Contents

List of Figures

List of Tables

List of Equations

List of Abbreviations

CC	Correlation Coefficient
KDE	Kernel Density Estimate
MR	Molecular Replacement
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

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 Template-Modelling score (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 (Root-Mean-Square Deviation (RMSD) < 1.5Å). Although challenging by current means to identify these decoys, it is of great interest to structural biologists to do so since these decoys might be sufficient by themselves as Molecular Replacement (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 [1] and De Oliveira et al. [2] 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 [3] 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 (**??**), PREDICTORS (**??**) and TRANSMEMBRANE (**??**) datasets. Top-L ($>$ 5 residues sequence separation) CCMPRED [4], PCONSC2 [5], METAPSICOV STAGE 1 [6] and MEMBRAIN [7] 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 **??**) 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:

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

INDIVIDUAL keep the top-5 decoys only

The fixed definition in the *CUTOFF* strategy was determined by De Oliveira et al. [2]. 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 Kernel Density Estimate (KDE) and max_{kde} to the maximum KDE obtained by applying the algorithm described by Sadowski [8] 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 [9]

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

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

6.2.4 Molecular Replacement

To evaluate the benefits of such subselection to MR in AMPLE, a subset of 35 decoy sets (spanning 35 unique targets) were processed as described 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 [10] was shown to be advantageous in most cases by Thomas [11], 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 spects: 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 [12] with identical settings to those used in AMPLE.

Each MR run was assessed using the criteria defined in **??**.

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 [1] 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 [1] was limited to 10 representative targets with a maximum

chain length of 158 residues. Furthermore, FRAGFOLD [13] was used for *ab initio* protein structure prediction, a method with inferior performance to ROSETTA [14] 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 Correlation Coefficient (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 ϱcoefficients 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).

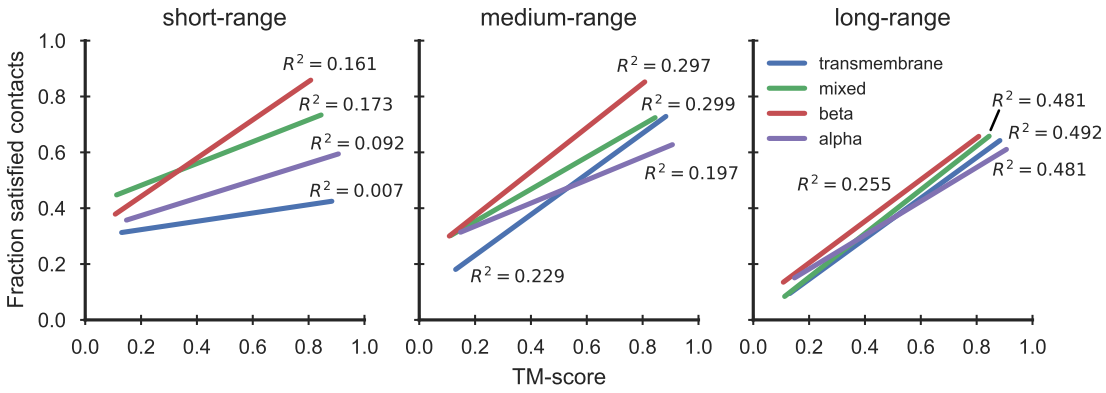


Figure 6.1: Linear regression model fitted to decoy TM-scores and corresponding fractions of satisfied, range-dependent contacts. Targets were further separated by fold classification. Coefficients of determination (R^2-values) added alongside each regression model.

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 Protein Data Bank (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 subdivide 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. Kosciolek and Jones [1] 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.

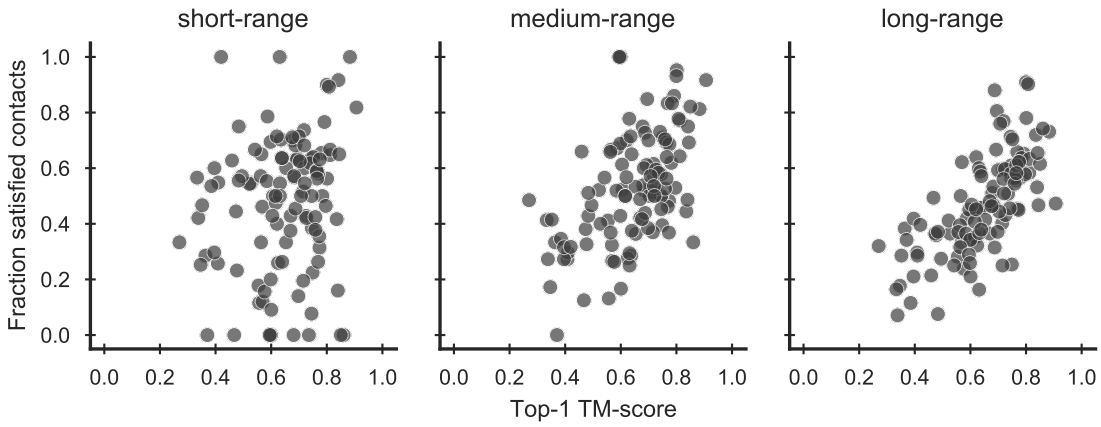


Figure 6.2: Analysis of the relationship between TM-score and contact satisfaction for the top-1 decoy (as ranked by TM-score) in each decoy set.

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.

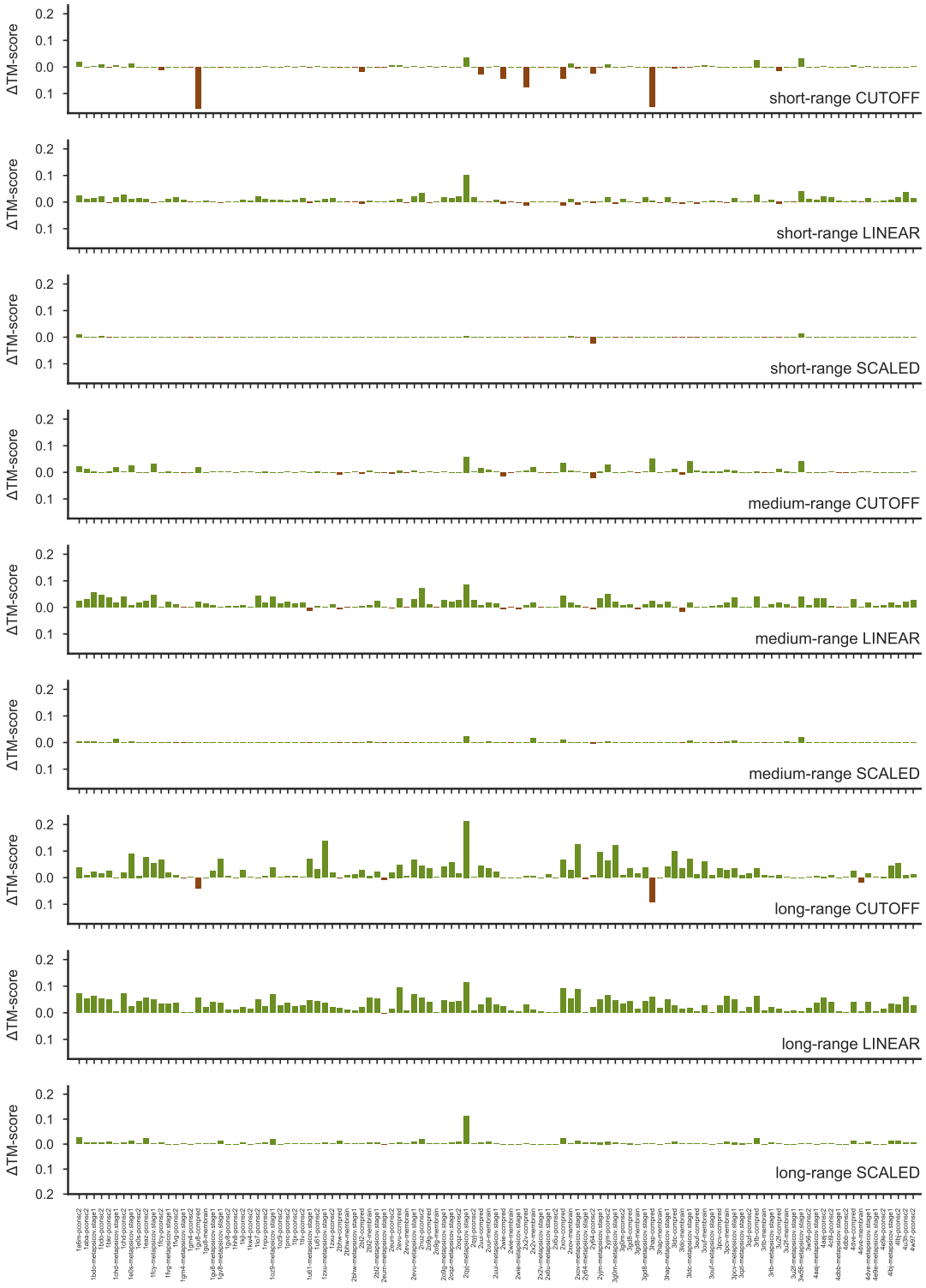


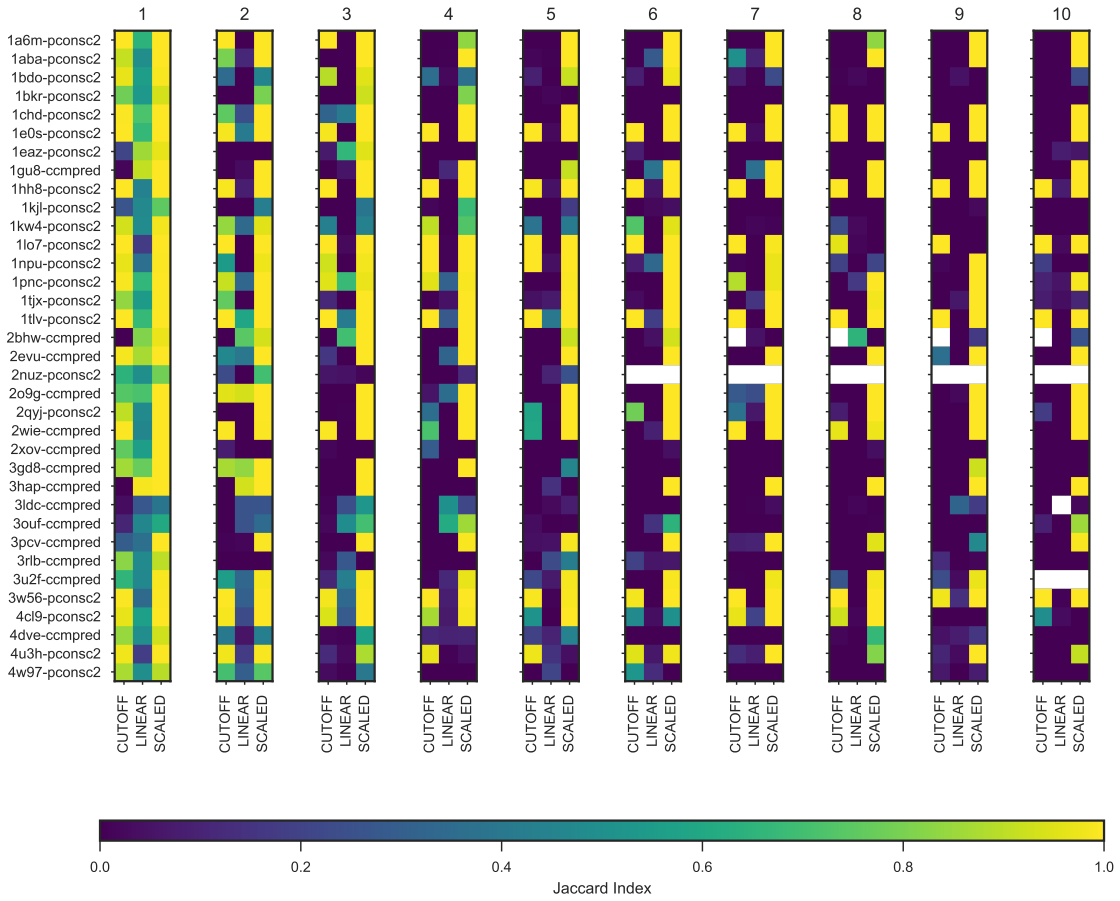
Figure 6.3: Differences in mean TM-score for decoy sets pre- and post-decoy subselection. Each subselection strategy is stated in each subplot along with the contact range used to establish decoy inclusion in the final set.

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

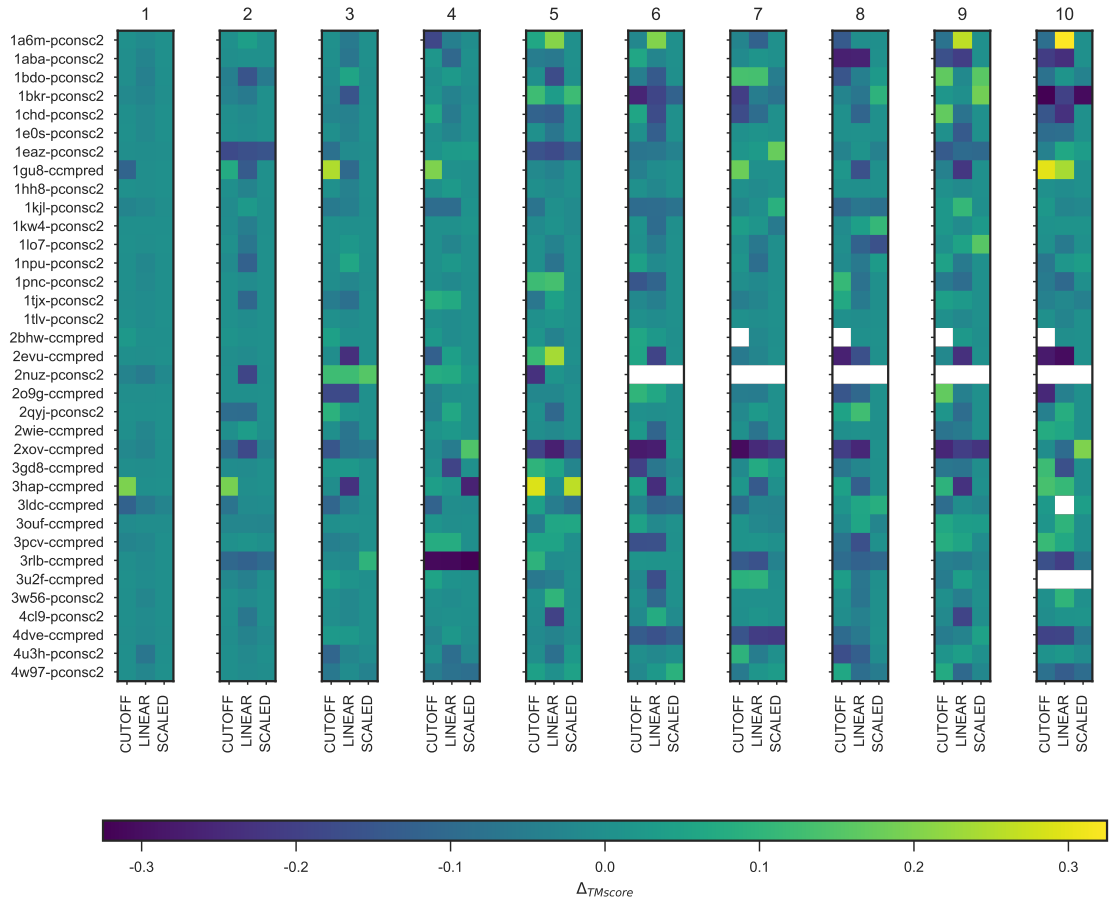
For evaluation of performance of filtered decoy sets in MR, a smaller sample of 35 decoy sets was selected spanning 35 unique targets (21 globular and 14 transmembrane targets). The contact prediction algorithm generating the restraints for the *ab initio* structure predictions was PCONSC2 (globular 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.


1
2
3
4
5
6
7
8
9
10
1a6m-pconsc2
1aba-pconsc2
1bdo-pconsc2
1bkr-pconsc2
1chd-pconsc2
1e0s-pconsc2
1eaz-pconsc2
1gu8-ccmpred
1hh8-pconsc2
1kjl-pconsc2
1kw4-pconsc2
1lo7-pconsc2
1npu-pconsc2
1pnc-pconsc2
1tjx-pconsc2
1tlv-pconsc2
2bhw-ccmpred
2evu-ccmpred
2nuz-pconsc2
2o9g-ccmpred
2qyj-pconsc2
2wie-ccmpred
2xov-ccmpred
3gd8-ccmpred
3hap-ccmpred
3ldc-ccmpred
3ouf-ccmpred
3pcv-ccmpred
3rlb-ccmpred
3u2f-ccmpred
3w56-pconsc2
4cl9-pconsc2
4dve-ccmpred
4u3h-pconsc2
4w97-pconsc2
CUTOFF
LINEAR
SCALED
0.0
0.2
0.4
0.6
0.8
1.0
Jaccard Index

(a)



(b)

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

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

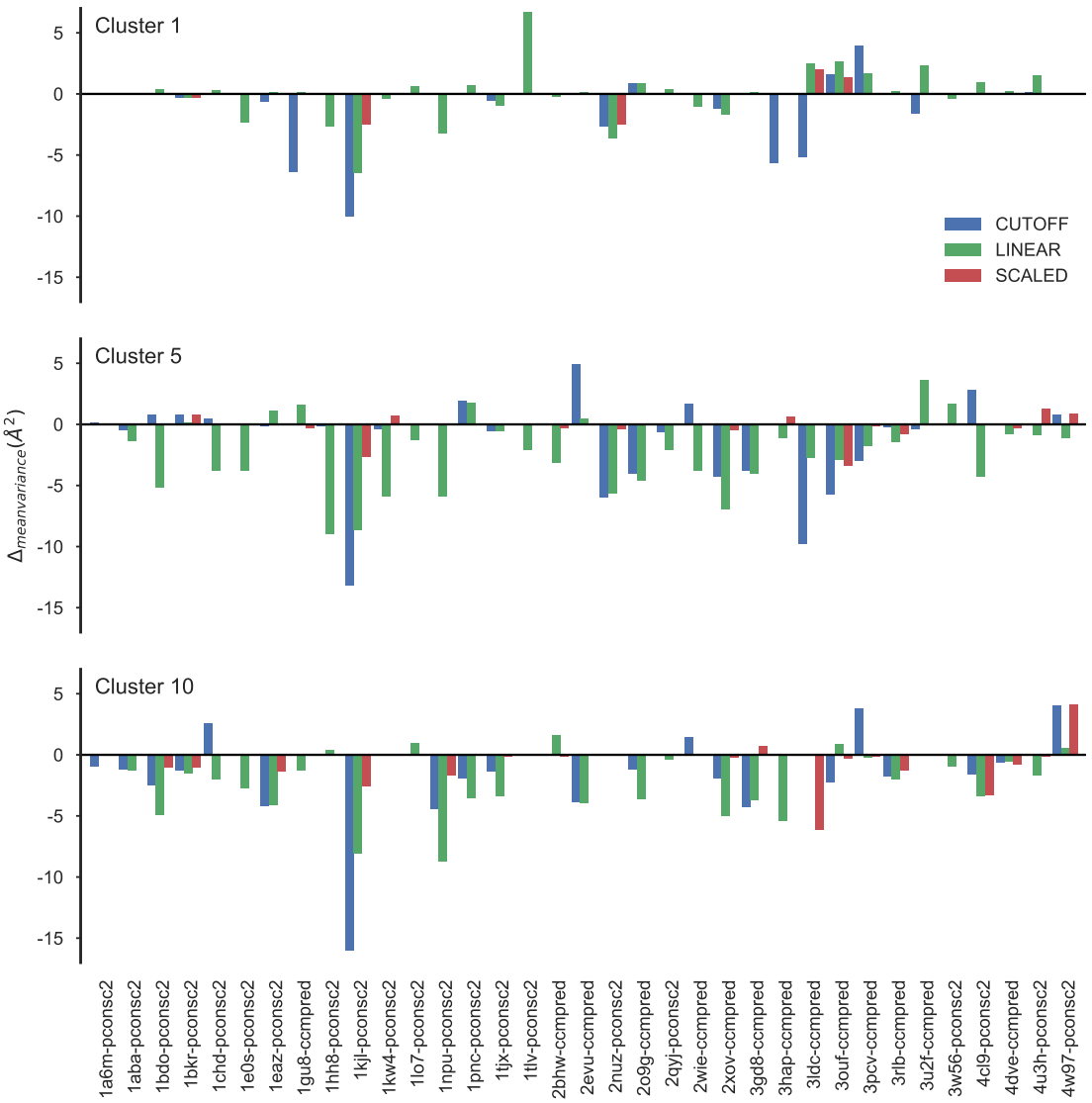


Figure 6.5: Effect of decoy subselection on mean inter-decoy THESEUS variance. Difference in mean variance calculated between the default and the three decoy subselection strategies *CUTOFF*, *LINEAR* and *SCALED*. Data for clusters 1, 5 and 10 shown as examples.

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 flects 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 [12, 15, 16], 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 80^{th} percentile or greater of TM-scores in each decoy set whilst 90% are in the at least the 40^{th} percentile (Fig. 6.6).

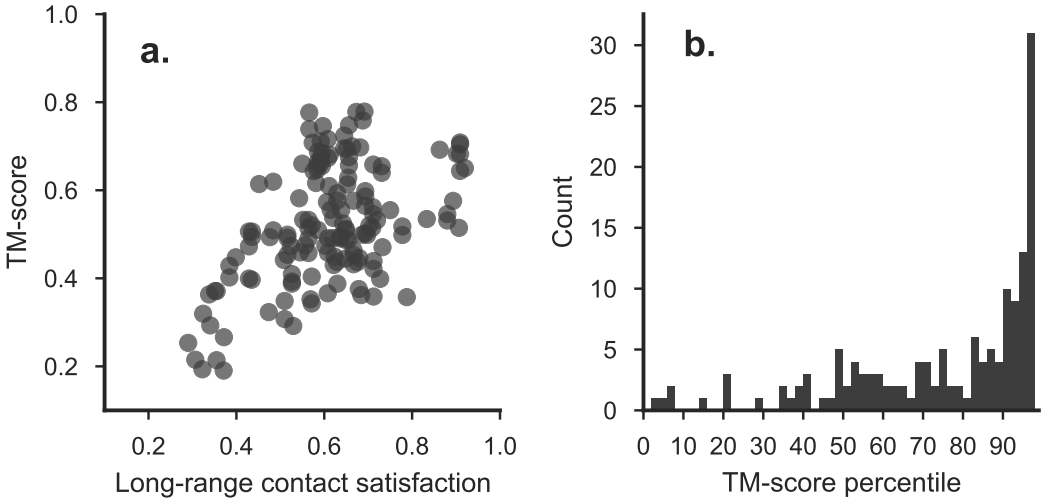


Figure 6.6: Analysis of long-range contact-satisfaction-based decoys with respect to the relationship between (a) the satisfaction and decoy quality and (b) the quality compared to the remaining, excluded decoy set.

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 4.0Å. This compares to a slightly lower RMSD change of 2.1Å for "DSSP"-treated decoys and 1.5Å for the "domain" treatment. Although almost all decoys were improved by either of the treatments, a small number of decoys 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 -1.6Å.

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

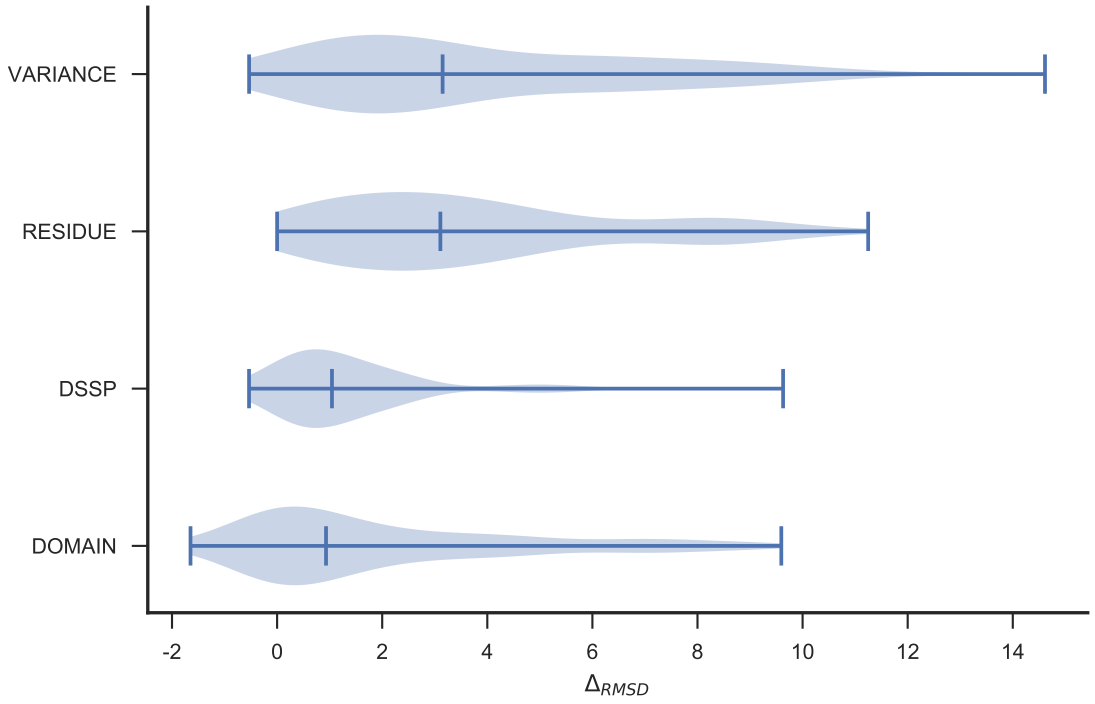


Figure 6.7: Distribution of differences in RMSD values between the initial and modified decoys under four different treatments. A positive Δ_{RMSD} value corresponds to a decrease in RMSD compared to the crystal structure.

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.

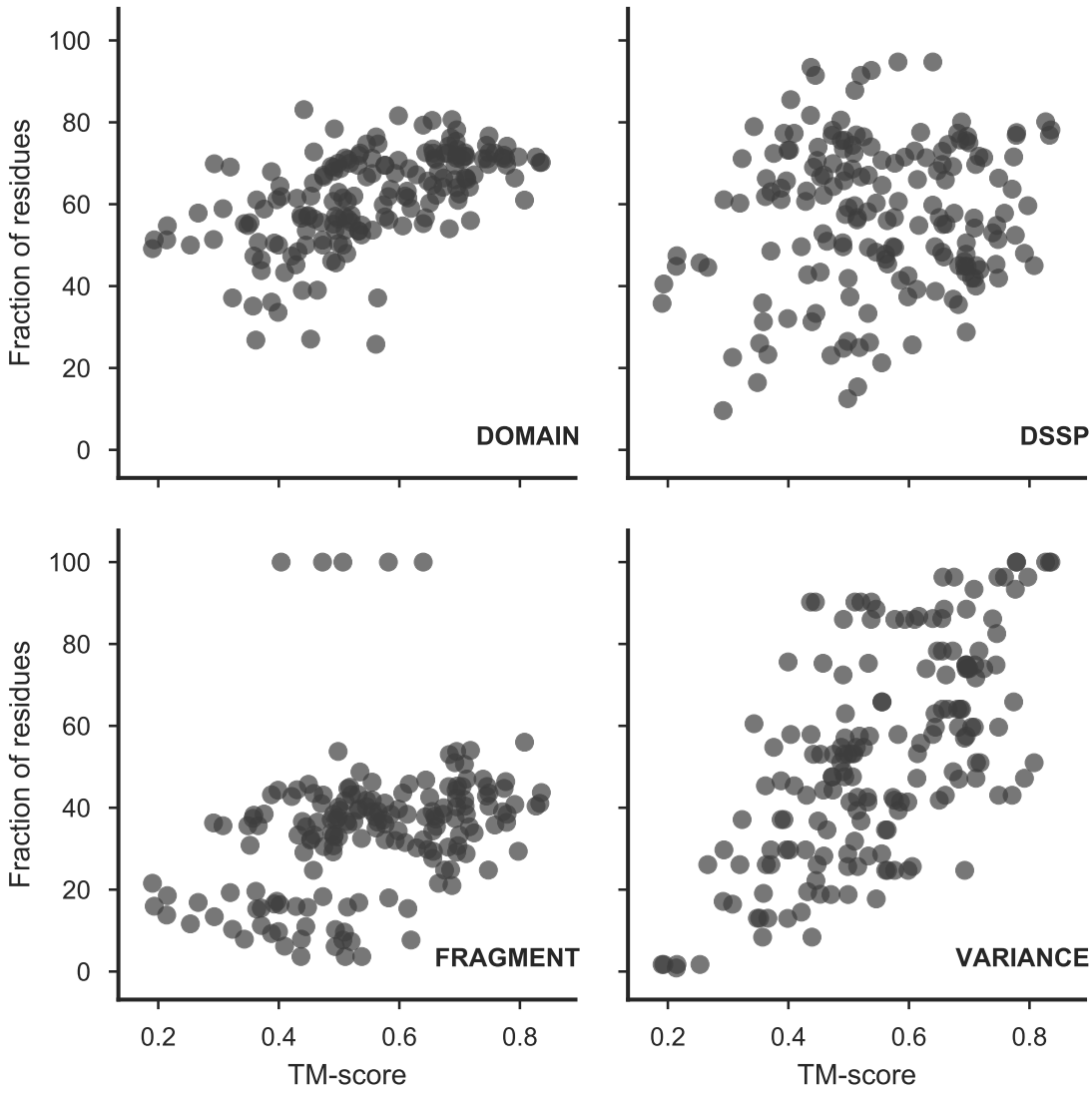


Figure 6.8: Scatter plots of initial decoy TM-score and the fraction of residues retained after one of four different residue removal treatments.

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.

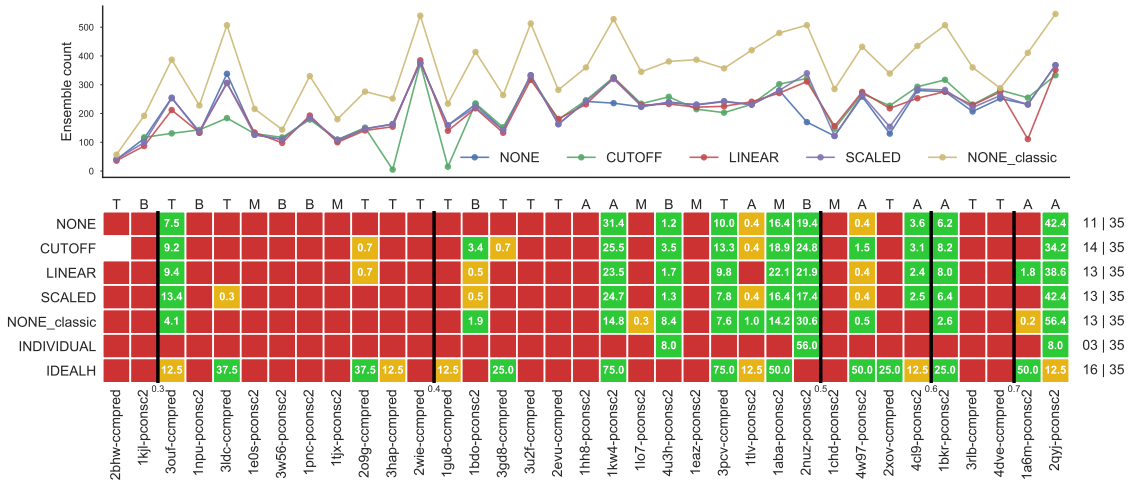


Figure 6.9: Molecular Replacement summary of decoy-subselected AMPLE ensembles. AMPLE-generated ensemble counts illustrated at the top with Molecular Replacement results in grid below: red cell equates to no solution; orange to a singleton solution; and green to multiple solutions. All *INDIVIDUAL* attempts 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 KstR2 (PDB ID: 4w97) sequence and contained one, two or three small helical fragments.

In certain cases the subselection of starting decoys made a subtle yet essential difference to 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.

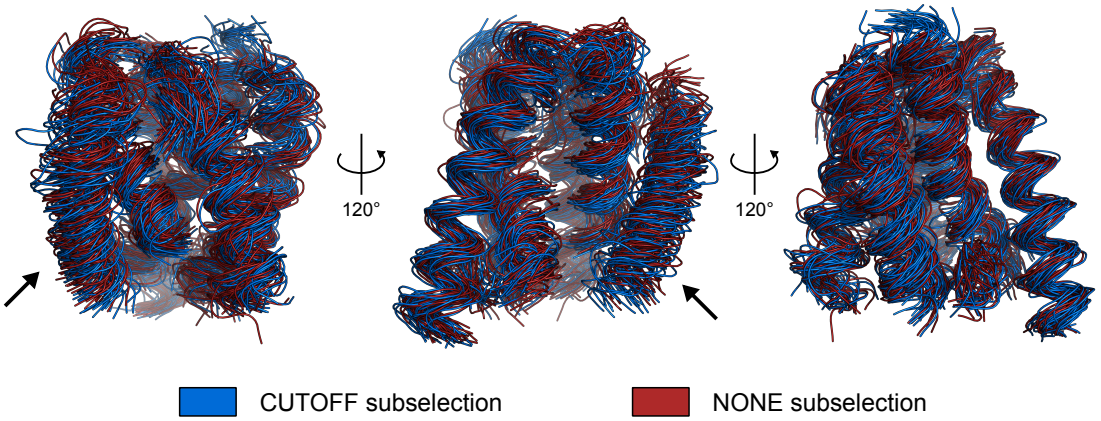


Figure 6.10: Example of the structural divergence of two ensemble search models with an identical AMPLE cluster-and-truncate path. Ensembles are based on the CCMPRED decoy set of PDB ID 2o9g and derived from cluster 1 with 59% truncation and subclustering radius of 3Å. The blue ensemble was derived from *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 [16] 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. [17], 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. [17] suggested that decoy-derived search models are essential since ideal helices provide insufficient scattering matter with low resolution (> 2Å) intensity data. In this study, these findings cannot be validated given that PDB ID 1gu8 (resolution of 2.27Å) was solved solely with ideal helices whilst being the target with the lowest resolution of all solved ones.

6.4 Discussion

The subselection of decoy sets by long-range contact satisfaction is a concept originally proposed by Kosciolek and Jones [1] and later confirmed and extended by De Oliveira et al. [2] and Adhikari and Cheng [3]. 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 AM-

PLE'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 constradicts the one reported by Thomas et al. [17], 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 [18] or FRAGON [19], 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 [20]. 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 [12, 15, 16], 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 po-

tentially 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 [1–3]. 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

Chapter 8

Conclusion

Appendix A

Appendix

Bibliography

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

[2] S. H. P. De Oliveira, J. Shi, C. M. Deane, en, *Bioinformatics* **Feb. 2017**, *33*, 373–381.

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

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

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

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

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

[8] M. I. Sadowski, en, *Proteins: Struct. Funct. Bioinf.* **Feb. 2013**, *81*, 253–260.

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

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

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

[12] 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.

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

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

[15] 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.

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

[17] 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.

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

[19] H. T. Jenkins, *Acta Crystallographica Section D Structural Biology* **Mar. 2018**, *74*, 205–214.

[20] 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.