced the most precise contact predictions. Furthermore, results showed that target-specific decoy quality may be affected by the ROSETTA distance-restraint energy function used, which also translated into MR structure solutions in AMPLE.

Beyond different contact prediction algorithms and ROSETTA distance-restraint energy functions, alternative protein structure prediction algorithms exist. In Chapter 5, a study to compare the most promising alternatives to ROSETTA was conducted to explore potential alternatives for AMPLE. However, ROSETTA remained the optimal structure prediction algorithm to maximise structure solutions in AMPLE. A promising fragment-independent alternative, CONFOLD2, generated similarly accurate decoys, but the resulting AMPLE ensembles did not produce successful MR structure solutions.

AMPLE's cluster-and-truncate routine was originally developed to process contact-unassisted decoys. However, more accurate starting decoys, such as those deriving from contact-assisted modelling, may require processing differently to generate the ensemble search models. The findings in Chapter 6 demonstrated that decoy quality could be reliably predicted by measuring the satisfaction of the long-range contact predictions used initially to restrain the folding procedure. Excluding the decoys that satisfied the fewest long-range contacts enabled further structure solutions of targets that were previously intractable.

Lastly, in Chapter 7, contact-driven selection of supersecondary structure elements or subfolds identified by fragment picking software was explored as a novel route to search models for unconventional MR. Preliminary results of this approach showed promise. Two out of four protein targets were solved with fragments extracted from unrelated protein targets which, crucially, satisfied many predicted residue contacts.

Acknowledgements

First and foremost, I would like to express my gratitude to Dr Daniel Rigden for his supervision throughout my PhD and previous projects. His experience and guidance was invaluable, and I strongly believe that I am going to benefit from what I have learned under his supervision in years to come. I would also like to thank him for his never-ending support and patience when computer programming seemed more important than research to me.

I would also like to thank my secondary supervisor Prof Olga Mayans. Six years ago, Olga gave me my first Bioinformatics research experience, and sparked an interest in computational work that lasts to this date. I would also like to thank her for her continued support and critical opinion throughout all of my projects.

Besides my supervisors, I would like to thank Dr Jens Thomas for his help throughout my PhD project. His patience was always appreciated when simple concepts proved difficult to grasp, and his critical mind and strive for simplicity also helped me to becoming a better researcher and Software Engineer, for which I am very grateful. I would also like to thank Dr Ronan Keegan, whose support was essential to understanding the mysterious and complex world of X-ray crystallography a little better. I would also like to thank all other members of the CCP4 and CCP-EM core teams for their help and support with all software related issues.

My thanks also go out to my funding bodies, the BBSRC and CCP4, without whose fincancial support this research could have not been conducted.

Last but not least I would like to thank my partner Joanna Lorek, my father Peter, stepmother Petra, grandparents Viera and Peter Šimkovic, and the rest of my family for their selfless support, help and patience whenever I needed it. All sacrificed significant parts of their lives to making my achievements possible, and thus this work is as much theirs as it is mine. My gratitude to all is beyond words.

To the memory of my grandmother
Dr Viera Šimkovic
(1940-2016)

Contents

List of Figures

List of Tables

List of Equations

List of Abbreviations

ACL	Average Chain Length
APC	Average Product Correction
CC	Correlation Coefficient
CMO	Contact Map Overlap
CNS	Crystallography & NMR System
DCA	Direct Coupling Analysis
EC	Evolutionary Coupling
eLLG	expected Log-Likelihood Gain
FP	False Positive
KDE	Kernel Density Estimate
LLG	Log-Likelihood Gain
M_{eff}	Number of Effective Sequences
MAE	Mean Absolute Error
MR	Molecular Replacement
MSA	Multiple Sequence Alignment
MX	Macromolecular Crystallography
NMR	Nuclear Magnetic Resonance
NOE	Nuclear Overhauser Effect
PDB	Protein Data Bank

PDBTM	Protein Data Bank of Transmembrane Proteins
RIO	Residue-Independent Overlap
RMSD	Root-Mean-Square Deviation
SML	Supervised Machine Learning
TFZ	Translation Function Z-score
TM-score	Template-Modelling score
TP	True Positive

Chapter 1

Introduction

1.1 Macromolecular X-ray crystallography

The discovery of X-ray diffraction by crystals by Max van Laue [1, 2] marked the origins of modern crystallography. However, it was not until the work of William Lawrence Bragg and William Henry Bragg that X-ray scattering could be interpreted as atomic positions [3–5]. Since then, X-ray crystallography and the determination of atomic positions in organic and inorganic molecules has come a long way and shaped the path for many 21st century discoveries. Amongst those groundbreaking discoveries are the earliest structural models of biological molecules including DNA [6], vitamin B12 [7], and the first protein structures [8–11]. These structure elucidations hallmarked the dawn of a new era in biological and biomedical research. At the time of writing, 124,551 structural models deposited in the Protein Data Bank (PDB) were determined by X-ray diffraction studies [12], and thus X-ray crystallography is a key method in biological research.

1.1.1 X-ray scattering

X-rays are high energy photons part of the electromagnetic spectrum with a wavelength of 0.1-100Å [13]. X-rays can be described as packets of travelling electromagnetic waves, whose electric field vector interacts with the charged electrons of matter [13]. Such interaction, typically termed scattering, results in the diffraction of the incoming wave, which X-ray crystallography relies on.

In its simplest form, scattering of X-ray radiation can be explained in the scenario of exposure to a single free electron. The resulting scattering can be classed as elastic (Thomson scattering) or inelastic (Crompton scattering) [13]. The latter — scattering that results in a loss of energy of the emitting photon due to energy transfer onto the electron — does not contribute to discrete scattering, the type of scattering X-ray diffraction relies on. In comparison, Thomson scattering does not result in a loss of energy of the emitting photon. This has significant effects, the incoming photon emits with the same frequency causing the electron to oscillate identically further enhancing the signal.

If we expand the example to include all electrons in an atom and expose the atom to X-ray radiation, our theory needs to be slightly expanded. Given that one or more electrons in an atom are not free but orbit around the atom's nucleus in a stable and defined manner, the distribution of these electrons around the nucleus determines the scattering of the incoming X-ray photons. The distribution of scattered photon waves is thus an overall representation of the probability distributions of each electron in the atom and is referred to as electron density $\rho(\boldsymbol{r})$. In X-ray scattering, it suffices to approximate the shape of the electron density to a sphere. If we now consider the emitting wave $\boldsymbol{s_1}$ of an X-ray photon scattered by any position $\boldsymbol{r}$ in the electron density

of an atom, then the phase difference $\Delta\varphi$ to the incoming wave $\boldsymbol{s_0}$ can be described by Eq. 1.1 [13].

$$\Delta\varphi = 2\pi\left(\boldsymbol{s_1} - \boldsymbol{s_0}\right)\boldsymbol{r} = 2\pi \cdot \boldsymbol{S}\boldsymbol{r} \tag{1.1}$$

If more than one electron in an atom's electron density scatter the incoming X-ray photon wave, then the emitting partial waves can be described by the atomic scattering factor f_s (Eq. 1.2), which describes the interference of all scattered waves [13]. The total scattering power of an atom is proportional to the number of electrons and element-specific with heavier atoms scattering more strongly. Given the approximation of a centrosymmetric electron density, the atomic scattering function is also symmetric.

$$f_s = \int_{\boldsymbol{r}}^{V(atoms)} \rho\left(\boldsymbol{r}\right) \cdot \exp\left(2\pi i \boldsymbol{S}\boldsymbol{r}\right) \cdot d\boldsymbol{r} \tag{1.2}$$

With an enhanced understanding of X-ray scattering of electrons orbiting a single atom, it is important to consider X-ray scattering of adjacent atoms, such as it is typically found in molecules. If the electromagnetic wave of an X-ray photon excites all electrons of adjacent atoms, then the resulting partial waves — amplified by oscillations of electrons of Thomson scattering — result in constructive or destructive interference. Maximal interference can be obtained when all partial waves are in-phase, and maximal destructive interference when out-of-phase. This leads to varying intensities of the emitting X-ray photon at different points in space. To obtain the overall scattering power F_s of all contributing atoms, Eq. 1.2 needs to be modified to include the sum over all atoms j as described in Eq. 1.3.

$$F_s = \sum_{j=1}^{atoms} f_{s,j}^{0} \cdot \exp\left(2\pi i \boldsymbol{S}\boldsymbol{r}_j\right) \tag{1.3}$$

If we now translate our hypothetical experiment into a crystal lattice then our understanding described in Eq. 1.3 needs to be expanded from a 1-dimensional distance vector $\boldsymbol{r}$ to the three dimensional lattice translation vectors $\boldsymbol{a}$, $\boldsymbol{b}$ and $\boldsymbol{c}$. The Laue equations (Eq. 1.4) do exactly that and ultimately determine the positions of the diffraction peaks in 3-dimensional space.

$$\boldsymbol{S} \cdot \boldsymbol{a} = n_1, \qquad \boldsymbol{S} \cdot \boldsymbol{b} = n_2, \qquad \boldsymbol{S} \cdot \boldsymbol{c} = n_3 \tag{1.4}$$

Such determination is possible through the findings made by Bragg and Bragg [3], who identified the relationship between the scattering vector $\boldsymbol{S}$ and the planes in the crystal lattice. Today, this relationship is defined by the Bragg equation (Eq. 1.5) [3], which allows us to interpret X-ray diffraction as reflections on discrete lattice planes, which relates the diffraction angle θ to the lattice spacing d_{hkl} (Fig. 1.1) [13]. For maximum diffraction n needs to be an integer multiple to result in maximum constructive interference of wavelength λ.

$$n\lambda = 2d_{hkl}sin\theta \tag{1.5}$$

If the hypothetical model is expanded to molecular crystals, then the total scattering from the unit cell is merely a summation of all molecular unit cell scattering contributions in the crystal. Mathematically, this results in Eq. 1.3 being generalised to Eq. 1.6 through the application of the Laue equations (Eq. 1.4). This allows us to express the scattering vector $\boldsymbol{Sr}_j$ as Miller indices of the reflection planes $\boldsymbol{hx}_j$.

$$F_h = \sum_{j=1}^{atoms} f_{s,j}^0 \cdot \exp\left(2\pi i \boldsymbol{hx}_j\right) \tag{1.6}$$

The structure factor equation defines the scattering power from a crystal in a given reciprocal lattice direction $\boldsymbol{h}$. The scattering is enhanced by the number of repeating units of lattice translation vectors $\boldsymbol{a}$, $\boldsymbol{b}$ and $\boldsymbol{c}$, and thus the overall scattering power is proportional to the number of unit cells in the crystal.

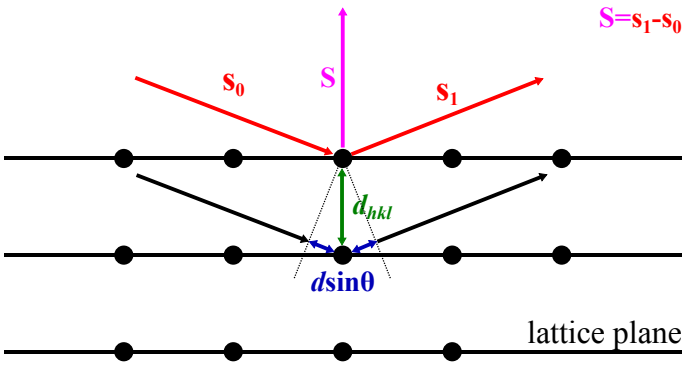


Figure 1.1: Schematic of Bragg scattering.

It should be noted that Eq. 1.6 is a simplification of the problem at hand. In reality, instrument and experimental corrections need to be applied to the structure factor equation. A correction factor for each experiment-dependent parameter needs

to be applied to the structure factor equation. However, in the scope of this work the details of such correction factors do not need to be discussed.

Since complex structure factors describe the molecular structure in the reciprocal space domain, the conversion to the real space domain in form of electron density is required. This can be conveniently done through the bijective Fourier transform, which allows the conversion of complex structure factors to electron density and vice versa without the loss of any information [13]. Thus, electron density can be obtained from the complex structure factors using Eq. 1.7. The normalisation factor $1/V$ (V represents the volume of the unit cell) provides the correct units for the electron density $\rho(x, y, z)$.

$$\rho(x,y,z) = \frac{1}{V} \sum_{h=0}^{+\infty} \sum_{k=-\infty}^{+\infty} \sum_{l=-\infty}^{+\infty} \boldsymbol{F}(hkl) \cdot \exp\left(-2\pi i(hx + ky + lz)\right) \tag{1.7}$$

1.1.2 From crystal to structure

In X-ray crystallographic experiments, X-ray radiation is measured using light detectors. However, the measurement taken is incomplete. Light detectors only capture the intensity of the scattered X-ray photons but crucially lose the phase information. The latter is essential for atomic reconstruction of the crystallised molecule, and thus needs to be obtained. In Macromolecular Crystallography (MX), experimentalists have a number of alternative techniques to compensate for the lost phase information.

Prior to the big advances in computing power and the successful elucidation of many protein structures, MX crystallographers primarily recovered the lost phase information through direct methods or experimental methods [14]. Today, the most popular method to recovering the lost phase information is Molecular Replacement (MR) [15, 16]. In a MR search, a known structure ('search model') similar to the unknown is relocated in the unit cell until the solution with the best fit between calculated and observed diffraction data is obtained [13]. A 6-dimensional search, i.e. a simultaneous rotation and translation search, is possible [17–19], but is computationally very expensive and less suitable for challenging cases. In comparison, most modern crystallographic applications opt for two distinct sub-searches, the rotation search to orient the search model within the unit cell followed by the translation search to locate it [13]. The benefits over a combined search include search-specific target functions that enable increased sensitivity and additional terms to compensate for imperfect data.

The most successful MR algorithms perform the rotation and translation searches using Patterson methods or Maximum Likelihood functions. Patterson methods — originally developed by Rossmann and Blow [20] — rely on the use of a map of vectors

between the scattering atoms, which can be determined for the calculated and observed structure factor amplitudes. Patterson vectors can be subclassed as intra- and inter-molecular vectors. A distinct separation of the observed vectors is impossible. However, inter-molecular vectors appear further away from the central peak of the self-vector (vector from atom to itself) in the Patterson map [13]. The calculated Patterson vectors for the search model allow for a clearer distinction between the intra- and inter-molecular vectors. If the search model is placed in a large unit cell, then inter-molecular vectors must scale with the unit cell dimension [13]. Ultimately, using the Patterson vectors around the origin (the set that predominantly comprises the intra-molecular Patterson vectors) the search model can be oriented against the experimentally determined Patterson vectors to identify the optimal overlap between the observed and model Patterson vectors [14]. In a similar manner, the inter-molecular vectors can be used to identify the correct translation of the search model. Patterson methods are very sensitive to small orientation errors of the search model [13]. Thus, orientations with the highest vector peak overlaps are trialled in the subsequent translation search. Given that Patterson methods operate by Patterson vector comparisons in rotation and translation searches, these methods do not require search-model-derived phases.

In comparison to the Patterson methods, Maximum Likelihood methods do not rely on inter-atomic vectors in Patterson maps. Instead, Maximum Likelihood methods make use of Bayes' theorem [21] to compare calculated structure factors and observed structure factor amplitudes directly [19]. Bayes' theorem in crystallographic Maximum Likelihood methods is applied to compute the likelihood that an experimental value is observed given the current search model. The maximal likelihood indicates the optimal orientation and translation of the search model given the observed experimental data [14]. Since the search model likelihood term is the product of many individual probabilities, which are difficult to represent computationally due to floating point representations, the log of the likelihood is commonly used [13]. The major advantage of Maximum Likelihood methods over Patterson methods centres on the more realistic target functions, which consider errors and incompleteness of the search model, apply bulk solvent correction and conduct multi-model searches [14]. The latter is of particular relevance since the Maximum Likelihood rotation function can thus consider already placed search models in a fixed position whilst trialling additional ones [22], which proves to be a major advantage over Patterson methods. Furthermore, likelihood target functions can consider the structural variance of multiple superposed models in an ensemble search model, which is used to weight structure factors at the various positions to improve the overall likelihood term [19].

The initial electron density map — regardless of its determination by MR, direct or experimental methods — is almost always inaccurate. In MR, inaccuracies arise from experimental errors, model incompleteness, low signal-to-noise or model bias. Thus, approaches for improving the phases used to calculated the initial electron density map have been developed and are routinely applied in MX. Density modification describes

a set of methods that improve the obtained electron density typically by applying statistical corrections to electron density distributions. These corrections are based on prior knowledge or assumptions of the physical properties of macromolecular structures [13]. This process can transform initially poor or uninterpretable electron density maps to high quality ones. Three predominant density modification approaches exist: solvent flattening, histogram matching and the "sphere-of-influence" method. Solvent flattening is an approach was first proposed by Wang [23]. In solvent flattening, the disorder in the solvent region in a protein crystal is exploited, which differs in electron density from macromolecule-containing regions. If solvent electron density is set to a constant, then it is essentially flattened which will result in improved structure factors with improved phases and thus improved electron density. Histogram matching [24] exploits the defined characteristics of an electron density distribution determined from sets of proteins at the same resolution, irrespective of individual structural details. The electron density distribution for noisy maps are Gaussian-shaped. In contrast, the electron density distribution of a feature-defined map is positively skewed. Thus, attempting to improve the Gaussian-shaped electron density distribution to better match the positively skewed shape results in overall improvements to the electron density. The "sphere-of-influence" method was introduced by Sheldrick [25] and classifies solvent and protein electron density by observing its variance across the shell surface of a 2.42Å sphere (dominant 1-3 atom distance in macromolecular structures). If the sphere is positioned in the disordered solvent region typically found in inter-molecular channels, the density variance will be low. Thus, this approach allows to smoothen solvent-containing regions of the electron density [25]. Independent of the density modification strategy applied, it is important to understand that improvements to the electron density map anywhere lead to improvements everywhere by transferral of information from one part of the map to another [26].

A second approach to improving the initial electron density is termed Refinement. Iteratively, the placed search model is optimised to better explain the experimentally observed data. This optimisation problem is typically broken down into three main steps: the definition of the model parameters, the scoring function and the optimisation method. The model parameters describe the crystal and its content and can be subdivided into atomic and non-atomic model parameters [27]. These parameters combined are used to score the current model. The scoring function relates the experimental data to the model parameters. The scoring function contains two primary terms, the refinement data target and an *a priori* knowledge term. The former defines a target function that assesses the similarity between calculated and experimental structure factors. The target function is commonly a Maximum Likelihood-based function that considers missing or incomplete data [27, 28]. The *a priori* knowledge term in the scoring function defines the properties of a good model by including stereochemical property terms. Lastly, optimisation methods provide tools to vary the model parameters to better fit the experimental data. Different optimisation techniques can be used depending on the severity of model parameter alteration, which generally depend

on the entrapment of states in local energy minima. Model parameterisation and its scoring against the predefined scoring function combined with model optimisation form a refinement macrocycle, which is iteratively used to optimise a model's fit to the experimental data. This ultimately improves both the electron density map interpretability and model quality. MX refinement can be performed in structure-factor-based reciprocal space and electron-density-based real space [27]. A combination allows global and local refinement strategies and enables grid-like searches to optimise the model parameters until convergence.

Once initial phase information is improved through refinement and/or density modification, attempts can be made to build atomic model coordinates into the electron density map. This process is typically coupled with refinement or density modification to iteratively improve the quality of the partially built model and the electron density map [13]. A small number of distinct algorithms are currently used to automatically build atomic coordinates into electron density: main-chain autotracing [29], fitting pseudo-atoms into electron density [30], or fitting reference coordinates with similar electron density maps [31, 32]. In essence, all algorithms attempt to maximise the number of correctly identified and placed atomic coordinates into available electron density. Whilst autotracing solely builds main-chain polypeptides, the other two approaches rely on sequence information to also build side-chains. Independent of the complexity of the model building task, the higher the resolution and the more complete the initial starting model, the less ambiguous and challenging this task becomes [13].

1.1.3 Unconventional Molecular Replacement

The process of macromolecular structure determination via conventional MR has been outlined previously. Search models are typically derived from structural homologs identified by sequence identity to the crystallised target [13]. However, homologous structures are not always available or impossible to identify by current approaches. Direct or experimental phasing approaches to circumvent the absence of MR templates can be expensive, unsuccessful and very challenging for certain protein targets, and thus remain infeasible to pursue at times. Under such circumstances, alternative approaches are required, which are referred to as "unconventional" MR approaches from here onwards. The unconventional MR approach most relevant to the work presented in this thesis utilises the 3-dimensional structure prediction of a protein target starting from its sequence [33–35]. Although two distinct methods exist to predict the protein structure of a target sequence, homology modelling and *ab initio* structure prediction, only the latter is relevant to this work since the former relies on homologous structures.

1.2 *Ab initio* protein structure prediction

The folding of protein structures is commonly described by the folding funnel hypothesis [36]. It assumes that the native state of a protein fold corresponds to its global minimum free-energy state along its energy surface (Fig. 1.2) [37]. *In silico* protein folding experiments attempt to find this lowest free-energy state of the protein fold. However, to unambiguously identify this state sampling of all polypeptide chain conformations is necessary. In theory, sampling of all conformations for a 100-residue protein takes in the order of approximately 10^{52} years (10^{7} configurations with 10^{-11} seconds per configuration), yet *in vivo* an equivalent polypeptide chain folds in milliseconds to seconds [38, 39]. This paradox — termed the Levinthal paradox [38] — created the basis for the folding funnel hypothesis.

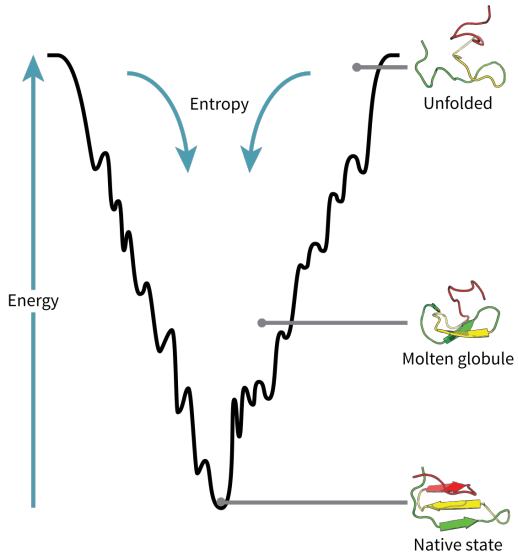


Figure 1.2: Schematic of the folding funnel hypothesis [36]. Diagram produced by Wikipedia contributors.

In *ab initio* protein structure prediction, the tertiary structure of a protein is predicted using its primary structure alone. This problem is in its nature identical to finding the lowest free-energy state along a protein's energy landscape. However, in an attempt to avoid the Levinthal paradox, different knowledge- and physics-based energy functions coupled with a variety of conformational search sampling algorithms are employed [40].

Physics-based energy functions use physiochemical force fields typically coupled with Molecular Dynamics simulations to sample the folding trajectory of a protein sequence (true physics-based approaches are computationally intractable because quantum mechanics models would need to be used). Force fields describe parameter sets used to calculate energy potentials for a system of atoms in a simulation run, and include potentials, such as van der Waals and electrostatic interactions [40]. In the context

of *ab initio* protein structure prediction, pure physics-based approaches are often less favourable, because the computational complexity to find the lowest free-energy state of a large protein structure remains intractable without the use of supercomputers.

Knowledge-based energy functions rely on empirical energy terms derived from statistics and regularities of experimentally determined structures [40]. These energy terms can be subdivided into two types, the generic or sequence-independent terms and amino-acid or sequence-dependent terms [41]. The former include terms to describe the backbone hydrogen-bonds and local backbone stiffness of a polypeptide chain. The sequence-dependent terms include terms such as pairwise residue contact potential, distance-dependent atomic contact potential, and secondary structure propensities. However, predicting local or global tertiary structure of a protein sequence using empirical energy terms alone is very difficult. Subtle differences in the local and global environment of a primary structure alongside the subtle differences in initial folds leading to common secondary structure features are very difficult to reproduce in a modelling scenario. Thus, knowledge-based energy functions are often coupled with the assembly of fragments extracted from other protein structures to predict the unknown tertiary structure of the target sequence [40].

The most successful *ab initio* structure prediction protocols use knowledge-based and physics-based energy functions combined with fragment-assembly-based conformational searches to find the lowest free-energy state [42–46]. Structural fragments of varying lengths (typically 3-20 residues) are extracted from existing protein structures [47–54]. These fragments are used in a Monte Carlo simulation to search the conformational space of the polypeptide chain for low free-energy states [55]. The insertion of overlapping fragments results in the replacement of torsion angles either at random positions or sequentially from predefined starting position (such as N- or C-termini). Each move is scored against the Metropolis criterion [55] consisting of knowledge-based and physics-based terms. The Metropolis criterion is typically defined to accept fragment insertions that lower the free-energy term of a decoy, whilst sometimes accepting insertions that increase the free-energy term to escape local energy minima. If the insertion of a fragment passed the Metropolis criterion, the related torsion angles are accepted and integrated in the polypeptide chain for the next fragment-insertion iteration. This process is repeated until convergence of the decoy, i.e. no lower free-energy state can be found. In all routines, these steps are independently repeated thousands of times to create a pool of decoys.

In order to identify the correct fold amongst the thousands of generated decoys, clustering approaches are often used in combination with *ab initio* protein structure prediction protocols. Shortle et al. [56] identified that the most-similar decoy to the native structure is most often the centroid (decoy with most neighbours in the cluster) of the largest cluster. Further studies showed that the selection of those centroid decoys helps to identify the most native-like folds amongst the many thousands generated [57–

59]. Some protocols use clustering as an intermediate or final step to identify decoys for which it will perform more computationally demanding all-atom refinement [58] or other decoy hybridisation techniques [43, 60, 61] to further approach the native-like fold [62].

Despite active research in *ab initio* protein structure prediction over decades, all approaches struggle with accurate predictions for larger protein domains (chain lengths > 150 residues) [58, 63–65]. The major challenge arises from the sampling of the conformational space since incorrect local changes influence the global structure. Furthermore, β-sheets are inherently difficult to predict given that β-strands in fragment-based approaches are inserted one at a time yet rely on the hydrogen-bond network typically found in β-sheets to reduce the overall energy of the decoy [42]. To address this issue, Lange et al. [66], Raman et al. [67] and Göbl et al. [68] started to use Nuclear Overhauser Effect (NOE) data as residue-residue distance restraints to reduce the sampling space of conformations, which enabled high-resolution predictions of tertiary structure for longer proteins. Although successfully applied in the aforementioned examples, experiments to collect NOE data are costly, challenging and intractable for larger multi-domain targets. To avoid this problem yet obtain similarly useful information on spatial proximity of amino acids in a protein fold, researchers started to exploit residue-residue contact information, which enables accurate *ab initio* structure prediction for longer polypeptide chains [e.g., 45, 46, 69–76].

1.3 Residue-residue contact prediction

The use of predicted residue-residue contact information to reduce the conformational search space in *ab initio* protein structure prediction relies on accurate identification of amino acids in close spatial proximity. Today, such identification can be detected from sequence information alone by either Direct Coupling Analysis (DCA) or Supervised Machine Learning (SML) algorithms.

1.3.1 Direct Coupling Analysis

Direct Coupling Analysis uses protein sequence information to identify coordinated changes of amino acids in sequences of a protein family (Fig. 1.3). These coordinated changes are caused by evolutionary pressure to maintain residue interactions important for protein structure and function. However, original attempts to detect covariation signal from sequences in a protein family were unsuccessful for many years [77–80]. The applied local statistical model suffered from numerous drawbacks, including the loss of covariation signal due to phylogenetic dependencies, limited availability of sequence data, and the potentially false assumption that truly coevolved residues are in close proximity in sequence space [81–83]. Implementations of the local statistical model

used raw covariation frequencies between pairs of positions in the sequence alignment. This further poses issues since successful distinction between "direct" casual (A-B and B-C) and "indirect" transitive (A-C) correlations is essential for successful protein structure prediction yet cannot be separated by frequency comparisons.

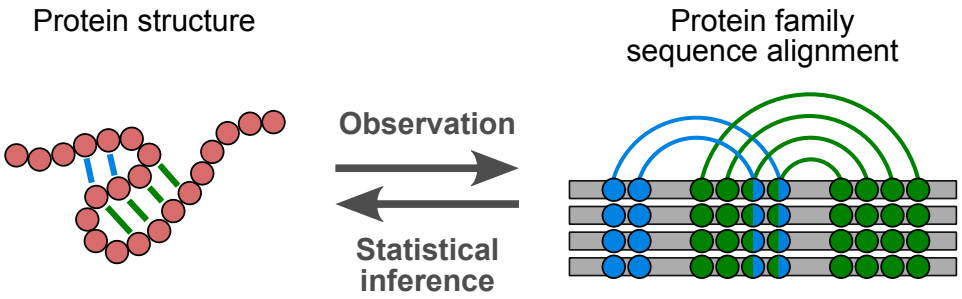


Figure 1.3: Schematic of inference of covariance signal originating from evolutionary pressure in protein tertiary structures and encoded in its family's sequence alignment (adapted from [84]).

Lapedes et al. [82] proposed the use of a global statistical model to infer correlations of residue pairs to circumvent the main problem of decoupling causal and transitive correlations. However, it was not until a decade later before first implementations of the global statistical model surfaced to successfully disentangle these types of correlations [69, 85–92]. The use of a global statistical model achieves successful disentanglement by inferring a probabilistic description of the sequence alignment that best explains observed correlations using underlying causal couplings between positions [93]. Such couplings can be inferred by maximising the likelihood of observing the sequences in the alignment under the maximum entropy probability model. In other words, by considering all amino acid pair positions simultaneously, causal and transitive couplings can be successfully disentangled [90].

The pairwise probabilistic model $P(\boldsymbol{\sigma})$ of the amino acid sequence $\boldsymbol{\sigma} = (\sigma_1, \sigma_2, \ldots, \sigma_N)$ of length N is defined in Eq. 1.8, which contains the amino acid configuration constraints σ_i and σ_j at positions i and j, the single-site conservation bias term h_i, and co-conservation term J_{ij} between position pairs i, j.

$$P(\boldsymbol{\sigma}) = \frac{1}{Z} \exp\left(\sum_{i=1}^{N} h_i(\sigma_i) + \sum_{1 \leqslant i < j \leqslant N} J_{ij}(\sigma_i, \sigma_j)\right) \tag{1.8}$$

$$Z = \sum_{\boldsymbol{\sigma}} \exp\left(\sum_{i=1}^{N} h_i(\sigma_i) + \sum_{1 \leqslant i < j \leqslant N} J_{ij}(\sigma_i, \sigma_j)\right) \tag{1.9}$$

The partition function Z (Eq. 1.9) acts as normalising constant, and additionally has the property to maximise the entropy in the probabilistic model. However, the computation of Z is intractable for the feature space found in DCA since the number of summations in Z exponentially increases with N for all 20 amino acid configurations. Thus, approximations of Z are typically used, which were shown to lead to precise covariance predictions [90].

Over the last decade, numerous approximations for the parameter inference of $P(\boldsymbol{\sigma})$ have been implemented, which include gradient ascent with Monte Carlo sampling [83], message passing [85], mean-field [69, 88, 89, 94], and pseudolikelihood maximisation [87, 90–92, 95]. However, it is the latter that has proven to be most successful, and is thus at the core of most widely-used applications. In pseudolikelihood maximisation DCA approaches, the full likelihood for each sequence position i in $\boldsymbol{\sigma}$ across all sequences in the alignment is approximated by a product of conditional likelihoods (Eq. 1.10) [93].

$$\mathcal{L}\left(\mathbf{h}, \mathbf{J}\right) = \prod_{\sigma \in \Sigma} P\left(\sigma | \mathbf{h}, \mathbf{J}\right) \approx \prod_{i=1}^{N} P\left(\sigma_i | \sigma \backslash \sigma_i, \mathbf{h}, \mathbf{J}\right) \tag{1.10}$$

Equation 1.10 describes the conditional probability of observing amino acid (σ_i) in position i given all other amino acids ($\sigma \backslash \sigma_i$) in $\boldsymbol{\sigma}$. This leads to the cancellation of the partition function Z, and instead normalises locally over all possible 20 amino acid configurations at each site i. The parameters $\mathbf{h}$ and $\mathbf{J}$, which minimise Eq. 1.10, are identified using iterative optimisation algorithms [93]. Typically, regularisation terms are also added to Eq. 1.10 to avoid overfitting of the input data [93].

The positional constraint matrices J_{ij} for all amino acid (k) pairs across all combinations of σ_i and σ_j in $\boldsymbol{\sigma}$ need be summarised to a coupling score between σ_i and σ_j. The Frobenius norm is the preferred summary statistic (Eq. 1.12), and applied to a row- and column-means-centred coupling matrix J'_{ij} (Eq. 1.11). Furthermore, Average Product Correction (APC) is applied to remove background couplings that arise due to noise from phylogenetic relationships between sequences to provide the final Evolutionary Coupling (EC) score (Eq. 1.13) [89–92, 96].

$$J'_{ij} = J_{ij}(k,l) - J_{ij}(\cdot,l) - J_{ij}(k,\cdot) + J_{ij}(\cdot,\cdot) \tag{1.11}$$

$$FN(i,j) = \sqrt{\sum_{k} \sum_{l} J'_{ij}(k,l)^2} \tag{1.12}$$

$$EC(i,j) = FN(i,j) - \frac{FN(i,\cdot)FN(\cdot,j)}{FN(\cdot,\cdot)} \tag{1.13}$$

Despite the great precision achievable by DCA algorithms, such algorithms suffer from one major drawback. All covariance-based algorithms rely on the availability of a sufficiently large and diverse Multiple Sequence Alignment (MSA) for the protein family of interest. Although the minimum number of sequences required per MSA might be target- and algorithm-dependent, early works suggested a minimum requirement of > 1000 sequence homologs [89, 97, 98]. Simultaneously, Marks et al. [69] and Kamisetty et al. [91] recommended a more sequence-specific length-dependent factor, whereby the sequence count in the alignment should exceed at least five times the protein length for precise predictions. Whilst those earlier suggestions permit crude estimations of the likelihood of obtaining precise contact predictions, researchers realised that highly redundant MSAs could surpass such a threshold yet not provide enough diversity typically required for covariance-signal detection. Thus, the measure of alignment depth (also termed Number of Effective Sequences (M_{eff})) was introduced to capture both the sequence count and diversity in a given alignment [88, 99–101]. Although target- and algorithm-dependent thresholds persist, a minimum of 100-200 effective sequences are typically required [100, 101]. Furthermore, individual weights used to calculate the alignment depth are widely used in covariance-based algorithms to reweight individual sequences [90]. The benefit is twofold: an important assumption of Eq. 1.8 that all samples are independent is satisfied and the phylogenetic effect of non-independently evolved sequences is simultaneously reduced [90]. Similar results may be achieved by removing redundant sequences prior to DCA. However, this may result in the loss of information by the requirement of selectively choosing a single representative sequence.

1.3.2 Supervised Machine Learning

Unlike DCA approaches, SML algorithms do not rely on the availability of homologous sequences to predict residue-residue contacts. Instead, SML models are trained on a variety of sequence-dependent and sequence-independent features to infer contacting residue pairs [102–107]. Broadly speaking, such SML algorithms rely on the analysis of sequence-based features, such as secondary structure, and sequence profiles. SML algorithms suffer from an inability to distinguish between residue pairs that form direct and indirect contact pairs, similar to earlier implementations of covariance-based methods. However, pure SML-based algorithms are not relevant to the work described in this thesis, and thus not further discussed. It is worth noting though that covariance-based algorithms outperform pure SML algorithms for protein families with many homologous sequences, whilst SML algorithms outperform DCA algorithms for families with fewer homologous sequences [100, 107, 108].

1.3.3 Contact metapredictors

The most recent approaches in residue-residue contact prediction use combinatorial approaches to exploit information from DCA and SML approaches. Metapredictors commonly use SML approaches as priors [71] or posteriors [75, 100, 101, 109–111] in addition to DCA algorithms. Furthermore, metapredictors use multiple input MSAs and/or DCA algorithms to further enhance the prediction precision. In most cases, metapredictors outperform their individual approaches and improvements are most noticeable for targets with lower alignment depths [75, 112, 113].

1.4 AMPLE

The major challenge in unconventional MR is to address cases where a search model cannot easily be derived from the PDB, because structures homologous to the target have not been determined or cannot be identified. The ensemble search model preparation pipeline AMPLE (***A**b initio* **M**odelling of **P**roteins for mo**LE**cular replacement) — based on the work of Rigden et al. [34] — attempts to tackle this challenge by utilising structural information from a variety of sources, such as *ab initio* structure predictions [114–118], Nuclear Magnetic Resonance (NMR) ensembles [119], and single [120] or multiple distant homologs [121, 122].

AMPLE's algorithm attempts to identify a structurally shared core amongst the initial starting structures. The idea is simple, if a shared core is present amongst a set of many structures, the likelihood of its presence in the unknown target is high. However, the rationale for identifying the shared core changes given the origin of the starting structures. In the case of clustered *ab initio* decoys, local regions inaccurately predicted can be determined by the structural divergence within each cluster. The removal of these regions reduces the error in the set of structure predictions, and if the prediction was accurate it should elucidate a conserved structural core [114]. Similarly, in NMR ensembles locally divergent regions are the result of greater flexibility in solution, and often these regions differ most from the corresponding crystal structure. Thus, removal of such flexible regions increases the likelihood of determining a structurally similar, conserved subfold suitable as MR search model [119]. If only a single distant homolog is available, a structural ensemble can be generated reflecting the innate flexibility of the starting structure. Since rigidity and evolutionary conservation are correlated [123, 124], this flexibility can be used as a proxy similar to NMR ensembles to drive trimming for identification of a more rigid, shared core [120]. Multiple distant homologs differ to the previous three examples because the shared core is most likely a small subfold present in all homologous structures. Successful identification of this subfold or super-secondary-structure motif, which often contains the functional unit of the protein family and is also likely to be present in the target, could be sufficient for structure

determination [121, 122].

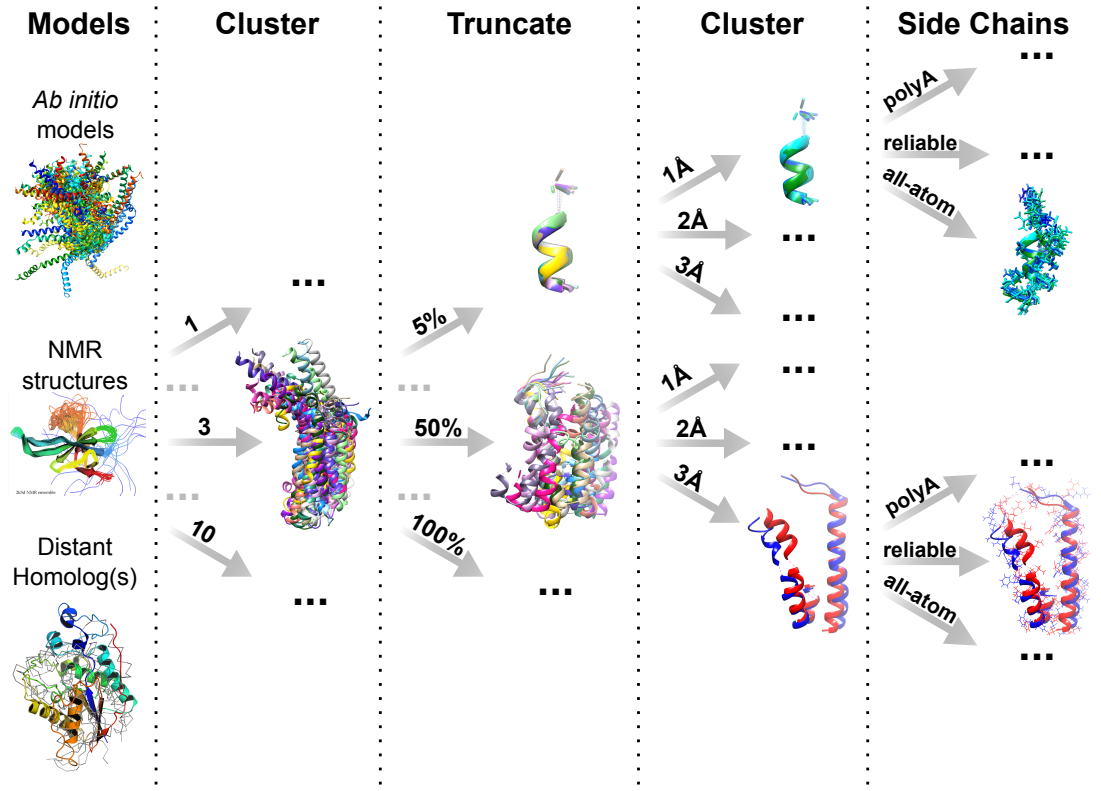


Figure 1.4: Schematic representation of the cluster-and-truncate approach employed by AMPLE.

In each case, AMPLE attempts to identify such a shared core by employing a cluster-and-truncate approach (Fig. 1.4) [114]. The latter can be separated into three main parts: (1) clustering of starting models to identify subsets of similar folds (only applicable for *ab initio* structure predictions), (2) incremental truncation of each cluster or collection of structures by its structural variance or other per-residue metric, and (3) subclustering of each truncated set of starting structures to create subgroups with varying levels of structural diversity. The incremental truncation of each cluster is typically done at 20 different levels (i.e. 5% intervals) based on the inter-residue variance score [125]. Sub-clustering is performed under three different Root-Mean-Square Deviation (RMSD) thresholds (1, 2 and 3Å). AMPLE requires each ensemble search model to contain at least two starting structures, and if this requirement is satisfied each ensemble search model is stripped to poly-alanine side-chains (all-atom and reliably-modelled side-chain [126] treatments are also available and were used by default in previous versions). This leads to the unbiased generation of a large number of ensemble search models, which cover a great diversity of its original structural information, and hopefully capture in one or more of those generated search models the shared core necessary for successful structure solution. Furthermore, AMPLE's unbiased ensemble search model generation protocol often identifies local features amongst sets of less accurate *ab initio* protein structure predictions, which are sufficient for structure determination.

Beyond the generation of ensemble search models, AMPLE also integrates the automated MR pipeline MRBUMP [127]. In AMPLE, MRBUMP's structure determination features are of particular interest. It employs PHASER [128] and MOLREP [129] for MR, refines the MR solutions with REFMAC5 [28], uses SHELXE for density modification and main-chain tracing [130], and attempts automated model building with ARP/wARP [131] and BUCCANEER [32]. These features enable the sampling of each AMPLE-generated ensemble for its suitability as MR search model.

1.5 Aims

In Sections 1.1 to 1.3, the fundamental theories behind three major areas of research were outlined: Molecular Replacement and the need for unconventional approaches, *ab initio* protein structure prediction, and residue-residue contact prediction. AMPLE, a well-established pipeline in MX, combines the former two to simplify structure solution of challenging or novel protein folds. However, the success of AMPLE's main idea, which generates ensemble search models from *ab initio* structure predictions and is the focus of the work presented in this thesis, is heavily dependent on the quality of the initial *ab initio* decoys, which are limited inherently by the computational complexity of finding the lowest free-energy state during sampling. Predicted residue-residue contact information, as described in Section 1.3, reduces the conformational search space in *ab initio* structure prediction.

Therefore, the primary aim of the work presented in this thesis focused on exploring benefits and applications of residue-residue contact prediction to improving the approach AMPLE takes in unconventional MR. Furthermore, work centred on the identification of other areas of application of residue-residue contact prediction to aid the structure solution process in unconventional MR. To address these aims, the following steps were taken:

1. In Chapter 3, an initial proof-of-principle study was conducted to highlight the benefits of residue-residue contact prediction to AMPLE's *ab initio* structure determination routine.

2. In Chapter 4, the proof-of-principle study was expanded to explore the newly defined boundaries of AMPLE by exploring a diversity of metapredictors and ROSETTA energy protocols for introducing distance restraints into the *ab initio* folding protocol.

3. In Chapter 5, work was carried out to identify potential alternatives to AMPLE's recommended structure prediction protocol ROSETTA. Three alternative protocols — SAINT2 [46], FRAGFOLD [45] and CONFOLD2 [132] — were explored for potential benefits over ROSETTA.

4. In Chapter 6, a study was carried out to explore the estimation of decoy quality by its satisfaction of predicted long-range contacts. Subsequent exclusion of the least accurate decoys was trialled as novel processing prior in AMPLE's cluster-and-truncate approach.

5. In Chapter 7, a pilot study was carried out to explore the potential of residue-residue contact prediction in identifying structural fragments or subfolds, specifically with the intend to use these identified structures as search models in MR.

Chapter 2

Materials & Methods

2.1 Selection of datasets

2.1.1 ORIGINAL dataset

A test set of 21 globular protein targets was manually selected to include a range of chain lengths, fold architectures, X-ray diffraction data resolutions and MSA depths for contact prediction (Table A.1). The targets were chosen to include the three main fold classes: all-α, all-β, and mixed α-β (α/β and α+β). The target chain lengths cover a range from 62 to 221 residues. Each crystal structure contains a single molecule per asymmetric unit and the resolution of the experimental data is in range of 1.0 to 2.3Å.

2.1.2 PREDICTORS dataset

An unbiased selection of 27 non-redundant protein targets was selected using the following protocol. For target-specific details, please refer to Table A.2.

The PFAM v29.0 [133] database was filtered for all protein families with at least one representative structure in the RCSB PDB [12] database. Each representative had to have monomeric protein stoichiometry and its fold classified in the SCOPe v2.05 database [134]. Targets with fold assignments other than "a" (all-α), "b" (all-β), "c" (mixed α+β) or "d" (mixed α/β) were excluded to focus on regular globular protein folds. Each resulting protein target was screened against the RESTful API of the RCSB PDB (`www.rcsb.org`) web server to identify targets meeting the following criteria: experimental technique is X-ray crystallography; chain length is ≥ 100 residues and ≤ 250 residues; resolution is between 1.3 and 2.3Å; structure factor amplitudes are deposited in the PDB [12] database; and there is only a single molecule in the asymmetric unit. The resulting protein structures were cross-validated against the Protein Data Bank of Transmembrane Proteins (PDBTM) [135] to exclude any possible matches. Subsequently, one representative entry was randomly selected for each PFAM family.

All PFAM family representatives obtained in the previous step were grouped by domain fold, target chain length and PFAM alignment depth (see Eq. 2.2). Each target was sorted in one of three fold bins depending on their SCOPe fold assignment: all-α, all-β, and mixed α-β (α+β and α/β). Each target was further binned by chain length (derived from the deposited sequence in the RCSB PDB) into three different bins: $[100, 150)$, $[150, 200)$, and $[200, 250]$. Lastly, each fold bin was also subgrouped by alignment depth, which was calculated for each sequence alignment of each PFAM family. Three bins were established: $[0, 100)$, $[100, 200)$, and $[200, \infty]$. Thus, all targets were grouped by fold class and further subgrouped by chain length or alignment depth.

In the final step, random targets were selected from each bin to obtain nine targets

per fold sampling a spectrum of target chain lengths and alignment depths. To ensure similar samples across the fold classes, random PFAM entries were continuously picked from the fold bins until the mean target chain length and alignment depth of nine representatives of each fold class were within ±15 of each other. This resulted in 27 targets in the final set.

2.1.3 TRANSMEMBRANE dataset

The selection of this dataset was done by Thomas et al. [118]. In summary, 13 non-redundant transmembrane protein targets were selected from the PDBTM [135], with a chain length of < 250 residues and resolution of < 2.5Å. The final selection is summarised in Table A.3. The target with PDB ID 3u2f was removed from the original dataset described by Thomas et al. [118] due to a high similarity with PDB ID 2wie.

2.2 Enhancement of β-sheet restraints

Structure prediction of β-strand containing protein targets *ab initio* is a notoriously challenging task. β-strands, potentially far in sequence space, form a β-sheet in 3-dimensions. Since fragment-assembly algorithms work on the basis of randomly inserting one fragment at the time, the probability of β-sheet formation is much lower compared to α-helices.

Recent advances in *ab initio* structure prediction have seen great improvements in structure prediction quality through the use of predicted residue-residue contacts as distance restraints (see Section 1.3). However, little research has specifically focused on improvements to the structure prediction of β-sheet formation [136]. To enhance the probability of β-sheet formation in *ab initio* structure prediction, part of this thesis focused on a more general model to enrich restraints between β-strands to attempt better super-secondary quality in the final decoys.

A more general approach, compared to Hayat et al. [136], which focused exclusively on β-barrel proteins, was developed combining a starting set of contact pairs with a specifically-prepared set obtained from BBCONTACTS [98]. A HHBLITS [137] MSA was constructed using two sequence-search iterations with an E-value cutoff of 10^{-3} against the `uniprot20` database [138]. Redundant sequences were removed from the MSA to 90% sequence identity using HHFILTER [137]. Subsequently, the MSA was subjected to CCMPRED [92] for coevolution based contact prediction, which was chosen to reproduce the approach published by Andreani and Söding [98]. Alternative starting predictions might be provided although such approach was not tested within the scope of this work. Additionally to the contact prediction, the BBCONTACTS algorithms requires a secondary-structure prediction, which can be obtained using the

`addss.pl` script [137] distributed with the HHSUITE [139]. BBCONTACTS also requires a descriptor for the diversity of the original MSA [98], which is calculated using Eq. 2.1. Ultimately, the BBCONTACTS algorithm yields β-strand-specific contact predictions identified from the starting set of contacts.

The BBCONTACTS contact pairs were added to a base set of contact pairs usually obtained from a separate (meta-)predictor. The first step included the filtering of the predicted BBCONTACTS contact list to exclude any one- or two-pair β-strand contacts, i.e. sequences of contacts with less than three consecutive contact pairs, because those typically show a high False Positive (FP) rate (Jessica Andreani, personal communication). The subsequent combination of the two sets of contact pairs was done by simple union of the lists; however, if a contact pair was in the intersection, a contact-pair related weight was doubled to allow subsequent modifications of the energy term in distance restraint creation. Furthermore, additional contact pairs were inferred if not present in the base set of contact pairs. The inference worked on the basis that any neighbouring contacts (i.e. $i, j \pm 1$; $i, j \pm 2$; $i \pm 1, j$; $i \pm 2, j$) to contact i, j must be present, and thus any missing such contacts were automatically added to the final set of contact pairs.

2.3 Evaluation of data

This section defines and describes data validation and verification procedures and equations used throughout one or more studies presented in this thesis. These definitions serve as a reference and define naming conventions where appropriate. All of the sequence- and contact-related analysis routines are implemented in CONKIT [140].

2.3.1 Sequence alignment data

2.3.1.1 Sequence alignment diversity

The sequence diversity in a MSA can be described by the number of sequences (M) it contains divided by the sequence length of the target (L). The diversity metric was used in BBCONTACTS as input parameter [98]. The MSA diversity η is defined in Eq. 2.1.

$$\eta = \frac{M}{L} \tag{2.1}$$

2.3.1.2 Sequence alignment depth

Co-evolution based residue-residue contact prediction is dependent on an input MSA ideally containing all homologous sequences found in the queried database. However, the MSA needs a certain level of sequence diversity amongst the homologs to accurately capture the coevolution signal. The alignment depth — often also referred to as M_{eff} — captures this diversity by computing the number of non-redundant sequences in the MSA.

$$M_{eff} = \sum_{i} \frac{1}{\sum_j S_{i,j}} \tag{2.2}$$

Various approaches exist for computing M_{eff} [88, 89, 101] yielding similar results [100]. In this thesis, the approach defined by Morcos et al. [88] was used. Morcos et al. [88] first described the approach by which sequence weights are computed by means of Hamming distances between all possible sequence combinations in the MSA (Eq. 2.2). All Hamming distances are then classed as determinant ($S_{i,j} = 1$) or not ($S_{i,j} = 0$) if their sequence-count-normalised value is more than a predefined identity threshold, which was set to 80% in this work. Subsequently, the contribution of each sequence to the overall alignment depth is defined as the reciprocal of its sum of determinant sequences ($\Sigma_j S_{i,j}$). The sum of all those contributions ultimately defines the alignment depth.

It is worth pointing out that M_{eff} and N_{eff} are both commonly used in literature to describe the alignment depth. Although the calculation might differ between cases — i.e. clustering-based or Hamming-distance-based — both refer to the same concept.

2.3.2 Contact prediction data

2.3.2.1 Contact map coverage

In the interpretation of a truncated contact map it is often of interest to identify the sequence coverage by the final set of contact pairs. In this particular context, coverage was defined by the number of residues (N_{map}) for which at least one contact pair existed proportional to the total number of residues in the target chain (L). Thus, the contact map coverage is defined by Eq. 2.3.

$$Coverage = \frac{N_{map}}{L} \tag{2.3}$$

2.3.2.2 Contact map precision

The precision of a set of contact pairs is equivalent to the proportion of True Positive (TP) contact pairs compared to the number of TP and FP ones (Eq. 2.4). A contact pair was considered a TP if the equivalent Cβ (Cα for glycine) atoms in the native structure were < 8Å apart, otherwise a FP. The precision value ranges from 0 to 1 with a value of 1 indicating all contact pairs are TPs.

$$Precision = \frac{TP}{TP - FP} \tag{2.4}$$

If contacts were unmatched between the target sequence and reference structure, they were not taken into account in the calculation of the precision score. This might affect a precision value; however, it also avoids inference of distances for residues absent in the native structure and therefore potentially incorrect results.

2.3.2.3 Range-dependent contact satisfaction

The range-dependent contact pair satisfaction score is computed identically to the precision of sets of contact pairs (Section 2.3.2.2). The main difference is that contact pairs are grouped by their sequence separation: short-range with < 12, medium-range with < 24 and long-range with ≥ 23 residues.

2.3.2.4 Contact map Jaccard index

The Jaccard index quantifies the similarity between two sets of contact pairs. It describes the proportion of contact pairs in the intersection compared to the union between the two sets (Eq. 2.5) [113].

$$J_{x,y} = \frac{|x \cap y|}{|x \cup y|} \tag{2.5}$$

The variables x and y are two sets of contact pairs. $|x \cap y|$ is the number of elements in the intersection of x and y, and the $|x \cup y|$ represents the number of elements in the union of x and y. The Jaccard index falls in the range [0,1], with a value of 1 corresponding to identical sets of contact pairs and 0 to non-identical ones.

It is worth noting that only exact matches are considered and the neighbourhood of a single contact is ignored.

2.3.2.5 Contact map singleton content

Most sets of residue-residue contact pairs contain a fraction of contact pairs that do not co-localise with others. These contact pairs — referred to as singleton contacts from here onwards — typically show a high FP rate and could be considered noise (although sometimes they encode TP contacts in an oligomeric interface) [100]. Such contacts are also often the ones to be down-weighted by neural network architectures of metapredictors, such as PCONSC2 [100] or METAPSICOV [101]. To quantify the fraction of singleton contacts, a distance-based clustering routine was defined to isolate singleton contact pairs, and thus describe the level of noise in the prediction.

To identify singleton contact pairs in a set of contacts, the neighbourhood of each pair was searched for the presence of other contacts. The search radius was defined by ± 2 residues in a 2D-representation of the contact map. If no other contact pair was identified under such constraint, the contact pair was classified as singleton.

2.3.3 Structure prediction data

2.3.3.1 Root-Mean-Square Deviation of atomic positions

The RMSD is a measure to quantify the average atomic distance between two protein structures (Eq. 2.6). The RMSD is sequence-independent, and measures the distance between Cα atoms.

$$RMSD = \sqrt{\frac{1}{n}\sum_{i,j}(x_i - x_j)^2 + (y_i - y_j)^2 + (z_i - z_j)^2} \tag{2.6}$$

2.3.3.2 Template-Modelling score

The Template-Modelling score (TM-score) is an alternative measure of the similarity between two protein structures [141]. Unlike the RMSD, the TM-score assigns a length-dependent weight to the distances between atoms, with shorter distances getting assigned stronger weights [141]. This results in the TM-score being less sensitive to local dissimilarities than the RMSD, and thus a better metric for overall fold similarity. The TM-score has widely been accepted as a standard for assessing the similarity between two structures, particularly in the field of *ab initio* structure prediction.

$$TMscore = max\left[\frac{1}{L_{target}}\sum_{i}^{L_{aligned}}\frac{1}{1+\left(\frac{d_i}{d_0}\right)^2}\right] \tag{2.7}$$

In Eq. 2.7, d_i describes the distance between the *ith* pair of residues. The distance scale d_0 to normalise the distances is defined by the equation $1.24\sqrt[3]{L_{target} - 15} - 1.8$. The TM-score value falls in the range (0, 1]. A TM-score value of < 0.2 indicates two random unrelated structures, and a value > 0.5 roughly the same fold [142]

2.3.4 Molecular Replacement data

2.3.4.1 Register-Independent Overlap score

The Residue-Independent Overlap (RIO) score [117] is a measure of structural similarity between two protein structures considering the total number of atoms within < 1.5Å. The RIO score can be separated into the in- (RIO_{in}) and out-of-register (RIO_{out}) score considering the sequence register between the search model and the target. The RIO score is primarily a measure for post-MR search models to assess the placement of search model atoms with respect to the previously solved crystal structure. To avoid the addition of single atoms place correctly purely by chance, the RIO metric requires at least three consecutive Cα atoms to be within the 1.5Å threshold.

2.3.4.2 Structure solution

MR structure solutions were assessed throughout all works presented in this thesis by the Correlation Coefficient (CC) [143] and Average Chain Length (ACL) scores computed by SHELXE. The latter performs density modification and main-chain tracing of the refined MR solution [130]. Thorn and Sheldrick [130] highlighted in their work that a CC of $\geq 25\%$ indicates a successful structure solution. Additionally, previous research with AMPLE [117] has shown that an ACL of the trace needs to be ≥ 10 residues.

In most studies in this thesis, additionally to the SHELXE metrics the post-SHELXE auto-built structures needed R values of ≤ 0.45. The R values had to be acquired by at least one of the Buccaneer [32] or ARP/wARP [131] solutions.

Lastly, the PHASER Translation Function Z-score (TFZ) and Log-Likelihood Gain (LLG) metrics were also considered when automatically judging a MR solution. Values of > 8 and > 120 were required, respectively. However, the PHASER metrics do not always indicate a structure solution — particularly for smaller fragments — and thus was not considered an essential metric to pass to be considered a successful solution.

Chapter 3

Evolutionary covariance in *ab initio* structure prediction-based Molecular Replacement

Note: *The majority of the work presented in this chapter was published in two independent pieces of work. All work relating to the globular targets was published by Simkovic et al. [116], and a great majority of work relating to the transmembrane targets by Thomas et al. [118]. As such, this chapter consists of extracts from both publications with additional information where appropriate. Text duplicated from either publication was written by Felix Šimkovic, all other elements were adapted.*

3.1 Introduction

The introduction of predicted residue-residue contact as distance restraints in *ab initio* protein structure prediction has proven to be a highly successful approach to limiting the conformation search space thereby enabling successful fold predictions of larger and more β-rich protein structures [e.g., 45, 46, 69–76]. In AMPLE, such proteins have historically proven the most difficult targets [114]. Furthermore, the initial AMPLE study by Bibby et al. [114] focused solely on globular targets, whilst Thomas [144] focused only much later on transmembrane protein targets. Predicted contact information was shown to be useful for both target classes, and thus should prove invaluable to AMPLE users.

Since the application of much more accurate *ab initio* protein structure prediction — obtained by restraining the conformational search space with predicted residue-residue contacts — has not yet been explored, this initial study examined the impact on AMPLE performance of contact predictions. The aim was to extend the target tractability with particular focus on larger and more β-rich protein structures.

3.2 Materials & Methods

3.2.1 Target selection

In this study, targets from the ORIGINAL and TRANSMEMBRANE datasets were used. This resulted in a final set of 21 globular and 17 transmembrane protein targets. For details on how the targets were selected refer to Sections 2.1.1 and 2.1.3, and for details on each target refer to Tables A.1 and A.3.

3.2.2 Contact prediction

For all globular targets, one contact map was predicted with the fully automated meta-predictor PCONSC2 v1.0 [100]. In summary, four MSAs were generated with each of JACKHMMER v3.1b2 [145] against the `uniref100` v2015-10 database and HHBLITS

v2.0.15 [137] against the `uniprot20` v2013-03 database [138] at E-value cutoffs of 10^{-40}, 10^{-10}, 10^{-4} and 1. Each MSA was analysed with PSICOV v2.13b3 [89] and PLMDCA v2 [146] to produce 16 individual contact predictions. All 16 predictions and per-target PSIPRED v3 [147] secondary structure prediction, NETSURFP v1.0 [148] solvent accessibility information and HHBLITS v2.0.15 [137] sequence profile were provided to the PCONSC2 deep learning algorithm [100] to identify protein-like contact patterns. The latter produced a final predicted contact map for each target sequence.

An additional contact map for β-structure containing targets was predicted using CCMPRED v0.3 [92] and reduced to β-sheet contact pairs using the CCMPRED-specific filtering protocol BBCONTACTS v1.0 [98]. Each MSA for CCMPRED contact prediction was obtained using HHBLITS v2.0.15 [137]. This entailed two sequence search iterations with an E-value cutoff of 10^{-3} against the `uniprot20` v2013-03 database [138] and filtering to 90% sequence identity using HHFILTER v2.0.15 [137] to reduce sequence redundancy in the MSA. Besides the contact matrix as input, BBCONTACTS requires a secondary structure prediction and an estimate of the MSA diversity. The secondary structure prediction was taken from the PCONSC2 step whilst the diversity factor was calculated using Eq. 2.1.

For each transmembrane protein target, a MSA was generated using HHBLITS v2.0.16 [137] against `uniprot20` v2016-02 database [138]. Three search iterations were selected at an E-value cutoff of $1e^{-3}$ and minimum coverage with the target sequence of 60%. Contact predictions for each transmembrane target were obtained using the metapredictor METAPSICOV v1.04 [101], which in turn used the contact prediction algorithms CCMPRED v0.3.2 [92], FREECONTACT v1.0.21 [149] and PSICOV v2.1b3 [89]. Additionally, a set of contacts was also predicted using the MEMBRAIN server v2015-03-15 [150].

3.2.3 Contact-to-restraint conversion

For all targets, the predicted contact maps were converted to ROSETTA restraints to guide *ab initio* structure prediction. The FADE energy function was used to introduce a restraint in ROSETTA's folding protocol. The implementation described by Michel et al. [70] was used, which defined a contact to be formed during folding if the participating Cβ atoms (Cα in case of glycine) were within 9Å of one another. The top-L (L corresponds to the number of residues in the target sequence) contact pairs were converted to ROSETTA restraints, and if satisfied a "squared-well" bonus of -15.00 added to the energy function.

Additionally, all β-containing targets were subjected to a further conversion step in a separate condition. The approach of adding BBCONTACTS restraints to a previous prediction is outlined in Section 2.2.

3.2.4 *Ab initio* structure prediction

Fragments for all targets were selected using the `make_fragments.pl` script shipped with ROSETTA. To ensure no closely homologous fragments were included in the fragment libraries, the `-nohoms` flag was set. This performs a PSIBLAST search to identify sequence homologs, whose corresponding PDB IDs are subsequently excluded from the fragment search. Each target's secondary structure prediction was provided to the fragment picker using the `-psipredfile` argument. The fragment libraries, contact restraints and secondary structure prediction were provided to the ROSETTA `AbinitioRelax` protocol [42] to predict 1,000 decoys per target. ROSETTA options were chosen according to the default protocol in AMPLE v1.0 [114]. ROSETTA v2015.05.57576 was used for globular targets and v2015.22.57859 for transmembrane ones for all ROSETTA-related protocols.

3.2.5 Molecular Replacement in AMPLE

All generated decoys were subjected to AMPLE v1.0 [114] for ensemble search model generation.

All transmembrane protein targets were processed using AMPLE's default parameters. MR trials were performed with software versions shipped in CCP4 v6.5.13 [151], with the exception of SHELXE v2014/14 [130] and ARP/wARP v7.5 [131].

All globular protein targets were subjected to AMPLE with two deviations from the default parameters. The `-use_scwrl` was set to subject all decoys to side-chain remodelling using SCWRL4 [126]. Furthermore, the number of clusters to trial was set increased from one to three via the `-num_clusters` parameter. All MR trials were performed with CCP4 v6.5.15 [151].

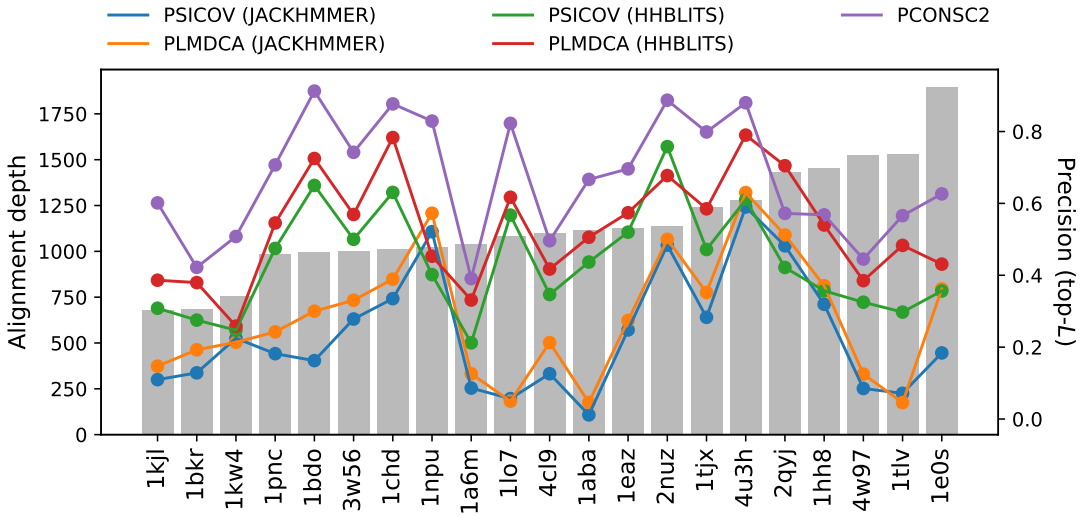
All MR solutions were assessed for success using the criteria described in Section 2.3.4.2.

3.3 Results

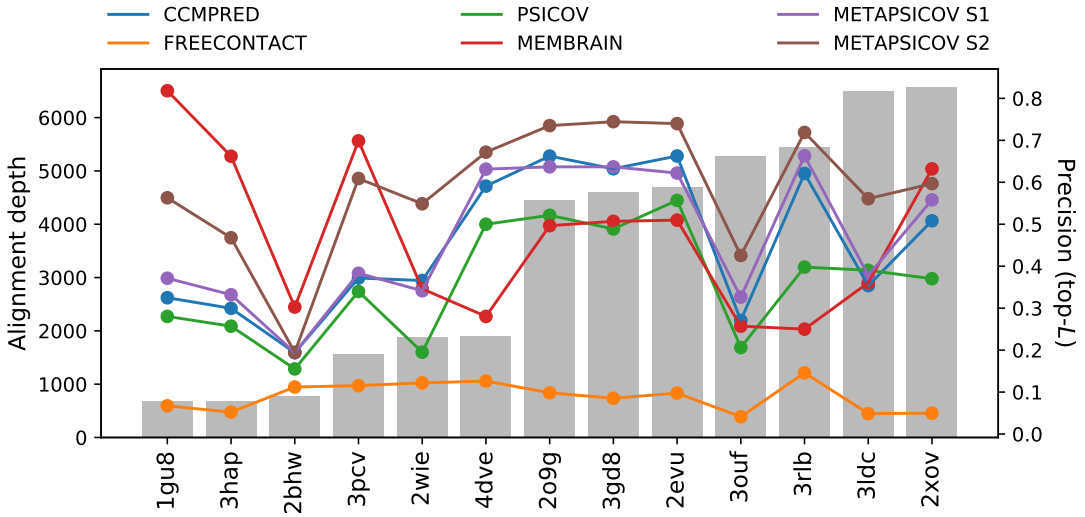
In this study, the application of residue-residue contact predictions to *ab initio* protein structure prediction and subsequently MR was investigated. This proof-of-concept work was based on two datasets covering a range of globular and transmembrane protein targets. At the time of conducting this study, state-of-the-art contact prediction algorithms were applied to obtain the best possible contact predictions to see how much AMPLE performance could be improved [114].

3.3.1 Residue-residue contact prediction

Accurate coevolution-based residue-residue contact prediction is highly dependent on the availability of many divergent homologous sequences [84]. As such, it was important to validate that the selected targets in this study satisfied such requirement.



(a)



(b)

Figure 3.1: Alignment depth and contact precision analysis of (a) globular and (b) transmembrane protein targets. Contact predictions were obtained with several contact prediction algorithms. Precision scores were calculated for the top-L contact pairs. JACKHMMER and HHBLITS alignments for PSICOV and PLMDCA contact predictions in (a) were obtained with E-value 10^{-4}.

The depth of MSAs obtained for each target sequence suggests that sufficient numbers of divergent homologous sequences were available. Across all globular targets, the minimum alignment depth was obtained for galectin-3 domain (PDB ID: 1kjl) with

679 effective sequences and the maximum for G-protein Arf6-GDP (PDB ID: 1e0s) with 1,897 effective sequences (Fig. 3.1a). The median alignment depth for all globular targets was over 1,000, which was beyond the often suggested threshold of 200 sequences [84]. The MSAs for all transmembrane protein targets also surpassed this threshold comfortably. The median alignment depth was much higher than for globular targets with 1,878 sequences (Fig. 3.1b). The minimum, which was obtained for Sensory rhodopsin II (PDB ID: 1gu8), was 692 sequences and the maximum for the sequence of Rhomboid protease GLPG (PDB ID: 2xov) was 6,583.

In coevolution-based contact prediction, the precision of predicted contacts depends on the depth of the starting MSA. Despite sufficient number of effective sequences across all targets, the data obtained as part of this study suggests that some (meta-)predictors were unable to fully utilise deeper alignments to more precisely predict contact pairs (Fig. 3.1).

PCONSC2 — a metapredictor using eight starting alignments and two contact predictors — outperformed its individual parts for almost all globular targets (Fig. 3.1a). Although only four individual components are shown in Fig. 3.1a, the pattern translated across all 16 individual predictions per target. Such results suggest that precision greatly depended, at the time of writing, on the tool used to identify and select homologous sequences for the MSA. A closer inspection of mean precision scores resulting from HHBLITS- and JACKHMMER-based alignments showed higher precision scores for top-L contact pairs based on the former alignments (Table 3.1). Nevertheless, the Machine Learning approach in PCONSC2 to combine more and less precise individual predictions resulted in superior precision in the output (Table 3.1). No correlation was observed between alignment depth and precision for either individual predictors or the metapredictor PCONSC2 (Fig. 3.1a).

Table 3.1: Summary of mean PCONSC2 raw contact prediction precision based on JACKHMMER and HHBLITS alignments and PSICOV, PLMDCA and PCONSC2 coevolution-based contact prediction.

Contact prediction		**Alignment E-value cutoff**			
		1^{0}	1^{-4}	1^{-10}	1^{-40}
PSICOV	JACKHMMER	0.240	0.239	0.213	0.167
	HHBLITS	0.439	0.435	0.354	0.209
PLMDCA	JACKHMMER	0.293	0.288	0.252	0.140
	HHBLITS	0.545	0.530	0.447	0.224
PCONSC2		0.667			

Contacts for transmembrane protein targets in this study were predicted with the metapredictor METAPSICOV and the transmembrane-specific predictor MEMBRAIN. METAPSICOV STAGE1 and STAGE2 predictions outperformed MEMBRAIN in nine and ten cases, respectively, whilst MEMBRAIN outperformed METAPSICOV for the

rest (Fig. 3.1b). The METAPSICOV algorithm utilises the raw predictions by CCMPRED, FREECONTACT and PSICOV to generate its STAGE1 and STAGE2 predictions. METAPSICOV STAGE1 predictions were near identical to CCMPRED, whereby 12 of 13 targets showed an absolute $\Delta_{precision}$ of less than 0.05 (Fig. 3.1b). This similarity did not propagate to METAPSICOV STAGE2 predictions with only a single target showing such similar precision values (Fig. 3.1b). Amongst the three raw predictors used by METAPSICOV, FREECONTACT performed by far the worst with a mean precision of 0.09 across all transmembrane targets. PSICOV showed similar trend to CCMPRED when assessed by target, which resulted in a mean absolute $\Delta_{precision}$ of 0.10.

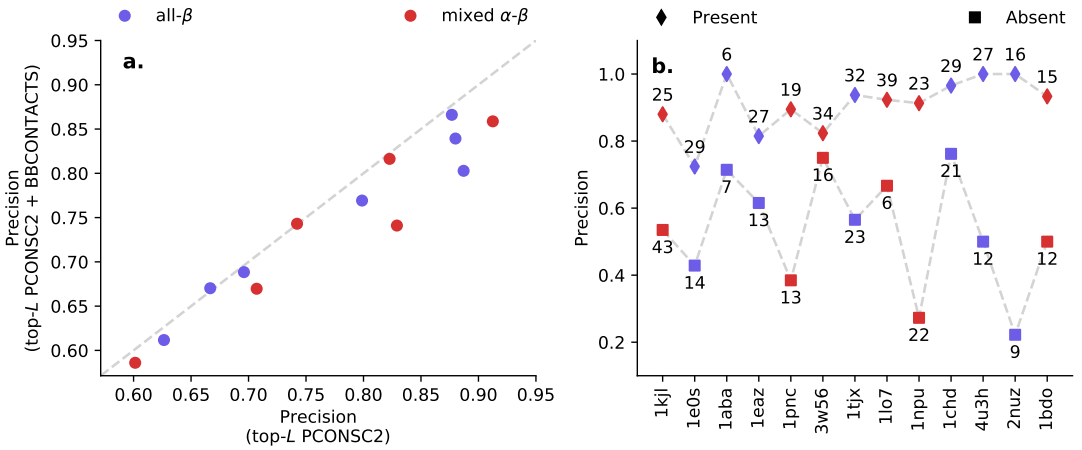


Figure 3.2: Evaluation of predicted BBCONTACTS contact pairs. (a) Precision evaluation of PCONSC2 predicted contact map with and without BBCONTACTS. (b) Precision evaluation of BBCONTACTS contact pairs split by status of presence or absence in the base PCONSC2 contact list. Numbers besides each marker indicate the number of contacts. The rank order of scatter points is identical between both subplots and based on the PCONSC2 precision values (x-axis) in (a).

The addition of BBCONTACTS contact pairs to improve structure prediction accuracy for β-structure containing targets was a novel aspect introduced in this study. The initial step of the addition of BBCONTACTS contact pairs included the filtering of predicted one- and two-pair β-strand contacts from the original BBCONTACTS list (for further details, see Section 2.2). The findings in this study confirmed this for all β-structure containing targets. Precision values improved for all targets with changes ranging from 0.01 to 0.14 whilst retaining on average 80% of all contacts. Filtered BBCONTACTS predicted contact maps were combined with other predicted contact maps, i.e. PCONSC2, to either upweight or add contact pairs. Findings in this study highlight that upweighted contact pairs were more precise than ones to be added. The minimum precision score for a set of upweighted contacts was 0.72 for 29 contact pairs and the maximum of 1.00 for up to 27 contact pairs. In comparison, contact pairs uniquely identified by BBCONTACTS ranged in precision scores from 0.22 (nine contacts) to 0.76 (21 contacts).

Despite the high precision of predicted BBCONTACTS contact pairs, the merge of such pairs with top-L PCONSC2 contact pairs resulted in an expected loss in precision for the resulting contact set (Fig. 3.2). TP contacts, which dominate the predicted BBCONTACTS contact set, were also predicted by PCONSC2, and thus upweighted (Fig. 3.2). Since upweighting does not affect the precision, the value remained unaffected after this procedure. However, contact pairs only predicted by BBCONTACTS contain more FP contacts. Once added to the base PCONSC2 contact list, these contacts therefore reduced the precision value (Fig. 3.2). Either subset of BBCONTACTS contacts did not show any correlation between the number it contained and its precision. The fold of the target did not show any clear distinction between better and worse sets of contacts either (Fig. 3.2).

3.3.2 Protein structure prediction

Predicted contact information is particularly useful to limit the conformation search space in *ab initio* protein structure prediction [40]. Since such predictions are the basis for AMPLE studies presented in this thesis, it is important to analyse the improvement in decoy quality.

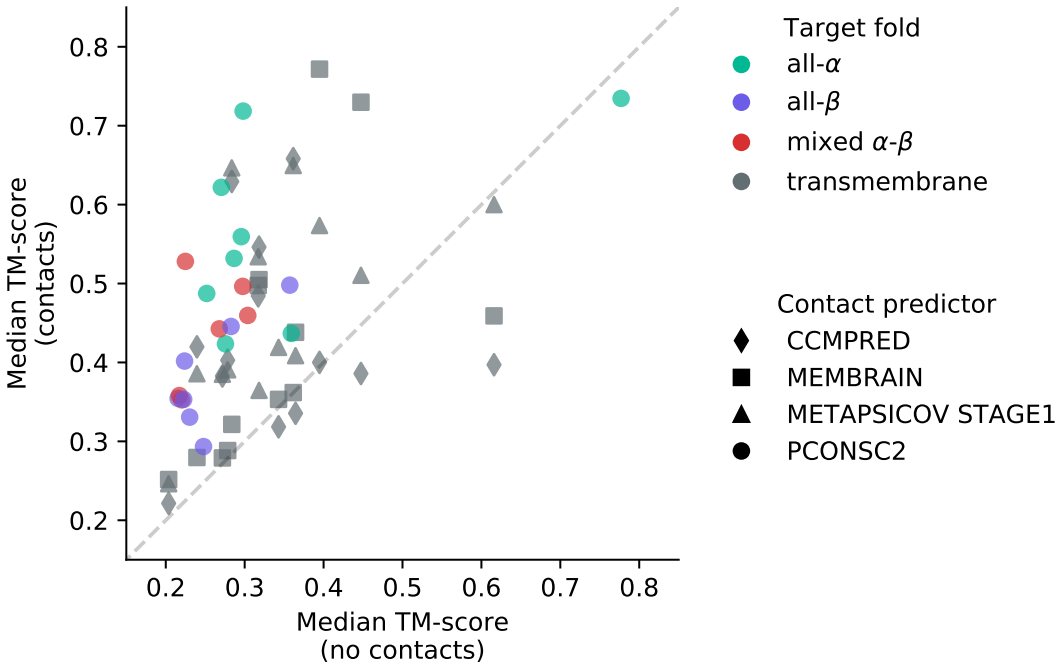


Figure 3.3: Effect of contact distance restraints on *ab initio* decoy quality by comparison of unrestrained (*no contacts*) and contact-restrained (*contacts*) median TM-scores for 1,000 decoys per target. Colours indicate the target fold and symbols the contact prediction algorithm.

Globular protein targets benefited greatly from the addition of PCONSC2 contact predictions. All but one target saw median TM-score improvements of at least

0.05 when comparing contact-assisted PCONSC2 decoys with simple ROSETTA decoys (Fig. 3.3). The greatest improvement over 1,000 decoys was achieved for Oxymyoglobin (PDB ID: 1a6m) with an improvement in median TM-score of 0.42. The decoys for ankyrin (PDB ID: 2qyj) showed a minor decrease in median TM-score of 0.04; however, the median TM-score for ROSETTA decoys was 0.78, and thus such minor decrease may be negligible.

Previously, *ab initio* protein structure prediction for globular targets was greatly limited by target fold and chain length. The addition of predicted residue-residue contacts enhanced decoy quality primarily for α-helical and mixed α-β protein targets (Fig. 3.4). Whilst only one all-α target had more than 50% native-like decoys in its ROSETTA decoy set, five targets surpassed this threshold when PCONSC2 contact data was used to restrain the folding procedure. Similarly, the median TM-score of no mixed α-β target decoy set surpassed the TM-score threshold of 0.5 with ROSETTA decoys compared to one for PCONSC2 decoys with three further ones greater than 0.4. All-β targets also benefited from the addition of predicted contact restraints, although decoy set quality did not surpass the native-like threshold in terms of their median TM-score (Fig. 3.4). Larger targets did not benefit any more than smaller targets from the addition of residue contacts to the structure prediction protocol. The only real exception to this were the decoys for the CheB methylesterase domain (PDB ID: 1chd), for which the majority of ROSETTA decoys were almost random-like whilst PCONSC2 decoys are native-like (Fig. 3.4).

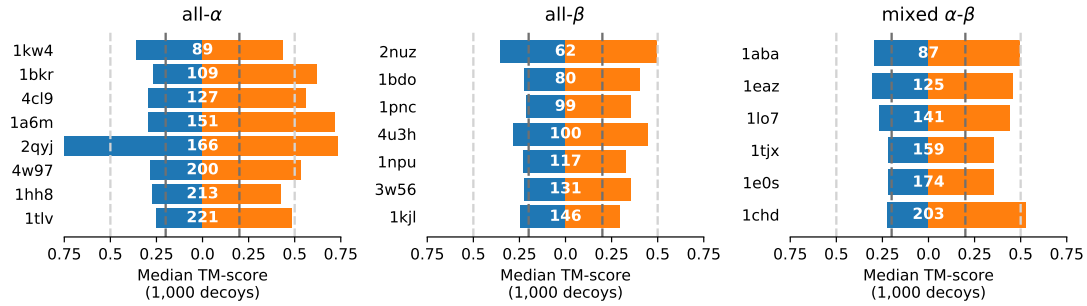
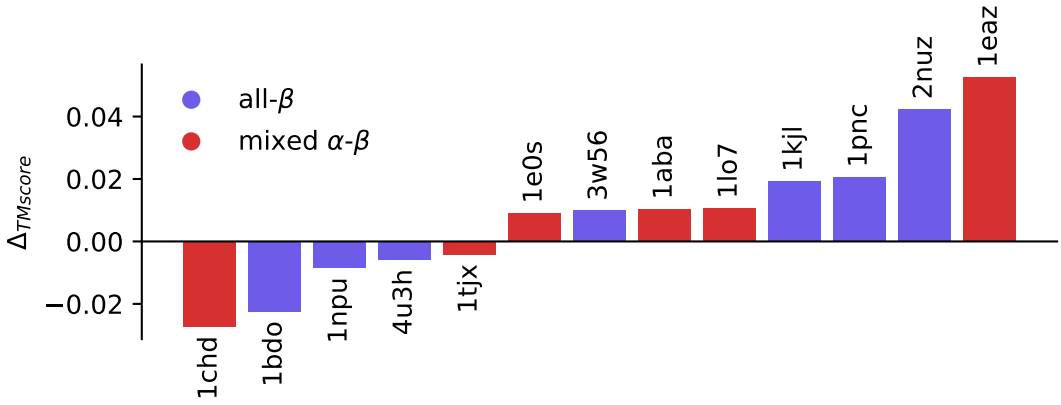


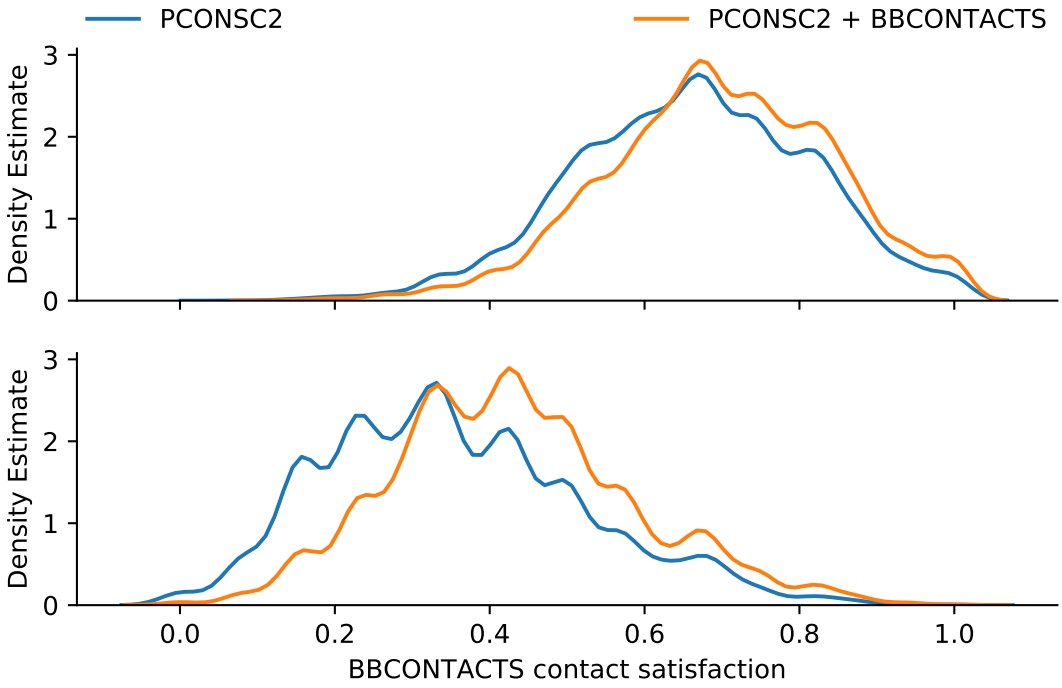
Figure 3.4: TM-score comparison for globular targets separated by fold and ordered by target chain length. Median TM-scores for 1,000 decoys generated with simple ROSETTA (orange) or contact-assisted ROSETTA (blue) runs. Numbers in each row correspond to the target chain length. Bars surpassing the dark grey line indicate that the majority of structures are better than random, whilst the light grey line indicates that the majority of structures are native-like [142].

The enhancement of β-structure specific contact pairs was an important part of this study. In Section 3.3.1, the high precision of added BBCONTACTS contact predictions was demonstrated. Thus, the next essential step was to explore how the BBCONTACTS supplement enhanced or degraded decoy quality after ROSETTA *ab initio* protein structure prediction. Given 13 β-structure containing targets, eight targets achieved better overall decoy quality with added BBCONTACTS (Fig. 3.5a). The

smallest improvement was observed for target 1e0s with 0.01 TM-score units, whilst the largest for target 1eaz with 0.05 units. The remaining five targets — PDB IDs 1chd, 1bdo, 1npu, 4u3h and 1tjx — saw decreases in median TM-score up to 0.03 when BBCONTACTS contact pairs were added as restraints (Fig. 3.5a). No clear difference between fold classes, i.e. mixed α-β or all-β targets, was observed, although mixed α-β targets did show slightly greater extremes (Fig. 3.5a).



(a)



(b)

Figure 3.5: (a) TM-score comparison for β-structure containing globular targets separated by fold and ordered by the difference in median TM-score between PCONSC2 and PCONSC2+BBCONTACTS decoys. Positive values indicate a better median TM-score in favour of PCONSC2+BBCONTACTS decoys, whilst negative values those for PCONSC2. PDB IDs are provided alongside each bar. (b) Satisfaction of BBCONTACTS contact predictions in decoys with added β-structure contact restraints (PCONSC2 + BBCONTACTS) and those without (PCONSC2).

An analysis of the satisfaction of BBCONTACTS contact predictions in decoys where extra β-sheet contact pairs were used as restraints compared those where they were not highlighted a greater satisfaction in the former. This indicates that the added and upweighted BBCONTACTS β-structure contact restraints enhanced the formation of β-sheets in the resulting decoys, which would explain the overall improved decoy quality for more than half the targets. A separation of contact satisfaction by upweighted and added BBCONTACTS contact pairs indicated that the upweighting had less effect compared to the addition (Fig. 3.5b). Although the former shows a marginal improvement in BBCONTACTS contact satisfaction of decoys without upweighted restraints, the difference was minimal. In comparison, PCONSC2 decoys without the added BBCONTACTS restraints showed less satisfaction for such contacts, indicating that they did not form as often compared to PCONSC2+BBCONTACTS. In combination with the upweighting, these resulted showed that β-rich regions are predicted more accurately when BBCONTACTS contact pairs supplement PCONSC2 contacts.

Transmembrane protein targets were modelled using residue-residue contact predictions derived with CCMPRED, MEMBRAIN and METAPSICOV STAGE1. A ROSETTA benchmark was also run to compare contact-assisted decoys to the current norm. Findings in this study highlight the much improved decoy quality for almost all targets when predicted contact information was used to reduce the conformational sampling space (Fig. 3.3). Across all methods, only the decoys for ATP synthase subunit C (PDB ID: 2wie) suffered from the addition of contact restraints during *ab initio* protein structure prediction. ROSETTA generated decoys with median TM-score of greater than 0.6 when no contact restraints were used. This contrasts strongly with contact-assisted decoy sets, for which only METAPSICOV STAGE1 predictions yielded overall native-like decoys, i.e. median TM-score of greater than 0.5.

A split for decoy quality comparison between no-contact and contact-assisted decoy sets by contact prediction algorithm showed that CCMPRED contact predictions were not sufficiently precise to always improve decoy quality. Four out of 13 targets are predicted more accurately without CCMPRED contact information (Fig. 3.6). In comparison, MEMBRAIN and METAPSICOV STAGE1 contact predictions resulted in enhanced decoy quality to the extend that only one decoy set was worse than their no-contact counterpart (Fig. 3.6). Most notably, either of the three contact predictions per target performed better for certain targets. The most extreme example may be the decoy sets for bacteriorhodopsin (PDB ID: 3hap) for which CCMPRED contacts resulted in decoy quality degradation of 0.06, METAPSICOV STAGE1 in a slight improvement of 0.06 and MEMBRAIN in an improvement of 0.28 TM-score units. This translated into absolute decoy counts with native-like fold — i.e., TM-score ≥ 0.5 — of the following: 274 for decoys without contact guidance, 289 for CCMPRED contact guidance, 538 for METAPSICOV STAGE1 contact guidance, and 996 for MEMBRAIN contact guidance. Similar examples exist (e.g., PDB IDs 1gu8, 3rlb or 4dve in Fig. 3.6) and highlight that no single method yielded the best decoys under all circumstances.

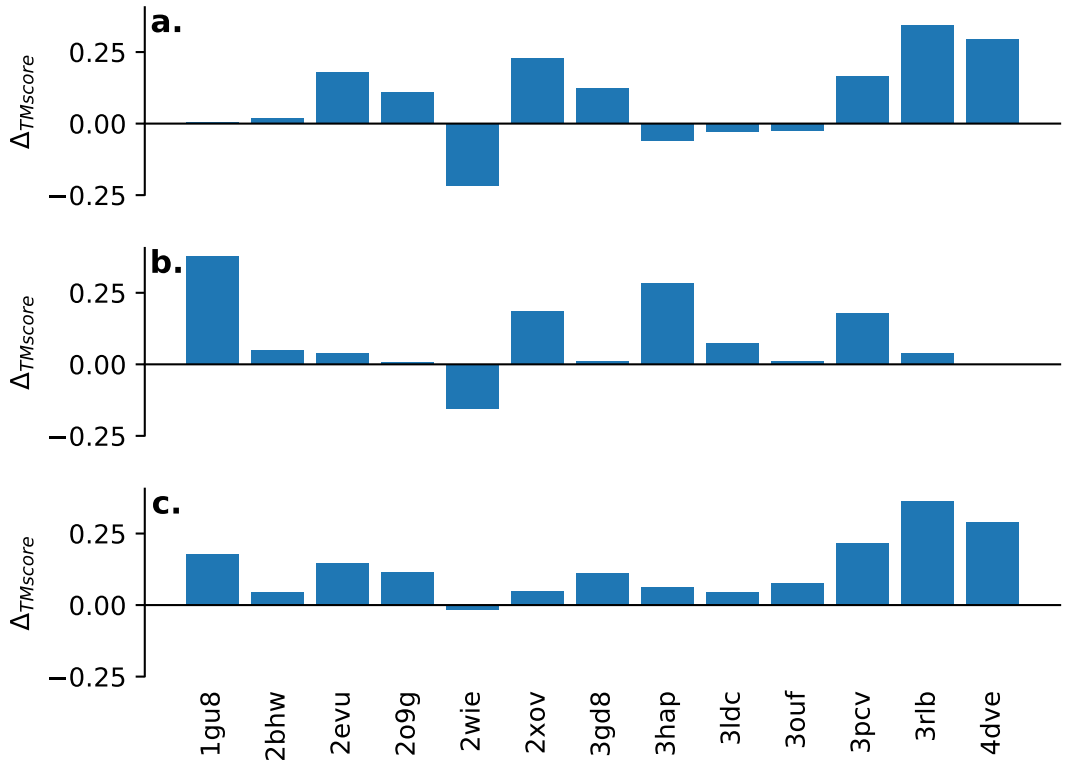


Figure 3.6: TM-score difference between contact-assisted and simple ROSETTA decoys for transmembrane protein targets. Positive $\Delta_{TMscore}$ values indicate more accurate contact-assisted decoys, whilst negative values better decoys without the addition of contacts. $\Delta_{TMscore}$ values were computed by median TM-score. Contact restraints were obtained with (a.) CCMPRED, (b.) MEMBRAIN, and (c.) METAPSICOV STAGE1.

3.3.3 Molecular Replacement

The most important aspect of this study was the impact of contact-assisted decoys in AMPLE-MR. Contact-unassisted AMPLE is primarily limited by a target's chain length and fold, which typically cannot exceed 150 residues, and performs poorly for β-rich folds [114]. Findings presented in Section 3.3.2 outlined improvements in overall decoy quality when predicted contact information was used as distance restraints in *ab initio* protein structure prediction. However, it is yet to be seen how the improved decoy quality translates into MR structure solutions.

3.3.3.1 Globular protein targets

Structure solutions were attempted for a total of 21 globular targets. Simple ROSETTA decoys — those without contact restraints and AMPLE's current default — resulted in nine structure solutions (Fig. 3.7). The addition of PCONSC2 contact-restraints to the structure prediction procedure improved decoy quality to achieve four additional

structure solutions. However, the structure of the N-terminal region of P67Phox (PDB ID: 1hh8) was not solved when PCONSC2-restrained decoys were used compared to simple ROSETTA ones. The addition of BBCONTACTS distance restraints to up-weight and supplement PCONSC2 contacts enabled a further unique solution for the Phosphoinositol (3,4)-bisphosphate PH domain (PDB ID: 1eaz) (Fig. 3.7).

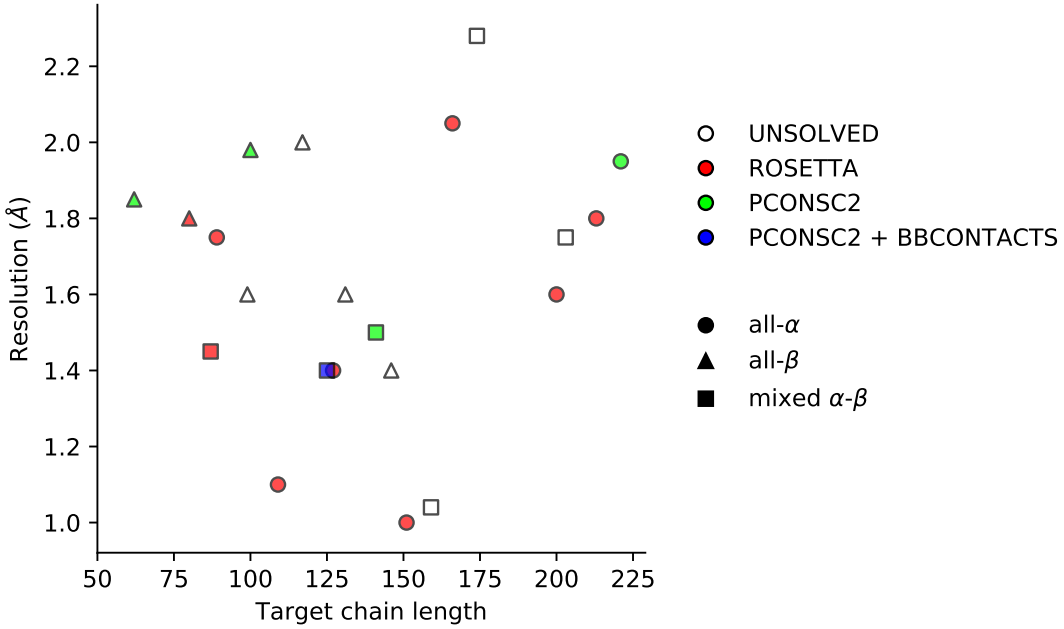
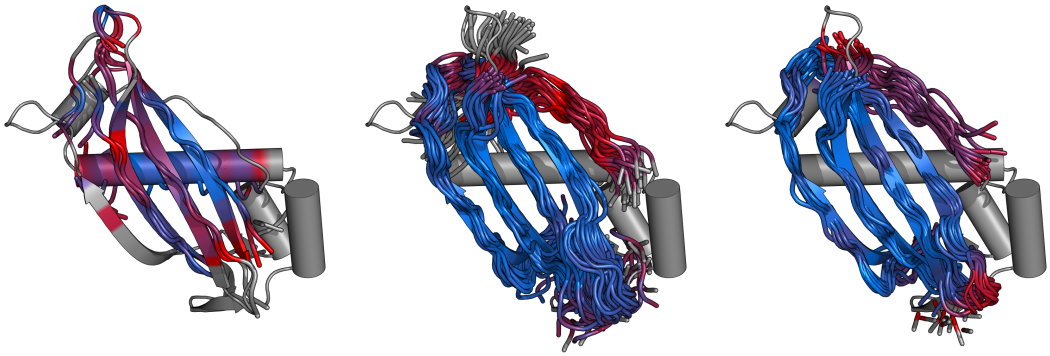


Figure 3.7: Summary of structure solutions obtained with AMPLE using no-contact ROSETTA, PCONSC2-contact-restraint-assisted ROSETTA and PCONSC2 + BBCONTACTS-assisted ROSETTA decoy sets. Empty markers indicate unsolved targets. Filled markers highlight the minimum decoy set by complexity of the prediction procedure to obtain MR structure solution. Marker shape distinguishes the fold class.

The majority of structure solutions were obtained for all-α targets in the dataset, with a total of eight structure solutions (Fig. 3.7). Seven of those eight structure solutions were achieved with unrestrained ROSETTA decoys, with target chain lengths up to 213 residues. The largest target in the globular dataset, and the only all-α target that required residue contacts, totals 221 residues in target chain length, which exceeded AMPLE's previously benchmarked limits for globular targets greatly [114]. In comparison to all-α targets, β-structure containing proteins required predicted contact restraints to result in sufficiently accurate decoys for MR. Across all-β and mixed α-β targets, only two structure solutions were obtained with unrestrained ROSETTA decoys. This contrasts to an additional three targets when PCONSC2 restraints were used during *ab initio* structure prediction. Furthermore, the addition of BBCONTACTS contact restraints enabled an additional structure solution, yielding much greater success for β-structure containing protein targets compared to the previous default. Structure solutions for β-containing targets were obtained for target chain lengths up to 141 residues (Fig. 3.7).

Two exceptional cases specifically exemplified the application of contact predictions and their benefit to MR. The first example were the structure solutions for 4-hydroxybenzoyl CoA thioesterase (PDB ID: 1lo7), which were based on AMPLE ensemble search models derived from the PCONSC2 and PCONSC2+BBCONTACTS decoy sets. Without contact restraints, AMPLE search models did not accurately represent the target fold (Fig. 3.8a). In comparison, precise residue-residue contact predictions primarily restraining the large β-sheet yielded decoys of sufficient quality to achieve MR structure solution with both PCONSC2 (Fig. 3.8b) and PCONSC2+BBCONTACTS (Fig. 3.8c) decoy sets.

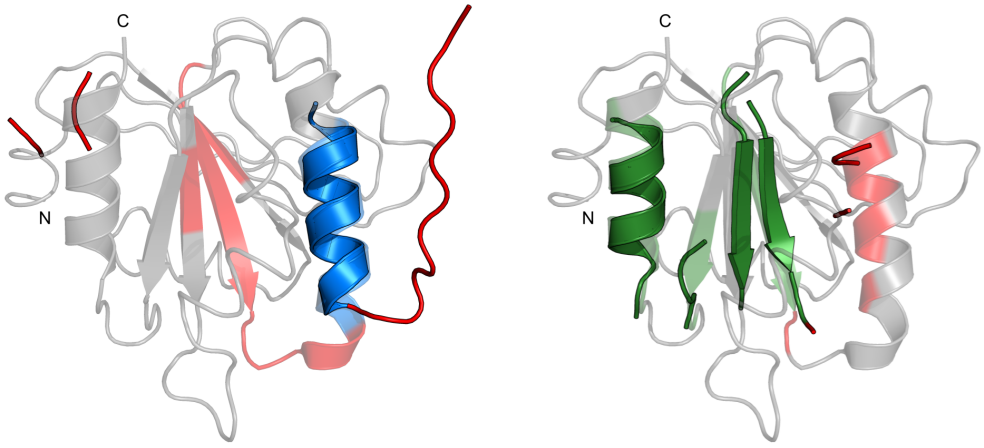


(a) (b) (c)

Figure 3.8: Structural superposition of the (a) ROSETTA (Cα RMSD 2.814 Å; ensemble contains two structures), (b) PCONSC2 (Cα RMSD 1.748 Å; 30 members) and (c) PCONSC2+BBCONTACTS (Cα RMSD 1.760 Å; 15 members) search-model ensembles for 4-hydroxybenzoyl CoA thioesterase (PDB ID: 1lo7). Examples are the highest scoring search models based on SHELXE CC score, with only (b) and (c) leading to successful MR structure solutions. Search models are shown as tubes and crystal structures as cartoons. (a) and (c) are 50% of the target sequence, while (b) is 55%. The colour scale illustrates the pairwise Cα RMSD between each search-model ensemble (represented by its first member) and the crystal structure, with blue representing the minimum Cα RMSD and red the maximum. Unaligned residues are coloured grey.

The second exceptional case, PDB ID 1e0s, did not yield any MR structure solution with either decoy set according to the stringent criteria for MR success applied in this study (see Section 2.3.4.2). However, a RIO analysis of PHASER solutions, i.e. after MR, indicated that some PCONSC2 and PCONSC2+BBCONTACTS AMPLE search models were placed partially correctly (Fig. 3.9). For the top PCONSC2 search model, 40% (12 residues) of the search model residues were correctly superimposed, albeit out of register on the target structure (PHASER TFZ=4.7, PHASER LLG=16) (Fig. 3.9a). For the top PCONSC2+BBCONTACTS search model, 77% (30 residues) of the search model were superimposed in an in-register fashion (PHASER TFZ=5.3, PHASER LLG=17) (Fig. 3.9b). For the latter, expert manual intervention might allow structure determination, but in this case the correct solution was not prominent in the list of MR placements. Nevertheless, it is clear that even when overall structure solution was not automatically achieved the PCONSC2+BBCONTACTS search model

provided better results which might be recoverable as successes in the future as MR and post-MR software improves still further.



(a) (b)

Figure 3.9: Top-PHASER solutions for PDB ID 1e0s based on RIO scores for (a) PCONSC2 (RIO score 12) and (b) PCONSC2+BBCONTACTS (RIO score 30) search models. Search-model colour coding indicates useful superposition of residues by in- (green) or out-of-sequence register (blue) residues as well as misplaced (red) residues. The addition of BBCONTACTS restraints produced a more accurate model with correctly placed β-strands that was placed correctly. Both structures are shown in cartoon representation with the crystal structure shown as a transparent cartoon. Unaligned reference crystal structure residues are coloured grey.

With much improved decoy quality deriving from the use of predicted contact restraints to guide *ab initio* structure prediction, the question arose whether AMPLE's existing cluster-and-truncate approach remained the most suitable for obtaining a conserved, native-like core from the decoys found in the largest clusters. For globular targets solved using simple ROSETTA decoys, certain features throughout AMPLE's cluster-and-truncate approach typically correlated with eventual success in structure solution [114]. In general, the greater the number of decoys in the largest cluster the more likely the success was with derived search models. Truncation removed structurally variant parts leading to smaller more accurate ensemble subsets of the cluster decoys. Although successful search models were found at every truncation interval, the majority were derived with search models containing around 30 residues. Lastly, each of the potential nine search models derived at each truncation level (three subclustering radii with three side chain treatments each) can lead to non-redundant structure solutions. Similar observations, particularly with respect to the most successful search model size range were made for other target classes [117, 118] and for *ab initio* decoys made with QUARK [115].

A size comparison of the largest clusters of ROSETTA and PCONSC2+BBCONTACTS

(or PCONSC2 for all-α) decoys indicated a median increase of 122 decoys per cluster in the latter. All cluster sizes increased except for target 2qyj. More accurate *ab initio* decoys are directly linked to larger cluster sizes because of the associated increase in convergence [152]. Here, as expected, the largest cluster contains better than average quality decoys but the size of the largest cluster does not link to the total number of successful search models.

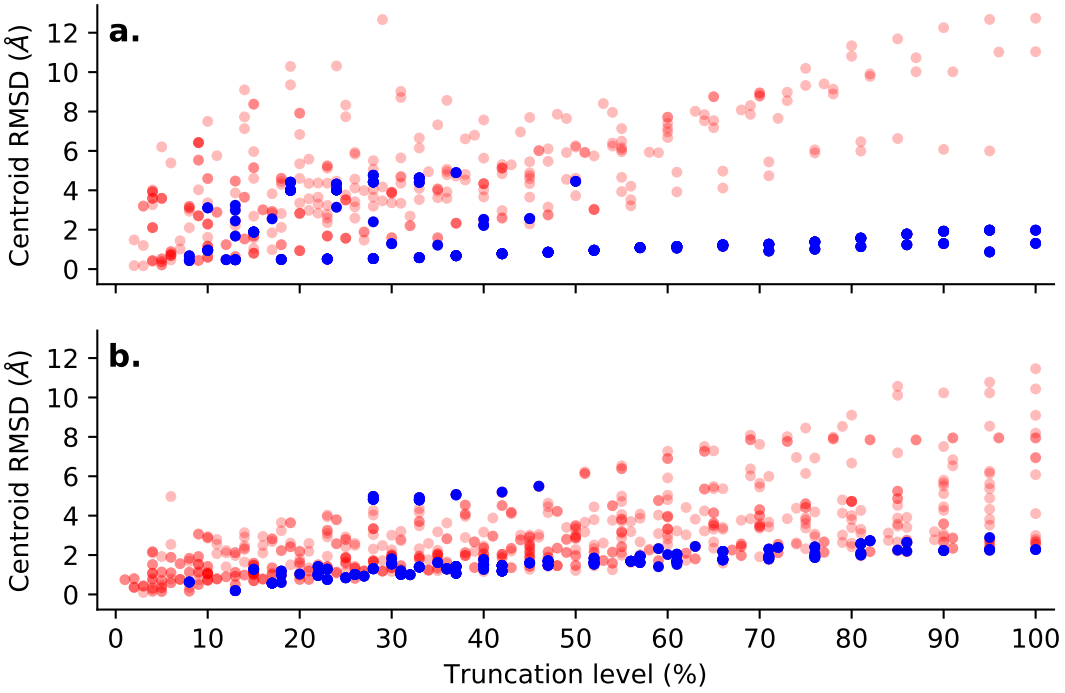


Figure 3.10: The percentage of sequence in the search model is mapped against the RMSD over all Cα atoms of the first representative of each search-model ensemble derived from the largest cluster against the native structure. Successful structure solutions of individual search models are highlighted in blue and unsuccessful solutions in red. Progressively darker shades of either colour correspond to increasing numbers of overlapping points. Progressive truncation is shown for (a) ROSETTA and (b) PCONSC2+BBCONTACTS decoys (or PCONSC2 decoys for all-α targets).

In comparison to the clustering step, the progressive truncation of decoys in the largest cluster at 20 different intervals directly affects the number of successful search models. An analysis of the progressive truncation and the effects on search model accuracy revealed that all successful search model ensembles had a Cα-RMSD better than 5.5Å compared to the native structure (Fig. 3.10). Although the latter cutoff is independent of whether predicted contact information was provided during *ab initio* modelling, a clear difference between the ROSETTA and PCONSC2+BBCONTACTS (or PCONSC2 for all-α) ensemble search models for all targets can be observed. In total, ROSETTA decoys for all targets produced 1,314 ensemble search models based on the largest clusters. In comparison, PCONSC2+BBCONTACTS decoys generated for the same targets 2,469 search model ensembles from the largest clusters. This

increase is the result of a more successful subclustering process due to the increased structural homogeneity across the decoys in the largest cluster. The most notable difference between the two sets is detected for the Small G-protein ARF6-GDP (PDB ID: 1e0s), which produced three ensemble search models based on ROSETTA decoys and 90 based on PCONSC2+BBCONTACTS decoys. Additionally, ensemble search models with structural fragments of 15-40 residues of the target sequence are more likely to succeed in MR phasing than larger or smaller search models [114]. Here we find that the same range is most successful for contact-assisted decoys (Fig. 3.11). Out of 246 successful search models for PCONSC2+BBCONTACTS decoys derived from the largest cluster (PCONSC2 for all-α), 101 successful search models contained 15-40 residues. Significantly, some cases like the PH domain of TAPP1 (PDB ID: 1eaz) and the N-terminal bromodomain of human BRD4 (PDB ID: 4cl9) only solved with truncated search models in this size range. Nevertheless, structure solutions were also achieved with larger or smaller search models. The smallest search model leading to a structure solution contained nine residues (8% of total sequence) and solved the Calponin Homology domain from human β-spectrin (PDB ID: 1bkr). In comparison, the largest successful search model in terms of residues was found for the designed full consensus ankyrin (PDB ID: 2qyj) domain with 158 residues (95% of total), and in terms of percentage of the total sequence the untruncated, 62 residue search model for α-spectrin SH3 domain (PDB ID: 2nuz) was successful. Therefore, although truncating the *ab initio* decoys at different levels remains essential for contact-assisted decoys, biasing sampling into the most successful size range may be advantageous in future runs.

The truncated decoys are further processed by subclustering at three different atomic radii, with the resulting subclusters previously found to be similarly successful [114]. Similar trends are seen here: 36% of structure solutions with ROSETTA decoys were achieved with a subclustering radius of 1Å, 36% at a radius of 2Å, and 28% at a radius of 3Å. For PCONSC2+BBCONTACTS (or PCONSC2 for all-α) decoy sets similar numbers were observed (35% at radius of 1Å; 40% at 2Å; 25% at 3Å). Nevertheless, in terms of number of targets solved all three subclustering radii were essential. Largest-cluster decoys for target 1eaz produced a total of 327 search models, but only one solved and this derived from a subclustering radius of 1Å. In comparison, contact-assisted decoys from the largest cluster for target 4u3h achieved structure solutions solely with decoys subclustered at 2Å. A single search model with subclustering radius of 3Å solved the target 4cl9 with ROSETTA decoys. The final step in search model creation is the side-chain processing of each subclustered ensemble. Similarly to the subclustering, no difference was observed between ROSETTA and PCONSC2+BBCONTACTS decoys. For both the polyalanine treatment is most successful, covering 37% of successful search models for ROSETTA decoys and 44% for PCONSC2+BBCONTACTS decoys. For almost all targets, the polyalanine side-chain treatment would be enough to obtain a structure solution. However, some cases, like the target 1eaz, only solve with either or both of the remaining treatments. Thus, relying solely on polyalanine side-chain

treatment may limit the overall success rate, although trialling polyalanine ensemble search models first might lead to structure solution faster.

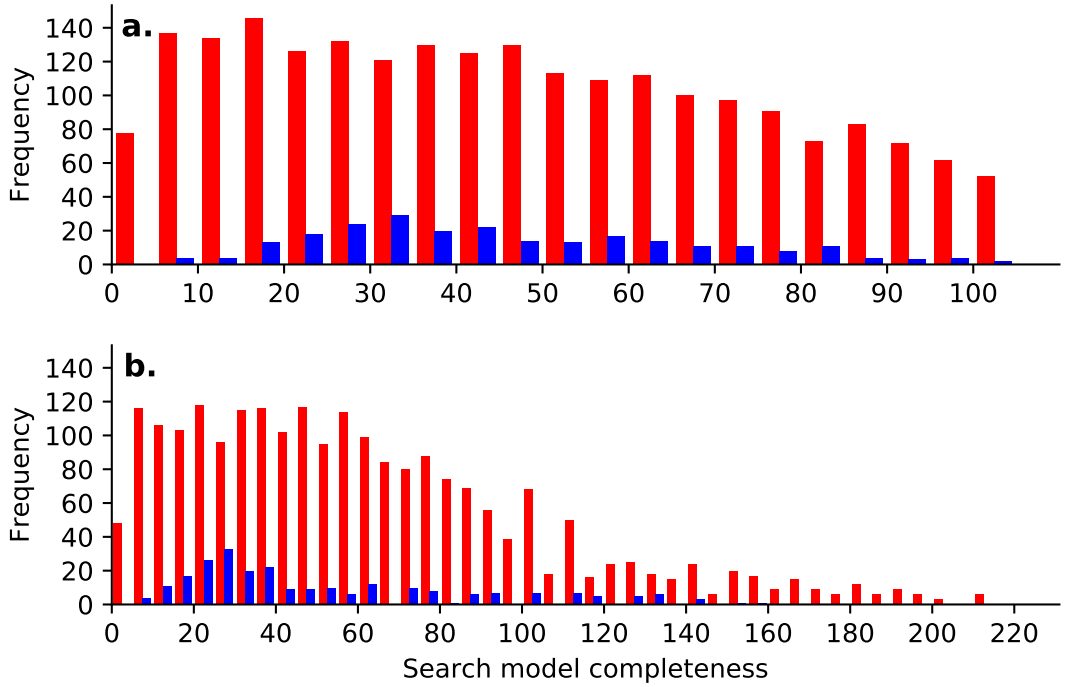


Figure 3.11: Summary of AMPLE truncation ranges for structure solution. (a) Percentage of residues and (b) number of residues per chain in search model mapped against the number of search models leading to structure solution (blue) or not (red).

3.3.3.2 Transmembrane protein targets

The MR structure solution attempts given the decoy sets for transmembrane protein targets was conducted by Dr Jens Thomas and is documented in Thomas [144] and Thomas et al. [118].

In summary, MR structure solution successes with decoys restrained by either of the three contact prediction protocols — CCMPRED, MEMBRAIN and METAPSICOV STAGE1 — were mixed. CCMPRED solved three targets, MEMBRAIN solved five and METAPSICOV STAGE1 decoys solved four. Simple ROSETTA decoys resulted in four structure solutions. CCMPRED and METAPSICOV STAGE1 both solved target 4dve, which could not be solved with any other method, and METAPSICOV STAGE1 also solved target 2o9g, which had previously only been solved with the AMPLE library of ideal helices.

3.4 Discussion

The change in statistical model for residue-residue contact prediction has enabled great improvements to its precision. Today, predicted contact information is often used to restrain the conformational search space to enable accurate *ab initio* protein structure prediction. In this study, the effect of such improved structure predictions was examined with a particular interest of their application in unconventional MR in AMPLE. The main focus of the presented work rested with the aim to extend AMPLE's target tractability, both for larger and more β-rich protein targets.

The addition of predicted residue-residue contacts unsurprisingly improved the quality of *ab initio* protein structure predictions, which is in line with numerous other studies [e.g., 45, 46, 69–76]. The improved decoy quality directly translated to further structure solutions with AMPLE. Contact-unassisted decoys, i.e. the current default, achieved nine and four solutions for globular and transmembrane protein targets, respectively. In comparison, contact-assisted decoys solved a further five globular targets, whilst contact-assisted decoys solved some different targets compared to contact-unassisted decoys for transmembrane protein targets.

The initial findings in this study highlighted the successful application of contact prediction to extend the target tractability with regards to the target chain length. Bibby et al. [114] previously benchmarked *ab initio* protein structure predictions up to chain lengths of 120 residues. However, the findings indicated that larger targets should be tractable with AMPLE, especially all-α ones [114]. The results in this study confirmed such extended target tractability, with contact-unassisted decoys leading to structure solutions up to 213 residues for globular targets and 223 residues for transmembrane protein targets. The addition of contacts to limit the conformational search space enabled structure solutions for the largest target in the globular target dataset with a 221-residue chain length and the transmembrane dataset with a 249-residue chain length. The fact that both of these targets are the largest in their sets is highly suggestive that contact-assisted decoys may enable solutions for much larger targets. In fact, recent research highlighted the successful *ab initio* structure prediction of globular and transmembrane protein targets with target chain lengths in excess of 300 residues [74], which further supports this claim.

AMPLE was previously also limited by the target fold [114]. Whilst the majority of all-α protein targets were comfortably tractable, mixed α-β and all-β targets were not [114]. This limitation primarily arose from upstream limitations in *ab initio* protein structure prediction but also the challenging task of tracing β-sheets in SHELXE, which was used to assess the successful structure solution. The use of contact-assisted decoys in AMPLE improved the target tractability for β-structure-containing protein targets. Structure solutions for four additional, β-structure-containing targets were obtained when contact-assisted decoys were used. A novel approach of combining β-sheet-specific

contact pairings with a normal base prediction enabled the structure solution of one further target. Although no MR structure solutions were lost when BBCONTACTS contact pairs were added to a base set of contact restraints, further studies are required to support routine application in AMPLE. Furthermore, BBCONTACTS contact pairs were identified by analysis of a CCMPRED predicted contact map, which is generally much noisier than metapredictor alternatives. Thus, further studies may explore the benefits or drawbacks of BBCONTACTS based on alternative predicted contact maps. Lastly, since the release of BBCONTACTS, other β-strand specific contact identification protocols have been developed [153], which may need to be explored too.

Beyond the proof-of-concept study outlined in this chapter, it is very important to appreciate new limitations and unexplored areas of this work. At the time of conducting this study, PCONSC2 proved to be the state-of-the-art metapredictor. However, numerous alternatives have since been developed with more advanced Machine Learning architectures to post-process multiple individual contact predictions [e.g., 71, 101]. Furthermore, the optimal introduction of contacts as distance restraints into *ab initio* protein structure prediction protocols is not yet clearly defined, and thus leaves the choice to the user without much comparison or guidance as to which works best. Lastly, contact information was exclusively used to restrain the *ab initio* protein structure prediction procedure despite other potential applications in the AMPLE cluster-and-truncate algorithm. Subsequent chapters therefore explore additional uses of contact information for obtaining more accurate structure predictions (Chapter 4), identifying the implications on different structure prediction protocols (Chapter 5), and establishing improved decoy selection for better AMPLE processing (Chapter 6).

Chapter 4

Evaluation of ROSETTA distance-restraint energy functions on contact-guided *ab initio* structure prediction

4.1 Introduction

The extended tractability of the AMPLE program for globular and transmembrane protein targets through the use of residue-residue contact predictions to restrain *ab initio* structure prediction was highlighted in Chapter 3. However, the study explored exclusively the metapredictor PCONSC2 for globular targets without considering any alternatives. It thus served solely as proof of principle for applications of contact predictions in unconventional MR.

Besides the individual contact prediction algorithms employed by the PCONSC2 protocol, numerous metapredictors have been developed exploiting different combinations of starting alignments and individual contact predictors to identify the strongest correlating pairs for optimal contact prediction [75, 91, 100, 101, 108, 110, 111]. Furthermore, each of those protocols typically includes its own post-prediction algorithms to find a consensus amongst individual predictions and/or further identify patterns characteristic for residue pairings between secondary structure elements in a protein fold. Thus, depending on the overall protocol, the resulting predictions may differ significantly despite the same underlying algorithms to generate starting alignments and to predict residue contact pairs.

Furthermore, the precision of contact predictions used as distance restraints in *ab initio* structure prediction improves the accuracy of the folding process significantly. However, a diversity of structure prediction protocols, whether fragment-based or not, have been applied and each with a unique integration of contact information as distance restraints [69–71, 101, 132, 154]. Such divergence results in three major problems: (1) researchers cannot directly compare results, and thus have to test each protocol against their own with every newly published approach; (2) novice users might find it difficult to make appropriate decisions given the diversity of algorithms and lack of comparative studies; and (3) users only interested in the information encoded in predicted contact pairs are at risk of picking the most readily available approach over the most accurate for their problem.

Thus, the presented work was aimed at extensively comparing state-of-the-art contact prediction and *ab initio* protein structure prediction protocols with a focus on the use of such resulting decoys in unconventional MR, with a particular focus on AMPLE users.

4.2 Materials & Methods

4.2.1 Target selection of PREDICTORS dataset

This study was conducted using 18 out of 27 targets in the PREDICTORS dataset (Section 4.2.1). All nine targets with PFAM alignment depths of less than 100 were excluded (Table A.2).

4.2.2 Contact prediction

Residue contacts for each target sequence were predicted using three different metapredictors, namely METAPSICOV [101], GREMLIN [91], and PCONSC2 [100]. Web servers for METAPSICOV v2016-02 (`http://bioinf.cs.ucl.ac.uk/METAPSICOV`) and GREMLIN v2015-12 (`http://gremlin.bakerlab.org`) were used to retrieve two sets of contact predictions. Web servers were preferred in this study over local installations to best imitate the typical behaviour of AMPLE users. Both servers were used with default settings.

The GREMLIN web server returns the raw contact prediction files as well as preformatted ROSETTA distance restraints. The raw contact prediction files were downloaded to allow different contact selection thresholds as well as local conversion into ROSETTA restraints files. The METAPSICOV web server returned two contact prediction files, one after STAGE1 and another after STAGE2 post-prediction processing. In this study, contact predictions after STAGE1 (referred to as METAPSICOV from here onwards) were chosen. The PCONSC2 contact prediction set was obtained using a local installation of PCONSC2 due to downtime of the web server at the time of this study. The settings and databases were identical to Section 3.2.2. Additionally to the three main contact predictions outlined above, a set of BBCONTACTS restraints was obtained for protein targets containing β-strands (Section 2.2).

The sequence-database versions of all three metapredictors, whether on- or offline, were identical to those outlined in Section 3.2.

4.2.3 Contact-to-restraint conversion

Contact restraints for *ab initio* protein structure prediction were generated by selecting the top-ranking contact pairs from each prediction and reformatting them into a ROSETTA-readable format. The number of top-ranking contact pairs varied according to the two energy functions used (FADE cutoff: L; SIGMOID cutoff: $3L/2$; where L corresponds to the number of residues in the protein chain). Both energy functions are sigmoidal functions and introduced into the ROSETTA folding protocol in the same

fashion.

Neither energy function enforces a specified distance between restrained atoms but reward those that meet it. The two energy functions (Fig. 4.1) differ in that the FADE function does not only have an upper but also a lower bound. Based on previous findings [70, 100], the FADE function was set to acknowledge a formed restraint if the participating Cβ atoms (Cα in case of Gly) were within 9Å. In comparison, the SIGMOID function was defined with amino acid specific distances for Cβ atoms (Cα in case of Gly) to recognise the different sizes of each amino acid [71, 91].

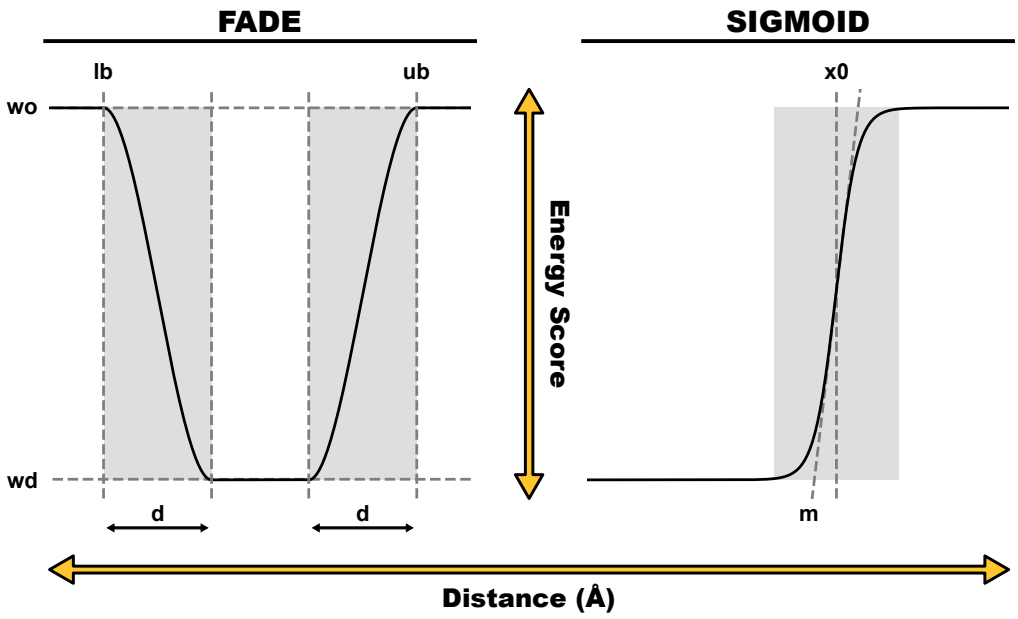


Figure 4.1: Schematic comparison of two ROSETTA energy functions. Abbreviations corresponds to input parameters.

To explore the effects of the varying ROSETTA energy function definitions, six separate contact restraint lists were created for each α-helical target and nine for each β-structure containing target. The top-ranking contact pairs per prediction were converted using the PCONSFOLD definition of the FADE function [70], the GREMLIN definition of the SIGMOID function [71], and additionally the PCONSC2+BBCONTACTS definition of the FADE function for β-structure containing targets (Section 3.2.3).

The conversion was handled in AMPLE and invoked with the keywords outlined in Table 4.1. The `-restraints_factor` keyword defines the factor used to select contact pairs based on the target chain length, i.e. a factor of 1.5 would correspond to $3L/2$ contact pairs. The `-distance_to_neighbour` keyword defines the minimum distance in sequence space between contact pair participating residues, which were set to 5 residues for the FADE function [70] and 3 for the SIGMOID function [71]. Additionally, all distance restraints were given an additional weight when introduced via the SIGMOID energy function to balance its energy term with all remaining terms in the ROSETTA

scoring function (Sergey Ovchinnikov, personal communication). This was achieved by using the `-restraints_weight` keyword and weights of 1.0 and 3.0 for the FADE and SIGMOID energy functions.

The addition of BBCONTACTS to existing sets of contacts was achieved with the FADE function in an identical manner as described in Section 2.2. In comparison, the `SCALARWEIGHTED` term in the GREMLIN implementation of the SIGMOID energy function [71] was multiplied by the number of occurrences of each contact pair in the combined map.

Table 4.1: AMPLE keyword arguments for FADE and SIGMOID ROSETTA energy functions.

Energy Function	**AMPLE keywords**

FADE	`-contact_file <FILENAME>`
	`-contact_format <FORMAT>`
	`-energy_function FADE`
	`-restraints_factor 1.0`
	`-distance_to_neighbour 5`
	`-restraints_weight 1.0`

FADE (BBCONTACTS)	`-contact_file <FILENAME>`
	`-contact_format <FORMAT>`
	`-energy_function FADE`
	`-restraints_factor 1.0`
	`-distance_to_neighbour 5`
	`-restraints_weight 1.0`

SIGMOID	`-contact_file <FILENAME>`
	`-contact_format <FORMAT>`
	`-energy_function SIGMOID`
	`-restraints_factor 1.5`
	`-distance_to_neighbour 3`
	`-restraints_weight 3.0`

SIGMOID (BBCONTACTS)	`-contact_file <FILENAME>`
	`-contact_format <FORMAT>`
	`-energy_function SIGMOID_bbcontacts`
	`-restraints_factor 1.5`
	`-distance_to_neighbour 3`
	`-restraints_weight 3.0`

4.2.4 *Ab initio* protein structure prediction

Six or nine individual lists of contact restraints generated for each target were used in separate ROSETTA *ab initio* protein structure prediction runs. Additionally, pro-

tein structures were predicted without any predicted contact restraints to acquire a control decoy set. Homologous fragments were excluded during fragment library generation to imitate the folding process of a target with unknown fold. Fragment libraries were generated for each target using a local installation of ROSETTA v2015.22.57859. PSIPRED secondary structure predictions were included from contact prediction runs and provided via the `-psipredfile` option. In total, 1,000 *ab initio* decoys were generated per run using ROSETTA v2015.22.57859 with default settings [42] and one of the seven or ten (six/nine plus control) contact conditions described in Section 4.2.3. In total, 162 sets of models were generated across 18 protein targets.

4.2.5 Molecular Replacement

Besides considering decoy quality, one key interest of this study was the assessment of the decoy sets created in the previous step as *ab initio* MR search model templates. To reduce the enormous computational cost linked to trialling 162 decoy sets, 108 sets were chosen from the following conditions: simple ROSETTA, PCONSC2 prediction and FADE function, GREMLIN prediction and SIGMOID function, METAPSICOV prediction and FADE function, and where applicable, PCONSC2+BBCONTACTS, GREMLIN+BBCONTACTS and METAPSICOV+BBCONTACTS predictions and FADE function. Overall, this resulted in four MR runs for the six α-helical targets, seven runs for the six all-β, and seven runs for the six mixed α-β targets. The resulting 108 model sets were trialled in AMPLE v1.1.0 and CCP4 v7.0.28. Structure solution success was assessed as described in Section 2.3.4.2.

4.3 Results

4.3.1 Direct comparison of three contact metapredictors

In this study, a direct comparison between three metapredictors — GREMLIN, METAPSICOV and PCONSC2 — was carried out. Residue-residue contact pairs were predicted for 18 protein target sequences with a range of chain lengths and numbers of effective sequences in their PFAM MSAs.

METAPSICOV was the most precise contact predictor across the protein target dataset in this study (Fig. 4.2). The difference between the three metapredictors was most evident in the highest-scoring contact pairs ($L/10$). The median precision values for METAPSICOV and PCONSC2 contact predictions were above 50% up to L contact pairs. GREMLIN, in comparison, predicted contacts with a median precision score at least 20% worse than that of METAPSICOV and 15% worse than PCONSC2. However, at $3L/2$ contact pairs the median precision scores were much more similar across the three different metapredictors: METAPSICOV and PCONSC2 were near

identical, and GREMLIN is at most 12% worse compared to the other two. Inspecting the mean precision scores over a continuous range of selection cutoff values illustrated further the difference between METAPSICOV, PCONSC2 and GREMLIN (Fig 4.3). The former two similarly high precision scores compared to the average precision scores for GREMLIN, which were approximately 0.2 precision score units lower. Added to the difference in precision scores was the difference in sequence coverage (Fig. 4.3). Although producing the on-average worst contact predictions out of the three metapredictors, GREMLIN contact predictions had the highest sequence coverage. However, an analysis of singleton contact pairs, usually with high degrees of FP predicted contacts, revealed a positive correlation ($\rho_{Pearson} = 0.47$; $p < 0.001$) between the fraction of singleton contact pairs and sequence coverage and hinted to a weak negative correlation ($\rho_{Pearson} = -0.27$; $p < 0.05$) between the fraction of singleton contact pairs and contact precision (Fig. 4.4).

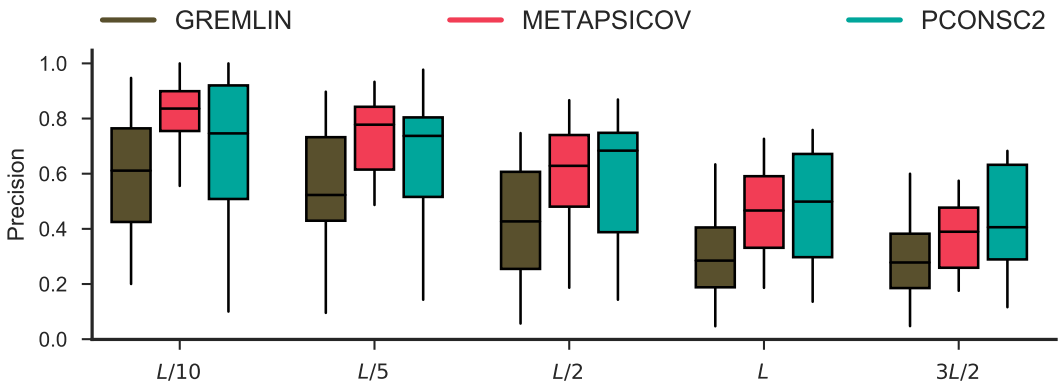


Figure 4.2: Distribution of precision values for three metapredictors computed at five contact selection cutoff values relative to the target chain length (L).

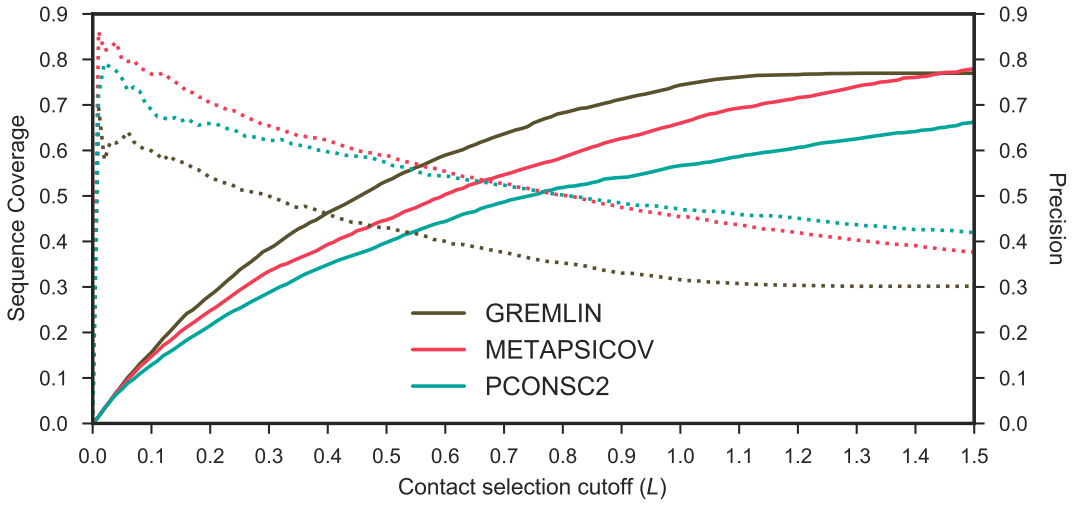


Figure 4.3: Average sequence coverage (line) and contact prediction precision scores (dashed) across a continuous range of contact selection cutoffs ranging from $[0.0, 1.5]$ for all targets.

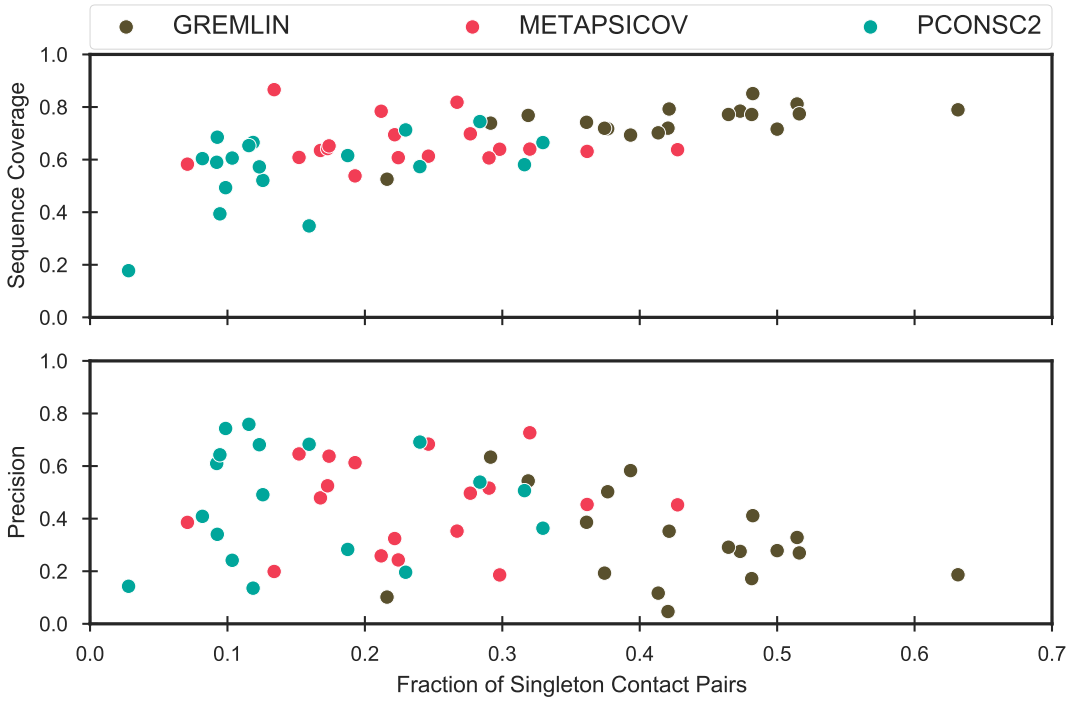


Figure 4.4: Contact singleton analysis compared against the precision of top-L contact pair lists for three metapredictors.

Given that the overall precision of contact pairs predicted by the three metapredictors differed, it was important to understand where the difference originated. To investigate this, a comparison of the precision values at different cutoff levels on a per-target basis was performed. For the majority of targets the precision scores were very similar across the three metapredictors (Fig. 4.5). However, the prediction precision of some targets differed significantly. For example, the METAPSICOV prediction for the human retinoic acid nuclear receptor HRAR (PDB: 1fcy) contained high precision in its highest scoring (top-$L/10$) contact pairs (Fig. 4.5). In comparison, GREMLIN and PCONSC2 predictions for the same target contained less precise contact pairs ($\Delta_{\text{METAPSICOV-GREMLIN}}\ L/10 = -0.522$; $\Delta_{\text{METAPSICOV-PCONSC2}}\ L/10 = -0.435$). However, the addition of further contact pairs up to $3L/2$ resulted in near-identical precision across the three metapredictors for this target. A second example illustrating such a difference were the contact predictions for the human galectin-3 CRD sequence (PDB: 4lbj). In contrast to the previous example, the data showed high precision scores for the METAPSICOV and PCONSC2 predictions for this target, yet low precision for the top GREMLIN contact pairs ($\Delta_{\text{METAPSICOV-GREMLIN}}\ L/10 = -0.231$; $\Delta_{\text{METAPSICOV-PCONSC2}}\ L/10 = 0.077$).

Figure 4.5: Contact prediction precision scores from three metapredictors for 18 targets at different contact pair selection thresholds (L, which is the target chain length). The PFAM alignment depth is given by means of M_{eff}. The colour scale corresponds to the precision in range $[0, 1]$.

The data presented in Fig. 4.5 also indicated that there was no direct link between chain length or M_{eff} and the precision of the resulting contact predictions. The N-(5'-phosphoribosyl)anthranilate isomerase sequence (PDB: 4aaj) with a chain length of 228 residues and 750 effective sequences in its PFAM MSA yielded a mean precision at $L/10$ contact pairs of 0.283 (top-L: 0.195) across the three metapredictors. This strongly contrasted with the sequence of sortase B (PDB: 2oqz), which showed similar characteristics yet obtained mean precision at $L/10$ contact pairs of 0.938 (top-L: 0.622).

Although the contact predictions differed in precision, an interesting question rested with the similarity of the predicted contact pairs amongst the sets. Thus, the similarity of contact predictions across the three metapredictors was an important metric to evaluate the most appropriate algorithm for AMPLE users. Using the Jaccard similarity index to evaluate the direct overlap of contact pairs across sets of predictions, the data suggested very little similarity between the contact predictions of the three metapredictors for each target (Fig. 4.6). As with the differences in precision scores at higher cutoff thresholds, the Jaccard index was also lower — indicating less overlap — at higher cutoff thresholds. However, it is worth noting that the Jaccard index only considers identical matches and did not consider the neighbourhood of a contact pair. Thus, the index does not highlight similar regions with contact pairs in both maps.

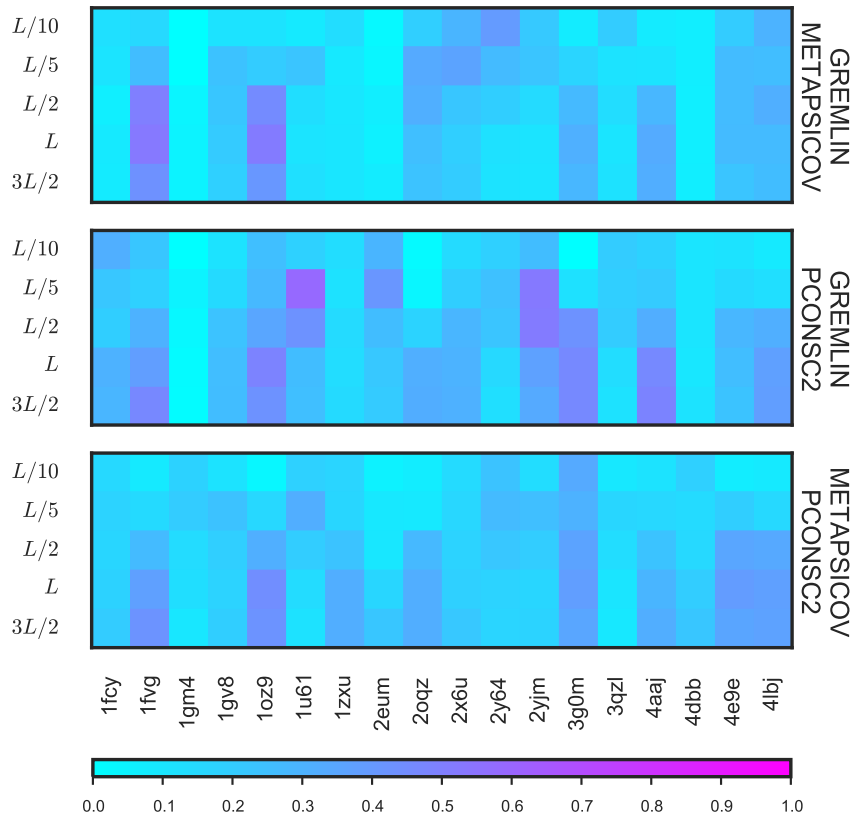


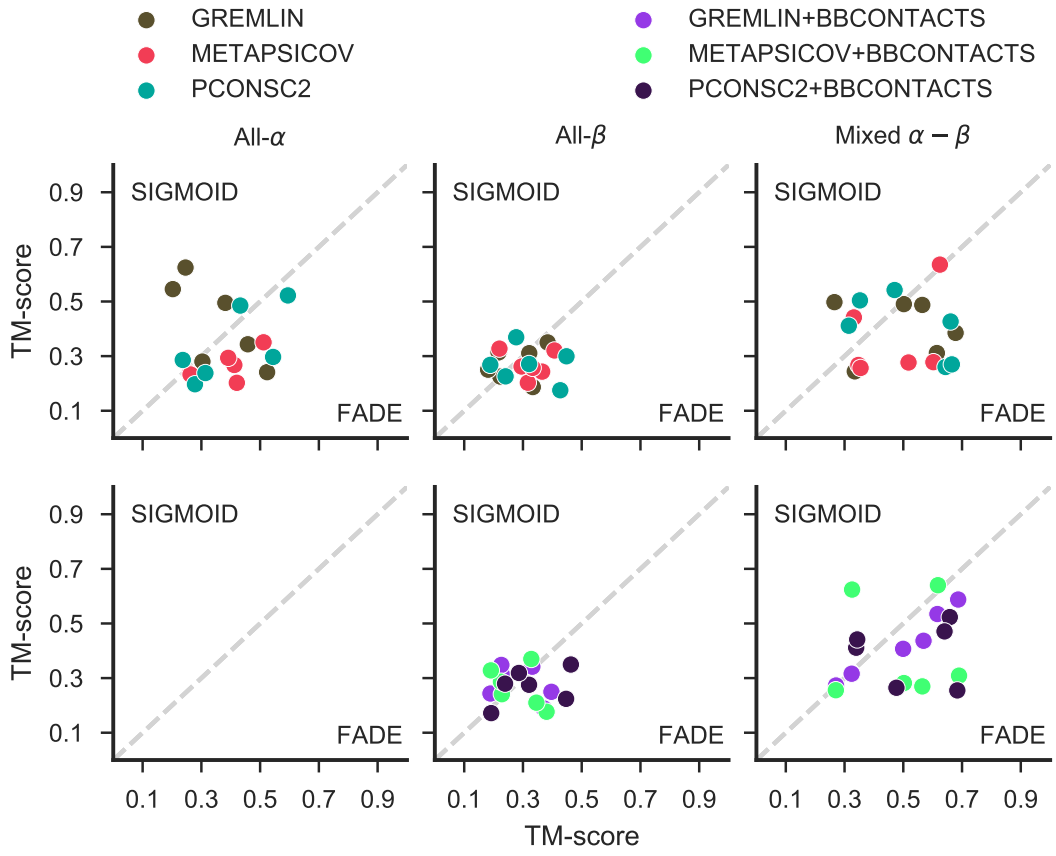
Figure 4.6: Jaccard similarity index illustrates a higher degree of overlap between metapredictor contact predictions with increasing numbers of contact pairs included in the calculation. The three panels show the different comparisons. The colour scale corresponds to the Jaccard index in range $[0, 1]$.

4.3.2 Protein structure prediction with two ROSETTA energy functions

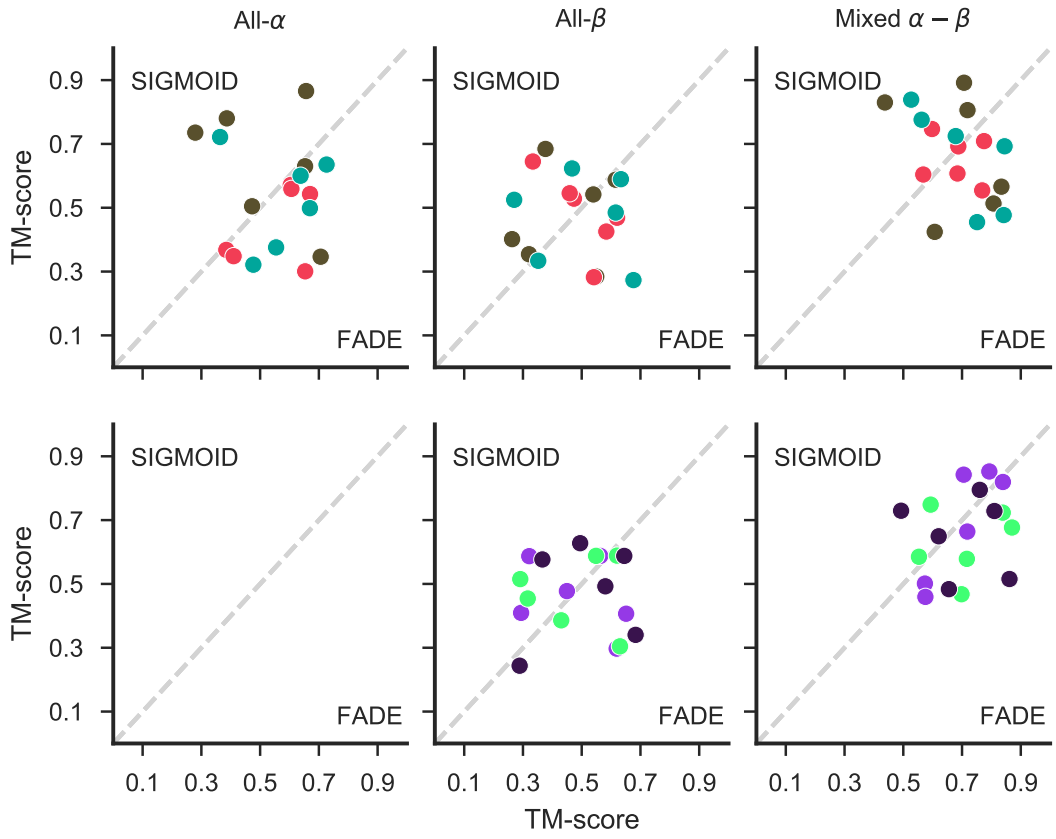
The accuracy of the starting decoys is a major factor for an AMPLE run to succeed [118]. Thus, the quality of the decoys was of great essence to this study. Given the two different ROSETTA energy functions, FADE and SIGMOID, all predicted contacts were subjected to individual *ab initio* structure prediction runs. Additionally, all contact predictions were enriched with BBCONTACTS for all β-containing targets in separate trials. A total of 234,000 individual decoys were generated in this study across all targets, contact predictions and ROSETTA energy function combinations.

Separating these individual decoys solely by the ROSETTA energy function (excluding unrestrained ROSETTA decoys) showed that the FADE energy function resulted in marginally more accurate decoys (median TM-score FADE: 0.3541; median TM-score SIGMOID: 0.2969). To further investigate which energy function was more suitable for the target dataset, the decoy sets were grouped by two additional characteristics: the fold of the target, and the source of distance restraints used. The results strongly suggested that the FADE energy function results in more accurate decoy sets

(Fig. 4.7a), outperforming the SIGMOID energy function by median TM-score in two-thirds of all decoys sets (FADE: 58; SIGMOID: 32). A split of the decoy sets into separate categories by fold and the addition of BBCONTACTS revealed that the SIGMOID energy function only yields similar results for all-β targets in combination with BBCONTACTS-supported distance restraints. Although the total count of decoy sets with higher accuracies between the two energy functions in this category were similar, the actual differences in TM-scores further supported the strength of the FADE energy function compared to the SIGMOID.



(a)



(b)

Figure 4.7: (a) Median and (b) top-1 decoy TM-score comparison of FADE and SIGMOID ROSETTA energy functions differentiated by fold and the addition of BBCONTACTS restraints.

Besides the structure prediction accuracy of each set of decoys, the single, most accurate decoy is also of great interest. If one energy function consistently predicted single decoys more accurately, it might be appropriate to reconsider the structure identification routine (i.e. clustering) in AMPLE for search model preparation. However, a similar difference to that of the decoy quality of entire sets was observed for the top-1 decoy in each set (Fig. 4.7b). The FADE energy function outperformed the SIGMOID function for the majority of target-contact prediction combinations (FADE: 51; SIGMOID: 39). However, the GREMLIN distance restraints in combination with the SIGMOID energy function produced better top-1 decoys than GREMLIN restraints with the FADE energy function. This suggested that GREMLIN restraints and the SIGMOID energy function were tailored to complement each other with the ultimate goal of predicting single decoys to high accuracy over entire sets of decoys. Additionally, the spread of decoy quality differences between the two energy functions widens when only looking at the best decoy in each predicted set (ΔMedian TM-score$_{\mathrm{ALL}}$: $min = 0.002, max = 0.429$; ΔMedian TM-score$_{\mathrm{TOP}}$: $min = 0.002, max = 0.456$).

A Kernel Density Estimate (KDE) of TM-scores using each predicted decoy was generated with the TM-scores of individual decoys separated only by fold class and ROSETTA energy function (Fig. 4.8). This density estimate further supported the results presented above: the FADE energy function generated more accurate decoys. However, a very important detail is highlighted by the KDEs. Distinct regions with high density are visible in the estimates of the TM-scores of individual decoys for all-α and mixed α-β targets (Fig. 4.8). The bimodal distribution of decoy TM-scores from both energy functions strongly suggests that predicted structures were either native-like or not (based on the TM-score threshold of ≤ 0.5). However, the number of correctly predicted decoys versus incorrectly predicted decoys was in favour of the latter. The decoy sets of all-β targets did not show such distinct regions of high density for decoys with TM-scores of less than 0.5 in any of its KDEs (Fig. 4.8). The generally poor decoy quality of decoys predicted without any predicted distance restraint information (ROSETTA) highlighted the benefit of contact predictions to *ab initio* protein structure prediction.

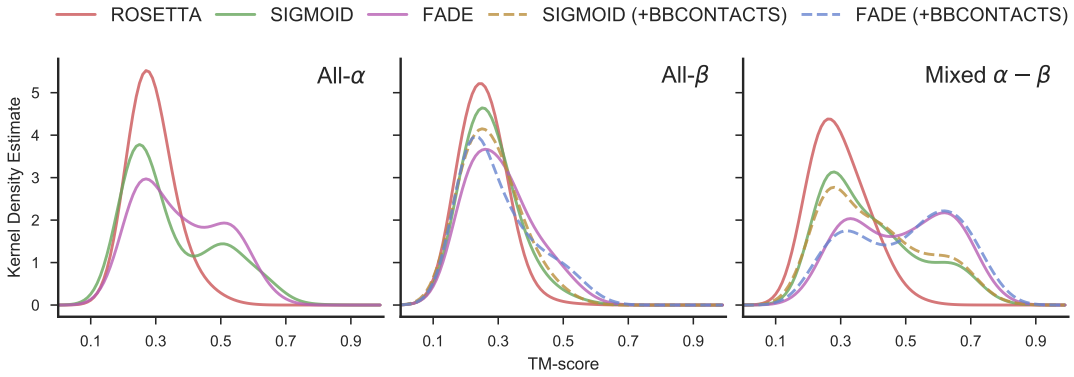
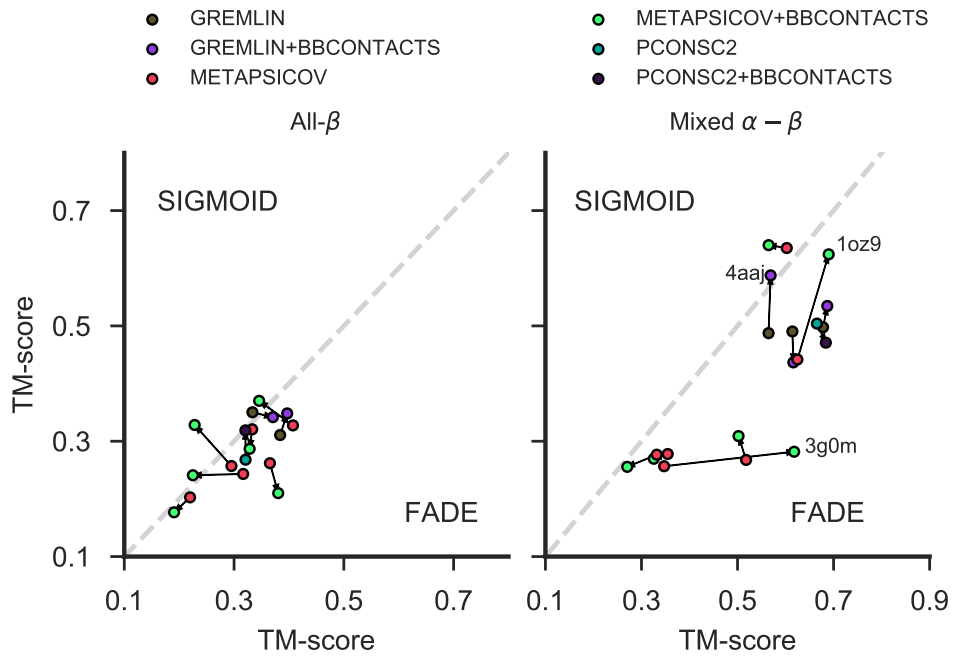


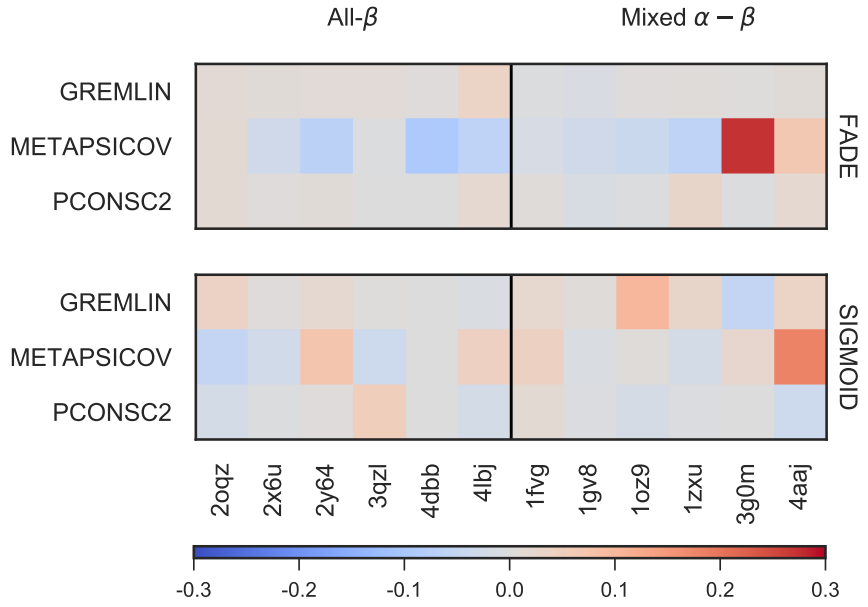
Figure 4.8: KDEs of TM-scores of all decoys in each respective fold class separating by ROSETTA energy function (SIGMOID or FADE) and no contact information used (ROSETTA). Dashed lines indicate decoys which were predicted with the addition of BBCONTACTS predictions.

A further important aspect of the presented work is the demonstration of the benefits of BBCONTACTS prediction addition to the *ab initio* protein structure prediction of β-containing targets. Although previous results described in Chapter 3 in combination with those presented above outlined overall improvements in decoy quality, it was essential to understand which targets benefit from this treatment. Figure 4.9a highlights the effects of adding BBCONTACTS restraints to the structure prediction strategies employed here. In summary, the addition of BBCONTACTS restraints hardly affected the decoy quality of most targets under the various contact prediction and energy function combinations. Nevertheless, three target, contact prediction and energy function combinations yielded TM-score improvements of at least 0.1 TM-score units compared to the same condition without the addition of BBCONTACTS restraints. In contrast, the addition of BBCONTACTS restraints did not lower the median TM-score by more

than 0.1 for any target (Fig. 4.9b).



(a)



(b)

Figure 4.9: Median TM-score comparison of FADE and SIGMOID ROSETTA energy functions differentiated by fold (excl. all-α). (a) Arrows indicate the effect on decoy quality through the addition of BBCONTACTS restraints. Targets with a distance of less than 0.03 TM-score units between normal and BBCONTACTS-added conditions were excluded from the scatter plots. (b) Effect on decoy quality through the addition of BBCONTACTS restraints highlighted by heatmap difference. The colour scale corresponds to the difference in median TM-score between normal and BBCONTACTS-added contact maps.

Two further aspects in understanding the differences in effects of the FADE and SIGMOID ROSETTA energy functions on decoy quality were the target chain length and restraints precision. The former appeared to affect the final decoy quality of all 1,000 decoys insignificantly (Fig. 4.10). However, the restraint precision resulted in some differences between the two ROSETTA energy functions (Fig. 4.10). The FADE energy function (top-L restraints) generally appeared to be less sensitive to restraint lists with higher FP contact pairs. In contrast, the SIGMOID function ($3L/2$ restraints) produced less accurate decoys than the FADE function with more accurate restraints. Most strikingly, the FADE energy function generated decoys with a median TM-score of 0.678 for the N-(5'-phosphoribosyl)anthranilate isomerase domain (PDB: 4aaj) compared to the SIGMOID function with a median TM-score of 0.498. Nevertheless, both energy functions appeared to broadly follow a positive linear trend, i.e. better restraint precision resulted in more accurate decoys.

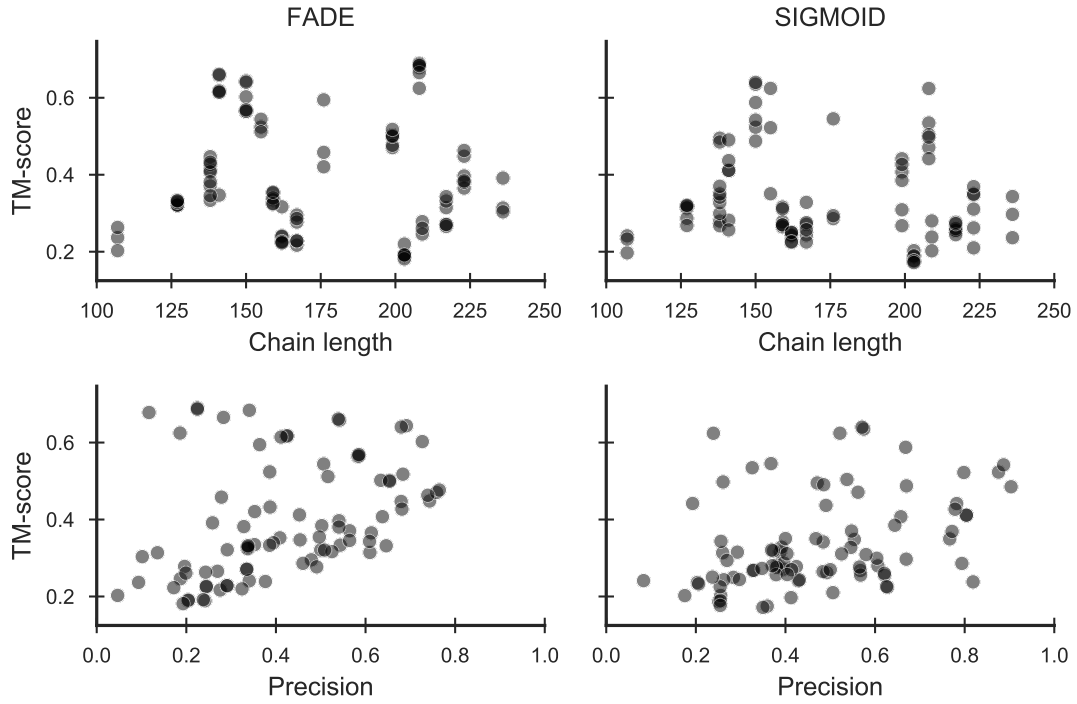


Figure 4.10: Effects of target chain length and restraint precision on the median TM-score for FADE and SIGMOID ROSETTA energy functions. Each scatter point represents a 1,000-decoy set.

4.3.3 Impact of metapredictors and energy functions on AMPLE

The results obtained from the decoy quality comparison outlined in Section 4.3.2 highlighted differences between the FADE and SIGMOID ROSETTA energy functions. This difference was more pronounced for some targets and did not generalise well in favour of one energy function. Thus, the next step in this study was to analyse the consequences of these differences for unconventional MR using the automated pipeline AMPLE.

Overall, the decoys restrained with GREMLIN distance restraints via the SIGMOID energy function throughout the *ab initio* protein structure prediction process yielded six out of 18 possible structure solutions (Fig. 4.11). This result was the highest of all trialled conditions and only resulted in one more structure solution compared to unrestrained ROSETTA decoys. All remaining conditions resulted in fewer structure solutions than those from ROSETTA decoys. Furthermore, the conditions METAPSICOV (FADE function), METAPSICOV BBCONTACTS (FADE function) and PCONSC2 BBCONTACTS (FADE function) yielded no more than half of the structure solutions achieved by GREMLIN (SIGMOID function). The remaining two conditions — PCONSC2 (FADE function) and GREMLIN+BBCONTACTS (FADE function) — resulted in four out of 18 structure solutions. The addition of BBCONTACTS did not improve decoy quality enough to increase the chances of structure solution success; however, the structure of the bovine peptide methionine sulfoxide reductase (PDB: 1fvg) was only solved with the GREMLIN+BBCONTACTS (FADE function) decoys further supporting the small but important value of BBCONTACTS restraint addition to separately determined contact predictions.

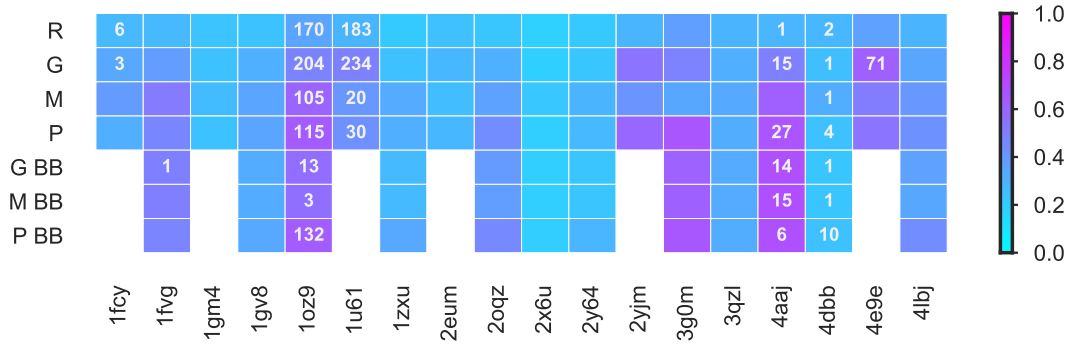
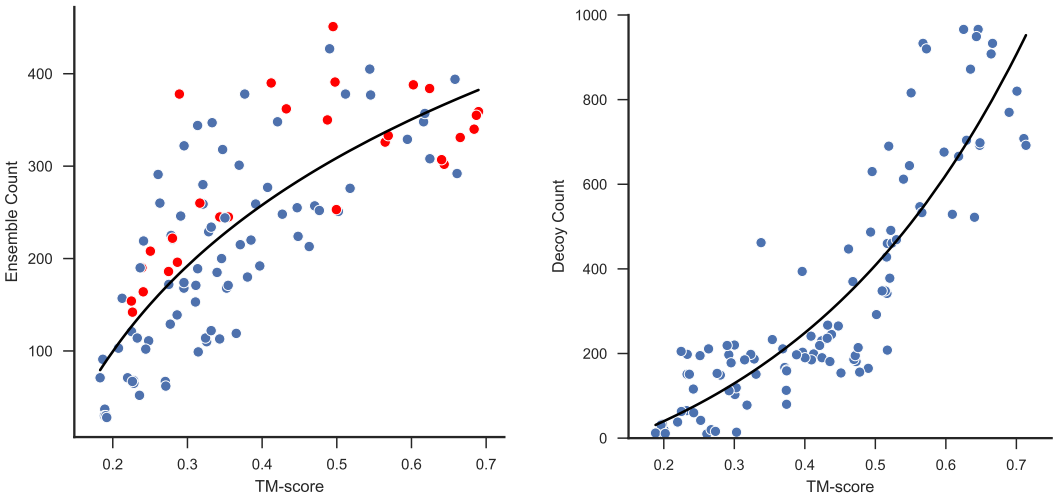


Figure 4.11: Structure solution count for AMPLE search models generated from decoys with varying contact prediction and ROSETTA energy function conditions: unrestrained ROSETTA (R); GREMLIN (G; SIGMOID function); METAPSICOV (M; FADE function); PCONSC2 (P; FADE function); GREMLIN+BBCONTACTS (G BB; FADE function); METAPSICOV+BBCONTACTS (M BB; FADE function); PCONSC2+BBCONTACTS (P BB; FADE function). The colour scale of each square indicates the median TM-score of all 1,000 starting decoys.

The number of structure solutions obtained from the decoy sets subjected to the AMPLE pipeline were somewhat surprising given that ROSETTA decoys resulted in the second-most structure solutions. These results suggest that AMPLE was unable to exploit the true value of more accurate decoy sets. This hypothesis was further supported when considering the decoy set quality and the number of structure solutions (Fig. 4.11). For example, PCONSC2 (FADE function) decoys predicted for the hypothetical protein AQ_1354 (PDB: 1oz9) showed high accuracy, and thus would generally be considered highly desirable starting structures for the AMPLE protocol. Nevertheless, the AMPLE protocol was unable to exploit these decoys for successful MR

structure solution. Similarly, the high-accuracy contact-assisted decoys sets predicted for other targets, e.g. cysteine desulferation protein SufE (PDB: 3g0m; median TM-score PCONSC2+BBCONTACTS (FADE function)=0.661) also failed to result in MR solutions. In comparison, the median TM-scores for all successful ROSETTA decoy sets did not exceed 0.355 TM-score units, which suggests that the AMPLE routine may be optimised for less accurate ROSETTA decoys.

Naturally, one would expect the best decoys to result in the most accurate ensemble search models, which in turn yield the highest number of structure solutions per target. However, here we demonstrated that the most accurate decoys did not guarantee structure solution, and in contrast some poorly predicted decoy sets achieved structure solution. Thus, it was essential to investigate the stage in AMPLE's cluster-and-truncate approach at which the higher decoy quality resulted in less suitable ensemble search models for MR.

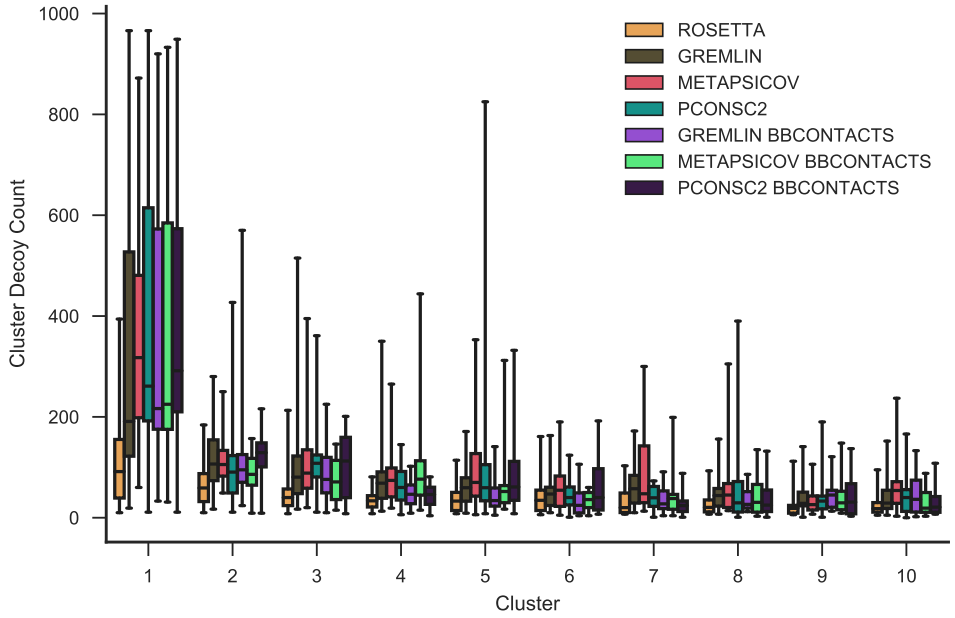


(a) (b)

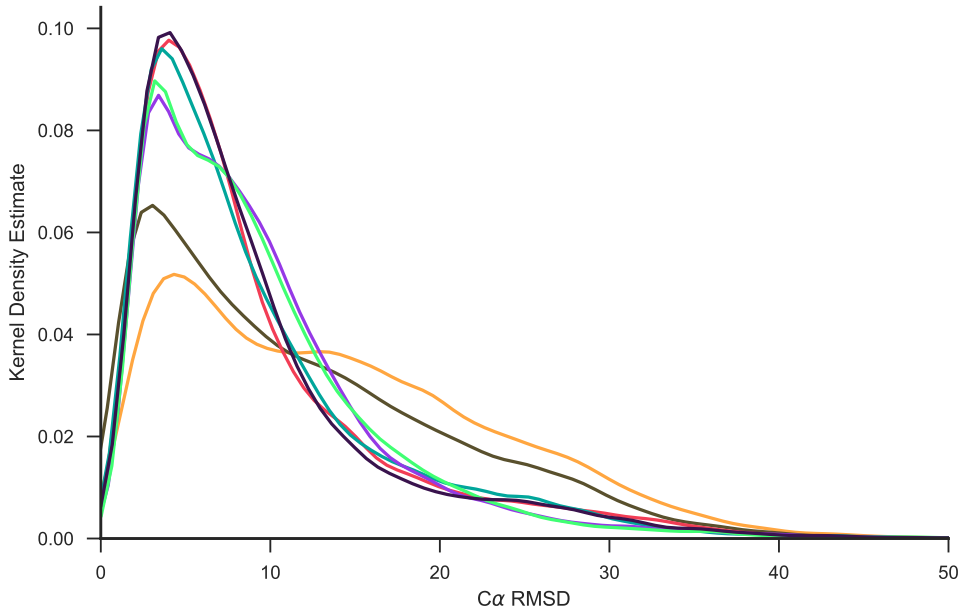
Figure 4.12: (a) Comparison of median TM-score (per 1,000 decoys) against the resulting AMPLE ensemble search model count. The equation of the line of best fit is defined by $y = 228.50 * \ln(20.96 * x) - 227.95$. Red dots indicate successful ensemble sets. (b) Relationship between cluster median TM-score and the number of cluster decoys. Blue line represents line of best fit with equation $y = 148.85 * \exp(2.90 * x) - 225.76$.

The data generated as part of this study revealed a positive correlation ($\rho_{Spearman} = 0.78$; $p < 0.001$) between the decoy quality and the number of resulting AMPLE ensemble search models. In Fig. 4.12a, the plotted data alongside a line of best fit further illustrate that small differences in decoy quality in the lower TM-score regions increased the total number of generated ensemble search models dramatically. However, once the threshold of 0.5 TM-score units was surpassed the number of generated ensemble search models plateaued at approximately 350-400 ensemble search models, which is close to the maximum AMPLE can generate from a starting set of 1,000 decoys. Furthermore, the data indicates that decoy sets containing fewer than 100 ensemble search models

do not lead to structure solution, although this result needs to be considered with care given the difficulty of predicting which search model solves a given structure.



(a)

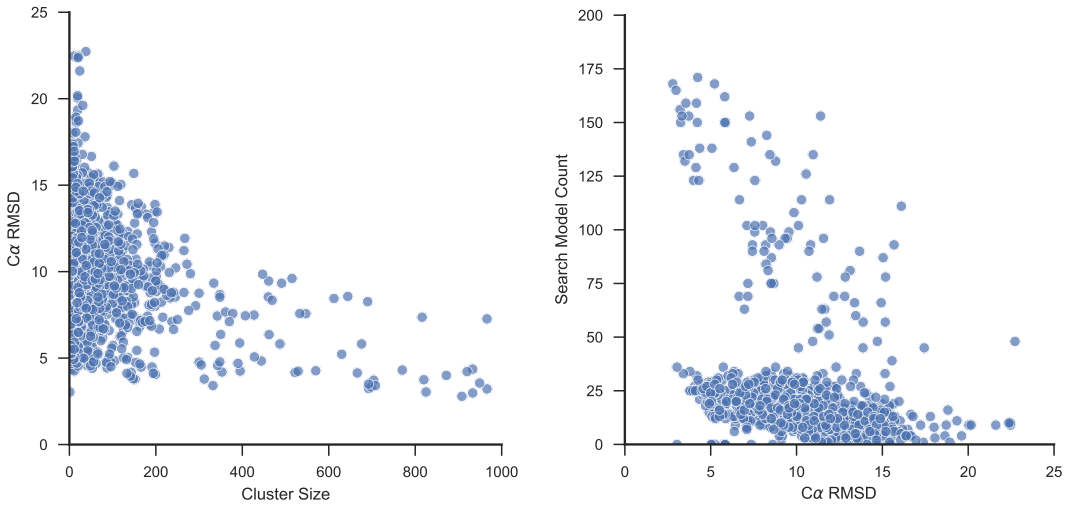


(b)

Figure 4.13: (a) SPICKER cluster sizes of each target grouped the restraint condition used during the structure prediction protocol. Whiskers span the range from the minimum to maximum counts. (b) KDE of Cα interatomic RMSD for SPICKER clusters.

Besides looking at the relationship between entire decoy sets and the resulting structure solutions on a per-target or per-condition basis, it was important to also consider individual ensemble search models, their origins and their properties in re-

lation to MR metrics. Findings outlined in Chapter 3 highlighted the relationship between the number of decoys in the first cluster and its decoy quality (see Chapter 3). Here, further support for these findings was given by means of the positive relationship between the median TM-scores and the corresponding size of the largest SPICKER cluster (Fig. 4.12b). An analysis of the cluster sizes demonstrated the downstream benefits of increased decoy quality through contact restraints in the folding process (Fig. 4.13a). The sizes of the first three clusters generated from most contact-restraint decoy sets greatly surpassed their equivalent cluster sizes for unrestrained ROSETTA decoys. Given that cluster sizes correlated with decoy quality, these results also supported the idea that the mean Cα RMSD — as calculated by THESEUS for cluster truncation — was directly related to better decoy quality via the larger number of decoys in each cluster (Fig. 4.14a). The same mean Cα RMSD was also related to the number of ensemble search models generated after subclustering (Fig. 4.14b), which hinted towards a direct relationship between increased quality of 1,000 decoys per set and the total number of ensemble search models generated. Interestingly, GREMLIN decoys showed similar Cα RMSD per cluster compared to unrestrained ROSETTA decoys (Fig. 4.13b), unlike all other contact-restraint-guided structure predictions. However, it is worth noting that almost no distinction could be made amongst the remaining contact restraint treatments despite some differences in cluster size distributions exist (Fig. 4.13a).



(a) (b)

Figure 4.14: (a) Number of decoys per SPICKER cluster plotted against the mean Cα-atom RMSD for all decoys in each cluster. (b) Mean Cα-atom RMSD for decoys per cluster plotted against the number of search models derived from the cluster.

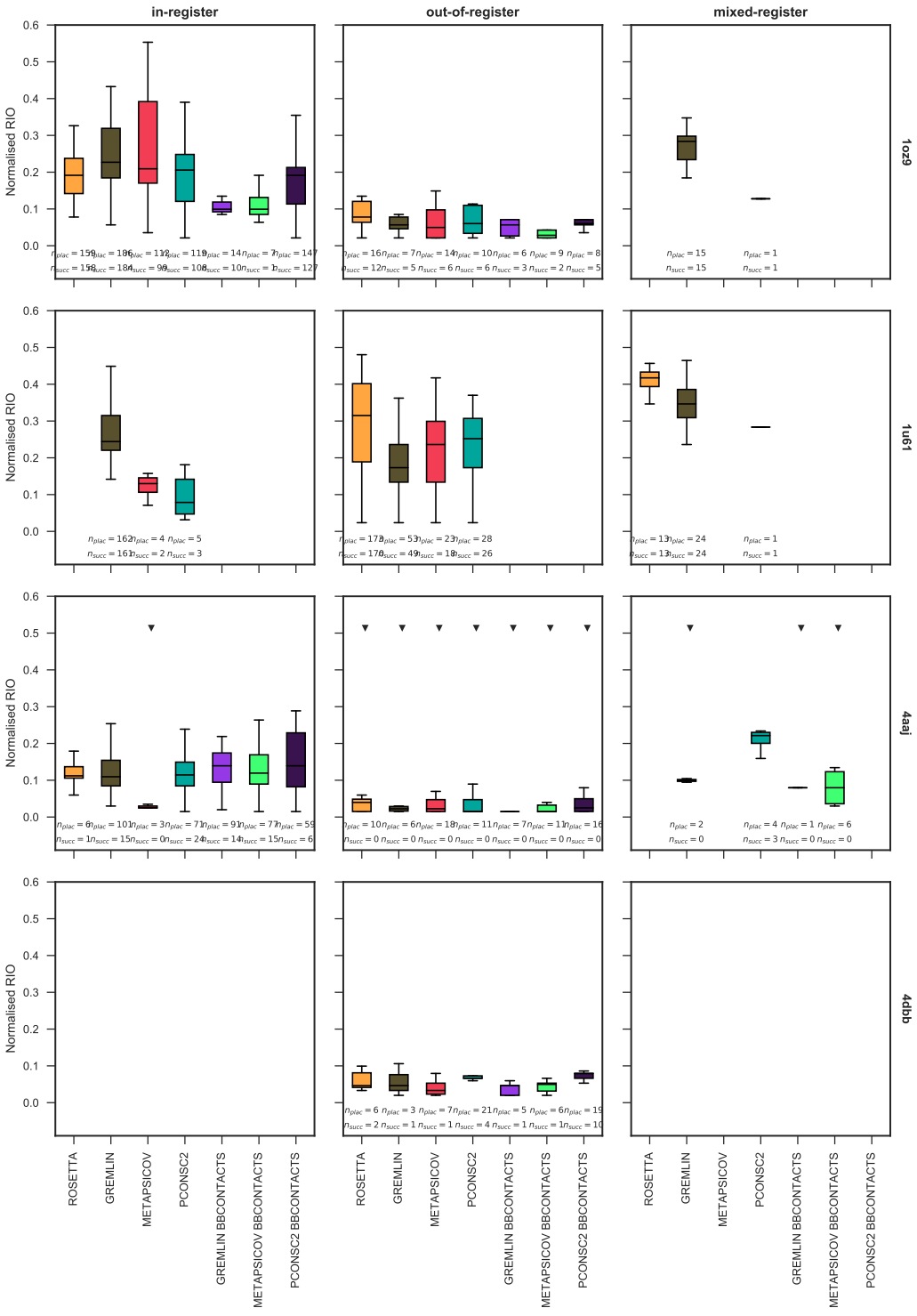
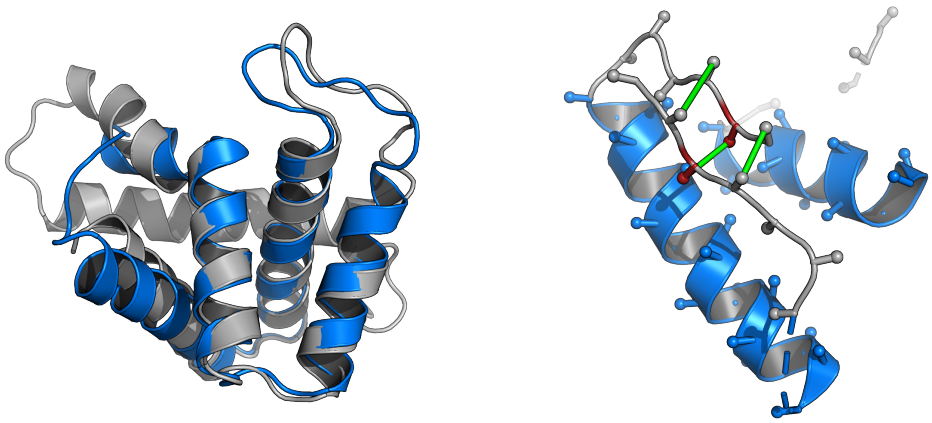


Figure 4.15: Normalised RIO score analysis of four successful targets in the MR dataset. Black triangles indicate AMPLE search model sets without a structure solution.

The structure solution through pipelines like AMPLE and other unconventional MR software [155, 156] can result from the placement of generated (ensemble) search models either in- or out-of-sequence register. The RIO metric [117] can reliably assess the register placement, and thus was used to analyse the MR placements of all search models of the seven targets with structure solutions from one or more decoy sets. The RIO

scores for the hypothetical protein AQ_1354 (PDB: 1oz9) strongly supported the high quality decoys used as input across all seven contact conditions (Fig. 4.15). Most search models were placed in-register and hardly any search models with out-of-register RIO scores failed either. In contrast, the search models of N-(5-phosphoribosyl)anthranilate isomerase (PDB: 4aaj) — derived from high quality decoys in most conditions — showed a low percentage of AMPLE search models with RIO scores leading to structure solution (Fig. 4.15). Furthermore, the RIO scores normalized by the target chain length indicated that search models, independent of MR structure solution, were relatively small only exceeding 20% of the total target sequence in a few cases.

One interesting target in this set with respect to the sequence register of the AMPLE search models leading to structure solution was the putative ribonuclease III (PDB: 1u61) domain. Although decoys from all contact conditions readily solved this target with at least 20 or more AMPLE search models, one important aspect arose from the RIO register analysis. Only GREMLIN decoys were primarily placed in-register (Fig. 4.15). AMPLE search models derived from the other three contact conditions, and in particular those from ROSETTA decoys, were primarily placed out-of-register with sequence coverage values of roughly 25%. An analysis of the diversity of AMPLE search models highlighted the accuracy of GREMLIN search models, which represented a closely-matched substructure of the target protein (Fig. 4.16a).



(a) (b)

Figure 4.16: AMPLE ensemble search models post-PHASER placement for (a) putative ribonuclease III (PDB: 1u61) and (b) peptide methionine sulfoxide reductase (PDB: 1fvg). Search models (blue) are superposed to their native crystal structures (grey). BBCONTACTS distance restraints are represented as green lines. Secondary structure assignment calculated with STRIDE [157]. In (b), red residues indicate β-strand residues.

Compared to all other targets with structure solutions in at least one condition, the PTB domain of Mint1 (PDB: 4dbb) produced similarly interesting yet somewhat surprising results. None of the search models, independent of their decoy source, had any residues placed in-register. All structure solutions were obtained from out-of-register

search model placements (Fig. 4.15). A visual inspection of all successful search models revealed that structure solutions were exclusively obtained with idealised fragments. ROSETTA, GREMLIN and METAPSICOV decoys resulted in one or more single-helix ensemble search models that led to structure solution (Fig. 4.17). More interestingly though, PCONSC2, GREMLIN+BBCONTACTS, METAPSICOV+BBCONTACTS and PCONSC2+BBCONTACTS decoys yielded one or more two-strand β-sheets which, after successful MR, yielded fully built structures (Fig. 4.17).

Lastly, three targets were solved with one or two decoy sets alone. The structures of the retinoic acid nuclear receptor HRAR (PDB: 1fcy) and the peptide methionine sulfoxide reductase (PDB: 1fvg) were only solved with a handful of AMPLE search models. Often singleton solutions like these are achieved through AMPLE's cluster-and-truncate procedure producing a single, idealised helix as search model. Here, the data confirmed this for target 1fcy, whereby single out-of-register helices derived from ROSETTA and GREMLIN decoys achieved structure solutions. However, the singleton search model derived from the GREMLIN+BBCONTACTS decoys for the peptide methionine sulfoxide reductase (PDB: 1fvg) was placed in-register. A closer inspection of this AMPLE ensemble search model highlighted a great success of the approach of adding BBCONTACTS distance restraints to separately predicted base contact maps. In this instance, the successful AMPLE ensemble search model had 77% of its 49 residues placed in-register. More importantly, the search model was made up of two β-strands packing against each other, which was supported by BBCONTACTS predictions (Fig. 4.16b). The last case, glycosylase domain of MBD4 (PDB: 4e9e), solved solely with GREMLIN decoys yielding 71 structure solutions. All successful AMPLE search models derived from the GREMLIN decoys were placed in-register.

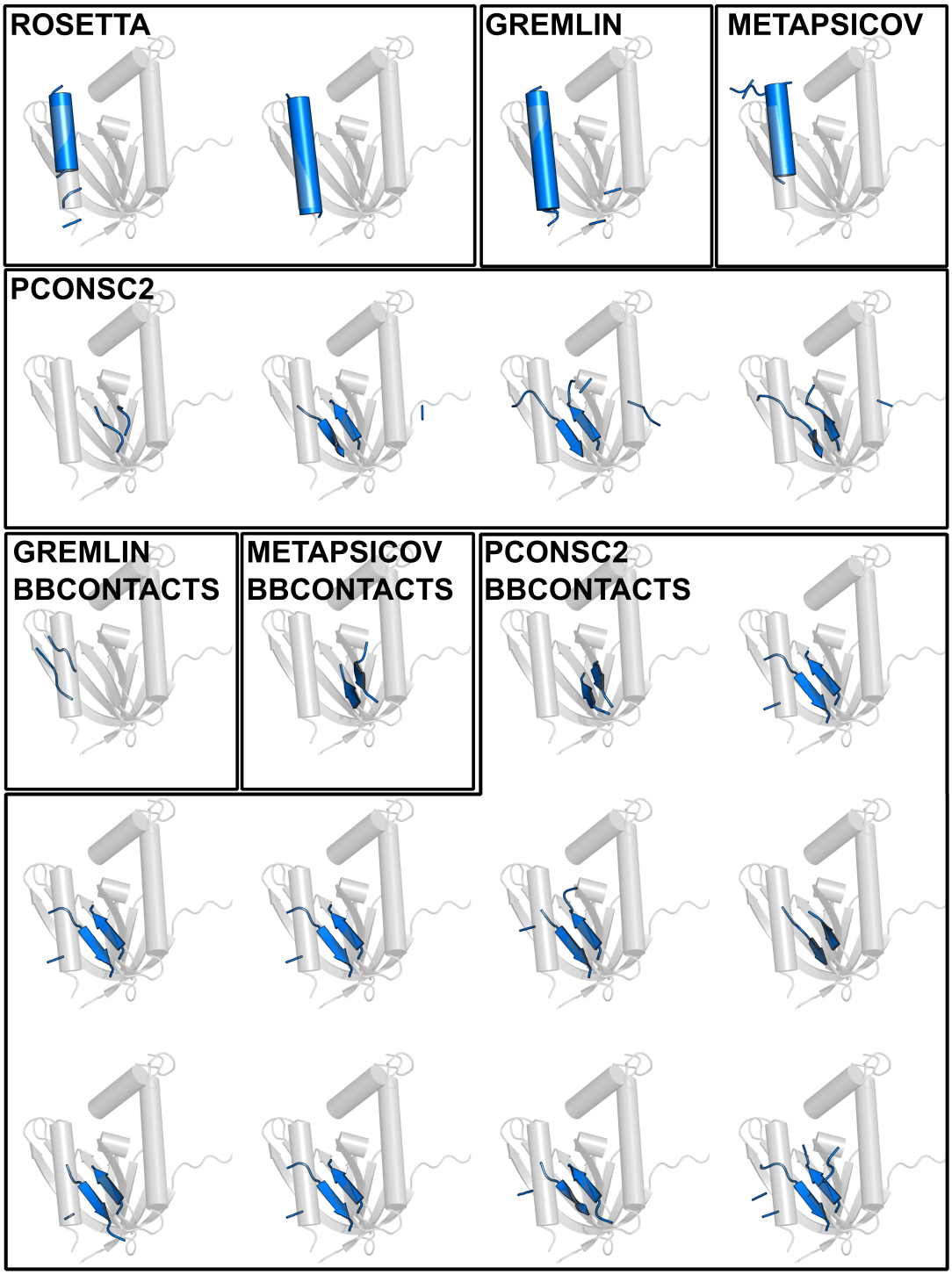


Figure 4.17: Successful search models post-PHASER placement (blue) superposed to the reference crystal structure (grey) for PTB domain of Mint1 (PDB: 4dbb).

4.4 Discussion

This study was designed to explore the state-of-the-art metapredictor pipelines for residue-residue contact prediction. The main focus of this work was to distinguish

differences in three key parts: raw contact predictions, their use in *ab initio* structure prediction and finally the effects on unconventional MR using AMPLE.

Key findings in this study revealed METAPSICOV and PCONSC2 metapredictors to yield the most precise contact predictions regardless of target fold or size. These results are in line with previous findings, which independently confirmed METAPSICOV contact predictions to yield the highest precision across numerous prediction algorithms [112, 113]. However, work in this study cannot confirm their findings, which demonstrated more precise contact predictions for all-β and mixed α-β protein targets compared to all-α ones. Several reasons might give insights into this discrepancy: (1) a much smaller sample size was trialled in this study (Wuyun et al. [113]: 680; Oliveira et al. [112]: 3500); (2) the targets were chosen to deliberately sample various alignment depths including relatively low alignment depth (< 200) values; (3) only final contact predictions were analysed, thus benefiting from post-prediction consensus finding and contact map processing through unsupervised machine-learning algorithms.

Furthermore, the results obtained in this study demonstrated that two similar ROSETTA energy functions yield different structure prediction results. The FADE function on average achieves more accurate structure predictions compared to the SIGMOID one. This result seems striking at first; however, a closer inspection of each of the energy function parameters gives possible insights into the reasons for the different outcomes. The FADE energy function defines both a maximum and minimum distance. The FADE energy function also does not consider amino acid-specific distances while the SIGMOID function does [91]. Furthermore, a custom weight factor is added for SIGMOID restraints to balance the restraint term in the overall energy term of each decoy (Sergey Ovchinnikov, personal communication). Thus, small changes in each of those definitions could have significant effects on the final structure prediction. Unfortunately, it is out of the scope of this study to explore all variations, and thus results aid primarily as guide for future work and AMPLE users. This study highlighted again the benefits of adding BBCONTACTS predictions to existing contact maps to further restrain β-rich regions during *ab initio* protein structure prediction.

Lastly, part of the comparison carried out in this study was aimed specifically at MX experimentalists and, in particular, AMPLE users. Beyond the proof-of-principle study described in Chapter 3, this work further illustrated how important additional restraint information is to increase the chances of unconventional MR success. However, this work also highlighted limitations in the AMPLE routine whereby decoys that were restrained by predicted residue-residue contacts achieved much higher decoy quality compared to unrestrained ROSETTA decoys, yet solved fewer targets. The idea that restrained decoys might benefit from a different kind of processing was further supported by the most successful decoy sets, which were obtained with GREMLIN contact predictions. Given that GREMLIN and ROSETTA decoys achieved similar decoy qualities across multiple targets, their structure solutions were identical for all of

ROSETTA's successful solutions. GREMLIN decoys outperformed ROSETTA decoys solely on the basis that it acquired highly accurate decoys for one further target, and thus achieved the most structure solutions in this study.

Therefore, further work was required to identify the optimal strategy for decoy sets with high structural similarities to the native fold. Such work could focus on the recent idea of selecting decoys based on their long-range contact precision [74, 112] to specifically eliminate the worst decoys, and thus enhance a more fine-grained clustering approach in SPICKER (Chapter 6). Alternatively, truncation could be guided by alternative means, such as the importance of each residue in the predicted contact map. Ultimately, it is key to improve the AMPLE protocol to exploit the much higher decoy quality to enhance the users chance of success.

Chapter 5

Alternative *ab initio* structure prediction algorithms for AMPLE

Acknowledgement: *I would like to thank Dr Saulo de Oliveira for his contributions to this chapter. He kindly provided his time and expertise to generating SAINT2 structure predictions included in the analysis presented in this chapter.*

5.1 Introduction

To-date, the recommended *ab initio* protein structure prediction protocol for optimal AMPLE performance is ROSETTA [114, 115, 117, 118]. This recommendation is based primarily on the superiority of the decoy quality compared to other modelling algorithms, which was recently reaffirmed by the CASP12 experiments [158, 159]. However, Keegan et al. [115] demonstrated that the alternative *ab initio* structure prediction protocol QUARK provides a suitable alternative to ROSETTA in AMPLE. Although inferior in the total number of structure solutions, QUARK decoys are a suitable ROSETTA alternative in most cases [115]. In particular, given ROSETTA's challenging installation procedure, availability limited to POSIX operating systems, requirement for large disk space and computationally expensive algorithm, QUARK's online server has been a very attractive alternative for AMPLE users.

Whilst ROSETTA and QUARK are amongst the best *ab initio* structure prediction algorithms currently available [158], other algorithms have been developed over the last two decades [e.g., 43, 69, 154, 160–162]. Although most of these algorithms utilise fragment-assembly algorithms similar to ROSETTA and QUARK, their procedures for fragment selection or assembly is substantially different [160, 161]. Furthermore, predicted contact information has recently seen a spike in accuracy. This invaluable source of information is introduced differently in each protocol, and thus might have profound effects on the resulting decoy quality. In particular, physics-based algorithms relying largely on this information are an interesting alternative to fragment-based approaches [69, 76, 154, 162].

CONFOLD2 [132], a distance-geometry based algorithm, uses predicted secondary structure and contact information to rapidly generate *ab initio* decoys. Unlike other algorithms, CONFOLD2's algorithm is driven almost entirely by the contact information to explore the fold space. Different contact selection thresholds are used to not limit the search space to a fixed, predefined selection. CONFOLD2 generates slightly less accurate decoys compared to ROSETTA, however outperforms it in speed and simplicity of installation [73, 132].

FRAGFOLD [160], a fragment-assembly based algorithm, generates decoys in a similar fashion to ROSETTA and QUARK. However, FRAGFOLD does not rely on large structural libraries for fragment extraction. Instead, it provides a relatively small library of supersecondary structural fragments and short length fragments, which were extracted from high resolution protein structures. Since the generalised fragment lib-

rary is shipped with FRAGFOLD, and target-specific fragments are extracted based on secondary structure and a sequence-based threading score, fragment library generation is fast and easy compared to ROSETTA [45].

SAINT2 [46], a further fragment-assembly based algorithm is substantially different to most others. SAINT2 attempts *ab initio* structure prediction sequentially, starting from either terminus of the target sequence [46]. Furthermore, SAINT2 uses FLIB [53] for fragment picking, an algorithm shown to outperform ROSETTA's equivalent NNMAKE [54] in precision with very similar coverage.

Since some of these algorithms are readily available and often easier to install without the overhead of large databases for fragment picking, the work in this study focused on exploring three alternative *ab initio* structure prediction algorithms and their value in unconventional MR. The *ab initio* structure prediction protocols CONFOLD2 [132], FRAGFOLD [160] and SAINT2 [46], were benchmarked in AMPLE given their substantially different approaches to AMPLE's current default ROSETTA [42].

5.2 Materials & Methods

5.2.1 Target selection

This study was conducted using all 27 targets from the PREDICTORS dataset (Section 4.2.1 and Table A.2).

5.2.2 Contact prediction

Residue-residue contact information was predicted for 18 out of 27 targets using METAPSICOV v1.04 [101]. Nine targets were left deliberately without contact prediction to trial the performance of each algorithm under such circumstances.

Secondary structure and solvent exposure were predicted using PSIPRED v4.0 [147] and SOLVPRED (shipped with METAPSICOV v1.04), respectively. The MSA for coevolution-based contact prediction was generated using HHBLITS v2.0.16 [137] against the `uniprot20` v2016-02 database. CCMPRED v0.3.2 [92], FREECONTACT v1.0.21 [149] and PSICOV v2.1b3 [89] were used by METAPSICOV to generate contact predictions.

METAPSICOV STAGE1 contact predictions were used in *ab initio* structure prediction since they result in more accurate *ab initio* protein structure predictions compared to METAPSICOV STAGE2 predictions [101].

5.2.3 *Ab initio* structure prediction

The ROSETTA 3- and 9-residue fragment libraries for each target were generated using the ROBETTA online server (`http://robetta.bakerlab.org/`). The option to "Exclude Homologues" was selected to avoid inclusion of homologous fragments. Each target sequence and its fragments were subjected to ROSETTA v2015.22.57859 [42] and 1,000 decoys per target generated with AMPLE v1.2.0 ROSETTA default options. Top-L (where L corresponds to the number of residues in the target chain) contact pairs were used in combination with the *FADE* ROSETTA energy function. For further details see Section 3.2.3 or Michel et al. [70].

The FRAGFOLD decoys were generated using FRAGFOLD v4.80 [160] with default options. Homologous fragments were removed from the shipped library by excluding all entries with PDB identifiers identical to those retrieved from the ROBETTA server. All contact pairs were used according to FRAGFOLD's internal protocol.

The fragment libraries for SAINT2 were generated using FLIB [53], which picks on average 30 fragments per target position. These fragments are typically six to 20 residues in length. Homologous fragments were removed from the final fragment list using the PDB identifiers obtained from the ROBETTA online server. The secondary structure prediction and solvent accessibility scores were identical to those obtained from the ROBETTA server. SAINT2 was used for decoy generation, and 1,000 decoys generated per target. The procedure and parameters were identical to those described in Supplementary Information (p. 16) by Oliveira et al. [46].

The CONFOLD2 decoys were generated using CONFOLD2 v2.0 [132], which uses Crystallography & NMR System (CNS) v1.3 [163] to drive the modelling. Default parameters were used except for the number of decoys per run, which was increased from 20 to 25 using the `-mcount` parameter. CONFOLD2 varies the number of contacts included in each separate modelling run, ranging from $L/10$ to $4L$ with increments of $L/10$. Thus, the CONFOLD2 protocol yields a total of 40 separate modelling runs generating 25 decoys each. Structure predictions for only 18 targets were done since nine targets were benchmarked without any contact predictions, which are an essential input in CONFOLD2.

5.2.4 Molecular Replacement

All decoy sets were subjected to AMPLE v1.2.0 and CCP4 v7.0.28. Default options were chosen with the following exceptions: decoys in all 10 clusters were used, subcluster radii thresholds were set to 1 and 3Å, and side-chain treatments were set to `polyala` only. This change in protocol from AMPLE's initial mode of operation [114] was shown to be advantageous in most cases by Thomas [144], and thus trialled in this context. Each MR run was assessed using the SHELXE criteria, where a minimum CC of 25.0

and ACL of 10 was required (Section 2.3.4.2). R-values of less than 0.45 after model building were not part of the success criteria in this study.

5.3 Results

The purpose of this study was to investigate the usefulness of alternative *ab initio* structure prediction algorithms in AMPLE. Three promising leads widely used in the *ab initio* modelling experiments were examined and compared against AMPLE's current algorithm of choice. This led to a direct comparison of the algorithms ROSETTA [42], CONFOLD2 [132], FRAGFOLD [160] and SAINT2 [46]. All four algorithms have recently seen great improvements through the use of residue-residue contact information, which was predicted for two-thirds of the targets using the METAPSICOV [101] algorithm.

5.3.1 Alignment depth and contact prediction precision

The first step in this study was the prediction of residue-residue contacts using the metapredictor METAPSICOV for 18 targets in the PREDICTORS dataset [101]. Since we attempted to test each of the structure prediction boundaries in extreme cases, a variety of targets with different alignment depths were chosen. The alignment depth — i.e., the number of effective sequences — of METAPSICOV-generated HHBLITS alignments ranged from 431 to 6,186 across all targets (Fig. 5.1). Six targets contained at least 200 and less than 1000 sufficiently-diverse sequences, whilst the remaining 16 targets contained more than 1,000 effective sequences.

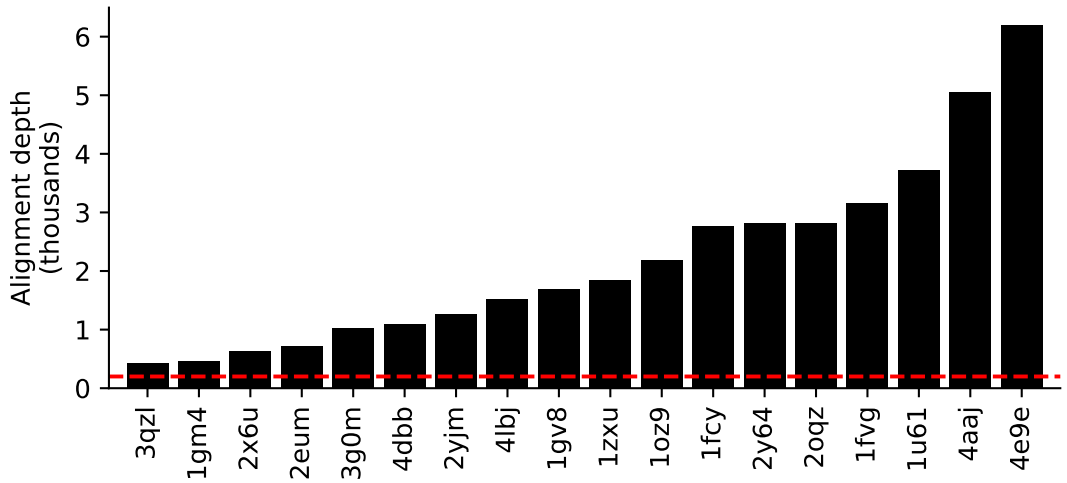


Figure 5.1: Distribution of HHBLITS alignment depth for subset of targets in the PREDICTORS dataset. Red line indicates the suggested alignment depth requirement for accurate coevolution-based contact prediction [84].

Coevolution-based contact predictors rely heavily on the alignment depth for accurate contact predictions. In this work, these findings were further confirmed. Sequence alignments with depths of less than 1,000 sequences produced contact predictions with lower precision scores across a number of cutoffs compared to those with deeper alignments (Fig. 5.2). Given the alignment depths and top-L contact predictions, a positive correlation between the two was found (Spearman's $\rho = 0.57$, p-value < 0.02). A moving average analysis showed that those contact predictions based on alignments with more than 1,000 effective sequences yielded better precision scores by at least 0.09 units up to 0.34. The difference between the two moving average curves in Fig. 5.2 highlights that the difference was greater at lower cutoff values, i.e. only the very best contacts were included in the selection. This difference declined more drastically for targets with deeper alignments (Fig. 5.2).

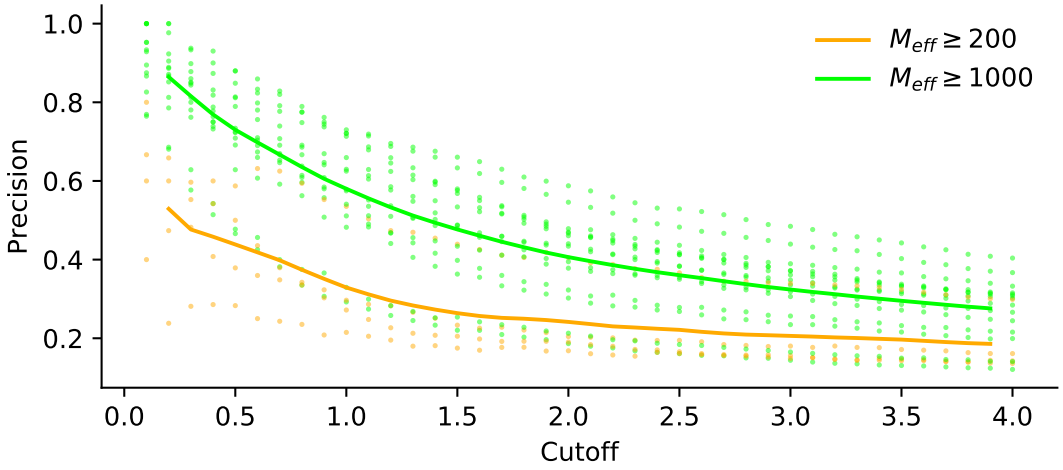


Figure 5.2: Contact precision analysis for numerous contact selection cutoffs for targets with alignment depths of more than 200 and more than 1,000 sequences. Lines indicate moving averages for both categories with a window size of three residues. M_{eff} refers to the alignment depth.

5.3.2 Comparison of decoy quality

One main interest of the work presented in this chapter was a direct comparison of the quality of decoys predicted with four *ab initio* structure prediction algorithms. At the time of writing, no such comparison existed on the same dataset, and thus might provide direct insights into the performance of each.

An initial comparison of overall performance highlighted that ROSETTA generated the highest quality decoys (Fig. 5.3). Across all modelling algorithms the distribution of TM-score values is right-skewed, which indicates a higher proportion of non-native-like folds within the sets. A TM-score quantile evaluation of each decoy set by algorithm showed that ROSETTA and CONFOLD2 contained only a single set with a lower quantile of less than 0.2 TM-score units. In comparison, FRAGFOLD predicted three

and SAINT2 eight decoy sets with a lower quantile of less than the aforementioned threshold. In comparison, ROSETTA, CONFOLD2 and FRAGFOLD predicted six, seven and five decoy sets with upper quantiles greater than 0.5 TM-score units, whilst SAINT2 predicted zero.

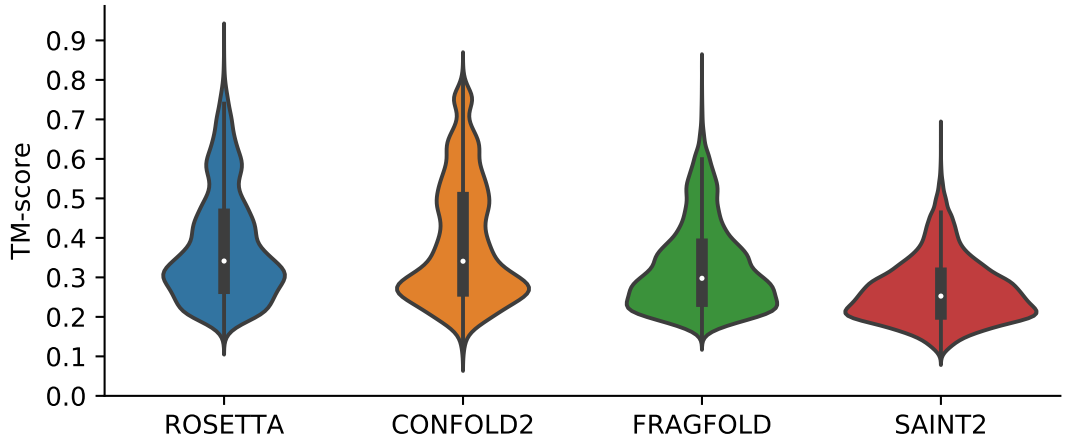


Figure 5.3: KDE of decoy TM-score for four different *ab initio* structure prediction algorithms, namely ROSETTA, CONFOLD2, FRAGFOLD and SAINT2. CONFOLD2 contains 9,000 less decoys than the remaining algorithms (for further details refer to Section 5.2.3).

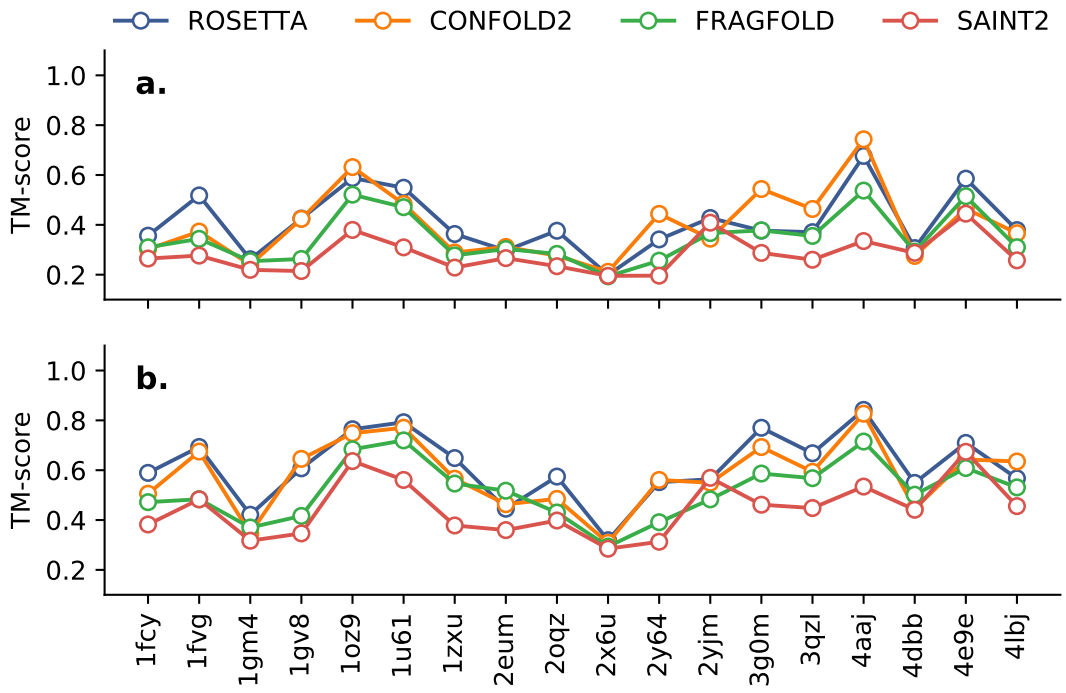


Figure 5.4: Per-target TM-score analysis for targets modelled with contact information and four separate *ab initio* structure prediction algorithms. Analysis is subdivided by (a) median TM-score of all decoys in each set and (b) TM-score of the top-1 decoy in each set.

A direct comparison of the methods by median TM-score of each contact-assisted decoy set reaffirmed ROSETTA's performance in predicting *ab initio* decoys accurately. Across 18 targets, ROSETTA decoy sets contained the best median TM-score for 11 targets (CONFOLD2 for remaining seven targets). This was further strengthened when comparing the top-1 decoy for which ROSETTA predicted the best in 13 cases (CONFOLD2 in three cases, FRAGFOLD and SAINT2 in one) (Fig. 5.4).

Abriata et al. [158] recently attributed the success in the CASP12 experiment to the improved precision of coevolution-based contact predictions and the availability of many more sequence homologs. Thus, it was of great interest to explore the structure prediction algorithms in this study with regards to their dependence on the availability of sequence homologs and precise contact predictions.

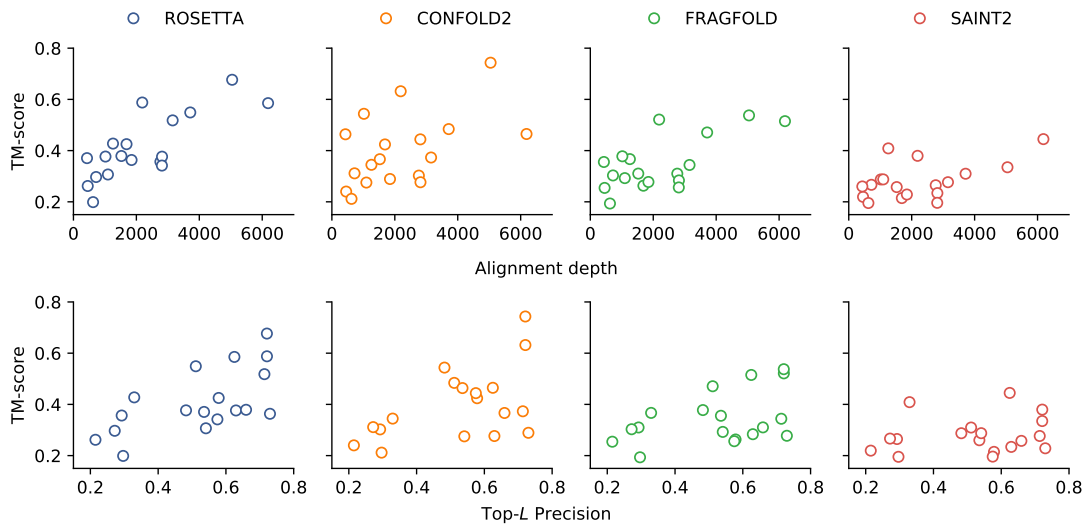


Figure 5.5: Analysis of median TM-score of the contact-based decoy sets and their dependence on alignment depth and top-L precision.

The results obtained in this study further supported the conclusions made by Abriata et al. [158] but only for the ROSETTA algorithm. A Spearman's rank-order CC analysis of alignment depth and median TM-score showed a significant positive correlation for ROSETTA-generated decoy sets (Spearman's $\rho = 0.68$, $p < 0.01$). This positive correlation was also found for ROSETTA-generated decoy sets with regards to their top-L precision and median TM-score (Spearman's $\rho = 0.61$, $p < 0.01$). All other modelling algorithms did not show a significant correlation, although better decoy sets were generally obtained with greater alignment depths and more precise top-L contacts (Fig. 5.5). Furthermore, the sample size for each correlation analysis was small ($n = 18$), and thus further test cases are required for a more confident inference.

Parts of this study also explored the performance of ROSETTA, FRAGFOLD and SAINT2 when no contact prediction was provided as distance restraint information in *ab initio* structure prediction (CONFOLD2 requires contact information, and thus

was excluded). ROSETTA performed best for seven of the nine contact-free decoy sets based on median TM-score of the entire decoy set and the TM-score of the top-1 decoy (Fig. 5.6). However, the difference was marginal for the majority of cases. The median values for eight ROSETTA and FRAGFOLD decoy sets differed by less than 0.10 TM-score units (seven ROSETTA and SAINT2 sets by less than 0.10 units). Furthermore, the top-1 decoys for only three targets differed greatly between the modelling algorithms, whilst the rest was near identical (Fig. 5.6).

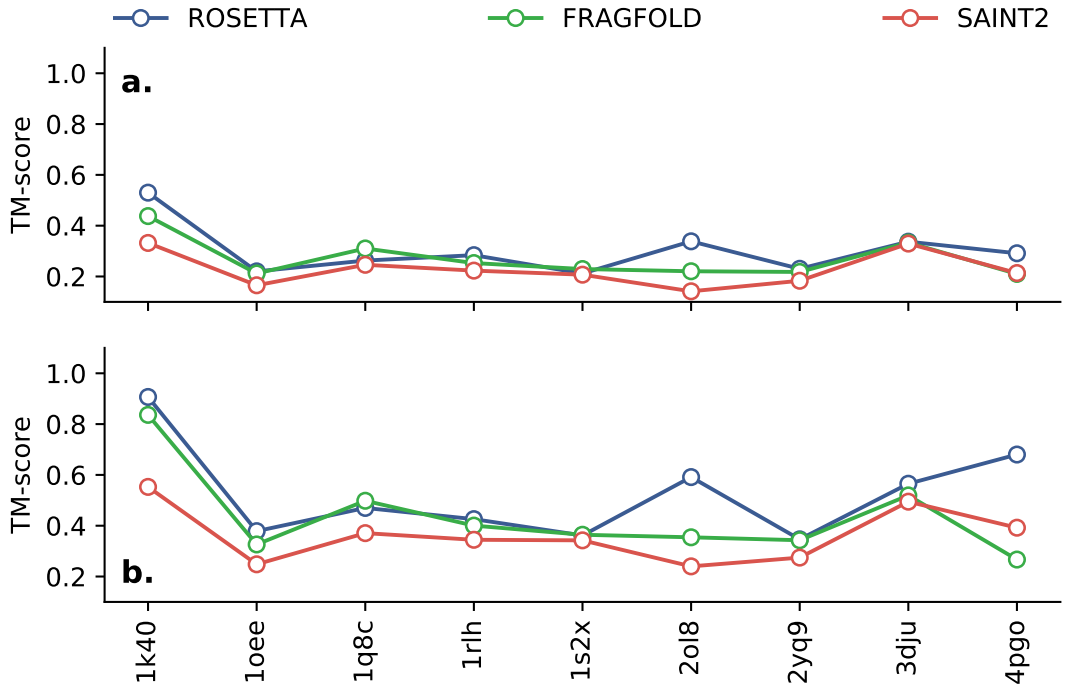


Figure 5.6: Per-target TM-score analysis for targets modelled without contact information and four separate *ab initio* structure prediction algorithms. Analysis is subdivided by (a) median TM-score of all decoys in each set and (b) TM-score of the top-1 decoy in each set.

The top decoy predicted by ROSETTA and SAINT2 based on the sequence of the FAT domain of focal adhesion kinase (PDB ID: 1k40) differed by 0.35 TM-score units. More significantly though, the top-1 decoy predicted by ROSETTA for the outer surface protein A (PDB ID: 2ol8) is considered native-like (TM-score = 0.59), whilst the FRAGFOLD (TM-score = 0.35) and SAINT2 (TM-score = 0.24) counterparts predicted incorrect folds. A near-identical scenario applies to the top-1 decoys of the Hypothetical protein PF0907 (PDB ID: 4pgo) (ROSETTA TM-score = 0.68; FRAGFOLD TM-score = 0.27; SAINT2 TM-score = 0.39).

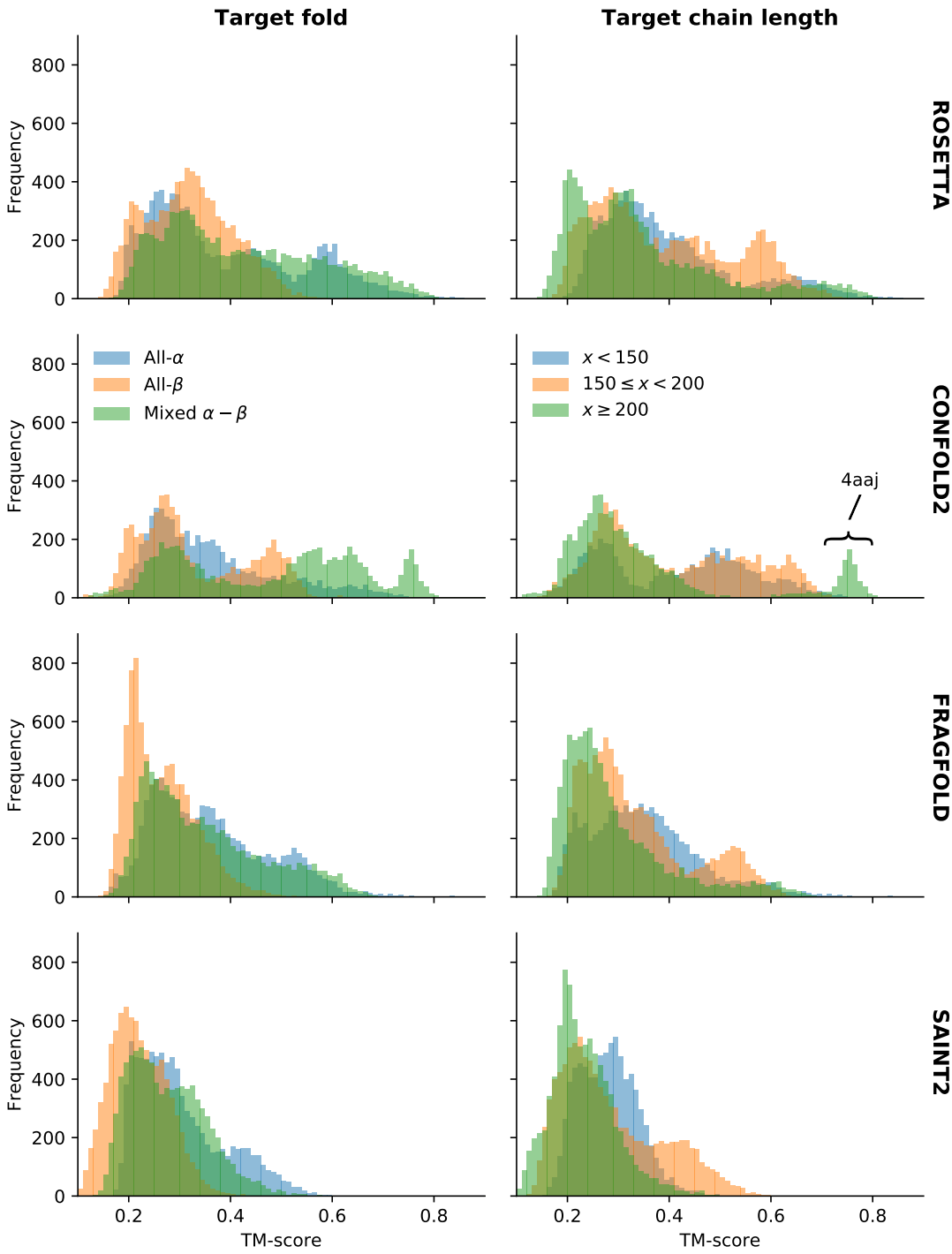


Figure 5.7: Decoy TM-scores by fold, chain length and algorithm.

An analysis of the modelling results by target fold showed that all-α and mixed α-β target folds were less challenging to predict than all-β targets (Fig. 5.7). The multimodal distributions of all-α and mixed α-β target decoys predicted by ROSETTA spans from 0.10 TM-score units to 0.80. In comparison, the approximately normal distribution for all-β targets by the same algorithm centres at 0.32 TM-score units (s.d.=0.08 TM-score units). Similarly, FRAGFOLD decoys showed a more-spread distribution of decoys for all-α and mixed α-β decoys compared to all-β. The TM-score distributions for

CONFOLD2 mixed α-β and all-β decoys follow multimodal distributions. Whilst this might indicate that CONFOLD2 either predicted the overall target fold correctly or incorrectly, the data might mislead because of the missing targets in the dataset. Lastly, the distributions of TM-score values for either fold class of SAINT2 decoys in Fig. 5.7 appear more similar than the others indicating less difference between the fold classes. However, similarly to the ROSETTA decoys the all-β distribution appears normal whilst the other two are right-skewed highlighting some more accurate decoys in the overall set (Fig. 5.7).

A further subdivision of all target decoys by target chain length was done. At the stage of target selection, three main bins were defined from which targets were randomly sampled (Section 4.2.1). These bins were defined with target chain length edges of 150 and 200 creating three bins: $0 < n < 150$ & $150 \leq n < 200$ & $n \geq 200$ (n refers to the target chain length). A grouping of the decoy TM-score by algorithm and target chain length indicated little difference in modelling difficulty (Fig. 5.7). The distributions in Fig. 5.7 show the largest spread across all modelling algorithms for chain lengths in the bin $150 \leq n < 200$. Surprisingly, only FRAGFOLD and SAINT2 performed better for targets in the smallest bin size whilst CONFOLD2 found those targets most challenging. CONFOLD2 also generated the best decoys for one of the largest targets in the dataset (n_{res}=216). The set of CONFOLD2 decoys for N-(5-phosphoribosyl)anthranilate isomerase (PDB ID: 4aaj) had a median TM-score of 0.74. ROSETTA decoys showed a comparable median TM-score of 0.68; however, FRAGFOLD (median TM-score=0.54) and SAINT2 (median TM-score=0.33) were unable to generate decoys of similarly high quality.

5.3.3 Molecular Replacement

The final step in this study was to explore the benefits or drawbacks of each *ab initio* structure prediction algorithm for MR.

Each *ab initio* modelling-algorithm generated at least two decoy sets sufficient for MR structure solution (Fig. 5.8). ROSETTA and SAINT2 decoy sets led to the solutions of five targets each, whilst FRAGFOLD decoys solved four and CONFOLD2 decoys just two. All four algorithms predicted decoys of good enough quality to solve the structures of the Hypothetical protein AQ_1354 (PDB ID: 1oz9) and Putative Ribonuclease III (PDB ID: 1u61), although SAINT2-based AMPLE search models yielded the highest ratio of successful search models compared to the total trialled in both cases (Fig. 5.8). Besides these two targets, little consensus exists amongst the targets for which structure solutions were obtained across the different modelling algorithms.

The chain length for targets with structure solutions ranged from 106 (PDB ID: 4pgo) to 236 (PDB ID: 1fcy) residues. Although statistics cannot reliable indicate the performance with such a small sample size, SAINT2 decoys solved on aver-

age the largest targets (mean target chain length ROSETTA=147, CONFOLD2=144, FRAGFOLD=136, SAINT2=162). The ROSETTA, FRAGFOLD and SAINT2 decoys achieved structure solutions for all three fold classifications, whilst CONFOLD2 decoys did not yield a structure solution for any all-β target. Nevertheless, successful AMPLE ensemble search models for all-β targets derived from the former three algorithms were scarce with only a single one leading to structure solution (Fig. 5.8).

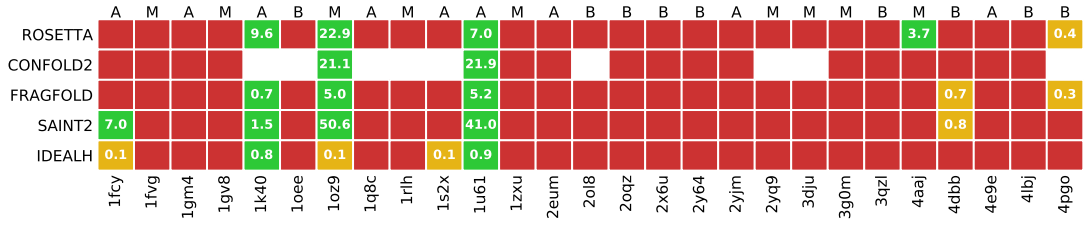


Figure 5.8: Summary of MR success with AMPLE ensemble search models. Search models are based on decoy sets generated with different *ab initio* structure prediction protocols. The colour coding indicates structure solution: no solution (red), one solution (orange), more than one solution (green). The number in cells with at least one solution states the percentage of successful search models. The one-letter codes above each column indicate the target fold: all-α (A); all-β (B); mixed α-β (M). The row labelled "IDEALH" refers to AMPLE's ideal helix run.

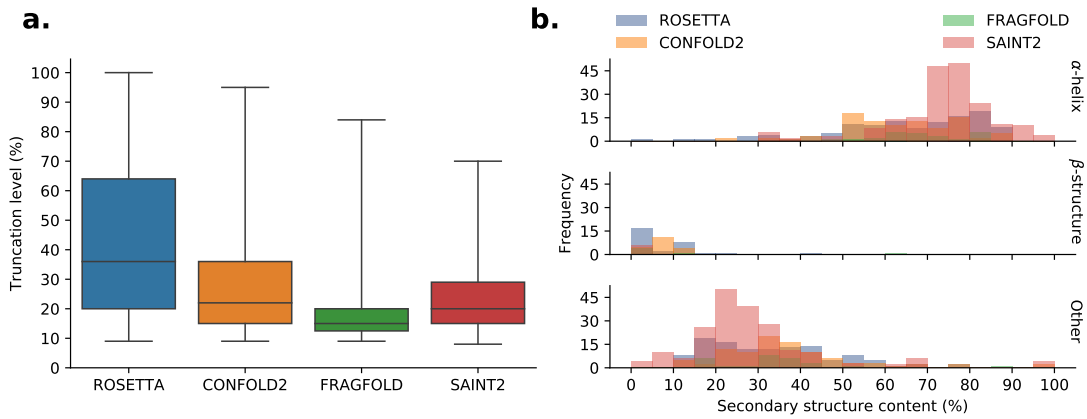


Figure 5.9: Distribution of (a) search model truncation and (b) secondary structure content for successful AMPLE ensemble search models given decoys from four *ab initio* structure prediction algorithms. Secondary structure for each ensemble search model evaluated with DSSP [157].

The difference in overall decoy quality between the four different *ab initio* structure prediction algorithms was further noticed in the successful AMPLE-generated ensemble search models. ROSETTA decoys resulted in more complete AMPLE ensemble search models, which led to structure solution (Fig. 5.9). Although CONFOLD2 had a similar maximum of just under 100% completeness, 75% of all successful search models contained at most 40% of the target sequence. Overall, FRAGFOLD decoys translated into the least complete successful AMPLE search models with 75% containing less than 20% of the target sequence. SAINT2 had the shortest range spanning from 8% to 70%

target completeness.

An inspection of the secondary structure content of all successful ensemble search models outlined an important difference between SAINT2 and the other three modelling algorithms. Successful search models derived from SAINT2 decoys were predominantly α-helical (Fig. 5.9). An analysis of the secondary structure makeup, as assigned by DSSP [157], showed that successful SAINT2 search models contained approximately 70-80% α-helices with the rest being unassigned secondary structure. In comparison, the successful ensemble search models from other modelling algorithms contained a range from 50-90% α-helices, whilst the remainder was either unstructured or β-structure (Fig. 5.9).

This important observation is crucial in assessing the structure solutions obtained since simple helices could be derived from idealised α-helix libraries, and thus save the great overhead of predicting, preparing and sampling decoys in AMPLE. A visual inspection of SAINT2 ensemble search models highlighted that the FAT domain of focal adhesion kinase (PDB ID: 1k40) and the amyloid-β A4 precursor protein-binding family A1 (PDB ID: 4dbb) were solved with single α-helices (Fig. 5.10). Trialling the experimental data of these targets against AMPLE's ideal helix library [117] showed that the former could have been solved without the modelling overhead (Fig. 5.8). In fact, SAINT2 decoys did not result in any additional structure solutions compared to AMPLE's ideal helix library except for the solution of the A4 precursor protein-binding family A1 (PDB ID: 4dbb) (Fig. 5.8). In comparison, the other modelling algorithms resulted in similar idealised fragments, especially in borderline cases (Fig. 5.10). However, these fragments were not strictly α-helical, and thus would require more sophisticated and computationally complex idealised-fragment library generation protocols [e.g., 164] or libraries of recurring tertiary structure motifs [e.g., 156]. Nevertheless, even the most sensitive MR ideal-fragment-selection algorithms could almost certainly not identify a search model of similar quality to that derived from ROSETTA decoys for the Hypothetical protein PF0907 (PDB ID: 4pgo) (Fig. 5.10), which might be essential in structure solution determination of some targets.

Whilst all *ab initio* structure prediction algorithms enabled structure solutions of at least two targets, the relationship between the quality of the starting decoys and MR structure solution success needed to be evaluated. ROSETTA and CONFOLD2 generated the highest quality decoys, followed by FRAGFOLD and SAINT2 (Fig. 5.3). Thus, most structure solutions would have been expected for the former two since more native-like decoys are generally considered better search models. However, an analysis of the RMSD of each ensemble search model's centroid showed that decoy quality may not always be the most reliable indicator. Although search models are often considered suitable once their RMSD to the native structure is better than 1.5Å [165], this threshold did not strictly apply to *ab initio* modelling-based AMPLE ensemble search models (Fig. 5.11). For example, a small number of SAINT2-derived

search models, which were prepared for the FAT domain of focal adhesion kinase (PDB ID: 1k40), exceeded this threshold greatly with RMSD values > 10Å (up to 28Å) yet resulted in PHASER LLG values in excess of the success threshold of 60 [166]. Additionally, nearly 25% of all successful ensemble search models had RMSD values ≥ 2Å and PHASER LLG scores of ≥60. Although striking at first, structure solutions in these situations were often achieved by out-of-sequence-register placement of search models. An analysis of the RIO score metric showed that the usefully placed parts of all but one AMPLE search model with RMSD value greater than 10Å was out-of-register. Furthermore, it is important to remember that RMSD values greatly differ based on the optimal superposition of the model and target.

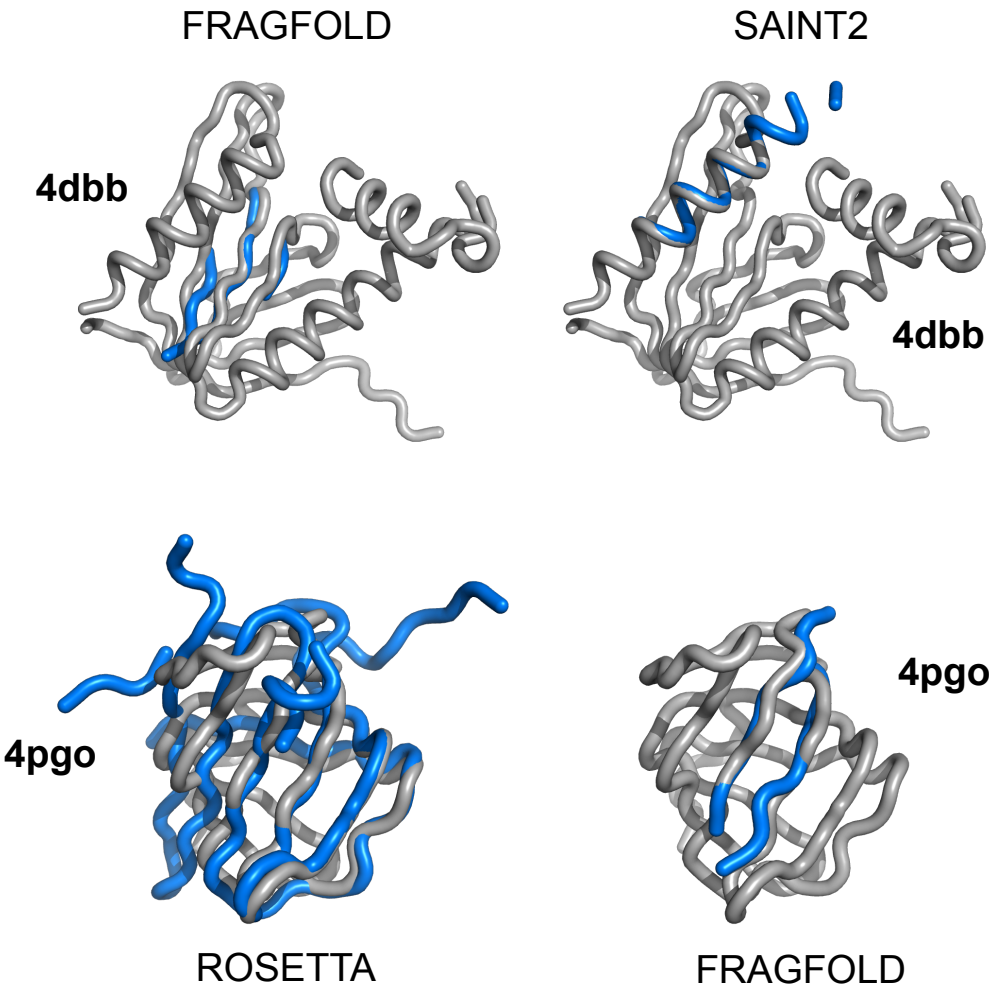


Figure 5.10: Examples of PHASER-placed AMPLE search models that led to structure solution. AMPLE search models are coloured blue and deposited native structures in grey. The PDB identifiers and modelling protocol are provided alongside each example.

Lastly, one characteristic of a good MR search model is good stereochemical geometry of its peptide-chain backbone, especially during refinement. Fragment-based structure prediction algorithms typically contained good stereochemistry, because the

template fragments are derived from refined protein structures. In comparison, CONFOLD2, which does not use fragments, relies on physics-based energy functions to identify good stereochemistry of the decoy backbone. Thus, it is important to understand if poor stereochemistry was present in CONFOLD2 ensemble search models, such that it might explain why good decoy quality did not translate to more MR structure solutions.

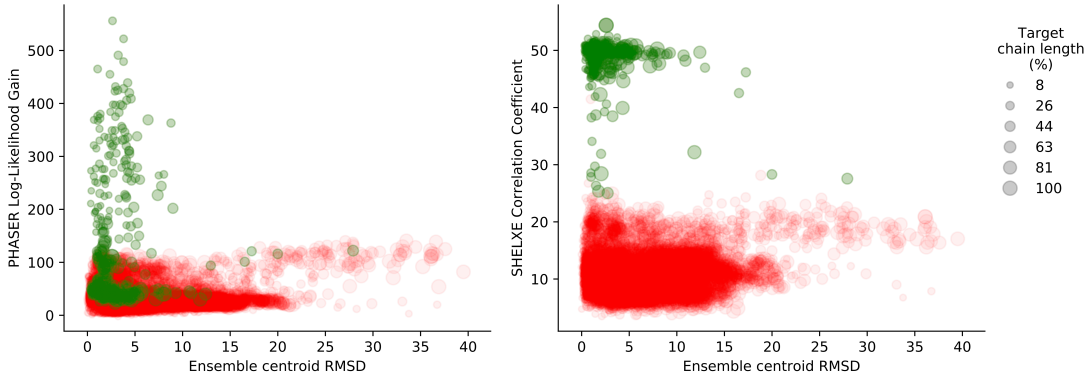


Figure 5.11: Relationship between ensemble quality, PHASER LLG and SHELXE CC. Data points are coloured based on the outcome of their MR trials: green indicate structure solution, red indicate no structure solution.

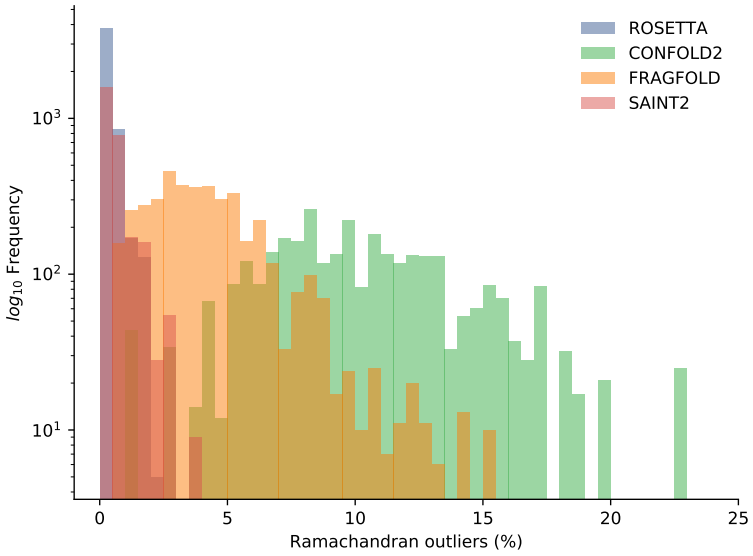


Figure 5.12: Distribution of Ramachandran outliers of AMPLE ensemble search model centroids based on decoys predicted with four *ab initio* structure prediction protocols. Outliers were calculated using PyRAMA (https://github.com/gerdos/PyRAMA).

Indeed, a Ramachandran analysis of φ and ψ peptide backbone angles outlined much poorer stereochemistry of ensemble search model centroids for CONFOLD2 compared to all fragment-assembly-based structure prediction algorithms (Fig. 5.12). ROSETTA search models, which were made up of crudely-refined decoys, possessed at most 2% Ramachandran outliers. SAINT2, which generated less accurate structure predictions

compared to other protocols, showed the second best stereochemistry of centroid models without any refinement. FRAGFOLD contained around 5% outliers for the majority of search models. In comparison to these statistics CONFOLD2 contained around 5-15% Ramachandran outliers in centroid decoys.

Further analysis of the centroids of each truncated AMPLE ensemble demonstrated the importance of good stereochemistry for success. Out of 94 successful ensemble search models only 17 contained outliers. This contrasts to over 390 unsuccessful ensemble search models with TM-scores greater than 0.5 units and on average 6% outliers ($\min_{\text{outliers}}$=1%; $\max_{\text{outliers}}$=23%). Therefore, perfect peptide backbone stereochemistry is still no guarantee for MR structure solution.

5.4 Discussion

In this chapter, work was conducted to explore *ab initio* protein structure prediction protocols as alternatives to ROSETTA and QUARK. Three algorithms — CONFOLD2, FRAGFOLD and SAINT2 — were trialled on a set of 27 globular targets to evaluate their performance with regards to structure prediction and subsequent MR trials.

The experiments in this study highlighted that ROSETTA remains the most accurate structure prediction protocol amongst the trialled ones. ROSETTA outperformed the other three algorithms across the majority of protein targets for entire decoy sets and the best decoy in each set. These findings were further confirmed in the latest CASP12 experiments, which outlined ROSETTA's success compared to other protocols [158, 159]. Furthermore, the findings describing the comparable performance of ROSETTA and CONFOLD2 [73, 132] are supported in this work. Given that the latter relies entirely on the predicted contact information, such performance emphasises the quality and importance of contact prediction in *ab initio* protein structure modelling. It is also to be expected that the increase in sequence availability will improve the decoy quality further [158, 167]. In this study, the alternative fragment-assembly based algorithms FRAGFOLD and SAINT2 were tested. Although both did predict native-like decoys for some targets, their performance was overall much worse than ROSETTA and CONFOLD2. SAINT2 did not generate decoys of native-like quality in cases where all other algorithms did. Beyond overall decoy quality, previous findings suggested a difference in difficulty based on the target fold. These findings are further manifested here. All algorithms predicted most native-like decoys for all-α and mixed α-β targets. Although previous studies also reported on greater difficulty for larger targets — especially in cases without contact prediction — such findings could not be confirmed here.

Given that the application of these decoys is primarily aimed at challenging targets in MR, the quality of decoys was not necessarily enough to predict the success of

AMPLE-generated search models. The results in this chapter clearly demonstrated that highly accurate decoys predicted by CONFOLD2 do not routinely translate into MR structure solutions. Despite a recent example of the successful application of CNS-generated decoys in MR [168], further research is required to identify the main bottleneck observed in this study. ROSETTA, FRAGFOLD and SAINT2 achieved structure solutions for a number of targets, despite poor decoy quality in cases of the latter two. CONFOLD2 decoys appear to suffer from poor stereochemistry, and results suggest that decoy refinement might be essential to exploit the underlying decoy quality [169].

In conclusion, ROSETTA remains the best modelling algorithm for unconventional MR in AMPLE. Although some of this success must be due to the fact that AMPLE's algorithm is tailored towards exploiting the cluster variance derived from ROSETTA decoys, it cannot be downplayed that ROSETTA generates the most accurate decoys overall. However, it is crucial to investigate whether CONFOLD2 decoys, potentially remodelled to improve the backbone stereochemistry, might provide a suitable routine alternative to ROSETTA, especially because fragment databases are not required and modelling time per decoy is reduced by approximately a factor of four.

Chapter 6

Decoy subselection to enhance MR search model creation

6.1 Introduction

Work presented in Chapters 3 to 5 highlighted the much improved *ab initio* decoy quality achievable by restraining the conformational search space with residue-residue contact predictions. Furthermore, the data also highlighted that this improvement extends AMPLE's performance of achieving structure solution for more challenging targets. However, the data in Chapters 4 and 5 also indicated that AMPLE's protocol is currently not tailored towards decoy sets with much improved accuracy. In some cases, decoy sets with correctly predicted folds — whereby the mean TM-score of the decoy set was greater than 0.5 score units — did not generate any successful ensemble search models.

Furthermore, *ab initio* decoy similarity to the crystal structure was exceptionally high in some cases (RMSD < 1.5Å). Although challenging by current means to identify these decoys, it is of great interest to structural biologists to do so since these decoys might be sufficient by themselves as MR search models. A contact prediction, which is typically used to restrain the folding protocol, might provide enough information to drive such identification. Indeed, Kosciolek and Jones [45] and Oliveira et al. [112] highlighted the usefulness of long-range residue-residue contact pair satisfaction for model selection since it correlates well with decoy quality. Additionally, Adhikari and Cheng [132] use long-range contact satisfaction routinely in CONFOLD2 to exclude the worst decoys amongst the set predicted ones.

Thus, this chapter focused on exploring alternative strategies of decoy selection in AMPLE. In particular, work presented here focused on exploiting long-range contact information to drive search model generation to extend AMPLE's performance on difficult cases further.

6.2 Materials & Methods

6.2.1 Target selection

The dataset for this study consisted of 113 ROSETTA decoy sets generated throughout the work outlined in Chapters 3 to 5. The 113 decoy sets covered all targets in the ORIGINAL (Table A.1), PREDICTORS (Table A.2) and TRANSMEMBRANE (Table A.3) datasets. Top-L (> 5 residues sequence separation) CCMPRED [92], PCONSC2 [100], METAPSICOV STAGE1 [101] and MEMBRAIN [150] contact pairs were used in combination with the *FADE* energy function to restrain the *ab initio* structure prediction process.

6.2.2 Computation of range-specific satisfaction scores

The satisfaction of short- (> 5 residues sequence separation), medium- (> 12 residues sequence separation) and long-range contact pairs (> 23 residues sequence separation) were computed for each decoy in each set (for further details, see Section 2.3.2.3). Hereby, the short-, medium- or long-range predicted contacts were extracted from the original predictions used to restrain the *ab initio* structure prediction, matched against the contact pairs observed in individual decoys and the range-specific contact satisfaction score evaluated.

6.2.3 Decoy subselection

Each set of decoys was ranked in descending order by their long-range contact pair satisfaction scores and the n decoys with the lowest scores removed from each set. The number of decoys to remove n was selected using a number of different strategies:

NONE	leave the original set unchanged
LINEAR	remove the worst 500 decoys
CUTOFF	remove all decoys with a score of < 0.287
SCALED	remove all decoys with a scaled score of < 0.5, where the scaled score is score divided by set average
INDIVIDUAL	keep the top-5 decoys only

The fixed definition in the *CUTOFF* strategy was determined by Oliveira et al. [112]. The scaled score used by the *SCALED* strategy was computed by dividing each decoy's long-range contact pair satisfaction by the set's average.

The *INDIVIDUAL* subselection strategy differed substantially from the others. The top-5 decoys by long-range contact satisfaction were selected and subjected to treatment outside of AMPLE. The per-decoy treatments were the following:

default	leave the decoy unchanged
domain	remove all residues with $kde < \frac{1}{2}max_{kde}$, where kde corresponds to the KDE and max_{kde} to the maximum KDE obtained by applying the algorithm described by Sadowski [170] to the top-$5L$ contact map
dssp	remove all residues with secondary structure of "helix turn (T)", "bend (S)" or "coil (C)", which were assigned using DSSP [157]

fragment remove all residues that do not satisfy the following condition: extract all contacts from a decoy [Cβ distance of < 8Å (Cα in case of Gly)] and reconstruct the decoy's sequence using the residue indices present in the set of contacts, then keep residues that are within a sequence fragment of at least three consecutive residues

variance remove all residues with variance of more than 5Å^2, which was extracted from the decoy's corresponding cluster in the *NONE* subselection strategy

6.2.4 Molecular Replacement

To evaluate the benefits of such subselection to MR in AMPLE, a subset of 35 decoy sets (spanning 35 unique targets) were processed as described in Section 6.2.3 and subjected to AMPLE v1.2.0 and CCP4 v7.0.28. Default options were chosen with the following exceptions: decoys in all 10 clusters were used, subcluster radii thresholds were set to 1 and 3Å, and side-chain treatments were set to `polyala` only. This change in protocol from AMPLE's initial mode of operation [114] was shown to be advantageous in most cases by Thomas [144], and thus trialled in this context.

To allow comparability of these results to previous AMPLE runs, an additional condition was added, namely *NONE_classic*. The decoy set from the *NONE* strategy was thereby subjected to the AMPLE protocol with default settings except `-num_clusters`, which was set to sample the three largest clusters. Thus, the *NONE_classic* strategy differed from the *NONE* one in three aspects: top-3 clusters are used instead of top-10, 1, 2 and 3Å subclustering radii are used instead of 1 and 3Å only, and the most-reliable and all-atom side-chain treatments are kept.

All individual decoys created under the *INDIVIDUAL* strategy were subjected as poly-alanine decoys to MRBUMP v0.9 [127] with identical settings to those used in AMPLE.

Each MR run was assessed using the criteria defined in Section 2.3.4.2.

6.3 Results

This chapter focused on identifying further uses of predicted residue-residue contact pairs in unconventional MR. In particular, the exclusion of *ab initio* decoys by their contact satisfaction scores was investigated. A total of 113 decoy datasets were used to identify potential means of identifying the best or worst decoys. Furthermore, three strategies were trialled alongside two standard approaches to test the consequences of

excluding the worst decoys in ensemble search model preparation in AMPLE.

6.3.1 Contact pair satisfaction correlates with decoy quality

Kosciolek and Jones [45] previously identified a correlation between the TM-score of a decoy and its fraction of satisfied contact pairs. Although reporting striking positive correlations (short-range:$\rho = 0.50$; medium-range:$\rho = 0.57$; long-range: $\rho = 0.87$) for top-1 decoys, the study by Kosciolek and Jones [45] was limited to 10 representative targets with a maximum chain length of 158 residues. Furthermore, FRAGFOLD [160] was used for *ab initio* protein structure prediction, a method with inferior performance to ROSETTA [42] when using the decoys in unconventional MR (see Chapter 5). Thus, the more diverse set of decoys generated in this study might be more representative in determining a correlation between decoy TM-scores and contact pair satisfaction.

A Pearson's CC analysis with 113 ROSETTA decoy sets representing 56 globular and transmembrane targets showed a positive linear correlations between a decoy's TM-score and short-, medium- and long-range contact satisfaction (Table 6.1). Furthermore, separating the correlation analysis of all targets by fold classification revealed that all-α, mixed α-β and transmembrane protein targets showed the strongest positive correlations for long-range contact satisfaction (Table 6.1). All-β and mixed α-β decoy sets showed the strongest correlations for short- and medium-range contact satisfaction, whereby the former highlighted a stronger positive correlation between the decoy's TM-score and its medium-range contact satisfaction than its long-range contact satisfaction (medium-range:$\rho = 0.54$; long-range:$\rho = 0.50$) (Table 6.1). Notably, the decoys of transmembrane protein targets showed no significant correlation between TM-score and short-range contact satisfaction ($\rho = 0.08$; Table 6.1).

Table 6.1: Pearson's CC analysis between a ROSETTA decoy's TM-score and short-, medium- and long-range contact satisfaction. Probability values for all ρ coefficients are < 0.01.

Target class	**Pearson's CC**		
	Short-range	Medium-range	Long-range
all	0.11	0.18	0.64
all-α	0.30	0.44	0.69
all-β	0.40	0.54	0.50
mixed α-β	0.42	0.55	0.69
transmembrane	0.08	0.48	0.70

Following on from the Pearson's CC analysis, a linear regression model was fitted to individual subsets of the data used for the correlation analysis to see if a decoy's TM-score could be predicted from its contact satisfaction score. However, weak coefficients of determination indicated that only some cases show models with reasonably

good fits to the data (Fig. 6.1). Nevertheless, all models further supported the positive linear correlations between a decoy's TM-score and its range-dependent contact satisfaction. Interestingly, the strongest and best fits of the linear regression model to its corresponding data was for long-range contact pairs, where the linear regression models were also near identical between the different fold categories (Fig. 6.1).

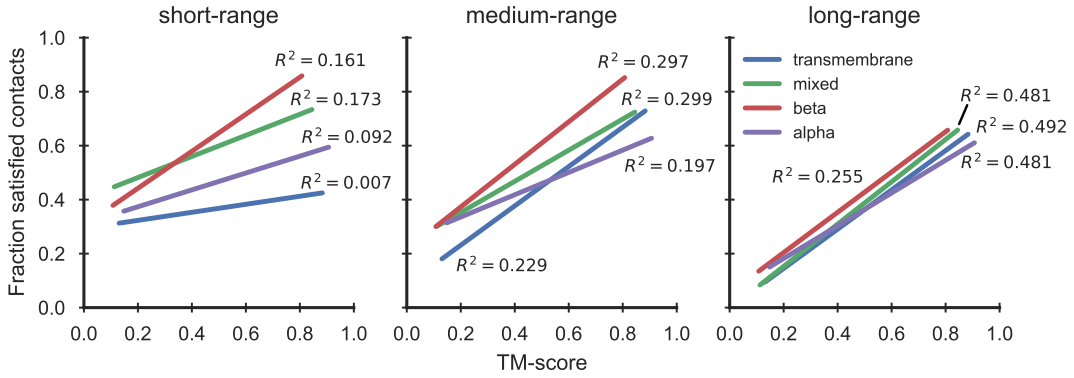


Figure 6.1: Linear regression model fitted to decoy TM-scores and corresponding fractions of satisfied, range-dependent contacts. Targets were further separated by fold classification. Coefficients of determination (R^2-values) added alongside each regression model.

An analysis of the correlation between the TM-score and long-range contact satisfaction of individual decoy sets further highlighted the potential to subselect decoy sets by their long-range contact satisfaction. One hundred and eight decoy sets showed statistically significant positive correlations between decoy TM-scores and their long-range contact satisfaction (ρ-values in range of 0.09 to 0.97 with p-value < 0.01). A single ROSETTA decoy set, derived for the glycolipid transfer protein (PDB ID: 2eum) and restrained with predicted METAPSICOV STAGE1 contact data, showed a weak negative correlation ($\rho = -0.10$, $p < 0.01$). The remaining four decoy sets, derived for targets with PDB IDs 1chd, 1gm4, 2x6u and 3ouf and restrained with predicted METAPSICOV STAGE1 contact data except for 2x6u (PCONSC2), showed no statistically significant correlation between the TM-score and long-range contact satisfaction of the decoy sets.

A further subdivide of the previously presented data by metapredictor highlighted that no predictor outperformed the others. Decoy sets calculated using predictions from all metapredictors exhibited a range of stronger to weaker correlations. Similarly, target chain length and fold did not show overall stronger or weaker correlations.

So far, all analyses focused on entire sets of decoys (1,000 decoys per set); however, it is often desirable to know if one could better estimate the accuracy of the best decoy by some measure. Kosciolek and Jones [45] demonstrated strong positive correlations for short-, medium- and long-range contact satisfaction with a decoy's corresponding TM-score (short-range:$\rho = 0.50$; medium-range:$\rho = 0.57$; long-range: $\rho = 0.87$). In

this work, some of these findings were confirmed (short-range: no correlation; medium-range:$\rho = 0.52$; long-range:$\rho = 0.69$) although the strength of the correlation for long-range contact satisfaction is much weaker than previously observed (Fig. 6.2). The weak positive correlation for short-range contact satisfaction is statistically non-significant, and thus could not be validated.

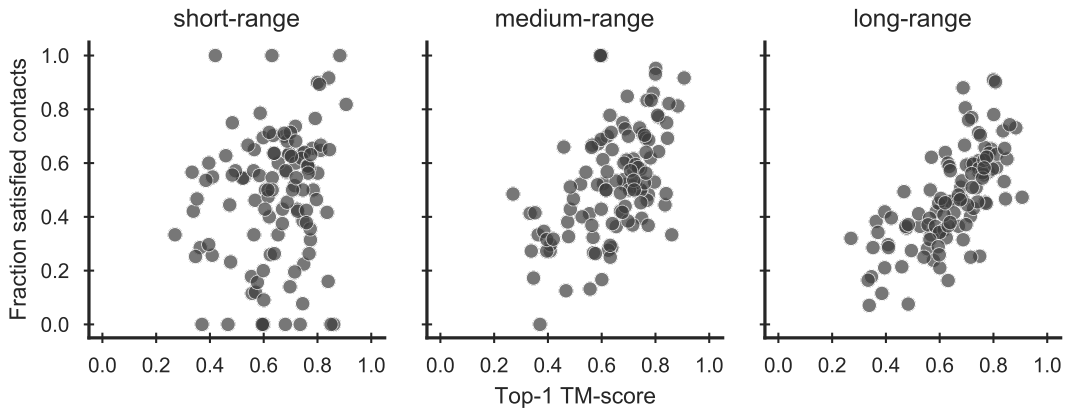


Figure 6.2: Analysis of the relationship between TM-score and contact satisfaction for the top-1 decoy (as ranked by TM-score) in each decoy set.

6.3.2 Long-range contact satisfaction metric to filter decoy sets

In Section 6.3.1, the data highlighted that decoy quality correlates positively with contact satisfaction. In particular, a strong positive correlation between long-range contact satisfaction and decoy quality could be established for almost all decoy sets in this study. A key ambition in this work was to determine if this correlation could be used to alter the starting decoy sets prior to the submission to the AMPLE cluster-and-truncate pipeline to enhance the chances of generating ensemble search models for more frequent MR success.

The difference in mean TM-score of each decoy set before and after applying a subselection strategy (see Section 6.2.3) is shown in Fig. 6.3. Estimating a decoy's quality by short-range contact satisfaction resulted in marginal mean TM-score changes of decoy sets ($\Delta_{CUTOFF} = -0.003$; $\Delta_{LINEAR} = 0.008$; $\Delta_{SCALED} = 0.001$). In comparison, medium- ($\Delta_{CUTOFF} = 0.005$; $\Delta_{LINEAR} = 0.015$; $\Delta_{SCALED} = 0.002$) and especially long-range ($\Delta_{CUTOFF} = 0.025$; $\Delta_{LINEAR} = 0.032$; $\Delta_{SCALED} = 0.005$) contact satisfaction were better metrics to use to improve the mean TM-scores of each decoy set. Notably, per-decoy long-range contact satisfaction provided the best estimate for identifying and excluding the least accurate decoys independent of the subselection strategy.

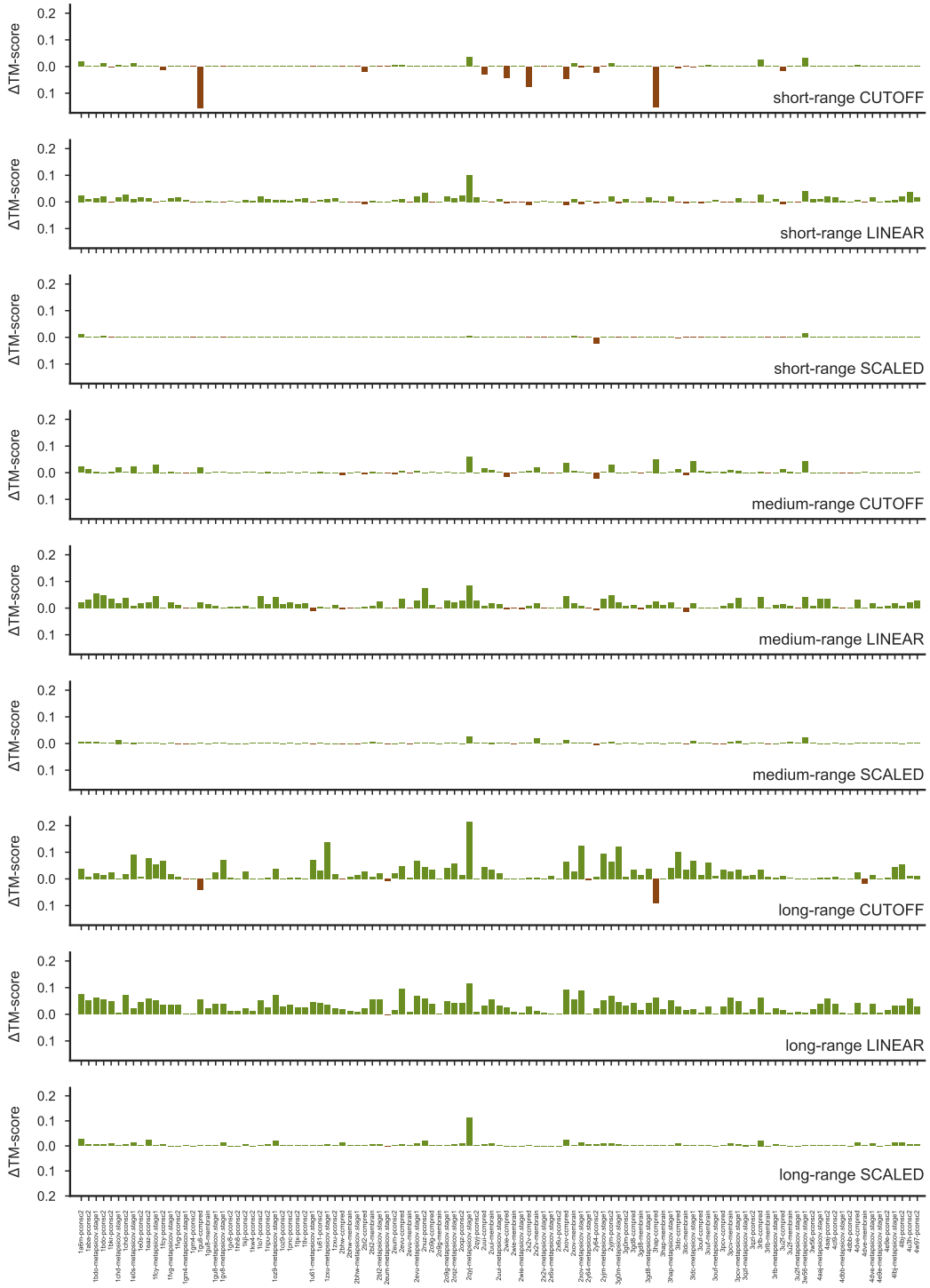


Figure 6.3: Differences in mean TM-score for decoy sets pre- and post-decoy subselection. Each subselection strategy is stated in each subplot along with the contact range used to establish decoy inclusion in the final set. Green bars indicate TM-score changes in favour of subselected decoy sets, red in favour of complete ones.

Given the improvement of TM-scores for each decoy set by decoy subselection, it was important to analyse the number of decoys left in each set after long-range contact-satisfaction subselection. This metric is important since too few decoys might

not generate any AMPLE ensemble search models due to AMPLE's filters after clustering and sub-clustering. For the decoy sets used in this study, the *LINEAR* strategy removed on average the most decoys from each set with a fixed number of 500 (median=500). In comparison, the *CUTOFF* subselection strategy removed on average 409 decoys (median=316) whilst the *SCALED* method only 56 (median=29). However, the sample-dependent strategies (*CUTOFF* and *SCALED*) may remove a much greater number of decoys from a set if the corresponding satisfaction scores fall below a certain threshold (maximum removed by *CUTOFF*=1000 and *SCALED*=497). Since these numbers varied drastically similarly to the changes in TM-score, it became apparent that the more decoys were removed, the better the overall score became, which further supported the linear correlation between long-range contact satisfaction and TM-score.

In certain cases, some subselection strategies greatly altered the overall size and quality of the resulting decoy set, which started with a set of 1,000 decoys. The METAPSICOV STAGE1 decoy set of the ankyrin sequence (PDB ID: 2qyj) showed overall quality improvements from 0.006 (short-range *SCALED*; $n_{models} = 958$) to 0.213 (long-range *CUTOFF*; $n_{models} = 218$). The CCMPRED decoy set of sensory rhodopsin II sequence (PDB ID: 1gu8) showed overall changes from -0.155 (short-range *CUTOFF*; $n_{models} = 2$) to 0.06 (long-range *LINEAR*; $n_{models} = 500$).

Overall, the optimal strategy to select or exclude decoys from a starting set of structures appeared to be long-range contact satisfaction driving the *LINEAR* strategy.

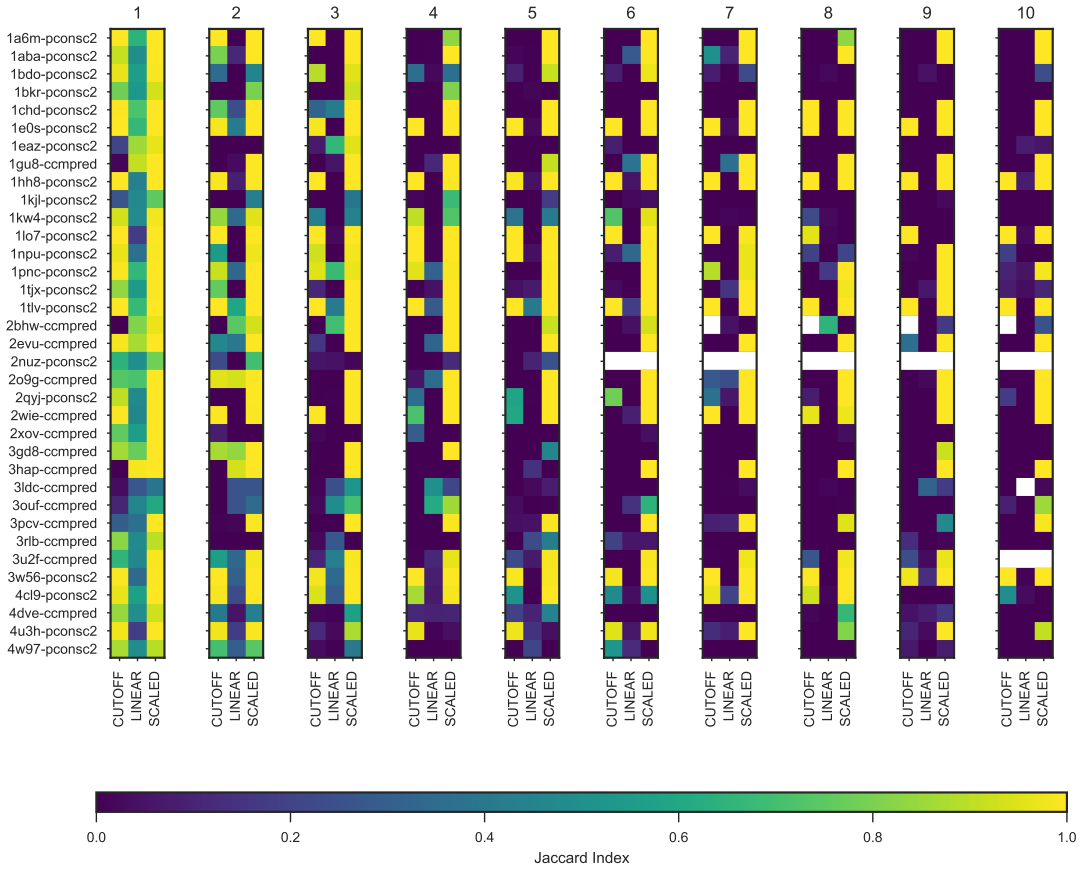
6.3.3 AMPLE's cluster-and-truncate approach with filtered decoy sets

For evaluation of performance of filtered decoy sets in MR, a smaller sample of 35 decoy sets was selected spanning 35 unique targets (21 globular and 14 transmembrane targets). The contact prediction algorithm generating the restraints for the *ab initio* structure predictions was PCONSC2 (globular targets) or CCMPRED (transmembrane targets). Each decoy set was subjected to the AMPLE pipeline with certain decoys removed according to one of four subselection strategies, namely *NONE*, *CUTOFF*, *LINEAR* and *SCALED*.

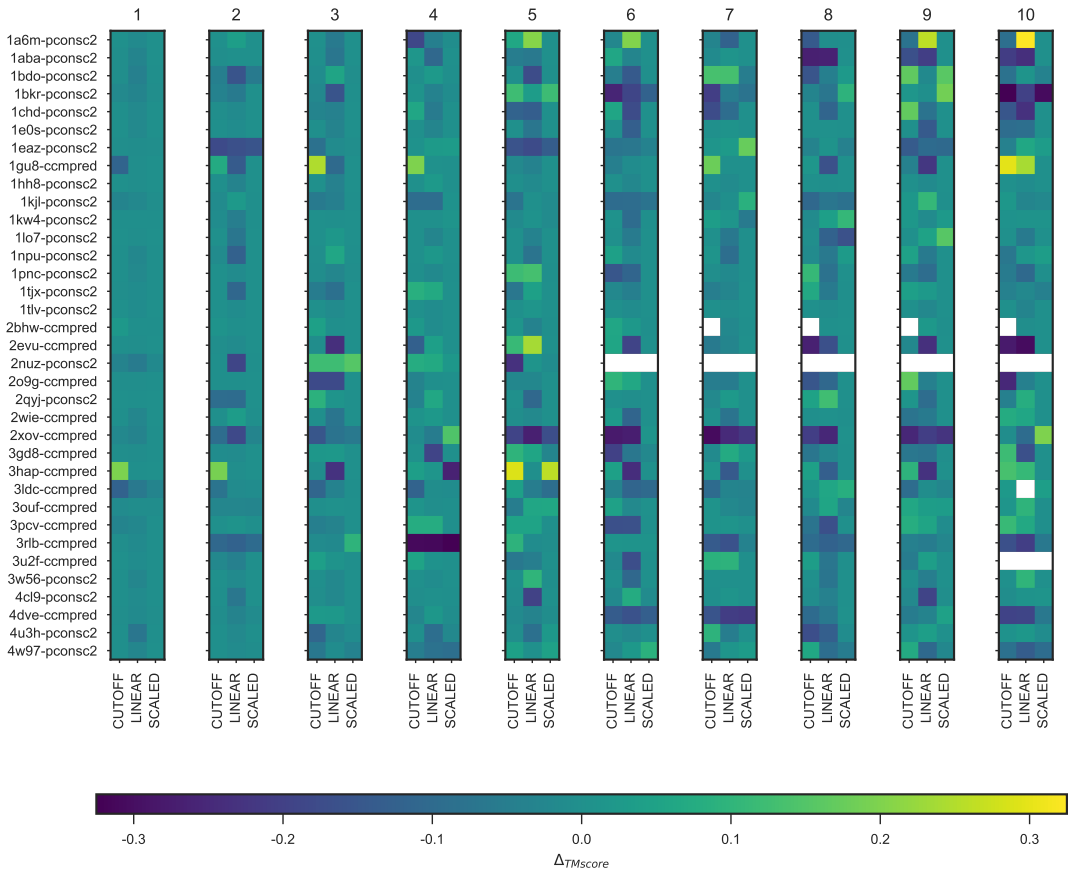
The initial step in the AMPLE pipeline is the clustering of decoys. A comparison of SPICKER clusters between the *NONE* default strategy and the *CUTOFF*, *LINEAR* and *SCALED* subselection strategies highlighted an important observation. Larger clusters — those ranked higher — showed higher similarity between a subselection strategy and the default (Fig. 6.4a). The top SPICKER cluster showed high similarities between the *NONE* strategy and all other subselection ones, whereby it has to be noted that the *LINEAR* strategy contained only 50% of the starting decoys, and thus can at best show a Jaccard index of 0.5. With increasing cluster index, the overall similarity degraded and most of the decoys in cluster 10 were non-identical between

each subselection strategy and the default. It is important to consider though that clusters might be swapped between subselection strategies, and thus the Jaccard index might not reliably indicate presence of individual decoys.

Furthermore, a similar analysis to compare the overall quality of each cluster to the target structure revealed less difference between the default and each subselection strategy for higher-ranked SPICKER clusters (Fig. 6.4b). With decreasing SPICKER cluster index, the difference in median TM-scores started to alternate without any particular pattern. Thus, pre-selecting decoys prior to AMPLE's cluster-and-truncate approach most certainly preserved the top cluster for the *CUTOFF* and *SCALED* subselection strategies, whereby lower clusters showed more deviation from the default.



(a)



(b)

Figure 6.4: Effect of decoy subselection on SPICKER clusters. Effect illustrated by (a) the Jaccard Index and (b) median TM-score difference. Values were calculated for clusters resulting from the full starting set of decoys and the *CUTOFF*, *LINEAR* and *SCALED* subselection strategies. Larger TM-score differences indicate that the subselection improved the TM-score of the cluster.

The mean of the inter-decoy variance computed by THESEUS — used in AMPLE to guide truncation of each cluster — is reduced in lower clusters compared to the *NONE* default strategy (Fig. 6.5). In other words, these clusters have become more structurally homogeneous. The clusters of decoys based on the galectin-3 domain (PDB ID: 1kjl) sequence show overall the highest reduction in mean inter-decoy variance up to -15Å^2 compared to the default strategy. Similarly, clusters 4 and 8 of the K^+-channel protein domain (PDB ID: 3ouf) show reductions in mean inter-decoy variance of up to -20Å^2. In general, clusters starting from *CUTOFF*-subselected decoys show the greatest mean inter-decoy variance reductions, followed by *LINEAR* and then *SCALED*-subselected decoys sets.

A comparisons of intermediate stages in the AMPLE pipeline resulting from differently subselected decoy sets is generally very difficult. Each strategy resulted in different starting sets, which resulted in different clusters. Since AMPLE's objective truncation procedure was based on the inter-decoy variance, it might be greatly affected by differing

clusters. Nevertheless, structure solution is more likely when AMPLE generated more ensemble search models because a greater number of search models reflects greater inter-cluster decoy similarity and trialling a greater number should provide a higher chance of success. A count of generated AMPLE ensemble search models revealed that the *SCALED* strategy generated the most search models ($n = 7,611$), which is roughly 300 more than the default *NONE* strategy ($n = 7,340$). The *CUTOFF* subselection strategy generated the least ensemble search models ($n = 7,237$), whilst the *LINEAR* strategy's count ($n = 7,401$) was very similar to the *NONE* one.

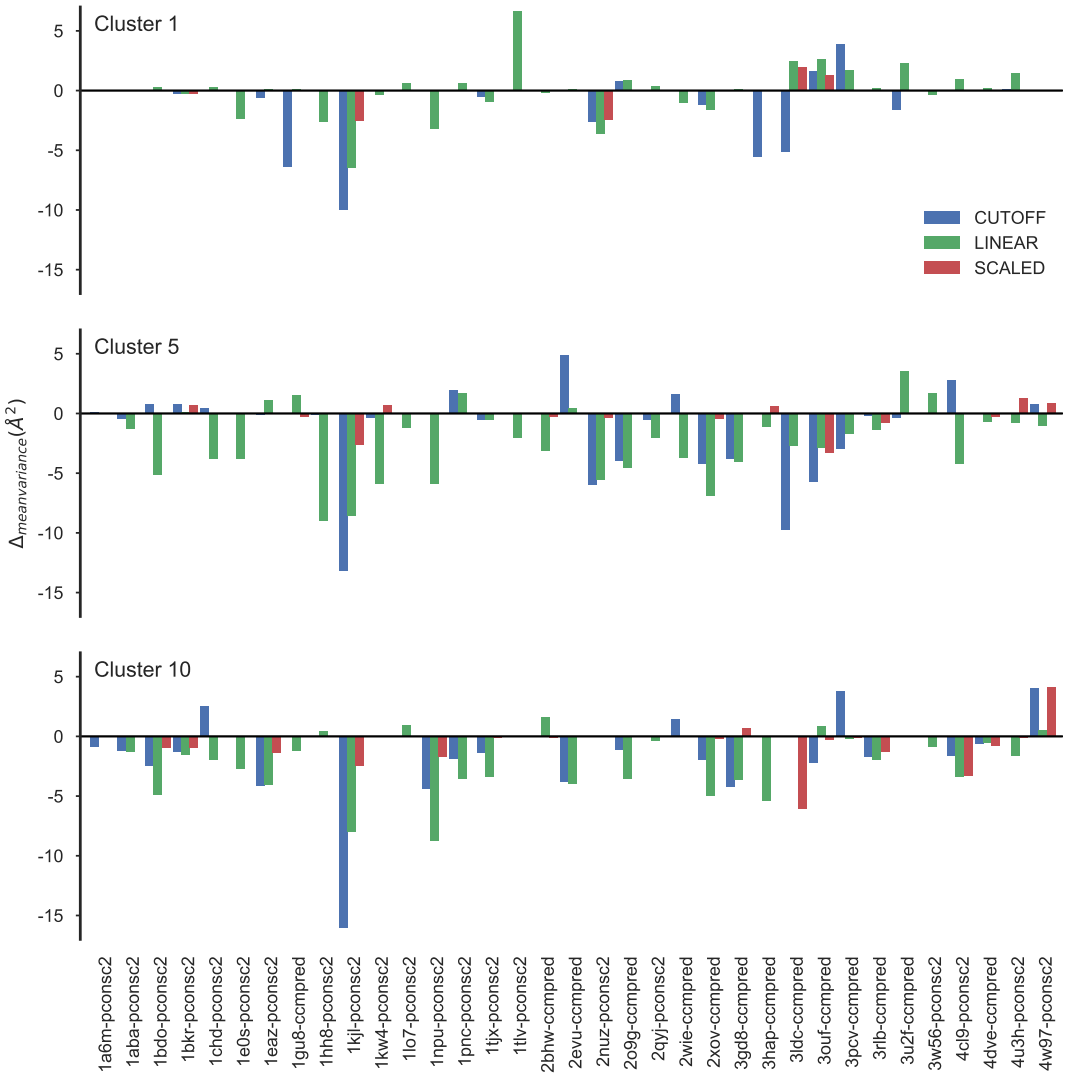


Figure 6.5: Effect of decoy subselection on mean inter-decoy THESEUS variance. Difference in mean variance calculated between the default and the three decoy subselection strategies *CUTOFF*, *LINEAR* and *SCALED*. Data for clusters 1, 5 and 10 shown as examples.

Further inspection of the number of AMPLE ensemble search models by target revealed near identical numbers between the *NONE*, *LINEAR* and *SCALED* strategies (Fig. 6.9). In fact, only few outliers for each of those methods distinguished them from

the others. The *CUTOFF* strategy showed greater deviation from the other three, especially for certain targets with differences up to approximately 200 ensemble search models (Fig. 6.9). A comparison of all these strategies to the previous default processing in AMPLE (*NONE_classic*; further details in Section 6.2.4) highlighted a reduction in the total number of generated ensemble search model count (Fig. 6.9). A comparison of the previous default (*NONE_classic*) with the new one (*NONE*) showed on average 144 fewer ensemble search models per target, whilst sampling a larger range of folds through all ten clusters.

6.3.4 MR search models by processing single decoys

In addition to the decoy set subselection, this study also attempted to identify single decoys of sufficient quality to be used directly as MR search models. Although ensembles are generally more desirable MR search models [117, 120, 127], individual decoys might be successful by themselves, and thus save the overhead of generating and trialling a great number of AMPLE ensemble search models. Thus, the top-5 decoys, as judged by long-range contact satisfaction, were selected from each decoy set. Four distinct processing approaches were applied to each decoy to eliminate less reliable parts, and subsequently compared against the unmodified initial decoy.

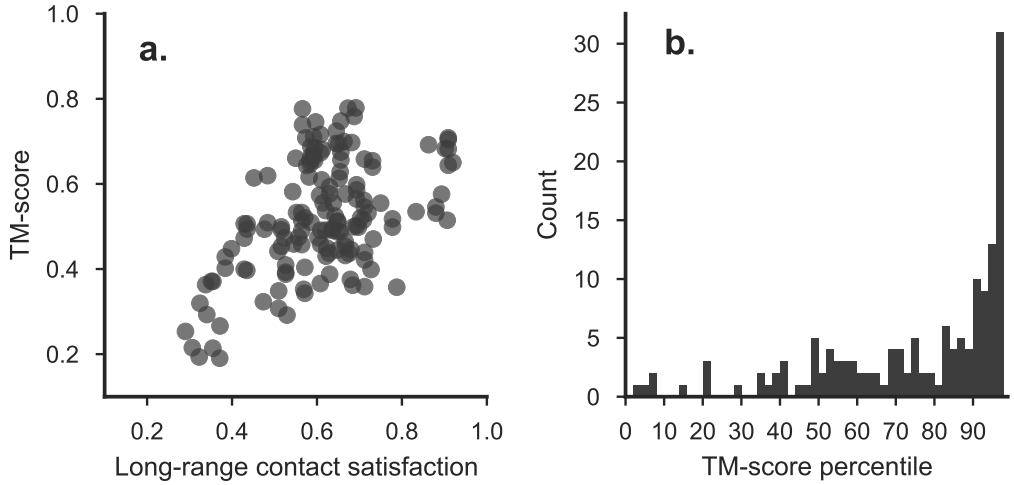


Figure 6.6: Analysis of long-range contact-satisfaction-based decoys with respect to the relationship between (a) the satisfaction and decoy quality and (b) the quality compared to the remaining, excluded decoy set.

The correlation between a decoy's long-range contact satisfaction and its TM-score has previously been outlined and was further confirmed here (Fig. 6.6). However, the positive correlation was dependent on the target's fold class and the overall accuracy of the decoy set. An analysis of the top-5 decoys by long-range contact satisfaction in each decoy set showed that 50% of selected decoys fall in the 80^{th} percentile or greater of

TM-scores in each decoy set whilst 90% are in the at least the 40th percentile (Fig. 6.6).

A comparison of RMSD value changes indicated that the "fragment" and "variance" metrics provided the best approximation to identifying less-reliable regions in each decoy. The average RMSD change compared to the original decoys was just under 4.0Å. This compared to a slightly lower RMSD change of 2.1Å for "DSSP"-treated decoys and 1.5Å for the "domain" treatment. Although almost all decoys were improved by either of the treatments, a small number of decoys worsened in terms of RMSD compared to its native structure. All treatments except the "fragment" one had worsened decoys in the final set, with changes up to -1.6Å.

A comparison of the range of RMSD values revealed much greater changes for the "fragment" and "variance" conditions (Fig. 6.7). However, these changes were not reflected in the fraction of residues retained in each decoy. Most residues were removed by the "domain" treatment (μ=61.6%), whilst the "fragment" one saw the least removal (μ=34.5%). Similarly to the ranged in RMSD values, the "fragment" and "variance" treatments resulted in the greatest spread of fraction of residues in the treated decoy. The values range for both treatments from retaining less than 5% of the initial decoy up to 100%.

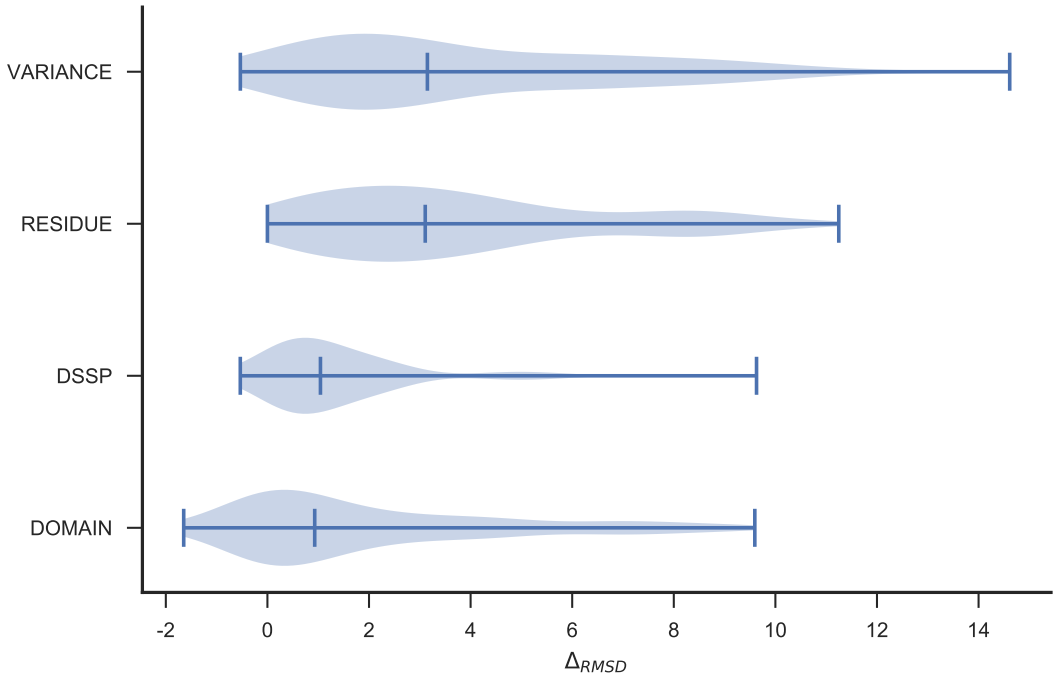


Figure 6.7: Distribution of differences in RMSD values between the initial and modified decoys under four different treatments. A positive Δ_{RMSD} value corresponds to a decrease in RMSD compared to the crystal structure.

A further aspect of the decoy treatments highlighted that the fraction of residues retained after decoy post-processing correlates with the cluster variance of the decoy, which were extracted from THESEUS results of each decoy's cluster in the *NONE*

strategy (Fig. 6.8). Unlike the variance metric, all other processing metrics did not show a correlation with the fraction of residues retained. This explains at least in part why much greater changes in RMSD value between the initial and processed decoy were observed for the "variance" treatment compared to the others. However, if a decoy was of particularly poor quality (TM-score < 0.3), the "variance" treatment retained as little as 0.87% and 1.7% of the initial decoy (2 and 4 residues) whilst the others retained a much larger fraction of at least 40% for equivalent decoys.

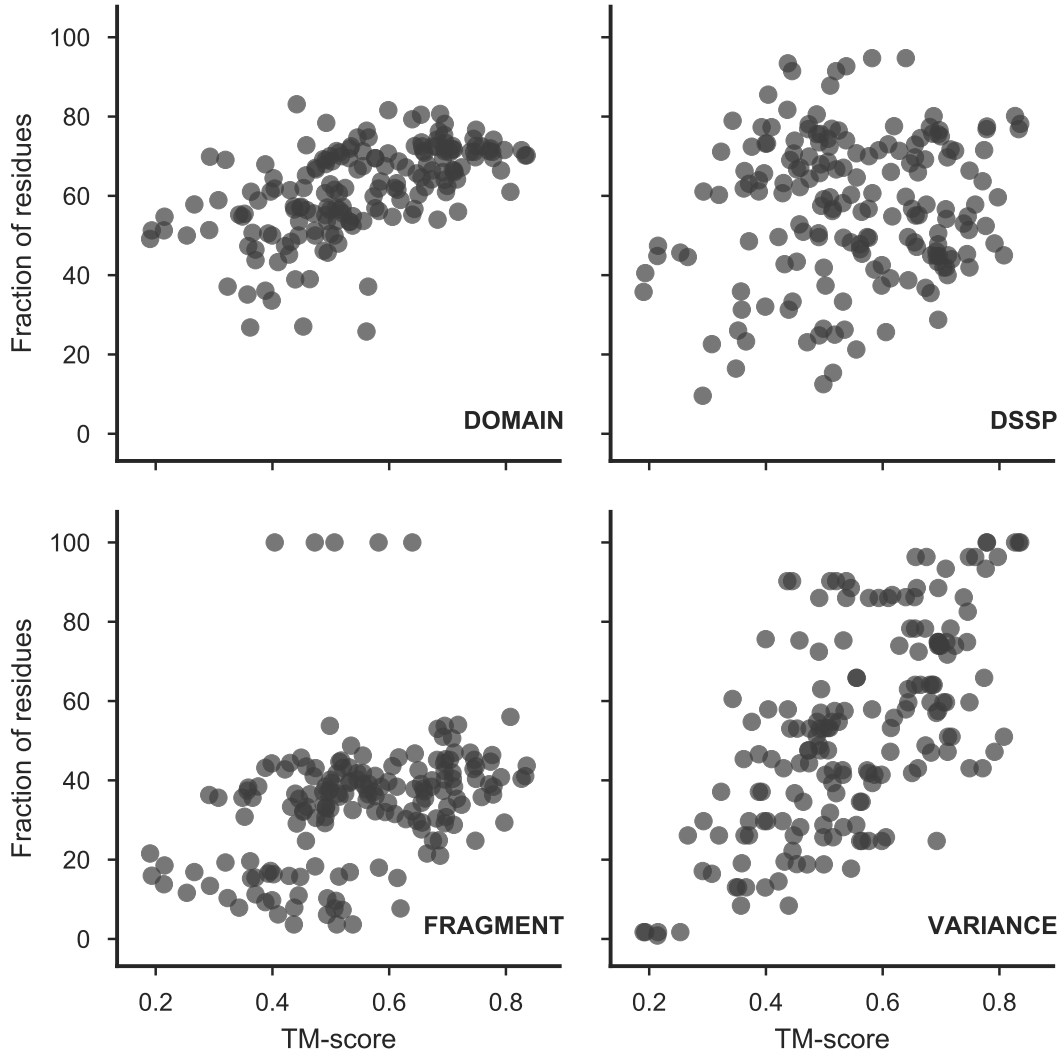


Figure 6.8: Scatter plots of initial decoy TM-score and the fraction of residues retained after one of four different residue removal treatments.

6.3.5 Decoy subselection extends AMPLE's performance

The final step in this study was the assessment of AMPLE ensemble search models and single-decoy-based search models in MR. In particular, the comparison of different decoy subselection strategies and individual decoy-processing treatments was of great interest since it might extend AMPLE's performance beyond that described in Chapters 3 to 5.

A comparison of the total number of targets solved by each subselection strategy showed that the *CUTOFF*-subselected decoys led to most structure solutions (14 out of 35) (Fig. 6.9). Although slightly less successful, the *LINEAR* and *SCALED* subselection strategies lead to structure solutions of two additional targets compared to the *NONE* strategy (11 out of 35). The *LINEAR* and *SCALED* strategies were on par with AMPLE's default, the *NONE_classic* strategy (Fig. 6.9). Although the *NONE_classic* strategy generated two version of each ensemble search model with poly-Alanine and all-atom side chain treatments, the former was enough to solve all targets outlined in (Fig. 6.9). Therefore, the *LINEAR* and *SCALED* subselection strategies would be the minimum processing requirement to solve the same number of targets with fewer search models and hence improved performance.

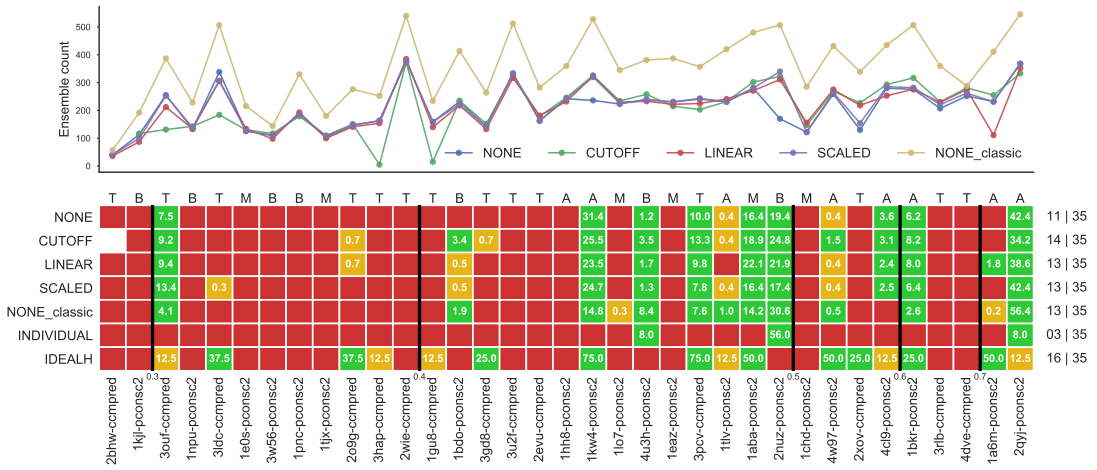


Figure 6.9: Molecular Replacement summary of decoy-subselected AMPLE ensembles. AMPLE-generated ensemble counts illustrated at the top with Molecular Replacement results in grid below: red cell equates to no solution; orange to a singleton solution; and green to multiple solutions. All *INDIVIDUAL* attempts were compressed to a single row per decoy set. The number in the orange and green cells indicates the percentage of ensemble search models leading to structure solutions. One letter code above each column indicates the target fold: "T" for transmembrane; "A" for all-α; "B" for all=β; "M" for mixed α-β. Values alongside each row indicate the number of targets with structure solutions and total number targets attempted. Targets are sorted from left to right with increasing median TM-score of the starting decoy set. The black lines highlight TM-score thresholds from 0.3 to 0.7 from left to right. The subselection strategy *IDEALH* refers AMPLE's ideal helix library.

The *CUTOFF* method yielded the highest number of structure solutions based on AMPLE ensemble search models whilst generating the fewest search models. In fact, this subselection strategy generated no ensemble search models for target 2bhw. Furthermore, the *CUTOFF* method achieved amongst the best ratio of search models leading to structure solution compared to the total number produced.

In few cases, only a single AMPLE search model led to a structure solution (orange cells in Fig. 6.9). Upon closer inspection, 71% of all singleton solutions were achieved

with AMPLE ensemble search models containing at least 30% of the target sequence. Twenty-nine percent of the singleton solutions contained at least 50% of the target sequence, whilst none contained more than 70%. Three out of four search models with less than 30% of the target sequence were derived from the PCONSC2 decoy set predicted for the ketosteroid transcriptional regulator KstR2 (PDB ID: 4w97) sequence and contained one, two or three small helical fragments.

In certain cases the subselection of starting decoys made a subtle yet essential difference to generating an AMPLE ensemble search model for successful structure solution. An example of such a case is the CCMPRED decoy set of the aquaporin Z domain (PDB ID: 2o9g). *CUTOFF* and *LINEAR* subselected decoys led to a single search model each (cluster 1; 59% truncation and subclustering radius of 3Å), which was sufficient for structure solution (Fig. 6.9). The *NONE* and *SCALED* subselection strategies generated an ensemble with identical AMPLE processing parameters, which did not lead to structure solution (Fig. 6.9). An analysis of the decoys in the ensembles reveals that 30% (9 out of 30) were different between the successful ensembles and the *NONE* strategy. However, only a single decoy was unique to either *CUTOFF* and *LINEAR* in a direct comparison. Ultimately, this resulted in a RMSD difference between the *NONE* and *CUTOFF* ensembles of 2.25Å (Fig. 6.10), whilst the *CUTOFF* and *LINEAR* ensembles are identical (RMSD=0.00Å). Thus, subselection showed crucial value in preparing decoy datasets prior to AMPLE's cluster-and-truncate approach.

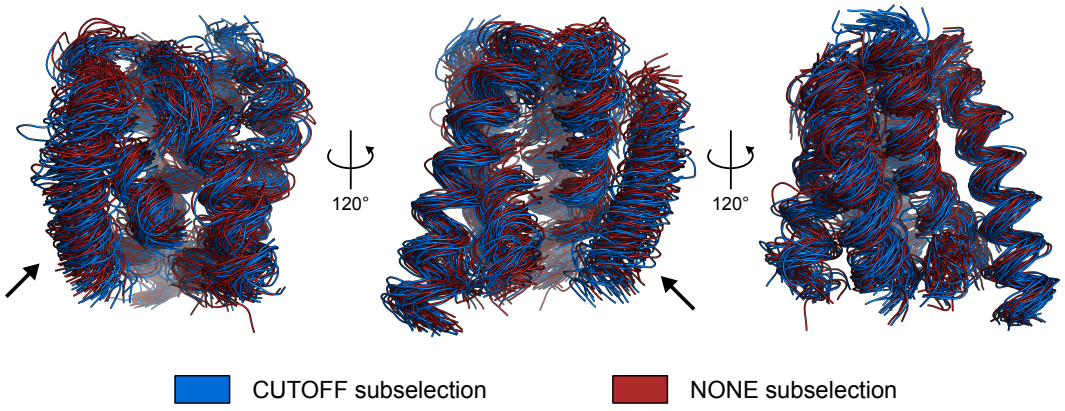


Figure 6.10: Example of the structural divergence of two ensemble search models with an identical AMPLE cluster-and-truncate path. Ensembles are based on the CCMPRED decoy set of PDB ID 2o9g and derived from cluster 1 with 59% truncation and subclustering radius of 3Å. The blue ensemble was derived from *CUTOFF* subselected decoys and the red ensemble without subselection. The blue ensemble search model is successful in deriving a MR structure solution, the red one is not. The arrow indicates the substructure with the highest degree of structural divergence.

The rank order of targets by median TM-score of the initial starting decoy set in Fig. 6.9 showed that no decoy set with median TM-score of less than 0.3 score units led to structure solution; however, only two such cases existed in the dataset, and therefore this threshold may only serve as indication. With increasing median TM-

score, i.e. increasing similarity between the decoy set and its reference target structure, the chances appear to increase to achieve structure solution. Beyond a threshold of 0.4 TM-score units, structure solutions were much more likely (over 50% of targets solved with one of the four subselection strategies), which highlights AMPLE's success in processing such accurate decoy sets appropriately.

The work in this study further explored whether individual decoys could be selected via their long-range contact satisfaction and trialled directly as MR search models. The *INDIVIDUAL* subselection strategy explored this aspect with a variety of post-selection processing approaches. However, structure solutions for only three targets could be obtained using this single-decoy approach (Fig. 6.9). All processing strategies obtained led to structure solutions based on the PCONSC2 decoy set of the α-spectrin SH3 domain (PDB ID: 2nuz). The other two targets with solutions, PDB IDs 2qyj and 4u3h with PCONSC2 decoy sets, solved at least once with a single decoy subjected to the "domain", "DSSP", "fragment" or "variance" treatments. Across the three targets, only five decoys (three based on the sequence of PDB ID 2nuz) with a minimum TM-score of 0.682 resulted in the 20 structure solutions (PDB ID 2nuz: 16 solutions; PDB ID 2qyj: 2 solutions; PDB ID 4u3h: 2 solutions).

Lastly, a comparison of decoy-derived search models and AMPLE's simplistic ideal helix library [117] in MR was done. Ideal helices achieved the most structure solutions solving 16 out of 35 targets (Fig. 6.9). In particular, ideal helices achieved structure solutions for more transmembrane targets. Eight out of 14 transmembrane targets were solved with at least one ideal helix, which compares to six out of 14 for all decoy-based search models combined. No transmembrane target was solved with decoy-based search models that could not be solved with ideal helices. The number of solved transmembrane targets was also increased by two compared to the work by Thomas et al. [118], which was exclusively due to improved MR software. Ideal helices also managed to achieve near identical results for all-α and mixed α-β targets in the set compared to decoy-derived search models. However, four targets remained intractable by ideal helices yet were solved with decoy-based search models. Three of these targets are all-β targets (PDB IDs: 1bdo, 2nuz and 4u3h) and the fourth a mixed α+β one (PDB ID: 1lo7). Lastly, Thomas et al. [118] suggested that decoy-derived search models were essential since ideal helices provide insufficient scattering matter with low resolution (> 2Å) intensity data. In this study, these findings could not be validated given that PDB ID 1gu8 (resolution of 2.27Å) was solved solely with ideal helices whilst being the target with the lowest resolution of all solved ones.

6.4 Discussion

The subselection of decoy sets by long-range contact satisfaction is a concept originally proposed by Kosciolek and Jones [45] and later confirmed and extended by Oliveira

et al. [112] and Adhikari and Cheng [132]. In this study, these findings were further confirmed by reanalysing all decoy sets generated in Chapters 3 to 5.

Furthermore, the benefit of subselecting decoys based on their long-range contact satisfaction pre-AMPLE was evaluated. Subselection extended the target tractability of AMPLE whilst reducing the number of generated search models, which effectively enhances AMPLE's performance. The *CUTOFF* subselection strategy proved to be most successful in flagging the worst decoys, which resulted in more accurate ensemble search models being generated. The data presented showed that subtle differences in clustering have significant effects on ensemble search model generation resulting in the loss or gain of structure solutions. Finally, given that the *NONE* strategy has become AMPLE's default since this study was conducted, the results are important for AMPLE users to improve the chances of structure solution.

Based on the results in this work, it also became apparent that decoy-based ensemble search models are inferior to AMPLE's simple ideal helix library, particularly for transmembrane protein targets. The latter was sufficient to solve the majority of transmembrane protein targets, which outperformed all decoy-based approaches combined. This result contradicts the one reported by Thomas et al. [118], who found that decoy-based search models are required when the resolution was worse than 2Å. Furthermore, it is expected that the application of more sophisticated ideal helix library approaches, such as ARCIMBOLDO [171] or FRAGON [164], would make decoy-based search models less needed for transmembrane targets. However, decoy-based search models are still required, especially for globular folds with little or no helical secondary structure. Decoy-based search models are also needed when the resolution of the experimental data is low ($<$ 2Å). In such cases, MR algorithms require higher proportions of scattering matter compared to the asymmetric unit content to detect the signal of a correctly placed search model [172]. Since it is easier to derive larger search models by truncating sequence-specific decoys than identifying larger fragments or even substructures, decoy-based search models are still needed.

Beyond subselecting decoys sets, some very preliminary work in this chapter aimed to explore the possibility of identifying, processing and trialling individual *ab initio* structure predictions as MR search models. Although previous work has extensively demonstrated the benefits of ensembles over individual search models in MR [117, 120, 127], interest in this approach remains. In particular, individual decoys with high similarity to the crystal structure are sometimes present amongst 1,000 non-native-like starting decoys. Although such decoys are included in AMPLE ensemble search models, trialling them individually might enhance the performance of AMPLE by avoiding the generation and trial of potentially hundreds of ensemble search models. As such, identification and MR trial could be crucial to solving a target, whose sequence was used to predict the decoys. However, findings in this work supported previous challenges in the field of identifying the very best decoys reliably by long-range contact satisfaction

[45, 112, 132]. Although a general correlation exists for most decoy sets, the best decoy by long-range contact satisfaction is not necessarily the very best by TM-score. Thus, the data suggests that AMPLE's ensembling routine remains the more successful. Nevertheless, further work needs to be conducted to explore alternate decoy processing options. These could include a combination of metrics used in this study, or alternatives such as solvent accessible surface. Furthermore, exploiting contact information to aid AMPLE's cluster-and-truncate approach could prove a promising alternative, too.

Chapter 7

Protein fragments as search models in Molecular Replacement

7.1 Introduction

Ab initio structure prediction algorithms typically start with a coarse grained search of conformational space through the assembly of previously picked structural fragments. As such, the accuracy of structure prediction is heavily dependent on the similarity of fragments to the target fold for each position [54]. Thus, the necessary structural information for accurate structure prediction must be encoded in the fragment library for a given target sequence. This approach allows the modelling of new protein folds by considering them as assemblies of already known building blocks, such as super-secondary structure motifs [173]. Furthermore, fragments similar to those typically selected for *ab initio* protein structure prediction were successfully used in other areas of Structural Biology including NMR [174, 175] and X-ray crystallography [176] studies to elucidate unknown protein folds. Despite their modest success, almost all attempts neglected target-specific information generally available to structural biologists obtainable through Bioinformatics software. This information includes the primary sequence of the target, torsion angle predictions, predicted solvent accessibility or coevolution information. In theory, all additional information should improve the generation of such fragment libraries by aiding the selection process or cross-validating the identified fragments.

Over the last decade, efforts have been made to improve the precision of structural fragment libraries used in *ab initio* structure prediction [47–54, 177]. Various different algorithms have been developed to generate static and dynamic fragment libraries. Static fragment libraries are those precomputed and generally consist of common super-secondary structure motifs. In comparison, dynamic fragment libraries consist of fragments of variable lengths acknowledging the fragment-dependent optimal length. Most commonly used in *ab initio* structure prediction are dynamic algorithms, such as FLIB [53], FLIB-COEVO [177], NNMAKE [54] or HHFRAG [50]. Dynamic-library producing algorithms differ in their definition of ideal fragment lengths, the default number of fragments used per position and the way in which fragments are extracted. However, these algorithms typically share the same additional sequence-based information used to aid the selection of target fragments, which usually includes sequence similarity, three-state secondary structure prediction and torsion angle prediction.

Given that fragment libraries selected to perform *ab initio* structure prediction can contain high quality fragments or super-secondary structure motifs, those fragments must sometimes be suitable as MR search models. Correct identification of very similar fragments should allow for dynamic fragment selection to achieve MR structure solution without the overhead of *ab initio* structure prediction. Furthermore, dynamic algorithms could pick fragments of varying lengths, possibly matching coevolution data or other externally obtainable restraints to validate fragments prior to any MR attempt.

As such, the work in this chapter focused on exploring this idea using FLIB-COEVO

[177], a dynamic fragment picking algorithm considering coevolution data to verify fragments during the picking procedure.

7.2 Materials & Methods

7.2.1 Target selection

Four targets were manually selected for this study. The crystallographic data needed a resolution of around 1.5Å with a single molecule in the asymmetric unit. The target chain length needed to be below 150 residues, and the fold of the protein structure to be either mixed α-β or all-β. A further target selection criterion was the availability of precise contact information for fragment selection.

The PDB identifiers of the selected targets were: 1aba, 1lo7, 1u06, and 5nfc. The former two are described in Table A.1. Target 1u06 is a more recently published structure of α-spectrin SH3 domain (PDB ID: 1kjl in Table A.1) with a resolution of 1.49Å. Target 5nfc is a more recently published structure of Galectin-3 (PDB ID: 1kjl in Table A.1) with a resolution of 1.59Å. This resulted in a dataset with similar attributes for each target: crystallographic data resolution of 1.5Å with a single molecule in the asymmetric unit, and the target chain length of less than 150 residues. Each fold class, mixed α-β and all-β, contained two targets.

7.2.2 Fragment picking using FLIB-COEVO

FLIB-COEVO [177] requires four inputs: the predicted secondary structure, predicted torsion angles, predicted or differently derived residue-residue contact pair data and a copy of the PDB. The secondary structure for each target was predicted using PSIPRED v4.0 [147] with default parameters. The torsion angles were predicted using SPIDER v2 [178] with default parameters, and residue-residue contact pairs predicted with METAPSICOV v1.04 [101] with default parameters. HHBLITS v2.0.16 [137] with `uniprot20` database v2016-02 was used by METAPSICOV to generate the MSA for contact prediction of each target sequence. BLASTP v2.2.31+ [179, 180] was used by PSIPRED with the `uniref90` database v2016-06. The local copy of the PDB for fragment picking was downloaded on August 11, 2016.

Two modifications were made to the default FLIB-COEVO v1.01 (`https://github.com/sauloho/Flib-Coevo`, commit `abade3b`) protocol. The first focused on exclusion of fragments with $> 90\%$ helical content (assigned by DSSP [157]). If fragments with $> 90\%$ helical content were allowed and residues were predicted to be part of an α-helix, fragment libraries tended to be overpopulated for these positions with short helices. This would generate fragment libraries very similar to ideal-helix libraries, which was

not the purpose of this work. The second modification was to allow fragments with RMSD > 10.0Å to the reference structure to be considered. This modification to the FLIB-COEVO algorithm was implemented for development purposes by the authors to validate the performance of the algorithm. However, to allow for the automatic calculation of RMSD value of each fragment without deliberately excluding less-similar fragments this constraint was lifted.

Two-hundred fragments were picked per target sequence position. Top-L or $L/2$ contact pairs were selected from both METAPSICOV STAGE1 and STAGE2 predictions with a minimum sequence separation of either 6 or 12 residues. Helical fragments were either included or excluded. The fragment length ranged from either 6 or 12 (dependent on minimum sequence separation) to 63 residues. In all instances the `-coevo_only` flag was set to exclude fragments with starting residues undefined by any contact pair in the set [1]. Overall, this generated 16 fragment libraries per target.

Each fragment library was then filtered to remove homologs of the target of interest to best replicate a blind study, in which this procedure may be relevant. BLASTP v2.2.31+ [179, 180] and HHPRED (HHBLITS v2.0.16 and HHSEARCH v2.0.16) [181] searches were conducted to identify homologous PDB entries. The BLASTP search was performed identically to Oliveira and Deane [177] against the `pdbaa` v2016-10 database using an E-value cutoff of 0.05. The HHPRED search parameters were identical to the MPI-Toolkit [182] web server version (`https://toolkit.tuebingen.mpg.de/`) and searches done against the `uniprot20` v2016-02 and `pdb70` v2016-09-14 databases. Fragments derived from PDB entries identified by BLASTP and HHPRED (probability score of $\geq$ 20.0) were excluded from the fragment libraries. This resulted in a much more rigorous homolog exclusion than similar *ab initio* protein structure prediction studies would typically employ. However, in this study fragments were used directly as search models, and thus excluding even distantly related protein structures made it truly blind.

All per-target fragments were then binned by their peptide lengths. Subsequently, they were ranked by FLIB-COEVO scores and RMSD values, and the best fragment from each length-dependent bin selected. Partially redundant fragments of the same template structure consisting of the same region with varying flanking residues were kept, if they were ranked top for each fragment length group. Finally, the coordinates of the fragment backbone atoms were extracted to create poly-alanine search models.

Note, the FLIB-COEVO score refers in this chapter to the predicted torsion angle score for a given fragment, which FLIB-COEVO uses in its default routine to rank fragments with lower scores being more favourable [53, 177].

[1] The `-coevo_only` flag was intended to select only fragments that satisfied at least one contact pair. This originally intended behaviour was not part of the source code throughout this study, and only detected post-analysis. The issue was reported to the developers and has since been fixed in the FLIB-COEVO source code (commit "`b3eb01d`").

7.2.3 Molecular Replacement in MRBUMP

The previously extracted fragments were subjected to the MR pipeline MRBUMP v0.9 shipped with CCP4 v7.0.28 [127]. This uses PHASER [128] for MR, REFMAC5 [28] for refinement and SHELXE [130] for density modification and main-chain tracing. MRBUMP default parameters were used with exception of the PHASER RMSD estimate. Each fragment was subjected to MRBUMP using PHASER RMSD values of 0.1, 0.6 and 1.0Å.

7.2.4 Assessment of FLIB-COEVO fragments

Fragment torsion angles — predicted by SPIDER v2 [178] — were assessed using the Mean Absolute Error (MAE), which evaluated the average absolute difference between the predicted and experimentally determined angles [178]. To account for the periodicity of an angle, the smaller value of the absolute difference d_i and $360 - d_i$ was used. The coverage of a fragment library was assessed by the proportion of residues present in at least one fragment in the library. The precision of a fragment library was defined by the fraction of TP fragments. All fragments with an RMSD of < 1.5Å were considered TP else FP. The equation used to calculated the precision score is Eq. 2.4. The RMSD value, as calculated by FLIB-COEVO [177], was computed between the aligned residues of the corresponding crystal structure and the fragment. The number of satisfied contact pairs in each fragment was calculated by scoring the number of TP contact pairs by using a contact's residue indices according to the sequence alignment provided by FLIB-COEVO. MR success for each search model was solely assessed by SHELXE scores, whereby a CC score of ≥ 25.0 combined with an ACL score of ≥ 10.0 was required.

7.3 Results

In this study, the main objective was to determine if peptide fragments derived from unrelated protein structures in the PDB could be reliably identified and trialled in MR to achieve structure solutions. The fragment picking algorithm FLIB-COEVO [177] was used to pick fragments given its novel approach of validating selected fragments against a set of predicted residue-residue contacts.

7.3.1 Precision of FLIB-COEVO input data

The FLIB-COEVO algorithm requires two sets of input data — the predicted secondary structure and per-residue torsion angles — for each target sequence alongside an optional third source of information in form of coevolution data. The first part of the

analysis in this study focused on these data given that the FLIB-COEVO fragment picking heavily relies on the individual features in the selection and scoring of each individual fragment [177]. Poor data at this stage could lead to poor fragments that would be unsuitable for MR trials given that high accuracy, i.e. a low RMSD value between the search model and target, is required [165].

The secondary structure prediction highlighted high precision between each target's prediction and the DSSP-assigned [157] secondary structure of the target reference structure (Fig. 7.1). The three targets with PDB identifiers 1aba, 1lo7 and 1u06 had secondary structure predictions with a precision of $> 89\%$. The fourth target, 5nfc, showed comparatively poor precision of 50.7% over all residues in the PSIPRED prediction and the DSSP assignment using the reference crystal structure. However, 11 out of 13 secondary structure features were correctly predicted with deviations primarily found in the flanking residues of each secondary structure feature.

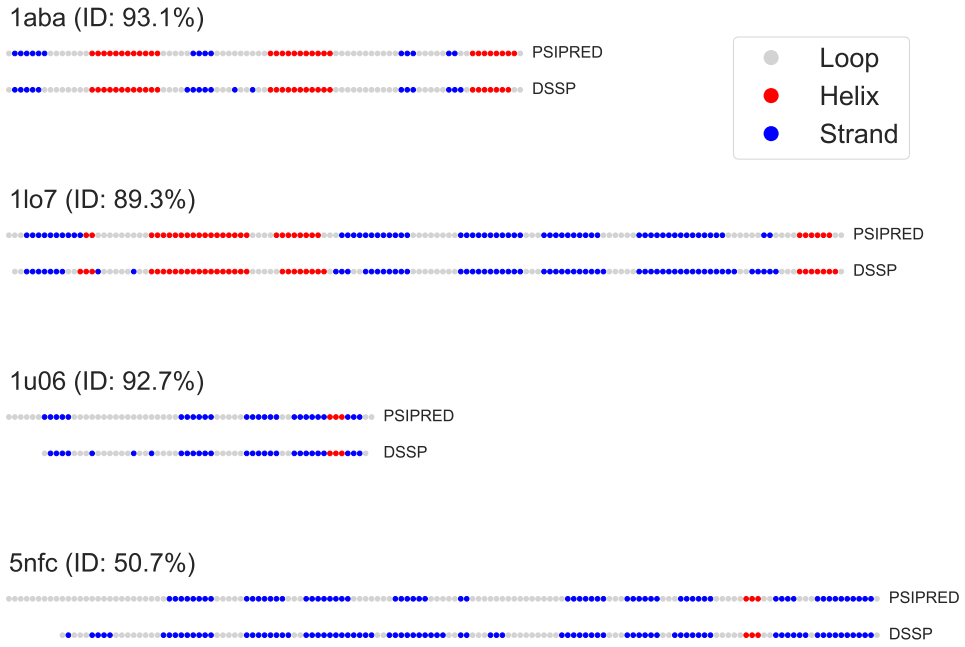


Figure 7.1: Schematic comparison of PSIPRED [147] secondary structure prediction and DSSP [157] assignment. Percentage identity is provided next to each identifier. The identity was computing using the Hamming distance over all positions present in the target sequence and reference structure.

The contact prediction data for METAPSICOV STAGE1 and STAGE2 predictions demonstrated the high precision achievable by this algorithm (Table 7.1). In this study, the top contact pairs at cutoffs L and $L/2$ were provided to the FLIB-COEVO algorithm. All targets had precision scores for both sets of predictions at both cutoff levels of more at least 0.6 (Table 7.1). A comparison of the sets of predicted contact pairs showed that only every third (for $L/2$ contacts) or every other (for L contacts)

contact pair is shared between both METAPSICOV predictions highlighting the importance of trialling both when selecting FLIB-COEVO fragments (Jaccard index in Table 7.1).

Table 7.1: Precision scores for METAPSICOV [101] STAGE1 and STAGE2 contact predictions. Jaccard index calculated for the same L-dependent selection of contact pairs between METAPSICOV STAGE1 and STAGE2 predictions.

Target	$L/2$ **contact pairs**			L **contact pairs**		
	$\text{Prec}_{\text{STAGE1}}$	$\text{Prec}_{\text{STAGE2}}$	Jaccard	$\text{Prec}_{\text{STAGE1}}$	$\text{Prec}_{\text{STAGE2}}$	Jaccard
1aba	0.884	0.884	0.303	0.713	0.759	0.513
1lo7	0.857	0.957	0.308	0.738	0.837	0.446
1u06	0.839	0.806	0.378	0.710	0.787	0.459
5nfc	0.822	0.836	0.327	0.619	0.762	0.434

Given the two METAPSICOV contact prediction files, both showed localised clusters of contact pairs characteristic for secondary structure features (Fig. 7.2). These clusters were more populated with contact pairs in METAPSICOV STAGE2 predictions. This behaviour is to-be-expected given that the second stage in METAPSICOV screens the first to remove singleton contact pairs whilst enriching the already existing clusters [101]. Besides the visual analysis, a cluster determination study on each of those contact maps further confirmed a higher singleton frequency in METAPSICOV STAGE1 predictions. The latter contained on average 9% more singleton contact pairs, and thus a higher degree of noise.

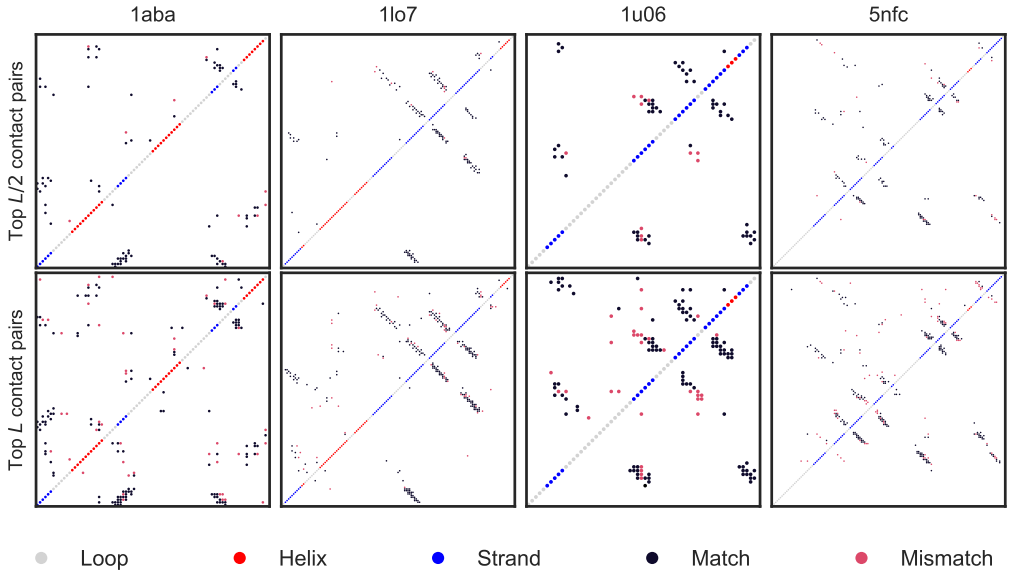


Figure 7.2: Comparison of $L/2$ and L correctly and incorrectly predicted contact pairs for four FLIB-COEVO targets. Contacts were predicted using METAPSICOV [101] STAGE1 (top left) and STAGE2 (bottom right). TP and FP contact pairs were identified using an 8Å cutoff between Cα (Cβ in case of Gly) atoms of a reference crystal structure. PSIPRED [147] secondary structure prediction provided along the diagonal.

An analysis of the MAE of torsion angles between the SPIDER2 [178] prediction and a corresponding reference crystal structure highlighted accurate predictions for three of four targets (Fig. 7.3). The largest MAE_{φ} across the four target sequences was 24.347°, and the largest MAE_{ψ} was 45.459° (MAE values for PDB entry 1u06). The smallest MAE_{φ} was 13.822° (PDB ID: 1aba) and smallest MAE_{ψ} was 17.273° (PDB ID: 1lo7). Segments in sequence space with regular secondary structure, as predicted by PSIPRED [147], resulted primarily in low MAE values of torsion angles. In contrast, unstructured regions highlighted much larger MAE values indicating the difficulty of predicting these regions. Noticeably, the MAE_{ψ} appeared to be much larger in those regions than the MAE_{φ} for the same residue.

In summary, all target sequences had FLIB-COEVO input data of good quality, which should allow FLIB-COEVO to select fragments of suitable accuracy for MR.

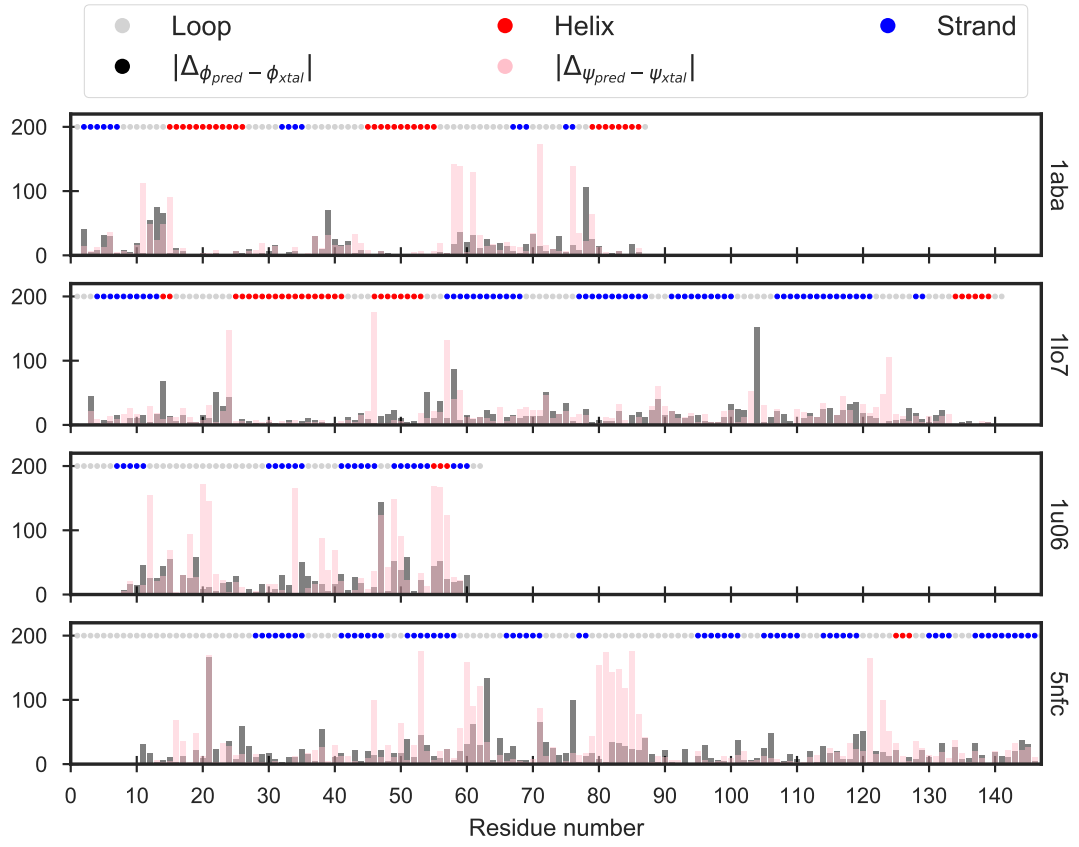


Figure 7.3: Comparison of MAE of torsion angles predicted by SPIDER2 [178] and extracted from a corresponding PDB structure. PSIPRED [147] secondary structure prediction provided alongside the MAE values.

7.3.2 FLIB-COEVO fragment picking

Sixteen FLIB-COEVO fragment libraries were created for each protein target in this study. Each fragment library consisted of one combination of one of two contact pre-

diction files and altering input parameters.

Across all four targets, the FLIB-COEVO algorithm selected a total of 8,535,458 fragments (Table 7.2). The fragment libraries showed similar statistics across the four protein targets despite the diversity in fold and chain lengths. The mean FLIB-COEVO score was 3,200 score units with a mean RMSD of 9.00Å. Fragments for the alpha-spectrin SH3 domain (PDB ID: 1u06) scored the lowest mean FLIB-COEVO score with 3,034 units; however, the same target scored the worst by mean RMSD with an average of 9.47Å. In contrast, fragments picked for the sequence of the bacteriophage T4 glutaredoxin (PDB ID: 1aba) achieved the best mean RMSD of 7.85Å given the second highest mean FLIB-COEVO score of 3,217 units (Table 7.2).

Table 7.2: Summary of fragment statistics for FLIB-COEVO libraries selected for four protein targets. Count_{H} corresponds to the count of fragments extracted from homologs.

Target	Count	Count_{H}	FLIB-COEVO score			RMSD		
			Median	Mean	Std Dev	Median	Mean	Std Dev
1aba	2,091,321	45,133	3,061	3,217	1,405	7.70	7.85	3.81
1lo7	2,497,813	23,396	3,187	3,371	1,497	9.00	9.43	4.61
1u06	1,133,517	60,159	2,901	3,034	1,306	9.51	9.47	3.94
5nfc	2,812,807	48,828	2,982	3,127	1,316	8.89	9.16	4.18
Total	8,535,458	177,516	3,049	3,208	1,397	8.68	8.96	4.25

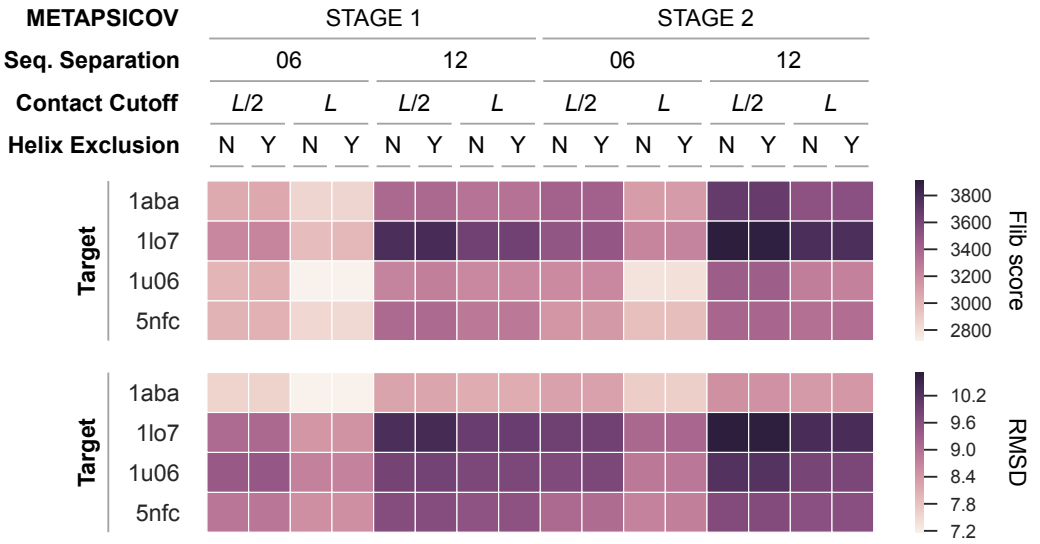


Figure 7.4: FLIB-COEVO fragment library comparison for four targets highlighting the differences in mean FLIB-COEVO score and RMSD by starting with different subsets of contact predictions. L refers to the number of residues per target sequence. Y refers to idealised α-helical fragment exclusion during fragment picking; N refers to treating those fragments like all others.

A split of the per-target fragment libraries by input options highlighted the better fragment library quality under certain conditions with regards to the mean FLIB-

COEVO score and RMSD (Fig. 7.4). In particular, top-L (6 residues sequence separation) METAPSICOV STAGE1 contact predictions yielded the lowest for both metrics across all targets. A comparison of the sequence separation, i.e. using all contact pairs or medium- and long-range ones only, strongly suggested much lower and thus more favourable scores for using short-, medium- and long-range contact pairs. A very similar difference was noticeable for METAPSICOV STAGE2 contact predictions (Fig. 7.4).

In this study, predicted contact information was used to further guide fragments selection. The FLIB-COEVO algorithm only selected fragment for positions of the target sequence with at least one contact pair. Given this scenario, an analysis of the coverage of the target sequence with respect to each picking strategy further demonstrated the benefits of starting with METAPSICOV STAGE1, i.e. noisier contact predictions (Fig. 7.5). Coverage was more evenly spread across the target sequences compared to missing regions especially for target 4-hydroxybenzoyl CoA thioesterase (PDB ID: 1lo7) when starting with METAPSICOV STAGE2 predictions. Noticeably, none of the picking strategies yielded any fragments for the C-termini of α-spectrin SH3 domain (PDB ID: 1u06) and galectin-3 CRD (PDB ID: 5nfc) (Fig. 7.5). Furthermore, an analysis of the precision of fragments in each library strongly supported the benefits of starting with top-L (6 residues sequence separation) METAPSICOV STAGE1 contact pairs. Across all four targets, the coverage of correct fragments (classed by RMSD < 1.5Å to the reference structure) was highest for this condition. This is of particular importance for α-spectrin SH3 domain (PDB ID: 1u06) and galectin-3 CRD (PDB ID: 5nfc), for which most strategies picked very few to no correct fragments. Excluding idealised α-helical fragments did not affect the quality of the FLIB-COEVO libraries greatly. A consideration of differences in mean FLIB-COEVO and RMSD scores showed differences of 25.68 and 0.06 between comparable libraries, i.e. with and without idealised α-helical fragments.

Given that FLIB-COEVO used coevolution data to help select fragments, it was little surprise that higher degrees of TP fragments colocalise with high-density contact pair regions along the target sequence (Fig. 7.5). This characteristic explained less TP fragments in top-$L/2$ fragment libraries because less contacts (compared to top-L) were available during fragment selection. The resulting selection was purely based on the FLIB-COEVO score which might not yield high-accuracy fragments (RMSD < 1Å) as frequently. Therefore, the co-localisation of TP FLIB-COEVO fragments and regions of high-density contact predictions highlighted the importance of adding this additional source of information to pick fragments.

Figure 7.5: Summary of the coverage and precision of FLIB-COEVO fragment libraries according to their target sequence. The coverage of all fragments with respect to their target-aligned sequence register are shown in red bars, and fragments with RMSD < 1.5Å to the reference structure in blue. The predicted secondary structure of each target sequence is given at the top: α-helices (red), β-strands (blue), and loops (grey). Contact prediction information is illustrated using black bars. The fragment frequency is shown using a log-scale.

7.3.3 FLIB-COEVO fragment selection for Molecular Replacement

One of the most challenging and important aspects of bypassing *ab initio* protein structure prediction to use the picked fragments directly as MR search models is the accurate identification of fragments with the highest similarity between fragment and target structure.

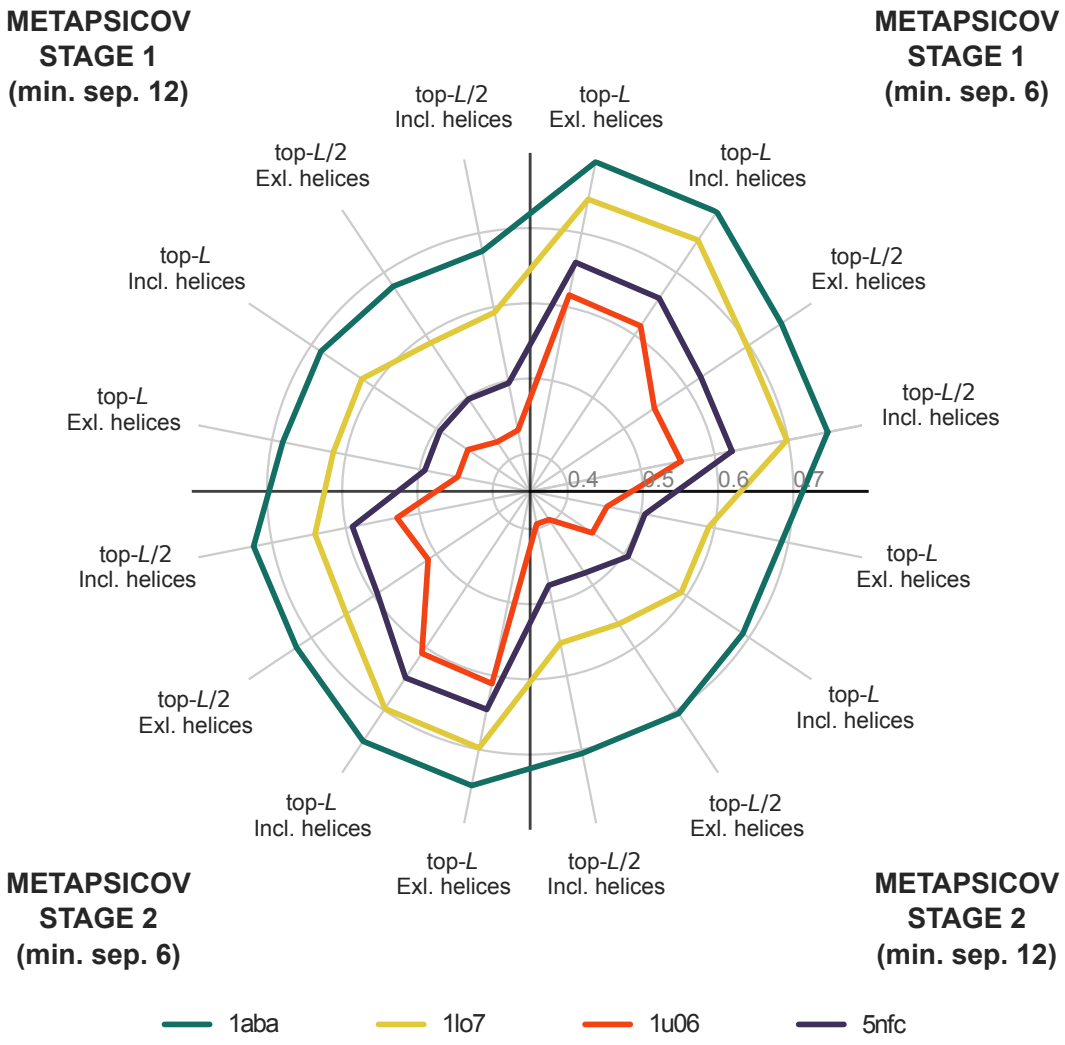


Figure 7.6: Spearman rank-order correlation coefficient analysis of FLIB-COEVO fragments' FLIB-COEVO score and RMSD value given the 16 unique fragment picking strategies across four targets. P-values of all Spearman correlations are < 0.001 and not shown for simplicity of the plot.

A fragment's FLIB-COEVO score — its cumulative absolute error of predicted torsion angles — has the highest correlation with the RMSD of a fragment compared to all other scores used in the FLIB-COEVO protocol [177]. To validate this finding, all non-homologous fragments in this study were tested for a correlation between their FLIB-COEVO scores and RMSD values. The Spearman's rank-order correlation coefficient analysis confirmed the correlation between a fragment's FLIB-COEVO

and RMSD scores (Fig. 7.6). However, the strength of the correlation varied greatly between different fragment libraries and targets. The optimal fragment picking strategy — top-L (6 residues sequence separation) METAPSICOV STAGE1 — resulted in the strongest correlations across all targets. The same contact pair selection with METAPSICOV STAGE2 predictions results in the second strongest correlation. Noticeably, the bacteriophage T4 glutaredoxin (PDB ID: 1aba) fragment libraries showed stronger positive correlations than the remaining targets. The fragments selected for α-spectrin SH3 domain (PDB ID: 1u06) showed the overall weakest correlations. It is worth noting that the two targets (PDB IDs: 1aba & 1lo7) were classed as mixed α-β targets, and thus the strength of this correlation might be fold-dependent.

Further inspection of the fragments and the relationship between each fragment's FLIB-COEVO score and RMSD value revealed a small subset of outliers in each fragment library. These fragments (hereafter referred to as outlier fragments) were sparse in each library with an overall mean count of less than 0.2%. An analysis for unique characteristics of these outliers, which would allow for their exclusion, revealed no unique feature. These fragments contained all secondary structure types, spanned across all target sequences and ranged over all peptide lengths. Furthermore, they occurred in all fragment libraries, irrelevant of their original picking strategy. The only characteristic setting these outlier fragments apart from the remaining set was a RMSD value of > 30Å. Nevertheless, it appeared that these outlier fragments with unusually high RMSD values were never included in the final fragment search model set, given that their overall FLIB-COEVO$_{\text{min}}$ score was 796 units (one order of magnitude more than the overall minimum for the remaining fragments).

An analysis of the fragment metrics in the final MR set (6,547 fragments) further supported the positively linear relationship between a fragment's FLIB-COEVO score and RMSD (Fig. 7.7a). However, the best FLIB-COEVO fragments by RMSD showed much less spread compared to the best fragments by FLIB-COEVO score (Fig. 7.7b). Furthermore, the size of the fragments also positively correlated with the FLIB-COEVO ($\rho_{Spearman} = 0.860$, $p < 0.001$) and RMSD ($\rho_{Spearman} = 0.697$, $p < 0.001$) values. Longer fragments with higher dissimilarity with respect to the target showed higher FLIB-COEVO scores and RMSD values (Fig. 7.7a).

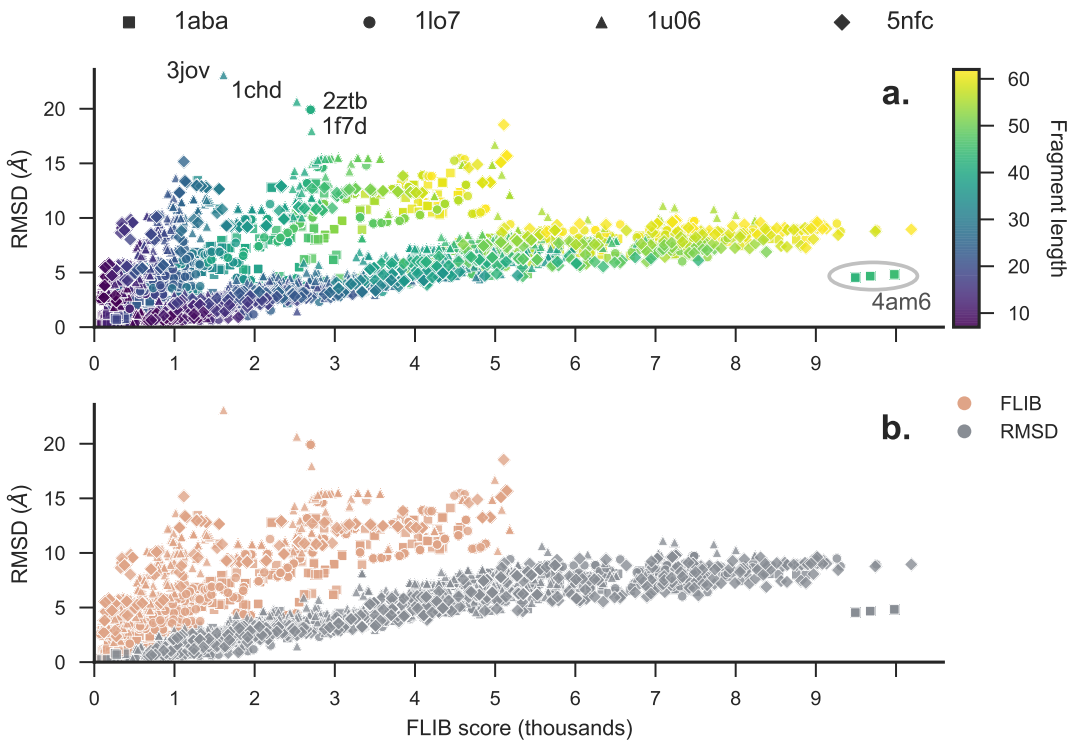


Figure 7.7: Scatter plot highlighting the positive correlation between fragment FLIB-COEVO scores and RMSD values. The plot contains all fragments independent of target or picking strategy. **a.** The colour of each scatter point illustrates the fragment length. All extreme outlier fragments are highlighted with their PDB identifiers as labels. **b.** The colour codes indicate the sorting strategy to select the top FLIB-COEVO fragments for each fragment peptide length bin.

Notably, a cluster of large fragments with some of the highest FLIB-COEVO scores in the set showed a reasonable similarity to their target structure (Fig. 7.7a). All fragments in this cluster were picked for the bacteriophage T4 glutaredoxin sequence (PDB ID: 1aba) and extracted from the same region of the crystal structure of the actin-related protein ARP8 (PDB ID: 4am6). In comparison, some smaller fragments with peptide lengths less than 50 residues and lower FLIB-COEVO scores of less than 3000 showed the highest RMSD values in the final set.

One further unique aspect of this study compared to other fragment-MR approaches was the use of predicted residue-residue contact information to select fragments during picking, only selecting fragments for target-sequence residues with at least one contact pair in the predicted set (Saulo de Oliveira, personal communication). In the final set, 39% of all fragments satisfied at least one, 26% at least two and 20% at least three contact pairs. Across the four targets, 50% of all fragments selected for 4-hydroxybenzoyl CoA thioesterase (PDB ID: 1lo7) satisfied at least one predicted contact pair (Fig. 7.8). In comparison, 28% of fragments selected for the α-spectrin SH3 domain (PDB ID: 1u06) satisfied at least one contact pair.

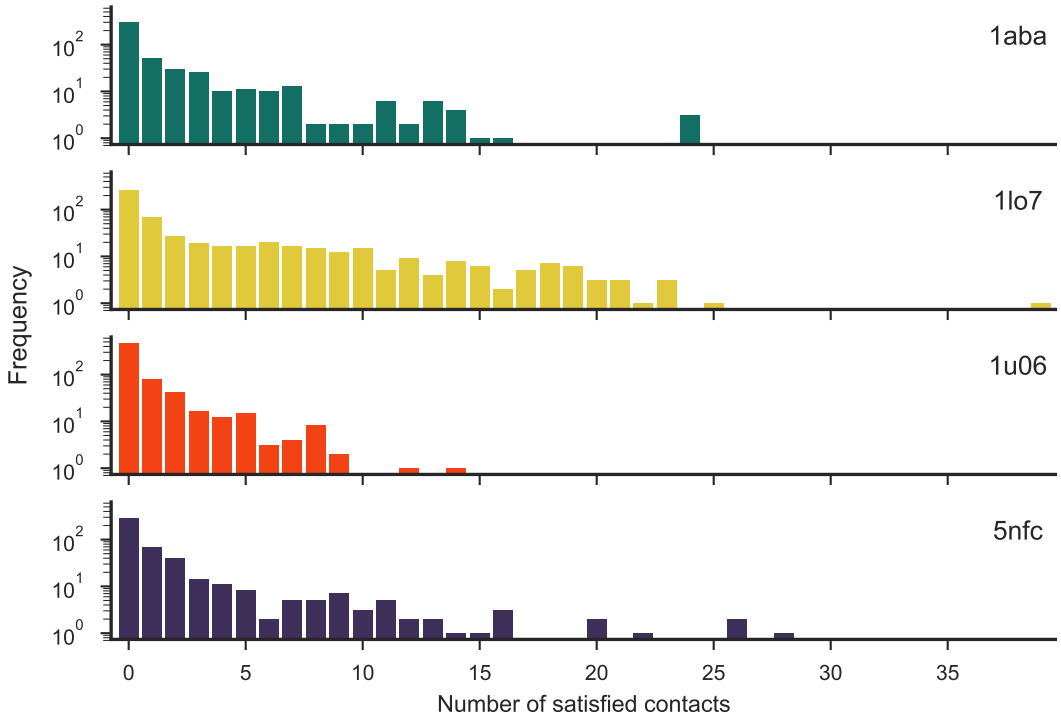


Figure 7.8: Distribution of contact precision for FLIB-COEVO fragments selected as MR search models separated on a per-target basis.

Thus, the final set of FLIB-COEVO fragment MR search models spanned a wide range of peptide lengths, RMSD values, predicted contact satisfaction scores, and generally secondary structure make-up. To illustrate the latter, a random selection of sample fragments is illustrated in Fig. 7.9. Importantly, not a single super-secondary structure motif dominated the set, increasing the sampling diversity to be undertaken during MR.

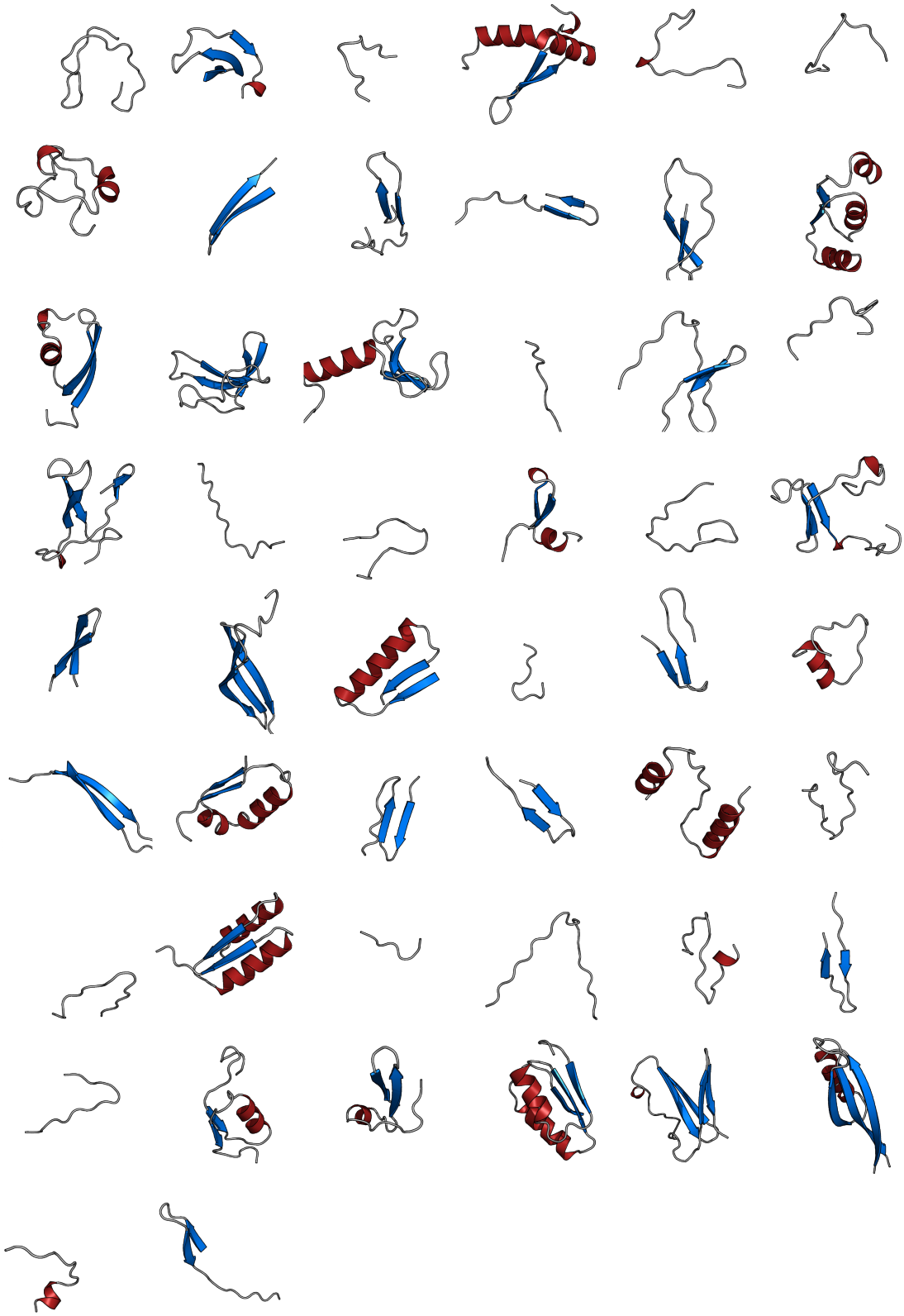


Figure 7.9: Non-redundant sample of FLIB-COEVO fragment search models selected for four different protein targets. Secondary structure defined by and visualisation done in PyMOL [183]. Unpaired β-strands rendered using the loop style.

7.3.4 Molecular Replacement using FLIB-COEVO fragments

FLIB-COEVO fragments picked for four target sequences using a variety of FLIB-COEVO input options generated more than 6,500 fragments, which were subjected to the MR pipeline MRBUMP with their corresponding target experimental data. Given that each fragment was trialled with three different PHASER RMSD values, a total of 19,716 MR attempts were made across four target structures. Out of nearly 20,000 MR attempts, 299 led to the successful structure solution of two targets, namely the T4 glutaredoxin (PDB ID: 1aba) and α-spectrin SH3 domain (PDB ID: 1u06) (Fig. 7.10).

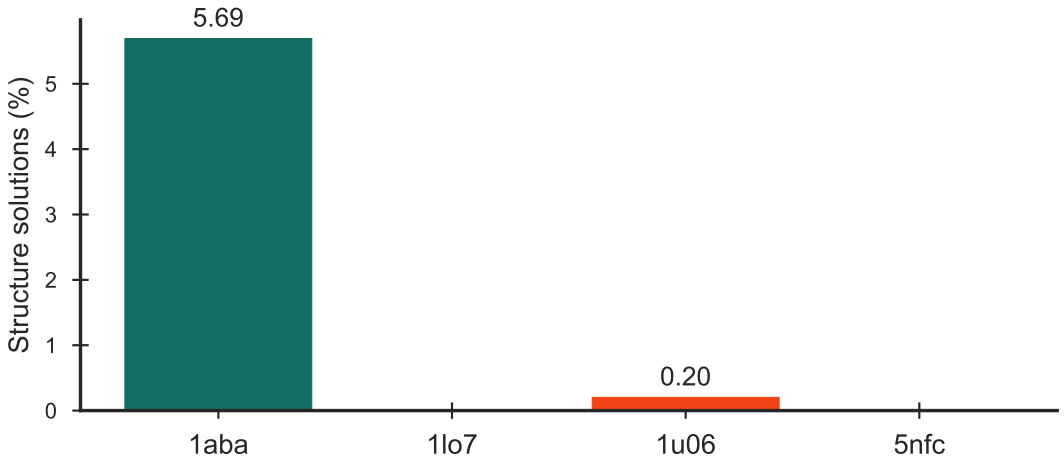


Figure 7.10: Distribution of structure solutions by FLIB-COEVO target. All MR attempts total to 19,716, out of which 299 are structure solutions. Values above each bar indicates percentage search models successful out of the corresponding set.

The total of 299 MR structure solutions were achieved by 70 sequence-unique fragments. Sixty-nine of those fragments were picked from 60 unique structures for the T4 glutaredoxin (PDB ID: 1aba) leading to 97% of all structure solutions. In comparison, a single fragment, selected from three different fragment libraries, led to nine structure solutions of the α-spectrin SH3 domain (PDB ID: 1u06). The largest FLIB-COEVO fragment leading to a structure solution contained 37 residues and the smallest ten.

A division of FLIB-COEVO-fragment search models by their respective origin libraries provided strong evidence that METAPSICOV STAGE1 contact predictions allows for the selection of the most accurate fragments (Fig. 7.4), which directly translated into the structure solution count (Fig. 7.11). Furthermore, this division also highlighted and supported the quality of fragment libraries picked with top-L (6 residues sequence separation) METAPSICOV STAGE1 predictions. Trialling the optimal fragment picking strategy with and without helical fragments ($>$ 90% α-helical content assigned using DSSP) resulted in the library without outperforming the other (Fig. 7.11, 3$^{\text{rd}}$ and 4$^{\text{th}}$ bars).

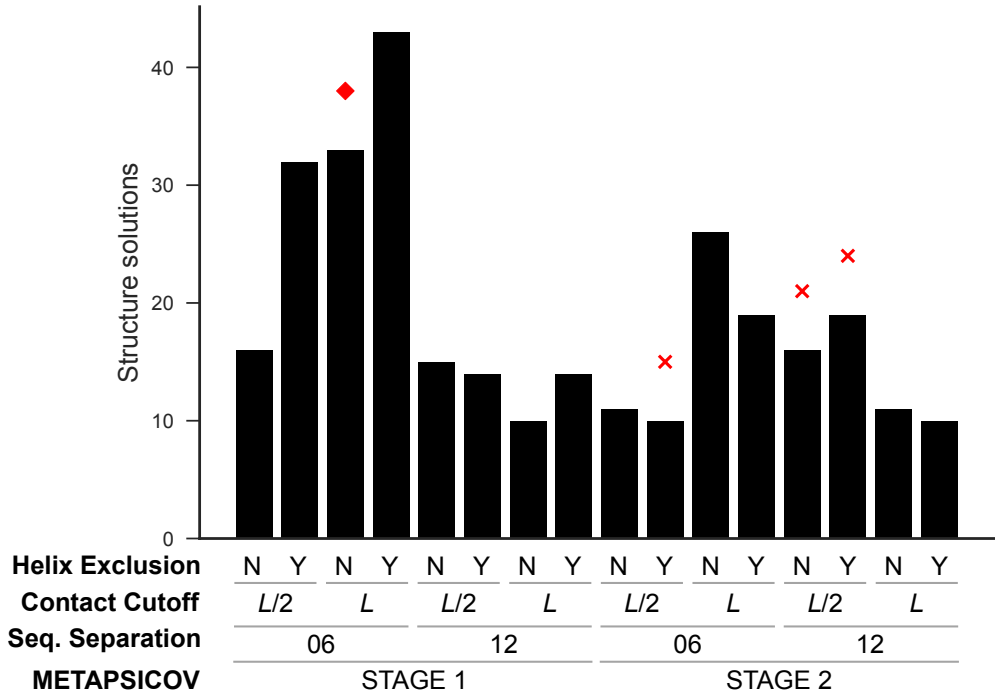


Figure 7.11: Distribution of structure solutions by FLIB-COEVO library configuration. The optimal fragment picking strategy, as assessed by FLIB-COEVO values, is highlighted with a red diamond to illustrate that the method that picks the best fragments is close to, but not the absolute best for ultimate structure solution. Fragment picking strategies leading to solutions of α-spectrin SH3 domain (PDB ID: 1u06) are highlighted with red crosses.

An analysis of the binned results by fragment-ranking or PHASER RMSD value confirmed the expected outcome: the top fragments selected by fragment RMSD score result in more structure solutions than their FLIB-COEVO score counterparts (Fig. 7.12). To reiterate, all FLIB-COEVO fragments were grouped by their peptide length, and the top fragment in each group selected when sorted by either FLIB-COEVO or RMSD values. When separating the total number of structure solutions by the score that made each fragment the best in its original library, it became clear that two-thirds of solutions were achieved with fragments scoring best by RMSD. However, the structure of α-spectrin SH3 domain (PDB ID: 1u06) was only solved with fragments that scored best in their FLIB-COEVO fragment libraries by FLIB-COEVO score. A further subdivision of successful fragments, sorted either by FLIB-COEVO scores or RMSD values, highlighted that a larger proportion of successful RMSD-sorted fragments satisfied at least one predicted contact (FLIB-COEVO-sorted: 7%; RMSD-sorted: 13%). A separation of attempts by PHASER RMSD value suggested a value of 0.1 to be the most favourable.

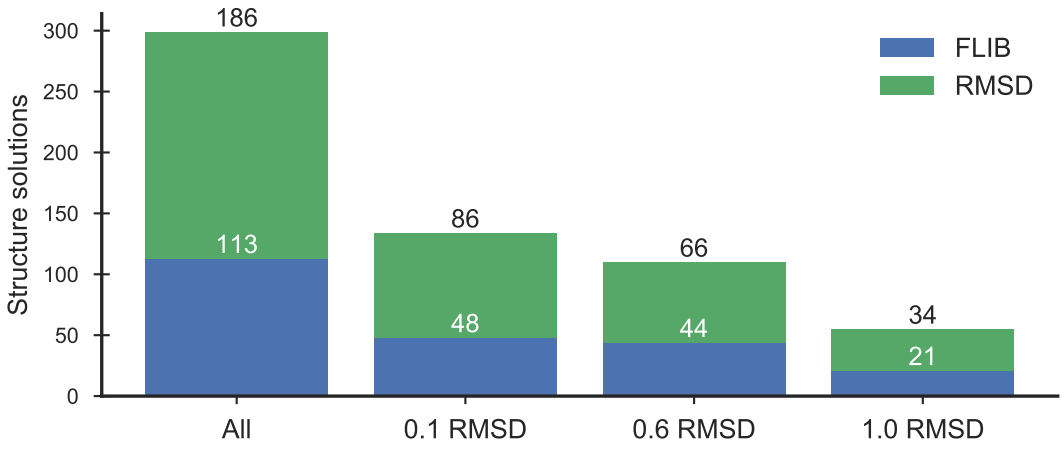


Figure 7.12: Distribution of structure solutions by fragment and MRBUMP configuration. The structure solution count is provided above each bar.

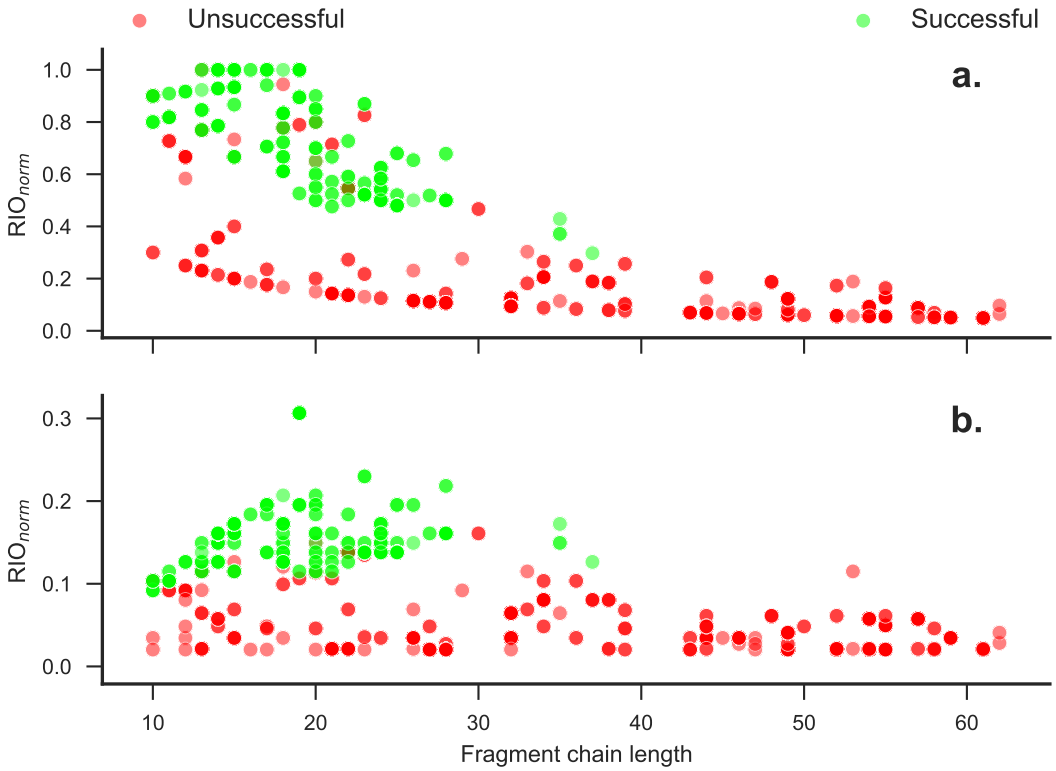


Figure 7.13: Dependence of normalised Residue-Independent Overlap (RIO_{norm}) score on the fragment chain length. The two plots show RIO scores normalised by the chain lengths of (a) the fragment and (b) the target. Colour coding indicates if the FLIB-COEVO-fragment search model resulted in a structure solution. Each plot contains 890 fragment points; however, not all points are visible due to the superposition of individual scatter points because the same fragment was scored under different MR conditions.

In MR, the correct placement of very small structural fragment may not always be detectable by inspecting output metrics of underlying software. In benchmarking

exercises, the RIO metric has shown to be a very useful and powerful metric to detect such situations [117, 118]. Given that the peptide lengths of FLIB-COEVO fragments in this study ranged from six to 63 residues, the RIO score was most suitable in validating the correct placement of FLIB-COEVO-fragment search models. Indeed, all fragments with SHELXE CC $\geq$ 25 and ACL $\geq$ 10 contained at least three correctly placed Cα atoms (i.e. a RIO score $\geq$ 3). Furthermore, the RIO metric indicated that more than 500 fragments had Cα atoms placed within 1.5Å of any atom in the target structure. However, only four residues were on average placed correctly, which was not enough to achieve structure solution (Fig. 7.13). All successful FLIB-COEVO fragments had a minimum model- and target-normalised RIO scores of 29.7% and 9.2% (Fig. 7.13, green markers).

In 33 MR attempts more than 60% of a fragment's residues were placed correctly, yet structure solution was not achieved. These trials affected exclusively fragments picked for the target sequences of T4 glutaredoxin (PDB ID: 1aba) and 4-hydroxybenzoyl CoA thioesterase (PDB ID: 1lo7). Overall, the 33 MR attempts made were done with 17 fragments extracted from 15 templates containing between ten and 23 amino acids. The fragments' RMSD values ranged from 0.19 to 2.72Å with a mean RMSD of 1.10Å. Surprisingly, almost all of these fragments contained primarily α-helices. Given the presence of helices in the fold of both targets (Fig. 7.1) and the success of idealised fragments to solve such targets with data resolution better than 2.0Å, it came as a surprise to not see more structure solutions from these fragments.

Finally, the coevolution data used in this study to select fragments was a novelty in the field. Thus, it was of great interest to identify if fragments leading to structure solution satisfied many predicted residue-residue contacts. Eighty-seven percent (n = 61) of all unique fragments leading to structure solutions for either target satisfied no predicted residue-residue contact. The remaining nine fragments, all of which led to structure solutions of T4 glutaredoxin (PDB ID: 1aba), satisfied either one ($n = 4$), two ($n = 4$) or 24 ($n = 1$) predicted contacts.

The fragment with 24 satisfied contacts is a particularly striking example of the value of the approach explored in this study (Fig. 7.14). The fragment was derived from the template structure of cobalt chelatase found in *Salmonella typhimurium* (PDB ID: 1qgo). The picked fragment contained 35 residues, its supersecondary structure consisted of a two-strand β-sheet packing against a single α-helix, and its FLIB-COEVO-calculated RMSD to the target is 3.39Å. The majority of satisfied contact pairs were between Cβ atoms of the β-strands; however, a small number of individual contact pairs also identified the packing of one β-strand against the α-helix (Fig. 7.14, top-right). Although not considered at this stage in the FLIB-COEVO algorithm, this particular fragment satisfied 75% of all top-predicted contact pairs. Most importantly though, this fragment was derived from an entirely unrelated protein structure, and thus illustrated the value in *ab initio* structure prediction fragments as MR search models.

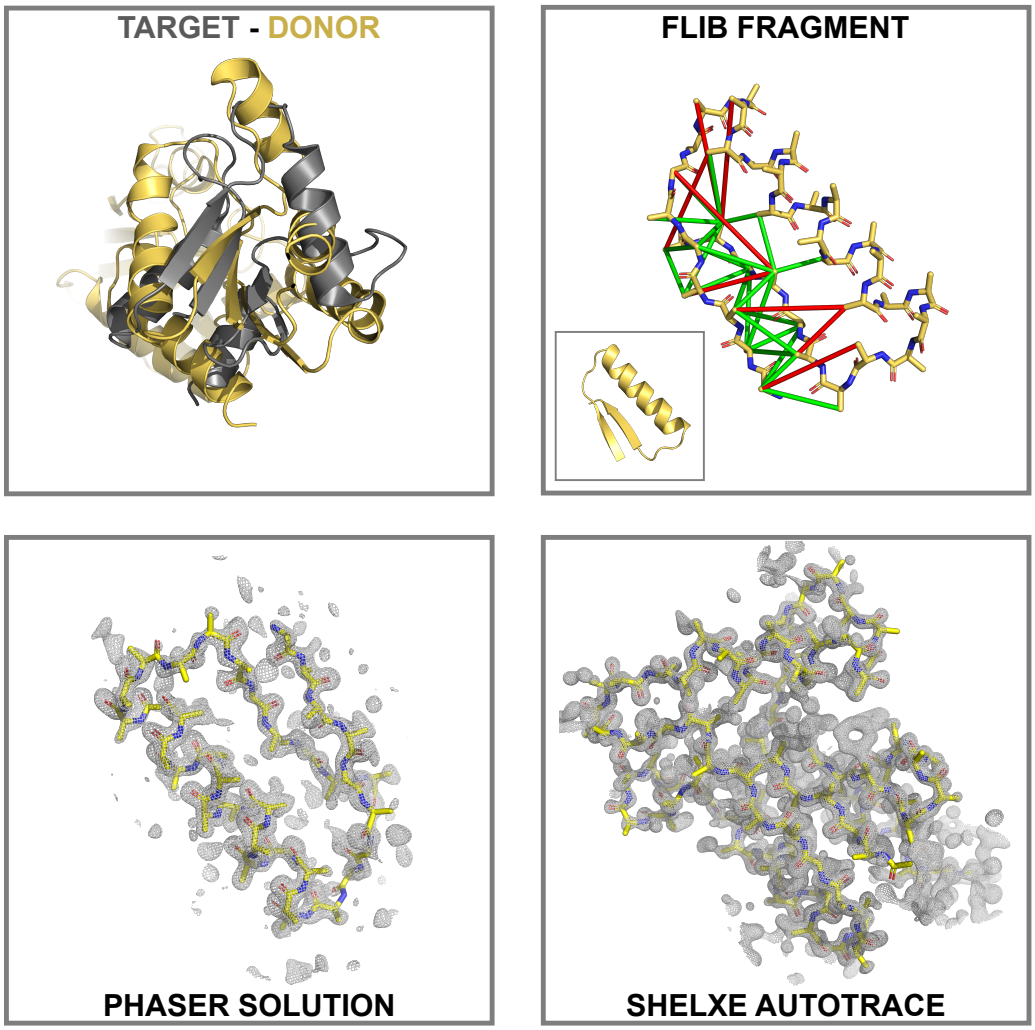


Figure 7.14: Intermediary steps from donor structure to SHELXE main-chain autotrace for a fragment derived from cobalt chelatase found in *Salmonella typhimurium* (PDB ID: 1qgo). The structure solution was obtained against the target crystallographic data of T4 glutaredoxin (PDB ID: 1aba). METAPSICOV STAGE2 predicted contacts, against which the fragment was selected, are illustrated with True Positive (green) and False Positive (red) contacts (distance cutoff of 8Å). 2mFo-DFc electron density maps shown at 2.0 sigma and radius around the peptide atoms of 5Å. The RMSD between the sequence-independently superposed structures of target and donor is 10.384Å (computed with the `super` command in PyMOL [183]).

7.4 Discussion

The main objective of this study was to investigate the application of FLIB-COEVO structural fragments to MR. Four experimental datasets were chosen and 16 FLIB-COEVO fragment libraries built per target sequence varying primarily in the predicted residue-residue contact information. A selection of highest scoring fragments were then forwarded to MRBUMP to trial each fragment as MR search model. The findings in

this study validated the concept of this approach. Firstly, a positive correlation between a fragment's FLIB-COEVO score and RMSD value was identified. These correlations were target-independent and found, with various strengths, in all FLIB-COEVO fragment libraries. Furthermore, this work identified top-L (6 residue sequence separation) METAPSICOV STAGE1 contact pairs to be the optimal selection of contact pairs for the FLIB-COEVO algorithm when starting with METAPSICOV predictions. The additional noise, typically filtered in the second stage of the METAPSICOV algorithm [101], allowed for the selection of more accurate fragments across the entire target sequence. Lastly, trialling a selection of high-scoring FLIB-COEVO fragments in routine MR showed the usefulness of such fragments in attempting to solve protein structures. Two out of four targets were successfully solved despite only trialling a small proportion of FLIB-COEVO fragments per library (mean MRBUMP runtime of 10.5 CPU hours per fragment).

Intuitively, most crystallographers would declare the limitations of this approach to be the size and quality of the selected FLIB-COEVO fragments as well as the resolution of the crystallographic data. Although the former was long-thought to be a major limitation, more recent work highlighted the success of likelihood-based MR methods (i.e., PHASER [128]) with very small search models. McCoy et al. [172] demonstrated the successful *ab initio* MR structure solution of aldose reductase starting from as little as two correctly placed atoms. Furthermore, automated MR pipelines, such as AMPLE [114], ARCIMBOLDO [155], BORGES [156], FRAGON [164] or FRAP [184], also successfully demonstrated MR successes with search models comprising a fraction of the target structure. Thus, MR structure solutions with FLIB-COEVO fragments as short as six residues should be considered possible, especially when high resolution data is available and the fragment size is proportionally large compared to target size.

MR search models need to be sufficiently accurate to derive phase information for successful structure solution. The findings in this study highlighted the success of identifying accurate fragments solely by the fragment's FLIB-COEVO score. Given that the FLIB-COEVO implementation used in this study only selected fragments for positions with at least one available contact pair, future research is required to identify the potential benefits of specifically selecting fragments that satisfy at least one contact pair. Furthermore, it is important to understand the potentially beneficial implications of using the contact satisfaction score in the FLIB-COEVO scoring metric of a given fragment. In theory, higher precision scores should imply a closer match of the overall tertiary structure of the trialled region. Alternatively, selecting secondary structure motifs or substructures of templates by means of searching with a predicted contact map could be an attractive alternative. Recent studies indicated success in identifying sub-folds by means of Contact Map Overlap (CMO) [74, 185]. Further work also needs to explore the benefits of considering the expected Log-Likelihood Gain (eLLG) as a conceptual framework to identify the linked effects of the fragment search model size, its accuracy and the resolution on the likelihood of the solution of a target structure

[172].

Nevertheless, FLIB-COEVO fragments with near-identical subfolds to the target might not be traceable by current means of assessing structure solutions. Commonly, automatic MR attempts and their successes are judged by the combination of SHELXE CC, ACL scores and R-values [130]. However, it is known that β-strands are notoriously difficult to trace, and thus SHELXE might not pick up on correctly placed search models, as it was observed for target 1e0s in Section 3.3.3.1. Although this study did not suffer from this problem for fragments containing primarily β-strands, it did have correctly placed α-helices without structure solutions. Thus, the approach taken in this study would benefit from improvements to the density modification and sequence tracing algorithm SHELXE.

Finally, this work served primarily as proof-of-concept study, and thus attempted to explore a diversity of options. With a better understanding of input parameters future work could build on the work presented here and use a large-scale analysis to assess the suitability of this concept more thoroughly. Furthermore, improvements to the FLIB-COEVO algorithm through the incorporation of coevolution data should also improve the quality of *ab initio* structure predictions, which should result in a greater success rate of other MR pipelines, such as AMPLE [114].

Chapter 8

Conclusion & Outlook

8.1 Conclusion

The successful disentanglement of direct and indirect residue contacts in contact prediction revolutionised many aspects of Structural Bioinformatics research [84]. Successful applications of predicted contact information range from accurately defining domain boundaries [170] to identifying druggable protein-protein interfaces [186]. Although many such applications have been highlighted over the last few years [84], few concerned the topic of MR in X-ray crystallography. In Chapters 3 to 7, work was presented that made the first attempts to apply predicted contact information to explore some of its applications in MR.

The use of contact prediction in *ab initio* protein structure prediction allowed researchers to predict the structure of many previously unknown protein folds based on their sequence alone [e.g., 45, 46, 69–75]. The major benefit of adding such information was to reduce the conformational search space, which allowed more challenging folds to be sampled correctly. Work presented in Chapters 3 to 5 further confirmed such findings. More importantly, the presented results highlighted that the modelling algorithm ROSETTA is very sensitive to the way contact predictions are introduced into its folding protocol. Two important examples included the up-weighting of β-strand contacts and the choice of energy function used to "reward" satisfied contacts. Furthermore, work in Chapter 5 highlighted that fragment-based structure prediction algorithms may no longer be essential for accurate structure prediction. CONFOLD2, a fragment-independent algorithm, predicts protein structures using secondary structure and contact information alone, which provided decoys of comparable accuracy to the state-of-the-art ROSETTA. Nevertheless, further research is required to establish the optimal routine to process CONFOLD2 decoys, since AMPLE's default routine cannot generate ensemble search models sufficient for MR solutions.

Beyond the prediction of protein structures, a major focus of the presented research centred on the benefit of improved structure predictions in unconventional MR. In line with prior expectations, better structure predictions yielded more MR structure solutions. In particular, previous weaknesses of the AMPLE approach — a target's chain length and fold — were partially overcome with contact-guided structure predictions. Some examples for which structure solutions were obtained exceed 200 residues in chain length, whilst many others contain large portions of β-structure. Nevertheless, simply adding contact predictions to *ab initio* protein structure prediction is not sufficient to solve all trialled targets. In part, this limitation resulted from a lack of precision of predicted contact information for some targets, since it depends significantly on the availability of divergent homologous sequences. Further research is also required to address new limitations in AMPLE resulting from suboptimal processing of much more native-like structure predictions. One approach, outlined in Chapter 6, explored the incorporation of contact information in the AMPLE processing pipeline to address the latter issue. Contact information was used to estimate the similarity

of a predicted decoy to its native structure by means of scoring its long-range contact satisfaction [45, 112, 132]. Exclusion of the worst decoys by this metric prior to clustering allowed more fine-grain sampling in AMPLE, which turned unsuccessful decoy sets into ones with which the native structure was solvable. However, key examples presented in Chapters 3 to 5 also highlighted the requirement for further developments in MR-related software to enable the automatic detection and subsequent processing of AMPLE ensemble search models, which were correctly placed but are undetectable as structure solutions by current metrics.

A further topic of research concerned the use of supersecondary structure elements or subfolds as MR search models. The default mode in AMPLE currently relies on computationally expensive *ab initio* structure predictions. Since contact predictions reached sufficient quality for protein families with many known sequences, such information could be used to identify matching subfolds in other, unrelated protein structures. In Chapter 7, a new hybrid approach demonstrated the successful implementation of such an idea. Although imperfect at this stage, several examples highlighted the successful identification of such subfolds and subsequently successful MR structure solution. Tied to this idea may also be recent research that attempts to identify subfolds by means of matching a predicted contact map to those extracted from protein structures [74, 185].

8.2 Outlook

In this thesis the first applications of predicted contact information in MR were presented. Despite the already promising results, this area of research is still in its infancy and a great number of potentially promising routes remain unexplored [84].

Earlier studies by Rigden [187] and Sadowski [170] demonstrated the successful application of predicted residue contacts to identify domain boundaries. Although unexplored to-date, precise domain boundary predictions could be applied for better domain boundary definitions prior to *ab initio* structure prediction to avoid sampling of terminal loops and linkers, and thus improve protein structure prediction quality. Furthermore, contact information was used to improve the AMPLE ensemble-generation pipeline with respect to identifying poorly predicted decoys. However, the AMPLE pipeline might additionally benefit from predicted contact information to drive the truncation procedure. For example, contact data could be used to rank individual residues by their contribution to a contact network, similar to Parente et al. [188], and truncation driven by the rank order or a hybrid score, which also includes the structural variance. Additionally, contact prediction might be used in the context of identifying alternative conformational states [99, 189–192]. AMPLE could exploit these to identify structurally conserved residues shared by both states, and thus truncate to this conserved core. Alternatively, AMPLE could attempt remodelling after successful disen-

tanglement of state-dependent contact pairs and try both conformations separately as ensemble search models. Simkovic et al. [84] outlined many further such applications of contact prediction in the field of Structural Biology. Ultimately, the precision of contact information improves daily with the increasing depth of sequence databases, thus enabling an ever-increasing number of applications with more precise outcomes. Furthermore, many more research groups start to identify the value in using predicted contact information in their own studies, and by means of pushing the boundaries new tools and applications are most likely going to emerge.

Despite the vast space of unexplored applications, predicted residue contacts with perfect precision may never solve all current or future challenges in unconventional MR. Despite the ability to limit the conformational space search in *ab initio* protein structure prediction greatly, sampling of larger protein targets will always remain difficult unless energy functions and force fields become true representations of all properties found *in vivo*. Furthermore, computational resources need to expand to allow many more sampling steps. Additionally, many protein targets exist in multiple conformations. Energy functions in fragment-based *ab initio* protein structure prediction may always favour one such conformation over all others, which may make conventional or unconventional MR very challenging.

Beyond limitations in Bioinformatics software to facilitate the generation of search models for unconventional MR, limits are also posed in the procedure of MR itself. The most prominent limitation may be the resolution of the experimental data, and the proportion of the search model compared to the content of the crystallographic unit cell. SHELXE [130], a popular and powerful algorithm to perform density modification and main-chain autotracing, is heavily limited by a lower resolution limit of 2.5Å. Thus, MR pipelines, such as AMPLE [114] or ARCIMBOLDO [193], may not be able to automatically detect correctly placed search models due to the current dependence on associated software metrics. Furthermore, MR is extremely challenging, if not impossible, when the scattering matter, i.e. a correctly placed search model, is particularly small in relation to the asymmetric unit content whilst the resolution of the experimental data is low. Heavily truncated AMPLE search models or other fragments may thus never suffice as MR probes regardless of the associated similarity to the native structure.

Finally, AMPLE and similar unconventional MR software pipelines try to enable MR when one or more sufficiently similar structures are unavailable to derive the essential phase information. Despite the relative rarity of such a scenario [134], it is essential to provide routes to structure solution when conventional approaches fail since those cases may often provide novel or unexpected findings. The current toolbox for unconventional MR provides idealised fragments [117, 164, 193], supersecondary structure motifs [156], and ensemble search models extracted from a diversity of different starting structures [114, 119, 120]. The former two are usually target-independent, and

thus limited by structural deviations between selected search probes and the target. In comparison, the latter depend much more on accurate and target-specific starting structures but provide a great alternative in lower resolution cases or scenarios whereby larger search models are required. Therefore, unconventional MR requires a diversity of approaches to attempt structure solutions of the most challenging cases. AMPLE and its improvements through predicted residue contacts should therefore be considered an important tool in this set of approaches.

Appendix A

Summary of datasets

Table A.1: Summary of the ORIGINAL dataset.

PDB ID	Molecule	Resolution (Å)	Space Group	Chain ID	Chain Length	Molecules per ASU	Matthew's Coefficient	Solvent Content (%)	Fold	Citation
1a6m	Oxy-myoglobin	1.00	$P2_1$	A	151	1	1.90	36.00	all-α	[194]
1aba	T4 glutaredoxin	1.45	$P2_12_12_1$	A	87	1	2.22	44.62	mixed α/β	[195]
1bdo	Biotinyl domain of acetyl-coenzyme A carboxylase	1.80	$P2_12_12$	A	80	1	2.48	49.00	all-β	[196]
1bkr	Calponin Homology (CH) domain from β-spectrin	1.10	$P2_1$	A	109	1	2.04	39.80	all-α	[197]
1chd	CheB methylesterase domain	1.75	$P3_221$	A	203	1	2.35	47.65	mixed α/β	[198]
1e0s	G-protein Arf6-GDP	2.28	$P6_122$	A	174	1	2.18	37.00	mixed α/β	[199]
1eaz	Phosphoinositol (3,4)-bisphosphate PH domain	1.40	$C222_1$	A	125	1	2.48	48.00	mixed α+β	[200]
1hh8	N-terminal region of P67Phox	1.80	$P3_1$	A	213	1	2.71	45.00	all-α	[201]
1kjl	Galectin-3 domain	1.40	$P2_12_12_1$	A	146	1	2.15	42.68	all-β	[202]
1kw4	Polyhomeotic SAM domain	1.75	$P6_5$	A	89	1	2.25	45.27	all-α	[203]
1lo7	4-hydroxybenzoyl CoA thioesterase	1.50	I222	A	141	1	2.06	40.22	mixed α+β	[204]
1npu	Extracellular domain of murine PD-1	2.00	$P2_12_12_1$	A	117	1	1.67	25.80	all-β	[205]
1pnc	Poplar plastocyanin	1.60	$P2_12_12_1$	A	99	1	1.82	32.48	all-β	[206]
1tjx	Synaptotagmin I C2B domain	1.04	$P3_221$	A	159	1	2.40	48.00	mixed α+β	[207]
1tlv	LicT PRD	1.95	$P3_221$	A	221	1	2.80	50.00	all-α	[208]
2nuz	α-spectrin SH3 domain	1.85	$P2_12_12_1$	A	62	1	2.57	52.16	all-β	
2qyj	Ankyrin	2.05	$P6_1$	A	166	1	2.28	45.99	all-α	[209]
3w56	C2 domain	1.60	I2	A	131	1	2.05	40.10	all-β	[210]
4cl9	N-terminal bromodomain of Brd4	1.40	$P2_12_12_1$	A	127	1	2.21	44.37	all-α	[211]
4u3h	FN3con	1.98	$P4_132$	A	100	1	2.47	50.27	all-β	[212]
4w97	KstR2	1.60	C2	A	200	1	2.75	55.25	all-α	[213]

Table A.2: Summary of the PREDICTORS dataset.

PDB ID	Molecule	Resolution (Å)	Space Group	Chain ID	Chain Length	Molecules per ASU	Matthew's Coefficient	Solvent Content (%)	Fold	Citation
1fcy	Retinoic acid nuclear receptor HRAR	1.30	$P4_12_12$	A	236	1	2.25	45.50	all-α	[214]
1fvg	Peptide methionine sulfoxide reductase	1.60	C121	A	199	1	2.10	41.55	mixed α+β	[215]
1gm4	Cytochrome C3	2.05	$P6_122$	A	107	1	2.48	50.43	all-α	[216]
1gv8	N-II domain of ovotransferrin	1.95	$P3_1$	A	159	1	2.24	45.00	mixed α/β	[217]
1k40	FAT domain of focal adhesion kinase	2.25	C121	A	126	1	2.21	44.40	all-α	[218]
1oee	Hypothetical protein YodA	2.10	C121	A	193	1	2.30	46.20	all-β	[219]
1oz9	Hypothetical protein AQ_1354	1.89	$P4_32_12$	A	150	1	2.76	55.07	mixed α+β	[220]
1q8c	Hypothetical protein MG027	2.00	$P4_1$	A	151	1	2.42	49.25	all-α	[221]
1rlh	Conserved hypothetical protein	1.80	$P6_3$	A	173	1	2.12	41.98	mixed α+β	
1s2x	Cag-Z	1.90	$P2_12_12_1$	A	206	1	2.74	54.70	all-α	[222]
1u61	Putative Ribonuclease III	2.15	$I4_132$	A	138	1	6.50	80.80	all-α	
1zxu	At5g01750 protein	1.70	$P2_12_12_1$	A	217	1	2.50	50.20	mixed α+β	
2eum	Glycolipid transfer protein	2.30	C121	A	209	1	2.25	45.39	all-α	[223]
2ol8	Outer surface protein A	1.90	$P12_11$	O	249	1	2.19	43.87	all-β	[224]
2oqz	Sortase B	1.60	$P12_11$	A	223	1	2.07	40.71	all-β	[225]
2x6u	T-Box transcription factor TBX5	1.90	$P2_12_12_1$	A	203	1	2.20	44.21	all-β	[226]
2y64	Xylanase	1.40	$P2_12_12_1$	A	167	1	2.15	43.00	all-β	[227]
2yjm	TtrD	1.84	C121	A	176	1	2.08	40.80	all-α	[228]
2yq9	2, 3-cyclic-nucleotide 3-phosphodiesterase	1.90	$P2_12_12_1$	A	221	1	2.10	41.70	mixed α+β	[229]
3dju	Protein BTG2	2.26	$P2_12_12_1$	B	122	1	1.98	37.73	mixed α+β	[230]
3g0m	Cysteine desulfuration protein sufE	1.76	$P12_11$	A	141	1	1.88	34.58	mixed α+β	
3qzl	Iron-regulated surface determinant protein A	1.30	$P2_12_12$	A	127	1	2.42	49.12	all-β	[231]
4aaj	N-(5-phosphoribosyl)anthranilate isomerase	1.75	$P6_1$	A	228	1	2.38	48.30	mixed α/β	[232]
4dbb	Amyloid-β A4 precursor protein-binding family A1	1.90	$P4_12_12$	A	162	1	3.25	62.10	all-β	[233]
4e9e	Methyl-CpG-binding domain protein 4	1.90	H3	A	161	1	2.42	49.23	all-α	[234]
4lbj	Galectin-3	1.80	$P2_12_12_1$	A	138	1	2.09	41.01	all-β	[235]
4pgo	Hypothetical protein PF0907	2.30	$P6_522$	A	116	1	3.25	62.10	all-β	[236]

Table A.3: Summary of the TRANSMEMBRANE dataset.

PDB ID	Molecule	Resolution (Å)	Space Group	Chain ID	Chain Length	Molecules per ASU	Matthew's Coefficient	Solvent Content (%)	Fold	Citation
1gu8	Sensory rhodopsin II	2.27	$C222_1$	A	239	1	2.75	53.00	all-α	[237]
2bhw	Chlorophyll A-B binding protein AB80	2.50	C121	A	232	3	4.10	69.00	all-α	[238]
2evu	Aquaporin aqpM	2.30	I4	A	246	1	3.38	63.57	all-α	[239]
2o9g	Aquaporin Z	1.90	I4	A	234	1	3.34	63.19	all-α	[240]
2wie	ATP synthase subunit C	2.13	$P6_322$	A	82	5	3.41	68.00	all-α	[241]
2xov	Rhomboid protease GLPG	1.65	H32	A	181	1	3.50	64.92	all-α	[242]
3gd8	Aquaporin 4	1.80	$P42_12$	A	223	1	2.73	54.97	all-α	[243]
3hap	Bacteriorhodopsin	1.60	$C222_1$	A	249	1	2.73	54.99	all-α	[244]
3ldc	Calcium-gated potassium channel mthK	1.45	$P42_12$	A	82	1	2.48	50.44	all-α	[245]
3ouf	Potassium channel protein	1.55	I2	A	97	2	2.40	48.76	all-α	[246]
3pcv	Leukotriene C4 synthase	1.90	F23	A	156	1	4.91	74.77	all-α	[247]
3rlb	ThiT	2.00	C121	A	192	2	3.89	68.39	all-α	[248]
4dve	Biotin transporter BioY	2.09	C121	A	198	3	3.27	62.40	all-α	[249]

Bibliography

[1] W Friedrich, P Knipping, M Laue, *Ann. Phys.* **1913**, *346*, 971–988.

[2] M Laue, *Ann. Phys.* **1913**, *346*, 989–1002.

[3] W. H. Bragg, W. L. Bragg, *Proceedings of the Royal Society A: Mathematical Physical and Engineering Sciences* **1913**, *88*, 428–438.

[4] W. L. Bragg, *Scientia* **1929**, *23*, 153.

[5] W. L. Bragg, *Nature* **1912**, *90*, 410.

[6] J. D. Watson, F. H. C. Crick, *Nature* **1953**, *171*, 737–738.

[7] D. C. Hodgkin, J Kamper, M Mackay, J Pickworth, K. N. Trueblood, J. G. White, *Nature* **1956**, *178*, 64–66.

[8] T. L. Blundell, J. F. Cutfield, S. M. Cutfield, E. J. Dodson, G. G. Dodson, D. C. Hodgkin, D. A. Mercola, M Vijayan, *Nature* **1971**, *231*, 506–511.

[9] C. C. Blake, D. F. Koenig, G. A. Mair, A. C. North, D. C. Phillips, V. R. Sarma, *Nature* **1965**, *206*, 757–761.

[10] M. F. Perutz, M. G. Rossmann, A. F. Cullis, H Muirhead, G Will, A. C. North, *Nature* **1960**, *185*, 416–422.

[11] J. C. Kendrew, G Bodo, H. M. Dintzis, R. G. Parrish, H Wyckoff, D. C. Phillips, *Nature* **1958**, *181*, 662–666.

[12] H. M. Berman, J Westbrook, Z Feng, G Gilliland, T. N. Bhat, H Weissig, I. N. Shindyalov, P. E. Bourne, *Nucleic Acids Res.* **2000**, *28*, 235–242.

[13] B. Rupp, K. Kantardjieff, *Biomolecular Crystallography: Principles, Practice, and Application to Structural Biology*, Garland Science, New York, **2010**.

[14] P. Evans, A. McCoy, *Acta Crystallogr. D Biol. Crystallogr.* **2008**, *64*, 1–10.

[15] M. G. Rossmann, *Acta Crystallogr. D Biol. Crystallogr.* **2001**, *57*, 1360–1366.

[16] M. G. Rossmann, *Acta Crystallogr. A* **1990**, *46 (Pt 2)*, 73–82.

[17] C. R. Kissinger, D. K. Gehlhaar, D. B. Fogel, *Acta Crystallogr. D Biol. Crystallogr.* **1999**, *55*, 484–491.

[18] N. M. Glykos, M Kokkinidis, *Acta Crystallogr. D Biol. Crystallogr.* **2000**, *56*, 169–174.

[19] R. J. Read, *Acta Crystallogr. D Biol. Crystallogr.* **2001**, *57*, 1373–1382.

[20] M. G. Rossmann, D. M. Blow, *Acta Crystallogr.* **1962**, *15*, 24–31.

[21] M. Bayes, M. Price, *Philosophical Transactions of the Royal Society of London* **1763**, *53*, 370–418.

[22] L. C. Storoni, A. J. McCoy, R. J. Read, *Acta Crystallogr. D Biol. Crystallogr.* **2004**, *60*, 432–438.

[23] B. C. Wang, *Methods Enzymol.* **1985**, *115*, 90–112.

[24] V. Y. Lunin, *Acta Crystallogr. A* **1988**, *44*, 144–150.

[25] G. M. Sheldrick, *Zeitschrift für Kristallographie - Crystalline Materials* **2002**, *217*, 371.

[26] T. C. Terwilliger, *Acta Crystallogr. D Biol. Crystallogr.* **2000**, *56*, 965–972.

[27] P. V. Afonine, R. W. Grosse-Kunstleve, N. Echols, J. J. Headd, N. W. Moriarty, M. Mustyakimov, T. C. Terwilliger, A. Urzhumtsev, P. H. Zwart, P. D. Adams, *Acta Crystallogr. D Biol. Crystallogr.* **2012**, *68*, 352–367.

[28] G. N. Murshudov, P. Skubák, A. A. Lebedev, N. S. Pannu, R. A. Steiner, R. A. Nicholls, M. D. Winn, F. Long, A. A. Vagin, *Acta Crystallogr. D Biol. Crystallogr.* **2011**, *67*, 355–367.

[29] G. M. Sheldrick, *Acta Crystallogr. D Biol. Crystallogr.* **2010**, *66*, 479–485.

[30] V. S. Lamzin, A Perrakis, K. S. Wilson, *International Tables for Crystallography* **2001**, 720–722.

[31] T. Terwilliger, *J. Synchrotron Radiat.* **2004**, *11*, 49–52.

[32] K. Cowtan, *Acta Crystallogr. D Biol. Crystallogr.* **2006**, *62*, 1002–1011.

[33] B. Qian, S. Raman, R. Das, P. Bradley, A. J. McCoy, R. J. Read, D. Baker, *Nature* **2007**, *450*, 259–264.

[34] D. J. Rigden, R. M. Keegan, M. D. Winn, *Acta Crystallogr. D Biol. Crystallogr.* **2008**, *64*, 1288–1291.

[35] R. Das, D. Baker, *Acta Crystallogr. D Biol. Crystallogr.* **2009**, *65*, 169–175.

[36] P. E. Leopold, M Montal, J. N. Onuchic, *Proc. Natl. Acad. Sci. U. S. A.* **1992**, *89*, 8721–8725.

[37] C. B. Anfinsen, *Science* **1973**, *181*, 223–230.

[38] C. Levinthal, *Mossbauer spectroscopy in biological systems* **1969**, *67*, 22–24.

[39] M. Karplus, *Nat. Chem. Biol.* **2011**, *7*, 401–404.

[40] J. Lee, P. L. Freddolino, Y. Zhang in *From Protein Structure to Function with Bioinformatics*, *Vol. 69*, (Ed.: D. J. Rigden), Springer Netherlands, Dordrecht, **2017**, pp. 3–35.

[41] J. Skolnick, *Curr. Opin. Struct. Biol.* **2006**, *16*, 166–171.

[42] C. A. Rohl, C. E. M. Strauss, K. M. S. Misura, D. Baker, *Methods Enzymol.* **2004**, *383*, 66–93.

[43] D. Xu, Y. Zhang, *Proteins* **2012**, *80*, 1715–1735.

[44] M. Blaszczyk, M. Jamroz, S. Kmiecik, A. Kolinski, *Nucleic Acids Res.* **2013**, *41*, W406–11.

[45] T. Kosciolek, D. T. Jones, *PLoS One* **2014**, *9*, e92197.

[46] S. H. P. de Oliveira, E. C. Law, J. Shi, C. M. Deane, *Bioinformatics* **2018**, *34*, 1132–1140.

[47] J. Abbass, J.-C. Nebel, *BMC Bioinformatics* **2015**, *16*, 136.

[48] Y. Shen, G. Picord, F. Guyon, P. Tuffery, *PLoS One* **2013**, *8*, e80493.

[49] S. C. Li, D. Bu, X. Gao, J. Xu, M. Li, *Bioinformatics* **2008**, *24*, i182–9.

[50] I. Kalev, M. Habeck, *Bioinformatics* **2011**, *27*, 3110–3116.

[51] D. Bhattacharya, B. Adhikari, J. Li, J. Cheng, *Bioinformatics* **2016**, *32*, 2059–2061.

[52] T. Wang, Y. Yang, Y. Zhou, H. Gong, *Bioinformatics* **2017**, *33*, 677–684.

[53] S. H. P. de Oliveira, J. Shi, C. M. Deane, *PLoS One* **2015**, *10*, e0123998.

[54] D. Gront, D. W. Kulp, R. M. Vernon, C. E. M. Strauss, D. Baker, *PLoS One* **2011**, *6*, e23294.

[55] N Metropolis, S Ulam, *J. Am. Stat. Assoc.* **1949**, *44*, 335–341.

[56] D Shortle, K. T. Simons, D Baker, *Proc. Natl. Acad. Sci. U. S. A.* **1998**, *95*, 11158–11162.

[57] Y. Zhang, J. Skolnick, *J. Comput. Chem.* **2004**, *25*, 865–871.

[58] P. Bradley, K. M. S. Misura, D. Baker, *Science* **2005**, *309*, 1868–1871.

[59] S Ołdziej, C Czaplewski, A Liwo, M Chinchio, M Nanias, J. A. Vila, M Khalili, Y. A. Arnautova, A Jagielska, M Makowski, H. D. Schafroth, R Kaźmierkiewicz, D. R. Ripoll, J Pillardy, J. A. Saunders, Y. K. Kang, K. D. Gibson, H. A. Scheraga, *Proc. Natl. Acad. Sci. U. S. A.* **2005**, *102*, 7547–7552.

[60] Y. Zhang, J. Skolnick, *Proc. Natl. Acad. Sci. U. S. A.* **2004**, *101*, 7594–7599.

[61] J. Yang, R. Yan, A. Roy, D. Xu, J. Poisson, Y. Zhang, *Nat. Methods* **2015**, *12*, 7–8.

[62] A. Kryshtafovych, A. Barbato, B. Monastyrskyy, K. Fidelis, T. Schwede, A. Tramontano, *Proteins* **2016**, *84 Suppl 1*, 349–369.

[63] C.-H. Tai, H. Bai, T. J. Taylor, B. Lee, *Proteins* **2014**, *82 Suppl 2*, 57–83.

[64] Z. He, M. Alazmi, J. Zhang, D. Xu, *PLoS One* **2013**, *8*, e74006.

[65] L. Kinch, S. Yong Shi, Q. Cong, H. Cheng, Y. Liao, N. V. Grishin, *Proteins* **2011**, *79 Suppl 10*, 59–73.

[66] O. F. Lange, P. Rossi, N. G. Sgourakis, Y. Song, H.-W. Lee, J. M. Aramini, A. Ertekin, R. Xiao, T. B. Acton, G. T. Montelione, D. Baker, *Proc. Natl. Acad. Sci. U. S. A.* **2012**, *109*, 10873–10878.

[67] S. Raman, O. F. Lange, P. Rossi, M. Tyka, X. Wang, J. Aramini, G. Liu, T. A. Ramelot, A. Eletsky, T. Szyperski, M. A. Kennedy, J. Prestegard, G. T. Montelione, D. Baker, *Science* **2010**, *327*, 1014–1018.

[68] C. Göbl, T. Madl, B. Simon, M. Sattler, *Prog. Nucl. Magn. Reson. Spectrosc.* **2014**, *80*, 26–63.

[69] D. S. Marks, L. J. Colwell, R. Sheridan, T. A. Hopf, A. Pagnani, R. Zecchina, C. Sander, *PLoS One* **2011**, *6*, e28766.

[70] M. Michel, S. Hayat, M. J. Skwark, C. Sander, D. S. Marks, A. Elofsson, *Bioinformatics* **2014**, *30*, i482–8.

[71] S. Ovchinnikov, L. Kinch, H. Park, Y. Liao, J. Pei, D. E. Kim, H. Kamisetty, N. V. Grishin, D. Baker, *Elife* **2015**, *4*, e09248.

[72] S. Ovchinnikov, D. E. Kim, R. Y.-R. Wang, Y. Liu, F. DiMaio, D. Baker, *Proteins* **2016**, *84 Suppl 1*, 67–75.

[73] M. Michel, D. Menéndez Hurtado, K. Uziela, A. Elofsson, *Bioinformatics* **2017**, *33*, i23–i29.

[74] S. Ovchinnikov, H. Park, N. Varghese, P.-S. Huang, G. A. Pavlopoulos, D. E. Kim, H. Kamisetty, N. C. Kyrpides, D. Baker, *Science* **2017**, *355*, 294–298.

[75] S. Wang, S. Sun, Z. Li, R. Zhang, J. Xu, *PLoS Comput. Biol.* **2017**, *13*, e1005324.

[76] R. N. Dos Santos, A. J. R. Ferrari, H. C. R. de Jesus, F. C. Gozzo, F. Morcos, L. Martínez, *Bioinformatics* **2018**, *34*, 2201–2208.

[77] W. R. Taylor, K Hatrick, *Protein Eng.* **1994**, *7*, 341–348.

[78] U Göbel, C Sander, R Schneider, A Valencia, *Proteins* **1994**, *18*, 309–317.

[79] E Neher, *Proc. Natl. Acad. Sci. U. S. A.* **1994**, *91*, 98–102.

[80] I. N. Shindyalov, N. A. Kolchanov, C Sander, *Protein Eng.* **1994**, *7*, 349–358.

[81] D. D. Pollock, W. R. Taylor, *Protein Eng.* **1997**, *10*, 647–657.

[82] A. S. Lapedes, B. Giraud, L. Liu, G. D. Stormo in *Statistics in molecular biology and genetics*, Institute of Mathematical Statistics, **1999**, pp. 236–256.

[83] A. Lapedes, B. Giraud, C. Jarzynski, *arXiv* **2012**, 1207.2484.

[84] F. Simkovic, S. Ovchinnikov, D. Baker, D. J. Rigden, *IUCrJ* **2017**, *4*, 291–300.

[85] M. Weigt, R. A. White, H. Szurmant, J. A. Hoch, T. Hwa, *Proc. Natl. Acad. Sci. U. S. A.* **2009**, *106*, 67–72.

[86] L. Burger, E. van Nimwegen, *PLoS Comput. Biol.* **2010**, *6*, e1000633.

[87] S. Balakrishnan, H. Kamisetty, J. G. Carbonell, S.-I. Lee, C. J. Langmead, *Proteins* **2011**, *79*, 1061–1078.

[88] F. Morcos, A. Pagnani, B. Lunt, A. Bertolino, D. S. Marks, C. Sander, R. Zecchina, J. N. Onuchic, T. Hwa, M. Weigt, *Proc. Natl. Acad. Sci. U. S. A.* **2011**, *108*, E1293–301.

[89] D. T. Jones, D. W. A. Buchan, D. Cozzetto, M. Pontil, *Bioinformatics* **2012**, *28*, 184–190.

[90] M. Ekeberg, C. Lövkvist, Y. Lan, M. Weigt, E. Aurell, *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* **2013**, *87*, 012707.

[91] H. Kamisetty, S. Ovchinnikov, D. Baker, *Proc. Natl. Acad. Sci. U. S. A.* **2013**, *110*, 15674–15679.

[92] S. Seemayer, M. Gruber, J. Söding, *Bioinformatics* **2014**, *30*, 3128–3130.

[93] T. A. Hopf, D. S. Marks in *From Protein Structure to Function with Bioinformatics*, (Ed.: D. J. Rigden), Springer Netherlands, Dordrecht, **2017**, pp. 37–58.

[94] R. R. Stein, D. S. Marks, C. Sander, *PLoS Comput. Biol.* **2015**, *11*, e1004182.

[95] T. A. Hopf, S. Morinaga, S. Ihara, K. Touhara, D. S. Marks, R. Benton, *Nat. Commun.* **2015**, *6*, 6077.

[96] S. D. Dunn, L. M. Wahl, G. B. Gloor, *Bioinformatics* **2008**, *24*, 333–340.

[97] D. S. Marks, T. A. Hopf, C. Sander, *Nat. Biotechnol.* **2012**, *30*, 1072–1080.

[98] J. Andreani, J. Söding, *Bioinformatics* **2015**, *31*, 1729–1737.

[99] T. A. Hopf, L. J. Colwell, R. Sheridan, B. Rost, C. Sander, D. S. Marks, *Cell* **2012**, *149*, 1607–1621.

[100] M. J. Skwark, D. Raimondi, M. Michel, A. Elofsson, *PLoS Comput. Biol.* **2014**, *10*, e1003889.

[101] D. T. Jones, T. Singh, T. Kosciolek, S. Tetchner, *Bioinformatics* **2015**, *31*, 999–1006.

[102] T. Du, L. Liao, C. H. Wu, B. Sun, *Methods* **2016**, *110*, 97–105.

[103] A. J. González, L. Liao, C. H. Wu, *Bioinformatics* **2013**, *29*, 1018–1025.

[104] G. Shackelford, K. Karplus, *Proteins* **2007**, *69 Suppl 8*, 159–164.

[105] J. Cheng, P. Baldi, *Bioinformatics* **2005**, *21 Suppl 1*, i75–84.

[106] H. Zhang, Q. Huang, Z. Bei, Y. Wei, C. A. Floudas, *Proteins* **2016**, *84*, 332–348.

[107] Z. Wang, J. Xu, *Bioinformatics* **2013**, *29*, i266–73.

[108] J. Ma, S. Wang, Z. Wang, J. Xu, *Bioinformatics* **2015**, *31*, 3506–3513.

[109] B. Adhikari, J. Hou, J. Cheng, *Bioinformatics* **2018**, *34*, 1466–1472.

[110] B. He, S. M. Mortuza, Y. Wang, H.-B. Shen, Y. Zhang, *Bioinformatics* **2017**, *33*, 2296–2306.

[111] M. Michel, M. J. Skwark, D. Menéndez Hurtado, M. Ekeberg, A. Elofsson, *Bioinformatics* **2017**, *33*, 2859–2866.

[112] S. H. P. de Oliveira, J. Shi, C. M. Deane, *Bioinformatics* **2016**, *87*, btw618.

[113] Q. Wuyun, W. Zheng, Z. Peng, J. Yang, *Brief. Bioinform.* **2018**, *19*, 219–230.

[114] J. Bibby, R. M. Keegan, O. Mayans, M. D. Winn, D. J. Rigden, *Acta Crystallogr. D Biol. Crystallogr.* **2012**, *68*, 1622–1631.

[115] R. M. Keegan, J. Bibby, J. Thomas, D. Xu, Y. Zhang, O. Mayans, M. D. Winn, D. J. Rigden, *Acta Crystallogr. D Biol. Crystallogr.* **2015**, *71*, 338–343.

[116] F. Simkovic, J. M. H. Thomas, R. M. Keegan, M. D. Winn, O. Mayans, D. J. Rigden, *IUCrJ* **2016**, *3*, 259–270.

[117] J. M. H. Thomas, R. M. Keegan, J. Bibby, M. D. Winn, O. Mayans, D. J. Rigden, *IUCrJ* **2015**, *2*, 198–206.

[118] J. M. H. Thomas, F. Simkovic, R. Keegan, O. Mayans, C. Zhang, Y. Zhang, D. J. Rigden, *Acta Crystallogr D Struct Biol* **2017**, *73*, 985–996.

[119] J. Bibby, R. M. Keegan, O. Mayans, M. D. Winn, D. J. Rigden, *Acta Crystallogr. D Biol. Crystallogr.* **2013**, *69*, 2194–2201.

[120] D. J. Rigden, J. M. H. Thomas, F. Simkovic, A. Simpkin, M. D. Winn, O. Mayans, R. M. Keegan, *Acta Crystallogr. D Biol. Crystallogr.* **2018**, *74*, 183–193.

[121] J. F. Bruhn, K. C. Barnett, J. Bibby, J. M. H. Thomas, R. M. Keegan, D. J. Rigden, Z. A. Bornholdt, E. O. Saphire, *J. Virol.* **2014**, *88*, 758–762.

[122] K. Hotta, R. M. Keegan, S. Ranganathan, M. Fang, J. Bibby, M. D. Winn, M. Sato, M. Lian, K. Watanabe, D. J. Rigden, C.-Y. Kim, *Angew. Chem. Int. Ed Engl.* **2014**, *53*, 824–828.

[123] S.-W. Yeh, J.-W. Liu, S.-H. Yu, C.-H. Shih, J.-K. Hwang, J. Echave, *Mol. Biol. Evol.* **2014**, *31*, 135–139.

[124] C.-H. Shih, C.-M. Chang, Y.-S. Lin, W.-C. Lo, J.-K. Hwang, *Proteins* **2012**, *80*, 1647–1657.

[125] D. L. Theobald, D. S. Wuttke, *Bioinformatics* **2006**, *22*, 2171–2172.

[126] G. G. Krivov, M. V. Shapovalov, R. L. Dunbrack, Jr, *Proteins* **2009**, *77*, 778–795.

[127] R. M. Keegan, S. J. McNicholas, J. M. H. Thomas, A. J. Simpkin, F. Simkovic, V. Uski, C. C. Ballard, M. D. Winn, K. S. Wilson, D. J. Rigden, *Acta Crystallogr. D Biol. Crystallogr.* **2018**, *74*, 167–182.

[128] A. J. McCoy, R. W. Grosse-Kunstleve, P. D. Adams, M. D. Winn, L. C. Storoni, R. J. Read, *J. Appl. Crystallogr.* **2007**, *40*, 658–674.

[129] A. Vagin, A. Teplyakov, *Acta Crystallogr. D Biol. Crystallogr.* **2010**, *66*, 22–25.

[130] A. Thorn, G. M. Sheldrick, *Acta Crystallogr. D Biol. Crystallogr.* **2013**, *69*, 2251–2256.

[131] S. X. Cohen, M. Ben Jelloul, F. Long, A. Vagin, P. Knipscheer, J. Lebbink, T. K. Sixma, V. S. Lamzin, G. N. Murshudov, A. Perrakis, *Acta Crystallogr. D Biol. Crystallogr.* **2008**, *64*, 49–60.

[132] B. Adhikari, J. Cheng, *BMC Bioinformatics* **2018**, *19*, 22.

[133] R. D. Finn, P. Coggill, R. Y. Eberhardt, S. R. Eddy, J. Mistry, A. L. Mitchell, S. C. Potter, M. Punta, M. Qureshi, A. Sangrador-Vegas, G. A. Salazar, J. Tate, A. Bateman, *Nucleic Acids Res.* **2016**, *44*, D279–85.

[134] J.-M. Chandonia, N. K. Fox, S. E. Brenner, *J. Mol. Biol.* **2017**, *429*, 348–355.

[135] G. E. Tusnády, Z. Dosztányi, I. Simon, *Nucleic Acids Res.* **2005**, *33*, D275–8.

[136] S. Hayat, C. Sander, D. S. Marks, A. Elofsson, *Proc. Natl. Acad. Sci. U. S. A.* **2015**, *112*, 5413–5418.

[137] M. Remmert, A. Biegert, A. Hauser, J. Söding, *Nat. Methods* **2011**, *9*, 173–175.

[138] The UniProt Consortium, *Nucleic Acids Res.* **2017**, *45*, D158–D169.

[139] J. Söding, *Bioinformatics* **2005**, *21*, 951–960.

[140] F. Simkovic, J. M. H. Thomas, D. J. Rigden, *Bioinformatics* **2017**, *33*, 2209–2211.

[141] Y. Zhang, J. Skolnick, *Proteins* **2004**, *57*, 702–710.

[142] J. Xu, Y. Zhang, *Bioinformatics* **2010**, *26*, 889–895.

[143] M Fujinaga, R. J. Read, *J. Appl. Crystallogr.* **1987**, *20*, 517–521.

[144] J. M. H. Thomas, PhD thesis, University of Liverpool, **2017**.

[145] L. S. Johnson, S. R. Eddy, E. Portugaly, *BMC Bioinformatics* **2010**, *11*, 431.

[146] M. Ekeberg, T. Hartonen, E. Aurell, *J. Comput. Phys.* **2014**, *276*, 341–356.

[147] D. T. Jones, *J. Mol. Biol.* **1999**, *292*, 195–202.

[148] B. Petersen, T. N. Petersen, P. Andersen, M. Nielsen, C. Lundegaard, *BMC Struct. Biol.* **2009**, *9*, 51.

[149] L. Kaján, T. A. Hopf, M. Kalaš, D. S. Marks, B. Rost, *BMC Bioinformatics* **2014**, *15*, 85.

[150] J. Yang, R. Jang, Y. Zhang, H.-B. Shen, *Bioinformatics* **2013**, *29*, 2579–2587.

[151] M. D. Winn, C. C. Ballard, K. D. Cowtan, E. J. Dodson, P. Emsley, P. R. Evans, R. M. Keegan, E. B. Krissinel, A. G. W. Leslie, A. McCoy, S. J. McNicholas, G. N. Murshudov, N. S. Pannu, E. A. Potterton, H. R. Powell, R. J. Read, A. Vagin, K. S. Wilson, *Acta Crystallogr. D Biol. Crystallogr.* **2011**, *67*, 235–242.

[152] K. T. Simons, C Kooperberg, E Huang, D Baker, *J. Mol. Biol.* **1997**, *268*, 209–225.

[153] W. Mao, T. Wang, W. Zhang, H. Gong, *BMC Bioinformatics* **2018**, *19*, 146.

[154] B. Adhikari, D. Bhattacharya, R. Cao, J. Cheng, *Proteins* **2015**, *83*, 1436–1449.

[155] D. Rodríguez, M. Sammito, K. Meindl, I. M. de Ilarduya, M. Potratz, G. M. Sheldrick, I. Usón, *Acta Crystallogr. D Biol. Crystallogr.* **2012**, *68*, 336–343.

[156] M. Sammito, C. Millán, D. D. Rodríguez, I. M. de Ilarduya, K. Meindl, I. De Marino, G. Petrillo, R. M. Buey, J. M. de Pereda, K. Zeth, G. M. Sheldrick, I. Usón, *Nat. Methods* **2013**, *10*, 1099–1101.

[157] D Frishman, P Argos, *Proteins* **1995**, *23*, 566–579.

[158] L. A. Abriata, G. E. Tamò, B. Monastyrskyy, A. Kryshtafovych, M. Dal Peraro, *Proteins* **2018**, *86 Suppl 1*, 97–112.

[159] S. Ovchinnikov, H. Park, D. E. Kim, F. DiMaio, D. Baker, *Proteins* **2018**, *86 Suppl 1*, 113–121.

[160] D. T. Jones, *Proteins* **2001**, *Suppl 5*, 127–132.

[161] J. J. Ellis, F. P. E. Huard, C. M. Deane, S. Srivastava, G. R. Wood, *BMC Bioinformatics* **2010**, *11*, 172.

[162] S. Wang, W. Li, R. Zhang, S. Liu, J. Xu, *Nucleic Acids Res.* **2016**, *44*, W361–6.

[163] A. T. Brünger, P. D. Adams, G. M. Clore, W. L. DeLano, P Gros, R. W. Grosse-Kunstleve, J. S. Jiang, J Kuszewski, M Nilges, N. S. Pannu, R. J. Read, L. M. Rice, T Simonson, G. L. Warren, *Acta Crystallogr. D Biol. Crystallogr.* **1998**, *54*, 905–921.

[164] H. T. Jenkins, *Acta Crystallogr D Struct Biol* **2018**, *74*, 205–214.

[165] G. Scapin, *Acta Crystallogr. D Biol. Crystallogr.* **2013**, *69*, 2266–2275.

[166] R. D. Oeffner, P. V. Afonine, C. Millán, M. Sammito, I. Usón, R. J. Read, A. J. McCoy, *Acta Crystallogr D Struct Biol* **2018**, *74*, 245–255.

[167] J. Schaarschmidt, B. Monastyrskyy, A. Kryshtafovych, A. M. J. J. Bonvin, *Proteins* **2018**, *86 Suppl 1*, 51–66.

[168] M. Sjodt, K. Brock, G. Dobihal, P. D. A. Rohs, A. G. Green, T. A. Hopf, A. J. Meeske, V. Srisuknimit, D. Kahne, S. Walker, D. S. Marks, T. G. Bernhardt, D. Z. Rudner, A. C. Kruse, *Nature* **2018**, *556*, 118–121.

[169] B. Mao, R. Tejero, D. Baker, G. T. Montelione, *J. Am. Chem. Soc.* **2014**, *136*, 1893–1906.

[170] M. I. Sadowski, *Proteins* **2013**, *81*, 253–260.

[171] C. Millán, M. Sammito, I. Usón, *IUCrJ* **2015**, *2*, 95–105.

[172] A. J. McCoy, R. D. Oeffner, A. G. Wrobel, J. R. M. Ojala, K. Tryggvason, B. Lohkamp, R. J. Read, *Proc. Natl. Acad. Sci. U. S. A.* **2017**, *114*, 3637–3641.

[173] N. Fernandez-Fuentes, J. M. Dybas, A. Fiser, *PLoS Comput. Biol.* **2010**, *6*, e1000750.

[174] F. Delaglio, G. Kontaxis, A. Bax, *J. Am. Chem. Soc.* **2000**, *122*, 2142–2143.

[175] G. Kontaxis, F. Delaglio, A. Bax, *Methods Enzymol.* **2005**, *394*, 42–78.

[176] T. A. Jones, S Thirup, *EMBO J.* **1986**, *5*, 819–822.

[177] S. H. P. de Oliveira, C. M. Deane, *Bioinformatics* **2018**, *34*, 2219–2227.

[178] R. Heffernan, K. Paliwal, J. Lyons, A. Dehzangi, A. Sharma, J. Wang, A. Sattar, Y. Yang, Y. Zhou, *Sci. Rep.* **2015**, *5*, 11476.

[179] S. F. Altschul, W Gish, W Miller, E. W. Myers, D. J. Lipman, *J. Mol. Biol.* **1990**, *215*, 403–410.

[180] C. Camacho, G. Coulouris, V. Avagyan, N. Ma, J. Papadopoulos, K. Bealer, T. L. Madden, *BMC Bioinformatics* **2009**, *10*, 421.

[181] J. Söding, A. Biegert, A. N. Lupas, *Nucleic Acids Res.* **2005**, *33*, W244–8.

[182] A. Biegert, C. Mayer, M. Remmert, J. Söding, A. N. Lupas, *Nucleic Acids Res.* **2006**, *34*, W335–9.

[183] Delano, W L, *The PyMOL Molecular Graphics System*, DeLano Scientific, San Carlos, **2002**.

[184] R. Shrestha, K. Y. J. Zhang, *Acta Crystallogr. D Biol. Crystallogr.* **2015**, *71*, 304–312.

[185] D. W. A. Buchan, D. T. Jones, *Bioinformatics* **2017**, *33*, 2684–2690.

[186] F. Bai, F. Morcos, R. R. Cheng, H. Jiang, J. N. Onuchic, *Proc. Natl. Acad. Sci. U. S. A.* **2016**, *113*, E8051–E8058.

[187] D. J. Rigden, *Protein Eng.* **2002**, *15*, 65–77.

[188] D. J. Parente, J. C. J. Ray, L. Swint-Kruse, *Proteins* **2015**, *83*, 2293–2306.

[189] B. Jana, F. Morcos, J. N. Onuchic, *Phys. Chem. Chem. Phys.* **2014**, *16*, 6496–6507.

[190] P. Sfriso, M. Duran-Frigola, R. Mosca, A. Emperador, P. Aloy, M. Orozco, *Structure* **2016**, *24*, 116–126.

[191] F. Morcos, B. Jana, T. Hwa, J. N. Onuchic, *Proc. Natl. Acad. Sci. U. S. A.* **2013**, *110*, 20533–20538.

[192] L. Sutto, S. Marsili, A. Valencia, F. L. Gervasio, *Proc. Natl. Acad. Sci. U. S. A.* **2015**, *112*, 13567–13572.

[193] M. Sammito, C. Millán, D. Frieske, E. Rodríguez-Freire, R. J. Borges, I. Usón, *Acta Crystallogr. D Biol. Crystallogr.* **2015**, *71*, 1921–1930.

[194] J Vojtechovský, K Chu, J Berendzen, R. M. Sweet, I Schlichting, *Biophys. J.* **1999**, *77*, 2153–2174.

[195] H Eklund, M Ingelman, B. O. Söderberg, T Uhlin, P Nordlund, M Nikkola, U Sonnerstam, T Joelson, K Petratos, *J. Mol. Biol.* **1992**, *228*, 596–618.

[196] F. K. Athappilly, W. A. Hendrickson, *Structure* **1995**, *3*, 1407–1419.

[197] S Bañuelos, M Saraste, K Djinović Carugo, *Structure* **1998**, *6*, 1419–1431.

[198] A. H. West, E Martinez-Hackert, A. M. Stock, *J. Mol. Biol.* **1995**, *250*, 276–290.

[199] J Ménétrey, E Macia, S Pasqualato, M Franco, J Cherfils, *Nat. Struct. Biol.* **2000**, *7*, 466–469.

[200] C. C. Thomas, S Dowler, M Deak, D. R. Alessi, D. M. van Aalten, *Biochem. J* **2001**, *358*, 287–294.

[201] S Grizot, F Fieschi, M. C. Dagher, E Pebay-Peyroula, *J. Biol. Chem.* **2001**, *276*, 21627–21631.

[202] P. Sörme, P. Arnoux, B. Kahl-Knutsson, H. Leffler, J. M. Rini, U. J. Nilsson, *J. Am. Chem. Soc.* **2005**, *127*, 1737–1743.

[203] C. A. Kim, M. Gingery, R. M. Pilpa, J. U. Bowie, *Nat. Struct. Biol.* **2002**, *9*, 453–457.

[204] J. B. Thoden, H. M. Holden, Z. Zhuang, D. Dunaway-Mariano, *J. Biol. Chem.* **2002**, *277*, 27468–27476.

[205] X. Zhang, J.-C. D. Schwartz, X. Guo, S. Bhatia, E. Cao, M. Lorenz, M. Cammer, L. Chen, Z.-Y. Zhang, M. A. Edidin, S. G. Nathenson, S. C. Almo, *Immunity* **2004**, *20*, 337–347.

[206] B. A. Fields, H. H. Bartsch, H. D. Bartunik, F Cordes, J. M. Guss, H. C. Freeman, *Acta Crystallogr. D Biol. Crystallogr.* **1994**, *50*, 709–730.

[207] Y. Cheng, S. M. Sequeira, L. Malinina, V. Tereshko, T. H. Söllner, D. J. Patel, *Protein Sci.* **2004**, *13*, 2665–2672.

[208] M. Graille, C.-Z. Zhou, V. Receveur-Bréchot, B. Collinet, N. Declerck, H. van Tilbeurgh, *J. Biol. Chem.* **2005**, *280*, 14780–14789.

[209] T. Merz, S. K. Wetzel, S. Firbank, A. Plückthun, M. G. Grütter, P. R. E. Mittl, *J. Mol. Biol.* **2008**, *376*, 232–240.

[210] D. A. K. Traore, A. J. Brennan, R. H. P. Law, C. Dogovski, M. A. Perugini, N. Lukoyanova, E. W. W. Leung, R. S. Norton, J. A. Lopez, K. A. Browne, H. Yagita, G. J. Lloyd, A. Ciccone, S. Verschoor, J. A. Trapani, J. C. Whisstock, I. Voskoboinik, *Biochem. J* **2013**, *456*, 323–335.

[211] S. J. Atkinson, P. E. Soden, D. C. Angell, M. Bantscheff, C.-W. Chung, K. A. Giblin, N. Smithers, R. C. Furze, L. Gordon, G. Drewes, I. Rioja, J. Witherington, N. J. Parr, R. K. Prinjha, *Med. Chem. Commun.* **2014**, *5*, 342–351.

[212] B. T. Porebski, A. A. Nickson, D. E. Hoke, M. R. Hunter, L. Zhu, S. McGowan, G. I. Webb, A. M. Buckle, *Protein Eng. Des. Sel.* **2015**, *28*, 67–78.

[213] A. M. Crowe, P. J. Stogios, I. Casabon, E. Evdokimova, A. Savchenko, L. D. Eltis, *J. Biol. Chem.* **2015**, *290*, 872–882.

[214] B. P. Klaholz, A Mitschler, D Moras, *J. Mol. Biol.* **2000**, *302*, 155–170.

[215] W. T. Lowther, N Brot, H Weissbach, B. W. Matthews, *Biochemistry* **2000**, *39*, 13307–13312.

[216] R. O. Louro, I Bento, P. M. Matias, T Catarino, A. M. Baptista, C. M. Soares, M. A. Carrondo, D. L. Turner, A. V. Xavier, *J. Biol. Chem.* **2001**, *276*, 44044–44051.

[217] P. Kuser, D. R. Hall, M. L. Haw, M. Neu, R. W. Evans, P. F. Lindley, *Acta Crystallogr. D Biol. Crystallogr.* **2002**, *58*, 777–783.

[218] I. Hayashi, K. Vuori, R. C. Liddington, *Nat. Struct. Biol.* **2002**, *9*, 101–106.

[219] G. David, K. Blondeau, M. Schiltz, S. Penel, A. Lewit-Bentley, *J. Biol. Chem.* **2003**, *278*, 43728–43735.

[220] V. Oganesyan, D. Busso, J. Brandsen, S. Chen, J. Jancarik, R. Kim, S.-H. Kim, *Acta Crystallogr. D Biol. Crystallogr.* **2003**, *59*, 1219–1223.

[221] J. Liu, H. Yokota, R. Kim, S.-H. Kim, *Proteins* **2004**, *55*, 1082–1086.

[222] L. Cendron, A. Seydel, A. Angelini, R. Battistutta, G. Zanotti, *J. Mol. Biol.* **2004**, *340*, 881–889.

[223] L. Malinina, M. L. Malakhova, A. T. Kanack, M. Lu, R. Abagyan, R. E. Brown, D. J. Patel, *PLoS Biol.* **2006**, *4*, e362.

[224] K. Makabe, S. Yan, V. Tereshko, G. Gawlak, S. Koide, *J. Am. Chem. Soc.* **2007**, *129*, 14661–14669.

[225] A. W. Maresso, R. Wu, J. W. Kern, R. Zhang, D. Janik, D. M. Missiakas, M.-E. Duban, A. Joachimiak, O. Schneewind, *J. Biol. Chem.* **2007**, *282*, 23129–23139.

[226] C. U. Stirnimann, D. Ptchelkine, C. Grimm, C. W. Müller, *J. Mol. Biol.* **2010**, *400*, 71–81.

[227] L. von Schantz, M. Håkansson, D. T. Logan, B. Walse, J. Osterlin, E. Nordberg-Karlsson, M. Ohlin, *Glycobiology* **2012**, *22*, 948–961.

[228] S. J. Coulthurst, A. Dawson, W. N. Hunter, F. Sargent, *Biochemistry* **2012**, *51*, 1678–1686.

[229] M. Myllykoski, A. Raasakka, M. Lehtimäki, H. Han, I. Kursula, P. Kursula, *J. Mol. Biol.* **2013**, *425*, 4307–4322.

[230] X. Yang, M. Morita, H. Wang, T. Suzuki, W. Yang, Y. Luo, C. Zhao, Y. Yu, M. Bartlam, T. Yamamoto, Z. Rao, *Nucleic Acids Res.* **2008**, *36*, 6872–6881.

[231] J. C. Grigg, C. X. Mao, M. E. P. Murphy, *J. Mol. Biol.* **2011**, *413*, 684–698.

[232] H. Repo, J. S. Oeemig, J. Djupsjöbacka, H. Iwaï, P. Heikinheimo, *Acta Crystallogr. D Biol. Crystallogr.* **2012**, *68*, 1479–1487.

[233] M. F. Matos, Y. Xu, I. Dulubova, Z. Otwinowski, J. M. Richardson, D. R. Tomchick, J. Rizo, A. Ho, *Proc. Natl. Acad. Sci. U. S. A.* **2012**, *109*, 3802–3807.

[234] S. Moréra, I. Grin, A. Vigouroux, S. Couvé, V. Henriot, M. Saparbaev, A. A. Ishchenko, *Nucleic Acids Res.* **2012**, *40*, 9917–9926.

[235] P. M. Collins, K. Bum-Erdene, X. Yu, H. Blanchard, *J. Mol. Biol.* **2014**, *426*, 1439–1451.

[236] T. Weinert, V. Olieric, S. Waltersperger, E. Panepucci, L. Chen, H. Zhang, D. Zhou, J. Rose, A. Ebihara, S. Kuramitsu, D. Li, N. Howe, G. Schnapp, A. Pautsch, K. Bargsten, A. E. Prota, P. Surana, J. Kottur, D. T. Nair, F. Basilico, V. Cecatiello, S. Pasqualato, A. Boland, O. Weichenrieder, B.-C. Wang, M. O. Steinmetz, M. Caffrey, M. Wang, *Nat. Methods* **2015**, *12*, 131–133.

[237] K. Edman, A. Royant, P. Nollert, C. A. Maxwell, E. Pebay-Peyroula, J. Navarro, R. Neutze, E. M. Landau, *Structure* **2002**, *10*, 473–482.

[238] J. Standfuss, A. C. Terwisscha van Scheltinga, M. Lamborghini, W. Kühlbrandt, *EMBO J.* **2005**, *24*, 919–928.

[239] J. K. Lee, D. Kozono, J. Remis, Y. Kitagawa, P. Agre, R. M. Stroud, *Proc. Natl. Acad. Sci. U. S. A.* **2005**, *102*, 18932–18937.

[240] D. F. Savage, R. M. Stroud, *J. Mol. Biol.* **2007**, *368*, 607–617.

[241] D. Pogoryelov, O. Yildiz, J. D. Faraldo-Gómez, T. Meier, *Nat. Struct. Mol. Biol.* **2009**, *16*, 1068–1073.

[242] K. R. Vinothkumar, K. Strisovsky, A. Andreeva, Y. Christova, S. Verhelst, M. Freeman, *EMBO J.* **2010**, *29*, 3797–3809.

[243] J. D. Ho, R. Yeh, A. Sandstrom, I. Chorny, W. E. C. Harries, R. A. Robbins, L. J. W. Miercke, R. M. Stroud, *Proc. Natl. Acad. Sci. U. S. A.* **2009**, *106*, 7437–7442.

[244] N. H. Joh, A. Oberai, D. Yang, J. P. Whitelegge, J. U. Bowie, *J. Am. Chem. Soc.* **2009**, *131*, 10846–10847.

[245] S. Ye, Y. Li, Y. Jiang, *Nat. Struct. Mol. Biol.* **2010**, *17*, 1019–1023.

[246] M. G. Derebe, D. B. Sauer, W. Zeng, A. Alam, N. Shi, Y. Jiang, *Proc. Natl. Acad. Sci. U. S. A.* **2011**, *108*, 598–602.

[247] H. Saino, Y. Ukita, H. Ago, D. Irikura, A. Nisawa, G. Ueno, M. Yamamoto, Y. Kanaoka, B. K. Lam, K. F. Austen, M. Miyano, *J. Biol. Chem.* **2011**, *286*, 16392–16401.

[248] G. B. Erkens, R. P.-A. Berntsson, F. Fulyani, M. Majsnerowska, A. Vujičić-Žagar, J. Ter Beek, B. Poolman, D. J. Slotboom, *Nat. Struct. Mol. Biol.* **2011**, *18*, 755–760.

[249] R. P.-A. Berntsson, J. ter Beek, M. Majsnerowska, R. H. Duurkens, P. Puri, B. Poolman, D.-J. Slotboom, *Proc. Natl. Acad. Sci. U. S. A.* **2012**, *109*, 13990–13995.