Contents

List of Figures

List of Tables

List of Equations

List of Abbreviations

ACL	Average Chain Length
CC	Correlation Coefficient
CNS	Crystallography & NMR System
KDE	Kernel Density Estimate
LLG	Log-Likelihood Gain
MR	Molecular Replacement
MSA	Multiple Sequence Alignment
PDB	Protein Data Bank
RMSD	Root-Mean-Square Deviation
TM-score	Template-Modelling score

Chapter 1

Introduction

Chapter 2

Materials & Methods

Chapter 3

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

Chapter 4

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

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 [1–4]. 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 [5, 6]. However, Keegan et al. [1] 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 [1]. 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 suitable alternative for AMPLE users.

Whilst ROSETTA and QUARK are amongst the best *ab initio* structure prediction algorithms currently available [5], other algorithms have been developed over the last two decades [e.g., 7–12]. Although most of these algorithms utelise fragment-assembly algorithms similar to ROSETTA and QUARK, their procedure to fragment selection or assembly is substantially different [7, 8]. 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 entirely on this information are an interesting alternative to fragment-based approaches [9, 11, 12].

CONFOLD2 [13], a distance-geometry based algorithm, utelises 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 pre-defined selection. CONFOLD2 generates slightly inferior decoys compared to ROSETTA, however outperforms it in speed and simplicity of installation [13, 14].

FRAGFOLD [7], 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 library is shipped with FRAGFOLD and target-specific fragments extracted based on

secondary structure and a sequence-based threading score, it enables fast and easy fragment library generation compared to ROSETTA [15].

SAINT2 [16], 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 [16]. Furthermore, SAINT2 uses FLIB [17] for fragment picking, an algorithm shown to outperform ROSETTA's equivalent NNMake [18] 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 Molecular Replacement (MR). The *ab initio* structure prediction protocols CONFOLD2 [13], FRAGFOLD [7] and SAINT2 [16], were explored given their substantially different approaches to AMPLE's current default ROSETTA [19].

5.2 Materials & Methods

5.2.1 Target selection

This study was conducted using all 27 targets from the PREDICTORS dataset (**??**).

5.2.2 Contact prediction

Residue-residue contact information was predicted for 18 out of 27 targets using METAPSICOV v1.04 [20].

Secondary structure and solvent exposure were predicted using PSIPRED v4.0 [21] and SOLVPRED (shipped with METAPSICOV v1.04), respectively. The Multiple Sequence Alignment (MSA) for coevolution-based contact prediction was generated using HHBLITS [22] against the `uniprot20_2016_02` database. CCMPRED v0.3.2 [23], FREECONTACT v1.0.21 [24] and PSICOV v2.1b3 [25] were used by METAPSICOV to generate contact predictions.

METAPSICOV STAGE 1 contact predictions were used in *ab initio* structure prediction since those result in more accurate structure predictions compared to METAPSICOV STAGE 2 ones [20].

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 [19] 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. [26].

The CONFOLD2 decoys were generated using CONFOLD2 v2.0 [13],which uses Crystallography & NMR System (CNS) v1.3 [27] 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`. This resulted in 40 seperated runs differing only in the number of contact pairs used, which was increased by 0.1 from $L/10$ to $4L$.

The FRAGFOLD decoys were generated using FRAGFOLD v4.80 [7] with default options. Homologous fragments were removed from the shipped library by excluding all entries with Protein Data Bank (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 [17], 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. [16].

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 [4] was shown to be advantageous in most cases by Thomas [28], and thus trialled in this context. Each MR run was assessed using the SHELXE criteria, where a minimum CC of 25.0 and Average Chain Length (ACL) of 10 was required (**??**). R-values of < 0.45 after model building were not part of the success criteria in this study.

5.3 Results

Tehe 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 [19], CONFOLD2 [13], FRAGFOLD [7] and SAINT2 [16]. 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 [20] 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 [20]. 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 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.

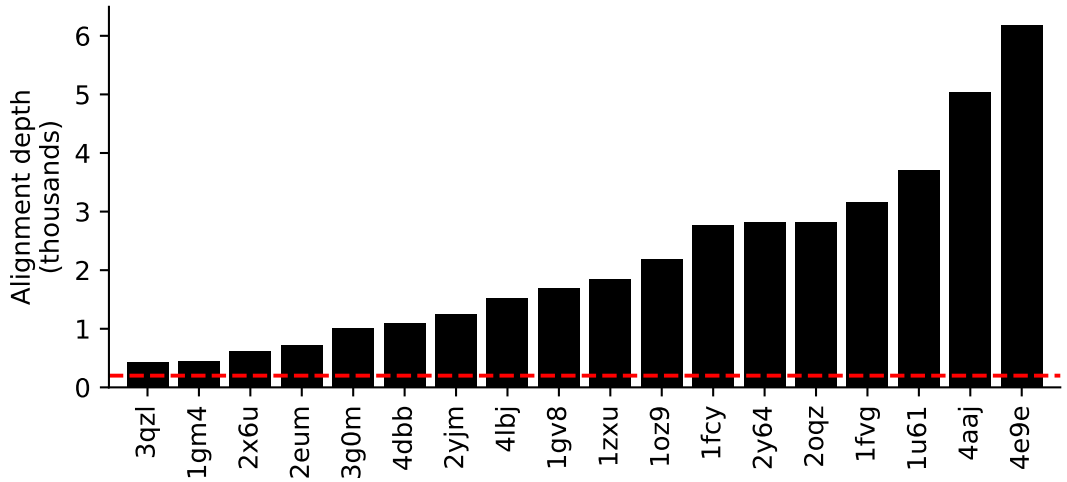

Figure 5.1: Distribution of HHBLITS alignment depth for subset of targets in the PREDICTORS dataset. Red line indicates alignment depth threshold of 200 sequences.

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 1000 sequences highlight 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 1000 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).

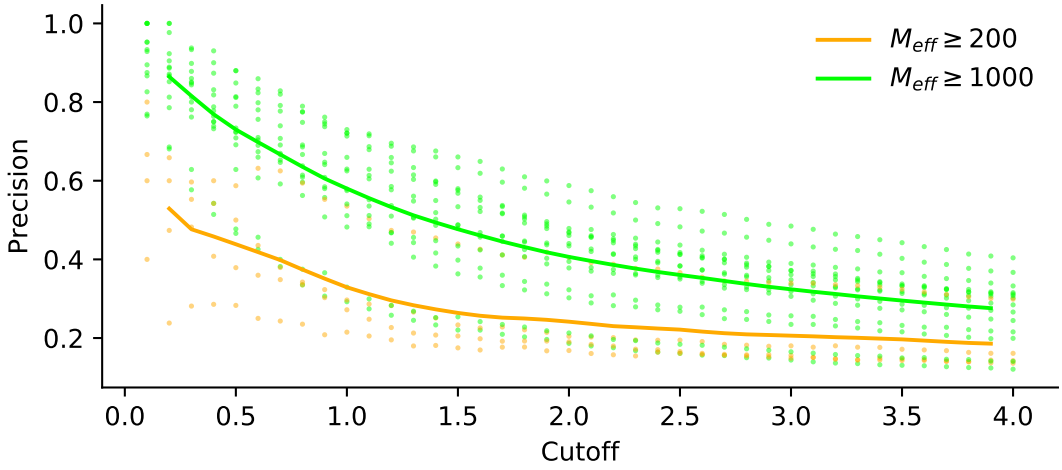

Figure 5.2: Contact precision analysis for numerous contact selection cutoffs for targets with alignment depths of more than 200 and more than 1000 sequences. Lines indicate moving averages for both categories with a window size of three residues. M_{eff} refers to the alignment depth (number of effective 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.

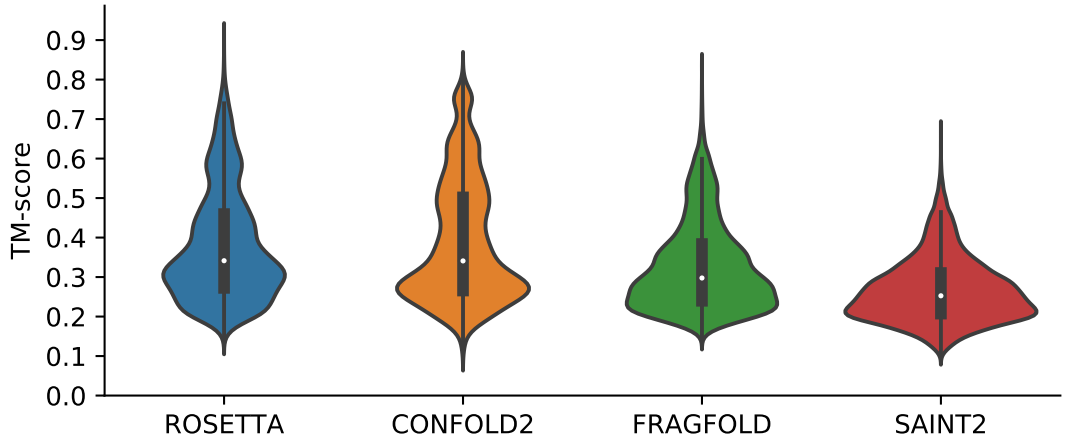

Figure 5.3: Kernel Density Estimate (KDE) of decoy Template-Modelling score (TM-score) for four different *ab initio* structure prediction algorithms, namely ROSETTA, CONFOLD2, FRAGFOLD and SAINT2. CONFOLD2 contains 9,000 less decoys than the remaining algorithms (for further details refer to Section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

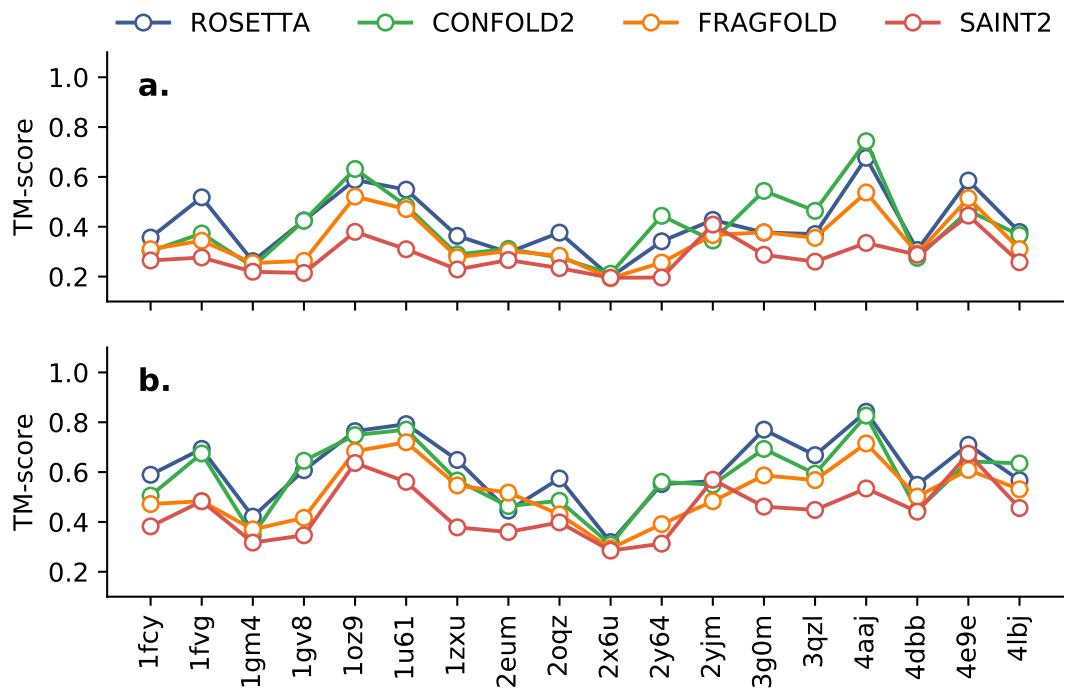

Figure 5.4: Per-target TM-score analysis for targets modelled with contact information and four separate *ab initio* structure prediction algorithms. Analysis is subdivided by (a) median TM-score of all decoys in each set and (b) TM-score of the top-1 decoy in each set.

Abriata et al. [5] 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. [5] 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.

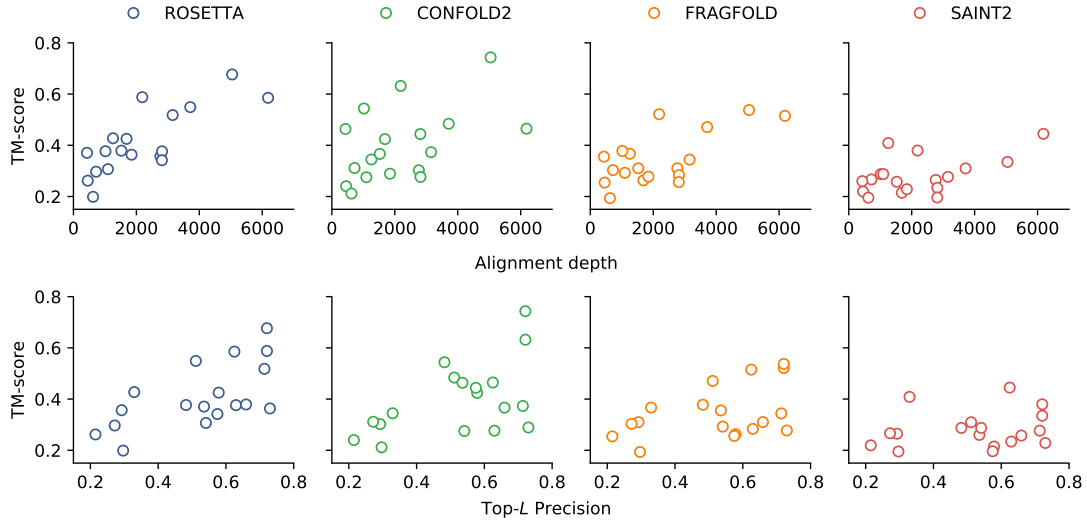

Figure 5.5: Analysis of median TM-score of the contact-based decoy sets and their dependence on alignment depth and top-L precision.

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).

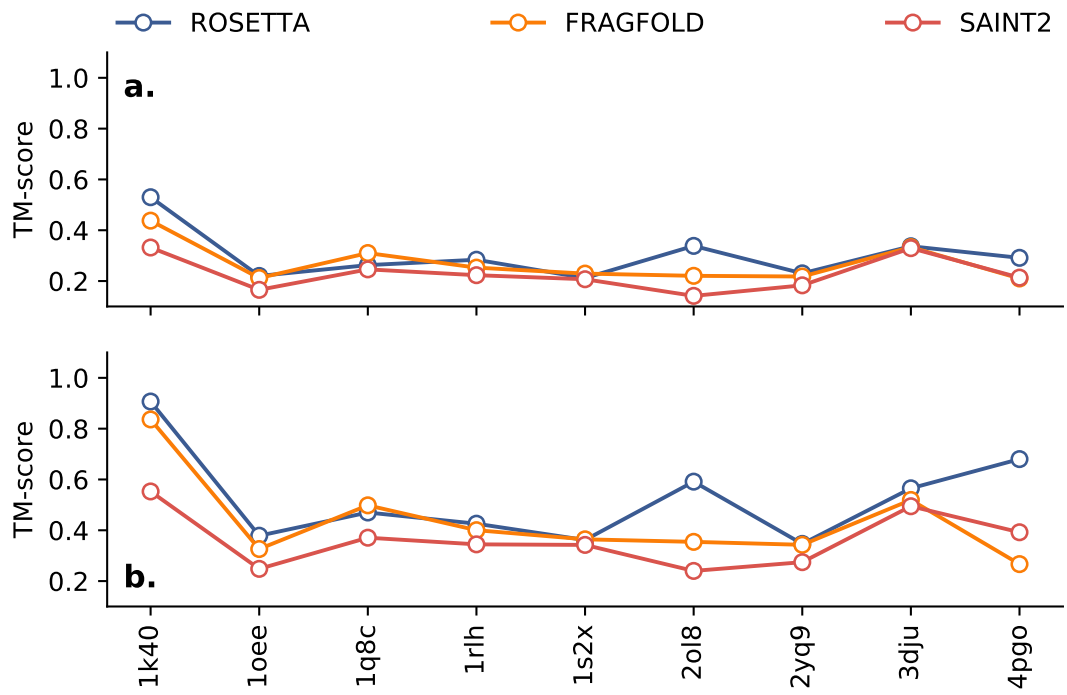

Figure 5.6: Per-target TM-score analysis for targets modelled without contact information and four separate *ab initio* structure prediction algorithms. Analysis is subdivided by (a) median TM-score of all decoys in each set and (b) TM-score of the top-1 decoy in each set.

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 roughly 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. Although the multimodal distribution of TM-score values for all-β target decoys might indicate better performance for some targets, it is most certainly misleading since three all-β targets are missing from the dataset in comparison to the other 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 **??**). 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_{res}=216). The set of CONFOLD2 decoys for N-(5-phosphoribosyl)anthranilate isomerase (PDB ID: 4aaj) 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.

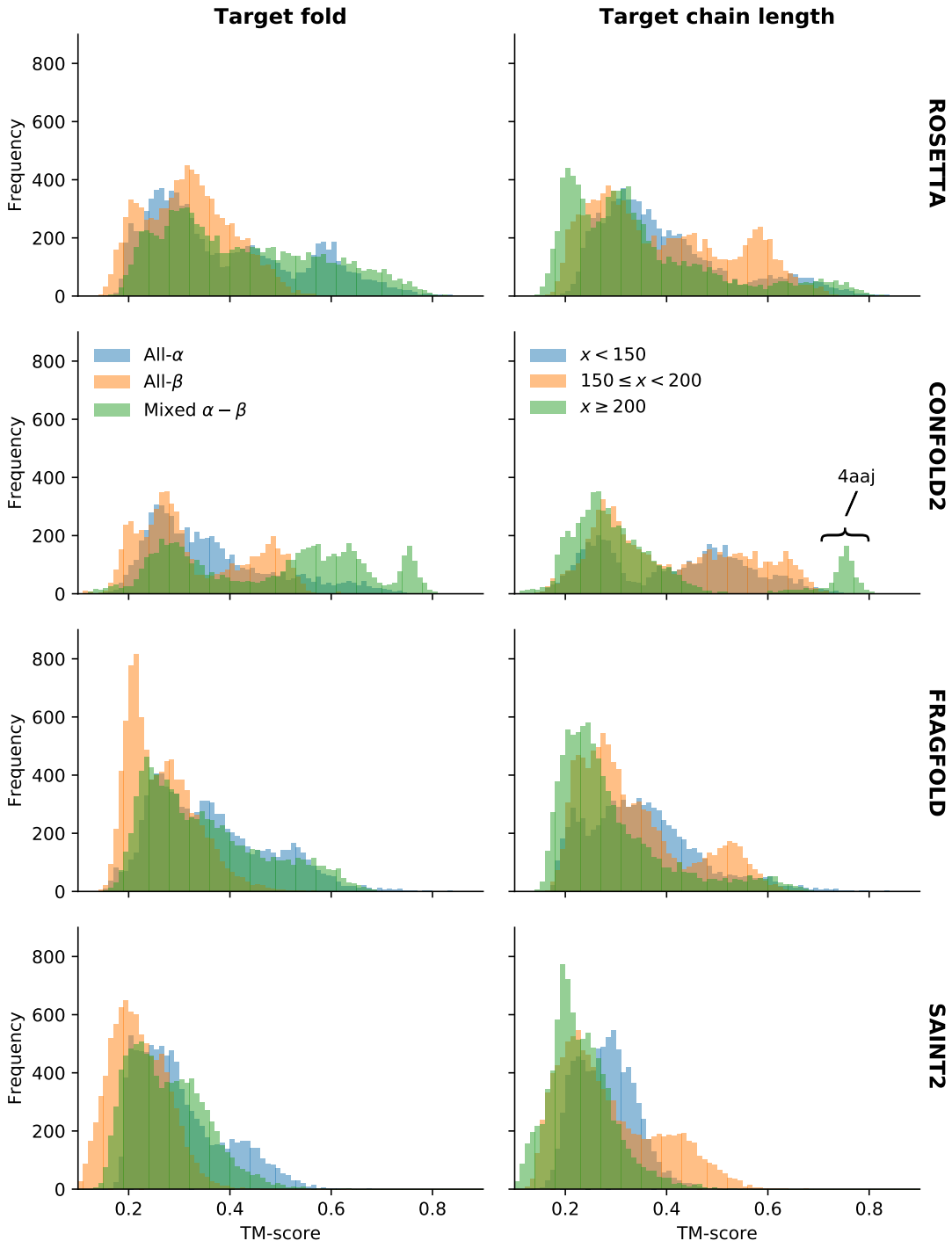

Figure 5.7: Distribution of decoy TM-scores by fold, chain length and algorithm.

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 to the AMPLE ensemble generation pipeline.

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 solu-

tions of five targets each, whilst FRAGFOLD decoys to four and CONFOLD2 decoys to 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.

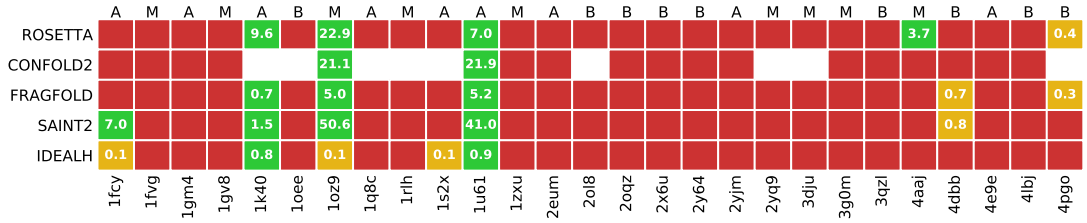

Figure 5.8: Summary of MR success with AMPLE ensemble search models. Search models are based on decoy sets generated with different *ab initio* structure prediction protocols. The colour coding indicates structure solution: no solution (red), one solution (orange), more than one solution (green). The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

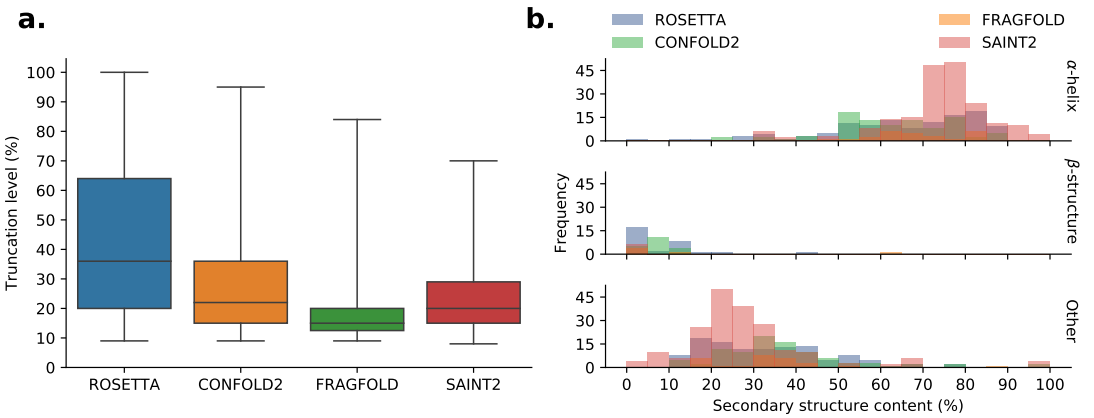

Figure 5.9: Distribution of (a) search model truncation and (b) secondary structure content for success AMPLE ensemble search models given decoys from four *ab initio* structure prediction algorithms. Secondary structure for each ensemble search model evaluated with DSSP [29].

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. Decoys from SAINT2 result in predominantly α-helical search models (Fig. 5.9). An analysis of the secondary structure makeup, as assigned by DSSP [29], shows that 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 and MRBUMP. 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 [3] 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, such as BORGES [30]. 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.

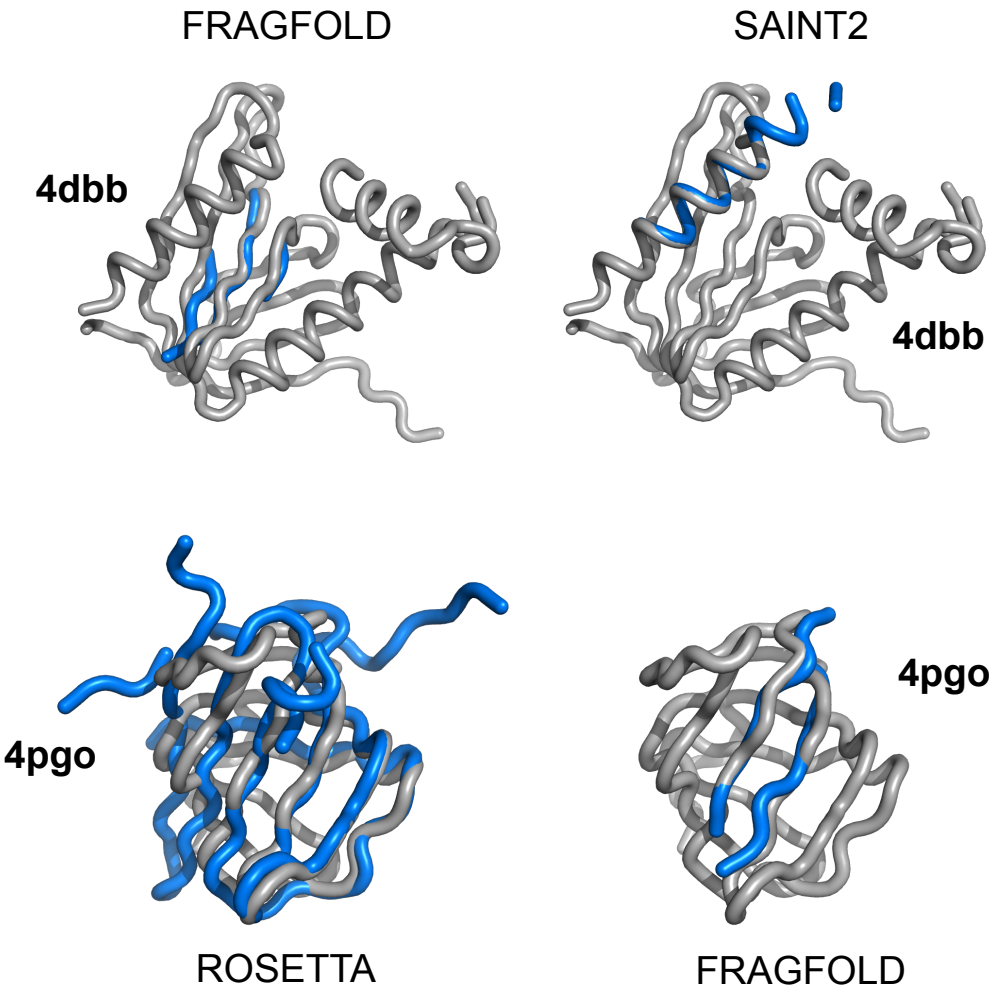

Figure 5.10: Examples of PHASER-placed AMPLE search models. AMPLE search models are coloured blue and deposited structures in gray. The PDB identifiers and modelling protocol given alongside each example.

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 Root-Mean-Square Deviation (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, data collected in this experiment may suggest different thresholds for *ab initio* modelling-derived search models (Fig. 5.11). Such differences need to be considered for some SAINT2 search models, which were prepared for the FAT domain of focal adhesion kinase (PDB ID: 1k40). These ensemble search models have RMSD values $>$ 10Å (up

to 28Å) yet result in PHASER LLG values in excess of the success threshold of 60 [31]. Furthermore, nearly 25% of all successful ensemble search models have RMSD values $\geq$ 2Å and PHASER LLG scores of $\geq$60. However, it is important to remember that RMSD values greatly differ based on the optimal superposition of the model and target, and whilst some scores might be inflated in this study compared to other superposition algorithms, this reflects the same superposition used for the TM-score calculation.

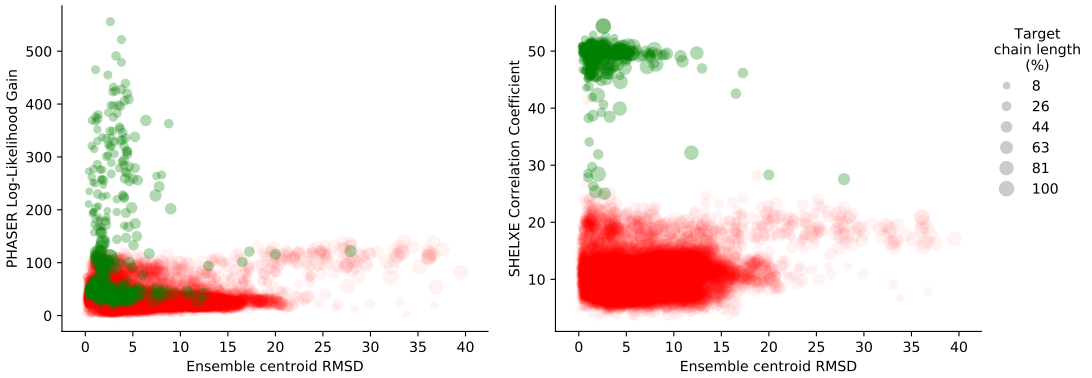

Figure 5.11: Relationship between ensemble quality, PHASER LLG and SHELXE CC.

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 of 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. Thus, poor stereochemistry might well give insights into the lack of success with highly accurate CONFOLD2 decoys.

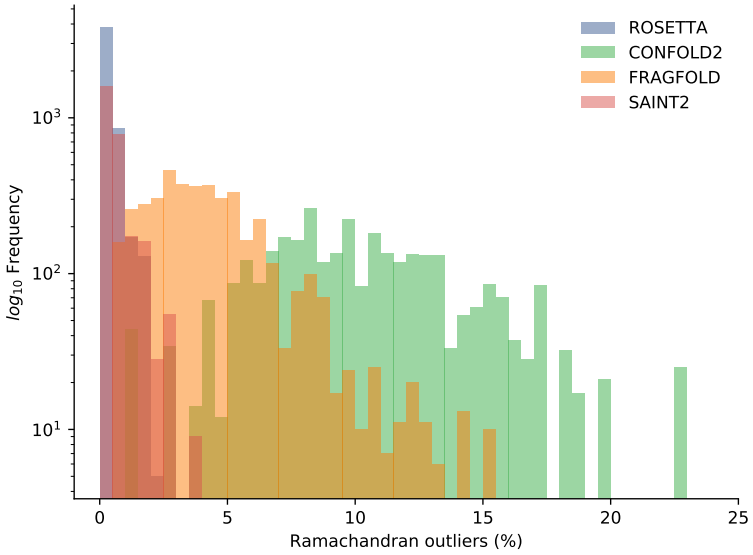

Figure 5.12: Distribution of Ramachandran outliers of AMPLE ensemble search model centroids based on decoys predicted with four *ab initio* structure prediction protocols. Outliers were calculated using PyRAMA (https://github.com/gerdos/PyRAMA).

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 globlar targets to evaluate the validity of either 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 [5, 6]. Furthermore, the findings describing the comparable performance of ROSETTA and CONFOLD2 [13, 14] 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.

In this study, alternative fragment-assembly based algorithms were tested, namely FRAGFOLD and SAINT2. 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 findinds are futher 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 larget 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 outline that although CONFOLD2 generates high quality decoys, this quality is lost and does not translate into a greater number of structure solutions. 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 ROSETTA decoys, it cannot be downplayed that ROSETTA generates the most accurate decoys overall. However, it is crucial to investigate whether CONFOLD2 decoys, potentially refined in the ROSETTA protocol, might provide a suitable alternative, especially given that fragment databases are not required.

Chapter 6

Decoy subselection using contact information to enhance MR search model creation

Chapter 7

Protein fragments as search models in Molecular Replacement

Chapter 8

Conclusion

Appendix A

Appendix

Bibliography

[1] 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.

[2] 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.

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

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

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

[6] S. Ovchinnikov, H. Park, D. E. Kim, F. Dimaio, D. Baker, en, *Proteins: Struct. Funct. Bioinf.* **Sept. 2017**, DOI `10.1002/prot.25390`.

[7] D. T. Jones, en, *Proteins: Structure Function and Genetics* **2001**, *Suppl 5*, 127–132.

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

[9] B. Adhikari, D. Bhattacharya, R. Cao, J. Cheng, en, *Proteins: Struct. Funct. Bioinf.* **Aug. 2015**, *83*, 1436–1449.

[10] D. Xu, Y. Zhang, en, *Proteins* **July 2012**, *80*, 1715–1735.

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

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

[13] B. Adhikari, J. Cheng, en, *BMC Bioinformatics* **Jan. 2018**, *19*, 22.

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

[15] T. Kosciolek, D. T. Jones, en, *PLoS One* **Mar. 2014**, *9*, e92197.

[16] S. H. P. de Oliveira, E. C. Law, J. Shi, C. M. Deane, en, *Bioinformatics* **Nov. 2017**, DOI `10.1093/bioinformatics/btx722`.

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

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

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

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

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

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

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

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

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

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

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

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

[29] D Frishman, P Argos, en, *Proteins* **Dec. 1995**, *23*, 566–579.

[30] 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.

[31] 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.