Contents

Chapter 1

Introduction

Chapter 2

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

Chapter 3

Approaches to ab initio molecular replacement of α-helical transmembrane proteins

Chapter 4

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

4.1 Introduction

The extended tractability of AMPLE 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 [4, 15, 3, 5, 2, 6, 18]. 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 impacts the folding process significantly (REFs). 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 (REFs). 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 Methods

4.2.1 Target selection

This study was conducted using 18 out of 27 targets from the KEENO dataset described in Chapter 2. The nine targets with effective sequence counts of less than 100 in the Pfam multiple sequence alignment were excluded.

4.2.2 Covariance-based contact prediction

Residue contacts for each target sequence were predicted using three different metapredictors, namely METAPSICOV [3], GREMLIN [4], and PCONSC2 [15]. Online servers for METAPSICOV (`http://bioinf.cs.ucl.ac.uk/METAPSICOV`) and GREMLIN (`http://gremlin.bakerlab.org`) were used to predict two sets of contact pairs. The choice of online servers over local installations was justified to directly imitate most AMPLE users. Both servers were used with default settings.

The GREMLIN web server returns the raw contact prediction files as well as preformatted ROSETTA distance restraints. The raw contact prediction files were downloaded to allow different contact selection thresholds as well as local conversion into

ROSETTA restraints files. The METAPSICOV web server returned two contact prediction files, one after 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. Additionally to the three main contact predictions outlined above, a set of BBCONTACTS restraints was obtained for protein targets containing β-strands. The approach was identical to that outlined in Chapter XYZ.

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

4.2.3 Contact pair to ROSETTA distance restraint formatting

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 [7, 15], the FADE function was set to acknowledge a formed restraint if the participating Cβ atoms (Cα in case of Gly) were within 9Å. In comparison, the SIGMOID function was defined with amino acid specific distances for Cβ atoms (Cα in case of Gly) to recognise the different sizes of each amino acid [4, 9].

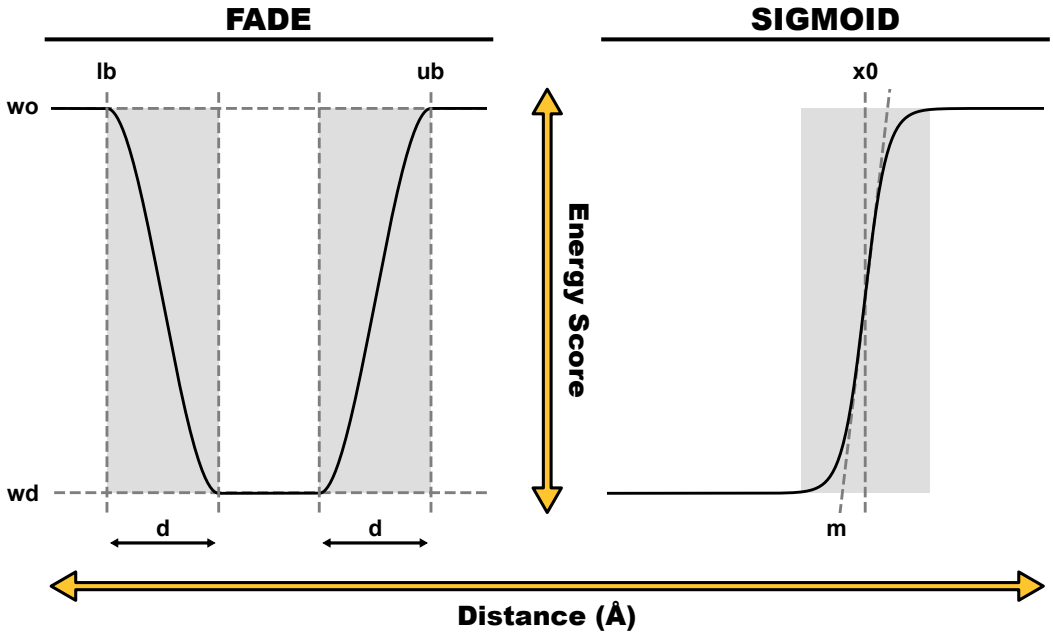


Figure 4.1: ROSETTA energy function comparison. Abbreviations corresponds to input parameters.

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 PCONSFOLD definition of the FADE function [7], the GREMLIN definition of the SIGMOID function [9], and additionally the PCONSC2 BBCONTACTS definition of the FADE function for β-structure containing targets (see Chapter XYZ).

The conversion was handled in AMPLE (see Chapter XYZ) 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 [7] and `3` for the SIGMOID function [9]. 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 Chapter XYZ. In comparison, the SCALARWEIGHTED term in the GREMLIN implementation of the SIGMOID energy function [9] was multiplied by the number of occurrences of each contact pair in the combined map.

Energy Function	**AMPLE keywords**

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

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

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

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

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

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 once per target and used throughout. In total, 1,000 ab initio decoys were generated per run using ROSETTAs default settings [12] 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 Molecular Replacement 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 successful structure solution assessed (see Chapter XYZ).

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 sequence alignments.

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).

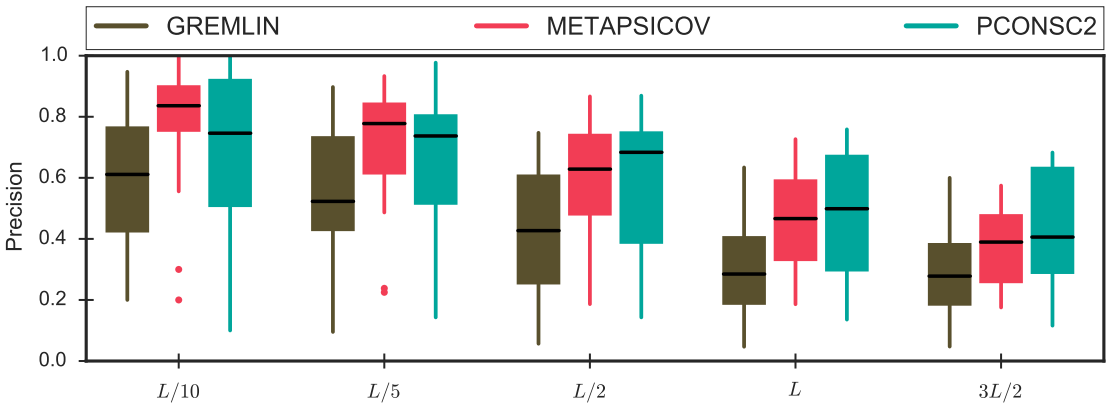


Figure 4.2: Precision spread for three metapredictors computed at five contact selection cutoff values relative to the target chain length (L).

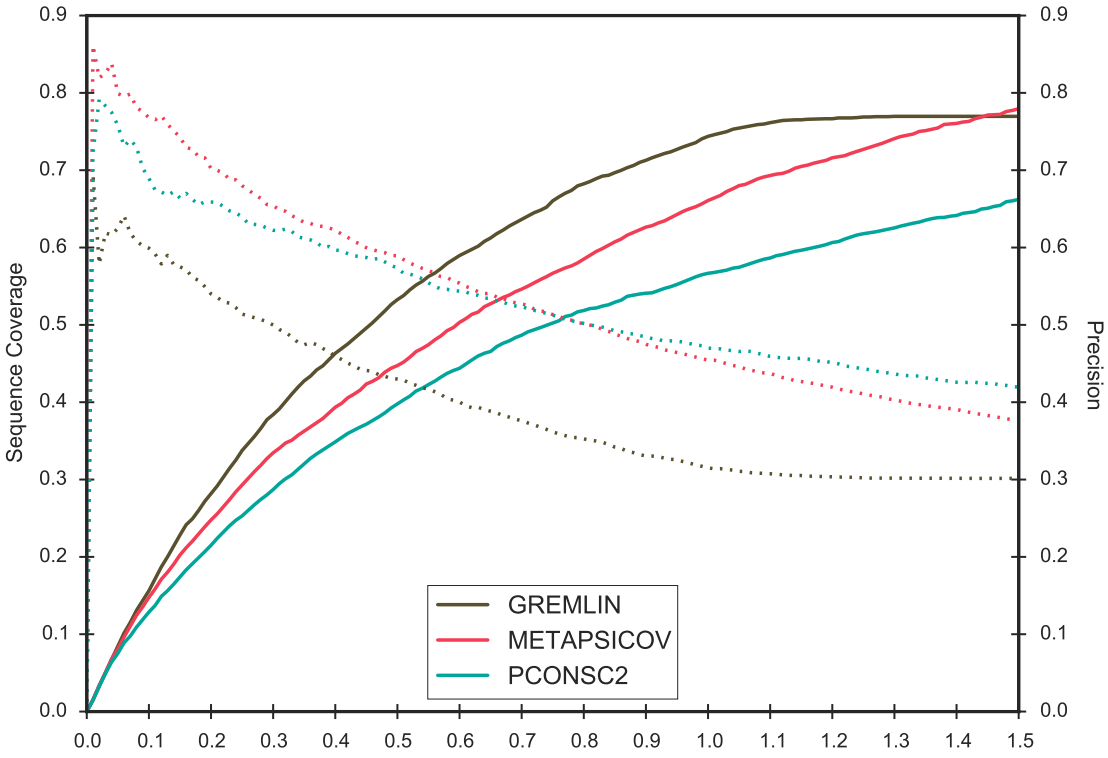


Figure 4.3: Average sequence coverage (line) and contact prediction precision scores (dashed) across a continuous range of contact selection cutoffs ranging from $[0.0, 1.5]$ for all targets.

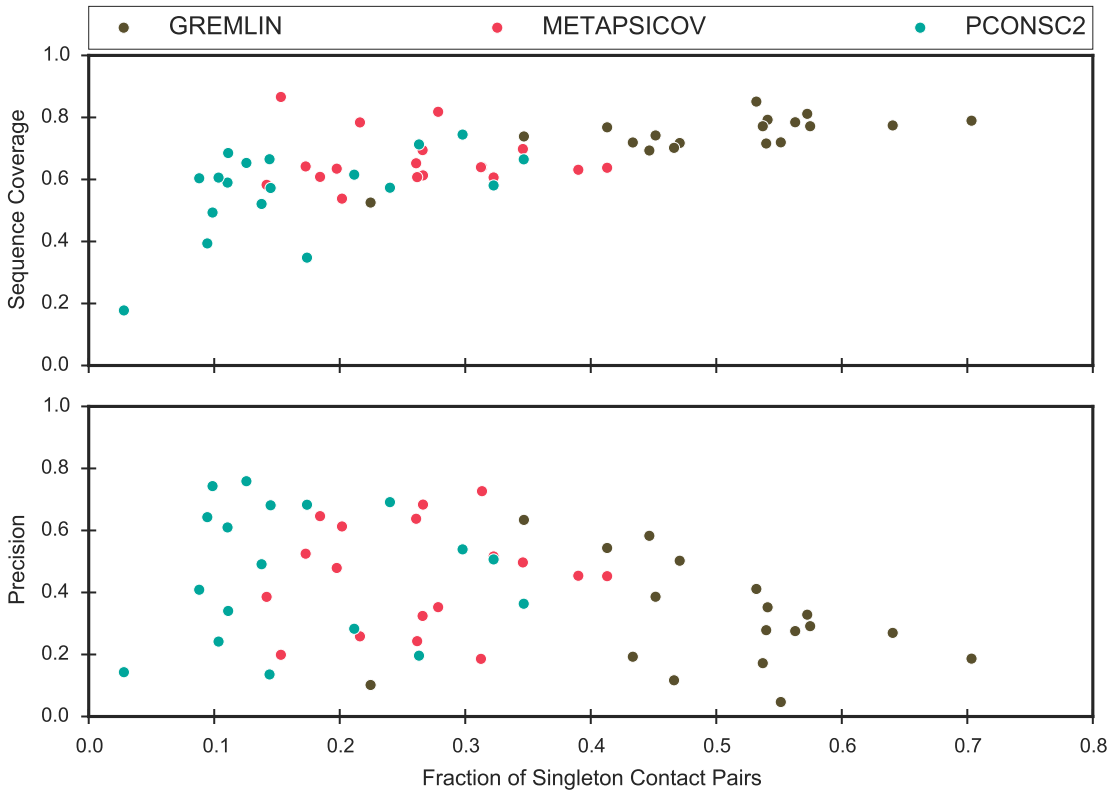


Figure 4.4: Contact singleton analysis compared against the precision of L contact pair lists for three metapredictors.

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 ($\Delta Precision_{METAPSICOV-GREMLIN} L/10 = -0.522$; $\Delta Precision_{METAPSICOV-PCONSC2} L/10 = -0.435$). 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

($\Delta Precision_{METAPSICOV-GREMLIN} L/10 = -0.231$; $\Delta Precision_{METAPSICOV-PCONSC2} L/10 = +0.077$).

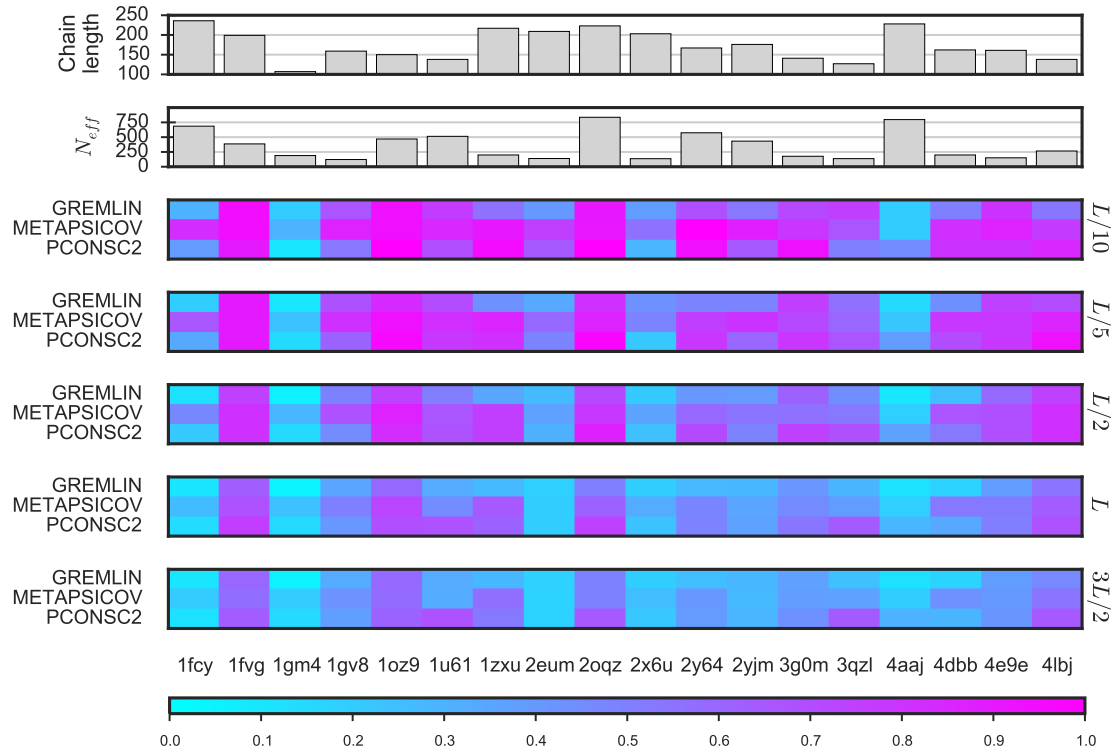


Figure 4.5: Contact prediction precision scores from three metapredictors for 18 targets at different contact pair selection thresholds. The Pfam alignment depth is given by means of number of effective sequences (*Neff*). The color scale corresponds to the precision in $[0, 1]$.

The data presented in Fig 4.5 also indicates that there is no direct link between chain length or Neff and the precision of the resulting contact predictions. The N-(5'-phosphoribosyl)anthranilate isomerase sequence (PDB: 4aaj) with a chain length of 228 residues and 750 effective sequences in its Pfam alignment 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.

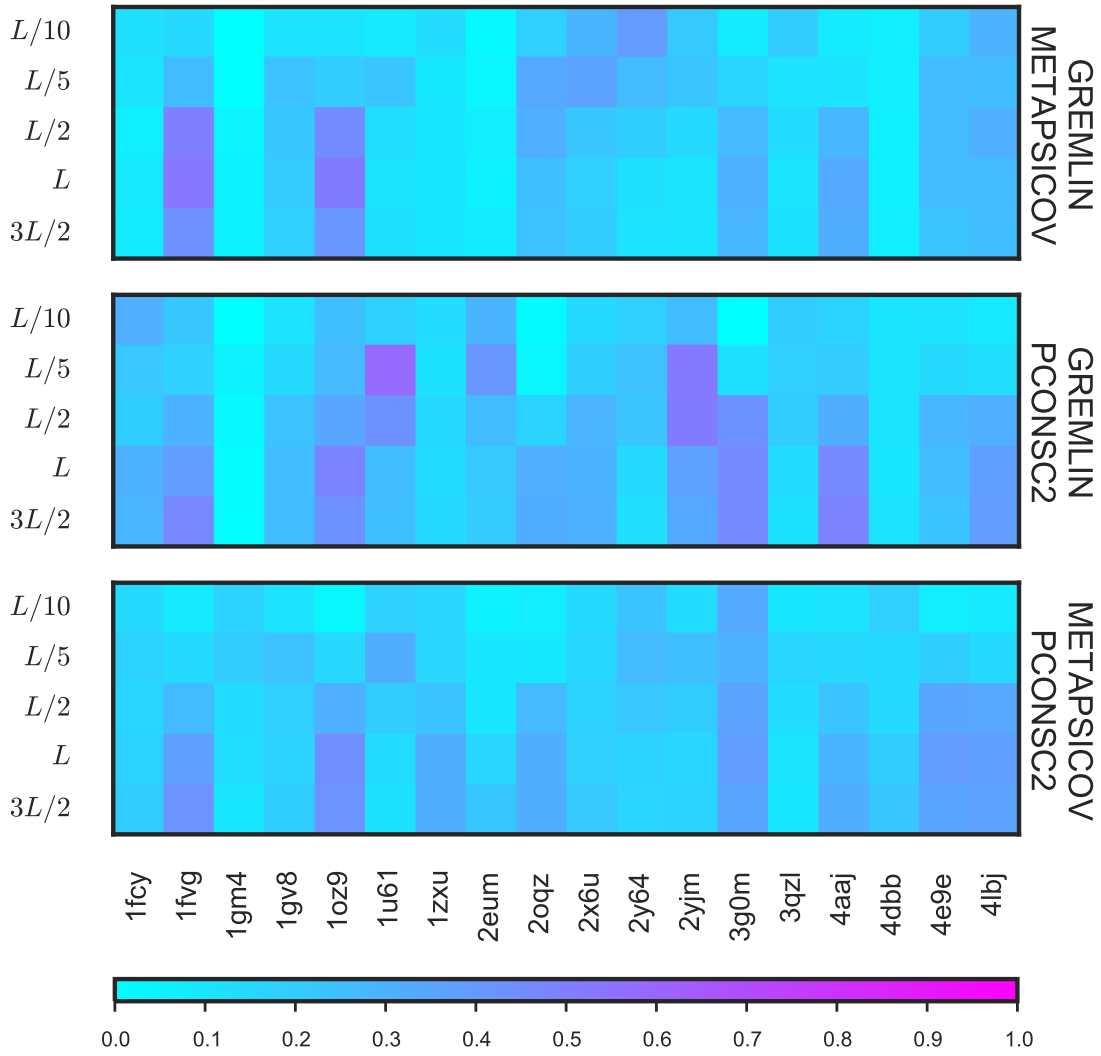


Figure 4.6: Jaccard similarity index illustrates a higher degree of overlap between metapredictor contact predictions with increasing numbers of contact pairs included in the calculation. The three panels show the different comparisons. The color scale corresponds to the Jaccard index in $[0, 1]$.

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 [14, 16]. 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 (excluding 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.

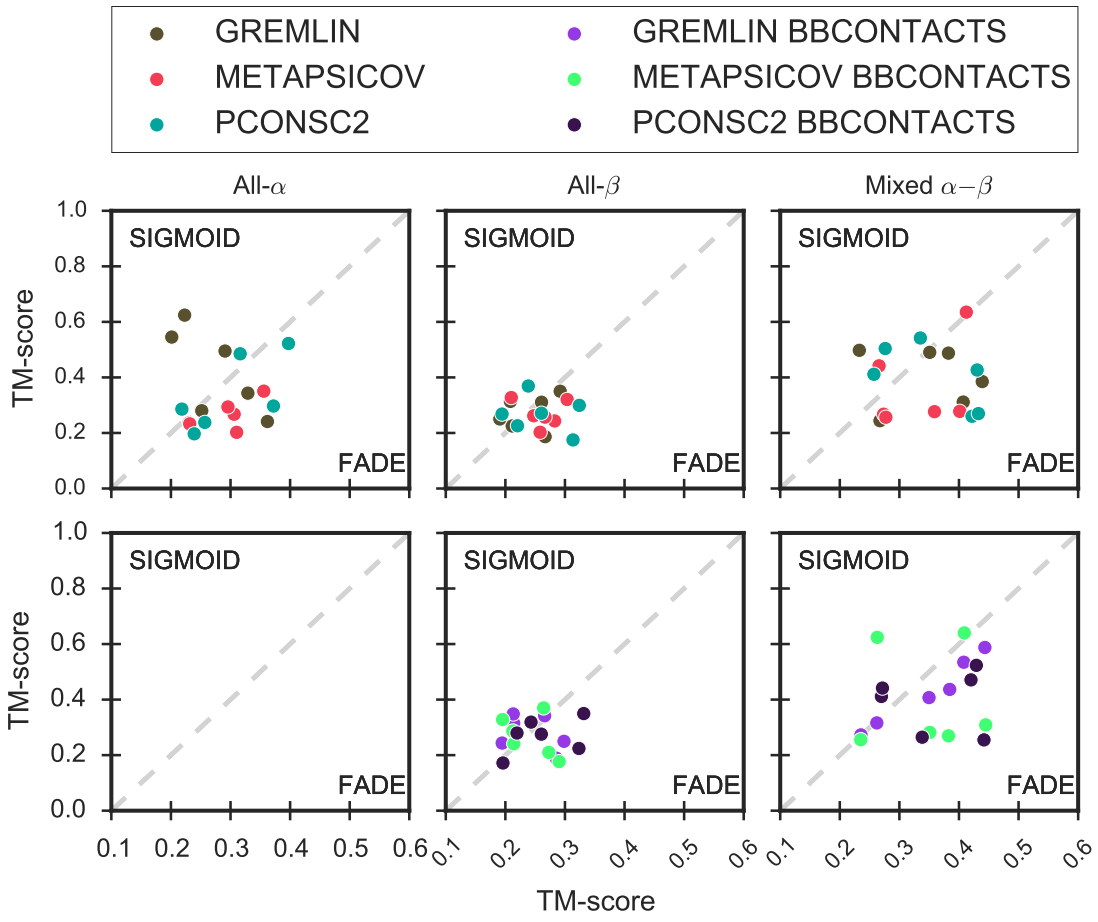


Figure 4.7: Median TM-score comparison of FADE and SIGMOID ROSETTA energy functions differentiated by fold and the addition of BBCONTACTS restraints.

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 identification 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 ($\Delta MedianTM - score_{ALL} : min = 0.002, max = 0.429$; $\Delta MedianTM - score_{TOP} : min = 0.002, max = 0.456$).

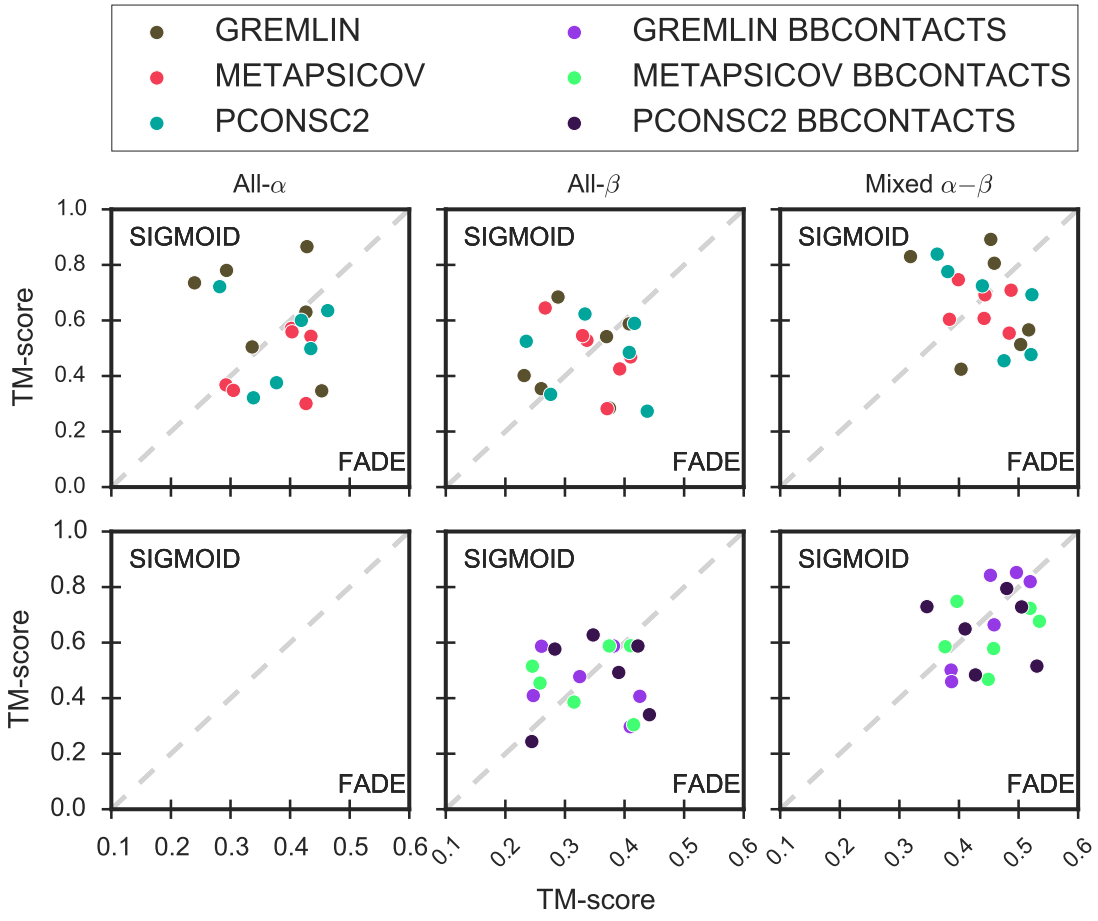


Figure 4.8: Top TM-score comparison of FADE and SIGMOID ROSETTA energy functions differentiated by fold and the addition of BBCONTACTS restraints.

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.

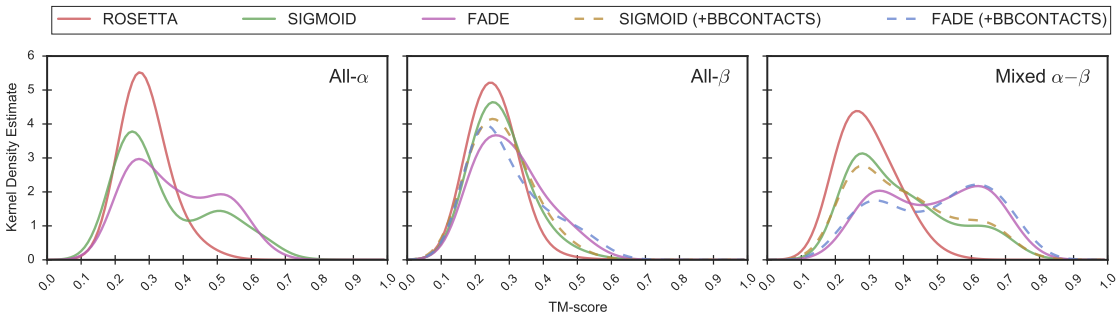
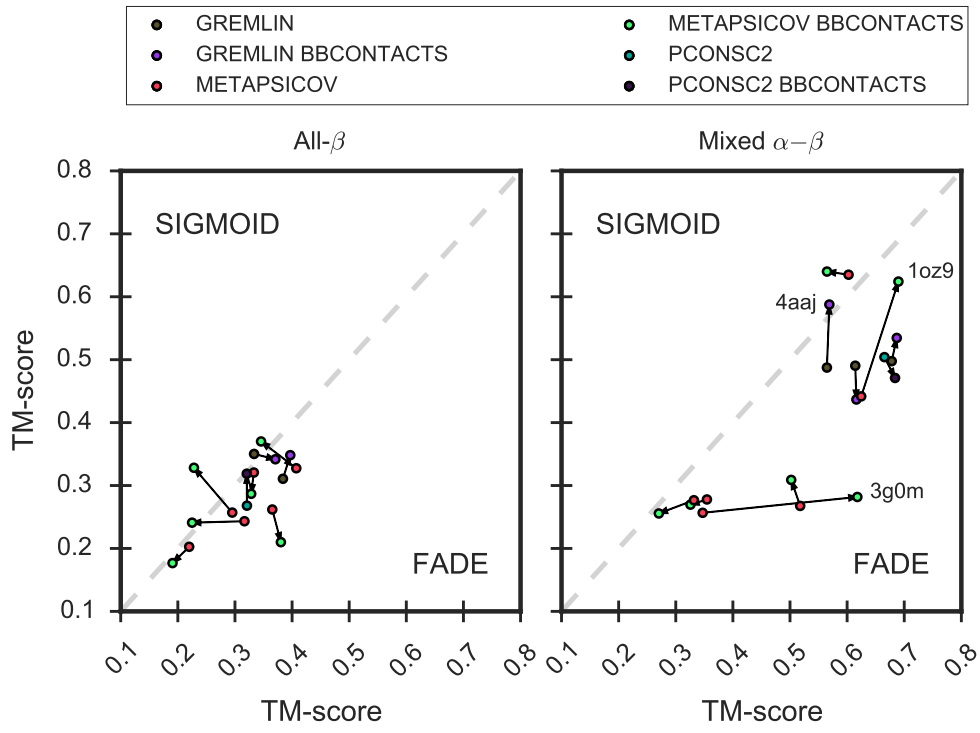
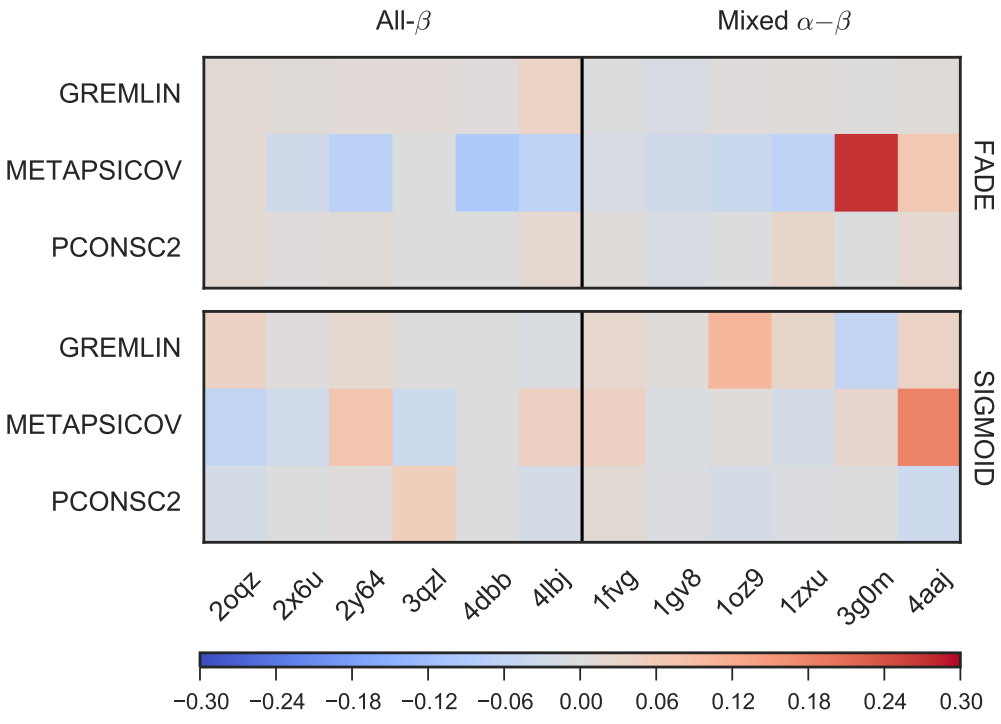


Figure 4.9: TM-score density estimate of all decoys in each respective fold class separating by ROSETTA energy function (SIGMOID or FADE) and no contact information used (ROSETTA). Dashed lines indicate decoys which were predicted with the addition of BBCONTACTS.

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 (see Chapter XYZ) 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).



(a)



(b)

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: 4aaj) 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.

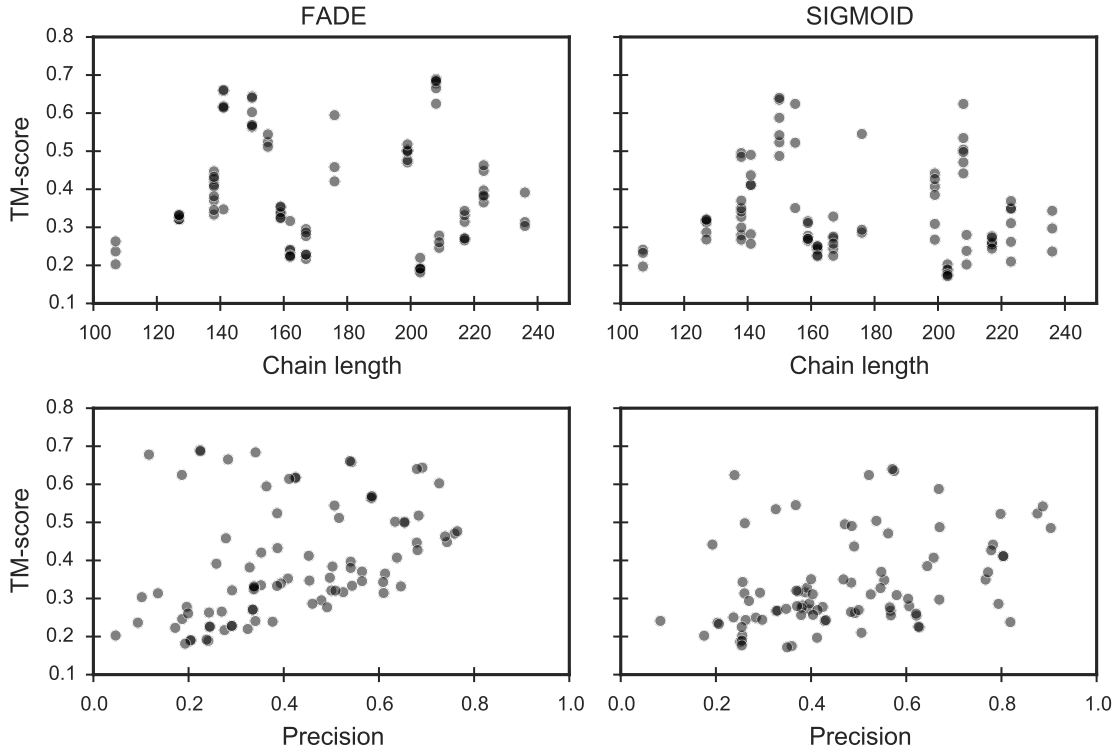


Figure 4.11: Effects of target chain length and restraint precision on the median TM-score for FADE and SIGMOID ROSETTA energy functions. Each scatter point represents a 1,000-decoy set.

4.3.3 Impact of metapredictors and energy functions on unconventional MR

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. Surprisingly, all remaining conditions resulted in fewer structure solutions than those from ROSETTA decoys. Furthermore, the conditions METAPSICOV (FADE function), METAPSICOV BBCONTACTS (FADE function) and PCONSC2 BBCONTACTS (FADE function) yielded no more than half of the structure solutions achieved by GREMLIN (SIGMOID function). The remaining two conditions - PCONSC2 (FADE function) and GREMLIN BBCONTACTS (FADE function) - resulted in four out of 18 structure solutions. The addition of BBCONTACTS did not improve decoy quality enough to increase the chances of structure solution success; however, the structure of the bovine peptide methionine sulfoxide reductase (PDB: 1fvg) was only solved with the GREMLIN BBCONTACTS (FADE function) decoys further supporting the small but important value of BBCONTACTS restraint addition to separately determined contact predictions (see Chapter XYZ).

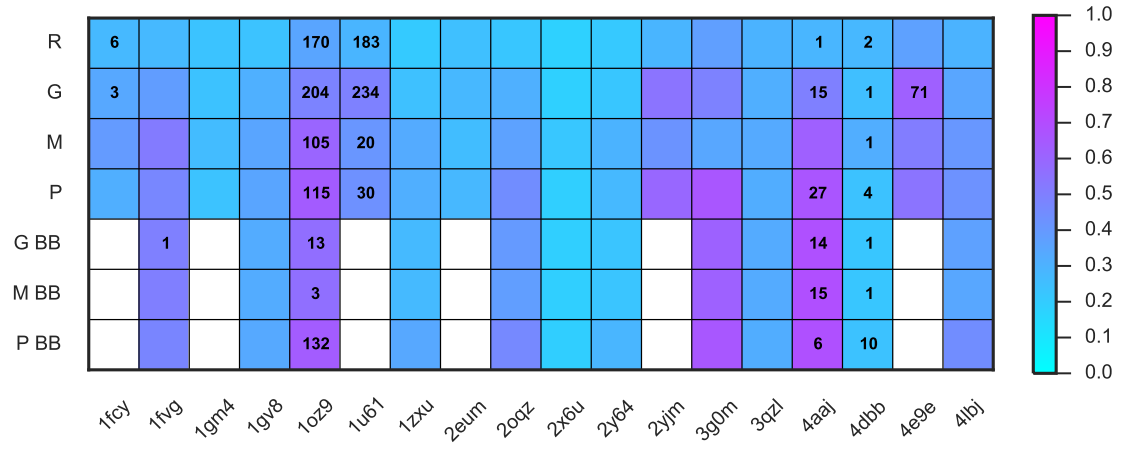


Figure 4.12: Structure solution count for AMPLE search models generated from decoys with varying contact prediction and ROSETTA energy function conditions: unrestrained ROSETTA (R); GREMLIN (G; SIGMOID function); METAPSICOV (M; FADE function); PCONSC2 (P; FADE function); GREMLIN BBCONTACTS (G BB; FADE 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 AMPLEs 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 fitted LOWESS function 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 [20] 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.

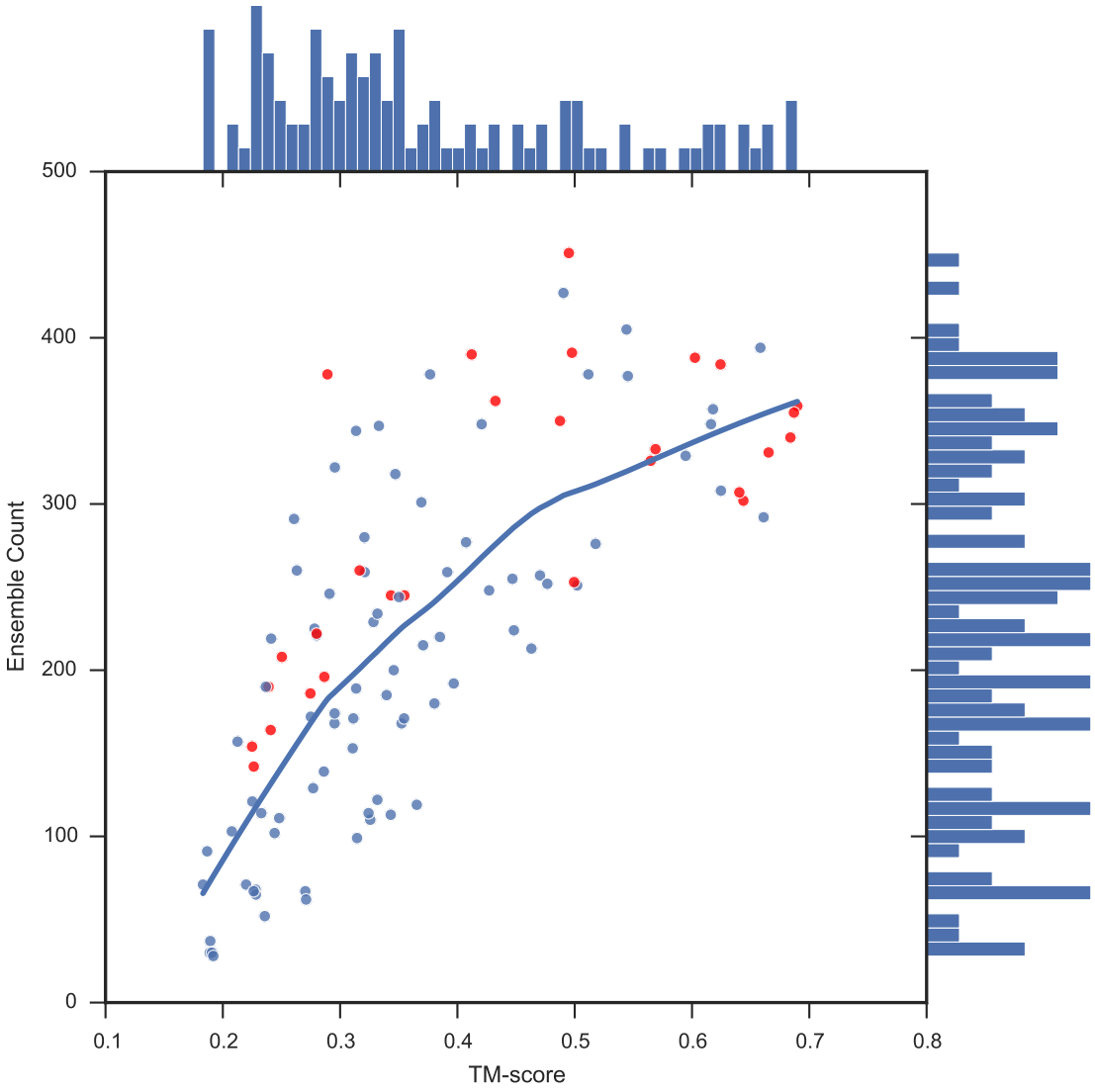


Figure 4.13: Comparison of median TM-score comparison (per 1,000 decoys) against the resulting AMPLE ensemble search model count. LOWESS function fitted to data to illustrate relationship. Red dots indicate successful ensemble sets.

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 XYZ). 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 Cα R.M.S.D. - 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 Cα R.M.S.D. 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 Cα R.M.S.D. 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).

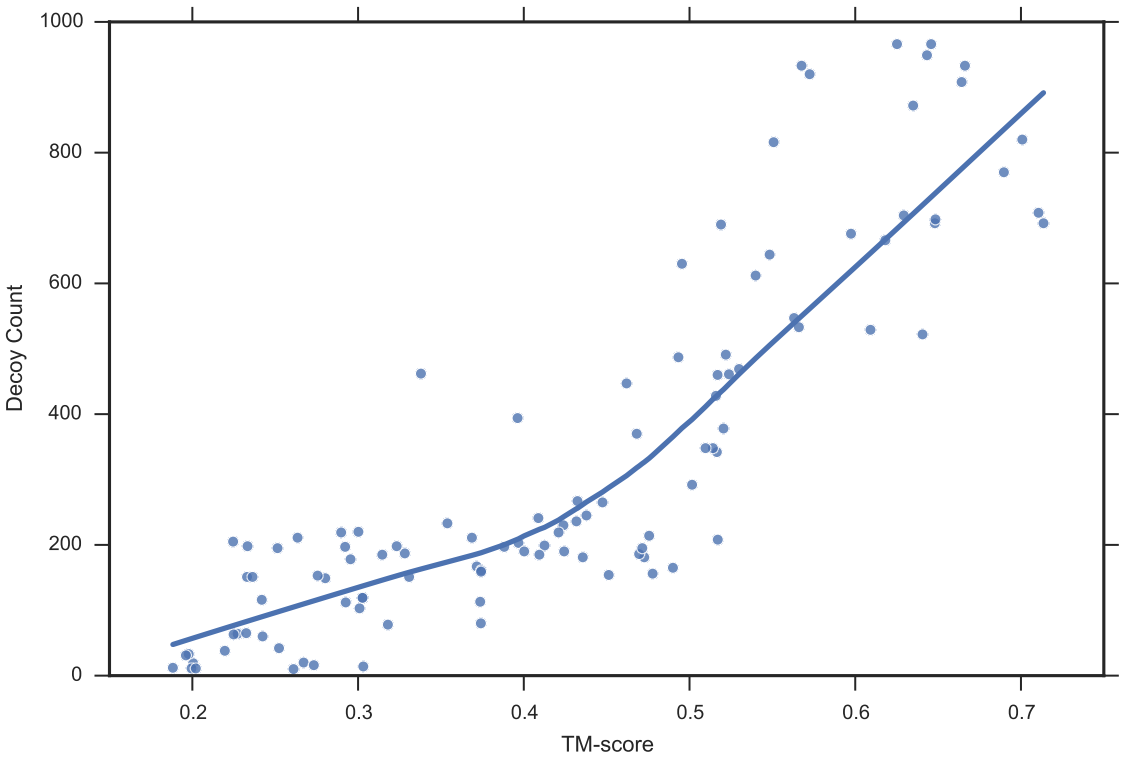


Figure 4.14: Relationship between cluster median TM-score and the number of cluster decoys. Blue line represents LOWESS relationship fitted to data.

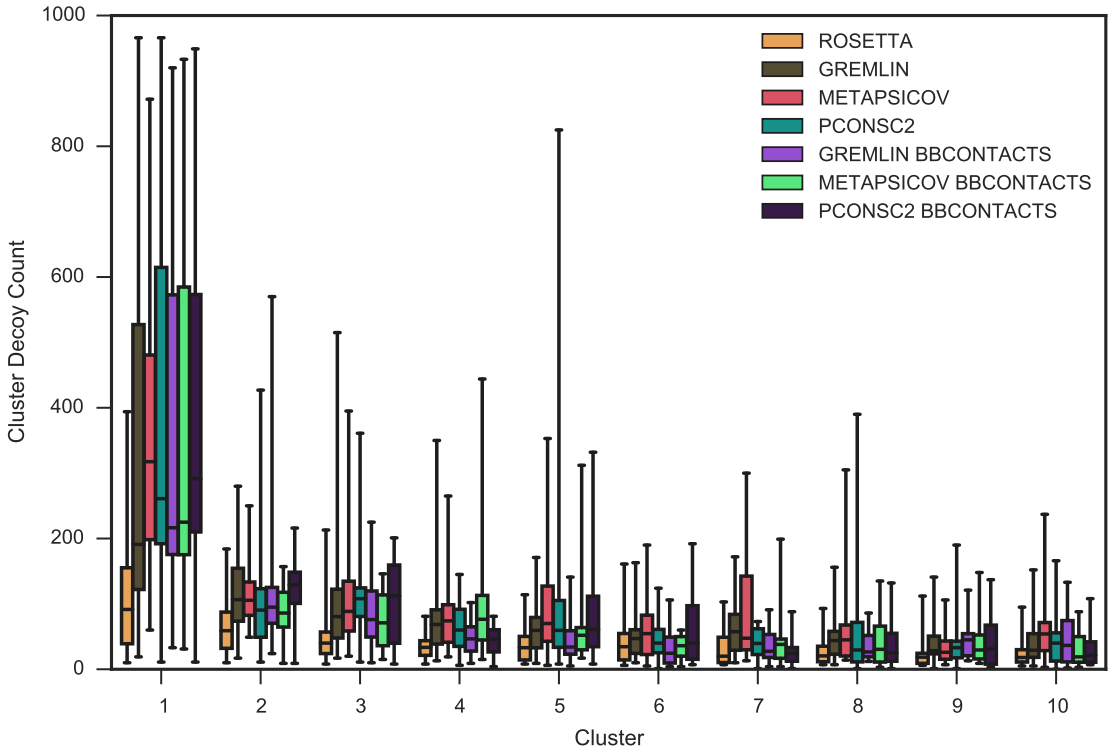
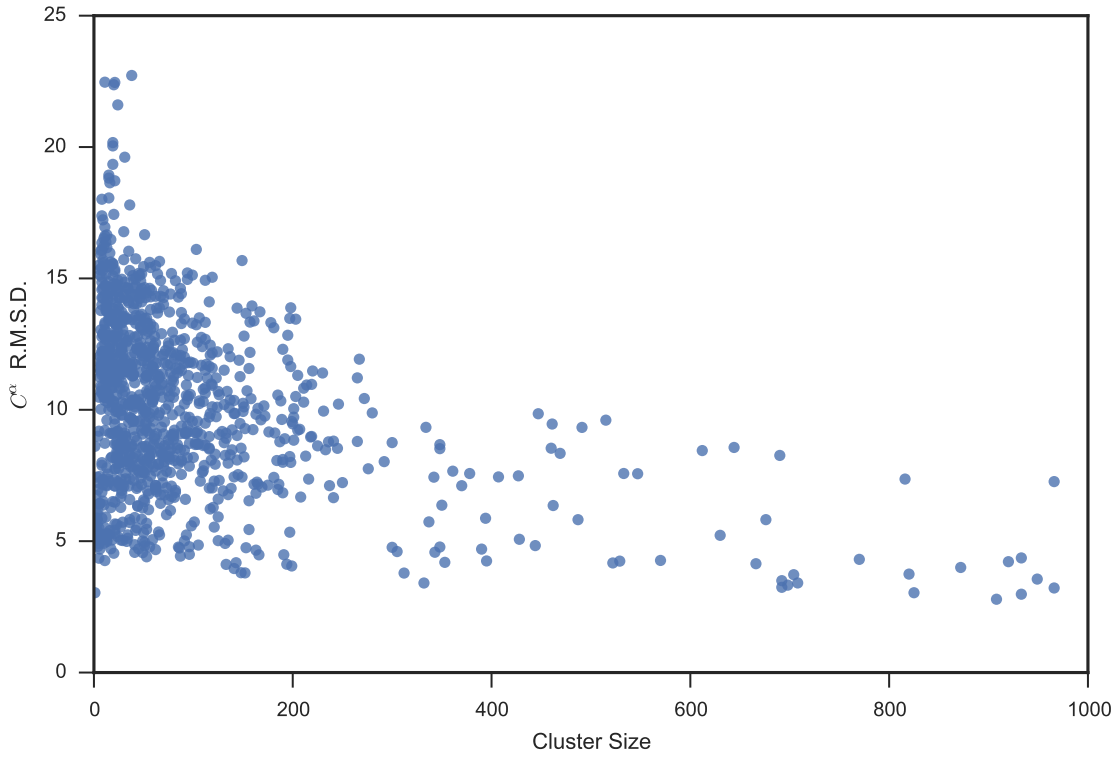
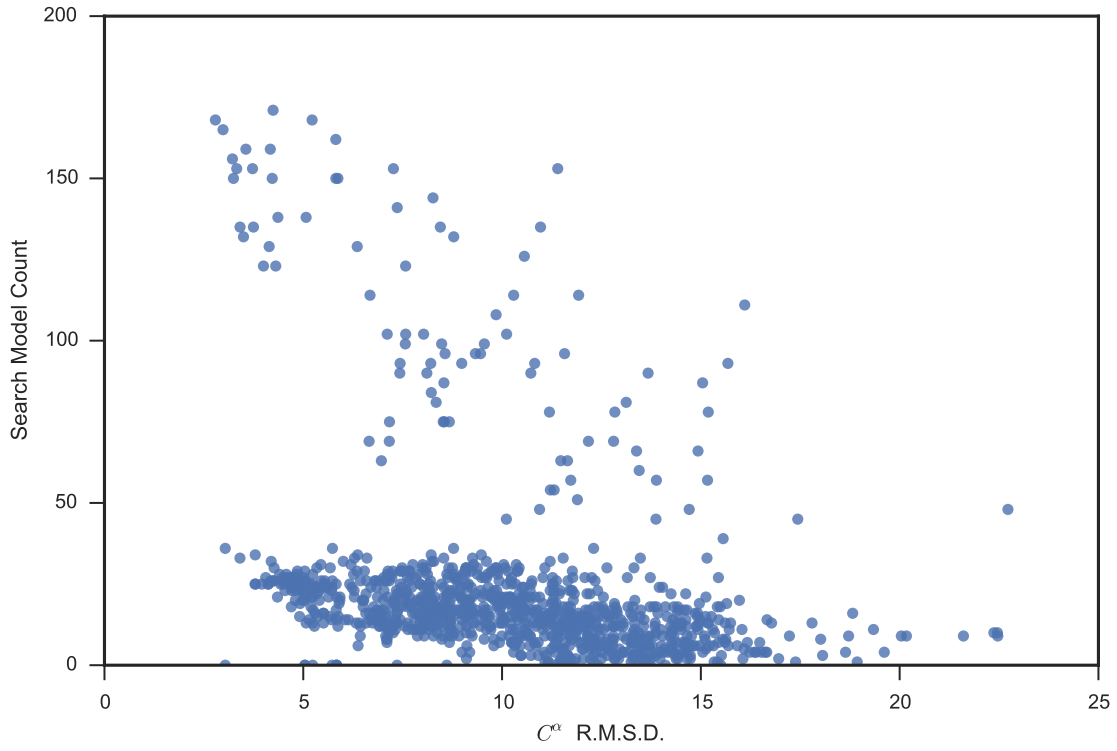


Figure 4.15: SPICKER cluster sizes of each target grouped the restraint condition used during the structure prediction protocol. Whiskers span the range from the minimum to maximum counts.



(a)



(b)

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

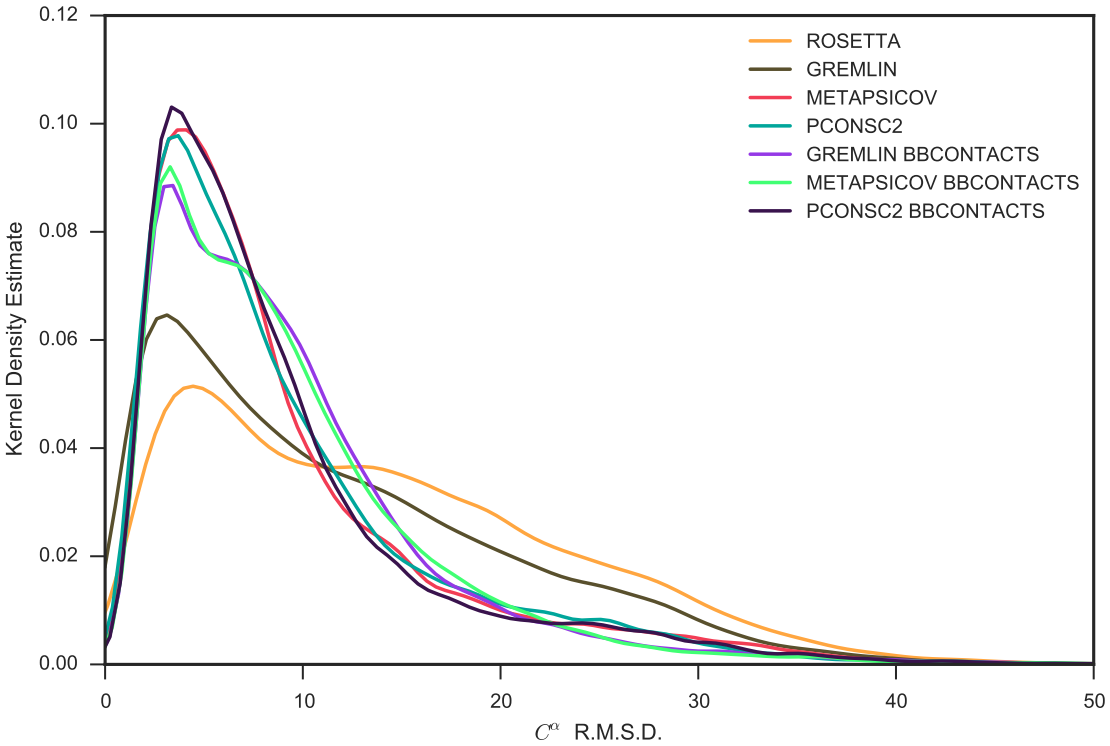


Figure 4.17: Kernel density estimate of Cα interatomic R.M.S.D. for SPICKER clusters.

The structure solution through pipelines like AMPLE and other unconventional MR software [11, 13] can result from the placement of generated (ensemble) search models either in- or out-of-sequence register. The RIO metric [17] 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.

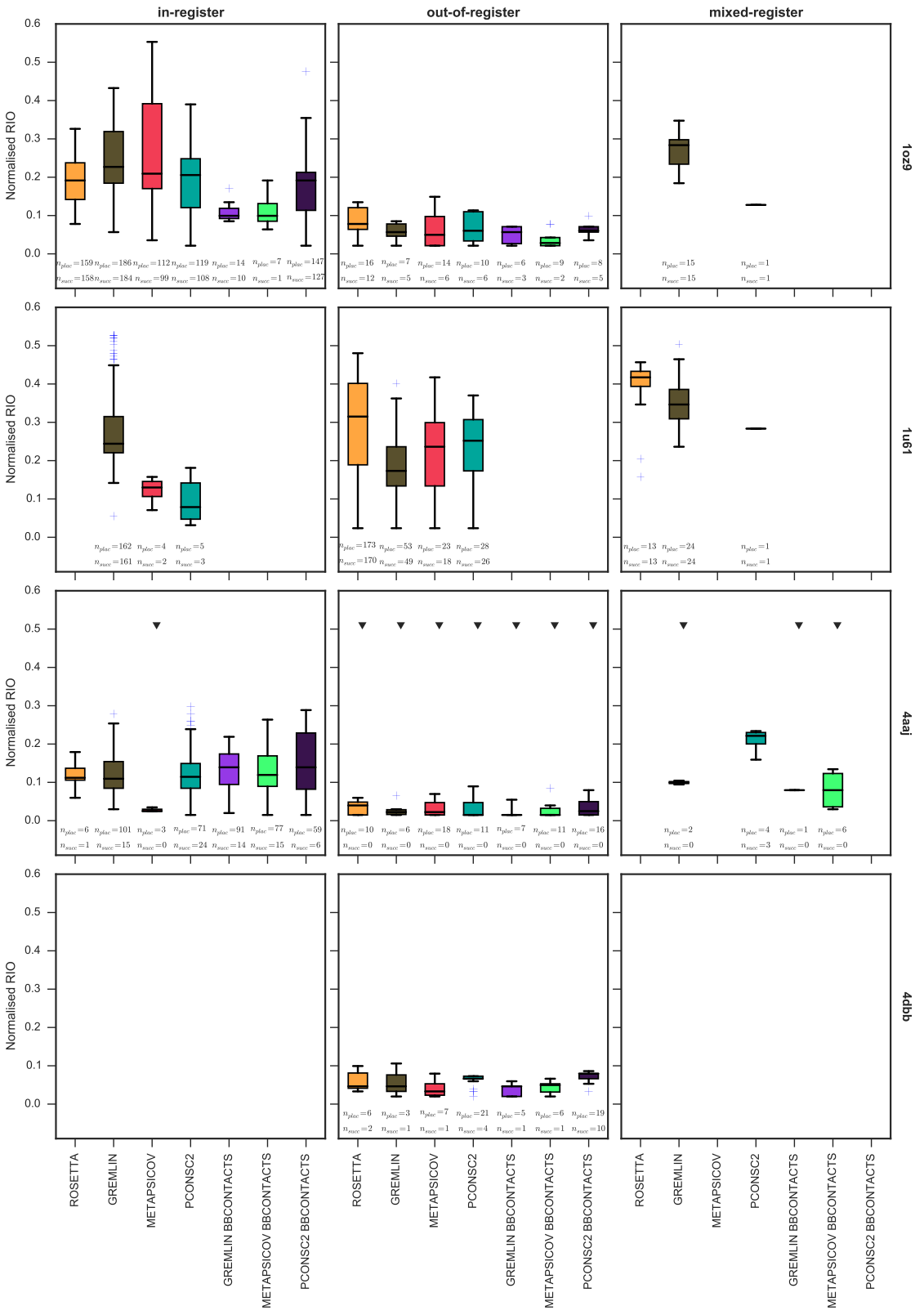


Figure 4.18: Normalised RIO score analysis of four successful targets in the MR dataset. Black triangles indicate AMPLE search model sets without a structure solution.

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). Al-

though 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).

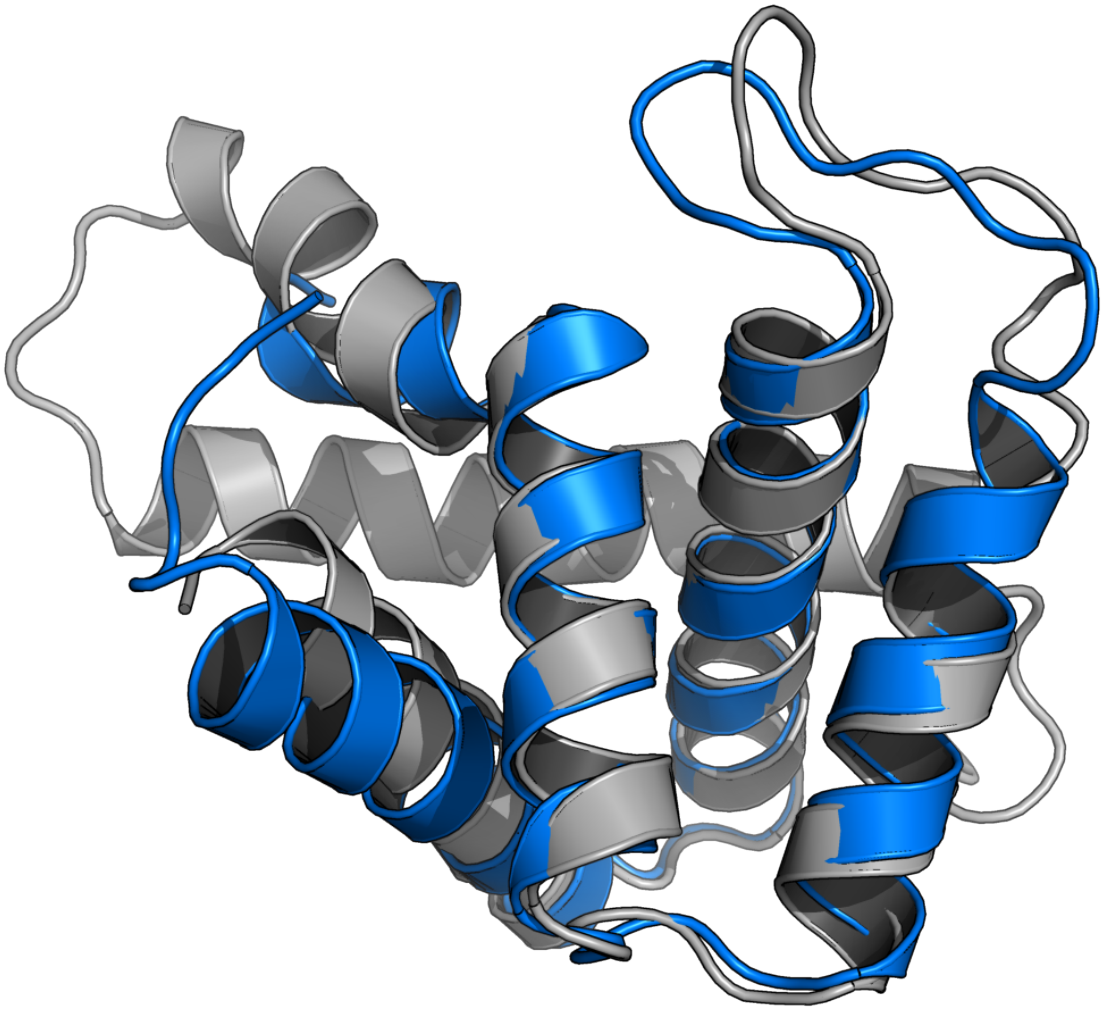


Figure 4.19: Successful search model (blue cartoon) post-PHASER placement superposed with the native structure (gray cartoon) for putative ribonuclease III (PDB: 1u61).

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 re-

sults. 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).

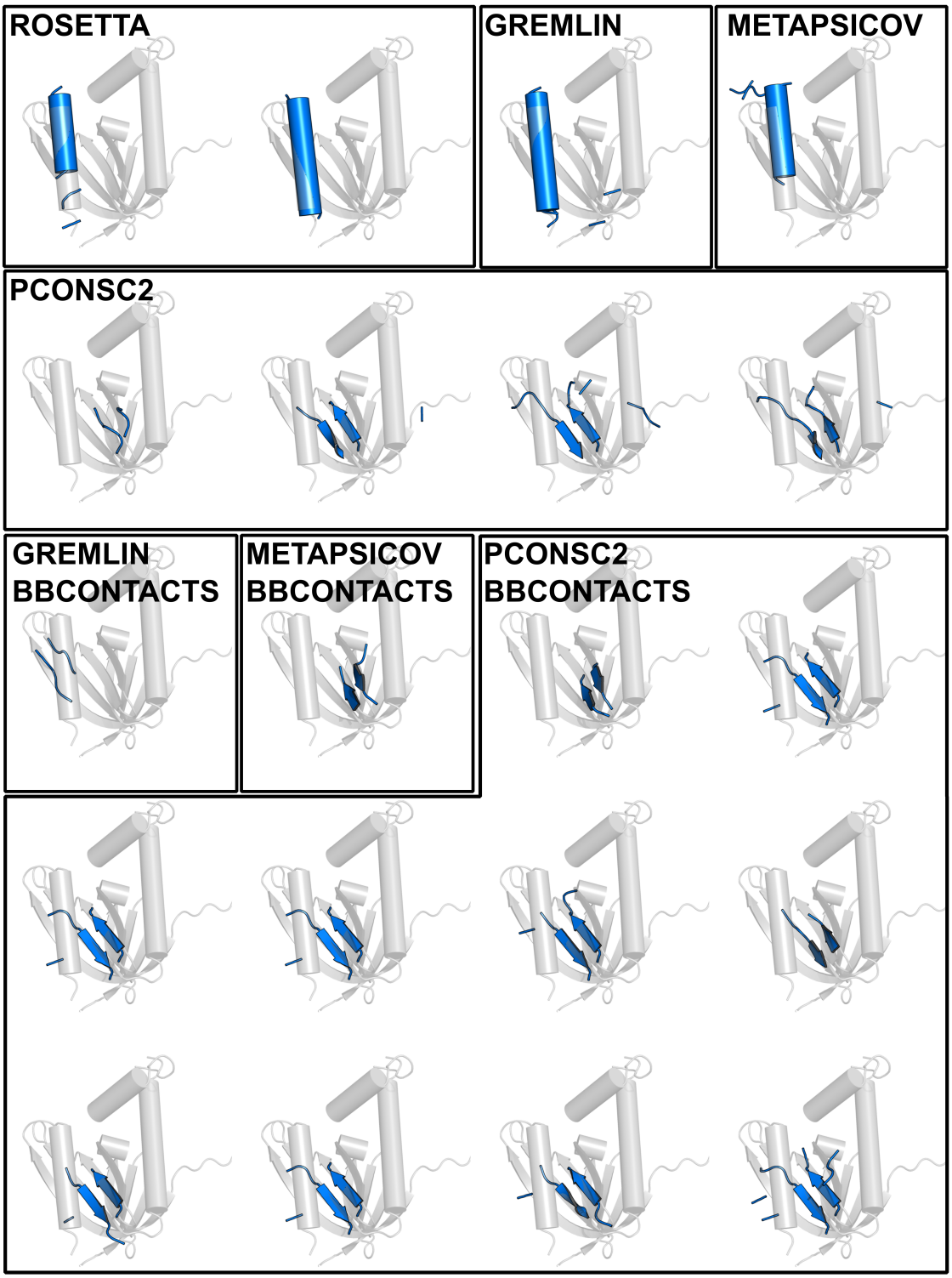


Figure 4.20: Successful search models post-PHASER placement (blue) superposed to the reference crystal structure (grey) for PTB domain of Mint1 (PDB: 4dbb).

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: 1fvg) 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.

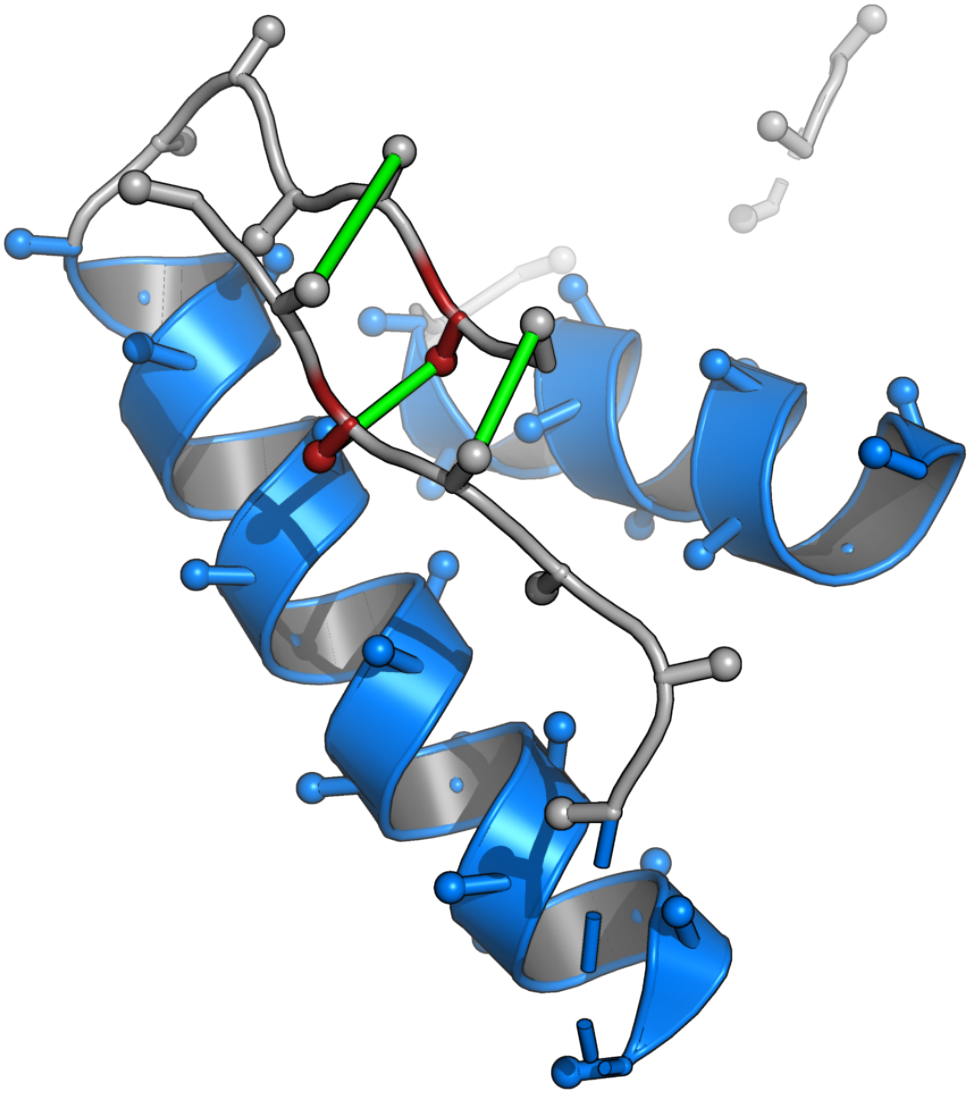


Figure 4.21: Successful search model post-PHASER placement for peptide methionine sulfoxide reductase (PDB: 1fvg). BBCONTACTS distance restraints are represented as green lines, α-helices in blue and β-strands in red. Secondary structure assignment calculated with STRIDE [1].

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 Molecular Replacement 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 [19, 8]. 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.: 680 [19]; de Oliveria et al.: 3500 [8]); (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 surprising 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 [4]. 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. This work provides further support to work outlined in Chapter XYZ.

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 XYZ, 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 ROSETTAs 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 satisfaction [8, 10] 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

Decoy subselection for ...

Chapter 6

Alternative *ab initio* structure prediction algorithms for AMPLE

Chapter 7

Fragments for MR ...

Chapter 8

Single model approach using AMPLE's cluster-and-truncate approach

Chapter 9

Software developments

Chapter 10

Conclusion

Bibliography

[1] D Frishman and P Argos. "Knowledge-based protein secondary structure assignment". en. In: *Proteins* 23.4 (Dec. 1995), pp. 566–579. ISSN: 0887-3585. DOI: 10.1002/prot.340230412.

[2] Baoji He et al. "NeBcon: Protein contact map prediction using neural network training coupled with naïve Bayes classifiers". en. In: *Bioinformatics* (Mar. 2017). ISSN: 1367-4803, 1367-4811. DOI: 10.1093/bioinformatics/btx164.

[3] David T Jones et al. "MetaPSICOV: combining coevolution methods for accurate prediction of contacts and long range hydrogen bonding in proteins". en. In: *Bioinformatics* 31.7 (Apr. 2015), pp. 999–1006. ISSN: 1367-4803, 1367-4811. DOI: 10.1093/bioinformatics/btu791.

[4] Hetunandan Kamisetty, Sergey Ovchinnikov, and David Baker. "Assessing the utility of coevolution-based residue–residue contact predictions in a sequence- and structure-rich era". In: *Proceedings of the National Academy of Sciences* 110.39 (Sept. 2013), pp. 15674–15679. DOI: 10.1073/pnas.1314045110.

[5] Jianzhu Ma et al. "Protein contact prediction by integrating joint evolutionary coupling analysis and supervised learning". en. In: *Bioinformatics* 31.21 (Nov. 2015), pp. 3506–3513. ISSN: 1367-4803, 1367-4811. DOI: 10.1093/bioinformatics/btv472.

[6] Mirco Michel et al. "Large-Scale Structure Prediction By Improved Contact Predictions And Model Quality Assessment". In: *bioRxiv* (2017). DOI: http://biorxiv.org/content/biorxiv/early/2017/04/18/128231.full.pdf.

[7] Mirco Michel et al. "PconsFold: improved contact predictions improve protein models". en. In: *Bioinformatics* 30.17 (Sept. 2014), pp. i482–8. ISSN: 1367-4803, 1367-4811. DOI: 10.1093/bioinformatics/btu458.

[8] Saulo Henrique Pires de Oliveira, Jiye Shi, and Charlotte M Deane. "Comparing co-evolution methods and their application to template-free protein structure prediction". en. In: *Bioinformatics* 33.3 (Feb. 2017), pp. 373–381. ISSN: 1367-4803, 1367-4811. DOI: 10.1093/bioinformatics/btw618.

[9] Sergey Ovchinnikov et al. "Large-scale determination of previously unsolved protein structures using evolutionary information". en. In: *Elife* 4 (Sept. 2015), e09248. ISSN: 2050-084X. DOI: 10.7554/eLife.09248.

[10] Sergey Ovchinnikov et al. "Protein structure prediction using Rosetta in CASP12". en. In: *Proteins* (Sept. 2017). ISSN: 0887-3585, 1097-0134. DOI: 10.1002/prot.25390.

[11] Dayté D Rodríguez et al. "Crystallographic ab initio protein structure solution below atomic resolution". en. In: *Nat. Methods* 6.9 (Sept. 2009), pp. 651–653. ISSN: 1548-7091, 1548-7105. DOI: 10.1038/nmeth.1365.

[12] Carol A Rohl et al. "Protein structure prediction using Rosetta". en. In: *Methods Enzymol.* 383 (2004), pp. 66–93. ISSN: 0076-6879. DOI: 10.1016/S0076-6879(04)83004-0.

[13] Massimo Sammito et al. "Exploiting tertiary structure through local folds for crystallographic phasing". en. In: *Nat. Methods* 10.11 (Nov. 2013), pp. 1099–1101. ISSN: 1548-7091, 1548-7105. DOI: 10.1038/nmeth.2644.

[14] Felix Simkovic et al. "Residue contacts predicted by evolutionary covariance extend the application of *ab initio* molecular replacement to larger and more challenging protein folds". en. In: *IUCrJ* 3.Pt 4 (July 2016), pp. 259–270. ISSN: 2052-2525. DOI: 10.1107/S2052252516008113.

[15] Marcin J Skwark et al. "Improved contact predictions using the recognition of protein like contact patterns". en. In: *PLoS Comput. Biol.* 10.11 (Nov. 2014), e1003889. ISSN: 1553-734X, 1553-7358. DOI: 10.1371/journal.pcbi.1003889.

[16] Jens M H Thomas et al. "Approaches to *ab initio* molecular replacement of α-helical transmembrane proteins". In: *Acta Crystallogr. D Biol. Crystallogr.* 73.12 (Dec. 2017), pp. 985–996. ISSN: 0907-4449. DOI: 10.1107/S2059798317016436.

[17] Jens M H Thomas et al. "Routine phasing of coiled-coil protein crystal structures with AMPLE". en. In: *IUCrJ* 2.Pt 2 (Mar. 2015), pp. 198–206. ISSN: 2052-2525. DOI: 10.1107/S2052252515002080.

[18] Tong Wang et al. “LRFragLib: an effective algorithm to identify fragments for de novo protein structure prediction”. en. In: *Bioinformatics* 33.5 (Mar. 2017), pp. 677–684. ISSN: 1367-4803, 1367-4811. DOI: 10.1093/bioinformatics/btw668.

[19] Qiqige Wuyun et al. “A large-scale comparative assessment of methods for residue–residue contact prediction”. In: *Brief. Bioinform.* (Oct. 2016). ISSN: 1467-5463. DOI: 10.1093/bib/bbw106.

[20] Jinrui Xu and Yang Zhang. “How significant is a protein structure similarity with TM-score = 0.5?” en. In: *Bioinformatics* 26.7 (Apr. 2010), pp. 889–895. ISSN: 1367-4803, 1367-4811. DOI: 10.1093/bioinformatics/btq066.