

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

Felix Simkovic

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



Institute of Integrative Biology
University of Liverpool
United Kingdom

Contents

List of Figures	v
List of Tables	vii
List of Equations	viii
List of Abbreviations	ix
1 Introduction	1
1.1 Macromolecular X-ray crystallography	2
1.1.1 X-ray scattering	2
1.1.2 From crystal to structure	5
1.1.3 Unconventional Molecular Replacement	8
1.2 <i>Ab initio</i> protein structure prediction	9
1.3 Residue-residue contact prediction	11
1.3.1 Direct Coupling Analysis	12
1.3.2 Supervised Machine Learning	14
1.3.3 Contact metapredictors	15
1.4 AMPLE	15
1.5 Aims	17
2 Materials & Methods	19
2.1 Selection of datasets	20
2.1.1 ORIGINAL dataset	20
2.1.2 PREDICTORS dataset	20
2.1.3 TRANSMEMBRANE dataset	21
2.2 Enhancement of β -sheet restraints	21
2.3 Evaluation of data	22
2.3.1 Sequence alignment data	22
2.3.2 Contact prediction data	23
2.3.3 Structure prediction data	25
2.3.4 Molecular Replacement data	26
3 Residue contacts predicted by evolutionary covariance extend the application of <i>ab initio</i> molecular replacement to larger and more challenging protein folds	27

3.1	Introduction	28
4	Evaluation of ROSETTA distance-restraint energy functions on contact-guided <i>ab initio</i> structure prediction	29
4.1	Introduction	30
4.2	Materials & Methods	30
4.2.1	Target selection of PREDICTORS dataset	30
4.2.2	Contact prediction	31
4.2.3	Contact-to-restraint conversion	31
4.2.4	<i>Ab initio</i> structure prediction	33
4.2.5	Molecular Replacement	34
4.3	Results	34
4.3.1	Direct comparison of three contact metapredictors	34
4.3.2	Protein structure prediction with two ROSETTA energy functions	38
4.3.3	Impact of metapredictors and energy functions on unconventional Molecular Replacement	43
4.4	Discussion	54
5	Alternative <i>ab initio</i> structure prediction algorithms for AMPLE	57
5.1	Introduction	58
5.2	Materials & Methods	59
5.2.1	Target selection	59
5.2.2	Contact prediction	59
5.2.3	<i>Ab initio</i> structure prediction	60
5.2.4	Molecular Replacement	60
5.3	Results	61
5.3.1	Alignment depth and contact prediction precision	61
5.3.2	Comparison of decoy quality	62
5.3.3	Molecular Replacement	68
5.4	Discussion	73
6	Decoy subselection using contact information to enhance MR search model creation	75
6.1	Introduction	76
6.2	Materials & Methods	76
6.2.1	Target selection	76
6.2.2	Computation of range-specific satisfaction scores	77
6.2.3	Decoy subselection	77
6.2.4	Molecular Replacement	78
6.3	Results	78
6.3.1	Contact pair satisfaction correlates with decoy quality	79
6.3.2	Long-range contact satisfaction metric to filter decoy sets	81
6.3.3	AMPLE’s cluster-and-truncate approach with filtered decoy sets	84

6.3.4	MR search models by processing single decoys	88
6.3.5	Decoy subselection extends AMPLE's performance	91
6.4	Discussion	94
7	Protein fragments as search models in Molecular Replacement	97
7.1	Introduction	98
7.2	Materials & Methods	99
7.2.1	Target selection	99
7.2.2	Fragment picking using FLIB	99
7.2.3	Molecular Replacement in MRBUMP	100
7.2.4	Assessment of FLIB fragments	101
7.3	Results	101
7.3.1	Precision of FLIB input data	101
7.3.2	FLIB fragment picking	104
7.3.3	FLIB fragment selection for Molecular Replacement	108
7.3.4	Molecular Replacement using FLIB fragments	113
7.4	Discussion	118
8	Conclusion	121
A	Appendix	123
	Bibliography	127

List of Figures

1.1	Schematic of Bragg scattering.	3
1.2	Schematic of the folding funnel hypothesis.	9
1.3	Schematic of inference of covariance signal	12
1.4	Cluster-and-truncate approach employed by AMPLE.	16
4.1	Schematic comparison of ROSETTA energy functions	32
4.2	Precision analysis of three metapredictors	35
4.3	Sequence coverage and contact precision analysis	35
4.4	Contact singleton analysis for three metapredictors	36
4.5	Comparison of contact precision for three metapredictors	37
4.6	Metapredictor contact pair similarity analysis	38
4.7	Median TM-score comparison between ROSETTA energy functions	39
4.8	Top TM-score comparison between ROSETTA energy functions	40
4.9	TM-score distribution by fold category and ROSETTA energy function	41
4.10	Median TM-score analysis by fold category and ROSETTA energy function	42
4.11	Influence of target chain length and restraint precision on median TM-score	43
4.12	Structure solution count from AMPLE-derived search models	44
4.13	Relationship between median TM-score and search model size of AMPLE ensembles	46
4.14	Relationship between median TM-score and decoy count in SPICKER cluster	47
4.15	SPICKER cluster sizes grouped by restraint condition	48
4.18	RIO score analysis of successful targets	51
4.19	Example of successfully placed AMPLE search model	52
4.20	Example of successfully placed AMPLE search model	53
4.21	Example of successfully placed AMPLE search model	54
5.1	Distribution of alignment depth for subset of targets in the PREDICTORS dataset.	61
5.2	Contact prediction analysis for numerous contact selection cutoffs	62
5.3	Distribution of decoy TM-scores for four modelling algorithms	63
5.4	Per-target TM-score analysis for four modelling algorithms with contacts	64
5.5	Analysis of alignment depth, precision and TM-scores	65
5.6	Per-target TM-score analysis for four modelling algorithms without contacts	66

5.7	Distribution of decoy TM-scores by fold, chain length and algorithm.	68
5.8	Summary of MR success with AMPLE ensemble search models.	69
5.9	Distribution of search model truncation and secondary structure content	70
5.10	Examples of PHASER-placed AMPLE search models.	71
5.11	Relationship between ensemble quality, PHASER LLG and SHELXE CC	72
5.12	Distribution of Ramachandran outliers of AMPLE ensemble search model centroids	73
6.1	Linear regression model between decoy TM-score and contact satisfaction	80
6.2	Top-1 decoy TM-score and contact satisfaction analysis	81
6.3	TM-score comparison pre- and post-decoy subselection	83
6.4	Effect of decoy subselection on SPICKER clusters	86
6.5	Effect of decoy subselection on THESEUS variance	87
6.6	Selection of single decoys by long-range satisfaction	89
6.7	Difference in RMSD for individually processed decoys	90
6.8	Relationship between decoy quality and fraction of residues retained . .	91
6.9	Molecular Replacement summary of decoy-subselected AMPLE ensembles	92
6.10	Comparison of ensembles derived from differently subselected decoys . .	93
7.1	PSIPRED schema for FLIB targets	102
7.2	Contact map comparison for FLIB targets	103
7.3	SPIDER2 torsion angle prediction analysis of FLIB targets	104
7.4	FLIB fragment library comparison	106
7.5	Coverage and precision of Flib fragment libraries	107
7.6	Spearman rank-order correlation coefficient analysis of FLIB fragments .	109
7.7	Correlation analysis for final FLIB Molecular Replacement (MR) fragments	110
7.8	Distribution of contact precision for FLIB fragments	111
7.9	Fragment search models derived from FLIB	112
7.10	MR structure solutions by FLIB target	113
7.11	MR structure solutions by FLIB library	114
7.12	MR structure solutions by input parameters	115
7.13	Relationship between fragment chain length and normalised RIO scores.	116
7.14	Example of FLIB fragment to MR solution	117

List of Tables

4.1	Summary of AMPLE keyword arguments for FADE and SIGMOID ROSETTA energy functions.	33
6.1	Correlation analysis between decoy TM-score and contact satisfaction	79
7.1	Contact prediction summary for FLIB targets	103
7.2	FLIB fragment characteristics across four protein targets	105
A.1	Summary of the ORIGINAL dataset.	124
A.2	Summary of the PREDICTORS dataset.	125
A.3	Summary of the TRANSMEMBRANE dataset.	126

List of Equations

1.1	Phase difference equation	3
1.2	Atomic Scattering Factor equation	3
1.3	Total Scattering Power equation	4
1.4	Laue equations	4
1.5	Bragg equation	4
1.6	Mathematical expression of a Structure Factor	5
1.7	Mathematical expression of Electron Density	5
1.8	Potts model	13
1.9	Partition function of Potts model	13
1.10	Covariance pseudo-likelihood approximation	13
1.11	Matrix centering	14
1.12	Frobenius norm	14
1.13	Evolutionary coupling score	14
2.1	Sequence Alignment diversity	22
2.2	Sequence Alignment depth	23
2.3	Contact map coverage	23
2.4	Precision score	23
2.5	Jaccard index	24
2.6	Root-Mean-Square Deviation	25
2.7	Template-Modelling score	25

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 hallmark 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.

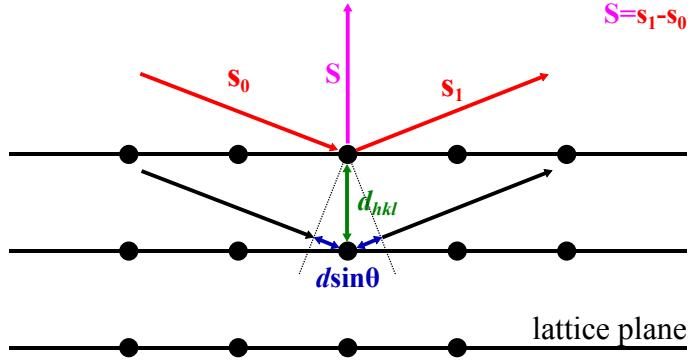


Figure 1.1: Schematic of Bragg scattering.

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 $\rho(\mathbf{r})$. In X-ray scattering, it suffices to approximate the shape of the electron density to a sphere. If we now consider the emitting wave s_1 of an X-ray photon scattered by any position \mathbf{r} in the electron density of an atom, then the phase difference $\Delta\varphi$ to the incoming wave s_0 can be described by Eq. 1.1 [13].

$$\Delta\varphi = 2\pi (s_1 - s_0) \cdot \mathbf{r} = 2\pi \cdot \mathbf{S} \cdot \mathbf{r} \quad 1.1$$

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.

$$f_s = \int_{\mathbf{r}}^{V(\text{atoms})} \rho(\mathbf{r}) \cdot \exp(2\pi i \mathbf{S} \cdot \mathbf{r}) \cdot d\mathbf{r} \quad 1.2$$

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 a 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.

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

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 \mathbf{r} to the three dimensional lattice translation vectors \mathbf{a} , \mathbf{b} and \mathbf{c} . The Laue equations (Eq. 1.4) do exactly that and ultimately determine the positions of the diffraction peaks in 3-dimensional space.

$$\mathbf{S} \cdot \mathbf{a} = n_1, \quad \mathbf{S} \cdot \mathbf{b} = n_2, \quad \mathbf{S} \cdot \mathbf{c} = n_3 \quad 1.4$$

Such determination is possible through the findings made by Bragg and Bragg [3], who identified the relationship between the scattering vector \mathbf{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 .

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

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 $\mathbf{S} \cdot \mathbf{r}_j$ as Miller indices of the reflection planes $\mathbf{h} \cdot \mathbf{x}_j$.

$$F_h = \sum_{j=1}^{atoms} f_{s,j}^0 \cdot \exp(2\pi i \hbar \mathbf{x}_j) \quad 1.6$$

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

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 $\rho(x, y, z)$.

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

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 Phasing [13]. Today, the most popular method to recovering the lost phase information is MR [14, 15]. 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 [16–18], 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 [19] — 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 sub-classed 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 intra-molecular Patterson vectors, the search model can be oriented against the experimentally determined Patterson vectors. 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 trialed 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 [20] to compare calculated structure factors and observed structure factor amplitudes directly [18]. 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. 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 [18]. 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 [21], 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 [18].

The initial electron density map — regardless of its determination by MR or Experimental Phasing — 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 calculate 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 first proposed by Wang [22], which exploits the fact that solvent regions in protein crystals are disordered, and thus differ 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 [23] 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 [24] 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 [24]. 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 [25].

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 [26]. 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 [26, 27]. 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 pre-defined 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 [26]. 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 [28], fitting pseudo-atoms into electron density [29], or fitting reference coordinates with similar electron density maps [30, 31]. 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. Experimental phasing approaches to circumvent the absence of MR templates can be expensive, unsuccessful and very challenging for certain protein targets, and thus re-

main 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 [32–34]. 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 [35]. 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) [36]. *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 [37, 38]. This paradox — termed the Levinthal paradox [37] — created the basis for the folding funnel hypothesis.

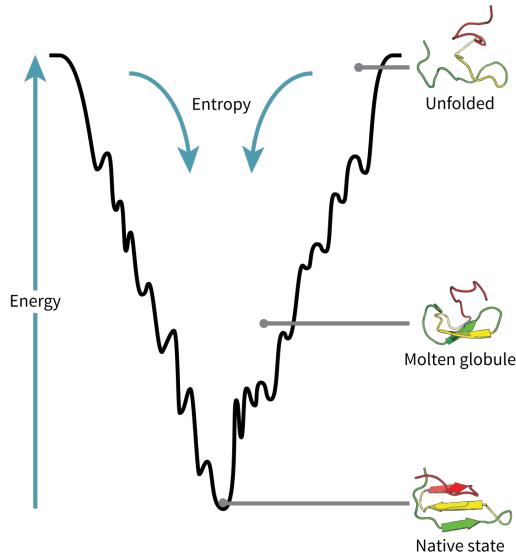


Figure 1.2: Schematic of the folding funnel hypothesis [35]. Diagram produced by Wikipedia [39] contributors.

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-energy state along a protein’s energy landscape. However, in an attempt to avoid the Levinthal paradox, different knowledge- and physics-based en-

ergy 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 pre-defined 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* 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. 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, 69–75].

1.3 Residue-residue contact prediction

The use of residue-residue contact information to reduce the conformational search space in 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 [76–79]. 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 [80–82]. 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” causal (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.

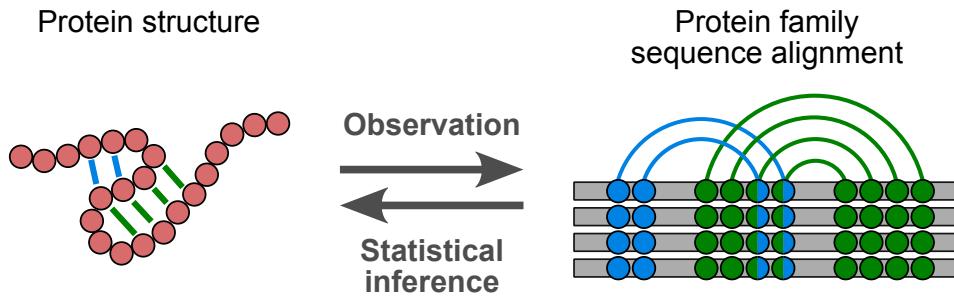


Figure 1.3: Schematic of inference of covariance signal originating from evolutionary pressure in protein tertiary structures and encoded in its family’s sequence alignment (adapted from [83]).

Lapedes et al. [81] 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, 84–91]. 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 [92]. 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 [89].

The pairwise probabilistic model $P(\sigma)$ of the amino acid sequence $\sigma = (\sigma_1, \sigma_2, \dots, \sigma_N)$

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 .

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

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 [89].

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

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 [82], message passing [84], mean-field [69, 87, 88, 93], and pseudolikelihood maximisation [86, 89–91, 94]. 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) [92].

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

Equation 1.10 describes the conditional probability of observing amino acid (σ_i) in position i given all other amino acids ($\sigma \setminus \sigma_i$) 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 [92]. Typically, regularisation terms are also added to Eq. 1.10 to avoid overfitting of the input data [92].

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-centered 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 Evolutionary Coupling (EC) score (Eq. 1.13) [88–91, 95].

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

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

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

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 required of > 1000 sequence homologs [88, 96, 97]. Simultaneously, Marks et al. [69] and Kamisetty et al. [90] recommended a more sequence-specific length-dependent factor, whereby the sequence count in the alignment should exceed at least five times 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*) was introduced to capture both the sequence count and diversity in a given alignment [87, 98–100]. Although target- and algorithm-dependent thresholds persist, a minimum of 100-200 effective sequences are typically required [99, 100]. Furthermore, individual weights used to calculate the alignment depth are widely used in covariance-based algorithms to reweight individual sequences [89]. 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 [89].

1.3.2 Supervised Machine Learning

Unlike DCA approaches, Supervised Machine Learning 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 fea-

tures to infer contacting residue pairs [101–106]. 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 [99, 106, 107].

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, 99, 100, 108–110] 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, 111, 112].

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 (Ab initio Modelling of Proteins for moLEcular replacement) — based on the work of Rigden et al. [33] — attempts to tackle this challenge by utilising structural information from a variety of sources, such as *ab initio* structure predictions [113–117], Nuclear Magnetic Resonance (NMR) ensembles [118], and single [119] or multiple distant homologs [120, 121].

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 models, and if the prediction was accurate it should elucidate a conserved structural core [113]. 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 [118]. 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 [122, 123], this flexibility can be used as a proxy similar to NMR ensembles to drive trimming for identification of a more rigid, shared core [119]. 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 [120, 121].

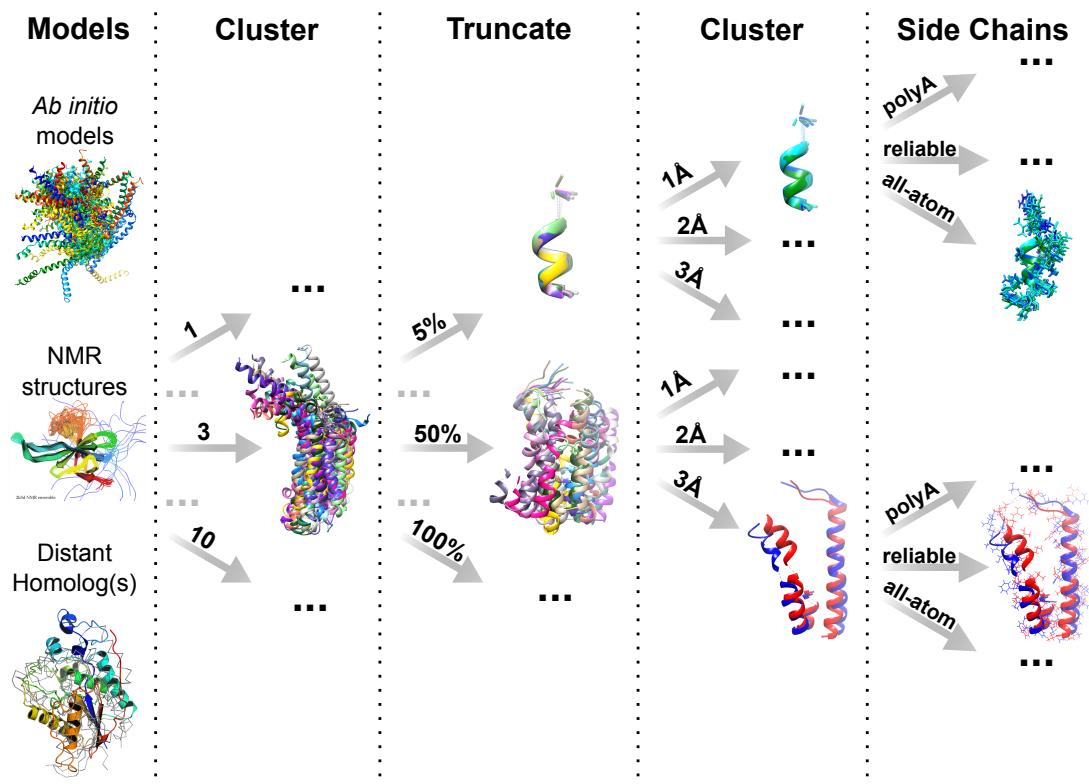


Figure 1.4: Cluster-and-truncate approach employed by AMPLE.

In each case, AMPLE attempts to identify such a shared core by employing a cluster-and-truncate approach (Fig. 1.4) [113]. The latter can be separated into three main parts: (i) clustering of starting models to identify subsets of similar folds (only applicable for *ab initio* structure predictions), (ii) incremental truncation of each cluster or collection of structures by its structural variance, and (iii) subclustering of each truncated set of models 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 [124]. 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 [125] 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 [126]. In AMPLE, MRBUMP’s structure determination features are of particular interest. It employs PHASER [127] and MOLREP [128] for MR, refines the MR solutions with REFMAC5 [27], uses SHELXE for density modification and main-chain tracing [129], and attempts automated model building with ARP/wARP [130] and BUCCANEER [31]. These features enable the sampling of each AMPLE-generated ensemble for its suitability as MR search model.

1.5 Aims

In the previous sections, the theory behind three major concepts was outlined: Molecular Replacement and the need for unconventional approaches, *ab initio* protein structure prediction, and residue-residue contact information. 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. Residue-residue contact information, as described above, 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 information to improving the approach AMPLE takes in unconventional MR. Furthermore, work centred on the identification of other areas of application of residue-residue contact information 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 information to AMPLE’s *ab initio* structure determination routine.

2. In Chapter 4, the proof-of-principle study is 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. Two alternative protocols, namely SAINT2 [46] and FRAGFOLD [45], were explored for potential benefits over ROSETTA.
4. In Chapter 6, a study was carried out to explore long-range contact satisfaction decoy subselection prior to submission to ensemble search model generation in AMPLE.
5. In Chapter 7, a pilot study was done to explore the potential use of contact information to determine structural fragments or substructures suitable for 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 $\alpha+\beta$). 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 2.3 \AA .

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 [131] 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 [132]. Targets with fold assignments other than "a" (all- α), "b" (all- β), "c" (mixed $\alpha+\beta$) 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) webserver 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 2.3 \AA ; 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) [133] 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 ($\alpha+\beta$ and α/β). Each target was further binned by chain length (derived from the deposited sequence in the RCSB PDB) with bin centres of 125, 175 and 225 residues. Lastly, each fold bin was also sub-grouped by alignment depth, which was calculated for the sequence alignments of each PFAM family. Three bins were established with bin edges of 0, 100 and 200 sequences. Thus, all targets were grouped by fold class and further grouped 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 units 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. [117]. In summary, 14 non-redundant transmembrane protein targets were selected from the PDBTM [133], with a chain length of < 250 residues and resolution of $< 2.5\text{\AA}$. The final selection is summarised in Table A.3.

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, only a single paper has specifically focused on improvements to the structure prediction of β -sheet formation [134]. 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. [134], which focused exclusively on β -barrel proteins, was developed combining a starting set of contact pairs with a specifically-prepared set obtained from BBCONTACTS [97]. A HHBLITS [135] MSA was constructed using two sequence-search iterations with an E-value cutoff of 10^{-3} against the uniprot20 database [136]. Redundant sequences were removed from the MSA to 90% sequence identity using HHFILTER [135]. Subsequently, the MSA was subjected to CCMPRED [91] for co-evolution based contact prediction. The BBCONTACTS algorithms also requires a secondary-structure prediction, which was obtained using the `addss.pl` script [135] distributed with the HHSUITE [137]. Lastly, BBCONTACTS requires a descriptor for the diversity of the input sequence alignment [97], which was calculated using Eq. 2.1. All three inputs were subjected to BBCONTACTS to obtain a set of β -strand specific contact pairs.

The BBCONTACTS contact pairs were added to a base set of contact pairs usually obtained from a separate (meta-)predictor. The 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 [138].

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 is used in BBCONTACTS as input parameter [97]. The MSA diversity nis defined in Eq. 2.1.

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

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 co-evolution signal. The alignment depth — often also referred to as Number of Effective Sequences (M_{eff}) — captures this diversity by computing the number of non-redundant sequences in the MSA.

$$M_{eff} = \sum_i \frac{1}{\sum_j S_{i,j}} \quad 2.2$$

Various approaches exist for computing M_{eff} [87, 88, 100] yielding similar results [99]. In this thesis, the approach defined by Morcos et al. [87] is used. Morcos et al. [87] 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). For each sequence in the MSA, the normalised Hamming distance to all other sequences is computed. Subsequently, all normalised distances of less than 0.2 (sequence identity of 80%) are assigned the binary value of one, whereas all others the value zero. The effectiveness of each sequence is thus described by its fractional weight, i.e. the sum of all binary distance values ($\sum_j S_{i,j}$), and the overall alignment depth as the sum of those fractional weights. Therefore, alignments with more diversity result in larger fractional weights per sequence yielding a higher sum of those weights (M_{eff}).

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 is defined by the number of residues (N_{map}) for which at least one contact pair exists proportional to the total number of residues in the target chain (L). Thus, the contact map coverage can be defined by Eq. 2.3.

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

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 False Positive (FP) ones (Eq. 2.4). A contact pair was considered a TP if the equivalent C β (C α for glycine) atoms in the native structure were $< 8\text{\AA}$ apart, otherwise a FP. The precision value ranges from 0 to 1 with a value of 1 indicating all contact pairs are TPs.

$$Precision = \frac{TP}{TP + FP} \quad 2.4$$

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) [112].

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

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) [99]. Such contacts are also often the ones to be down-weighted by neural network architectures of metapredictors, such as PCONSC2 [99] or METAPSICOV [100]. 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 Ca atoms.

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

2.3.3.2 Template-Modelling score

The Template-Modelling score (TM-score) is an alternative measure of the similarity between two protein structures [139]. Unlike the RMSD, the TM-score score assigns a length-dependent weight to the distances between atoms, with shorter distances getting assigned stronger weights [139]. 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.

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

d_i describes the distance between the i th 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 [140]

2.3.4 Molecular Replacement data

2.3.4.1 Register-Independent Overlap

The Residue-Independent Overlap (RIO) score [116] is a measure of structural similarity between two protein structures considering the total number of atoms within $< 1.5\text{\AA}$. The RIO can be separated into the in- (RIO_{in}) and out-of-register (RIO_{out}) score considering the sequence register between the 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 Ca atoms to be within the 1.5\AA threshold.

2.3.4.2 Structure solution

MR structure solutions were assessed throughout all works presented in this thesis by the Correlation Coefficient (CC) [141] and Average Chain Length (ACL) scores computed by SHELXE. SHELXE performs density modification and main-chain tracing of the refined MR solution [129]. Thorn and Sheldrick [129] highlighted in their work that a CC of $\geq 25\%$ indicates a successful structure solution. Additionally, previous research with AMPLE [116] 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 [31] or ARP/wARP [130] 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

Residue contacts predicted by evolutionary covariance extend the application of *ab initio* molecular replacement to larger and more challenging protein folds

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. [115], and a great majority of work relating to the transmembrane targets by Thomas et al. [117]. As such, this chapter consists of extracts from both publications with additional information where appropriate. Text duplicated from either publication was written by Felix Simkovic, all other elements were adapted.*

3.1 Introduction

The introduction of residue-residue contacts 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 prediction of larger and more β -rich protein structures [e.g., 45, 46, 69–75]. In AMPLER, these two domains are the major limitation for a more successful approach [113]. This typically results in user success being limited to small globular and primarily α -helical folds, or time- and resource-demanding attempts most likely going to be unsuccessful for larger targets

With the advent of contact information, it has thus become essential to identify the extend to which this invaluable bit of information is going to help AMPLER users in the future.

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 protein targets through the use of residue-residue contact information to restrain *ab initio* structure prediction has been highlighted in chapter XYZ. However, that study only focused on PCONSC2 as a metapredictor without considering alternatives, and thus served only as a proof-of-principle work for applications of contact information 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, 90, 99, 100, 107, 109, 110]. 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, 100, 142, 143]. 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 work presented in this chapter was aimed at extensively comparing state-of-the-art contact- and structure-prediction protocols with a focus on the use of such decoys for 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 alignment depths of < 100 in the Pfam MSA were excluded (Table A.2).

4.2.2 Contact prediction

Residue contacts for each target sequence were predicted using three different metapredictors, namely METAPSICOV [100], GREMLIN [90], and PCONSC2 [99]. Online 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. Online 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 pre-formatted 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 Stage 1 and another after Stage 2 post-prediction processing. In this study, contact predictions after Stage 1 (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 ???. Additionally to the three main contact predictions outlined above, a set of BBCONTACTS restraints was obtained for protein targets containing β -strands (see Section 2.2).

The sequence-database versions of all three metapredictors, whether on- or offline, were identical to those used in Chapter 3.

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, 99], the FADE function was set to acknowledge a formed restraint if the participating C β atoms (C α in case of Gly) were within 9Å. In comparison, the SIGMOID function was defined with amino acid specific distances for C β atoms (C α in case of Gly) to recognise the different sizes of each amino acid [71, 90].

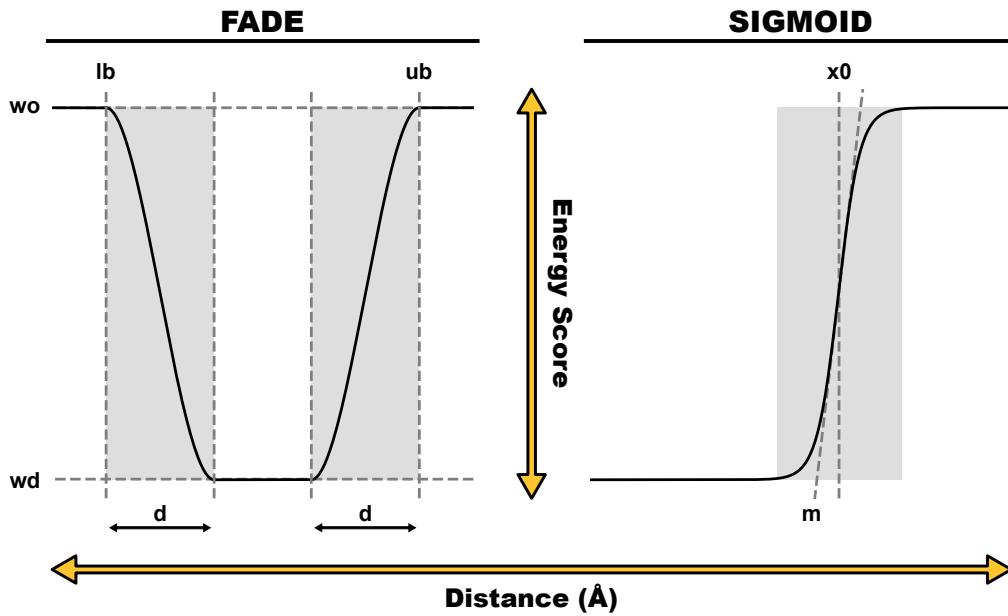


Figure 4.1: ROSETTA energy function comparison. Abbreviations corresponds to input parameters.

To explore the effects of the varying energy function definitions, we created six lists of contact restraints for each α -helical target and nine lists for each β -structure containing one. The top-ranking contact pairs per prediction were converted using the PCONS FOLD 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 (see Chapter 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 ???. 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: Summary of AMPLE keyword arguments for FADE and SIGMOID ROSETTA energy functions.

Energy Function	AMPLE keywords
FADE	<pre>-contact_file <FILENAME> -contact_format <FORMAT> -energy_function FADE -restraints_factor 1.0 -distance_to_neighbour 5 -restraints_weight 1.0</pre>
FADE (BBCONTACTS)	<pre>-contact_file <FILENAME> -contact_format <FORMAT> -energy_function FADE -restraints_factor 1.0 -distance_to_neighbour 5 -restraints_weight 1.0</pre>
SIGMOID	<pre>-contact_file <FILENAME> -contact_format <FORMAT> -energy_function SIGMOID -restraints_factor 1.5 -distance_to_neighbour 3 -restraints_weight 3.0</pre>
SIGMOID (BBCONTACTS)	<pre>-contact_file <FILENAME> -contact_format <FORMAT> -energy_function SIGMOID_bbcontacts -restraints_factor 1.5 -distance_to_neighbour 3 -restraints_weight 3.0</pre>

4.2.4 *Ab initio* 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, protein structures were predicted without any contact restraints to acquire a control set of decoys. 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 contact conditions described previously. In total, 162 sets of models were generated across 18 protein targets.

4.2.5 Molecular Replacement

Besides considering model quality, one key interest of this study was the assessment of the model sets created in the previous step as *ab initio* MR search model templates. To reduce the enormous computational cost linked to trialling 162 sets of models, 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 STAGE 1 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 is the most precise contact predictor across the protein target dataset in this study (Fig. 4.2). The difference between the three metapredictors is most evident in the highest-scoring contact pairs ($L/10$). The median precision values for METAPSICOV and PCONSC2 contact predictions are above 50% up to L contact pairs. GREMLIN, in comparison, predicts 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 are much more similar across the three different metapredictors: METAPSICOV and PCONSC2 are 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 illustrates 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 are 0.2 precision score units lower. Added to the difference in precision scores is the difference in sequence coverage (Fig. 4.3). Although producing the on-average worst contact predictions out of the three metapredictors used in this study, GREMLIN contact predictions have the highest sequence coverage. However, an analysis of singleton contact pairs, usually with high degrees of false positives, revealed a positive correlation ($\rho_{Pearson} = 0.47$; $p < 0.001$) between the fraction of

singleton contact pairs and sequence coverage and hints 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).

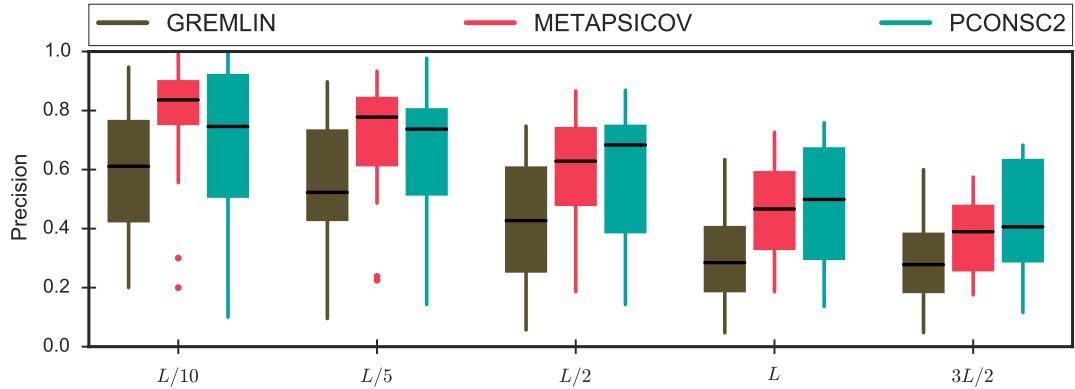


Figure 4.2: Precision spread for three metapredictors computed at five contact selection cutoff values relative to the target chain length (L).

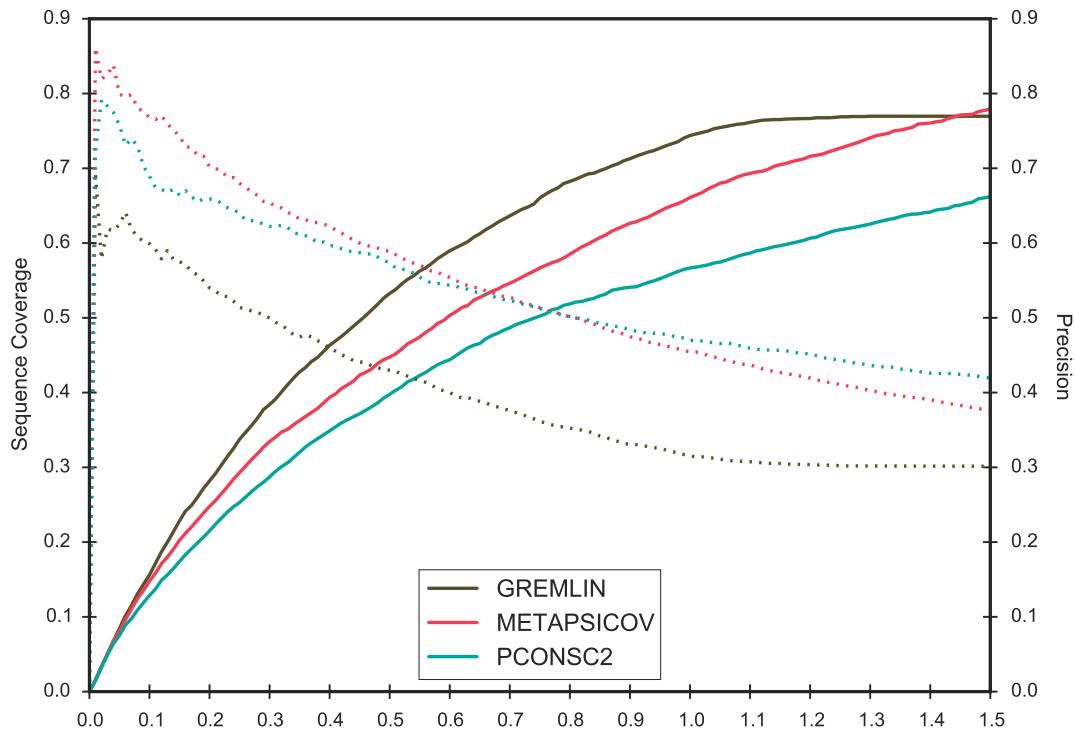


Figure 4.3: Average sequence coverage (line) and contact prediction precision scores (dashed) across a continuous range of contact selection cutoffs ranging from $[0.0, 1.5]$ for all targets.

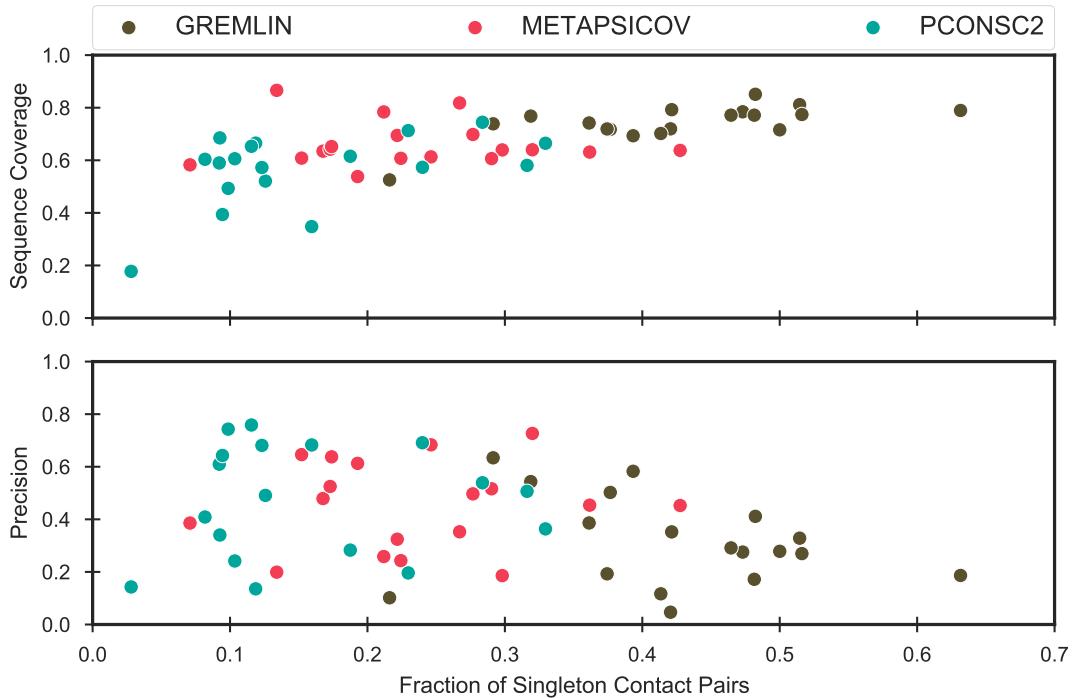


Figure 4.4: Contact singleton analysis compared against the precision of L contact pair lists for three metapredictors.

Given that the overall precision of contact pairs predicted by the three metapredictors differs, it is important to understand where the difference originates. 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 are very similar across the three metapredictors (Fig. 4.5). However, the prediction precision of some targets differs significantly. For example, the METAPSICOV prediction for the human retinoic acid nuclear receptor HRAR (PDB: 1fcy) contains high precision in its highest scoring (top- $L/10$) contact pairs (Fig. 4.5). In comparison, GREMLIN and PCONSC2 predictions for the same target contain less precise contact pairs ($\Delta\text{Precision}_{\text{METAPSICOV-GREMLIN}} L/10 = -0.522$; $\Delta\text{Precision}_{\text{METAPSICOV-PCONSC2}} L/10 = -0.435$). However, the addition of further contact pairs up to $3L/2$ results in near-identical precision across the three metapredictors for this target. A second example illustrating such a difference are the contact predictions for the human galectin-3 CRD sequence (PDB: 4lbj). In contrast to the previous example, the data shows high precision scores for the METAPSICOV and PCONSC2 predictions for this target, yet low precision for the top GREMLIN contact pairs ($\Delta\text{Precision}_{\text{METAPSICOV-GREMLIN}} L/10 = -0.231$; $\Delta\text{Precision}_{\text{METAPSICOV-PCONSC2}} L/10 = +0.077$).

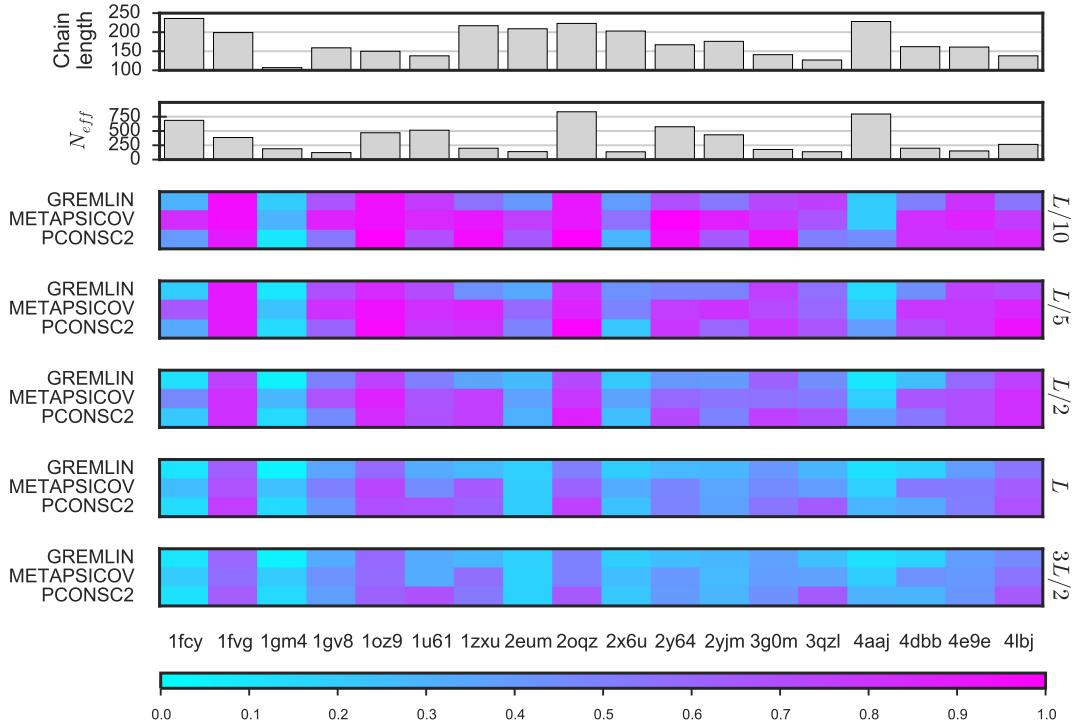


Figure 4.5: Contact prediction precision scores from three metapredictors for 18 targets at different contact pair selection thresholds. The Pfam alignment depth is given by means of number of effective sequences (N_{eff}). The color scale corresponds to the precision in $[0, 1]$.

The data presented in Fig. 4.5 also indicates that there is no direct link between chain length or N_{eff} and the precision of the resulting contact predictions. The N-(5'-phosphoribosyl)anthranilate isomerase sequence (PDB: 4aaaj) 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 contrasts with the sequence of sortase B (PDB: 2oqz), which shows similar characteristics yet obtained mean precision at $L/10$ contact pairs of 0.938 (top- L : 0.622).

Although the contact predictions differ in precision, an interesting question rests with the similarity of the predicted contact pairs amongst the sets. Thus, the similarity of contact predictions across the three metapredictors is 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 suggests 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 is also lower — indicating less overlap — at higher cutoff thresholds. However, it is worth noting that the Jaccard index only considers identical matches and does not consider the neighbourhood of a contact pair. Thus, the index does not highlight similar regions with contact pairs in both maps.

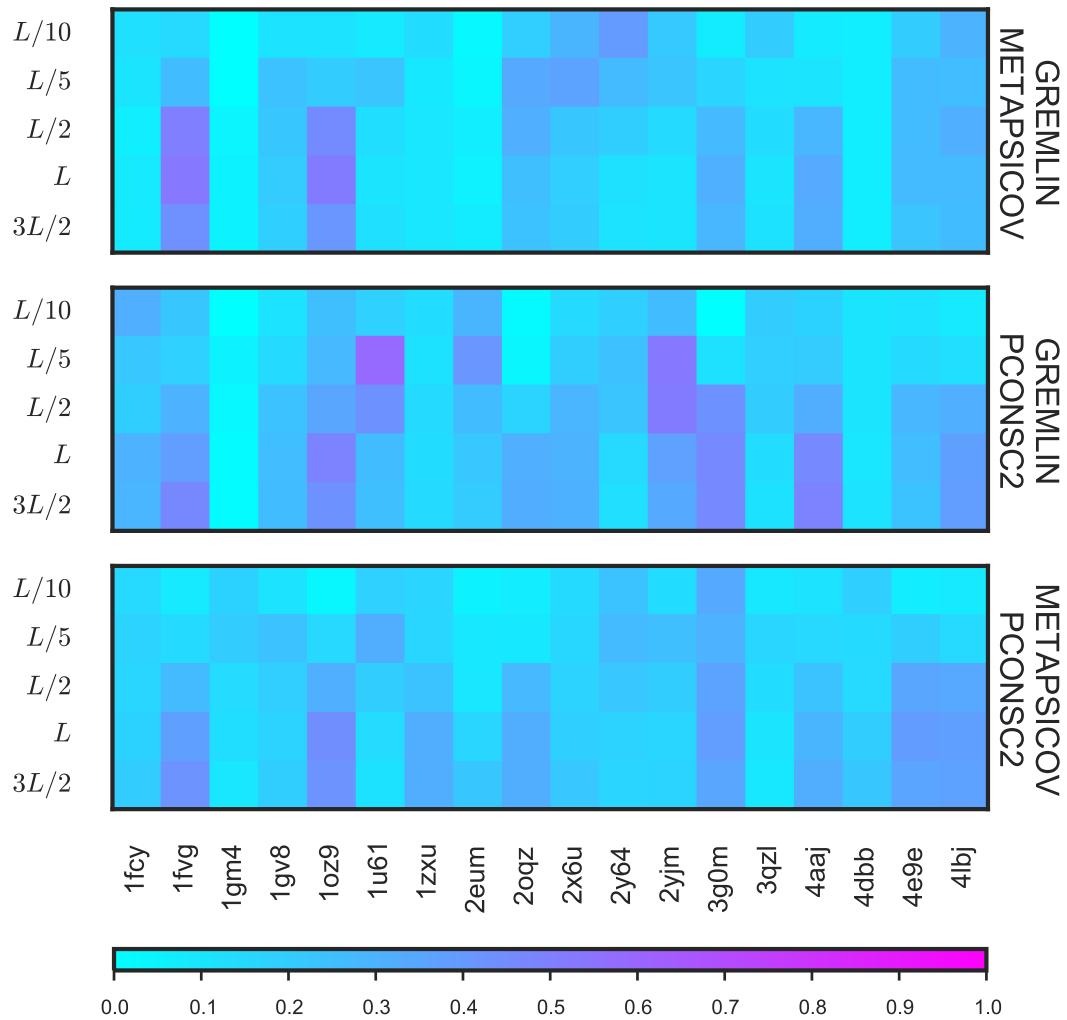


Figure 4.6: Jaccard similarity index illustrates a higher degree of overlap between metapredictor contact predictions with increasing numbers of contact pairs included in the calculation. The three panels show the different comparisons. The color scale corresponds to the Jaccard index in $[0, 1]$.

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 [117]. Thus, the quality of the decoys is of great essence to this study. Given the two different ROSETTA energy functions, FADE and SIGMOID, all contacts predicted 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 through all permutations of targets, contact predictions and ROSETTA energy function combinations.

Separating these individual decoys solely by the ROSETTA energy function (ex-

cluding unrestrained ROSETTA decoys) shows that the FADE energy function results in marginally more accurate decoys (median TM-score FADE: 0.3541; median TM-score SIGMOID: 0.2969). To further investigate which energy function is more suitable for the target dataset used in this study, the decoy sets were grouped by two additional characteristics: the fold of the target, and the source of distance restraints used. The results strongly suggest that the FADE energy function results in more accurate decoy sets (Fig. 4.7), 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 reveals 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 are similar, the actual differences in TM-scores further supports the strength of the FADE energy function compared to the SIGMOID.

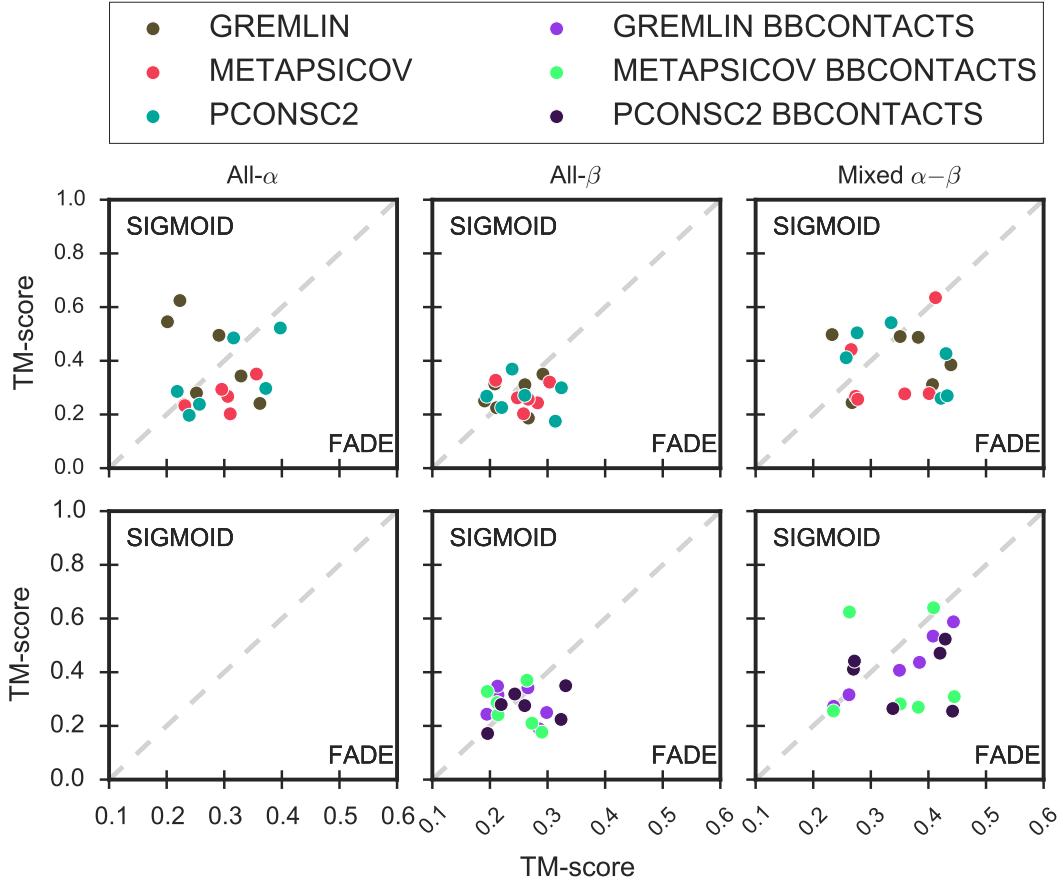


Figure 4.7: Median TM-score comparison of FADE and SIGMOID ROSETTA energy functions differentiated by fold and the addition of BBCONTACTS restraints.

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 predicts single decoys more accurately, it might be appropriate to reconsider the structure identifica-

tion routine (i.e. clustering) in AMPLE for search model preparation. However, a similar difference to that of the decoy quality of entire sets is observed for the top-1 decoy in each set (Fig. 4.8). The FADE energy function outperforms the SIGMOID function for the majority of target-contact prediction permutations (FADE: 51; SIGMOID: 39). However, the GREMLIN distance restraints in combination with the SIGMOID energy function produce better top-1 decoys than GREMLIN restraints with the FADE energy function. This suggests 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 (Δ Median TM-score_{ALL}: $\min = 0.002$, $\max = 0.429$; Δ Median TM-score_{TOP}: $\min = 0.002$, $\max = 0.456$).

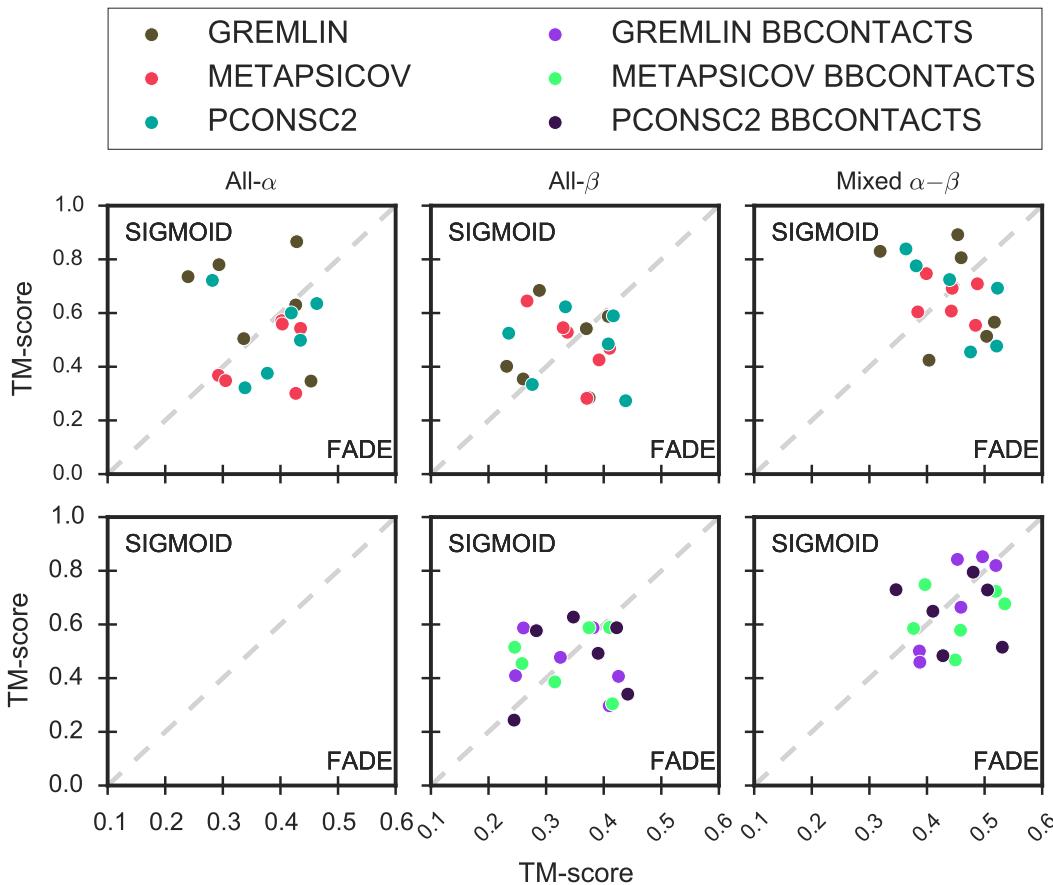


Figure 4.8: Top TM-score comparison of FADE and SIGMOID ROSETTA energy functions differentiated by fold and the addition of BBCONTACTS restraints.

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. 4.9). This density estimate further supports the results presented above: the FADE energy function generates more accurate decoys. However, a very important detail is highlighted by the estimates. 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. 4.9). The bimodal distribution of decoy TM-scores from both energy functions strongly suggests that predicted structures are 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 is in favour of the latter. The decoy sets of all- β targets do not show such distinct regions of high density for decoys with TM-scores < 0.5 units in any of its density estimates (Fig. 4.9). The generally poor decoy quality of decoys predicted without any distance restraint information (ROSETTA) highlights the benefit of contact predictions to *ab initio* protein structure prediction.

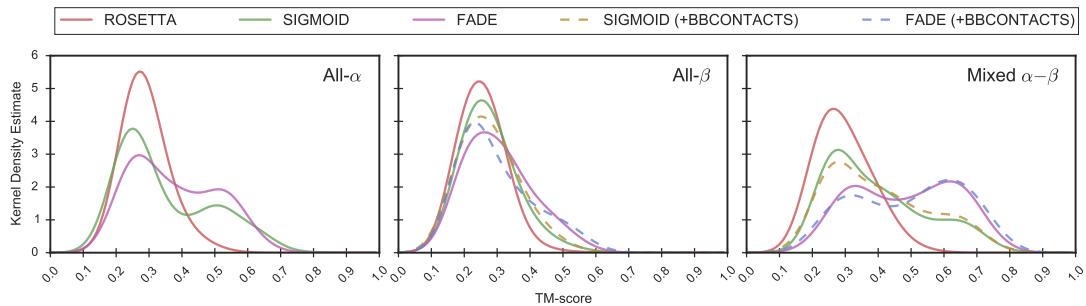


Figure 4.9: TM-score density estimate of all decoys in each respective fold class separating by ROSETTA energy function (SIGMOID or FADE) and no contact information used (ROSETTA). Dashed lines indicate decoys which were predicted with the addition of BBCONTACTS.

A further important aspect of this study is to explore the benefits of adding BBCONTACTS restraints to the structure prediction of β -containing targets. Although previous results ?? in combination with those presented above outline overall improvements in decoy quality, it is essential to understand which targets benefit from this treatment. Figure 4.10a highlights the effects of adding BBCONTACTS restraints to the structure prediction strategies employed here. In summary, the addition of BBCONTACTS restraints hardly affects 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 units for any target (Fig. 4.10b).

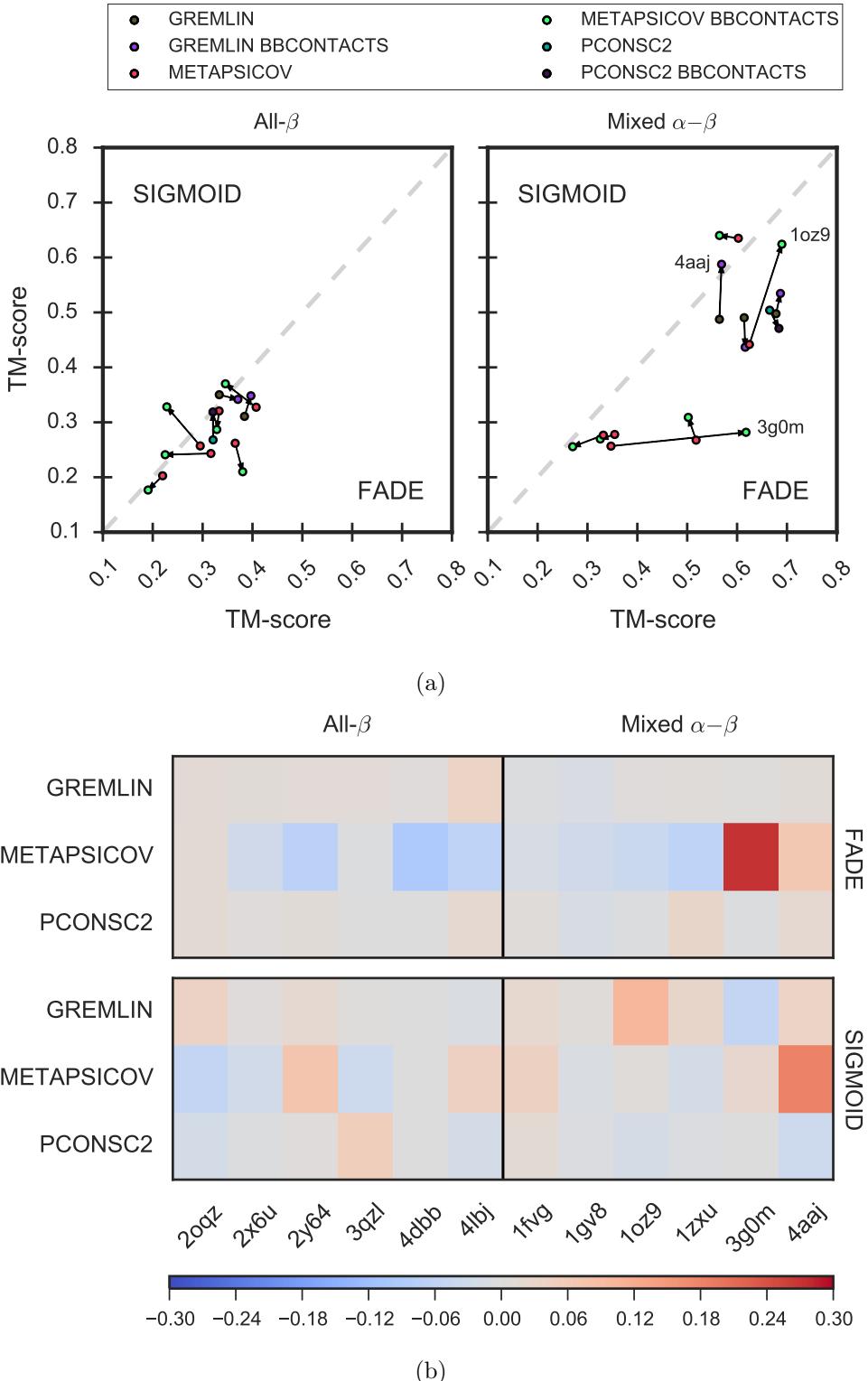


Figure 4.10: Median TM-score comparison of FADE and SIGMOID ROSETTA energy functions differentiated by fold (excl. all- α). (a) Arrows indicate the effect on decoy quality through the addition of BBCONTACTS restraints. Targets with a distance < 0.03 TM-score units between normal and BBCONTACTS-added conditions were excluded from the scatter plots. (b) Effect on decoy quality through the addition of BBCONTACTS restraints highlighted by heatmap difference. The color 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 are the target chain length and restraints precision. The former appears to affect the final decoy quality of all 1,000 decoys insignificantly (Fig. 4.11). However, the restraint precision results in some differences between the two ROSETTA energy functions (Fig. 4.11). The FADE energy function (L restraints) generally appears to be less sensitive to restraint lists with higher false positive contact pairs. In contrast, the SIGMOID function ($3L/2$ restraints) produces 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: 4aaaj) compared to the SIGMOID function with a median TM-score of 0.498. Nevertheless, both energy functions appear to broadly follow a positive linear trend, i.e. better restraint precision results in more accurate decoys.

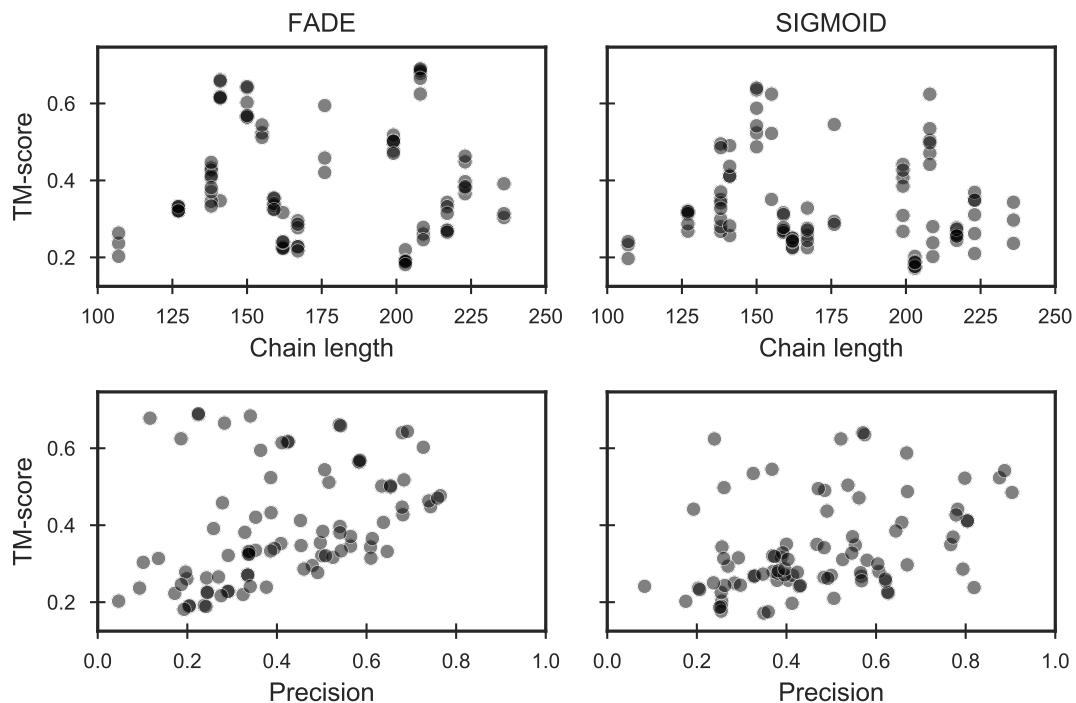


Figure 4.11: Effects of target chain length and restraint precision on the median TM-score for FADE and SIGMOID ROSETTA energy functions. Each scatter point represents a 1,000-decoy set.

4.3.3 Impact of metapredictors and energy functions on unconventional Molecular Replacement

The results obtained from the decoy quality comparison outlined above highlighted differences between the FADE and SIGMOID ROSETTA energy functions. This difference is more pronounced for some targets and less so for others. 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 structure prediction process yielded six out of 18 possible structure solutions (Fig. 4.12). 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: 1fgv) 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.

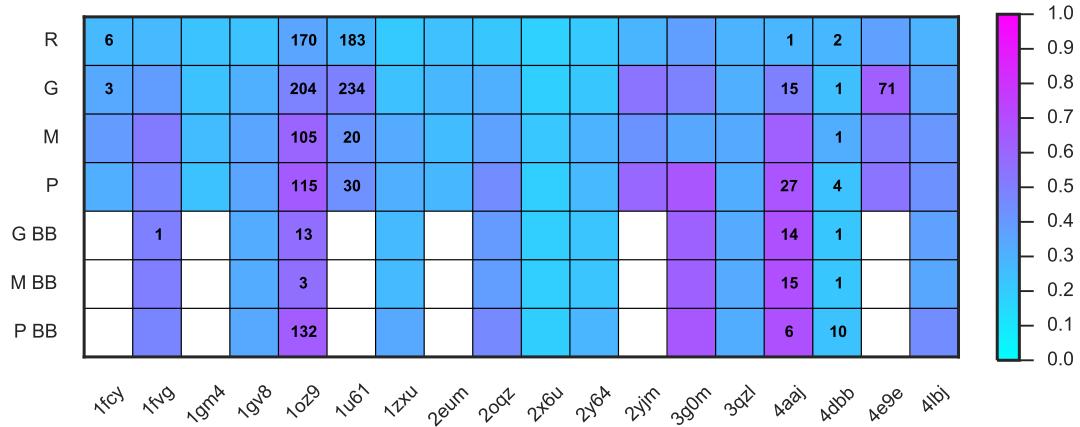


Figure 4.12: Structure solution count for AMPLE search models generated from decoys with varying contact prediction and ROSETTA energy function conditions: unrestrained ROSETTA (R); GREMLIN (G; SIGMOID function); METAPSICOV (M; FADE function); PCONSC2 (P; FADE function); GREMLIN BBCONTACTS (G BB; FADE function); METAPSICOV BBCONTACTS (M BB; FADE function); PCONSC2 BBCONTACTS (P BB; FADE function). The color 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 are somewhat surprising given that ROSETTA decoys result in the second-most structure solutions. These results suggest that the current implementation cannot exploit the true value of more accurate decoy sets. This hypothesis is further supported when considering the decoy set quality and the number of structure solutions (Fig. 4.12). For example, PCONSC2 (FADE function) decoys predicted for

the hypothetical protein AQ_1354 (PDB: 1oz9) yield high accuracy, and thus would generally be considered highly desirable starting structures for the AMPLE protocol; nevertheless, the AMPLE protocol was unable to exploit such highly accurate decoys for successful structure solutions of other targets, e.g. cysteine desulferation protein SufE (PDB: 3g0m; median TM-score PCONSC2 BBCONTACTS (FADE function)=0.661). In comparison, the median TM-scores for all successful ROSETTA decoy sets do not exceed 0.355 TM-score units.

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 demonstrate that the most accurate decoys do not guarantee structure solution, and in contrast some poorly predicted decoy sets achieve structure solution. Thus, it is essential to investigate the stage in AMPLE’s cluster-and-truncate approach at which the higher decoy quality results in less suitable ensemble search models for MR.

The data generated as part of this study reveals a positive correlation ($\rho_{Spearman} = 0.78$; $p < 0.001$) between the decoy quality and the number of resulting AMPLE ensemble search models (Fig. 4.13). The plotted data alongside a line of best fit further illustrate that small differences in decoy quality in the lower TM-score regions increases the total number of generated ensemble search models dramatically. However, once the threshold of 0.5 TM-score units [140] is surpassed the number of generated ensemble search models plateaus at around 350-400 ensemble search models, approaching the maximum number of search models generatable by AMPLE. Furthermore, the data suggests that 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 will lead to structure solution.

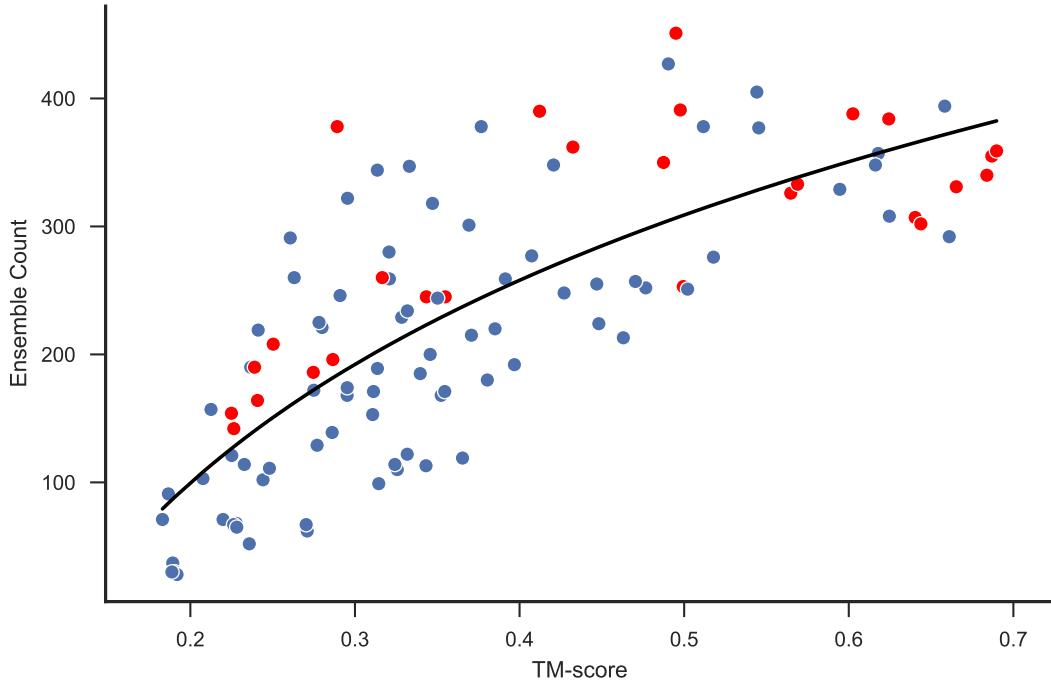


Figure 4.13: Comparison of median TM-score (per 1,000 decoys) against the resulting AMPLE ensemble search model count. The equation of the line of best fit is defined by $y = 228.50 * \ln(20.96 * x) - 227.95$. Red dots indicate successful ensemble sets.

Besides looking at the relationship between entire decoy sets and the resulting structure solutions on a per-target or per-condition basis, it is important to also consider individual ensemble search models, their origins and their properties in relation to MR metrics. Previous findings highlighted the relationship between the number of decoys in the first cluster and the quality of the decoys it contains (see Chapter 3). Here, we further support these findings given the positive relationship between the median TM-scores and the corresponding size of the largest SPICKER cluster (Fig. 4.14). An analysis of the cluster sizes demonstrates the downstream benefits of increased decoy quality through contact restraints in the folding process (Fig. 4.15). The sizes of the first three clusters generated from most contact-restraint decoy sets greatly surpass their equivalent cluster sizes for unrestrained ROSETTA decoys. Given that cluster sizes correlate with decoy quality, the findings in this study also support that the mean C α RMSD — as calculated by THESEUS for cluster truncation — is directly related to better decoy quality via the larger number of decoys in each cluster (Fig. 4.16a). The same mean C α RMSD is also related to the number of ensemble search models generated after subclustering (Fig. 4.16b), which hints 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 show similar C α RMSD per cluster compared to unrestrained ROSETTA decoys (Fig. 4.17), unlike all other contact restraint guided structure predictions. However, it is worth noting that almost no distinction can be made amongst the remaining contact restraint treatments albeit

some differences in cluster size distributions exist (Fig. 4.15).

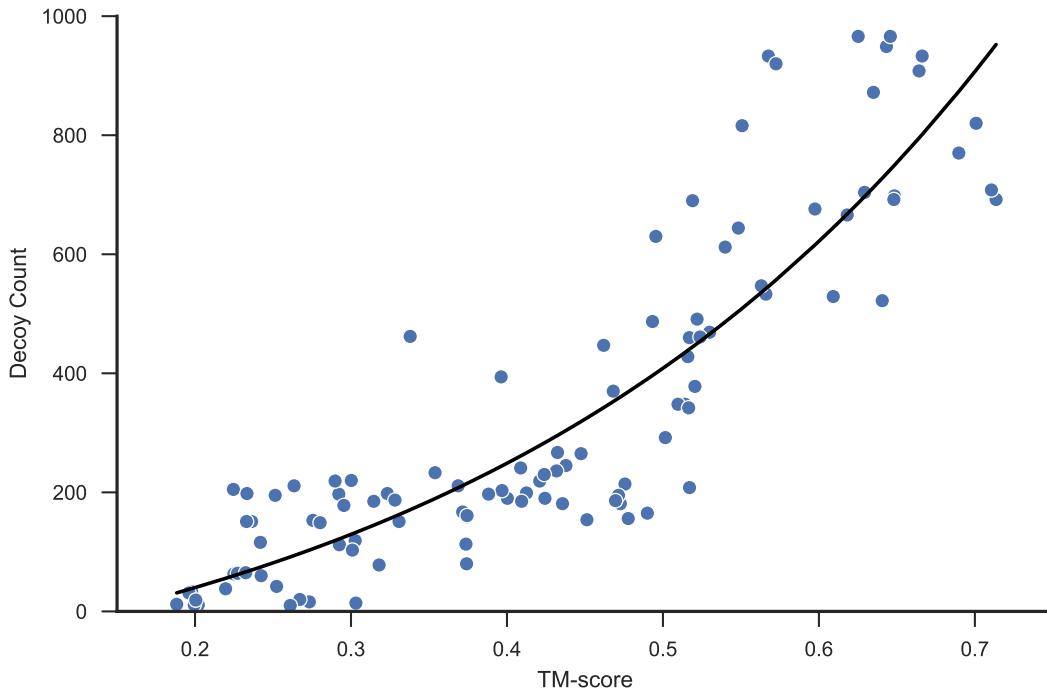


Figure 4.14: Relationship between cluster median TM-score and the number of cluster decoys. Blue line represents line of best fit with equation $y = 148.85 * \exp(2.90 * x) - 225.76$.

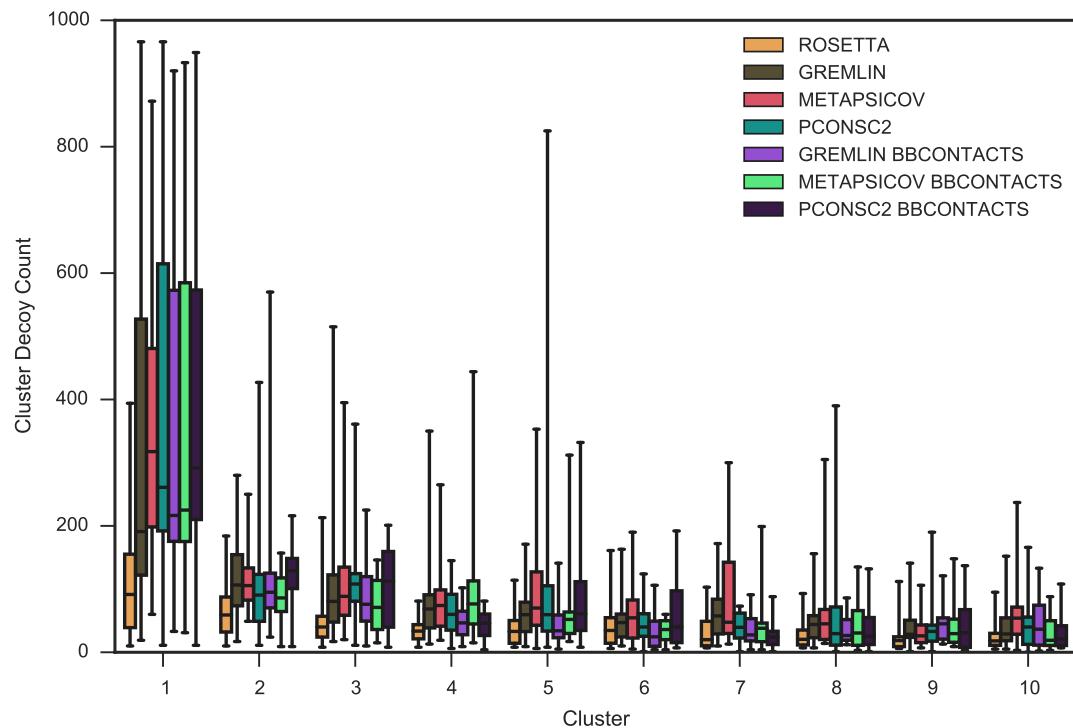


Figure 4.15: SPICKER cluster sizes of each target grouped the restraint condition used during the structure prediction protocol. Whiskers span the range from the minimum to maximum counts.

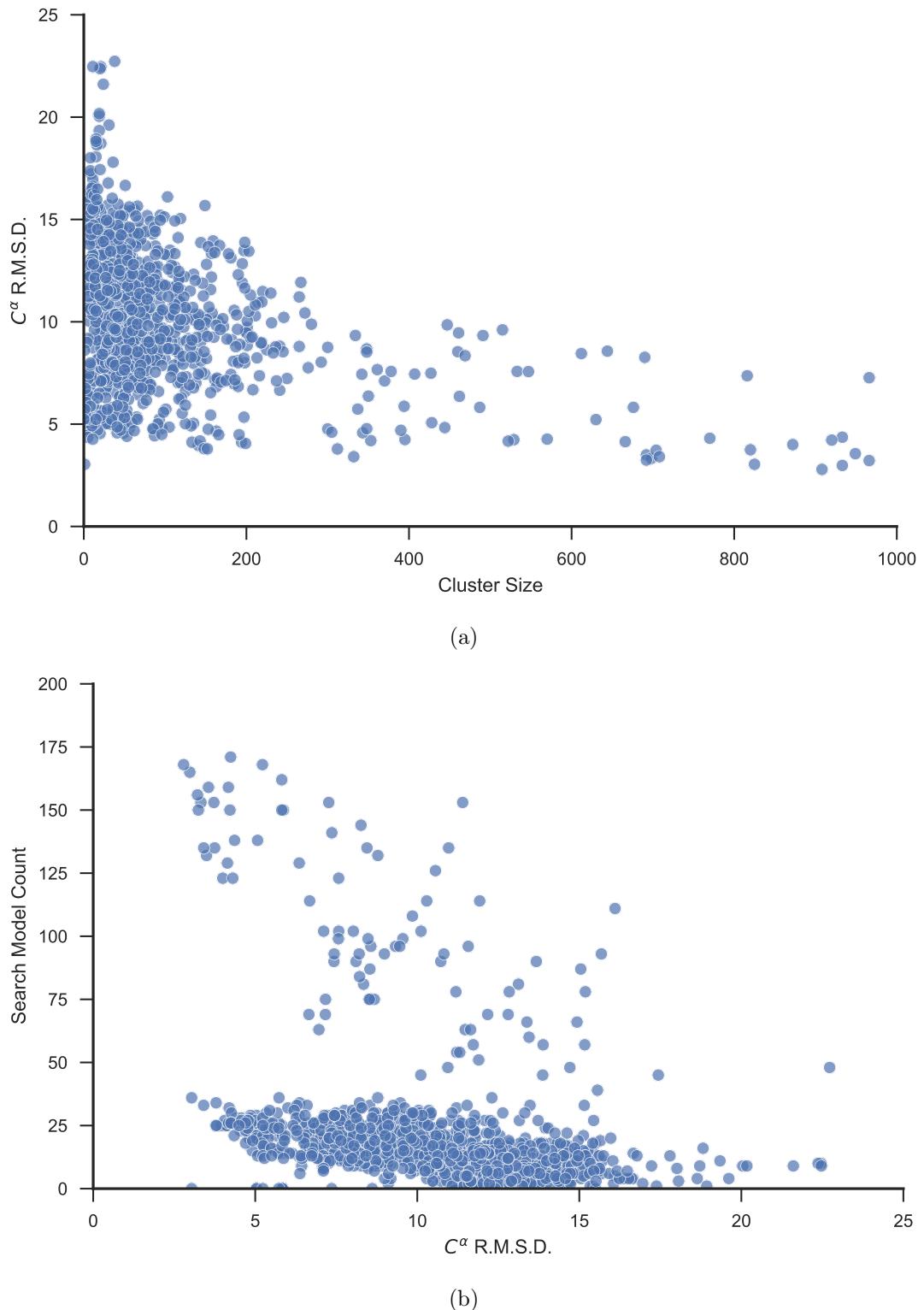


Figure 4.16: (a) Number of decoys per SPICKER cluster plotted against the mean Ca-atom RMSD for all decoys in each cluster. (b) Mean Ca-atom RMSD for decoys per cluster plotted against the number of search models derived from the cluster.

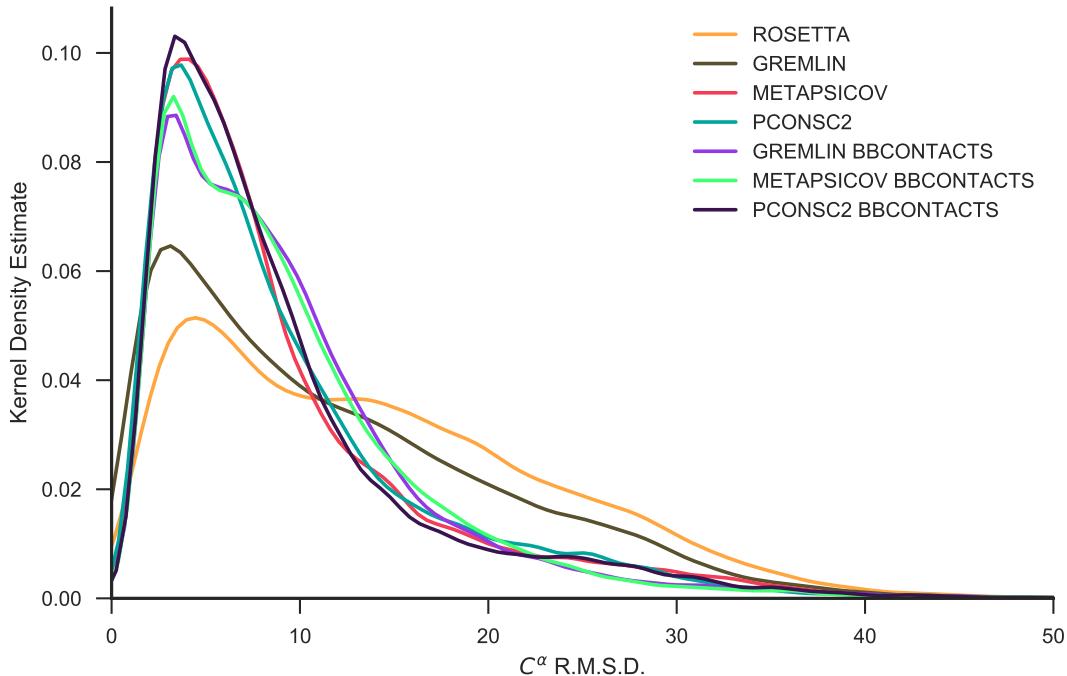


Figure 4.17: Kernel density estimate of $C\alpha$ interatomic RMSD for SPICKER clusters.

The structure solution through pipelines like AMPLE and other unconventional MR software [144, 145] can result from the placement of generated (ensemble) search models either in- or out-of-sequence register. The RIO metric [116] 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 support the high quality decoys used as input across all seven contact conditions (Fig. 4.18). Most search models are 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 — shows a low percentage of AMPLE search models with RIO scores leading to structure solution (Fig. 4.18). Furthermore, the RIO scores normalized by the target chain length indicate that search models, independent of MR structure solution, were relatively small only exceeding 20% of the total target sequence in a few cases.

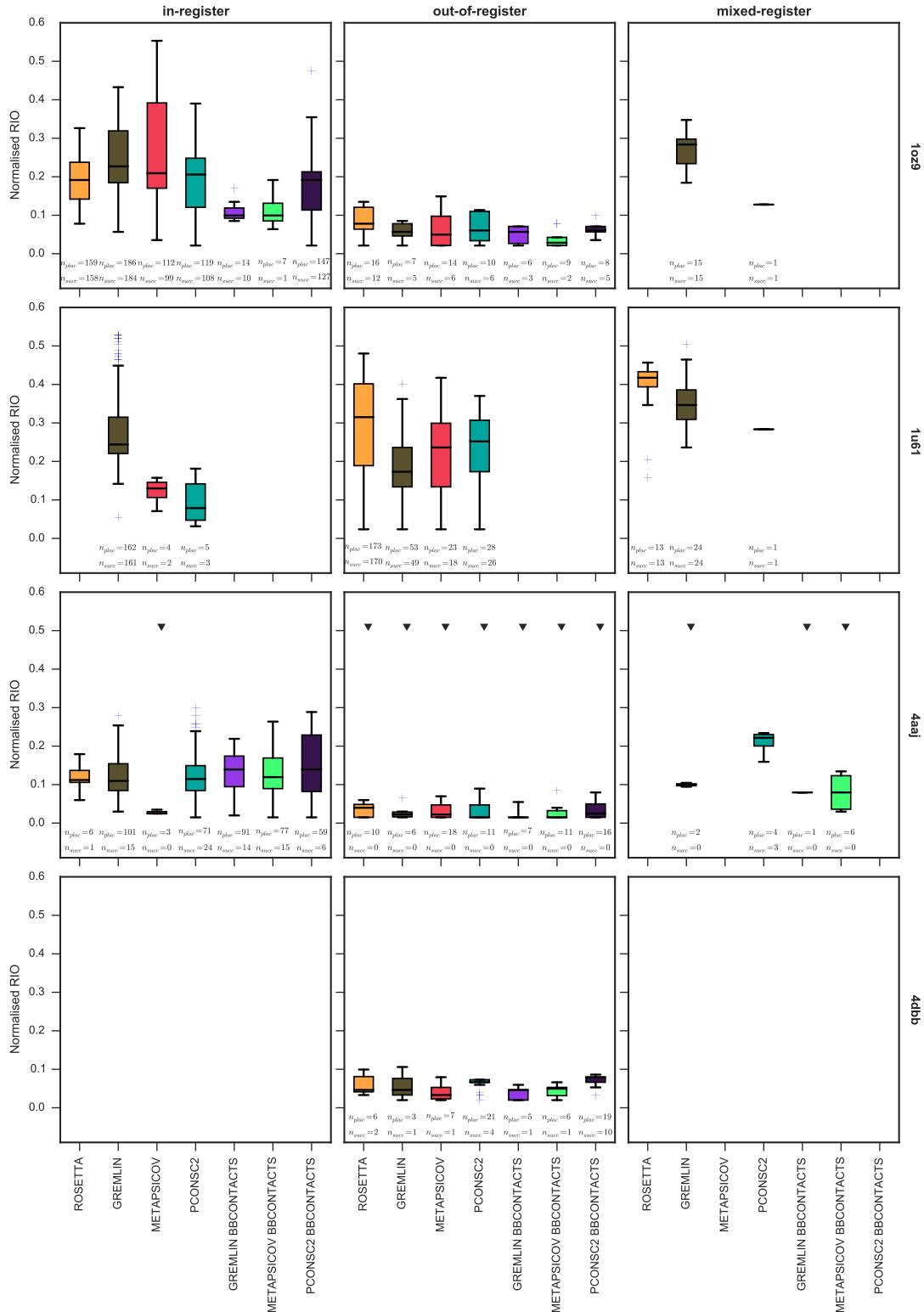


Figure 4.18: Normalised RIO score analysis of four successful targets in the MR dataset. Black triangles indicate AMPLE search model sets without a structure solution.

One interesting target in this set with respect to the sequence register of the AMPLE search models leading to structure solution is putative ribonuclease III (PDB: 1u61). Although decoys from all contact conditions readily solved this target with at least 20

or more AMPLE search models, one interesting aspect arises from the RIO register analysis. Only GREMLIN decoys are primarily placed in-register (Fig. 4.18). AMPLE search models derived from the other three contact conditions, and in particular those from ROSETTA decoys, are primarily placed out-of-register with sequence coverage values of roughly 25%. In fact, a close analysis of the diversity of AMPLE search models highlights the accuracy of GREMLIN search models which represent a closely-matched substructure of the target protein (Fig. 4.19).

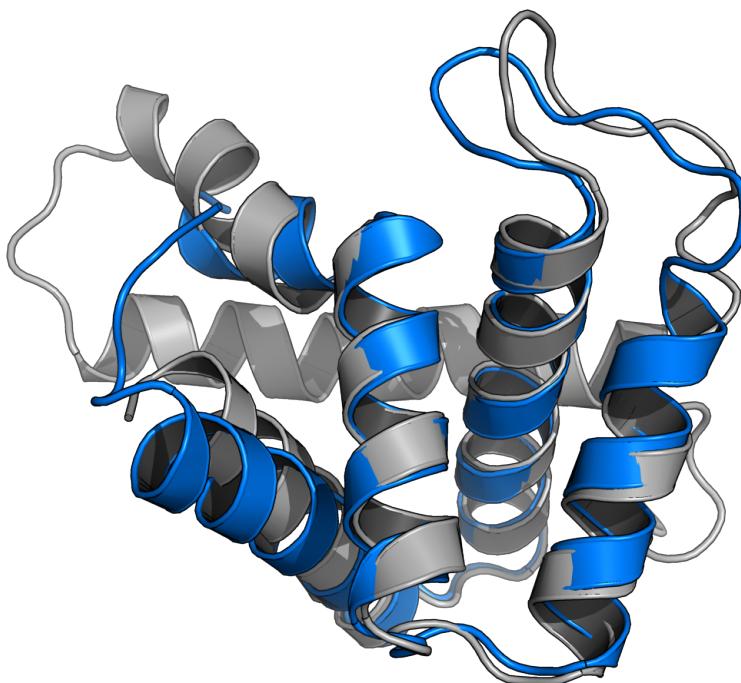


Figure 4.19: Successful search model (blue cartoon) post-PHASER placement superposed with the native structure (gray cartoon) for putative ribonuclease III (PDB: 1u61).

Compared to all other targets with structure solutions in at least one condition, the PTB domain of Mint1 (PDB: 4dbb) produced interesting yet somewhat surprising results. None of the search models, independent of their decoy source, achieved correct placement with any residue being in register. All structure solutions were obtained from out-of-register search model placements (Fig. 4.18). 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. 4.20). 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. 4.20).

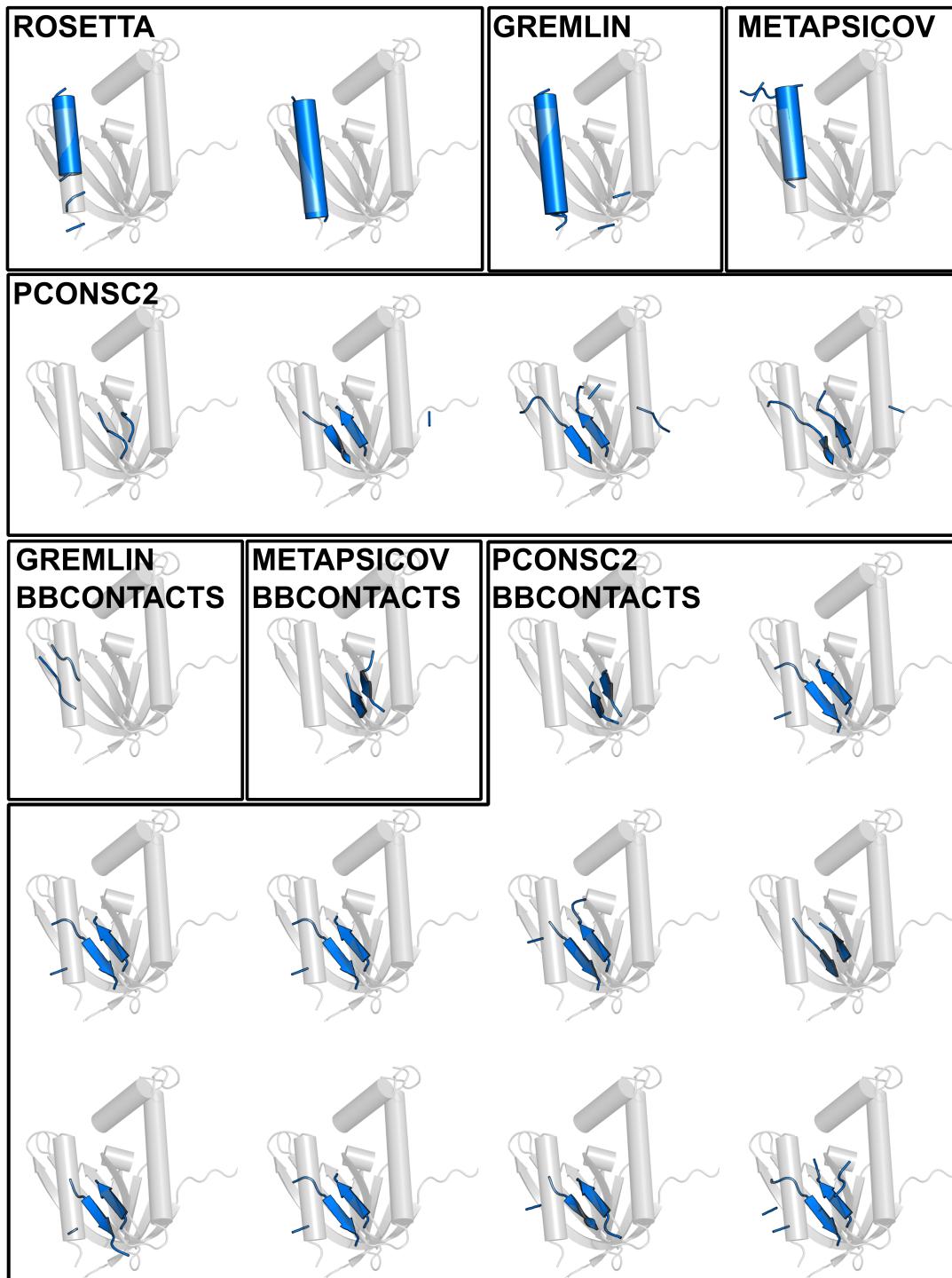


Figure 4.20: Successful search models post-PHASER placement (blue) superposed to the reference crystal structure (grey) for PTB domain of Mint1 (PDB: 4dbb).

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 AMPLEs cluster-and-truncate procedure producing a single, idealised helix as search model. Here, we

confirm such findings 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: 1fgv) was placed in-register. A closer inspection of this AMPLE ensemble search model highlights a great success of the approach of adding BBCONTACTS distance restraints to separately predicted contact maps. In this instance, the successful AMPLE ensemble search model has 77% of its 49 residues placed in-register. More importantly, the search model is made up of two β -strands packing against each other, which was supported by BBCONTACTS predictions (Fig. 4.21). 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.

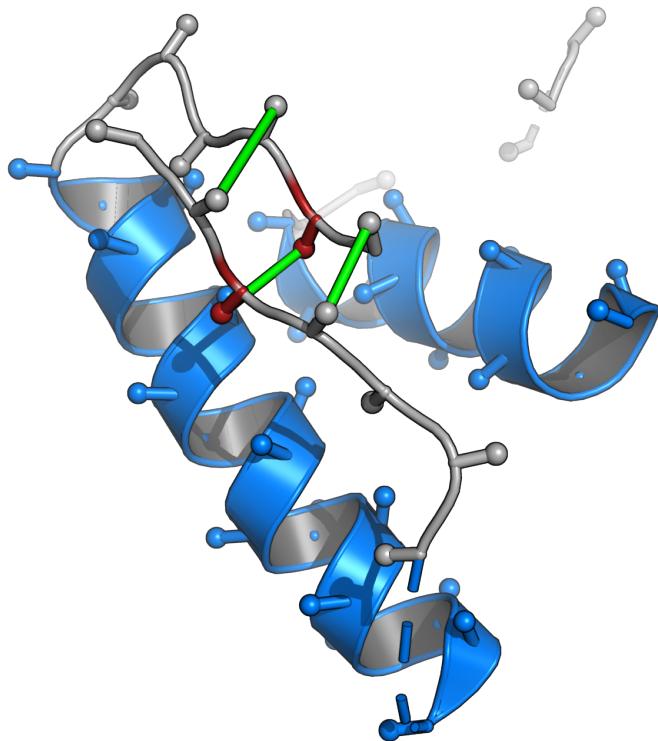


Figure 4.21: Successful search model post-PHASER placement for peptide methionine sulfoxide reductase (PDB: 1fgv). BBCONTACTS distance restraints are represented as green lines, α -helices in blue and β -strands in red. Secondary structure assignment calculated with STRIDE [146].

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 [111, 112]. However, work in this study cannot confirm their findings, which demonstrate 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. [112]: 680; De Oliveira et al. [111]: 3500); (2) the targets were chosen to deliberately sample various alignment depths including relatively low Neff (< 200) values; (3) only final contact predictions were analysed as part of this work, thus benefiting from post-prediction consensus finding and contact map processing through unsupervised machine-learning algorithms.

Furthermore, we demonstrated in this study 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 [90]. 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 structure prediction.

Lastly, part of the comparison carried out in this study was aimed specifically at macromolecular crystallographers and, in particular, AMPLE users. Beyond the proof-of-principle study described in Chapter 3, this work further illustrates how important additional restraint information can be to increase the chances of unconventional MR success. However, this work also highlighted limitations in the AMPLE routine whereby decoys that were restrained by 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 for a large set, 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 is 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, 111] to specifically eliminate the worst decoys, and thus enhance a more fine-grained clustering approach in SPICKER. 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 [113, 114, 116, 117]. This recommendation is based primarily on the superiority of the decoy quality compared to other modelling algorithms, which was recently reaffirmed by the lastest CASP12 experiments [147, 148]. However, Keegan et al. [114] 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 [114]. 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 [147], other algorithms have been developed over the last two decades [e.g., 43, 69, 142, 149–151]. Although most of these algorithms utilise fragment-assembly algorithms similar to ROSETTA and QUARK, their procedures for fragment selection or assembly is substantially different [149, 150]. 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, 142, 151].

CONFOLD2 [143], 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, 143].

FRAGFOLD [149], 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 li-

brary 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 [143], FRAGFOLD [149] and SAINT2 [46], were explored 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 [100].

Secondary structure and solvent exposure were predicted using PSIPRED v4.0 [152] and SOLVPRED (shipped with METAPSICOV v1.04), respectively. The MSA for coevolution-based contact prediction was generated using HHBLITS v2.0.16 [135] against the uniprot20 v2016-02 database. CCMPRED v0.3.2 [91], FREECONTACT v1.0.21 [153] and PSICOV v2.1b3 [88] were used by METAPSICOV to generate contact predictions.

METAPSICOV STAGE 1 contact predictions were used in *ab initio* structure prediction since they result in more accurate structure predictions than METAPSICOV STAGE 2 results [100].

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 Chapter 3 or Michel et al. [70].

The CONFOLD2 decoys were generated using CONFOLD2 v2.0 [143], which uses Crystallography & NMR System (CNS) v1.3 [154] to drive the modelling. Default parameters were used except for the number of decoys per run, which was increased from 20 to 25 with `-mcount 25`. 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.

The FRAGFOLD decoys were generated using FRAGFOLD v4.80 [149] 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 generates on average 30 fragments per target position that are 6 to 20 residues long. 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].

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 [113] was shown to be advantageous in most cases by Thomas [155], 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 < 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 [143], FRAGFOLD [149] 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 [100] 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 [100]. 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 ranges from 431 to 6,186 across all targets (Fig. 5.1). Six targets contain at least 200 and less than 1000 sufficiently-diverse sequences, whilst the remaining 16 targets contain more than 1,000 effective sequences.

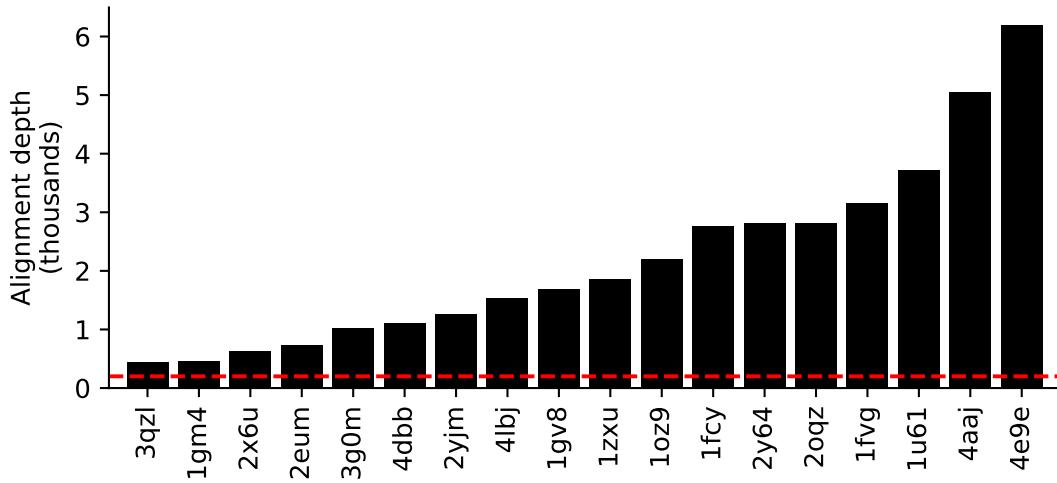


Figure 5.1: Distribution of HHBLITS alignment depth for subset of targets in the PREDICTORS dataset. Red line indicates alignment depth threshold of 200 sequences [83].

Coevolution-based contact predictors rely heavily on the alignment depth for accurate contact predictions. In this work, these findings are further confirmed. Sequence alignments with depths of less than 1,000 sequences produce 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 is found (Spearman's $\rho = 0.57$, p-value < 0.02). A moving average analysis shows that those contact predictions based on alignments with more than 1,000 effective sequences yield better precision scores of at least 0.09 units up to 0.34 units. The difference between the two moving average curves highlights that the difference is greater at lower cutoff values, i.e. only the very best contacts are included in the selection. This difference declines more drastically for targets with deeper alignments (Fig. 5.2).

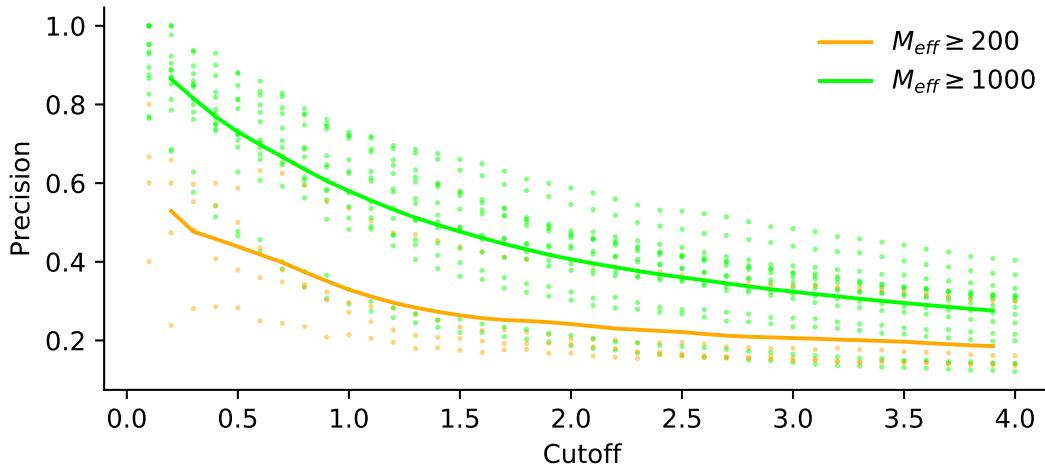


Figure 5.2: Contact precision analysis for numerous contact selection cutoffs for targets with alignment depths of more than 200 and more than 1,000 sequences. Lines indicate moving averages for both categories with a window size of three residues. M_{eff} refers to the alignment depth (number of effective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

5.3.2 Comparison of decoy quality

One main interest of the work presented in this chapter is the comparison of the quality of decoys predicted with four *ab initio* structure prediction algorithms. To-date, no such direct comparison exists on the same dataset, and thus might provide direct insights into the performance of each.

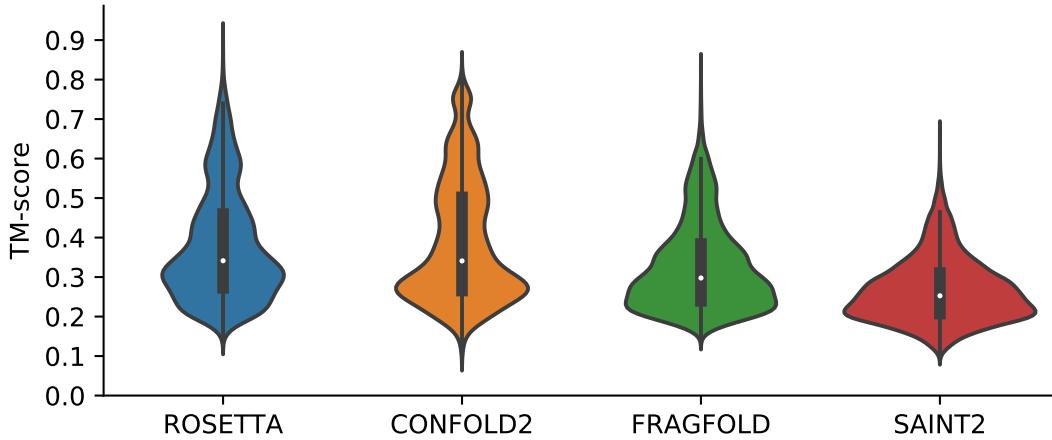


Figure 5.3: Kernel Density Estimate (KDE) of decoy TM-score for four different *ab initio* structure prediction algorithms, namely ROSETTA, CONFOLD2, FRAGFOLD and SAINT2. CONFOLD2 contains 9,000 less decoys than the remaining algorithms (for further details refer to Section 5.2.3).

An initial comparison of overall performance highlights that ROSETTA generates 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 shows that ROSETTA and CONFOLD2 contain 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.

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

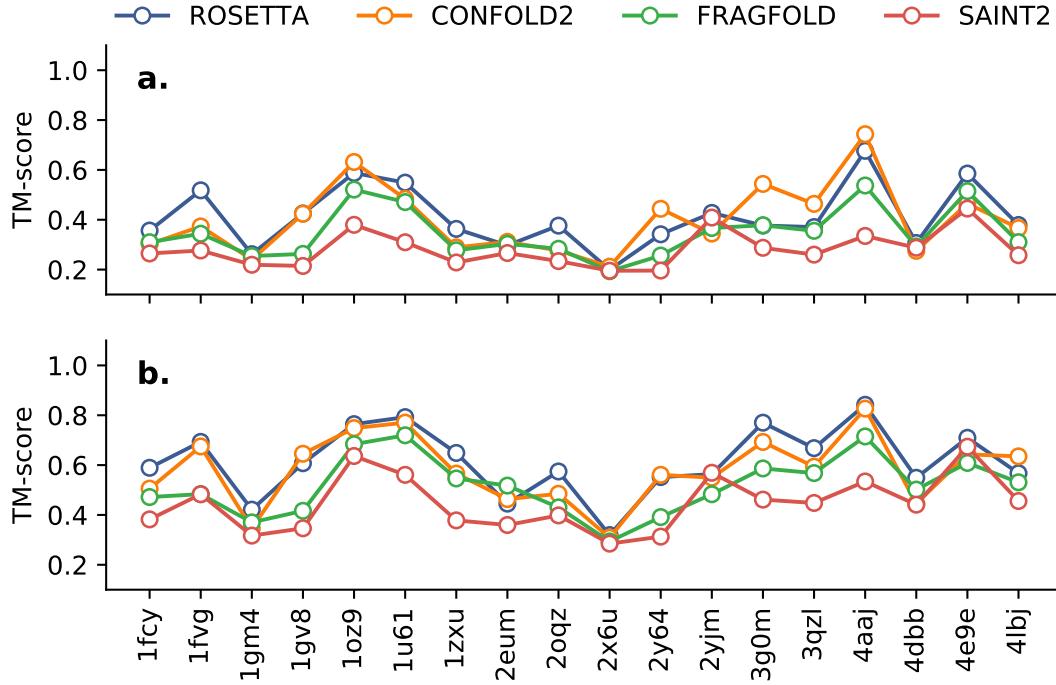


Figure 5.4: Per-target TM-score analysis for targets modelled with contact information and four separate *ab initio* structure prediction algorithms. Analysis is subdivided by (a) median TM-score of all decoys in each set and (b) TM-score of the top-1 decoy in each set.

Abriata et al. [147] recently attributed the success in the CASP12 experiments to improved accuracy of coevolution-based contact predictors and the availability of many more sequence homologs than ever before. Thus, it is 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.

The results obtained in this study further support the conclusions made by Abriata et al. [147] but only for the ROSETTA algorithm. A Spearman's rank-order CC analysis of alignment depth and median TM-score shows a significant positive correlation for ROSETTA-generated decoy sets (Spearman's $\rho = 0.68$, $p < 0.01$). This positive correlation is 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 do not show a significant correlation, although better decoy sets are 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.

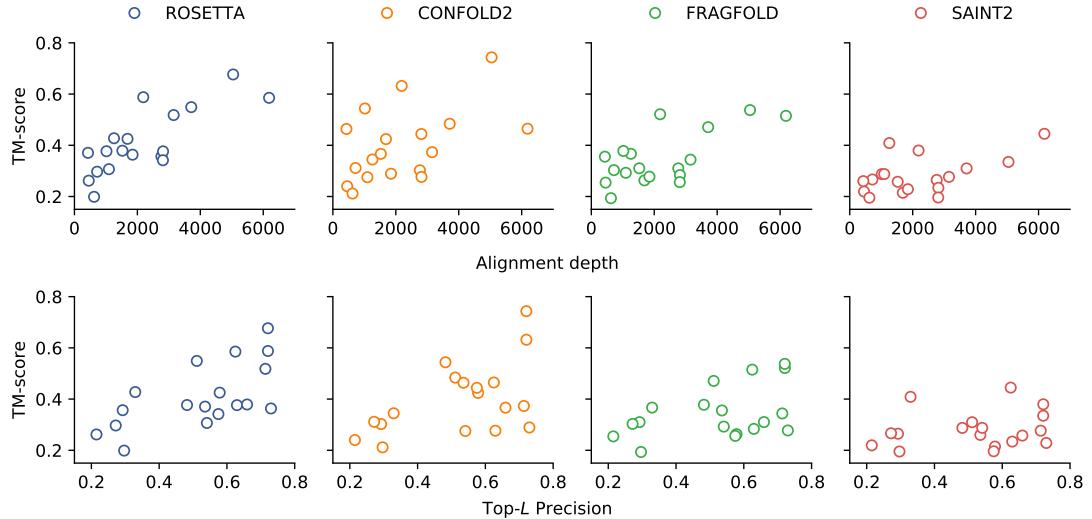


Figure 5.5: Analysis of median TM-score of the contact-based decoy sets and their dependence on alignment depth and top- L precision.

Beyond the use of contact information, parts of this study explored the performance of ROSETTA, FRAGFOLD and SAINT2 when no contact information is provided as distance restraints in *ab initio* structure prediction (CONFOLD2 requires contact information, and thus was excluded). ROSETTA performs 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 is marginal for the majority of cases. The median values for eight ROSETTA and FRAGFOLD decoy sets differ 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 differ greatly between the modelling algorithms, whilst the rest is near identical (Fig. 5.6).

The top decoy predicted by ROSETTA and SAINT2 based on the sequence of the FAT domain of focal adhesion kinase (PDB ID: 1k40) differs by 0.35 TM-score units. More significantly though, the top 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 predict incorrect folds. A near-identical scenario applies to the top decoys of the Hypothetical protein PF0907 (PDB ID: 4pgo) (ROSETTA TM-score = 0.68; FRAGFOLD TM-score = 0.27; SAINT2 TM-score = 0.39).

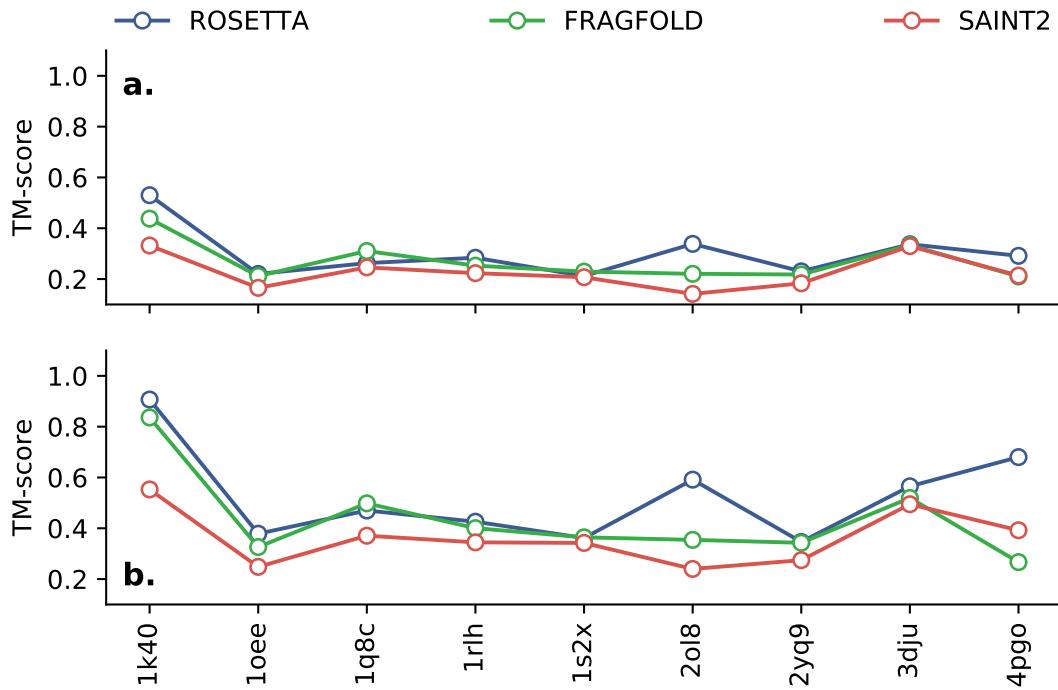


Figure 5.6: Per-target TM-score analysis for targets modelled without contact information and four separate *ab initio* structure prediction algorithms. Analysis is subdivided by (a) median TM-score of all decoys in each set and (b) TM-score of the top-1 decoy in each set.

An analysis of the modelling results by target fold shows that all- α and mixed α - β target folds are 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 show 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 predicts the overall target fold correctly or incorrectly, the data might be misleading due to missing targets in the dataset in comparison to the other three modelling algorithms. Lastly, the distributions of TM-score for either fold class of SAINT2 decoys 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.

A further subdivision of all target decoys is by target chain length. At the stage of target selection, three main bins were defined from which targets were randomly sampled (see Section 4.2.1). These bins were defined with target chain length edges of 150 and 200 creating three bins: $0 < x < 150$ & $150 \leq x < 200$ & $x \geq 200$ (x refers to the target chain length). A grouping of the decoy TM-score by algorithm and target

chain length indicates little difference in modelling difficulty (Fig. 5.7). Each of the modelling algorithms shows the largest spread for targets with chain lengths in the bin $150 \leq x < 200$. 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 ($n_{\text{res}}=216$). The set of CONFOLD2 decoys for N-(5-phosphoribosyl)anthranilate isomerase (PDB ID: 4aaaj) have a median TM-score of 0.74 units. ROSETTA decoys show 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.

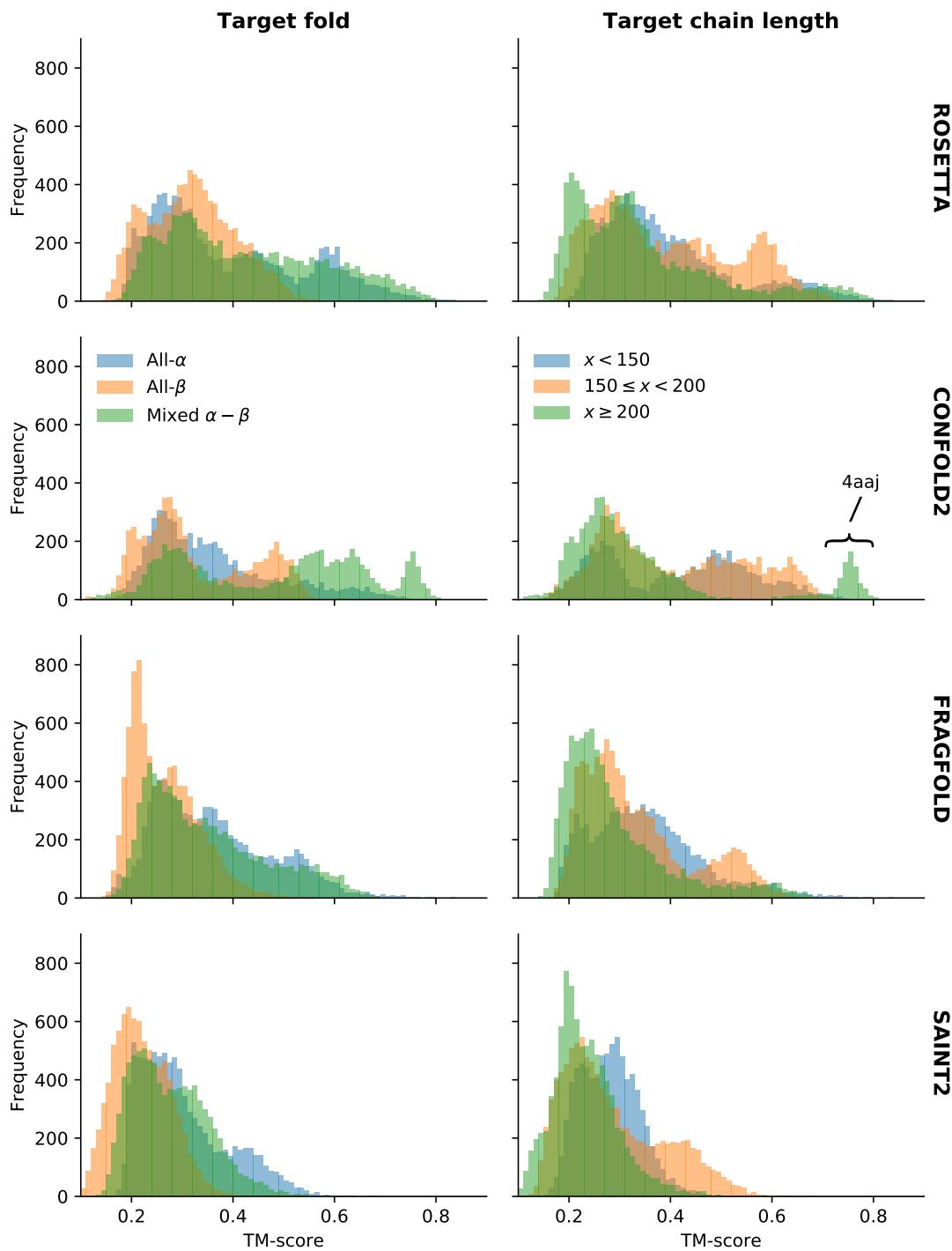


Figure 5.7: Distribution of decoy TM-scores by fold, chain length and algorithm.

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 so-

lutions 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.

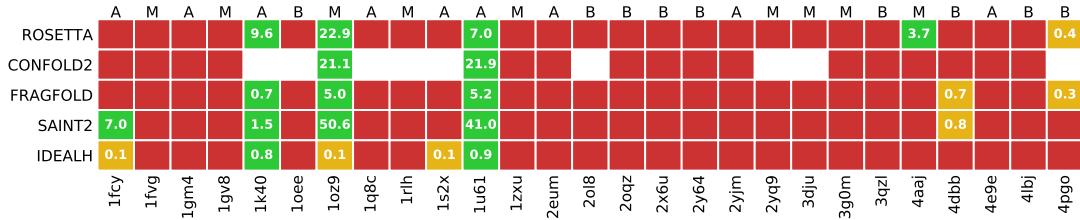


Figure 5.8: Summary of MR success with AMPLE ensemble search models. Search models are based on decoy sets generated with different *ab initio* structure prediction protocols. The colour coding indicates structure solution: no solution (red), one solution (orange), more than one solution (green). The 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.

The chain length for targets with structure solutions ranges 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 solve on average 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 did not solve 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).

The difference in overall decoy quality between the four different *ab initio* structure prediction algorithms is further noticed in the successful AMPLE-generated ensemble search models. ROSETTA decoys result in more complete AMPLE ensemble search models, which lead to structure solution (Fig. 5.9). Although CONFOLD2 has 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 has the shortest range spanning from 8% to 70% target completeness.

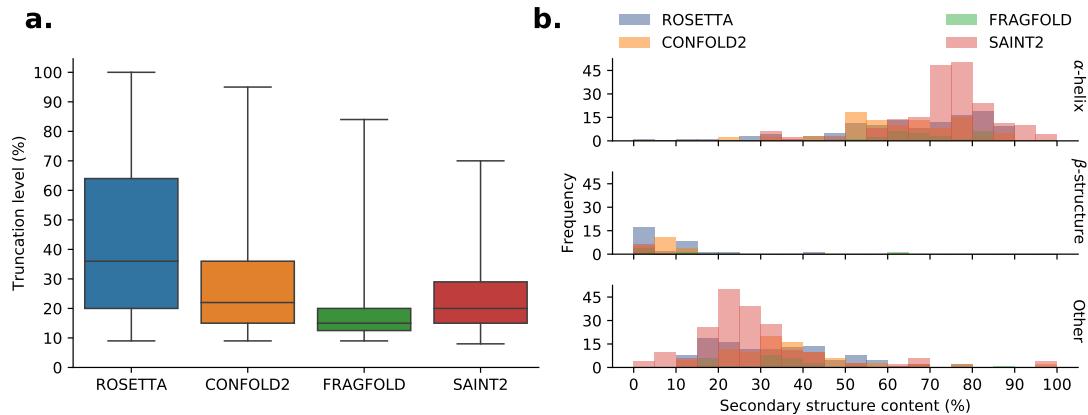


Figure 5.9: Distribution of (a) search model truncation and (b) secondary structure content for successful AMPLE ensemble search models given decoys from four *ab initio* structure prediction algorithms. Secondary structure for each ensemble search model evaluated with DSSP [146].

An inspection of the secondary structure content of all successful ensemble search models outlines an important difference between SAINT2 and the other three modelling algorithms. Successful search models derived from SAINT2 decoys are predominantly α -helical (Fig. 5.9). An analysis of the secondary structure makeup, as assigned by DSSP [146], shows that successful SAINT2 search models contain approximately 70-80% α -helices with the rest being unassigned secondary structure. In comparison, the successful ensemble search models from other modelling algorithms contain a range from 50-90% α -helices, whilst the remainder is 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 save the great overhead of predicting, preparing and sampling decoys in AMPLE. A visual inspection of SAINT2 ensemble search models highlights 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 [116] shows that the former could have been solved without the modelling overhead. 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 are not strictly α -helical, and thus would require more sophisticated and computationally complex idealised-fragment library generation protocols [e.g., 156] or libraries of recurring tertiary structure motifs [e.g., 145]. 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.

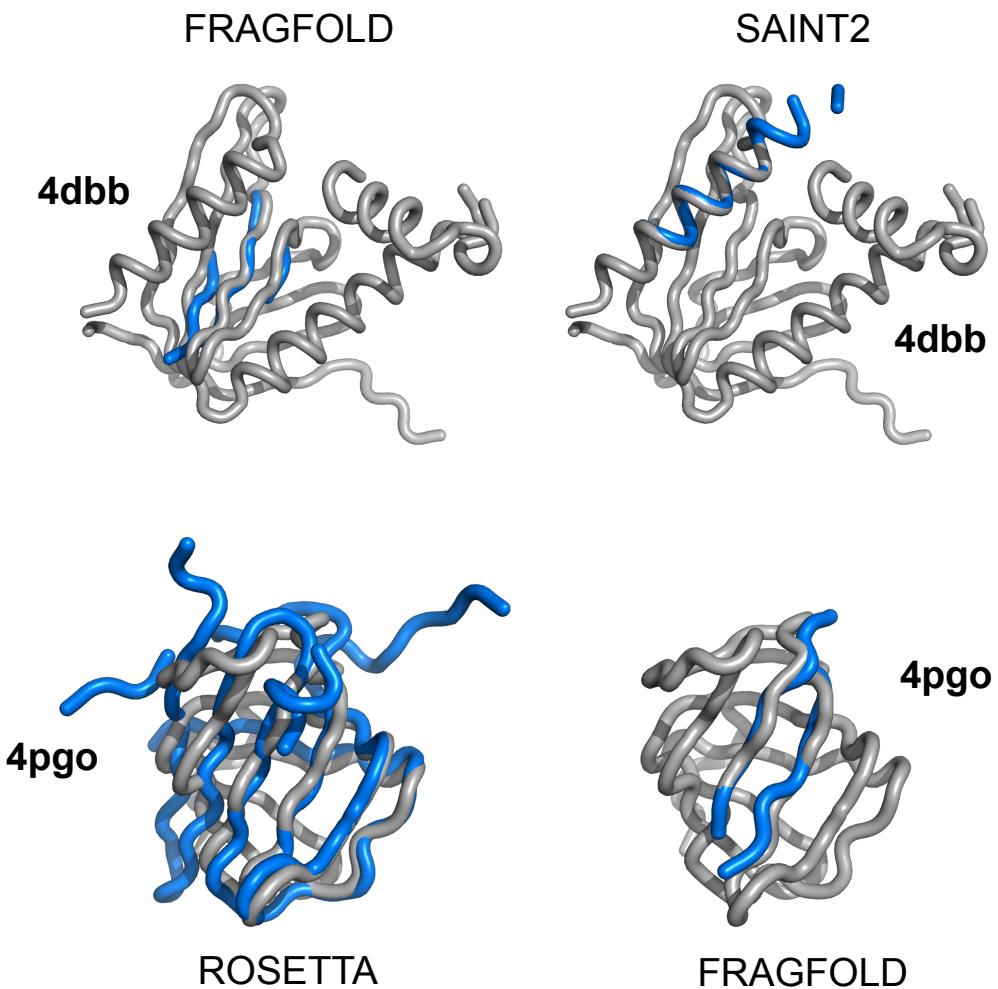


Figure 5.10: Examples of PHASER-placed AMPLEx search models that led to structure solution. AMPLEx search models are coloured blue and deposited structures in gray. The PDB identifiers and modelling protocol given alongside each example.

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 needs 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 better search models. However, an analysis of the RMSD of each ensemble search model's centroid shows 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[157], this threshold does not strictly apply to *ab initio* modelling-based AMPLEx 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),

exceed this threshold greatly with RMSD values $> 10\text{\AA}$ (up to 28\AA) yet resulting in PHASER LLG values in excess of the success threshold of 60 [158]. Additionally, nearly 25% of all successful ensemble search models have RMSD values $\geq 2\text{\AA}$ and PHASER LLG scores of ≥ 60 . Although striking at first, structure solutions in these situations are often achieved by out-of-sequence-register placement of search models. An analysis of the RIO score metric shows that the usefully placed parts of all but one AMPLE search model with RMSD value greater than 10\AA is out-of-register. Furthermore, it is important to remember that RMSD values greatly differ based on the optimal superposition of the model and target.

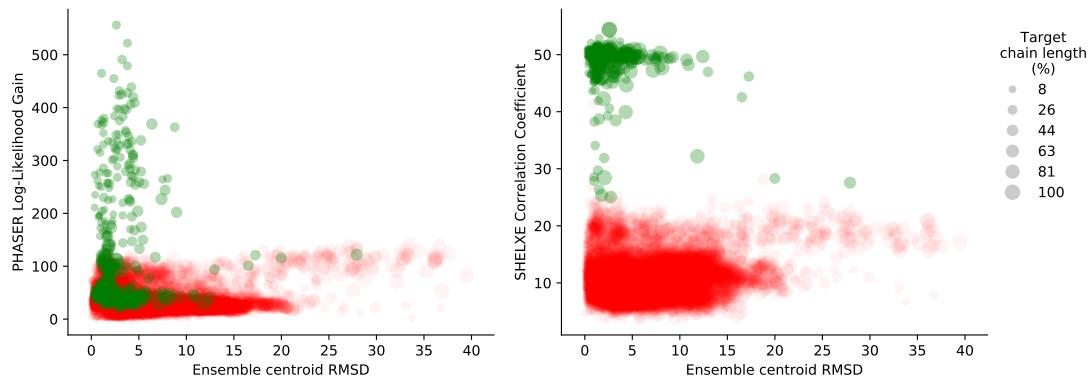


Figure 5.11: Relationship between ensemble quality, PHASER LLG and SHELXE CC. Data points are coloured based on the outcome of their MR trials: green indicate structure solution, red indicate no structure solution.

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 contain 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 is present CONFOLD2 ensemble search models, such that it might explain why good decoy quality does not translate to more MR structure solutions.

Indeed, a Ramachandran analysis of φ and ψ peptide backbone angles outlines 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 are made up of crudely-refined decoys, possess at most 2% of residues as outliers. SAINT2, which might not outcompete other protocols in overall quality, shows the second best stereochemistry of centroid models without any refinement. FRAGFOLD contains around 5% outliers for the majority of search models. In comparison to these statistics CONFOLD2 contains around 5-15% Ramachandran outliers in centroid decoys.

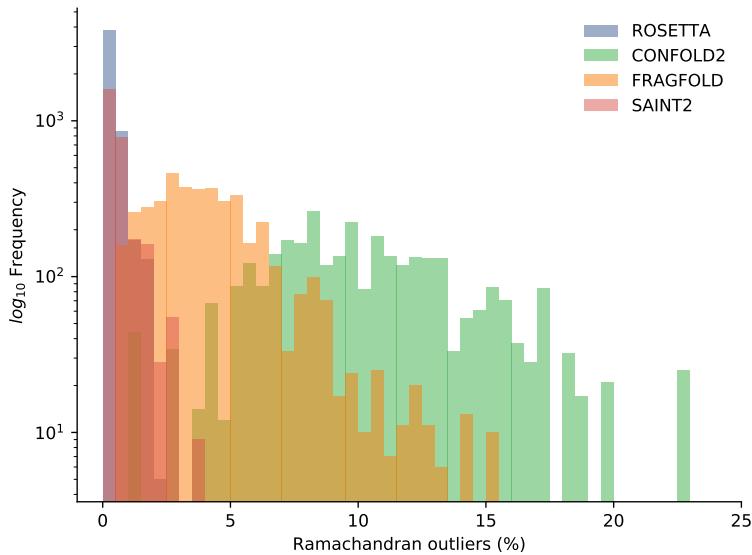


Figure 5.12: Distribution of Ramachandran outliers of AMPLE ensemble search model centroids based on decoys predicted with four *ab initio* structure prediction protocols. Outliers were calculated using PyRAMA (<https://github.com/gerdos/PyRAMA>).

Further analysis of the centroids of each truncated AMPLE ensemble demonstrates the importance of good stereochemistry for success. Out of 94 successful ensemble search models only 17 contain outliers. This contrasts to over 390 unsuccessful ensemble search models with TM-scores greater than 0.5 units and on average 6% outliers ($\text{min}_{\text{outliers}}=1\%$; $\text{max}_{\text{outliers}}=23\%$). However, perfect peptide backbone stereochemistry is still no guarantee for MR structure solution. The centroids of over 200 unsuccessful CONFOLD2 ensemble search models have TM-scores of at least 0.5 units with residue counts ranging from 6 to 106.

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 [147, 148]. Furthermore, the findings describing the comparable performance of ROSETTA and CONFOLD2 [73, 143] are supported in this work. Given that the latter relies entirely on the provided contact information, such performance emphasises the

quality and importance of contact prediction in protein structure modelling. It is also to be expected that the increase in sequence availability will improve the decoy quality further [147, 159].

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. In particular, 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 information — 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 is not necessarily enough to predict the success of AMPLE-generated search models. The results in this chapter clearly show that CONFOLD2 generates high quality decoys, which do not translate into MR search models suitable for routine structure solution. Despite a recent example of the successful application of CNS-generated decoys in MR [160], 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.

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 using contact
information to enhance MR
search model creation**

6.1 Introduction

Work presented in previous chapters highlighted the much improved *ab initio* decoy quality achievable by restraining the conformational search space with residue-residue contact information. Furthermore, the data also highlighted that this improvement extends AMPLE’s performance of achieving structure solution for more challenging targets. However, the data also indicated that AMPLE’s current protocol is not tailored towards decoy sets with overall much higher accuracy. In some cases, decoy sets with correctly predicted folds — whereby the mean TM-score of the decoy set is greater than 0.5 score units — did not generate any successful ensemble search models. It also became apparent that certain decoy sets contained some very high-quality decoys, but that these were often lost in the process of clustering due to the lack of other similar predictions.

Furthermore, *ab initio* decoy similarity to the crystal structure was exceptionally high in some cases ($\text{RMSD} < 1.5\text{\AA}$). 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. Contact information, which is typically used to restrain the folding protocol might provide enough information to drive such identification. Indeed, Kosciolek and Jones [45] and De Oliveira et al. [111] 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 [143] use long-range contact satisfaction routinely in CONFOLD2 to exclude the worst decoys amongst the set predicted ones.

Thus, this chapter focuses on exploring alternative strategies of decoy selection in AMPLE. In particular, work presented here focuses 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 previous chapters. 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 [91], PCONSC2 [99], METAPSICOV STAGE 1 [100] and MEMBRAIN [161] 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; see Section 2.3.2.3) were computed for each decoy in each set. 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:

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

The fixed definition in the *CUTOFF* strategy was determined by De Oliveira et al. [111]. 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 differs 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 [162] 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 [146]

fragment	remove all residues that do not satisfy the following condition: extract all contacts from a decoy [C β distance of < 8Å (C α 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 <i>NONE</i> 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 above 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 [113] was shown to be advantageous in most cases by Thomas [155], 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 [126] 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 focuses 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 [149] 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 shows 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 reveals that all- α , mixed α - β and transmembrane protein targets show the strongest positive correlations for long-range contact satisfaction (Table 6.1). All- β and mixed α - β decoy sets show the strongest correlations for short- and medium-range contact satisfaction, whereby the former shows 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 show 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 is < 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 indicate that only some cases show models with reasonably

good fits to the data (Fig. 6.1). Nevertheless, all models further support 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 is for long-range contact pairs, where the linear regression models are also near identical between the different fold categories (Fig. 6.1).

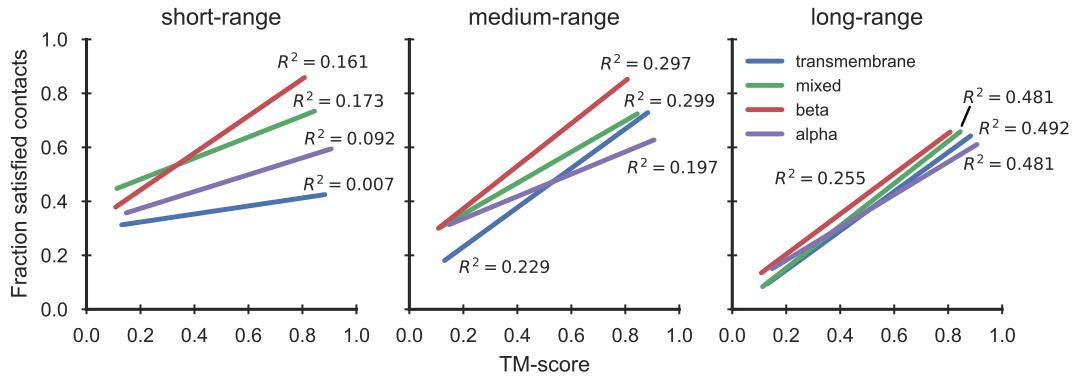


Figure 6.1: Linear regression model fitted to decoy TM-scores and corresponding fractions of satisfied, range-dependent contacts. Targets were further separated by fold classification. Coefficients of determination (R^2 -values) added alongside each regression model.

An analysis of the correlation between the TM-score and long-range contact satisfaction of individual decoy sets further highlights the potential to subselect decoy sets by their long-range contact satisfaction. One hundred and eight decoy sets show 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 with PDB ID 2eum and restrained with METAPSICOV STAGE 1 contact data, shows 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 METAPSICOV STAGE 1 contact data except for 2x6u (PCONSC2), show no statistically significant correlation between the TM-score and long-range contact satisfaction of the decoy sets.

A further subdivision of the previously presented data by metapredictor highlights that no predictor outperforms the others. Decoy sets calculated using predictions from all metapredictors exhibit a range of stronger to weaker correlations. Similarly, target chain length and fold do 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 we could better estimate the accuracy of the best decoy by some measure. Kosciolak 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 are 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 cannot be validated.

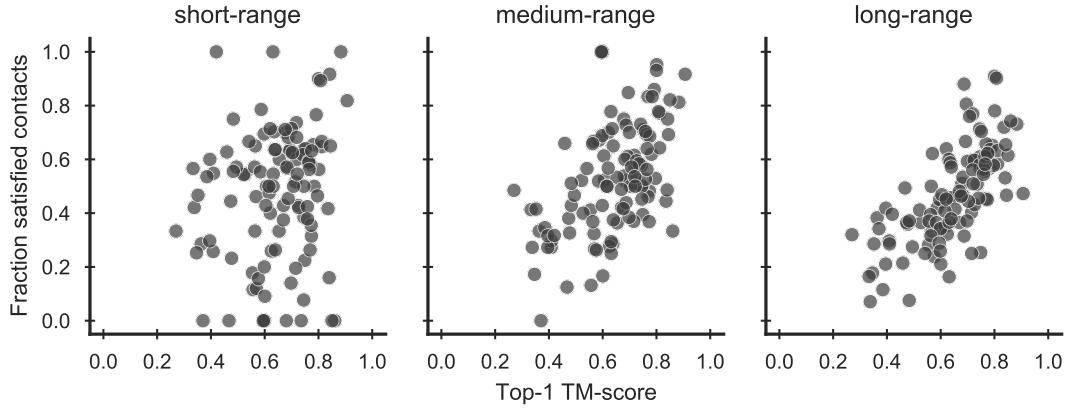


Figure 6.2: Analysis of the relationship between TM-score and contact satisfaction for the top-1 decoy (as ranked by TM-score) in each decoy set.

6.3.2 Long-range contact satisfaction metric to filter decoy sets

In the previous section, 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 is to determine if this correlation could be used to alter the starting decoy set prior to the submission to the AMPLE cluster-and-truncate pipeline to enhance the chances of generating more successful ensemble search models for MR.

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 results 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 are better values to use to improve the mean TM-scores of each decoy set. Notably, per-decoy long-range contact satisfaction provides the best estimate for identifying and excluding the least accurate decoys independent of the subselection strategy.

Given the improvement of TM-scores for each decoy set by decoy subselection, it is 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 requirements after

clustering and sub-clustering. For the decoy sets used in this study, the *LINEAR* strategy removes on average the most decoys from each set with a fixed number of 500 (median=500). In comparison, the *CUTOFF* subselection strategy removes 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 vary drastically similarly to the changes in TM-score, it becomes apparent that the more decoys are removed, the better the overall score becomes, which further supports 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 STAGE 1 decoy set of the ankyrin sequence (PDB ID: 2qyj) shows 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) shows 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 appears to be long-range contact satisfaction driving the *LINEAR* strategy.

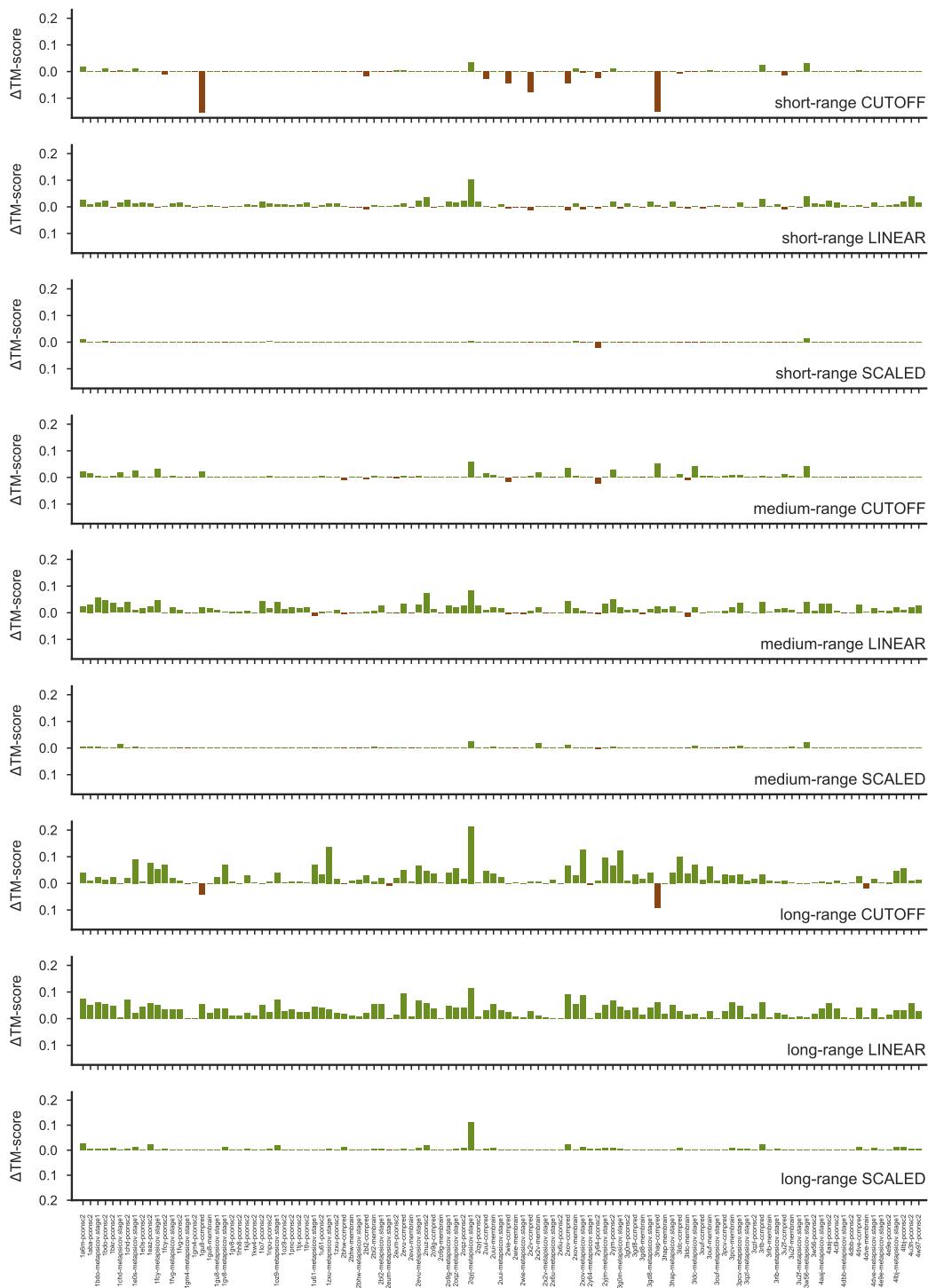


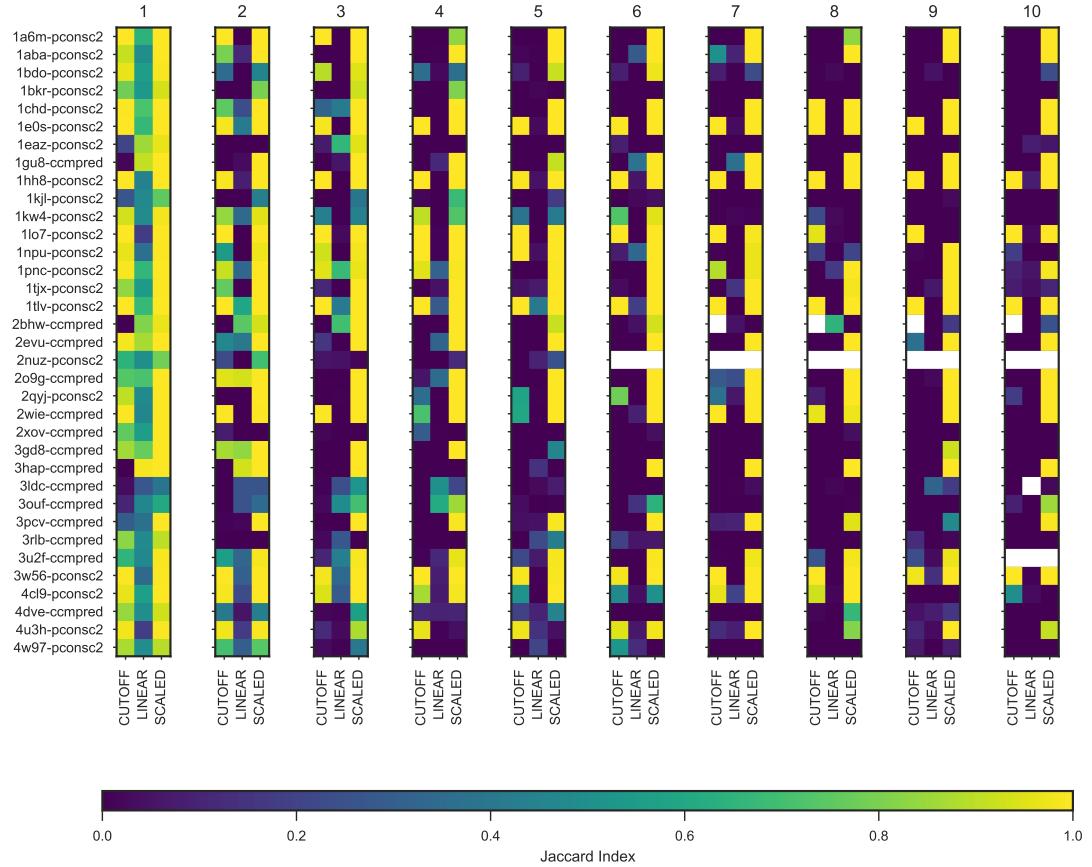
Figure 6.3: Differences in mean TM-score for decoy sets pre- and post-decoy subselection. Each subselection strategy is stated in each subplot along with the contact range used to establish decoy inclusion in the final set.

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 highlights an important observation. Larger clusters — those ranked higher — show higher similarity between a subselection strategy and the default (Fig. 6.4a). The top SPICKER cluster shows high similarities between the *NONE* strategy and all other subselection ones, whereby it has to be noted that the *LINEAR* strategy contains only 50% of the starting decoys and thus can at most show a Jaccard index of 0.5. With increasing cluster index, the overall similarity degrades and most of the decoys in cluster 10 are 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 compared 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 starts to alternate without any particular pattern. Thus, pre-selecting decoys prior to AMPLE’s cluster-and-truncate approach most certainly preserves the top cluster for the *CUTOFF* and *SCALED* subselection strategies, whereby lower clusters show more deviation from the default.



(a)

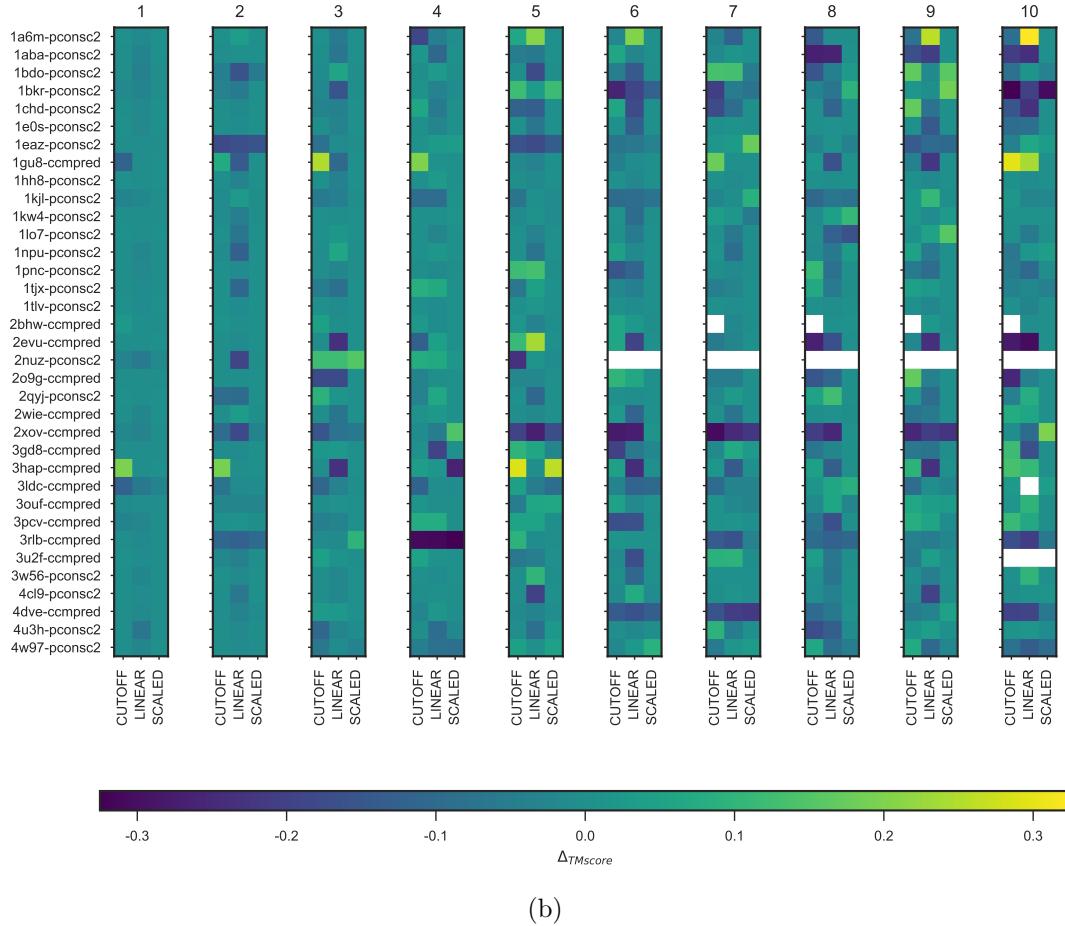


Figure 6.4: Effect of decoy subselection on SPICKER clusters. Effect illustrated by (a) the Jaccard Index and (b) median TM-score difference. Values were calculated for clusters resulting from the full starting set of decoys and the *CUTOFF*, *LINEAR* and *SCALED* subselection strategies. Larger TM-score 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 -15\AA^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 -20\AA^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.

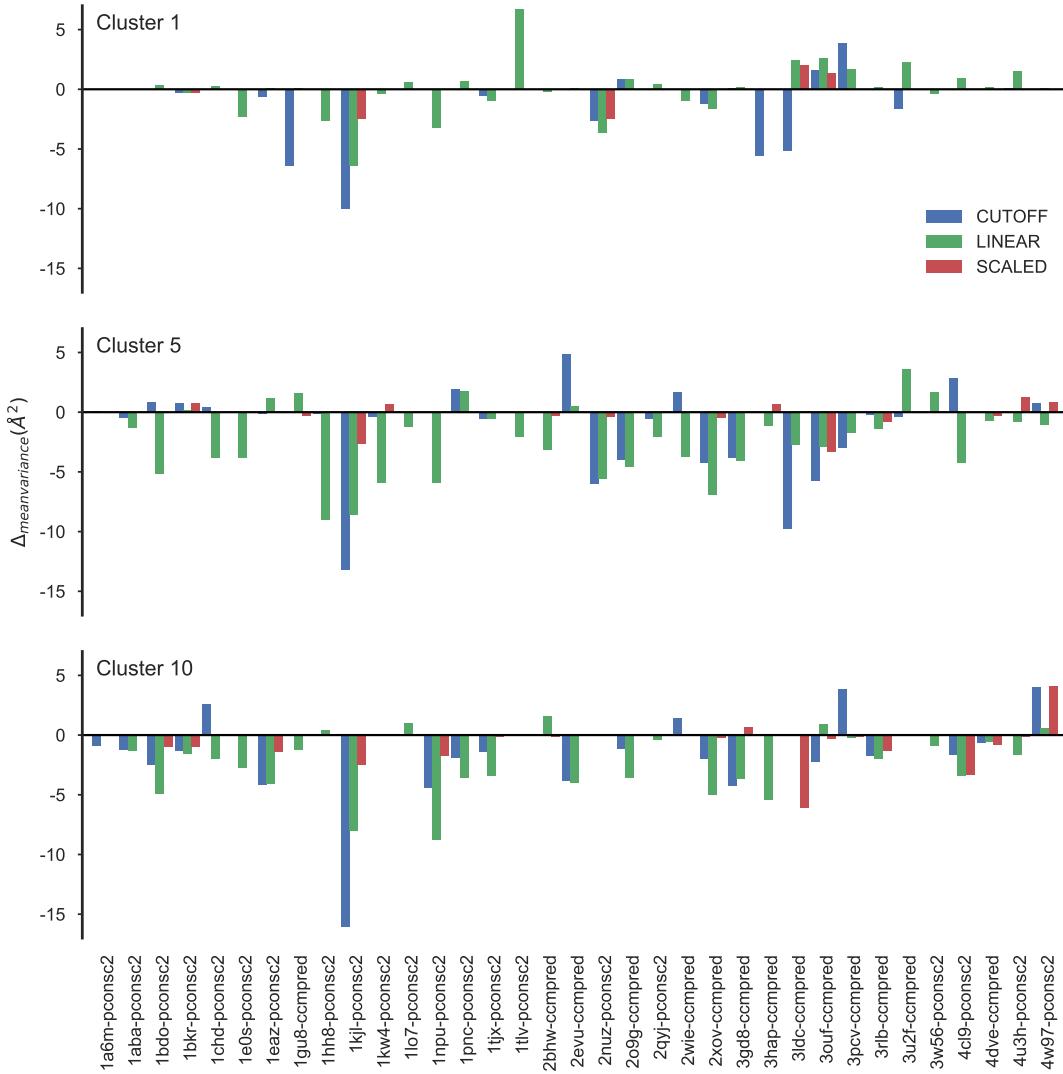


Figure 6.5: Effect of decoy subselection on mean inter-decoy THESEUS variance. Difference in mean variance calculated between the default and the three decoy subselection strategies *CUTOFF*, *LINEAR* and *SCALED*. Data for clusters 1, 5 and 10 shown as examples.

A comparisons of intermediate stages in the AMPLE pipeline resulting from differently subselected decoy sets is very difficult. Each strategy results in different starting sets, which result in different clusters. Since AMPLE’s objective truncation procedure is based on the inter-decoy variance, it might be greatly affected by differing clusters. Nevertheless, structure solution is more likely when AMPLE generates 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 reveals that the *SCALED* strategy generates the most search models ($n = 7,611$), which is roughly 300 more than the default *NONE* strategy ($n = 7,340$). The *CUTOFF* subselection strategy generates the least ensemble search models ($n = 7,237$), whilst the *LINEAR* strategy’s count ($n = 7,401$) is very similar to the *NONE* one.

Further inspection of the number of AMPLE-generated ensemble search models by target reveal near identical numbers between the *NONE*, *LINEAR* and *SCALED* strategies (Fig. 6.9). In fact, only a few outliers for each of those methods distinguish them from the others. The *CUTOFF* strategy shows greater deviation from the other three, especially for certain targets with differences up to approximately 200 ensemble search models (Fig. 6.9). If we compare all these strategies to the previous default processing in AMPLE (*NONE_classic*; further details in Section 6.2.4), we can see that the number of search models is greatly reduced (Fig. 6.9). A comparison of the previous default (*NONE_classic*) with the new one (*NONE*) shows on average 144 fewer ensemble search models per target, whilst sampling a larger range of folds through all 10 clusters.

6.3.4 MR search models by processing single decoys

In addition to the decoy set subselection, this study also attempts to identify single decoys of sufficient quality to be used directly as MR search models. Although ensembles are generally more desirable MR search models [116, 119, 126], 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.

The correlation between a decoy's long-range contact satisfaction and its TM-score has previously been outlined and is 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. Here, an analysis of the top-5 decoys by long-range contact satisfaction in each decoy set shows that 50% of selected decoys fall in the 80th percentile or greater of TM-scores in each decoy set whilst 90% are in the at least the 40th percentile (Fig. 6.6).

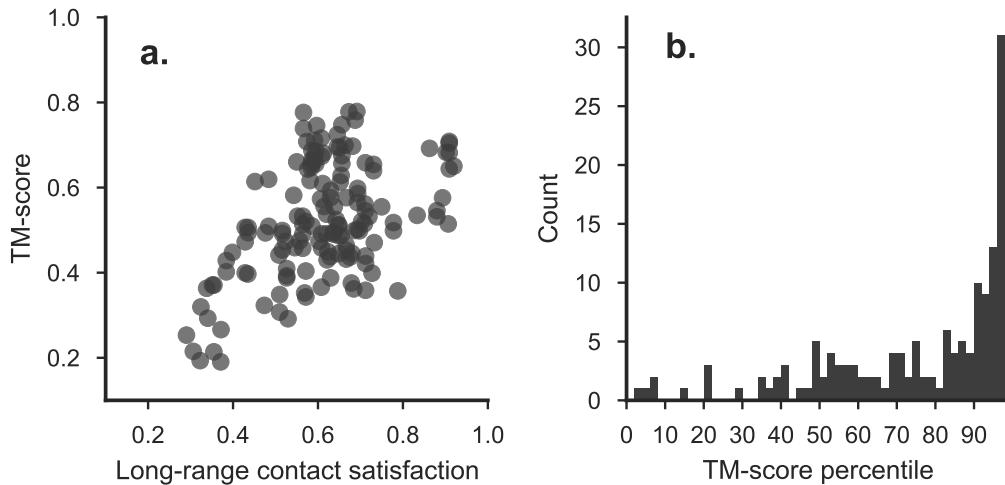


Figure 6.6: Analysis of long-range contact-satisfaction-based decoys with respect to the relationship between (a) the satisfaction and decoy quality and (b) the quality compared to the remaining, excluded decoy set.

A comparison of RMSD value changes indicated that the “fragment” and “variance” metrics provide the best approximation to identifying less-reliable regions in each decoy. The average RMSD change compared to the original decoys is just under 4.0\AA . This compares to a slightly lower RMSD change of 2.1\AA for “DSSP”-treated decoys and 1.5\AA for the “domain” treatment. Although almost all decoys were improved by either of the treatments, a small number of decoys worsen 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 -1.6\AA .

A comparison of the range of RMSD values revealed much greater changes for the “fragment” and “variance” conditions (Fig. 6.7). However, these changes are not reflected in the fraction of residues retained in each decoy. Most residues were removed by the “domain” treatment ($\mu=61.6\%$), whilst the “fragment” one saw the least removal ($\mu=34.5\%$). Similarly to the range 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%.

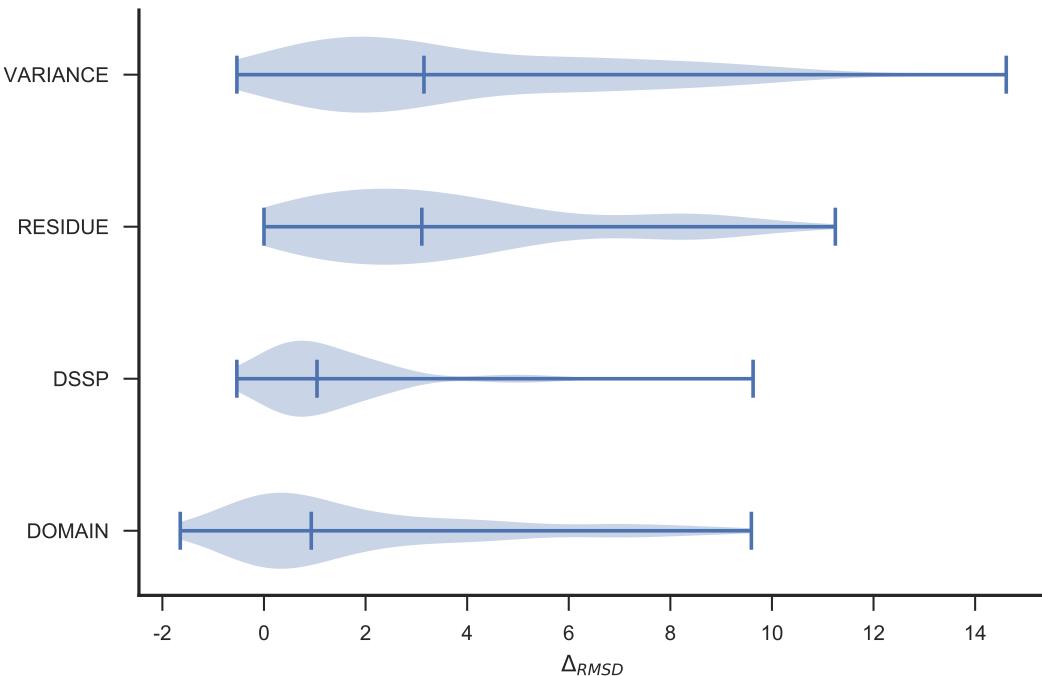


Figure 6.7: Distribution of differences in RMSD values between the initial and modified decoys under four different treatments. A positive Δ_{RMSD} value corresponds to a decrease in RMSD compared to the crystal structure.

A further aspect of the decoy treatments highlights that the fraction of residues retained after decoy post-processing correlates with the cluster variance of the decoy, extracted from THESEUS results of each decoy’s cluster in the *NONE* strategy (Fig. 6.8). Unlike the variance metric, all other processing metrics do not show a correlation with the fraction of residues retained. This explains at least in part why we see much greater changes in RMSD value between the initial and processed decoy for the “variance” treatment compared to the others. However, if a decoy is of particularly poor quality (TM-score < 0.3), the “variance” treatment retains as little as 0.87% and 1.7% of the initial decoy (2 and 4 residues) whilst the others retain a much larger fraction of at least 40% for equivalent decoys.

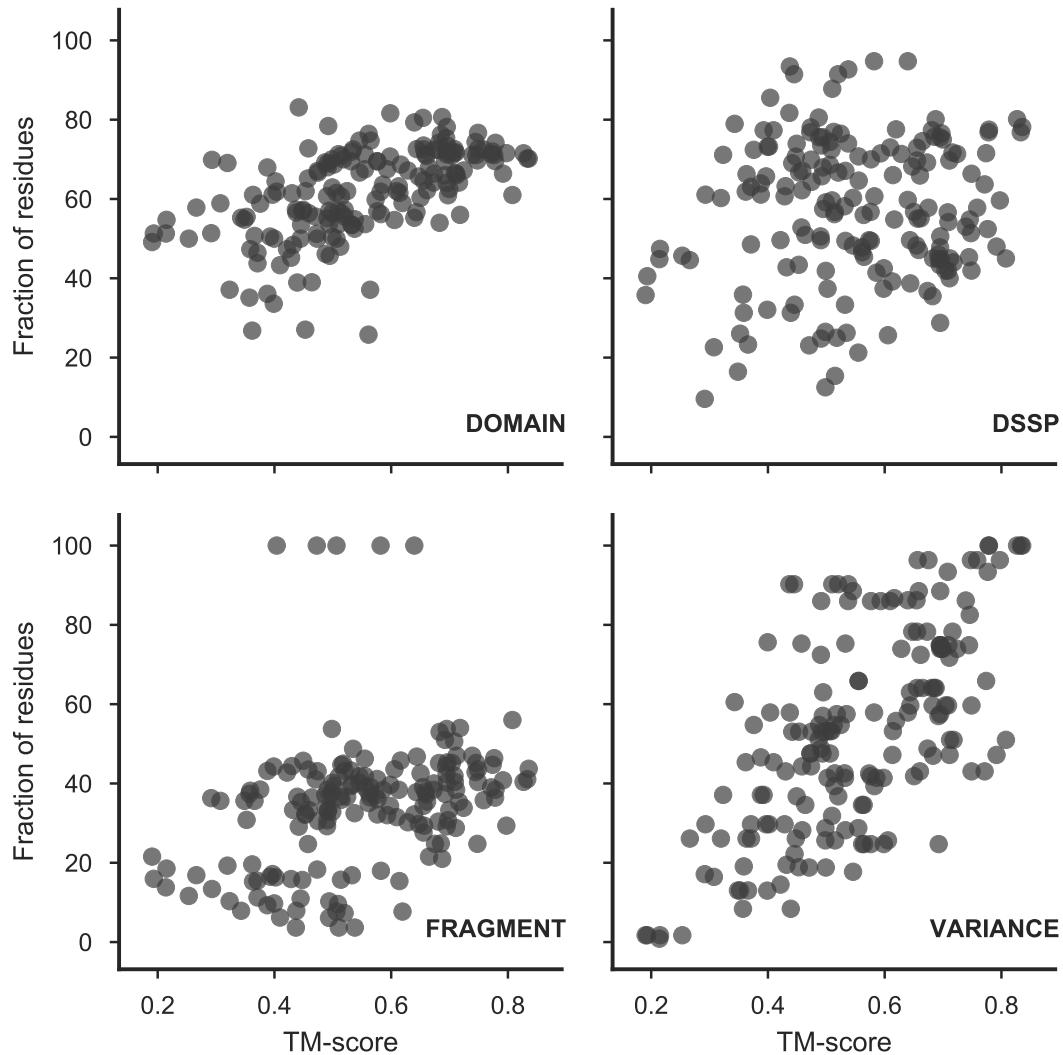


Figure 6.8: Scatter plots of initial decoy TM-score and the fraction of residues retained after one of four different residue removal treatments.

6.3.5 Decoy subselection extends AMPLE's performance

The final step in this study is the assessment of AMPLE-generate 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 is of great interest since it might allow us to extend AMPLE's performance beyond that described in previous chapters.

A comparison of the total number of targets solved by each subselection strategy shows that the *CUTOFF*-subselected decoys lead 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 are on par with

AMPLE’s default, the *NONE_classic* strategy (Fig. 6.9). Although the *NONE_classic* strategy generates two version of each ensemble search model with poly-Alanine and all-atom side chain treatment, 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.

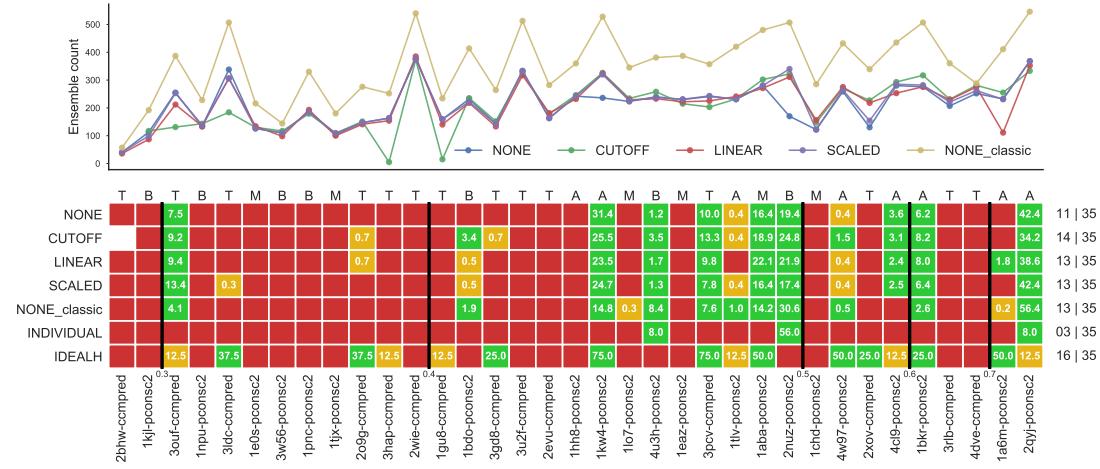


Figure 6.9: Molecular Replacement summary of decoy-subselected AMPLE ensembles. AMPLE-generated ensemble counts illustrated at the top with Molecular Replacement results in grid below: red cell equates to no solution; orange to a singleton solution; and green to multiple solutions. All *INDIVIDUAL* attempts 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 yields the highest number of structure solutions based on AMPLE-generated ensemble search models whilst generating the fewest search models. In fact, this subselection strategy has generated no ensemble search models for target 2bhw. Furthermore, the *CUTOFF* method achieves amongst the best ratio of search models leading to structure solution compared to the total number generated.

In a 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 contain at least 50% of the target sequence, whilst none contain 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 KstrR2 (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 generate 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 with 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 the same ensemble, 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) are different between the successful ensembles and the *NONE* strategy. However, only a single decoy is unique to either *CUTOFF* and *LINEAR* in a direct comparison. Ultimately, this results 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 shows crucial value in preparing decoy datasets prior to AMPLE’s cluster-and-truncate approach.

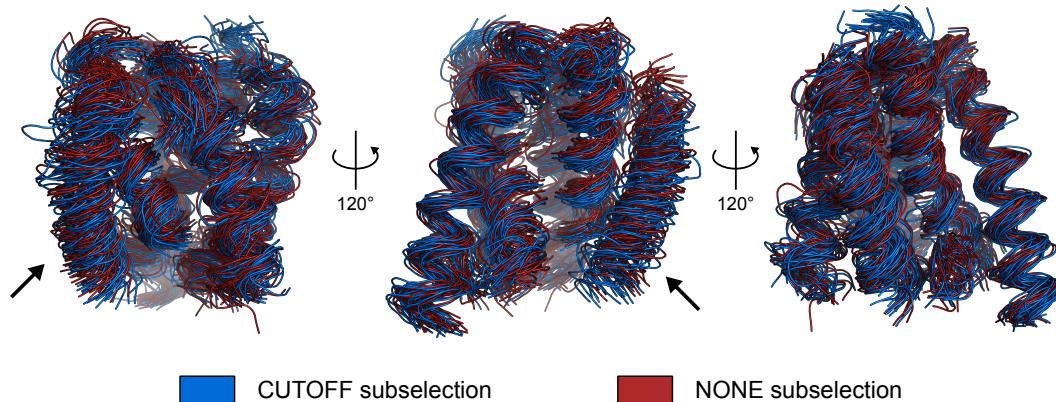


Figure 6.10: Example of the structural divergence of two ensemble search models with an identical AMPLE cluster-and-truncate path. Ensembles are based on the CCMPRED decoy set of PDB ID 2o9g and derived from cluster 1 with 59% truncation and subclustering radius of 3Å. The blue ensemble was derived from *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 shows that no decoy set with median TM-score of less than 0.3 score units led to structure solution; however, only two such cases exist 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 are 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 lead 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).

A comparison of decoy-derived search models and AMPLE’s simplistic ideal helix library [116] 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 is also increased by two compared to the work by Thomas et al. [117], which is 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. [117] suggested that decoy-derived search models are essential since ideal helices provide insufficient scattering matter with low resolution ($> 2\text{\AA}$) intensity data. In this study, these findings cannot be validated given that PDB ID 1gu8 (resolution of 2.27\AA) 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 Kosciolak and Jones [45] and later confirmed and extended by De Oliveira et al. [111] and Adhikari and Cheng [143]. In this study, these findings are further confirmed by re-analysing all decoy sets previously presented in this thesis.

Furthermore, the benefit of subselecting decoys based on their long-range contact satisfaction pre-AMPLE was evaluated. Subselection extends the target tractability of AMPLE whilst reducing the number of generated search models, which effectively

enhances AMPLE’s performance. The *CUTOFF* subselection strategy has proven to be most successful in flagging the worst decoys, which results in more accurate ensemble search models being generated. The data presented shows that subtle differences in clustering can 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 has also become apparent that decoy-based ensemble search models are inferior to AMPLE’s simple ideal helix library, particularly for transmembrane protein targets. The latter is sufficient to solve the majority of transmembrane protein targets, which outperforms all decoy-based approaches combined. This result contradicts the one reported by Thomas et al. [117], 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 [163] or FRAGON [156], will 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 [164]. 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 [116, 119, 126], 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, 111, 143]. 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 approach 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 supersecondary structure motifs [165]. Furthermore, fragments similar to those typically selected for *ab initio* structure prediction were successfully used in other areas of structural biology including NMR [166, 167] and X-ray crystallography [168] 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 co-evolution 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]. Various different algorithms have been developed to generate static and dynamic fragment libraries. Static fragment libraries are those pre-computed and generally consist of common supersecondary 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], 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 true positives 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 co-evolution data or other externally obtainable restraints to validate fragments prior to any MR attempt. As such, the work in this chapter focuses on exploring this idea using FLIB [53], a dynamic fragment picking algorithm considering co-evolution data to verify fragments

during the picking process.

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 are: 1aba, 1lo7, 1u06, and 5nfc. The former two are described in Table A.1. Target 1u06 is a 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 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 < 150 residues. Each fold class, mixed α - β and all- β , contained two targets.

7.2.2 Fragment picking using FLIB

FLIB [53] requires four inputs: the predicted secondary structure, predicted torsion angles, residue-residue contact pair data and a copy of the PDB. The secondary structure for each target was predicted using PSIPRED v4.0 [152] with default parameters. The torsion angles were predicted using SPIDER2 v[169] with default parameters, and residue-residue contact pairs using METAPSICOV v1.04 [100] with default parameters. HHBLITS v2.0.16 [135] with uniprot20 database v2016-02 was used by METAPSICOV to generate the MSA for contact prediction of each target sequence. BLASTP v2.2.31+ [170, 171] 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 v1.01 (<https://github.com/sauloho/FLIB-Coevo>, commit abade3b) protocol. The first focuses on exclusion of fragments with > 90% helical content (assigned by DSSP [146]). If fragments with > 90% helical content are allowed and residues are predicted to be part of an α -helix, fragment libraries tend to be overpopulated for these positions with short helices. This would generate fragment libraries similar to ideal helix libraries, which is 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

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 modification was lifted.

Two-hundred fragments were picked per target sequence position. Top- L or $L/2$ contact pairs were selected from both METAPSICOV STAGE 1 and STAGE 2 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 to be solved. BLASTP v2.2.31+ [170, 171] and HHPRED (HHBLITS v2.0.16 and HH-SEARCH v2.0.16) [172] searches were conducted to identify homologous PDB entries. The BLASTP search was performed identically to Oliveira et al. [53] against the pdbaa v2016-10 database using an E-value cutoff of 0.05. The HHPRED search parameters were identical to the MPI-Toolkit [173] webserver 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 ≥ 20.0) were excluded from the fragment libraries.

All per-target fragments were then binned by their peptide lengths. Subsequently, they were ranked by FLIB 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 score refers in this chapter to the predicted torsion angle score for a given fragment, which FLIB uses in its default routine to rank fragments with lower scores being more favourable [53].

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 [126]. This uses PHASER [127] for MR, REFMAC5 [27] for refinement and SHELXE [129] for density modification and main-chain tracing. MR-BUMP default parameters were used with exception of the PHASER RMSD estimate.

¹The `-coevo_only` flag was intended to select only fragments that satisfied at least one contact pair. This 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 source code (commit "b3eb01d").

Each fragment was subjected to MRBUMP using PHASER RMSD values of 0.1, 0.6 and 1.0Å.

7.2.4 Assessment of FLIB fragments

Fragment torsion angles — predicted by SPIDER2 [169] — were assessed using the Mean Absolute Error (MAE), which evaluates the average absolute difference between the predicted and experimentally determined angles [169]. 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 $< 1.5\text{\AA}$ were considered TP else FP. The equation used to calculate the precision score is Eq. 2.4. The RMSD value, as calculated by FLIB [53], 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 indexes according to sequence alignment provided by FLIB. 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 protein structures in the PDB could be reliably selected and trialed in MR to achieve structure solutions. The fragment picking algorithm FLIB [53] 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 input data

The FLIB 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 co-evolution data. The first part of the analysis in this study focuses on these data given that the FLIB fragment picking heavily relies on the individual features in the selection and scoring of each individual fragment [53]. 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.

The secondary structure prediction highlighted high precision between each tar-

get's prediction and the DSSP-assigned [146] secondary structure of the target reference structure (Fig. 7.1). The three targets with PDB identifiers 1aba, 1lo7 and 1u06 have secondary structure predictions with a precision of > 89%. The fourth target, 5nfc, shows 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 are correctly predicted, suggesting successful fragment picking is possible.

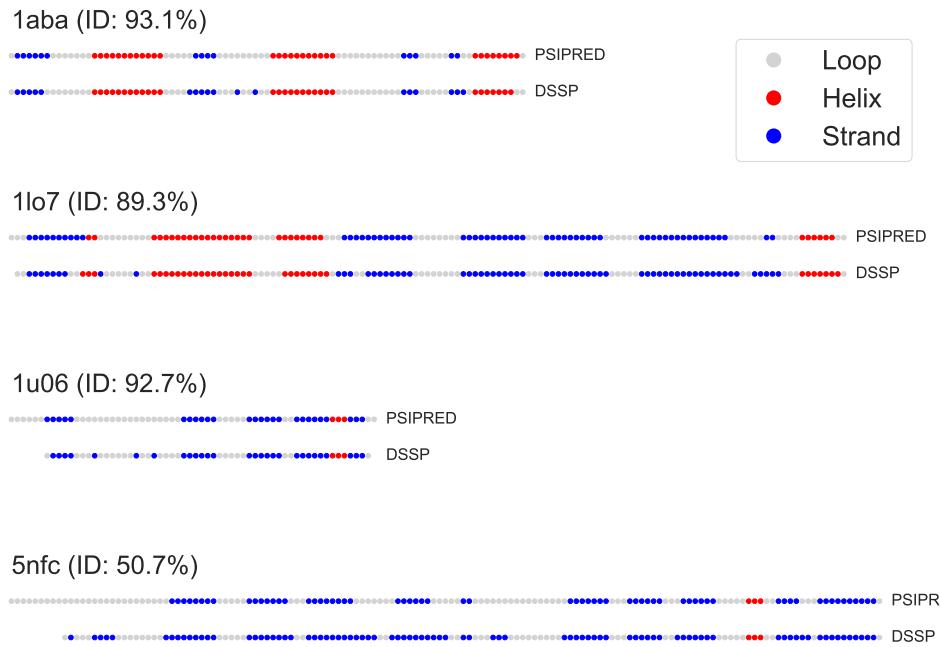


Figure 7.1: Schematic comparison of PSIPRED [152] secondary structure prediction and DSSP [146] assignment. Percentage identity is provided next to each identifier. The identity was computing using the Hamming distance over all positions present in the target sequence and reference structure.

The contact prediction data for METAPSICOV STAGE 1 and STAGE 2 predictions demonstrate the high precision scores 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 algorithm. All targets have precision scores for both sets of predictions at both cutoff levels of > 0.6 (Table 7.1). A comparison of the sets of contact pairs shows that only every third (for $L/2$ contacts) or every other (for L contacts) contact pair is shared between both METAPSICOV STAGE predictions highlighting the importance of trialling both when selecting FLIB fragments (Jaccard index in Table 7.1).

Table 7.1: Precision scores for METAPSICOV [100] STAGE 1 and STAGE 2 contact predictions. Jaccard index calculated for the same L -dependent selection of contact pairs between METAPSICOV STAGE 1 and STAGE 2 predictions.

Target	$L/2$ contact pairs			L contact pairs		
	Prec _{STAGE 1}	Prec _{STAGE 2}	Jaccard	Prec _{STAGE 1}	Prec _{STAGE 2}	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 show localised clusters of contact pairs characteristic for secondary structure features (Fig. 7.2). These clusters are more populated with contact pairs in METAPSICOV STAGE 2 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 [100]. Besides the visual analysis, a cluster determination study on each of those contact maps further confirmed a higher singleton frequency in METAPSICOV STAGE 1 predictions. The latter contain on average 9% more singleton contact pairs, and thus a higher degree of noise.

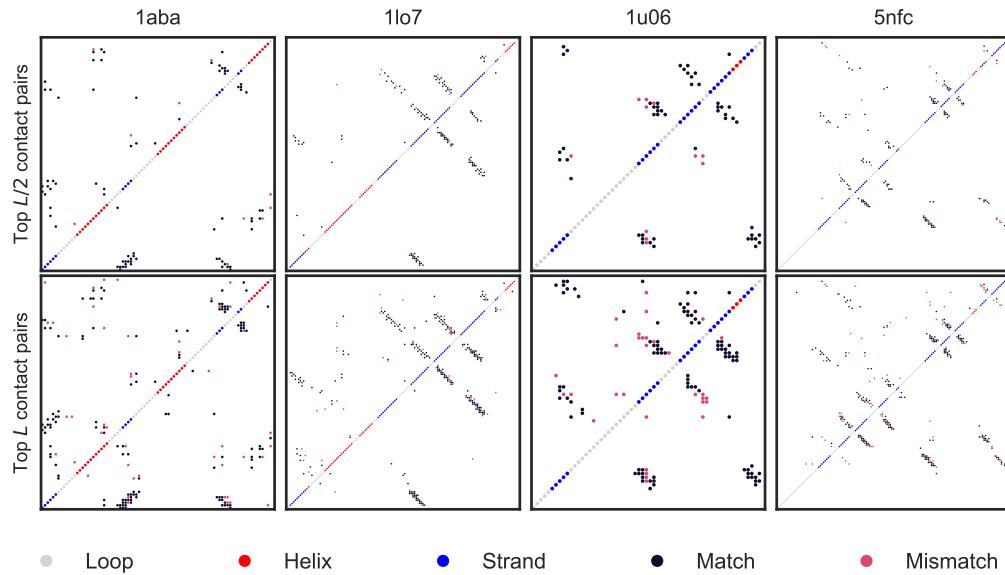


Figure 7.2: Comparison of $L/2$ and L correctly and incorrectly predicted contact pairs for four FLIB targets. Contacts were predicted using METAPSICOV [100] STAGE 1 (top left) and STAGE 2 (bottom right). True and false positive contact pairs were identified using a 8Å cutoff between Ca ($\text{C}\beta$ in case of GLY) atoms of a reference crystal structure. PSIPRED [152] secondary structure prediction provided along the diagonal.

An analysis of the MAE of torsion angles between the SPIDER2 [169] prediction and a corresponding reference crystal structure highlights accurate predictions for three of

four targets (Fig. 7.3). The largest MAE_ϕ across the four target sequences is 24.347° , and the largest MAE_ψ is 45.459° (MAE values for PDB entry 1u06). The smallest MAE_ϕ is 13.822° (PDB ID: 1aba) and smallest MAE_ψ is 17.273° (PDB ID: 1lo7). Segments in sequence space with regular secondary structure, as predicted by PSIPRED [152], result primarily in low MAE values of torsion angles. In contrast, unstructured regions highlight much larger MAE values indicating the difficulty of predicting these regions. Noticeably, the MAE_ψ appears to be much larger in those regions than the MAE_ϕ for the same residue.

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

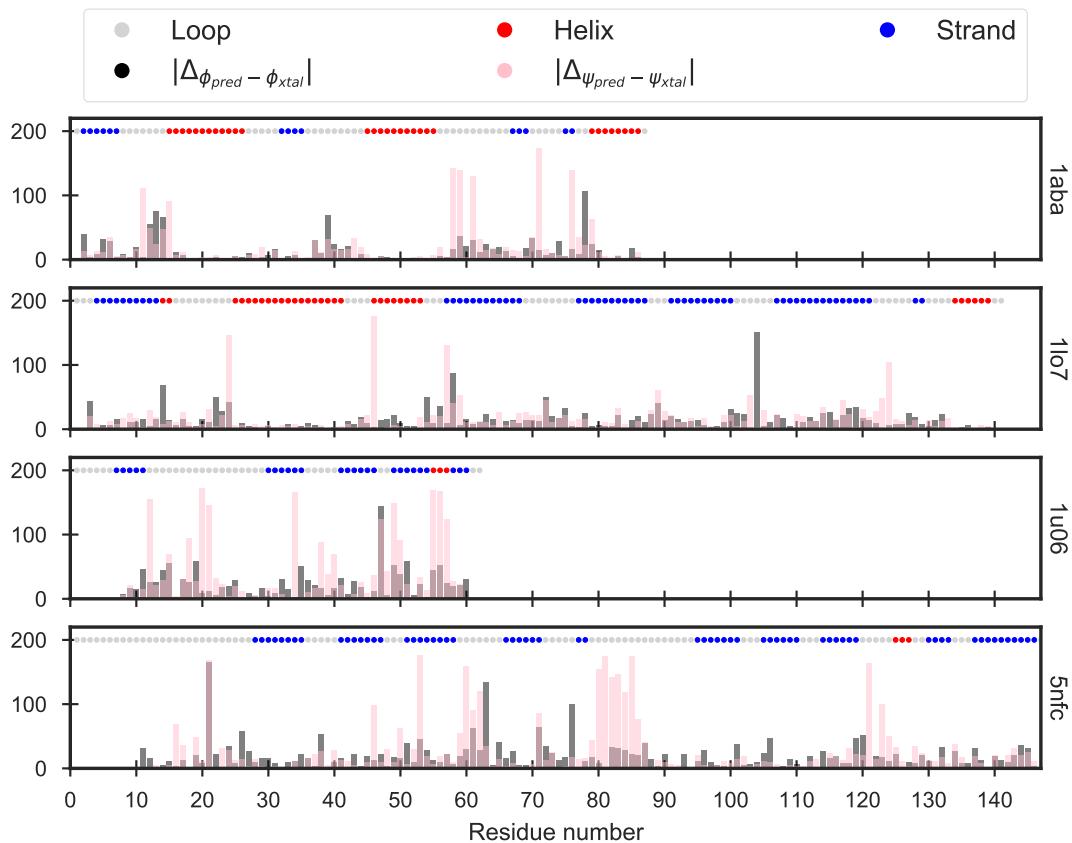


Figure 7.3: Comparison of MAE of torsion angles predicted by SPIDER2 and extracted from a corresponding PDB structure. PSIPRED [152] secondary structure prediction provided alongside the MAE values.

7.3.2 FLIB fragment picking

Sixteen FLIB fragment libraries were picked for each protein target in this study. Each fragment library consisted of one permutation of one of two contact prediction files and altering input parameters.

Across all four targets, the FLIB algorithm selected a total of 8,535,458 fragments

(Table 7.2). The fragment libraries show similar statistics across the four protein targets despite the diversity in fold and chain lengths. The mean FLIB score is 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 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 score of 3,217 units (Table 7.2).

Table 7.2: Summary of fragment statistics for FLIB libraries selected for four protein targets. Count_H corresponds to the count of fragments extracted from homologs.

Target	Count	Count_H	FLIB 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

A split of the per-target fragment libraries by input options highlights the better fragment library quality under certain conditions with regards to the mean FLIB score and RMSD (Fig. 7.4). In particular, top- L (6 residues sequence separation) METAPSICOV STAGE 1 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 suggests much lower and thus more favourable scores for using short-, medium- and long-range contact pairs. A very similar difference is noticeable for METAPSICOV STAGE 2 contact predictions (Fig. 7.4).

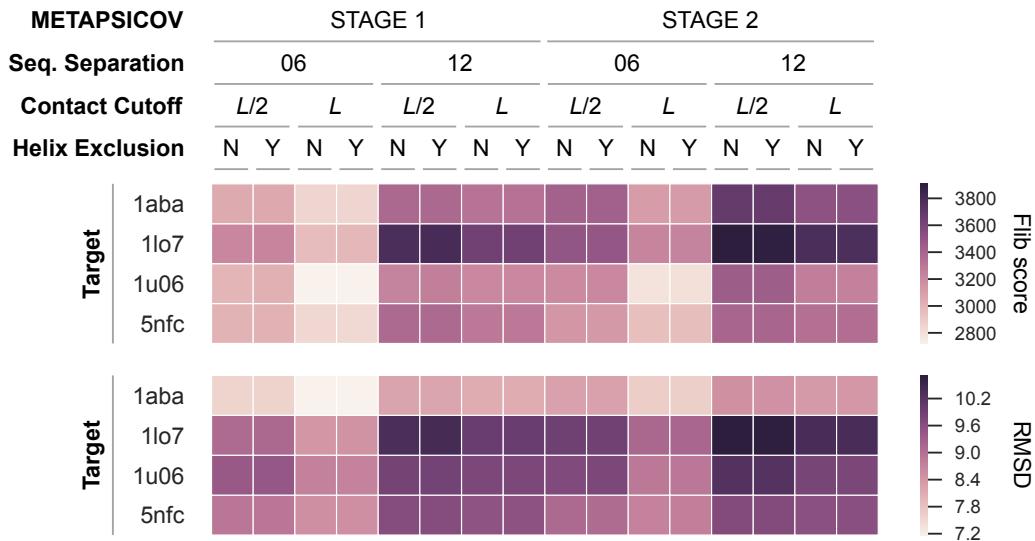


Figure 7.4: FLIB fragment library comparison for four targets highlighting the differences in mean FLIB score and RMSD by starting with different subsets of contact predictions. L refers to the number of residues per target sequence. Y refers to idealised α -helical fragment exclusion during fragment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

In this study, predicted contact information was used to further guide fragment selection. The FLIB 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 demonstrates the benefits of starting with METAPSICOV STAGE 1, i.e. noisier contact predictions (Fig. 7.5). Coverage is 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 STAGE 2 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 supports the benefits of starting with top- L (6 residues sequence separation) METAPSICOV STAGE 1 contact pairs. Across all four targets, the coverage of correct fragments (classed by RMSD $< 1.5\text{\AA}$ to the reference structure) is 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 does not affect the quality of the FLIB libraries greatly. A consideration of differences in mean FLIB and RMSD scores shows Δ differences of 25.68 and 0.06 between the comparable libraries, i.e. with and without idealised α -helical fragments.



Figure 7.5: Summary of the coverage and precision of FLIB 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 $< 1.5\text{\AA}$ 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 (gray). Contact prediction information is illustrated using black bars. The fragment frequency is shown using a log-scale.

Given that FLIB uses co-evolution data to help select fragments, it is little surprise

that higher degrees of TP fragments co-localise with high-density contact pair regions along the target sequence (Fig. 7.5). This characteristic explains less TP fragments in top- $L/2$ fragment libraries because less contacts (compared to top- L) are available during fragment selection. The resulting selection is purely based on the FLIB score which might not yield high-accuracy fragments ($\text{RMSD} < 1\text{\AA}$) as frequently. Therefore, the co-localisation of TP FLIB fragments and regions of high-density contact predictions highlights the importance of adding this additional source of information to pick fragments.

7.3.3 FLIB fragment selection for Molecular Replacement

One of the most important aspects of bypassing *ab initio* structure prediction and using the relevant fragments directly as MR search models is the selection of the fragments with the highest similarity between fragment and target structure.

A fragment’s FLIB 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 protocol [53]. To validate this finding, all non-homologous fragments in this study were tested for a correlation between their FLIB scores and RMSD values. The Spearman’s rank-order correlation coefficient analysis confirms the correlation between a fragment’s FLIB and RMSD scores (Fig. 7.6). However, the strength of the correlation varies greatly between different fragment libraries and targets. The optimal fragment picking strategy — top- L (6 residues sequence separation) METAPSICOV STAGE 1 — results in the strongest correlations across all targets. The same contact pair selection with METAPSICOV STAGE 2 predictions results in the second greatest correlations. Noticeably, the bacteriophage T4 glutaredoxin (PDB ID: 1aba) fragment libraries show much more positive correlations than the remaining targets. The fragments selected for α -spectrin SH3 domain (PDB ID: 1u06) show the overall weakest correlations. It is worth noting that the both targets, PDB IDs 1aba and 1lo7, are classed as mixed α - β targets, and thus the strength of this correlation might be fold dependent.

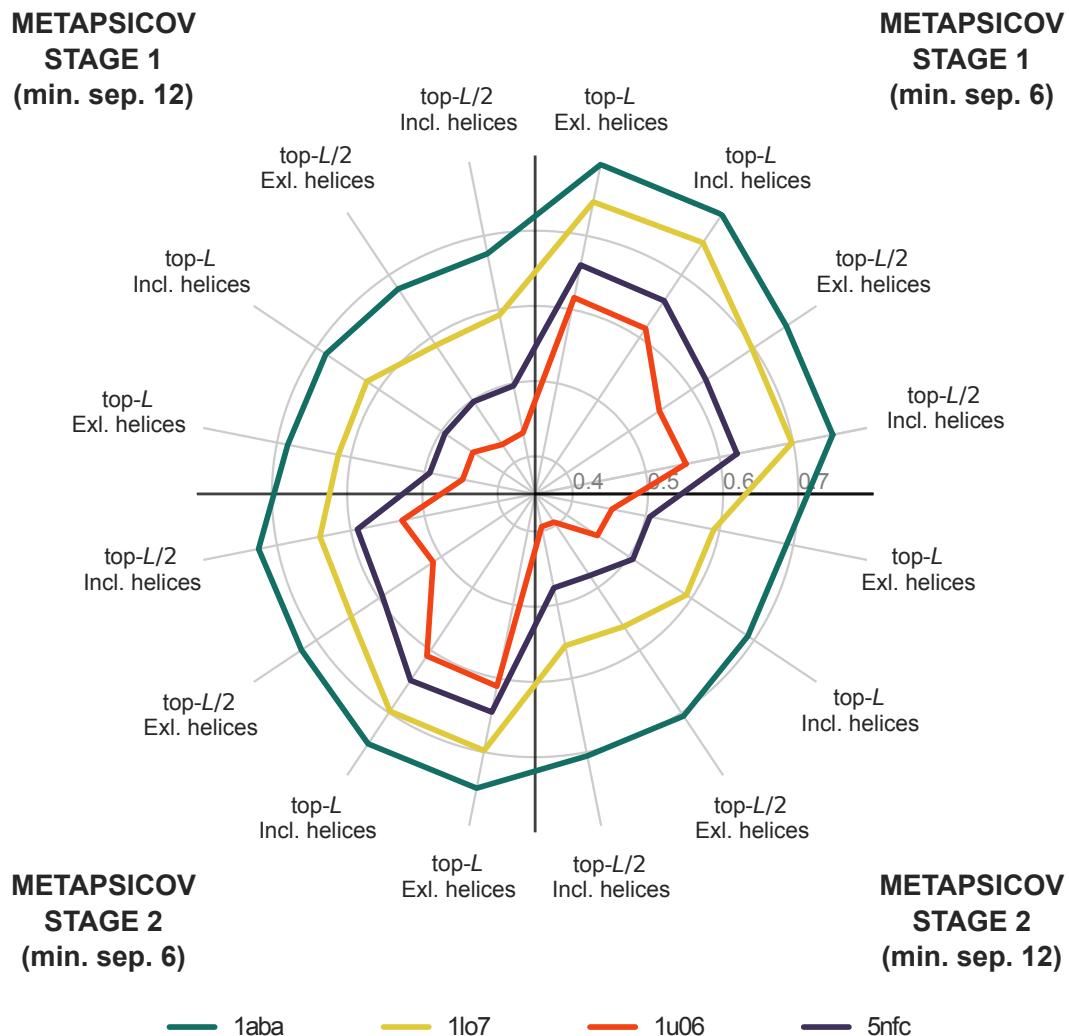


Figure 7.6: Spearman rank-order correlation coefficient analysis of FLIB fragments' FLIB score and RMSD value given the 16 unique fragment picking strategies across four targets. P-values of all Spearman correlations are < 0.001 and not shown for simplicity of the plot.

Further inspection of the fragments and the relationship between each fragment's FLIB score and RMSD value reveals a small subset of outliers in each fragment library. These fragments (hereafter referred to as outlier fragments) are sparse in each library with an overall mean count of $< 0.2\%$. An analysis for unique characteristics of these outliers, which would allow for their exclusion, reveals no unique feature. These fragments contain all secondary structure types, span the entire target sequence and range over all peptide lengths. Furthermore, they occur in all fragment libraries, irrelevant of their original picking strategy. The only characteristic setting these outlier fragments apart from the remaining set is a RMSD value of $> 30\text{\AA}$. Nevertheless, it appears that these outlier fragments with unusually high RMSD values are never included in the final fragment search model set, given that their overall FLIB_{\min} score is 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 supports the positively linear relationship between a fragment's FLIB score and RMSD (Fig. 7.7a). However, the best FLIB fragments by RMSD show much less spread compared to the best fragments by FLIB score (Fig. 7.7b). Furthermore, the size of the fragments also positively correlates with the the FLIB ($\rho_{Spearman} = 0.860$, $p < 0.001$) and RMSD ($\rho_{Spearman} = 0.697$, $p < 0.001$) values. Longer fragments with higher dissimilarity with respect to the target show higher FLIB scores and RMSD values (Fig. 7.7a).

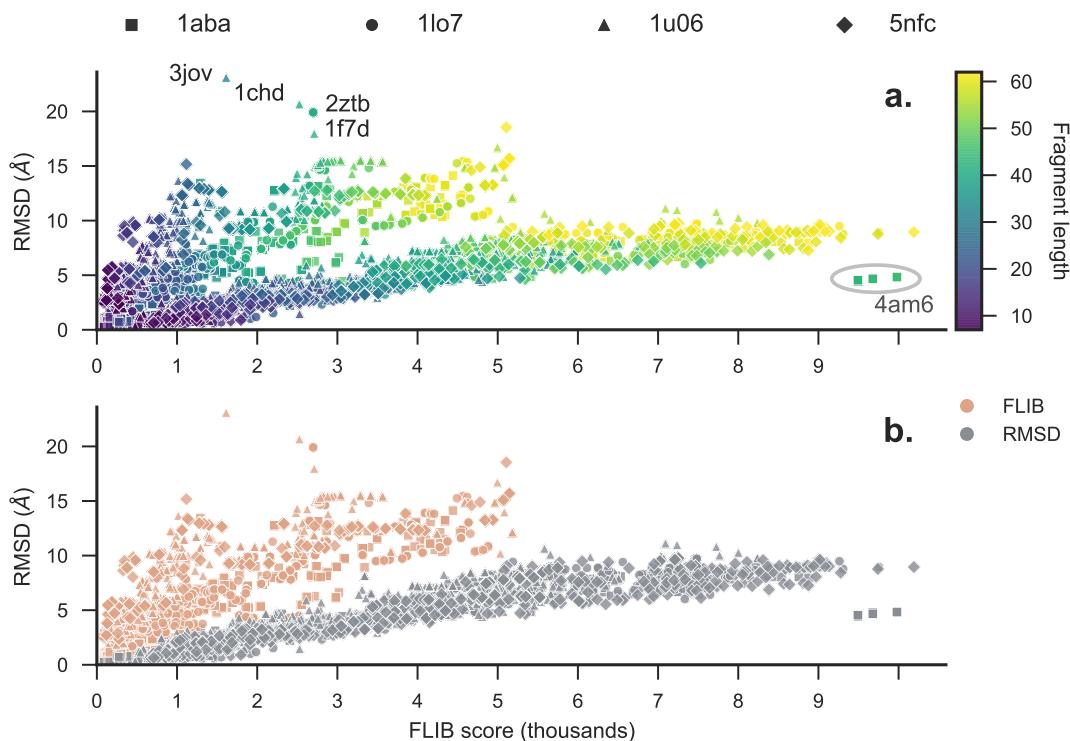


Figure 7.7: Scatterplot highlighting the positive correlation between fragment FLIB scores and RMSD values. The plot contains all fragments independent of target or picking strategy. **a.** The colour of each scatter point illustrates the fragment length. All extreme outlier fragments are highlighted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

Notably, a cluster of large fragments with some of the highest FLIB scores in the set show 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 < 50 residues and lower FLIB scores of < 3000 show the highest RMSD values in the final set.

One further unique aspect of this study compared to other fragment-MR approaches

is the use of 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 satisfy 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) satisfy at least one predicted contact pair (Fig. 7.8). In comparison, 28% of fragments selected for the α -spectrin SH3 domain (PDB ID: 1u06) satisfy at least one contact pair.

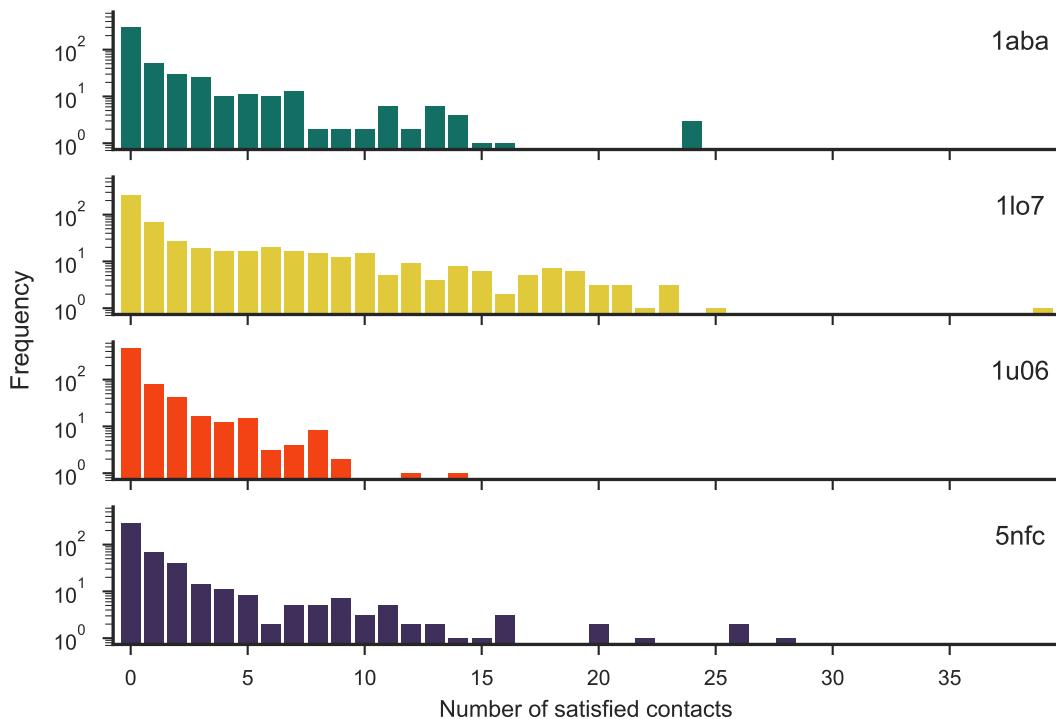


Figure 7.8: Distribution of contact precision for FLIB fragments selected as MR search models separated on a per-target basis.

Thus, the final set of FLIB fragment MR search models spans a wide range of peptide lengths, RMSD values, contact precision 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 dominates the set, increasing the sampling diversity to be undertaken during MR.

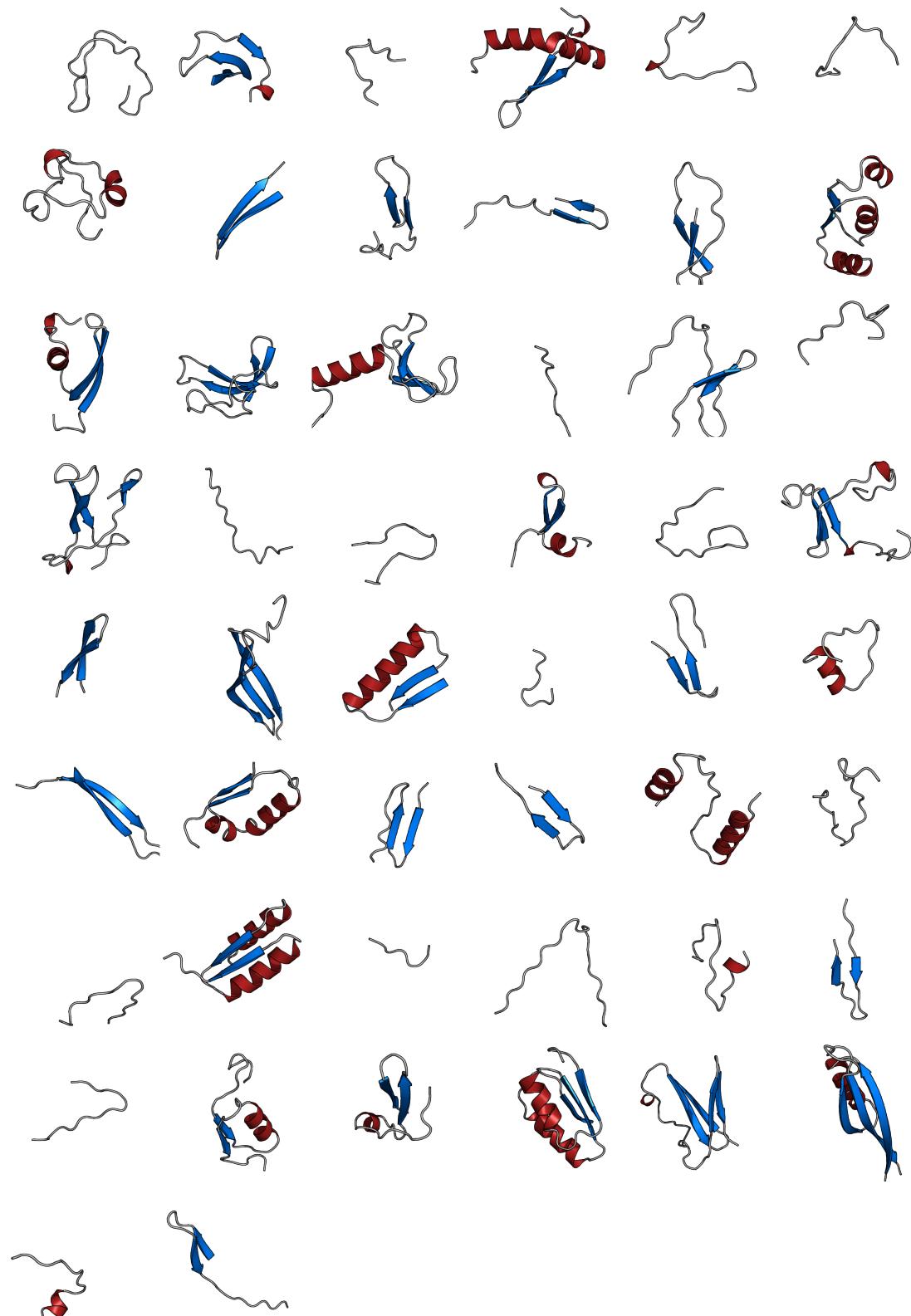


Figure 7.9: Non-redundant sample of FLIB fragment search models selected for four different protein targets. Secondary structure defined by and visualisation done in PyMOL [174]. Unpaired β -strands rendered using the loop style.

7.3.4 Molecular Replacement using FLIB fragments

FLIB fragments picked for four target sequences using a variety of FLIB input options generated $> 6,500$ 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 structure solutions of two targets, namely the T4 glutaredoxin (PDB ID: 1aba) and α -spectrin SH3 domain (PDB ID: 1u06) (Fig. 7.10).

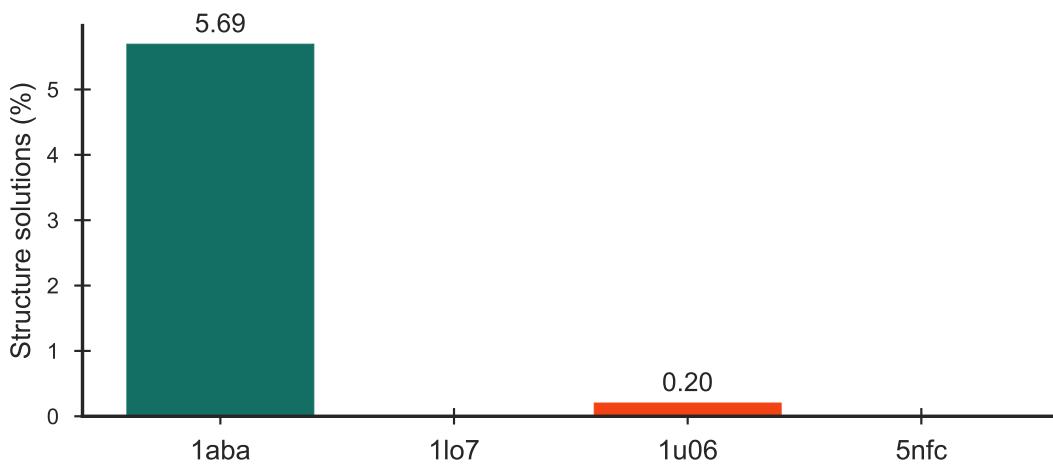


Figure 7.10: Distribution of structure solutions by FLIB target. All MR attempts total to 19,716, out of which 299 are structure solutions. Values above each bar indicates percentage search models successful out of the corresponding set.

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 9 structure solutions of the α -spectrin SH3 domain (PDB ID: 1u06). The largest FLIB fragment leading to a structure solution contained 37 residues and the smallest 10.

A division of FLIB-fragment search models by their respective origin libraries provides strong evidence that METAPSICOV STAGE 1 contact predictions allows for the selection of the most accurate fragments (Fig. 7.4), which directly translates into the structure solution count (Fig. 7.11). Furthermore, this division also highlights and supports the quality of fragment libraries picked with top- L (6 residues sequence separation) METAPSICOV STAGE 1 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, 3rd and 4th bars).

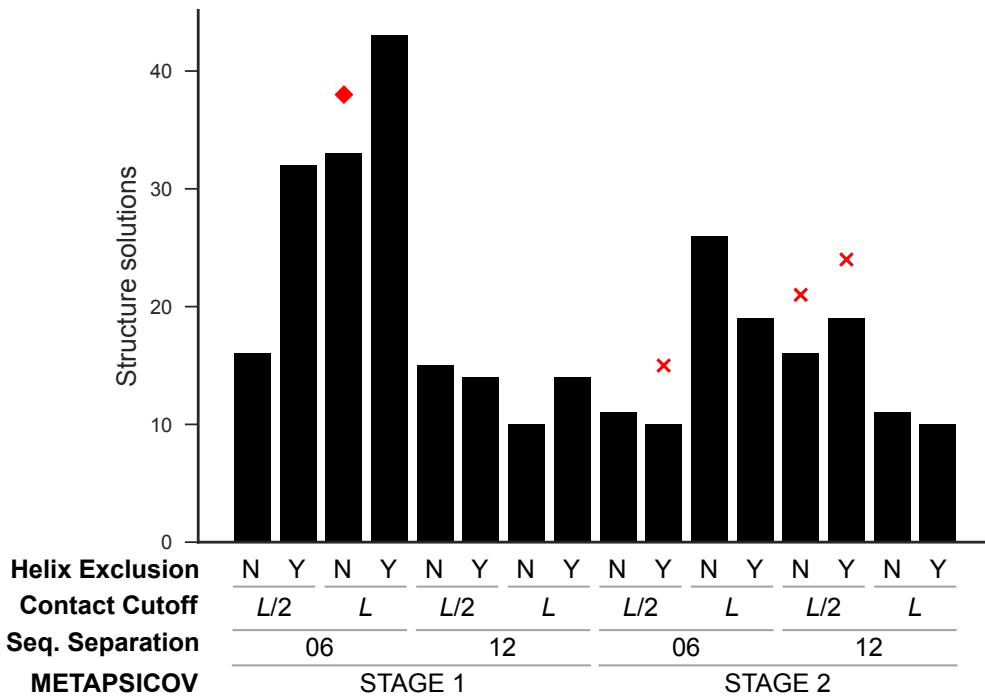


Figure 7.11: Distribution of structure solutions by FLIB library configuration. The optimal fragment picking strategy, as assessed by FLIB values, is highlighted with a red diamond to illustrate that the method that picks the best fragments is close to, but not the absolute best for ultimate structure 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 confirms the expected outcome: the top fragments selected by fragment RMSD score result in more structure solutions than their FLIB score counterparts (Fig. 7.12). To reiterate, all FLIB fragments were grouped by their peptide length, and the top fragment in each group selected when sorted by either FLIB 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 becomes 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 fragment libraries by FLIB score. A further subdivision of successful fragments, sorted either by FLIB scores or RMSD values, highlights that a larger proportion of successful RMSD-sorted fragments satisfied at least 1 contact (FLIB-sorted: 7%; RMSD-sorted: 13%). A separation of attempts by PHASER input RMSD value suggests a value of 0.1 to be the most favourable.

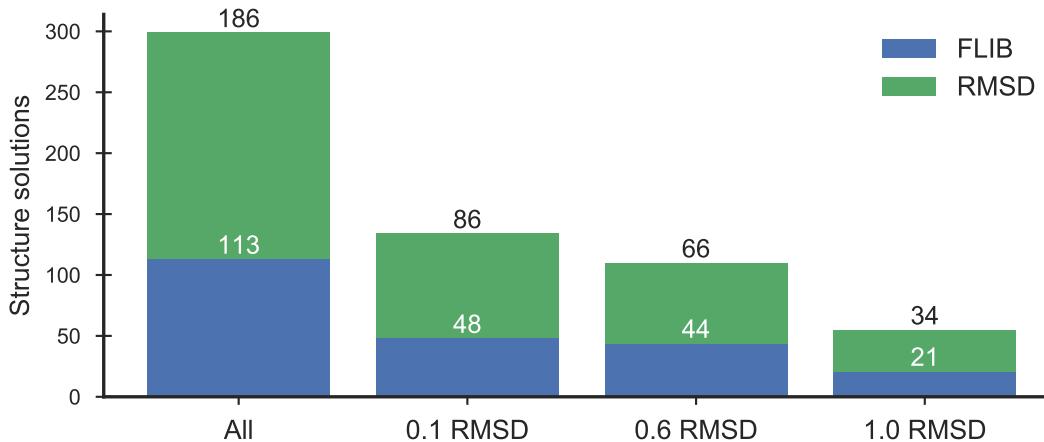


Figure 7.12: Distribution of structure solutions by fragment and MRBUMP configuration. The structure solution count is provided above each bar.

In MR, the correct placement of very small structural fragment may not always be detectable by the 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 [116, 117]. Given that the peptide lengths of FLIB fragments in this study range from 6 to 63 residues, the RIO score is most suitable in validating the correct placement of FLIB-fragment search models. Indeed, all fragments with SHELXE CC ≥ 25 and ACL ≥ 10 contain at least 3 correctly placed C α atoms (i.e. a RIO score ≥ 3). Furthermore, the RIO metric indicates that more than 500 fragments have C α atoms placed within 1.5 \AA of any atom in the target structure. However, only 4 residues are on average placed correctly, which was not enough to achieve structure solution (Fig. 7.13). All successful FLIB fragments have a minimum model- and target-normalised RIO scores of 29.7% and 9.2% (Fig. 7.13, green markers).

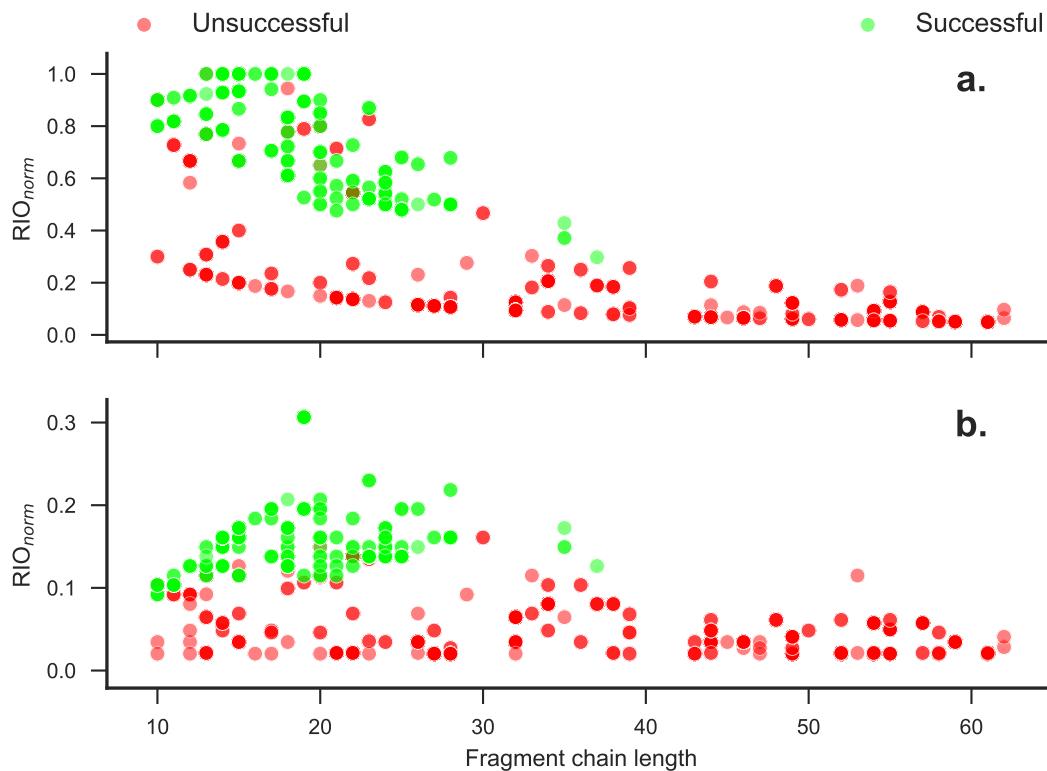


Figure 7.13: Dependence of normalised Residue-Independent Overlap (RIO_{norm}) score on the fragment chain length. The two plots show RIO scores normalised by the chain lengths of (a) the fragment and (b) the target. Colour coding indicates if the FLIB-fragment search model resulted in a structure solution. 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 33 MR attempts more than 60% of a fragment's residues were placed correctly, yet structure solution was not achieved. These trials affect 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 10 and 23 amino acids. The fragments' RMSD values range from 0.19 to 2.72 Å with a mean RMSD of 1.10 Å. Surprisingly, almost all of these fragments contain 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 < 2.0 Å, it is a surprise to not see more structure solutions from these fragments.

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

or 24 ($n = 1$) predicted contacts.

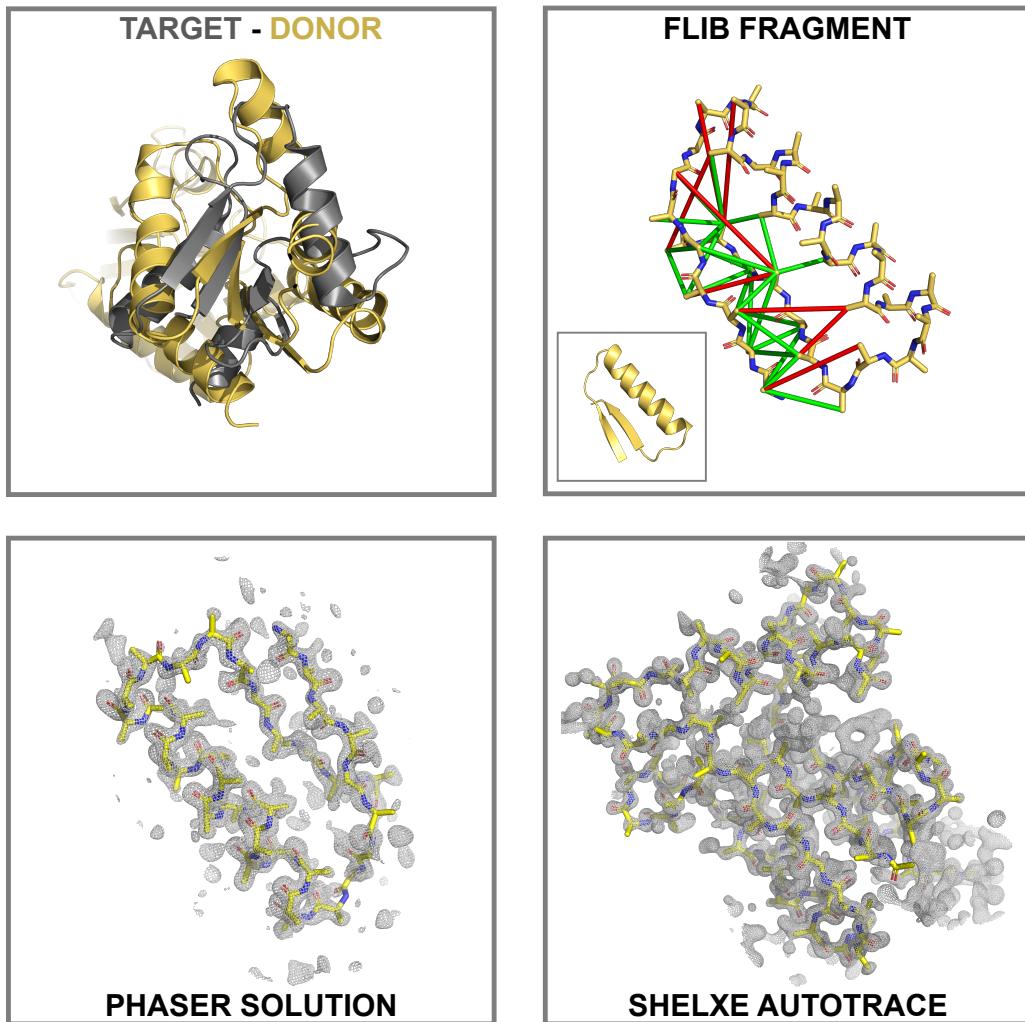


Figure 7.14: Intermediary steps from donor structure to SHELXE main-chain auto-trace for a fragment derived from cobalt chelatase found in *Salmonella typhimurium* (PDB ID: 1qgo). The structure solution was obtained against the target crystallographic data of T4 glutaredoxin (PDB ID: 1aba). METAPSICOV STAGE 2 predicted contacts, against which the fragment was selected, are illustrated with True Positive (green) and False Positive (red) contacts (distance cutoff of 8 Å). 2mFo-DFc electron density maps shown at 2.0 sigma and radius around the peptide atoms of 5 Å. The RMSD between the sequence-independently superposed structures of target and donor is 10.384 Å (computed with the `super` command in PyMOL [174]).

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 contains 35 residues and its supersecondary structure consists of a two-strand β -sheet packing against a single α -helix. The majority of satisfied contact pairs are between C β atoms of the β -strands; however, a small number

of individual contact pairs also identifies the packing of one β -strand against the α -helix (Fig. 7.14, top-right). Although not considered at this stage in the FLIB algorithm, this particular fragment satisfies 75% of all relevant 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.

7.4 Discussion

The main objective of this study was to investigate the application of FLIB structural fragments to MR. Four experimental datasets were chosen and 16 FLIB 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 validate the concept of this approach. Firstly, a positive correlation between a fragment's FLIB score and RMSD value was identified. These correlations were target-independent and found, with various strengths, in all FLIB fragment libraries. Furthermore, this work has identified top- L (6 residue sequence separation) METAPSICOV STAGE 1 contact pairs to be the optimal selection of contact pairs for the FLIB algorithm when starting with METAPSICOV predictions. The additional noise, typically filtered in the second STAGE of the METAPSICOV algorithm [100], allowed for the selection of more accurate fragments across the entire target sequence. Lastly, trialling a selection of high-scoring FLIB fragments in routine MR showed the usefulness of such fragments in attempting to solve protein structures. Two out of four targets were successfully solved albeit only trialling a small proportion of FLIB 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 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 [127]) with very small search models. McCoy et al. [164] 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 [113], ARCIMBOLDO [144], BORGES [145], FRAGON [156] or FRAP [175], also successfully demonstrated MR successes with search models comprising a fraction of the target structure. Thus, MR structure solutions with FLIB fragments as short as 6 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 highlight the success of iden-

tifying accurate fragments solely by the fragment's FLIB score. Given that the FLIB 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 score 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, 176]. 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 solvability of a target structure McCoy et al. [164].

Nevertheless, FLIB fragments with near-identical subfolds to the target might not be traceable by current means of assessing structure solutions. Commonly, MR success is judged by the combination of SHELXE CC and ACL scores [129]. However, it is known that β -strands are notoriously difficult to trace, and thus SHELXE might not pick up on correctly placed search models. 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 algorithms.

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 algorithm through the incorporation of co-evolution 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 [113].

Chapter 8

Conclusion

Appendix A

Appendix

Table A.1: Summary of the ORIGINAL dataset.

PDB ID	Molecule	Resolution (Å)	Space Group	Chain ID	Chain Length	Molecules per ASU	Matthew's Coeffi- cient	Solvent Content (%)	Fold	Citation
1a6m	Oxy-myoglobin	1.00	P2 ₁	A	151	1	1.90	36.00	all- α	[177]
1aba	T4 glutaredoxin	1.45	P2 ₁ 2 ₁ 2 ₁	A	87	1	2.22	44.62	mixed α/β	[178]
1bdo	Biotinyl domain of acetyl-coenzyme A carboxylase	1.80	P2 ₁ 2 ₁ 2	A	80	1	2.48	49.00	all- β	[179]
1bkr	Calponin Homology (CH) domain from β -spectrin	1.10	P2 ₁	A	109	1	2.04	39.80	all- α	[180]
1chd	CheB methyltransferase domain	1.75	P3 ₂ 21	A	203	1	2.35	47.65	mixed α/β	[181]
1e0s	G-protein Arf6-GDP	2.28	P6 ₁ 22	A	174	1	2.18	37.00	mixed α/β	[182]
1eaz	Phosphoinositol (3,4)-bisphosphate	1.40	C222 ₁	A	125	1	2.48	48.00	mixed $\alpha+\beta$	[183]
1hh8	N-terminal region of P67Phox	1.80	P3 ₁	A	213	1	2.71	45.00	all- α	[184]
1kjl	Galectin-3 domain	1.40	P2 ₁ 2 ₁ 2 ₁	A	146	1	2.15	42.68	all- β	[185]
1kw4	Polyhomeotic SAM domain	1.75	P6 ₅	A	89	1	2.25	45.27	all- α	[186]
1l07	4-hydroxybenzoyl CoA thioesterase	1.50	I222	A	141	1	2.06	40.22	mixed $\alpha+\beta$	[187]
1npu	Extracellular domain of murine PD-1	2.00	P2 ₁ 2 ₁ 2 ₁	A	117	1	1.67	25.80	all- β	[188]
1pnc	Poplar plastocyanin	1.60	P2 ₁ 2 ₁ 2 ₁	A	99	1	1.82	32.48	all- β	[189]
1pjx	Synaptotagmin I C2B domain	1.04	P3 ₂ 21	A	159	1	2.40	48.00	mixed $\alpha+\beta$	[190]
1tlv	LicT PRD	1.95	P3 ₂ 21	A	221	1	2.80	50.00	all- α	[191]
2nuz	α -spectrin SH3 domain	1.85	P2 ₁ 2 ₁ 2 ₁	A	62	1	2.57	52.16	all- β	
2qyj	Ankyrin C2 domain	2.05	P6 ₁	A	166	1	2.28	45.99	all- α	[192]
3w56	N-terminal bromodomain of Brd4	1.60	I2	A	131	1	2.05	40.10	all- β	[193]
4cl9	FN3con	1.40	P2 ₁ 2 ₁ 2 ₁	A	127	1	2.21	44.37	all- α	[194]
4t3h		1.98	P4 ₁ 32	A	100	1	2.47	50.27	all- β	[195]
4w97	KstR2	1.60	C2	A	200	1	2.75	55.25	all- α	[196]

Table A.2: Summary of the PREDICTORS dataset.

PDB ID	Molecule	ResolutionSpace (Å)	Space Group	Chain ID	Chain Length	Molecules per ASU	Matthew's Solvent Content (%)	Fold	Citation	
1fcy	Retinoic acid nuclear receptor HRAR	1.30	P4 ₁ 2 ₁ 2	A	236	1	2.25	45.50	all-α	
1fgv	Peptide methionine sulfoxide reductase	1.60	C121	A	199	1	2.10	41.55	mixed α+β [198]	
1gm4	Cytochrome C3	2.05	P6 ₁ 22	A	107	1	2.48	50.43	all-α	
1gy8	N-II domain of ovotransferrin	1.95	P3 ₁	A	159	1	2.24	45.00	mixed α/β [200]	
1k40	FAT domain of focal adhesion kinase	2.25	C121	A	126	1	2.21	44.40	all-α	
1oee	Hypothetical protein YodA	2.10	C121	A	193	1	2.30	46.20	all-β	
1oz9	Hypothetical protein AQ_1354	1.89	P4 ₃ 2 ₁ 2	A	150	1	2.76	55.07	mixed α+β [202]	
1o8c	Hypothetical protein MG027	2.00	P4 ₁	A	151	1	2.42	49.25	all-α	
1lh	Conserved hypothetical protein	1.80	P6 ₃	A	173	1	2.12	41.98	mixed α+β [204]	
1s2x	Cag-Z	1.90	P2 ₁ 2 ₁ 2 ₁	A	206	1	2.74	54.70	all-α	
1u61	Putative Ribonuclease III	2.15	I4 ₁ 32	A	138	1	6.50	80.80	all-α	
1zxu	At5g01750 protein	1.70	P2 ₁ 2 ₁ 2 ₁	A	217	1	2.50	50.20	mixed α+β [206]	
2eum	Glycolipid transfer protein	2.30	C121	A	209	1	2.25	45.39	all-α	
2o18	Outer surface protein A	1.90	P12 ₁ 1	O	249	1	2.19	43.87	all-β [207]	
2eqz	Sortase B	1.60	P12 ₁ 1	A	223	1	2.07	40.71	all-β [208]	
2x6u	T-Box transcription factor TBX5	1.90	P2 ₁ 2 ₁ 2 ₁	A	203	1	2.20	44.21	all-β [209]	
2y64	Xylanase	1.40	P2 ₁ 2 ₁ 2 ₁	A	167	1	2.15	43.00	all-β [210]	
2yjm	TtrD	1.84	C121	A	176	1	2.08	40.80	all-α [211]	
2yq9	2, 3-cyclic-nucleotide phosphodiesterase	3-	I4 ₁ 90	P2 ₁ 2 ₁ 2 ₁	A	221	1	2.10	41.70	mixed α+β [212]
3dju	Protein BTG2	2.26	P2 ₁ 2 ₁ 2 ₁	B	122	1	1.98	37.73	mixed α+β [213]	
3g0m	Cysteine desulfurase protein suffE	1.76	P12 ₁ 1	A	141	1	1.88	34.58	mixed α+β [214]	
3cqz	Iron-regulated surface determinant protein A	1.30	P2 ₁ 2 ₁ 2	A	127	1	2.42	49.12	all-β [215]	
4aaj	N-(5-phosphoribosyl)anthranilate isomerase	1.75	P6 ₁	A	228	1	2.38	48.30	mixed α/β [216]	
4dbb	Amyloid-β A4 precursor protein-binding family A1	1.90	P4 ₁ 2 ₁ 2	A	162	1	3.25	62.10	all-β [217]	
4e9e	Methyl-CpG-binding domain protein 4	1.90	H3	A	161	1	2.42	49.23	all-α [218]	
4bjj	Galectin-3	1.80	P2 ₁ 2 ₁ 2 ₁	A	138	1	2.09	41.01	all-β [219]	
4pg0	Hypothetical protein PF0907	2.30	P6 ₅ 22	A	116	1	3.25	62.10	all-β [219]	

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
Igu8	Sensory rhodopsin II	2.27	C2221	A	239	1	2.75	53.00	all- α	[220]
2bhw	Chlorophyll A-B binding protein	2.50	C121	A	232	3	4.10	69.00	all- α	[221]
AB80	Aquaporin aquaporin Z	2.30	I4	A	246	1	3.38	63.57	all- α	[222]
2evu	ATP synthase C chain	1.90	I4	A	234	1	3.34	63.19	all- α	[223]
2o9g	Rhomboid protease GLPG	2.13	P6 ₃ 22	A	82	5	3.41	68.00	all- α	[224]
2wie	Aquaporin 4	1.65	H32	A	181	1	3.50	64.92	all- α	[225]
2xov	Bacteriorhodopsin	1.80	P42 ₁ 2	A	223	1	2.73	54.97	all- α	[226]
3gd8	Calcium-gated potassium channel	1.60	C2221	A	249	1	2.73	54.99	all- α	[227]
3hap	Leukotriene C4 synthase	1.45	P42 ₁ 2	A	82	1	2.48	50.44	all- α	[228]
3ldc	ThiT	1.55	I2	A	97	2	2.40	48.76	all- α	[229]
3ouf	ATP synthase subunit C	1.90	F23	A	156	1	4.91	74.77	all- α	[230]
3pcv	Biotin transporter BioY	2.00	C121	A	192	2	3.89	68.39	all- α	[231]
3rlb	3u2f	2.00	P42 ₂ 2	K	76	5	2.32	46.92	all- α	[232]
4dve		2.09	C121	A	198	3	3.27	62.40	all- α	[233]

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* **July 1913**, *88*, 428–438.
- [4] W. L. Bragg, *Scientia* **1929**, *23*, 153.
- [5] W. L. Bragg, *Nature* **Dec. 1912**, *90*, 410.
- [6] J. D. Watson, F. H. C. Crick, Others, *Nature* **1953**, *171*, 737–738.
- [7] D. C. Hodgkin, J Kamper, M Mackay, J Pickworth, K. N. Trueblood, J. G. White, en, *Nature* **July 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, en, *Nature* **June 1971**, *231*, 506–511.
- [9] C. C. Blake, D. F. Koenig, G. A. Mair, A. C. North, D. C. Phillips, V. R. Sarma, en, *Nature* **May 1965**, *206*, 757–761.
- [10] M. F. Perutz, M. G. Rossmann, A. F. Cullis, H Muirhead, G Will, A. C. North, en, *Nature* **Feb. 1960**, *185*, 416–422.
- [11] J. C. Kendrew, G Bodo, H. M. Dintzis, R. G. Parrish, H Wyckoff, D. C. Phillips, en, *Nature* **Mar. 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.* **Jan. 2000**, *28*, 235–242.
- [13] B. Rupp, *Biomolecular crystallography : principles, practice, and application to structural biology*, English, Garland Science, New York, **2010**.
- [14] M. G. Rossmann, *Acta Crystallogr. D Biol. Crystallogr.* **2001**, *57*, 1360–1366.
- [15] M. G. Rossmann, en, *Acta Crystallogr. A* **Feb. 1990**, *46* (Pt 2), 73–82.
- [16] C. R. Kissinger, D. K. Gehlhaar, D. B. Fogel, en, *Acta Crystallogr. D Biol. Crystallogr.* **Feb. 1999**, *55*, 484–491.
- [17] N. M. Glykos, M Kokkinidis, en, *Acta Crystallogr. D Biol. Crystallogr.* **Feb. 2000**, *56*, 169–174.
- [18] R. J. Read, en, *Acta Crystallogr. D Biol. Crystallogr.* **Oct. 2001**, *57*, 1373–1382.
- [19] M. G. Rossmann, D. M. Blow, *Acta Crystallogr.* **Jan. 1962**, *15*, 24–31.

- [20] M. Bayes, M. Price, *Philosophical Transactions of the Royal Society of London* **Jan.** **1763**, *53*, 370–418.
- [21] L. C. Storoni, A. J. McCoy, R. J. Read, en, *Acta Crystallogr. D Biol. Crystallogr.* **Mar.** **2004**, *60*, 432–438.
- [22] B. C. Wang, en, *Methods Enzymol.* **1985**, *115*, 90–112.
- [23] V. Y. Lunin, *Acta Crystallogr. A* **Mar.** **1988**, *44*, 144–150.
- [24] G. M. Sheldrick, *Zeitschrift für Kristallographie - Crystalline Materials* **Jan.** **2002**, *217*, 371.
- [25] T. C. Terwilliger, en, *Acta Crystallogr. D Biol. Crystallogr.* **Aug.** **2000**, *56*, 965–972.
- [26] 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, en, *Acta Crystallogr. D Biol. Crystallogr.* **Apr.** **2012**, *68*, 352–367.
- [27] 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, en, *Acta Crystallogr. D Biol. Crystallogr.* **Apr.** **2011**, *67*, 355–367.
- [28] G. M. Sheldrick, en, *Acta Crystallogr. D Biol. Crystallogr.* **Apr.** **2010**, *66*, 479–485.
- [29] V. S. Lamzin, A Perrakis, K. S. Wilson, *International Tables for Crystallography* **2001**, 720–722.
- [30] T. Terwilliger, en, *J. Synchrotron Radiat.* **Jan.** **2004**, *11*, 49–52.
- [31] K. Cowtan, en, *Acta Crystallogr. D Biol. Crystallogr.* **Sept.** **2006**, *62*, 1002–1011.
- [32] B. Qian, S. Raman, R. Das, P. Bradley, A. J. McCoy, R. J. Read, D. Baker, en, *Nature* **Nov.** **2007**, *450*, 259–264.
- [33] D. J. Rigden, R. M. Keegan, M. D. Winn, en, *Acta Crystallogr. D Biol. Crystallogr.* **Dec.** **2008**, *64*, 1288–1291.
- [34] R. Das, D. Baker, en, *Acta Crystallogr. D Biol. Crystallogr.* **Feb.** **2009**, *65*, 169–175.
- [35] P. E. Leopold, M Montal, J. N. Onuchic, en, *Proc. Natl. Acad. Sci. U. S. A.* **Sept.** **1992**, *89*, 8721–8725.
- [36] C. B. Anfinsen, en, *Science* **July** **1973**, *181*, 223–230.
- [37] C. Levinthal, *Mossbauer spectroscopy in biological systems* **1969**, *67*, 22–24.
- [38] M. Karplus, en, *Nat. Chem. Biol.* **June** **2011**, *7*, 401–404.
- [39] Wikipedia, Folding Funnel — Wikipedia, The Free Encyclopedia, [Online; accessed 09-April-2018], **2004**.
- [40] J. Lee, P. L. Freddolino, Y. Zhang in *From Protein Structure to Function with Bioinformatics* (*2nd Ed.*) Vol. 69, (Ed.: D. J. Rigden), Springer Netherlands, Dordrecht, **2017**, pp. 3–35.
- [41] J. Skolnick, en, *Curr. Opin. Struct. Biol.* **Apr.** **2006**, *16*, 166–171.
- [42] C. A. Rohl, C. E. M. Strauss, K. M. S. Misura, D. Baker, en, *Methods Enzymol.* **2004**, *383*, 66–93.
- [43] D. Xu, Y. Zhang, en, *Proteins* **July** **2012**, *80*, 1715–1735.
- [44] M. Blaszczyk, M. Jamroz, S. Kmiecik, A. Kolinski, en, *Nucleic Acids Res.* **July** **2013**, *41*, W406–11.

- [45] T. Kosciolek, D. T. Jones, en, *PLoS One* **Mar. 2014**, *9*, e92197.
- [46] S. H. P. de Oliveira, E. C. Law, J. Shi, C. M. Deane, en, *Bioinformatics* **Nov. 2017**, DOI 10.1093/bioinformatics/btx722.
- [47] J. Abbass, J.-C. Nebel, en, *BMC Bioinformatics* **Apr. 2015**, *16*, 136.
- [48] Y. Shen, G. Picord, F. Guyon, P. Tuffery, en, *PLoS One* **Nov. 2013**, *8*, e80493.
- [49] S. C. Li, D. Bu, X. Gao, J. Xu, M. Li, en, *Bioinformatics* **July 2008**, *24*, i182–9.
- [50] I. Klev, M. Habeck, en, *Bioinformatics* **Nov. 2011**, *27*, 3110–3116.
- [51] D. Bhattacharya, B. Adhikari, J. Li, J. Cheng, en, *Bioinformatics* **July 2016**, *32*, 2059–2061.
- [52] T. Wang, Y. Yang, Y. Zhou, H. Gong, en, *Bioinformatics* **Mar. 2017**, *33*, 677–684.
- [53] S. H. P. de Oliveira, J. Shi, C. M. Deane, en, *PLoS One* **Apr. 2015**, *10*, e0123998.
- [54] D. Gront, D. W. Kulp, R. M. Vernon, C. E. M. Strauss, D. Baker, en, *PLoS One* **Aug. 2011**, *6*, e23294.
- [55] N Metropolis, S Ulam, en, *J. Am. Stat. Assoc.* **Sept. 1949**, *44*, 335–341.
- [56] D Shortle, K. T. Simons, D Baker, en, *Proc. Natl. Acad. Sci. U. S. A.* **Sept. 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, en, *Science* **Sept. 2005**, *309*, 1868–1871.
- [59] S Oldziej, C Czaplewski, A Liwo, M Chinchio, M Nania, 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, en, *Proc. Natl. Acad. Sci. U. S. A.* **May 2005**, *102*, 7547–7552.
- [60] Y. Zhang, J. Skolnick, en, *Proc. Natl. Acad. Sci. U. S. A.* **May 2004**, *101*, 7594–7599.
- [61] J. Yang, R. Yan, A. Roy, D. Xu, J. Poisson, Y. Zhang, en, *Nat. Methods* **Jan. 2015**, *12*, 7–8.
- [62] A. Kryshtafovych, A. Barbato, B. Monastyrskyy, K. Fidelis, T. Schwede, A. Tramontano, en, *Proteins* **Sept. 2016**, *84 Suppl 1*, 349–369.
- [63] C.-H. Tai, H. Bai, T. J. Taylor, B. Lee, en, *Proteins* **Feb. 2014**, *82 Suppl 2*, 57–83.
- [64] Z. He, M. Alazmi, J. Zhang, D. Xu, en, *PLoS One* **Sept. 2013**, *8*, e74006.
- [65] L. Kinch, S. Yong Shi, Q. Cong, H. Cheng, Y. Liao, N. V. Grishin, en, *Proteins* **Oct. 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, en, *Proc. Natl. Acad. Sci. U. S. A.* **July 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, en, *Science* **Feb. 2010**, *327*, 1014–1018.
- [68] C. Göbl, T. Madl, B. Simon, M. Sattler, en, *Prog. Nucl. Magn. Reson. Spectrosc.* **July 2014**, *80*, 26–63.
- [69] D. S. Marks, L. J. Colwell, R. Sheridan, T. A. Hopf, A. Pagnani, R. Zecchina, C. Sander, en, *PLoS One* **Dec. 2011**, *6*, e28766.

- [70] M. Michel, S. Hayat, M. J. Skwark, C. Sander, D. S. Marks, A. Elofsson, en, *Bioinformatics* **Sept. 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, en, *Elife* **Sept. 2015**, *4*, e09248.
- [72] S. Ovchinnikov, D. E. Kim, R. Y.-R. Wang, Y. Liu, F. DiMaio, D. Baker, en, *Proteins Sept. 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. Kyripides, D. Baker, en, *Science* **Jan. 2017**, *355*, 294–298.
- [75] S. Wang, S. Sun, Z. Li, R. Zhang, J. Xu, en, *PLoS Comput. Biol.* **Jan. 2017**, *13*, e1005324.
- [76] W. R. Taylor, K. Hatrick, en, *Protein Eng. Mar.* **1994**, *7*, 341–348.
- [77] U Göbel, C. Sander, R. Schneider, A. Valencia, en, *Proteins Apr.* **1994**, *18*, 309–317.
- [78] E. Neher, en, *Proc. Natl. Acad. Sci. U. S. A. Jan.* **1994**, *91*, 98–102.
- [79] I. N. Shindyalov, N. A. Kolchanov, C. Sander, en, *Protein Eng. Mar.* **1994**, *7*, 349–358.
- [80] D. D. Pollock, W. R. Taylor, en, *Protein Eng. June* **1997**, *10*, 647–657.
- [81] A. S. Lapedes, B. Giraud, L. Liu, G. D. Stormo, en in *Statistics in molecular biology and genetics*, Institute of Mathematical Statistics, **1999**, pp. 236–256.
- [82] A. Lapedes, B. Giraud, C. Jarzynski, **July 2012**.
- [83] F. Simkovic, S. Ovchinnikov, D. Baker, D. J. Rigden, *IUCrJ May* **2017**, *4*, 291–300.
- [84] M. Weigt, R. A. White, H. Szurmant, J. A. Hoch, T. Hwa, en, *Proc. Natl. Acad. Sci. U. S. A. Jan.* **2009**, *106*, 67–72.
- [85] L. Burger, E. van Nimwegen, en, *PLoS Comput. Biol. Jan.* **2010**, *6*, e1000633.
- [86] S. Balakrishnan, H. Kamisetty, J. G. Carbonell, S.-I. Lee, C. J. Langmead, en, *Proteins Apr.* **2011**, *79*, 1061–1078.
- [87] F. Morcos, A. Pagnani, B. Lunt, A. Bertolino, D. S. Marks, C. Sander, R. Zecchina, J. N. Onuchic, T. Hwa, M. Weigt, en, *Proceedings of the National Academy of Sciences Dec. 2011*, *108*, E1293–E1301.
- [88] D. T. Jones, D. W. A. Buchan, D. Cozzetto, M. Pontil, *Bioinformatics Jan.* **2012**, *28*, 184–190.
- [89] M. Ekeberg, C. Lökvist, Y. Lan, M. Weigt, E. Aurell, en, *Phys. Rev. E Stat. Nonlin. Soft Matter Phys. Jan.* **2013**, *87*, 012707.
- [90] H. Kamisetty, S. Ovchinnikov, D. Baker, *Proceedings of the National Academy of Sciences Sept. 2013*, *110*, 15674–15679.
- [91] S. Seemayer, M. Gruber, J. Söding, en, *Bioinformatics Nov.* **2014**, *30*, 3128–3130.
- [92] 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.
- [93] R. R. Stein, D. S. Marks, C. Sander, en, *PLoS Comput. Biol. July* **2015**, *11*, e1004182.
- [94] T. A. Hopf, S. Morinaga, S. Ihara, K. Touhara, D. S. Marks, R. Benton, en, *Nat. Commun. Jan.* **2015**, *6*, 6077.

- [95] S. D. Dunn, L. M. Wahl, G. B. Gloor, en, *Bioinformatics* **Feb. 2008**, *24*, 333–340.
- [96] D. S. Marks, T. A. Hopf, C. Sander, en, *Nat. Biotechnol.* **Nov. 2012**, *30*, 1072–1080.
- [97] J. Andreani, J. Söding, en, *Bioinformatics* **June 2015**, *31*, 1729–1737.
- [98] T. A. Hopf, L. J. Colwell, R. Sheridan, B. Rost, C. Sander, D. S. Marks, en, *Cell* **June 2012**, *149*, 1607–1621.
- [99] M. J. Skwark, D. Raimondi, M. Michel, A. Elofsson, en, *PLoS Comput. Biol.* **Nov. 2014**, *10*, e1003889.
- [100] D. T. Jones, T. Singh, T. Kosciolek, S. Tetchner, en, *Bioinformatics* **Apr. 2015**, *31*, 999–1006.
- [101] T. Du, L. Liao, C. H. Wu, B. Sun, en, *Methods* **Nov. 2016**, *110*, 97–105.
- [102] A. J. González, L Liao, C. H. Wu, *Bioinformatics* **2013**.
- [103] G. Shackelford, K. Karplus, en, *Proteins* **2007**, *69 Suppl 8*, 159–164.
- [104] J. Cheng, P. Baldi, en, *Bioinformatics* **June 2005**, *21 Suppl 1*, i75–84.
- [105] H. Zhang, Q. Huang, Z. Bei, Y. Wei, C. A. Floudas, en, *Proteins: Struct. Funct. Bioinf.* **Mar. 2016**, *84*, 332–348.
- [106] Z. Wang, J. Xu, en, *Bioinformatics* **July 2013**, *29*, i266–73.
- [107] J. Ma, S. Wang, Z. Wang, J. Xu, en, *Bioinformatics* **Nov. 2015**, *31*, 3506–3513.
- [108] B. Adhikari, J. Hou, J. Cheng, en, *Bioinformatics* **Dec. 2017**, DOI 10.1093/bioinformatics/btx781.
- [109] B. He, S. M. Mortuza, Y. Wang, H. B. Shen, Y. Zhang, en, *Bioinformatics* **Mar. 2017**, *33*, 2296–2306.
- [110] M. Michel, M. J. Skwark, D. M. Hurtado, M. Ekeberg, A. Elofsson, en, *Bioinformatics* **Sept. 2017**, *33*, 2859–2866.
- [111] S. H. P. De Oliveira, J. Shi, C. M. Deane, en, *Bioinformatics* **Feb. 2017**, *33*, 373–381.
- [112] Q. Wuyun, W. Zheng, Z. Peng, J. Yang, *Brief. Bioinform.* **Oct. 2016**, bbw106.
- [113] J. Bibby, R. M. Keegan, O. Mayans, M. D. Winn, D. J. Rigden, en, *Acta Crystallogr. D Biol. Crystallogr.* **Dec. 2012**, *68*, 1622–1631.
- [114] R. M. Keegan, J. Bibby, J. M. H. Thomas, D. Xu, Y. Zhang, O. Mayans, M. D. Winn, D. J. Rigden, en, *Acta Crystallogr. D Biol. Crystallogr.* **Feb. 2015**, *71*, 338–343.
- [115] F. Simkovic, J. M. H. H. Thomas, R. M. Keegan, M. D. Winn, O. Mayans, D. J. Rigden, en, *IUCrJ* **July 2016**, *3*, 259–270.
- [116] J. M. H. Thomas, R. M. Keegan, J. Bibby, M. D. Winn, O. Mayans, D. J. Rigden, en, *IUCrJ* **Mar. 2015**, *2*, 198–206.
- [117] J. M. H. Thomas, F. Simkovic, R. M. Keegan, O. Mayans, Y. Zhang, D. J. Rigden, C. Zhang, Y. Zhang, D. J. Rigden, *Acta Crystallographica Section D Structural Biology* **Dec. 2017**, *73*, 985–996.
- [118] J. Bibby, R. M. Keegan, O. Mayans, M. D. Winn, D. J. Rigden, en, *Acta Crystallogr. D Biol. Crystallogr.* **Nov. 2013**, *69*, 2194–2201.
- [119] D. J. Rigden, J. M. H. Thomas, F. Simkovic, A. Simpkin, M. D. Winn, O. Mayans, R. M. Keegan, *Acta Crystallographica Section D Structural Biology* **Mar. 2018**, *74*, 183–193.

- [120] 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.
- [121] 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, en, *Angewandte Chemie - International Edition* **Jan.** **2014**, *53*, 824–828.
- [122] S.-W. Yeh, J.-W. Liu, S.-H. Yu, C.-H. Shih, J.-K. Hwang, J. Echave, en, *Mol. Biol. Evol.* **Jan.** **2014**, *31*, 135–139.
- [123] C.-H. Shih, C.-M. Chang, Y.-S. Lin, W.-C. Lo, J.-K. Hwang, en, *Proteins* **June** **2012**, *80*, 1647–1657.
- [124] D. L. Theobald, D. S. Wuttke, en, *Bioinformatics* **Sept.** **2006**, *22*, 2171–2172.
- [125] G. G. Krivov, M. V. Shapovalov, R. L. Dunbrack, *Proteins: Struct. Funct. Bioinf.* **2009**, *77*, 778–795.
- [126] 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 Crystallographica Section D Structural Biology* **Mar.** **2018**, *74*, 167–182.
- [127] A. J. McCoy, R. W. Grosse-Kunstleve, P. D. Adams, M. D. Winn, L. C. Storoni, R. J. Read, en, *J. Appl. Crystallogr.* **Aug.** **2007**, *40*, 658–674.
- [128] A. Vagin, A. Teplyakov, en, *Acta Crystallogr. D Biol. Crystallogr.* **Jan.** **2010**, *66*, 22–25.
- [129] A. Thorn, G. M. Sheldrick, en, *Acta Crystallogr. D Biol. Crystallogr.* **Nov.** **2013**, *69*, 2251–2256.
- [130] 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, en, *Acta Crystallogr. D Biol. Crystallogr.* **Jan.** **2007**, *64*, 49–60.
- [131] 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, en, *Nucleic Acids Res.* **Jan.** **2016**, *44*, D279–D285.
- [132] J. M. Chandonia, N. K. Fox, S. E. Brenner, en, *J. Mol. Biol.* **Feb.** **2017**, *429*, 348–355.
- [133] G. E. Tusnády, Z. Dosztányi, I. Simon, en, *Nucleic Acids Res.* **Jan.** **2005**, *33*, D275–8.
- [134] S. Hayat, C. Sander, D. S. Marks, A. Elofsson, en, *Proc. Natl. Acad. Sci. U. S. A.* **Apr.** **2015**, *112*, 5413–5418.
- [135] M. Remmert, A. Biegert, A. Hauser, J. Söding, en, *Nat. Methods* **Dec.** **2011**, *9*, 173–175.
- [136] A. Bateman, M. J. Martin, C. O'Donovan, M. Magrane, E. Alpi, R. Antunes, B. Bely, M. Bingley, C. Bonilla, R. Britto, B. Bursteinas, H. Bye-AJee, A. Cowley, A. Da Silva, M. De Giorgi, T. Dogan, F. Fazzini, L. G. Castro, L. Figueira, P. Garmiri, G. Georghiou, D. Gonzalez, E. Hatton-Ellis, W. Li, W. Liu, R. Lopez, J. Luo, Y. Lussi, A. MacDougall, A. Nightingale, B. Palka, K. Pichler, D. Poggioli, S. Pundir, L. Pureza, G. Qi, S. Rosanoff, R. Saidi, T. Sawford, A. Shypitsyna, E. Speretta, E. Turner, N. Tyagi, V. Volynkin, T. Wardell, K. Warner, X. Watkins, R. Zaru, H. Zellner, I. Xenarios, L. Bougueret, A. Bridge, S. Poux, N. Redaschi, L. Aimo, G. ArgoudPuy, A. Auchincloss, K. Axelsen, P. Bansal, D. Baratin, M. C. Blatter, B. Boeckmann, J. Bolleman, E. Boutet, L. Breuza, C. Casal-Casas, E. De Castro, E. Coudert, B. Cuche, M. Doche, D. Dornevile, S. Duvaud,

- A. Estreicher, L. Famiglietti, M. Feuermann, E. Gasteiger, S. Gehant, V. Gerritsen, A. Gos, N. Gruaz-Gumowski, U. Hinz, C. Hulo, F. Jungo, G. Keller, V. Lara, P. Lemercier, D. Lieberherr, T. Lombardot, X. Martin, P. Masson, A. Morgat, T. Neto, N. Nouspikel, S. Paesano, I. Pedruzzi, S. Pilbout, M. Pozzato, M. Pruess, C. Rivoire, B. Roechert, M. Schneider, C. Sigrist, K. Sonesson, S. Staehli, A. Stutz, S. Sundaram, M. Tognolli, L. Verbregue, A. L. Veuthey, C. H. Wu, C. N. Arighi, L. Arminski, C. Chen, Y. Chen, J. S. Garavelli, H. Huang, K. Laiho, P. McGarvey, D. A. Natale, K. Ross, C. R. Vinayaka, Q. Wang, Y. Wang, L. S. Yeh, J. Zhang, en, *Nucleic Acids Res.* **Jan. 2017**, *45*, D158–D169.
- [137] J. Söding, en, *Bioinformatics* **Apr. 2005**, *21*, 951–960.
- [138] F. Simkovic, J. M. H. Thomas, D. J. Rigden, en, *Bioinformatics* **Mar. 2017**, *33*, 2209–2211.
- [139] Y. Zhang, J. Skolnick, en, *Proteins: Structure Function and Genetics* **Dec. 2004**, *57*, 702–710.
- [140] J. Xu, Y. Zhang, en, *Bioinformatics* **Apr. 2010**, *26*, 889–895.
- [141] M Fujinaga, R. J. Read, *J. Appl. Crystallogr.* **Dec. 1987**, *20*, 517–521.
- [142] B. Adhikari, D. Bhattacharya, R. Cao, J. Cheng, en, *Proteins: Struct. Funct. Bioinf.* **Aug. 2015**, *83*, 1436–1449.
- [143] B. Adhikari, J. Cheng, en, *BMC Bioinformatics* **Jan. 2018**, *19*, 22.
- [144] D. Rodríguez, M. Sammito, K. Meindl, I. M. de Ilarduya, M. Potratz, G. M. Sheldrick, I. Usón, en, *Acta Crystallogr. D Biol. Crystallogr.* **Apr. 2012**, *68*, 336–343.
- [145] 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, en, *Nat. Methods* **Nov. 2013**, *10*, 1099–1104.
- [146] D Frishman, P Argos, en, *Proteins* **Dec. 1995**, *23*, 566–579.
- [147] L. A. Abriata, G. E. Tamò, B. Monastyrskyy, A. Kryshtafovych, M. Dal Peraro, en, *Proteins* **Mar. 2018**, *86 Suppl 1*, 97–112.
- [148] S. Ovchinnikov, H. Park, D. E. Kim, F. Dimaio, D. Baker, en, *Proteins: Struct. Funct. Bioinf.* **Sept. 2017**, DOI 10.1002/prot.25390.
- [149] D. T. Jones, en, *Proteins: Structure Function and Genetics* **2001**, *Suppl 5*, 127–132.
- [150] J. J. Ellis, F. P. E. Huard, C. M. Deane, S. Srivastava, G. R. Wood, en, *BMC Bioinformatics* **Apr. 2010**, *11*, 172.
- [151] S. Wang, W. Li, R. Zhang, S. Liu, J. Xu, en, *Nucleic Acids Res.* **July 2016**, *44*, W361–6.
- [152] D. T. Jones, en, *J. Mol. Biol.* **Sept. 1999**, *292*, 195–202.
- [153] L. Kaján, T. A. Hopf, M. Kalaš, D. S. Marks, B. Rost, en, *BMC Bioinformatics* **Mar. 2014**, *15*, 85.
- [154] A. T. Brünger, P. D. Adams, G. M. Clore, W. L. DeLano, P. Gros, R. W. Grossesse-Kunstleve, J. S. Jiang, J. Kuszewski, M. Nilges, N. S. Pannu, R. J. Read, L. M. Rice, T Simonson, G. L. Warren, en, *Acta Crystallogr. D Biol. Crystallogr.* **Sept. 1998**, *54*, 905–921.
- [155] J. M. H. Thomas, PhD thesis, University of Liverpool, **Jan. 2017**.

- [156] H. T. Jenkins, *Acta Crystallographica Section D Structural Biology* **Mar. 2018**, *74*, 205–214.
- [157] G. Scapin, en, *Acta Crystallogr. D Biol. Crystallogr.* **Nov. 2013**, *69*, 2266–2275.
- [158] R. D. Oeffner, P. V. Afonine, C. Millán, M. Sammito, I. Usón, R. J. Read, A. J. McCoy, *Acta Crystallogr. D Biol. Crystallogr.* **Apr. 2018**, *74*, 245–255.
- [159] J. Schaarschmidt, B. Monastyrskyy, A. Kryshtafovych, A. M. J. J. Bonvin, en, *Proteins Mar. 2018, 86 Suppl 1*, 51–66.
- [160] 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, en, *Nature Apr. 2018*, *556*, 118–121.
- [161] J. Yang, R. Jang, Y. Zhang, H. B. Shen, en, *Bioinformatics Oct. 2013*, *29*, 2579–2587.
- [162] M. I. Sadowski, en, *Proteins: Struct. Funct. Bioinf. Feb. 2013*, *81*, 253–260.
- [163] C. Millán, M. Sammito, I. Usón, en, *IUCrJ Jan. 2015*, *2*, 95–105.
- [164] A. J. McCoy, R. D. Oeffner, A. G. Wrobel, J. R. M. Ojala, K. Tryggvason, B. Lohkamp, R. J. Read, en, *Proceedings of the National Academy of Sciences Apr. 2017*, *114*, 3637–3641.
- [165] N. Fernandez-Fuentes, J. M. Dybas, A. Fiser, en, *PLoS Comput. Biol. Apr. 2010*, *6*, e1000750.
- [166] F. Delaglio, G. Kontaxis, A. Bax, *J. Am. Chem. Soc. Mar. 2000*, *122*, 2142–2143.
- [167] G. Kontaxis, F. Delaglio, A. Bax, en, *Methods Enzymol. 2005*, *394*, 42–78.
- [168] T. A. Jones, S Thirup, en, *EMBO J. Apr. 1986*, *5*, 819–822.
- [169] R. Heffernan, K. Paliwal, J. Lyons, A. Dehzangi, A. Sharma, J. Wang, A. Sattar, Y. Yang, Y. Zhou, en, *Sci. Rep. June 2015*, *5*, 11476.
- [170] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, D. J. Lipman, en, *J. Mol. Biol. Oct. 1990*, *215*, 403–410.
- [171] C. Camacho, G. Coulouris, V. Avagyan, N. Ma, J. Papadopoulos, K. Bealer, T. L. Madden, en, *BMC Bioinformatics Dec. 2009*, *10*, 421.
- [172] J. Söding, A. Biegert, A. N. Lupas, en, *Nucleic Acids Res. July 2005*, *33*, W244–8.
- [173] A. Biegert, C. Mayer, M. Remmert, J. Söding, A. N. Lupas, en, *Nucleic Acids Res. July 2006*, *34*, W335–9.
- [174] W. L. DeLano, The PyMOL Molecular Graphics System, <http://www.pymol.org>, **Nov. 2002**.
- [175] R. Shrestha, K. Y. J. Zhang, *Acta Crystallogr. D Biol. Crystallogr.* **2015**, *71*, 304–312.
- [176] D. W. A. Buchan, D. T. Jones, *Bioinformatics Sept. 2017*, *33*, 2684–2690.
- [177] J. Vojtěchovský, K. Chu, J. Berendzen, R. M. Sweet, I. Schlichting, en, *Biophys. J. Oct. 1999*, *77*, 2153–2174.
- [178] H. Eklund, M. Ingelman, B. O. Söderberg, T. Uhlin, P. Nordlund, M. Nikkola, U. Sonnerstam, T. Joelson, K. Petratos, en, *J. Mol. Biol. Nov. 1992*, *228*, 596–618.
- [179] F. K. Athappilly, W. A. Hendrickson, en, *Structure Dec. 1995*, *3*, 1407–1419.
- [180] S. Bañuelos, M. Saraste, K. D. Carugo, en, *Structure Nov. 1998*, *6*, 1419–1431.

- [181] A. H. West, E. Martinez-Hackert, A. M. Stock, en, *J. Mol. Biol.* **July 1995**, *250*, 276–290.
- [182] J. Ménétrey, E. Macia, S. Pasqualato, M. Franco, J. Cherfils, en, *Nat. Struct. Biol.* **June 2000**, *7*, 466–469.
- [183] C. C. Thomas, S Dowler, M Deak, D. R. Alessi, D. M. van Aalten, en, *Biochem. J. Sept.* **2001**, *358*, 287–294.
- [184] S. Grizot, F. Fieschi, M. C. Dagher, E. Pebay-Peyroula, en, *J. Biol. Chem.* **June 2001**, *276*, 21627–21631.
- [185] P. Sörme, P. Arnoux, B. Kahl-Knutsson, H. Leffler, J. M. Rini, U. J. Nilsson, en, *J. Am. Chem. Soc.* **Feb. 2005**, *127*, 1737–1743.
- [186] C. A. Kim, M. Gingery, R. M. Pilpa, J. U. Bowie, en, *Nat. Struct. Biol.* **June 2002**, *9*, 453–457.
- [187] J. B. Thoden, H. M. Holden, Z. Zhuang, D. Dunaway-Mariano, en, *J. Biol. Chem.* **July 2002**, *277*, 27468–27476.
- [188] 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, en, *Immunity Mar. 2004*, *20*, 337–347.
- [189] B. A. Fields, H. H. Bartsch, H. D. Bartunik, F. Cordes, J. M. Guss, H. C. Freeman, en, *Acta Crystallogr. D Biol. Crystallogr.* **Sept. 1994**, *50*, 709–730.
- [190] Y. Cheng, S. M. Sequeira, L. Malinina, V. Tereshko, T. H. Söllner, D. J. Patel, en, *Protein Sci. Oct. 2004*, *13*, 2665–2672.
- [191] M. Graille, C. Z. Zhou, V. Receveur-Bréchot, B. Collinet, N. Declerck, H. Van Tilbeurgh, en, *J. Biol. Chem.* **Apr. 2005**, *280*, 14780–14789.
- [192] T. Merz, S. K. Wetzel, S. Firbank, A. Plückthun, M. G. Grütter, P. R. E. Mittl, en, *J. Mol. Biol.* **Feb. 2008**, *376*, 232–240.
- [193] D. A. K. Traore, A. J. Brennan, R. H. P. Law, C. Dogovski, M. A. Perugini, N. Lukyanova, 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, en, *Biochem. J. Dec. 2013*, *456*, 323–335.
- [194] 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, en, *Med. Chem. Commun.* **Feb. 2014**, *5*, 342–351.
- [195] B. T. Porebski, A. A. Nickson, D. E. Hoke, M. R. Hunter, L. Zhu, S. McGowan, G. I. Webb, A. M. Buckle, en, *Protein Eng. Des. Sel. Mar. 2015*, *28*, 67–78.
- [196] A. M. Crowe, P. J. Stogios, I. Casabon, E. Evdokimova, A. Savchenko, L. D. Eltis, en, *J. Biol. Chem.* **Jan. 2015**, *290*, 872–882.
- [197] B. P. Klaholz, A. Mitschler, D. Moras, en, *J. Mol. Biol.* **Sept. 2000**, *302*, 155–170.
- [198] W. T. Lowther, N Brot, H Weissbach, B. W. Matthews, en, *Biochemistry Nov. 2000*, *39*, 13307–13312.
- [199] 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, en, *J. Biol. Chem.* **Nov. 2001**, *276*, 44044–44051.
- [200] P. Kuser, D. R. Hall, L. H. Mei, M. Neu, R. W. Evans, P. F. Lindley, en, *Acta Crystallogr. D Biol. Crystallogr.* **May 2002**, *58*, 777–783.

- [201] I. Hayashi, K. Vuori, R. C. Liddington, en, *Nat. Struct. Biol.* **2002**, *9*, 101–106.
- [202] G. David, K. Blondeau, M. Schiltz, S. Penel, A. Lewit-Bentley, en, *J. Biol. Chem.* **Oct. 2003**, *278*, 43728–43735.
- [203] V. Oganesyan, D. Busso, J. BrandSEN, S. Chen, J. Jancarik, R. Kim, S. H. Kim, en, *Acta Crystallographica - Section D Biological Crystallography* **July 2003**, *59*, 1219–1223.
- [204] J. Liu, H. Yokota, R. Kim, S. H. Kim, en, *Proteins: Structure Function and Genetics* **June 2004**, *55*, 1082–1086.
- [205] L. Cendron, A. Seydel, A. Angelini, R. Battistutta, G. Zanotti, en, *J. Mol. Biol.* **July 2004**, *340*, 881–889.
- [206] L. Malinina, M. L. Malakhova, A. T. Kanack, M. Lu, R. Abagyan, R. E. Brown, D. J. Patel, en, *PLoS Biol.* **Nov. 2006**, *4*, 1996–2011.
- [207] K. Makabe, S. Yan, V. Tereshko, G. Gawlak, S. Koide, en, *J. Am. Chem. Soc.* **Nov. 2007**, *129*, 14661–14669.
- [208] A. W. Maresso, R. Wu, J. W. Kern, R. Zhang, D. Janik, D. M. Missiakas, M. E. Duban, A. Joachimiak, O. Schneewind, en, *J. Biol. Chem.* **Aug. 2007**, *282*, 23129–23139.
- [209] C. U. Stirnimann, D. Ptchelkine, C. Grimm, C. W. Müller, en, *J. Mol. Biol.* **July 2010**, *400*, 71–81.
- [210] L. Von Schantz, M. Håkansson, D. T. Logan, B. Walse, J. Österlin, E. Nordberg-Karlsson, M. Ohlin, M. Hkansson, D. T. Logan, B. Walse, J. Österlin, E. Nordberg-Karlsson, M. Ohlin, en, *Glycobiology* **July 2012**, *22*, 948–961.
- [211] S. J. Coulthurst, A. Dawson, W. N. Hunter, F. Sargent, en, *Biochemistry* **Feb. 2012**, *51*, 1678–1686.
- [212] M. Myllykoski, A. Raasakka, M. Lehtimäki, H. Han, I. Kursula, P. Kursula, en, *J. Mol. Biol.* **Nov. 2013**, *425*, 4307–4322.
- [213] X. Yang, M. Morita, H. Wang, T. Suzuki, W. Yang, Y. Luo, C. Zhao, Y. Yu, M. Bartlam, T. Yamamoto, Z. Rao, en, *Nucleic Acids Res.* **Dec. 2008**, *36*, 6872–6881.
- [214] J. C. Grigg, C. X. Mao, M. E. P. Murphy, en, *J. Mol. Biol.* **Oct. 2011**, *413*, 684–698.
- [215] H. Repo, J. S. Oeemig, J. Djupsjöbacka, H. Iwai, P. Heikinheimo, en, *Acta Crystallogr. D Biol. Crystallogr.* **Nov. 2012**, *68*, 1479–1487.
- [216] M. F. Matos, Y. Xu, I. Dulubova, Z. Otwinowski, J. M. Richardson, D. R. Tomchick, J. Rizo, A. Ho, en, *Proc. Natl. Acad. Sci. U. S. A.* **Mar. 2012**, *109*, 3802–3807.
- [217] S. Moréra, I. Grin, A. Vigouroux, S. Couvé, V. Henriot, M. Saparbaev, A. A. Ishchenko, en, *Nucleic Acids Res.* **Oct. 2012**, *40*, 9917–9926.
- [218] P. M. Collins, K. Bum-Erdene, X. Yu, H. Blanchard, en, *J. Mol. Biol.* **Apr. 2014**, *426*, 1439–1451.
- [219] 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, en, *Nat. Methods* **Feb. 2015**, *12*, 131–133.
- [220] K. Edman, A. Royant, P. Nollert, C. A. Maxwell, E. Pebay-Peyroula, J. Navarro, R. Neutze, E. M. Landau, en, *Structure* **Apr. 2002**, *10*, 473–482.

- [221] J. Standfuss, A. C. T. Van Scheltinga, M. Lamborghini, W. Kühlbrandt, en, *EMBO J.* **Mar. 2005**, *24*, 919–928.
- [222] J. K. Lee, D. Kozono, J. Remis, Y. Kitagawa, P. Agre, R. M. Stroud, en, *Proc. Natl. Acad. Sci. U. S. A.* **Dec. 2005**, *102*, 18932–18937.
- [223] D. F. Savage, R. M. Stroud, en, *J. Mol. Biol.* **May 2007**, *368*, 607–617.
- [224] D. Pogoryelov, Ö. Yildiz, J. D. Faraldo-Gómez, T. Meier, en, *Nat. Struct. Mol. Biol.* **Oct. 2009**, *16*, 1068–1073.
- [225] K. R. Vinothkumar, K. Strisovsky, A. Andreeva, Y. Christova, S. Verhelst, M. Freeman, en, *EMBO J.* **Nov. 2010**, *29*, 3797–3809.
- [226] J. D. Ho, R. Yeh, A. Sandstrom, I. Chorny, W. E. C. Harries, R. A. Robbins, L. J. W. Miercke, R. M. Stroud, en, *Proceedings of the National Academy of Sciences May 2009*, *106*, 7437–7442.
- [227] N. H. Joh, A. Oberai, D. Yang, J. P. Whitelegge, J. U. Bowie, en, *J. Am. Chem. Soc.* **Aug. 2009**, *131*, 10846–10847.
- [228] S. Ye, Y. Li, Y. Jiang, en, *Nat. Struct. Mol. Biol.* **Aug. 2010**, *17*, 1019–1023.
- [229] M. G. Derebe, D. B. Sauer, W. Zeng, A. Alam, N. Shi, Y. Jiang, en, *Proceedings of the National Academy of Sciences Jan. 2011*, *108*, 598–602.
- [230] H. Saino, Y. Ukita, H. Ago, D. Irikura, A. Nisawa, G. Ueno, M. Yamamoto, Y. Kanaoka, B. K. Lam, K. F. Austen, M. Miyano, en, *J. Biol. Chem.* **May 2011**, *286*, 16392–16401.
- [231] G. B. Erkens, R. P. A. Berntsson, F. Fulyani, M. Majsnerowska, A. Vujičić-Žagar, J. Ter Beek, B. Poolman, D. J. Slotboom, en, *Nat. Struct. Mol. Biol.* **June 2011**, *18*, 755–760.
- [232] J. Symersky, V. Pagadala, D. Osowski, A. Krah, T. Meier, J. D. Faraldo-Gómez, D. M. Mueller, en, *Nat. Struct. Mol. Biol.* **Apr. 2012**, *19*, 485–91, S1.
- [233] R. P.-A. Berntsson, J. ter Beek, M. Majsnerowska, R. H. Duurkens, P. Puri, B. Poolman, D.-J. Slotboom, en, *Proceedings of the National Academy of Sciences Aug. 2012*, *109*, 13990–13995.