# Improving Biologically Inspired Protein Structure Prediction using Satisfied Predicted Contacts




Laura Depner

St Cross College

University of Oxford

Thesis submitted for the Degree of

*Master of Science by Research*

Trinity Term 2018

*"It is hard to make the boat go as fast as you want to. The enemy of course, is resistance of the water, as you have to displace the amount of water equal to the weight of the men and equipment, but that very water is what supports you and that very enemy is your friend. So is life: the very problems you must overcome also support you and make you stronger in overcoming them."*

**- George Yeoman Pocock**

# Acknowledgements

Firstly, I would like to thank my supervisors, who guided me through my time at Oxford. Saulo, thank you for being patient, kind and always on hand to answer my questions, regardless of location. Charlotte, thank you for supporting me even when you did not agree with my decisions, and when things did not go to plan. I have learnt more from both of you than you could ever imagine.

To the Oxford Protein Informatics Group, thank you for being some of the most welcoming, enthusiastic, clever people I know. It has been great to be part of such a friendly research group, who all know how to keep a healthy work-life-cake-crossword balance.

Clare, thank you for being my buddy throughout the last three years - always there with advice and an understanding ear, together on this confusing journey out of the lab. I will always treasure my alcohol dehydrogenase monomer, and I hope you don't forget me when you are running the world.

Ellie and Hannah, two of the mightiest lighties out there. I am so, so unbelievably grateful to you for always being there.

Thank you to my family, for supporting me whenever I have needed it (mostly through endless visits to the Trout) and coming to visit me when I couldn't make it home.

Wolfson College Boat Club - thank you for changing my life. Thank you for the people I have met, and the experiences I have had. I couldn't possibly mention everyone, but thank you to Stef, Abi and Elo for being some of the most crazy, driven people I have ever met, and for always motivating me and showing me that hard work can get you wherever you want to be.

Finally, to the Oxford University Women's Boat Club of 2018 - Abi, Alice, Anna, Beth, Ellie, Jess, Juliette, Kate, Linda, Liv P, Liv W, Mads, Morgan, Rachel, Renee, Sara, Sanja, SPR, Stef, Sophie and Tilly. You are some of the strongest, most inspirational people I know - I have been so lucky to share the past year with you. Thank you for being my family through some of the highest and lowest points. I'm excited to see where we all end up.

# Abstract

The number of published protein structures is vastly outnumbered by protein sequences - this gap is only increasing as sequencing technologies continue to advance. Experimental methods of protein structure determination are costly and slow; computational protein structure prediction could be a solution.

A major problem with current methods of protein structure prediction is the large amount of computational time needed to generate an accurate structure for one protein. For some cases, notably smaller proteins, structures can now be reliably and accurately predicted. However, for a large number of proteins, especially longer proteins, the computational cost means that structure prediction remains infeasible using current methods.

SAINT2 is a sequential, fragment-based *de novo* structure prediction method that has been demonstrated to be faster and comparably successful at predicting correct structures than other methods. The aim of this study is to increase the number of protein families that SAINT2 can predict correct structures for by utilising predicted contacts in a unique way.

This thesis finds that in some cases, evaluating the long-range satisfied predicted contacts in a partial model can give an indication as to the final model quality. Discarding partial models predicted to lead to an incorrect final structure leads to overall enrichment of the decoy pool for these proteins. However, it is still unclear how to identify cases that our new version of SAINT2 would be better at predicting.

We also find that during extrusion, correct and incorrect models can satisfy and break predicted contacts differently. Correct models may be forming stable sub-structures early on in extrusion, whereas incorrect models appear to be breaking contacts throughout extrusion. It may be possible to identify incorrect predicted contacts based on their rank by satisfaction in models. This may lead to an increase in model quality if contacts predicted to be incorrect are down-weighted in further SAINT2 runs.

# Contents

**List of Figures** **iv**

**1 Introduction** **1**
- 1.1 Project Aim . . . 4
- 1.2 *De novo* Protein Structure Prediction . . . 5
  - 1.2.1 Critical Assessment of protein Structure Prediction . . . 6
  - 1.2.2 Conformational Space . . . 6
  - 1.2.3 Fragment Libraries . . . 7
  - 1.2.4 Scoring Functions . . . 8
  - 1.2.5 Model Selection . . . 9
  - 1.2.6 Refinement . . . 9
- 1.3 Evolutionary Couplings . . . 10
  - 1.3.1 Early Approaches . . . 11
  - 1.3.2 Improved Co-evolutionary Coupling Algorithms . . . 12
  - 1.3.3 Metapredictors . . . 12
  - 1.3.4 Using Evolutionary Couplings to Predict Structure . . . 14
- 1.4 Biologically Inspired Protein Structure Prediction: SAINT2 . . . 15
  - 1.4.1 Cotranslational Folding . . . 15
  - 1.4.2 How SAINT2 works . . . 16
- 1.5 Outline . . . 17

**2 Methods for Improving the SAINT2 Search Strategy** **20**
- 2.1 Hardware . . . 20
- 2.2 Definitions . . . 20
  - 2.2.1 Model Quality Assessment . . . 21
- 2.3 Data Sets . . . 22
  - 2.3.1 Targets with Correct Models Data Set . . . 22
  - 2.3.2 100+ Residues Data Set . . . 23

2.3.3 N-terminal Contacts Data Set . . . 24
2.3.4 Testing cut-off versions . . . 24
2.3.4.1 Reduced Correct Set . . . 24
2.3.5 Good and Bad Proteins Data Set . . . 25
2.4 Larger Dataset . . . 25
2.5 SAINT2 . . . 26
2.5.1 Secondary structure and torsion angle prediction . . . 26
2.5.2 Fragment Library generation . . . 27
2.5.3 Residue-residue contact prediction . . . 27
2.5.4 Scoring Function . . . 28
2.6 Implementation of Satisfied Predicted Contacts cut-off . . . 28
2.6.1 Version 1: strict cut-off . . . 29
2.6.2 Version 2: stepped cut-off . . . 30
2.6.3 Version 3: stepped cut-off to optimum . . . 32
2.6.4 Version 3.2: stepped down cut-off to optimum with time restriction . . . 33

**3 Improving the SAINT2 Search Strategy 35**
3.1 Model Generation Timings . . . 36
3.2 Satisfied Predicted Contacts during Decoy Generation . . . 38
3.3 Design of an SPC cut-off . . . 41
3.3.1 Identification of suitable cut-off value for SPCs . . . 41
3.4 Testing SPC Cut-off Implementations . . . 43
3.4.1 Version 1: Strict Cut-off . . . 43
3.4.2 Version 2: Stepped Cut-off . . . 44
3.4.3 Version 3: Stepped to Optimum Cut-off . . . 46
3.4.4 Time Regression . . . 48
3.4.5 Version 3.2: Stepped to Optimum Cut-off with Time Limit . . . 50
3.4.6 Further Testing of Version 3.2 . . . 51
3.4.6.1 Number of decoys needed for correct model . . . 52
3.5 SAINT2 Diagnostics . . . 54
3.6 Testing Protocol on Larger Datasets . . . 55
3.6.1 Dataset 1 . . . 55
3.6.2 Dataset 2 . . . 61
3.7 Discussion . . . 66

**4 Investigating the Behaviour of Predicted Contacts** **68**

4.1 Methods for Investigating the Behaviour of Predicted Contacts . . . . . 70

4.1.1 Analysis of Contacts Present in Correct and Incorrect Models . . . . . 70

4.1.2 Tracking Forming and Breaking of Contacts . . . . . 70

4.1.3 Analysing the Satisfaction of Individual Predicted Contacts . . . . . 71

4.2 Satisfaction of Predicted Contacts in Correct and Incorrect Models . . . . . 71

4.3 Decoy Tracking and Lost Contacts . . . . . 75

4.4 Satisfaction of Individual Predicted Contacts . . . . . 77

4.5 Discussion for Investigating the Behaviour of Predicted Contacts . . . . . 85

4.5.1 Behaviour of Predicted Contacts during Model Extrusion . . . . . 85

4.5.2 Satisfaction of Individual Predicted Contacts . . . . . 87

**5 Conclusions and Future Work** **89**

**References** **91**

**A SI** **99**

# List of Figures

1.1 The growth of GenBank (including WGS project) and PDB databases from 1982-2018. . . . . 2
1.2 Prediction of residues in contact (within 8Å) in 3-D protein structure, from correlated mutations identified from a multiple sequence alignment. . . . . 10
1.3 A flowchart describing how SAINT2 works to predict the structure of a protein given its sequence. An N-terminal fragment to represent the N-terminus of the sequence is used as input into SAINT2 and move steps or extrusion steps are then carried out until a complete model is produced. . . . . 16
1.4 A cartoon representation of a model of the soluble protein 1AIU, produced sequentially by SAINT2, being extruded in increments of 10% of protein length. Each length of model is comprised of the previous, N-terminal fragment (blue), and the newly extruded residues (purple). . . . . 18

2.1 Version 1 of our cut-off implemented into SAINT2 to filter out partial models predicted to lead to incorrect final structures. . . . . 29
2.2 Version 2 of our cut-off implemented into SAINT2 to filter out partial models predicted to lead to incorrect final structures. . . . . 30
2.3 Version 3 of our cut-off implemented into SAINT2 to filter out partial models predicted to lead to incorrect final structures. . . . . 32
2.4 Version 3.2 of our cut-off implemented into SAINT2 to filter out partial models predicted to lead to incorrect final structures. . . . . 33

3.1 Time in seconds to generate one model as extrusion occurs for 1AIU. . . . . 37
3.2 Satisfied Predicted Contacts at each percentage length for 10,000 decoys for the Targets with Correct Models data set, split into correct (blue) and incorrect (red) decoys. (top left) Total predicted contacts satisfied. (top right) LR predicted contacts only. (bottom left) SR contacts only. . . . . 39

3.3 Satisfied Predicted Contacts at each % length for 10,000 decoys for the all membrane proteins in the data set, split into correct (blue) and incorrect (red) decoys. (top left) Total predicted contacts satisfied. (top right) LR predicted contacts only. (bottom left) SR contacts only. . . . . 40
3.4 Density of LR-SPCs for correct (blue) and incorrect (red) models in a run of 10,000 models for proteins from the N-Terminal Contacts Data Set (Table S2). 42
3.5 Density of LR-SPCs for correct (blue) and incorrect (red) models from a run of 10,000 models for the soluble proteins 1NAT, 1SFE and 1VL1. . . . . 42
3.6 A comparison of structure prediction and generation of decoys with no cut-off (blue), and the strict cut-off protocol (purple). Decoys were generated as detailed in Table 2.3. The values compared were running time (limited to a maximum of 24 hours) (top left); Total number of complete decoys produced (top left); Total number of correct decoys produced for each protein (top right); the proportion of decoys produced that are correct for each protein, with TM-score of top scoring decoy labelled (bottom left); the enrichment of correct decoys using the cut-off (bottom middle). . . . . 44
3.7 Comparing runs of 1,000 models for standard SAINT2 (blue), to the strict cut-off (purple), stepped cut-off (orange), stepped cut-off to optimal value (green) and stepped cut-off to optimal value with time restriction (pink) versions (as described in Table 2.3). Proportion of correct models produced (top left), the corrected running time in seconds (top right), enrichment of correct decoys produced for each method as compared to standard SAINT2 (represented by red line) (bottom left). . . . . 45
3.8 Looking at the decoys generated for each protein from the test set described in Table 2.3, using version 2. The TM-scores of decoys produced at each step-down of cut-off (top), the total number of complete decoys produced at each step-down in cut-off (middle), the proportion of correct decoys produced at each step-down in cut-off (bottom). . . . . 46
3.9 Looking at the decoys generated for each protein from the test set described in Table 2.3, using version 3. The TM-scores of decoys produced at each step-down of cut-off (top), the total number of complete decoys produced at each step-down in cut-off (middle), the proportion of correct decoys produced at each step-down in cut-off (bottom). . . . . 48
3.10 Length of the proteins in Table 3.2 and the time in seconds to generate 60 models on 60 cores (left). Time regression from linear model to describe the time in seconds to generate a model for a protein using the square of its length (right) . . . . . 49

3.11 Comparing the stepped cut-off to optimal value protocol with an added time restriction (purple) to no cut-off (blue), as described in Table 2.3. Proportion of correct decoys produced, with top TM-score of decoys labelled (top left); number of decoys produced that were correct (top right); the total number of decoys produced for each protein (middle left); running time in seconds (middle right); Enrichment of correct decoys for each protein (bottom left). . . . . 52
3.12 Minimum sample size for proteins of varying lengths so that best model is correct (left) or at least five models are correct (right), using version 3.2 (blue) and SAINT2 (red). . . . . 53
3.13 Comparison of 1,000 models generated with the new protocol (purple) with a sample of 1,000 models generated with standard SAINT2 (blue). Proportion of models produced that were correct (top), and zoomed in (second from top), the total number of correct models produced (middle), and zoomed in (second from bottom), and the enrichment of correct models produced using the improved protocol over standard SAINT2, for the 44 soluble proteins in the data set. Cases where standard SAINT2 produces no correct models, but the new protocol produces one or more correct models are marked with *. Cases where standard SAINT2 produces one or more correct models and the new protocol produce no correct models are marked with **. . . . . 58
3.14 Comparison of 10,000 models generated with the new protocol (purple) with 10,000 models generated with standard SAINT2 (blue). Proportion of models produced that were correct (top), and zoomed in (second from top), the total number of correct models produced (middle), and zoomed in (second from bottom), and the enrichment of correct models produced using the improved protocol over standard SAINT2, for the 44 soluble proteins in the data set. Cases where standard SAINT2 produces one or more correct models and the new protocol produce no correct models are marked with **. . . . . 60
3.15 The density of TM-scores for 1,000 models produced for all of the 245 proteins in the larger data set, produced using standard SAINT2 (red) and the new protocol (blue). Generated using the default kernel density estimation method in R. . . . . 61
3.16 The distribution of TM-scores of the best scoring model for each of the 245 proteins in the larger data set produced using standard SAINT2 compared to the new protocol. . . . . 62
3.17 The number of cases from the larger data set of 245 proteins where using the new protocol enriched the number of correct models produced (green), produced the same number of correct models (blue) and produced less correct models (red). . . . . 62

3.18 The distribution of lengths (top), $N_{eff}$ values (middle), and number of predicted contacts (bottom) for those cases the new protocol can enrich (green), produce the same number of correct models for (blue) and produces fewer correct models for (red). . . . . 63
3.19 The distribution of lengths (top), $N_{eff}$ values (middle), and number of predicted contacts (bottom) for those cases the new protocol can predict at least one correct model for (red), and cannot predict correctly (blue). . . . . 64
3.20 The number of cases the new SAINT2 can produce at lease one correct model for, at incrementally increasing run sizes, from 1 to 1,000 models. . . . . 65
3.21 The SCOP classes - $\alpha/\beta$ (green), $\alpha/\beta$ (blue), all $\beta$ (red), and all $\alpha$ (purple) for the cases from the larger dataset of 245 proteins where the new protocol enriched the number of correct models produced, produced the same number of correct predictions or produced less correct predictions. . . . . 67

4.1 Heatmap displaying the predicted contacts for 1AIU, ordered by position in the sequence, scaled based on the percentage of correct models with each contact satisfied, minus the percentage satisfaction in incorrect models for each extrusion length from a run of 10,000 decoys. Grey areas on the figure represent extrusion lengths where a contact is not yet available. . . . . 73
4.2 Heatmap displaying the predicted contacts for 1AIU, ordered by PPV (highest PPV at the bottom), scaled based on the percentage of correct models with each contact satisfied, minus the percentage satisfaction in incorrect models for each extrusion length from a run of 10,000 decoys. Grey areas on the figure represent extrusion lengths where a contact is not yet available. . . . . 74
4.3 Proportion of predicted contacts available that were satisfied at each 10% length of decoy generated. Data from 20 Decoys, 10 correct (red) and 10 incorrect (blue), for the soluble protein 1TZT; (top left) All contacts; (top right) Short-range contacts and (bottom) long-range contacts. . . . . 75
4.4 Lost contacts (Equation 4.1) at each 10% length of decoy generated. Data from 20 Decoys, 10 correct (red) and 10 incorrect(blue) for the soluble protein 1TZT. All contacts (top left); Short-range contacts (top right) and long-range contacts (bottom). . . . . 77
4.5 Heatmap displaying the predicted contacts for 1TZT, ordered from low to high PPV for 1,000 models, ordered from low to high TM-score. Predicted contacts marked in blue indicate satisfaction, white indicated unsatisfied. Models within the red box are incorrect and those outside the red box are correct. Predicted contacts marked with $*$ are incorrect. . . . . 79

4.6 Heatmap displaying the predicted contacts for 1TZT, ordered from early to late position in protein sequence for 1,000 models, ordered from low to high TM-score. Predicted contacts marked in blue indicate satisfaction, white indicated unsatisfied. Models within the red box are incorrect and those outside the red box are correct. Predicted contacts marked with ∗ are incorrect. . . . . 80
4.7 Heatmap displaying the predicted contacts for 1TZT, ordered from low to high residue range for 1,000 models, ordered from low to high TM-score. Predicted contacts marked in blue indicate satisfaction, white indicated unsatisfied. Models within the red box are incorrect and those outside the red box are correct. Predicted contacts marked with ∗ are incorrect. . . . . 81
4.8 Heatmap displaying the predicted contacts for 1TZT, ordered from low to high satisfaction for 1,000 models, ordered from low to high TM-score. Predicted contacts marked in blue indicate satisfaction, white indicated unsatisfied. Models within the red box are incorrect and those outside the red box are correct. Predicted contacts marked with ∗ are incorrect. . . . . 82
4.9 ROC curve varying the threshold of satisfaction rank of predicted contacts for all 245 proteins in the Large Dataset. . . . . 84
4.10 Boxplots describing the satisfaction ranks for correct predicted contacts (red) and incorrect predicted contacts (blue) for 1,000 models (left) and a filtered subset of models (right) for 245 proteins in the Large Dataset. . . . . 85

S1 TM scores of the decoys produced at each step of cut-off for 11 soluble proteins, using the stepped cut-off to optimal value protocol with an added time regression 99
S2 Total number of complete decoys produced at each step of cut-off for 11 soluble proteins, using the stepped cut-off to optimal value protocol with an added time regression . . . . . 100
S3 Proportion of correct models produced at each step of cut-off for 11 soluble proteins, using the stepped cut-off to optimal value protocol with an added time regression . . . . . 100

# 1

# Introduction

A major barrier in *de novo* protein structure prediction is the large amount of computational resources required to generate an accurate structure for one protein. This is especially the case with longer proteins, where there is significantly larger conformational space to sample. If we are to predict the structures of the many families with no structural information, we need to improve the efficiency of sampling techniques.

Determination of the structure of a protein is a key step in elucidating its function and how it may interact with other molecules, including potential drugs. Recent advances in DNA sequencing technologies mean that sequences can be obtained increasingly quickly and easily, leading to rapid expansion of sequence databases. Techniques for experimental determination of a protein structure such as NMR (Nuclear Magnetic Resonance) spectroscopy, electron microscopy and X-ray crystallography have improved over the last few decades, but remain

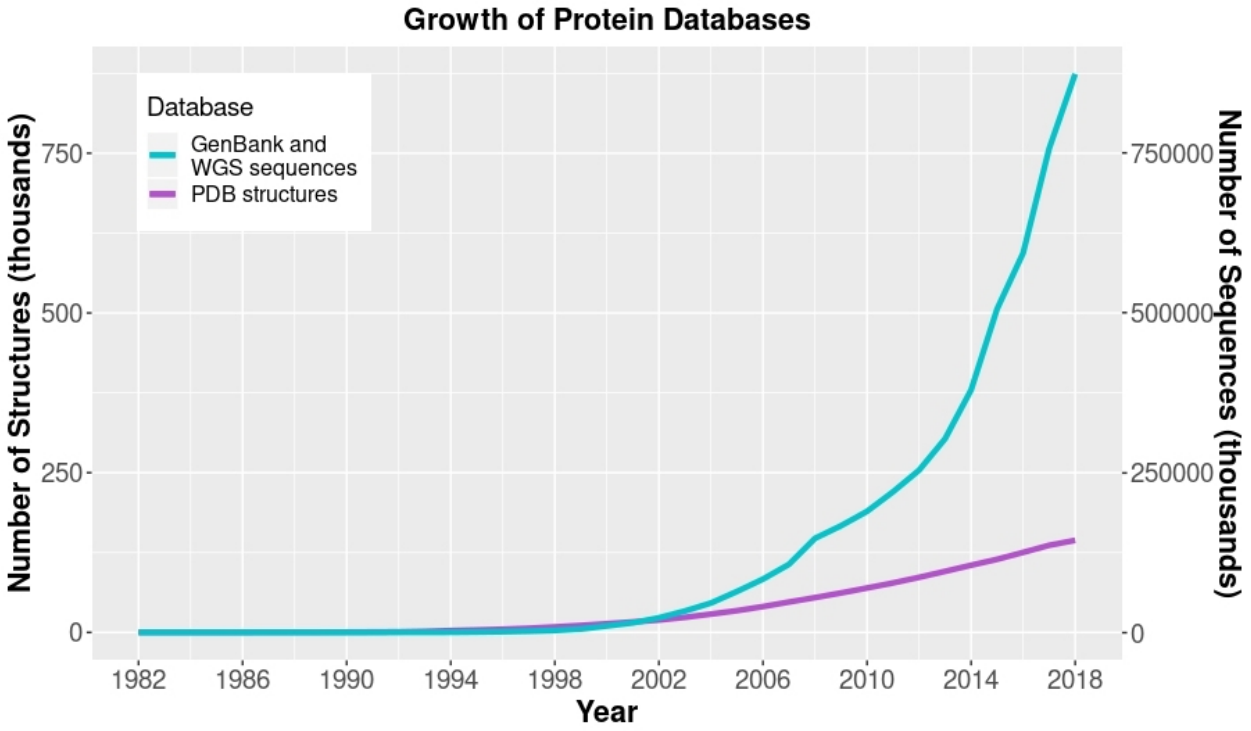



**Figure 1.1:** The growth of GenBank (including WGS project) and PDB databases from 1982-2018.

expensive, time-consuming and challenging [1]. As a result, there is a large disparity between the number of sequences in databases such as GenBank [2], and the number of published structures in the Protein Data Bank (PDB) - a public global archive for experimentally-determined, three-dimensional structures of biological macromolecules (Figure 1.1). The number of structures in the PDB currently lies at just over 130,000 [3], whereas GenBank currently contains over 208 million sequences, and that figure rises to over 850 million when sequences from the WGS (Whole Genome Shotgun) project are included. Sequencing technologies are continuing to advance, so these numbers are rapidly growing and the division between sequences and structures is only widening. Accurate protein structure prediction from sequence is one way that we can close this gap.

The most succesful approaches for predicting protein structures are template-based. Template-based ( or homology) modeling, can provide accurate structures for many proteins

[4], however this is reliant on the presence of existing structures (templates) with sufficient sequence similarity to and coverage of the target sequence. In template-based methods, the target and template sequences are aligned, and a structural framework built from the overlapping regions. The unaligned loop regions are then predicted to give an overall structure.

The Pfam database organises protein sequences into families, where the sequences are evolutionarily-related, and are thought to have a similar 3-D structure [5]. Currently, there are over 14,500 Pfam families. Of these, over 4,500 have experimentally determined structures; approximately 4,800 have homologous structures sufficient to build reliable models using template-based modeling. However, this leaves over 5,000 families with no structural information [6]; we are still relatively far from achieving the original aim of the Protein Structure Initiative [7] - to solve at least one structure for every protein family.

This means that for a large number of protein families, modelling has to be done in a template-free fashion, using *de novo* protein structure prediction methods. These methods do not rely on structural information from homologous proteins, but instead use information derived from physics-based and statistical observations relating to protein structures and/or folds. Until recently, template-free modeling was not consistently accurate; recent developments, notably the incorporation of predicted contacts from co-evolution in alignments of related sequences, have increased the accuracy of *de novo* structure prediction [4, 8], making it possible to reliably predict topologies of protein families of unknown structure.

The inclusion of constraints inferred from correlated mutations between residues in multiple sequence alignments has improved the field of *do novo* protein structure prediction. Increased information on residues likely to be in close proximity within a structure allows the conformational search space to be reduced, enabling the prediction of structures that were previously hard to predict, or even infeasible. Residues that are close in 3-D space within a protein are likely to undergo compensatory mutations in order to maintain biological function. Therefore, information on co-variance between residue positions within an alignment

of sequences can lead to predictions for residue-residue contacts. These evolutionary couplings can provide information about intra-protein residue interactions, as well as protein-protein interaction sites, identification of evolutionarily related proteins for use as templates and functionality of proteins [9]. Protein structure prediction methods such as EV-Fold [10], FRAGFOLD [11], ROSETTA [12] and SAINT2 [13] utilise contact prediction to improve the quality of the models they produce.

Contact prediction methods often yield many predicted interactions, and there have been multiple theories on how best to include these constraints [13–15]. Different methods produce different numbers of contacts; however, the more contacts that are considered, the less precise the predictions tend to be. It is important to include enough contacts to enable sufficient coverage of the target protein, however, including too many contacts could affect the quality of the fold as incorrect constraints are more likely to be used. A contact is defined as occuring between two residues, over five residues apart in sequence, with $C_\beta$ (or $C_\alpha$ in the case of glycine) atoms less than 8Å apart in 3-D space. The maximum number of feasible contacts per residue is $6.3L$, where $L$ is the length of the protein [16]. In reality, over 50% of structures in the PDB contain less than $2L$ contacts [17]. Methods for selecting the best number of contacts to use for modeling have included using one contact every 12 residues [14], using the top $L$ contacts (or some division of $L$) [15], or using a simple cut-off of the mutation correlation score to select those contacts most likely to be correct [13].

## 1.1 Project Aim

In this project, we aim to increase the number of protein families that the protein structure prediction software SAINT2 can predict correct structures for. We plan to do this by improving the efficiency of the modelling algorithm used, and the accuracy of the models (decoys) produced. For this, we utilise the predicted contacts of the target sequence in a

novel fashion, making use of the sequential nature of the algorithm. We investigate the information provided by the satisfaction of these predicted contacts in partially extruded decoys and explore whether it is possible to predict the quality of the final model using this information, and focus sampling efforts towards models most likely to be correct. We also investigate how contacts are satisfied during decoy generation and extrusion. If we can distinguish between correct and incorrect models during extrusion based on the specific contacts satisfied, using features such as the position in a sequence or the probability of being correct to evaluate each predicted contact, we may be able to more efficiently use computational time to sample the vast search space, or bias contact formation towards contacts more likely to produce better quality models.

## 1.2 *De novo* Protein Structure Prediction

While there are a number of methods for *de novo* protein structure prediction, the most successful approaches are fragment-based structure predictors - as exemplified by ROSETTA [4, 12]. These methods use a library of fragments (usually <20 residues long) compiled from known protein structures that share local sequence similarity to the target protein. Fragments are selected and assembled to create complete candidate structures, by optimising a pre-determined scoring function. Many thousands of decoys are built for each target [8](one of the causes of the large computational burden of *de novo* protein structure prediction). Models may then be refined to bring the structure closer to the native conformation. Once a number of candidate structural models have been generated, model quality assessment can then be can then be carried out to select the best models that will be most representative of the native conformation. Given the increasing complexity of protein structure prediction, community-wide efforts have been made to assess the state-of-the-art.

### 1.2.1 Critical Assessment of protein Structure Prediction

The Critical Assessment of protein Structure Prediction (CASP) experiments, are bienniel blind assessments, run to establish the current state of protein structure prediction, and highlight future areas of work. So far, there have been 12 CASP experiments. In recent CASP rounds, about 100 protein sequences whose structures have been experimentally solved, but not yet published, are released to all CASP participants. Structures are then predicted using methods from research groups worldwide. These blind predictions are independently assessed to determine the accuracy of each predictor tested. There are multiple categories in CASP. In our work, the category of interest is Template-Free Modeling (TFM), which assesses methods for models where there are no homologues with known structure.

It is relatively difficult to compare CASP results from year to year; the blind prediction cases have differing levels of difficulty - those selected for CASP11 had lower sequence similarity to known structures and higher radius of gyration (more non-globular targets are usually harder to model) [4] than targets from previous years. While the TFM category remained stagnantly inaccurate for many years, recent advances such as the addition of evolutionary couplings as contact predictions, have led to improvements in this category in the last few rounds, especially from 2012 onwards [4, 18]. Most success is usually seen for short, single-domain proteins [18] that come from families with sufficient sequence information, where the co-evolving residues can be predicted accurately [6]. Prediction of long proteins, multi-domain targets, and atypical folds remains challenging. This is mostly due to the large conformational space and the lack of accurate scoring functions for these cases.

### 1.2.2 Conformational Space

There are a large number of degrees of freedom in an unfolded polypeptide chain and it would take a physically impossible amount of time to sample all available conformations [19]. As the

size of a protein increases, so does the conformational search space ( a discrete representation of all of the possible conformations of a sequence of amino acids); this is one of the main reasons protein structure prediction fails for longer proteins [13]. It is therefore necessary to improve sampling heuristics in order to attempt to narrow and refine the conformational search space. A common way to restrict the search space is to use fragment libraries [20, 21] to restrict the search to folds that are known to be possible. Scoring functions direct the folding toward the more likely area of conformational space, based on statistical observations, or physics-based energy rules alongside predicted contacts [8, 9]. A Monte-Carlo sampling method is used in SAINT2 when predicting protein structure; introducing an element of randomness into perturbation of the chain by fragment replacement helps to avoid local minimum entrapment. Inaccurate or insufficient sampling heuristics can steer you away from the correct area of conformational space and therefore it is important to use an effective combination of heuristics.

### 1.2.3 Fragment Libraries

As described above, fragment-based approaches are the standard method for *de novo* structure prediction [22]. Fragments are selected from fragment libraries primarily based on their sequence similarity to local areas of target sequence - the underlying principle is that local amino acid sequence biases the chain towards a small number of shapes [23]. Fragment libraries are compiled from short segments (usually <20 residues) of known structures, often with hundreds of fragments representing each position in the target sequence. One limitation with the use of fragment libraries is that if the library is poor, or there are no fragments that represent structural segments that occur in the target structure, incorrect models will be generated [20].

The exact approach used to select fragments varies between different methods; NNMake (used by ROSETTA) outputs 9-residue fragments based on sequence similarity, predicted

secondary structure and predicted torsion angles [24]. FRAGFOLD uses super-secondary structure fragments (formed by taking two or three sequential secondary structures from a library of protein structures), selected using pair potential and solvation terms [11]. HHFold uses profile Hidden Markov Models, it does not emphasise fragment coverage, but concentrates on fragment precision [25]. FRazor builds on NNMake, but also incorporates solvent accessibility and contact capacity [26]. The method used with SAINT2 is Flib. This uses the predominant predicted secondary structure of fragments as well as sequence and predicted torsion angles to select fragments (see methods section for more details). As fragments are assembled to give candidate structures, these conformations are scored using a scoring function. As there are often hundreds of fragments per position in the sequence; it is not feasible to test all combinations. A heuristic search is therefore used, which relies on an accurate scoring function.

### 1.2.4 Scoring Functions

A perfect scoring function at its minimum value would represent the native state of the protein. Usually a combination of different potentials are used to measure how close a particular conformation is to the native state.

Scoring functions can evaluate the positions of all of the atoms in a structure, or a reduced representation of the protein. An example of a reduced representation involves including the heavy backbone atoms, and using only one atom to represent the side chains. The latter is particularly popular in *de novo* protein structure prediction as it saves computational time. Some methods, including ROSETTA, begin sampling using reduced representation to create the backbone structure and then switch to all-atom scoring to refine the final structure [27].

There are two main classes of scoring function: physics-based, and knowledge-based. Physics-based functions use the known energetics of molecular interactions; for example the Lennard-Jones potential [28] models the interactions between two molecules, addressing

attractive and repulsive forces.

Knowledge-based functions are currently more popular in *de novo* protein structure prediction methods; these are based on statistical observations from known structures present in the PDB. For example the Residue specific All-atom conditional Probability Function (RAPDF) [29], that scores the interactions between atoms based on statistical probabilities of the distances between residue pairs. Other knowledge-based potentials score features such as the backbone torsion angles, protein radius of gyration, and the angle of curvature between three residue fragments [30, 31].

### 1.2.5 Model Selection

Protein structure prediction methods usually implement some form of model ranking system after decoys have been generated to assess the quality of the models and identify those most likely to be correct. Some methods of model quality assessment involve using information from an individual model, for example, ranking based on the score from a scoring function. Other methods use information from multiple models, for example clustering based on pairwise structural similarity between decoys. The theory behind this approach is that larger clusters are more likely to contain the correct fold and thus the centre of the cluster is most likely to represent the highest quality model [32].

### 1.2.6 Refinement

As mentioned in the previous section, some protein structure prediction methods include a refinement stage, to bring an approximate initial model closer to the native structure. Refinement strategies can be categorised into three types [33]. Firstly, methods that refine local structures, for example by maximising side chain packing, or hydrogen bond interactions. Secondly, methods based on molecular dynamics (MD) that reorganise backbone atoms and

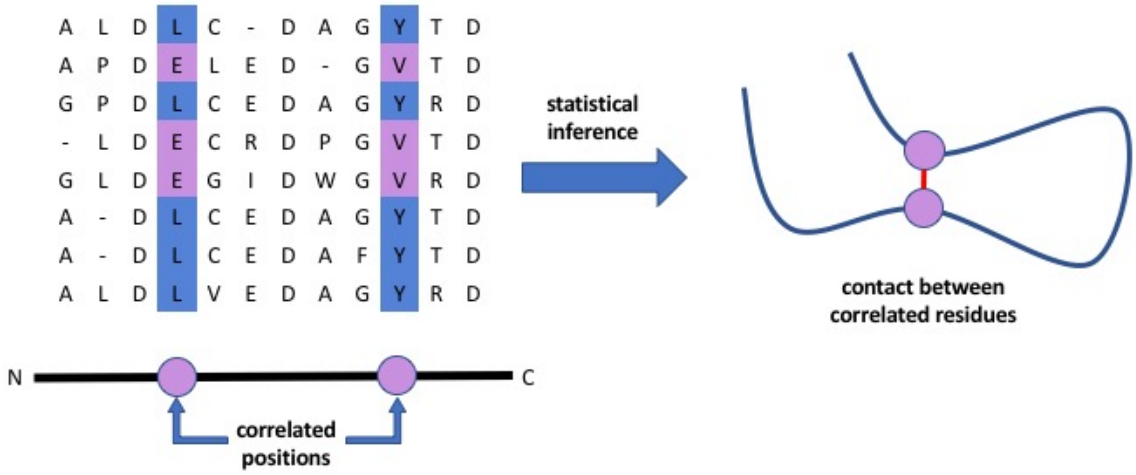



**Figure 1.2:** Prediction of residues in contact (within 8Å) in 3-D protein structure, from correlated mutations identified from a multiple sequence alignment.

thus the overall structure. Finally, there has been investigation into concentrating refinement on regions of large variability among models.

CASP8 was the first round to include a refinement assessment category and since then, refinement techniques have improved. In CASP12 MD-based techniques were most successful; the top refinement methods could consistently improve the quality of nearly all models tested, although improvement was small [34].

## 1.3 Evolutionary Couplings

*De novo* protein structure prediction has been revolutionised in recent years by the use of evolutionary couplings. Correlated mutations between positions in multiple sequence alignments (MSAs) of a protein family can be used to predict residue-residue contacts for a target sequence. The theory is that between interacting residues, if there is a mutation

in one residue then in order for the protein to maintain a functional structure, there must be a compensatory mutation in the second residue. Identifying co-variance, or correlations in mutations between two columns in a MSA as shown in Figure 1.2 can provide relatively accurate predictions when sufficient sequence information is available as to pairs of residues that interact in a sequence. These can then be used as constraints when modeling the structure of the target protein, narrowing the conformational search space.

### 1.3.1 Early Approaches

Some of the first contact prediction methods used mutual information (MI) [35] between positions in the sequence. After the creation of a MSA, the similarity of the residues present at each position is recorded in a set of $N$x$N$ matrices, where $N$ is the number of sequences in the MSA. These similarity scores are taken from a 20x20 matrix, with each value representing the similarity between two amino acids. The correlation coefficient is then measured between all of the combinations of two similarity matrices, down-weighting information from proteins with high sequence-similarity present in the MSA. This gives a set of mutational correlation scores, measuring the co-variation between two positions in the MSA (representing two residues). If the mutational correlation score for two positions exceeds a cut-off, then the residues are predicted to be in contact.

However, MI was found to be a relatively poor predictor of residue contacts [10]. MI treats correlated pairs of residues independently of each other so there is high recognition of transitive correlations which may result in contacts being predicted between two residues that are not close in 3-D space. For example, if residue $a$ interacts with residues $b$ and $c$ MI will also detect correlation between residues $b$ and $c$ and may predict they form a contact despite there being no direct interaction. Other reasons include the lack of information harvested from conserved positions, statistical noise caused by small numbers of sequences

utilised in MSAs, and phylogenetic bias [10].

### 1.3.2 Improved Co-evolutionary Coupling Algorithms

Newer contact prediction methods have overcome some of the problems encountered in early approaches. Correlated pairs of residues are treated as dependent on each other; when investigating the covariance between two columns, information from all of the other columns in the alignment is used to identify direct rather than indirect couplings.

Two methods that are different in terms of deriving and estimating correlations, but both use a maximum entropy model and a resulting inverse covariance matrix to describe the most likely set of direct couplings are mean field direct coupling analysis (mfDCA) [36], implemented in EV-Fold, and PSICOV [37], which is implemented in FRAGFOLD.

A different class of contact prediction uses pseudo-likelihood maximisation and attempts to learn the direct couplings as parameters of a Markov random field model. CCMPred is one implementation of this and appears to be faster to predict couplings for larger MSAs, but with smaller MSAs methods utilising estimation of inverse covariance matrices are faster [38].

In a comparison of contact prediction methods on a data set comprised of 3458 proteins, the average precision of the top L (where L is the length of the protein) contacts predicted by EV-Fold (an implementation of DCA) was around 38%, PSICOV was around 30% and CCMPred was the highest, at around 60% [8]. However, the contacts predicted by these methods were not always overlapping. Therefore, as is described below, methods have been developed to combine these predictions.

### 1.3.3 Metapredictors

Contact prediction methods have been found to generate significant numbers of non-overlapping predicted contacts [39]. Metapredictors have been created to combine these methods, in an

attempt to increase the number of correct residue-residue contact predictions. Two examples that have been found to be more accurate than individual predictors are metaPSICOV [40] and the PconsC family (PconsC [41], PconsC2 [42], PconsC3 [43]).

MetaPSICOV uses a combination of CCMPred, PSICOV, and an implementation of mfDCA, in a 2-stage predictor where the second stage filters the output of the first stage [40]. PconsC3 combines two DCA methods with a machine learning approach; it builds upon two previous methods: PConsC2 [42] and PConsC [32] by analysing residue pairs in the context of neighbouring residue pairs and improving contact predictions for smaller protein families.

These methods rely on the information gained from MSAs. Often, however, there are not enough sequences available in these alignments to accurately and reliably derive evolutionary relationships between residues. In an attempt to overcome this, RaptorX has been developed, which uses machine learning to produce a predicted contact map [44].

The average precision for the top L contacts (where L is the length of the protein) was found to be around 55% and 60% for metaPSICOV stage 1 and stage 2 respectively. The precision of the top L contacts predicted by PconsC was comparable to metaPSICOV stage 2 [8]. A separate study showed the precision of the top L contacts predicted by RaptorX, a machine learning protein structure prediction method, to be better than all of the above methods [44]. For example, RaptorX outperformed metaPSICOV by 20% when predicting long-range contacts, 7% when predicting medium-range contacts, and 3% when predicting short-range contacts on 105 CASP11 test proteins. However, at the time of writing this thesis, RaptorX is not freely available.

A study carried out in 2016 comparing a number of contact prediction methods (excluding RaptorX), shows that metaPSICOV stage 2 was most accurate when predicting contacts. However, contact predictions from metaPSICOV stage 1 and CCMPred most often led to the production of correct structural models [8]. This suggests that correct contact predictions do not automatically lead to accurate models; there are other factors that influence the

production of correct prediction of structures.

More recently, MetaPSICOV2 [45] has been released, with an improved neural net from MetaPSICOV. Results from CASP12 show improved performance, however it is thought that this increase in predictive power may be inflated due to the ease of the targets in the most recent CASP experiment.

### 1.3.4 Using Evolutionary Couplings to Predict Structure

Once evolutionary couplings have been calculated for a target protein, those that have a high co-variance score are assumed to represent residue-residue contacts. The distance between these pairs of residues can be constrained during model folding. Scoring potentials can reward the presence of these contacts, or penalise if these contacts are not present in a model as it folds.

One issue when using many current contact prediction methods is the lack of diverse sequences in a number of protein families. This means that for some families, especially smaller protein families, the MSAs assembled are not of sufficient quality to accurately derive covariation between residues and distinguish statistical noise.

$N_{eff}$ (the number of sequence clusters at an 80% sequence identity-clustering threshold divided by the square root of the protein length [46]) is a measure of sequence diversity; ROSETTA defines an $N_{eff}$ value of 32 or above as sufficient to predict contacts that produce fold-level accuracy in models [6], although other methods require more (or less) sequence information from the MSA.

Including sequences from metagenomic data in MSAs can greatly increase the number of protein families with sufficient sequence diversity for accurate contact prediction. In a recent study, the authors found that the proportion of protein families in the Pfam database [5] with a $N_{eff}$ value of 32 or above increased from 16% to 33% [6] when metagenomic data were incorporated. It is thought that an exponential increase in sequencing of metagenome data

will allow us to eventually be able to accurately predict structures for most protein families.

Evolutionary couplings and their presence in a model can provide information about its quality, both during folding and after. Usually, predicted contacts are applied to the full protein chain as conformational sampling occurs. However, using SAINT2, a sequential protein structure prediction method, we can use information from predicted contacts in a unique way.

## 1.4 Biologically Inspired Protein Structure Prediction: SAINT2

### 1.4.1 Cotranslational Folding

SAINT2 is a sequential, fragment-based *de novo* protein structure prediction method, based on the observation that some proteins fold cotranslationally (as they are extruded) in the cell, under the influence of translational machinery [47–50]. SAINT2 has been shown to be faster and less computationally intensive than other, non-sequential methods. When predicting the targets from the most recent CASP experiment, it could produce models of a similar quality to the most successful template-free predictor (Baker Group) [13]. However, the Baker Group selected their top models from a pool of hundred of thousands of predictions produced by their protocol - this is far in excess of the ten thousand models produced using the SAINT2 protocol.

Proteins are formed after the genetic code that makes up DNA is transcribed into mRNA, which in turn is translated to produce a chain of amino acids. This peptide chain then folds to give the final, functional 3-D protein structure. Post-translational modifications may then occur, such as the addition of other molecules including lipids to the surface of the protein. However, there is experimental evidence that in some cases proteins fold cotranslationally, as the chain is being extruded by the ribosome [51, 52]. The majority of current protein structure prediction methods do not take this into account and begin

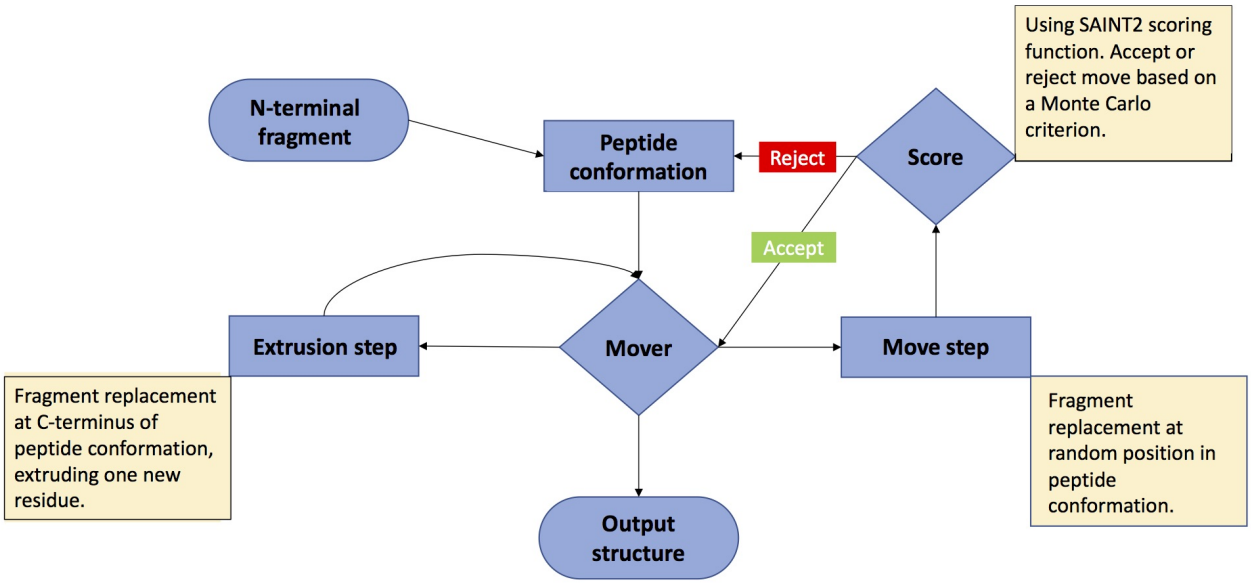



**Figure 1.3:** A flowchart describing how SAINT2 works to predict the structure of a protein given its sequence. An N-terminal fragment to represent the N-terminus of the sequence is used as input into SAINT2 and move steps or extrusion steps are then carried out until a complete model is produced.

conformational sampling with a complete extended chain of residues.

SAINT2 incorporates a sequential folding mechanism, and has been shown to improve conformational sampling, and structure prediction [13]. Sequential prediction allows the exploration of folding pathways that may not be possible when starting with a complete peptide chain; the N-terminus can enter local energy minima rather than immediately attempting to reach the overall global energy minimum.

### 1.4.2 How SAINT2 works

SAINT2 is a fragment-assembly *de novo* structure prediction method [13]. Figure 1.3 schematically describes how SAINT2 predicts protein structure. Its forward mode mimics sequential cotranslational folding, working from the N- to the C-terminus (the direction a protein is extruded from ribosomes as it is translated). It utilises a fragment from a fragment library of known structures to represent the N-terminus of the protein and it builds

upon this starting point to predict the full structure. At each step, there are two potential actions: to perform a move step or an extrusion. A move step assigns a fragment to a random position within the existing conformation and performs a fragment replacement; this new conformation is then either accepted or rejected based on its score. An extrusion step performs a random fragment replacement at the C-terminal end of the existing structure and simultaneously extrudes one new residue. Many move steps are performed between extrusions so that a wide range of folds can be explored. This allows the conformation to explore the energy minima at each stage of extrusion, as would happen with cotranslational protein folding. A demonstration of a decoy being produced is shown in Figure 1.4. Each snapshot shown is a further 10% of the decoy length.

SAINT2 builds models using the 3-D cartesian coordinates of the four main backbone atoms (C, N, O, $C_\alpha$), as well as $C_\beta$. SAINT2 does not explicitly consider side-chains, however $C_\beta$ atoms are included to represent the orientation of the side-chains. The backbone coordinates are calculated using torsion and bond angles from the fragment library, and completed using ideal bond lengths.

## 1.5 Outline

One of the major limitations of protein structure prediction is that accurate modeling, especially of longer proteins, requires vast computational resources due to the large conformational search space. For example, using SAINT2 to produce one model for a protein of 105 residues takes 41 seconds. For a protein of length 495 residues, SAINT2 takes 1094 seconds. When following the full SAINT2 protocol and generating 10,000 models for each target this difference in computational time for longer proteins becomes more significant.

In this project, I aim to improve the efficiency and accuracy of the protein structure prediction method SAINT2. The sequential nature of SAINT2 means that we can utilise

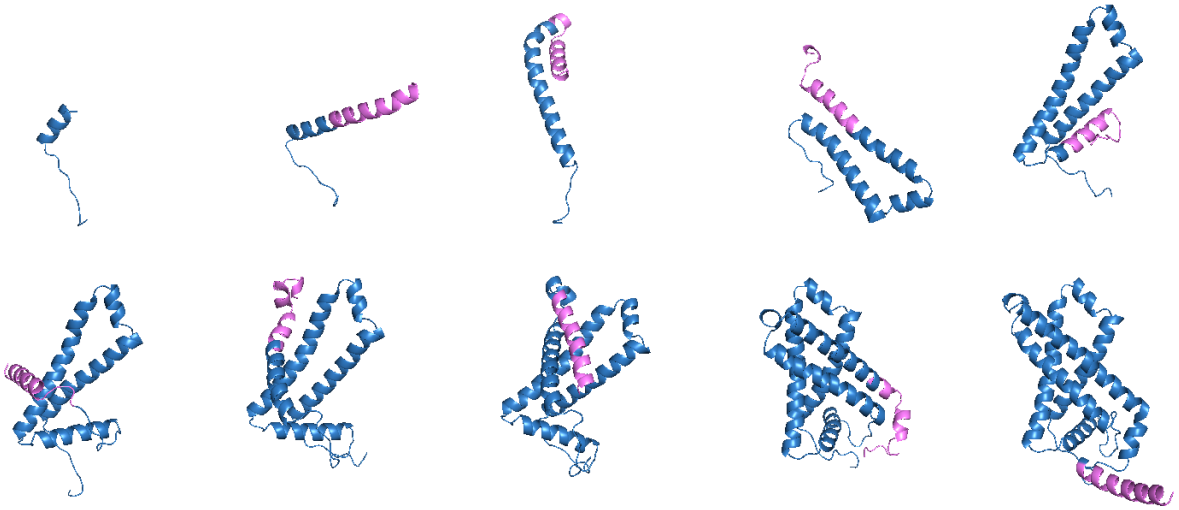


**Figure 1.4:** A cartoon representation of a model of the soluble protein 1AIU, produced sequentially by SAINT2, being extruded in increments of 10% of protein length. Each length of model is comprised of the previous, N-terminal fragment (blue), and the newly extruded residues (purple).

the information provided by predicted contacts in a unique way.

Previous research has shown that the proportion of predicted contacts that are satisfied in a final model can be indicative of the model quality [8]. Therefore incorrect models could be identified by evaluation of predicted contacts satisfied in the final structure. Here, we investigate whether it is possible to use the proportion of predicted contacts that are satisfied in partial models to predict the quality of the final model. This may allow the identification of incorrect models, which could be discarded, saving the computational time that would have been taken to build the remaining section.

Secondly, we study the behaviour of the formation and breaking of predicted contacts as models are built. This may give us insight into the mechanism of folding into or predicting a correct structure, and may also provide us with other markers to detect models likely to be incorrect early on in decoy generation.

If the efficiency of correct protein structure prediction by SAINT2 can be improved, we can use the computational time that has been saved to further sample conformational

search space. This will allow us to solve structures for more complicated proteins (notably, longer proteins where the computational time required to solve the structure is currently prohibitive), that at present have no structural information.

The third chapter of this thesis will explore how satisfied predicted contacts (SPCs) can be used to evaluate the quality of a model during its generation, and how we can use this to enrich the correct models produced in a run of SAINT2. In the fourth chapter, I will investigate how individual contacts are formed and broken in models as they are produced, and whether some predicted contacts are more important for the correct prediction of a protein structure than others.

# 2

# Methods for Improving the SAINT2 Search Strategy

## 2.1 Hardware

All simulations were performed on an Intel(R) Xeon(R) CPU E5-2699 v4 at 2.20GHz with 44 virtual cores and 0.49TB of RAM.

## 2.2 Definitions

A true contact is defined as two residues where their $C_\beta$ (or $C_\alpha$ in the case of glycine) atoms are less than 8Å apart in 3-D space.

Following [37], a long-range (LR) contact occurs between two amino acids that are over 23 residues apart in the sequence, whereas a short-range (SR) contact occurs between two amino acids that are from 5 to 23 residues apart in sequence.

We define the satisfied predicted contacts (SPCs) as the proportion of the possible predicted contacts that are present in either a partial or a full model. These are sub-classified as long-range SPCs (LR-SPCs) - the proportion of possible LR predicted contacts that are satisfied, and short-range SPCs (SR-SPCs) - the proportion of possible SR predicted contacts that are satisfied.

### 2.2.1 Model Quality Assessment

We used the Template Modeling score, or TM-score [53] to measure model quality, by comparing the predicted conformation to the native structure. TM-score is defined as:

$$\text{TM-score} = \max \left[ \frac{1}{L_N} \sum_{i=1}^{L_T} \frac{1}{1 + (\frac{d_i}{d_0})^2} \right] \tag{2.1}$$

Equation 2.1 gives the definition of TM-Score, where 'max' is the maximum value after optimal spatial superposition, $L_N$ is the length of the native structure, $L_T$ is the length of the aligned residues to the template structure, $i$ is a position in the alignment, $d_i$ is the distance between the $i$th pair of residues and $d_0$ is a scaling factor, calculated using $1.24\sqrt[3]{L_N - 15} - 1.8$.

Models with TM-score>0.5 are generally in the same topological classification and we refer to them as correct; models with TM-score$\leqslant$0.5 are mostly in different folds and we therefore refer to these as incorrect [53, 54].

## 2.3 Data Sets

Two main data sets were used during my project, with subsets of these proteins taken for further testing. Initially, a set of 44 structurally diverse soluble proteins and a set of 24 transmembrane proteins (shown in Table S2) was used. The soluble set was selected to evenly represent each of the four SCOP [55] protein classes and contained single chain, single domain proteins with an even distibution of lengths ranging from 54 to 504 residues. The transmembrane set contained polytopic $\alpha$-helical transmembrane chains taken from the Orientations of Proteins in Membranes (OPM) [56] database. Only unbroken chains were selected, with no soluble domains, forming a single transmembrane domain consisting of at least four helices. The chains selected ranged from 132 to 394 residues in length. The resulting data set had a maximum sequence identity of 20% between chains.

The methods used in this project to predict protein structure require a large amount of computational time: to generate a single model for 1AIU (a 105 residue protein) it takes 41 seconds. Therefore to generate 10,000 models over 4.5 days of computational time are required. The computational time needed for model generation increases exponentially with protein length, as the number of atoms to be considered by the scoring function increases. A longer protein, 1SMD, with 495 residues requires 1094 seconds to generate a single model, and over 125 days of computational time to generate 10,000 models. We therefore use small subsets of proteins from the data set in Table S2 for initial testing - these are described below.

### 2.3.1 Targets with Correct Models Data Set

The 12 soluble and 8 membrane proteins for which SAINT2 produced at least 20 correct models were selected to test the levels of predicted contacts that were satisfied during extrusion in correct and incorrect models. The soluble proteins in this set comprised of 3 $\alpha/\beta$, 3 $\alpha+\beta$, 1

| | PDB code | Length | Predicted Contacts | | Contact Precision (%) | SCOP class |
|---|---|---|---|---|---|---|
| | | | Long-range | Short-range | | |
| **Soluble** | 1AIU | 105 | 35 | 40 | 74 | $\alpha/\beta$ |
| | 1CSP | 67 | 15 | 43 | 73 | All $\beta$ |
| | 1EKG | 119 | 3 | 44 | 72 | $\alpha+\beta$ |
| | 1ENH | 54 | 0 | 10 | 100 | All $\alpha$ |
| | 1JWF | 139 | 3 | 11 | 100 | All $\alpha$ |
| | 1NAT | 119 | 72 | 55 | 88 | $\alpha/\beta$ |
| | 1SFE | 165 | 33 | 64 | 70 | All $\alpha$ |
| | 1VIN | 252 | 30 | 24 | 81 | All $\alpha$ |
| | 1WM3 | 72 | 9 | 31 | 83 | $\alpha+\beta$ |
| | 1XWY | 260 | 106 | 72 | 81 | $\alpha/\beta$ |
| | 1Z2U | 150 | 14 | 33 | 81 | $\alpha+\beta$ |
| | 2J9V | 98 | 0 | 2 | 50 | All $\alpha$ |
| **Membrane** | 1KQF_C | 216 | 46 | 14 | 72 | - |
| | 1ORS_C | 132 | 4 | 5 | 70 | - |
| | 2XOW_A | 179 | 38 | 13 | 82 | - |
| | 3KLY_A | 257 | 57 | 22 | 62 | - |
| | 3RLB_A | 176 | 36 | 19 | 82 | - |
| | 4B4A_A | 225 | 54 | 22 | 66 | - |
| | 4O6Y_A | 210 | 37 | 19 | 59 | - |
| | 4OD5_A | 274 | 58 | 30 | 80 | - |

**Table 2.1:** PDB code, length, number of short-range and long-range predicted contacts, precision of the predicted contacts, and SCOP class for each of the soluble and membrane proteins that generated correct decoys.

all $\beta$ and 5 all $\alpha$. These proteins are shown in Table 2.1, along with the numbers of predicted LR and SR predicted contacts, the precision of these contacts, and the SCOP class.

### 2.3.2 100+ Residues Data Set

This data set was comprised of the soluble proteins from Table S2 with length over 100 residues. This included 39 proteins, and maintained a relatively even spread over the SCOP classes.

### 2.3.3 N-terminal Contacts Data Set

We selected the proteins which had at least 10 predicted contacts in the N-terminal half, from the original test set of 44 soluble proteins in Table S2. Cases where SAINT2 had not produced enough correct decoys to give a reliable density plot were omitted. This gave 21 proteins: 9 $\alpha/\beta$, 6 $\alpha + \beta$, 3 all $\beta$ and 3 all $\alpha$.

### 2.3.4 Testing cut-off versions

To test different versions of our method, we subdivided the proteins from the data set described in Table S2 into three different categories:

- Proteins that SAINT2 predicted well (100 or more correct models from a run of 10,000). This set is shown in Table 2.1.
- Proteins that SAINT2 predicted a few correct models for (between 0 and 99 correct models from a run of 10,000).
- Proteins for which SAINT2 did not correctly predict the structure.

#### 2.3.4.1 Reduced Correct Set

The Reduced Correct data set is shown in Table 2.2. It is comprised of a selection of proteins from the first two sets described above: proteins that SAINT2 predicted well and proteins it only predicted a few correct models for. This is a small data set of five proteins, it is intended to give an idea as to the effect of any modifications to SAINT2 while not consuming too much computational time.

| PDB code | SCOP class | Length | Full protein | | | 50% extrusion | | |
|---|---|---|---|---|---|---|---|---|
| | | | All | LR | SR | All | LR | SR |
| 1NAT | $\alpha/\beta$ | 119 | 88 | 90 | 87 | 95 | 100 | 89 |
| 1SFE | All $\alpha$ | 165 | 70 | 66 | 72 | 80 | 76 | 82 |
| 1SL8 | All $\alpha$ | 191 | 65 | 87 | 57 | 52 | 84 | 36 |
| 1VIN | All $\alpha$ | 252 | 81 | 86 | 75 | 79 | 80 | 77 |
| 206L | $\alpha+\beta$ | 164 | 53 | 42 | 64 | 56 | 44 | 75 |

**Table 2.2:** Precision of predicted contacts (All contacts, long-range contacts, short-range contacts) for the full protein, and first 50% extruded for the proteins in the reduced correct data set.

### 2.3.5 Good and Bad Proteins Data Set

We selected 14 soluble proteins from the original data set of 44 proteins; 7 proteins that SAINT2 could predict correct structures for and 7 that SAINT2 could not predict correct structures for.

The 7 good cases (where correct structures could be predicted) were: 1VIN, 1SFE, 1SL8, 1VL1, 1FTG, 1XWY and 1JKS.

The 7 bad cases (where no correct structures could be predicted) were: 1SDI, 1SQW, 1W66, 2HVM, 1MSK, 1BMG and 1NEP.

## 2.4 Larger Dataset

We needed to test the new versions of SAINT2 on more proteins and so also used a second, larger data set (Table S3), comprised of 245 structurally diverse proteins ranging from 50 to 249 residues. This set also contained a higher proportion of proteins that SAINT2 could predict correctly, and thus was useful when investigating how individual contacts were formed and broken. There were 59 proteins classified as $\alpha/\beta$, 47 $\alpha+\beta$, 65 all $\beta$ and 74 all $\alpha$. For each of these proteins, 1,000 models were generated using standard SAINT2, and version 3.2 (see section 2.6.4) with no time regression.

## 2.5 SAINT2

SAINT2 is a fragment-based *de novo* protein structure prediction method. A target sequence is taken as input, PSIPRED [57] is used to generate secondary structure predictions, SPINE-X [58] generates torsion angle predictions, Flib [20] produces a fragment library and metaPSICOV [40] is used to generate residue-residue contact predictions. The output from these steps is then used by SAINT2 (here used in Forward mode) to predict structures for the target sequence.

A fragment is selected from the fragment library to represent the N-terminus of the protein. A linear mover carries out 10,000 moves as the model is being extruded; the number of moves performed between extrusion increases in a linear fashion as length increases. A move step involves a random fragment replacement at any position in the extruded model. This new conformation is then scored using the SAINT2 scoring function and either accepted or rejected based on this score using a Monte Carlo strategy. An extrusion step involves a fragment replacement at the C-terminus of the extruded length of model; this is always accepted and results in the extrusion of one new residue. After these initial 10,000 moves and full extrusion of the model, a further 1,000 random fragment replacement moves are carried out.

Structures are represented using the Cartesian co-ordinates for the four main backbone atoms (C, N, O, $C_\alpha$), as well as $C_\beta$ to represent the side-chain orientation. These co-ordinates are calculated by combining fragments that represent segments of structure of known proteins, from the fragment library and are completed using ideal bond lengths.

### 2.5.1 Secondary structure and torsion angle prediction

The first stage of SAINT2 involves using PSIPRED V3 [57], a two-stage neural network, to predict secondary structure for the input sequence. A multiple sequence alignment is generated by PSI-BLAST [59] using the following parameters: inclusion E-value threshold =

0.001, number of iterations = 3. This MSA is then used by PSIPRED, along with default weights, to generate secondary structure predictions.

SPINE-X (version 2.0) [58], a neural network that combines information from a position specific scoring matrix and several physical parameters, is used to predict torsion angles from the target sequence.

### 2.5.2 Fragment Library generation

Information from the initial stages of SAINT2 (secondary structure and torsion angle predictions) is used to generate a fragment library using Flib [20]. Flib extracts around 30 fragments per target position, each of 6-20 residues in length from a database of known structures. The database used is a subset of the PDB which is non-redundant (with sequence identity $< 90\%$) and high quality (with resolution $< 2.5$ Å). Fragment selection is based on sequence similarity to the target, predicted secondary structure and predicted torsion angles.

### 2.5.3 Residue-residue contact prediction

MetaPSICOV [40] is used to generate residue-residue contact predictions for target sequences. MetaPSICOV combines three different contact predictors: PSICOV [37], mfDCA (implemented in FreeContact[60]) and CCMPred [38]. Only predictions that were output with an estimated precision (PPV) of 0.5 or above were considered.

The accuracy of contact predictions generated using metaPSICOV depends on the size and variety of sequences in the MSA. $N_{eff}$ was used as a measure of sequence diversity and is calculated as shown in Equation 2.2 below:

$$N_{eff} = \frac{\text{no. sequence clusters at 80\% threshold}}{\sqrt{\text{protein length}}} \tag{2.2}$$

Sequence clustering is carried out using a greedy algorithm that compares the Hamming

distances of sequences and forms clusters based on the similarity between sequences. It is, however, dependent on the order of sequences within the MSA, as each sequence will be investigated in order. An $N_{eff}$ value of 32 or above is usually indicative of sufficient accuracy of predicted contacts to generate fold-level accuracy in models [6].

### 2.5.4 Scoring Function

The scoring function used in SAINT2 contains a weighted combination of knowledge-based and physics-based potentials, normalised based on protein length. The potentials included are: Lennard-Jones potential; RAPDF (residue-specific all-atom conditional probability discriminatory function); solvation potential; predicted residue-residue contact potential and predicted secondary structure potential.

## 2.6 Implementation of Satisfied Predicted Contacts cut-off

Four different versions of SAINT2 were tested (described in Table 2.3) and compared to the original version to investigate the best way to implement our SPC cut-off. The cut-off implemented was based on the proportion of predicted LR-SPCs available in the extruded length of decoy that were satisfied. As shown in Figure 2.1, partial models were evaluated for SPCs and those that did not pass the cut-off were discarded.

| Method | Description | Dataset | No. decoys per protein | Time |
|---|---|---|---|---|
| SAINT2 | no cut-off | all proteins in Table 2.1 | 10,000 | 24h |
| Version 1 | cut-off of 30% LR-SPCs at 50% extrusion | Reduced Correct | 10,000 | 24h |
| Version 2 | cut-off steps down from 100% to 0% | Reduced Correct | 1,000 | 24h |
| Version 3 | cut-off steps down from 100% to optimum value | Reduced Correct | 1,000 | 24h |
| Version 3.2 | cut-off steps down from 80% to optimum, with time restriction | Reduced Correct | 10,000 | based on time regression |

**Table 2.3:** A brief description of each version of SAINT2 that was tested, which protein data set was tested, the number of decoys produced per protein, and the maximum time limit implemented.

### 2.6.1 Version 1: strict cut-off

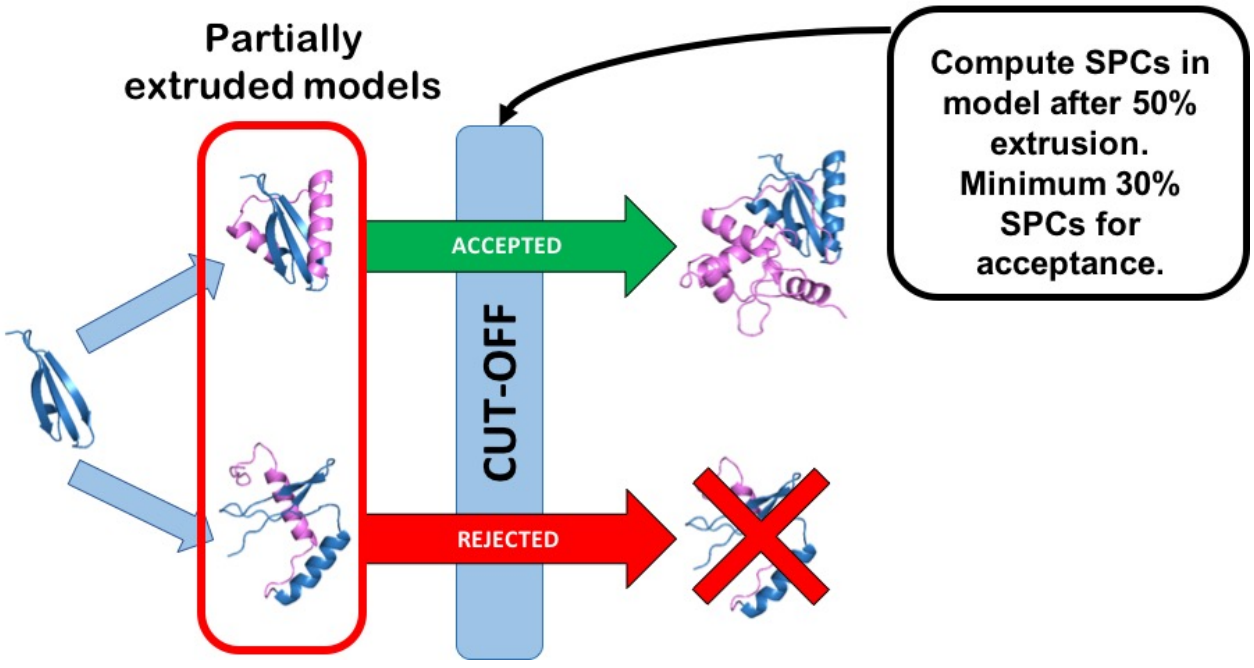



**Figure 2.1:** Version 1 of our cut-off implemented into SAINT2 to filter out partial models predicted to lead to incorrect final structures.

In version 1, a strict cut-off was added at 50% model extrusion, with a value of 30% LR-SPCs - models with LR-SPCs under this value were discarded.

1: **while** [no. complete decoys < 10,000] and [running time < 24 hours] **do**
2: start SAINT2 run
3: **if** LR-SPCs < 30% at 50% decoy extrusion **then**
4: end SAINT2 run
5: **else**
6: continue to generate full decoy
7: end SAINT2 run
8: **end if**
9: **end while**

### 2.6.2 Version 2: stepped cut-off

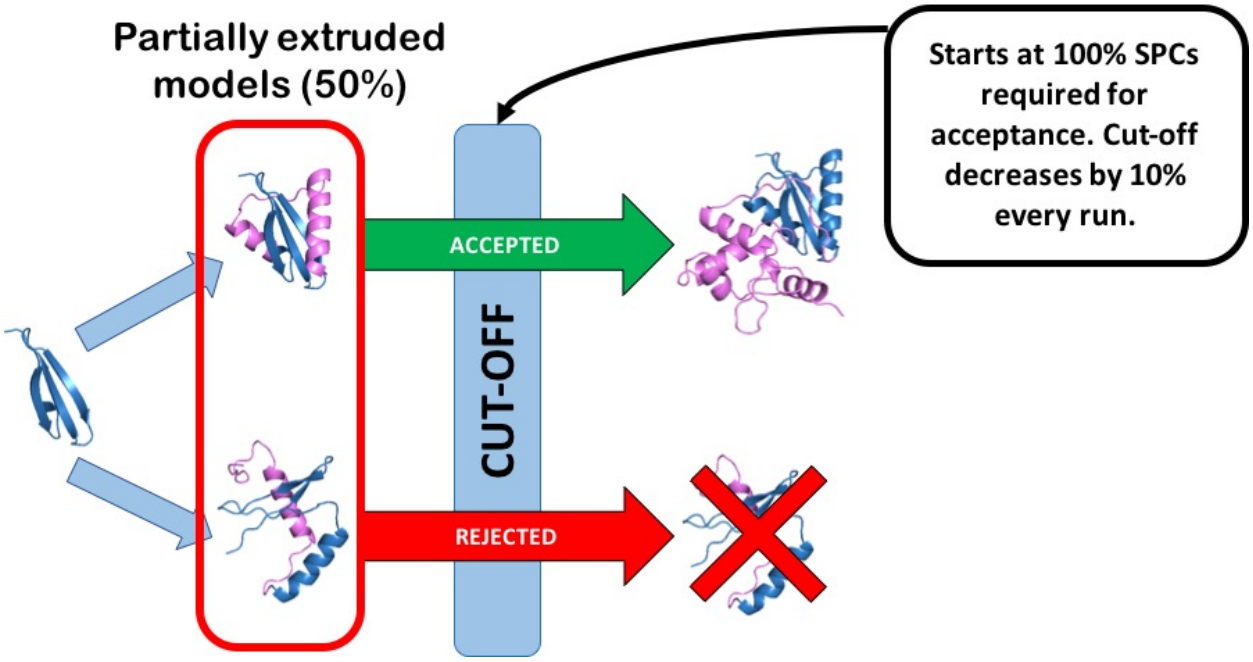



**Figure 2.2:** Version 2 of our cut-off implemented into SAINT2 to filter out partial models predicted to lead to incorrect final structures.

In order to optimise enrichment of correct decoys for individual proteins, in version 2 we modified the cut-off so that it stepped down from 100% LR-SPCs by decrements of 10%.

1: cut-off = 100%
2: **while** [no. complete decoys < 1,000] and [running time < 24 hours] **do**
3: start SAINT2 run for min(1000, 1000 - no. full decoys)
4: **if** LR-SPCs < cut-off at 50% decoy production **then**
5: end SAINT2 run
6: **else**
7: continue to generate full decoy
8: end SAINT2 run
9: **end if**
10: decrease cut-off by 10%▷ Cut-off loops down by 10% each round of runs, until 0% is reached
11: **end while**

### 2.6.3 Version 3: stepped cut-off to optimum

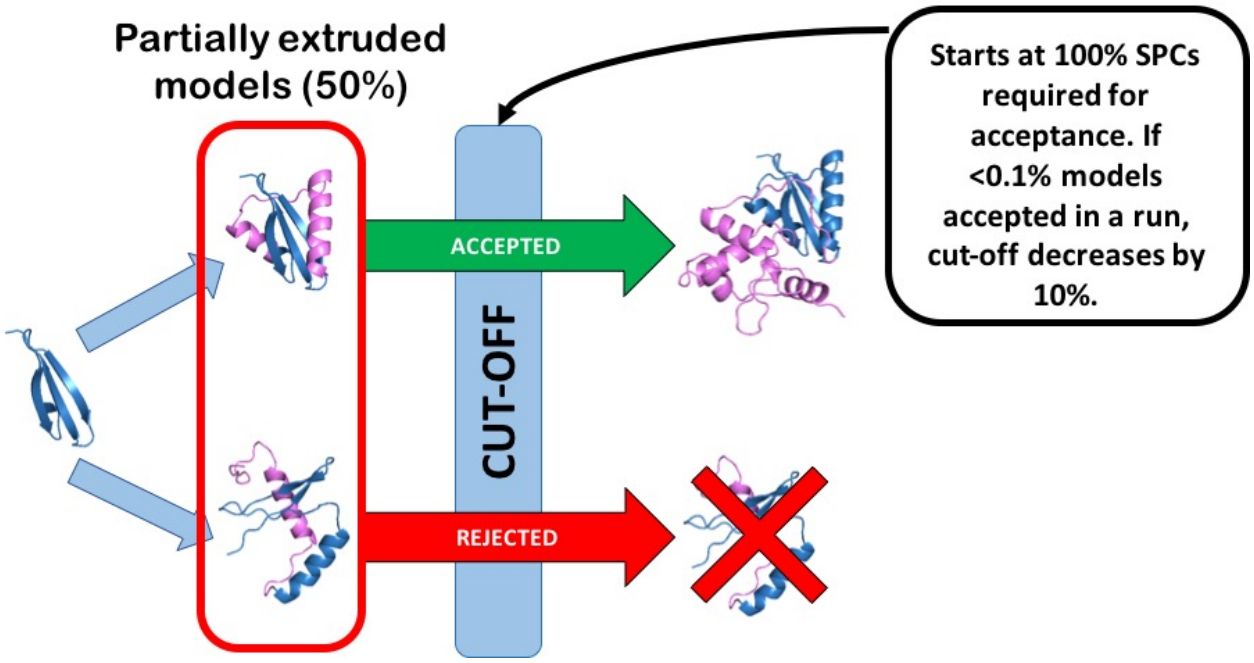



**Figure 2.3:** Version 3 of our cut-off implemented into SAINT2 to filter out partial models predicted to lead to incorrect final structures.

In version 3 we further built upon this stepped cut-off by only reducing the cut-off value while the proportion of models accepted by the cut-off was low. Once a level was reached where an acceptable proportion of attempted models passed the cut-off checkpoint and were completed, the cut-off remained at this value.

1: cut-off = 100%
2: **while** [no. complete decoys < 1,000] and [running time < 24 hours] **do**
3: start SAINT2 run for min(1000, 1000 - no. full decoys)
4: **if** LR-SPCs < cut-off at 50% decoy production **then**
5: end SAINT2 run
6: **else**
7: continue to generate full decoy

8: end SAINT2 run

9: **end if**

10: **if** no. complete decoys produced this round of SAINT2 runs $< 10$ **then**

11: decrease cut-off by 10% ▷ If more than 10 full decoys produced this round of SAINT2 runs, stay at this 'optimal' cut-off

12: **end if**

13: **end while**

### 2.6.4 Version 3.2: stepped down cut-off to optimum with time restriction

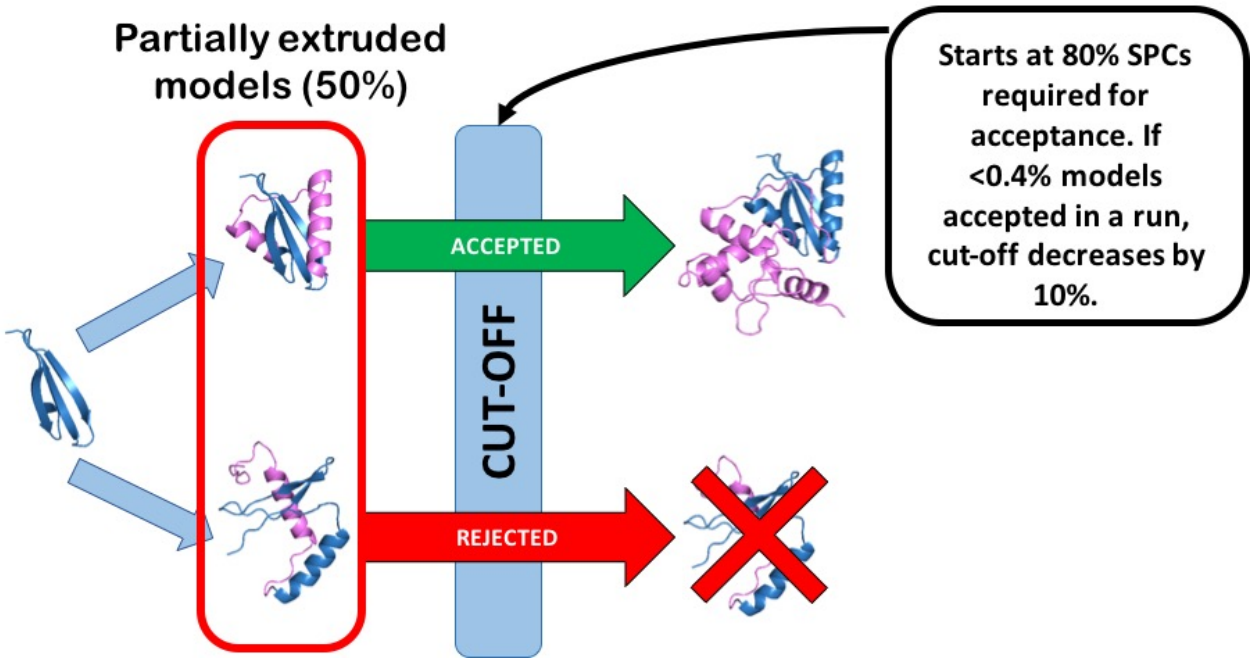



**Figure 2.4:** Version 3.2 of our cut-off implemented into SAINT2 to filter out partial models predicted to lead to incorrect final structures.

The final version that we tested (stepped to optimum cut-off with time restriction) attempted to speed up the algorithm by starting the cut-off at a lower value than previously tested, and

stepping down from there to an optimum value. We also implemented a time limit based on the time it would take SAINT2 to generate a decoy for a protein of equivalent length to the target.

1: cut-off = 80%
2: TIME = time for SAINT2 to produce 10,000 decoys for protein of equivalent length to target
3: **while** [no. complete decoys < 10,000] and [running time < TIME] **do**
4: start SAINT2 run for min(1000, 1000 - no. full decoys)
5: **if** LR-SPCs < cut-off at 50% decoy production **then**
6: end SAINT2 run
7: **else**
8: continue to generate full decoy
9: end SAINT2 run
10: **end if**
11: **if** no. complete decoys produced this round of SAINT2 runs < 40 **then**
12: decrease cut-off by 10% ▷ If more than 10 full decoys produced this round of SAINT2 runs, stay at this 'optimal' cut-off
13: **end if**
14: **end while**

# 3

# Improving the SAINT2 Search Strategy

Despite recent improvements in protein structure determination and prediction methods, there remains a large proportion of protein families with no representative structure. The computational time required to adequately sample the conformational space, especially for longer proteins, is infeasibly large and current methods of sampling are not sufficient to narrow this search space.

One strategy developed over recent years in an attempt to narrow the search space has involved the use of predicted contacts. However, as contact prediction accuracy improves with the introduction of new methods, we do not always see a corresponding increase in the quality of structure prediction [43]. We investigated how information from predicted contacts could be used to improve protein structure prediction, in particular improving the efficiency and speed when producing a correct model.

| Decoy extrusion (%) | Time (secs) |
|---|---|
| 10 | ∼0 |
| 20 | ∼0 |
| 30 | ∼0 |
| 40 | 1 |
| 50 | 2 |
| 60 | 5 |
| 70 | 9 |
| 80 | 16 |
| 90 | 26 |
| 100 | 41 |

**Table 3.1:** Time (seconds) for each snapshot (taken every 10% length of the final model) to be produced during generation of one model for the protein 1AIU.

Previously it has been shown that the satisfied predicted contacts (the proportion of available predicted contacts that are present) in a final model can be used as an indicator of the quality of the model [8]. Using sequential protein structure prediction, we can use the information from predicted contacts uniquely, evaluating the presence of contacts during extrusion of a model.

We first tested whether satisfied predicted contacts in partial models could be used as an indicator of the quality of the final model. We then used this information to investigate and implement a cut-off during extrusion, to identify and discard models that were likely to be incorrect in order to avoid using unnecessary computational time.

## 3.1 Model Generation Timings

If computational time can be used more efficiently when sampling conformational space, this could allow the prediction of more complicated protein structures, for families that currently have no solved structures. To demonstrate the potential time savings that could be made by discarding an incorrect decoy during generation, we calculated the time taken by SAINT2

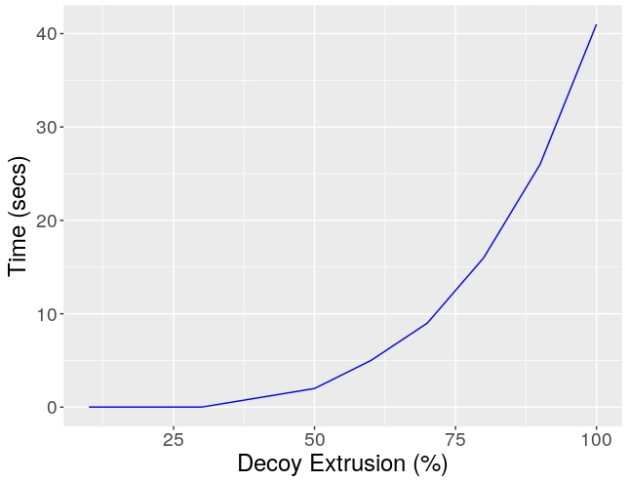



**Figure 3.1:** Time in seconds to generate one model as extrusion occurs for 1AIU.

to produce different snapshots (at 10% length intervals of the full decoy) for the soluble protein 1AIU. Table 3.1 and Figure 3.1 show that as decoy extrusion increases, the time for the next snapshot to be generated increases quadratically. The scoring function used in SAINT2 scores each atom against every other atom in the structure; as the length of the protein increases, as does the number of atoms that are scored, and thus the time required for scoring. Therefore, if we can predict that a run of SAINT2 is likely to lead to an incorrect model, the computational time saved by not completing this model could be large. Even if the quality of models produced does not improve, if we discard decoys at 50%, we would save 39 seconds per model discarded, a speed up of 20-fold in this case.

Producing 10,000 models for the protein 1AIU using standard SAINT2 takes roughly 113 CPU hours. Applying a strict cut-off at 50% extrusion that required 30% SPCs to be present would discard enough partially complete models to save roughly 3 hours. increasing this cut-off value to 50% SPCs would also increase the time saved to 13 hours. 1AIU is a short protein that is relatively easy to predict; therefore these time savings would likely be higher with longer and more complicated proteins as less predicted contacts are satisfied.

## 3.2 Satisfied Predicted Contacts during Decoy Generation

If we can identify incorrect decoys as they are being produced, we can utilise the computational time saved, by better searching conformational space and improve the performance of structure prediction. Given previous results that show that SPCs could be used to discriminate between correct and incorrect final models, we assessed if this was also true for partial conformations.

I generated 10,000 models for the Targets with Correct Models data set (Table 2.1). For each model, a snapshot was taken just before extrusion at 10% length intervals. Here, we look at All, long-range (LR), and short-range (SR) SPCs. Long-range contacts have previously been identified as being more useful for identifying whether final models are correct [8] but this may not be the case for partial models. We wanted to investigate whether partial correct models satisfied more predicted contacts than partial incorrect models. We considered membrane and soluble proteins separately as SAINT2 is generally worse at predicting membrane proteins than soluble proteins [8].

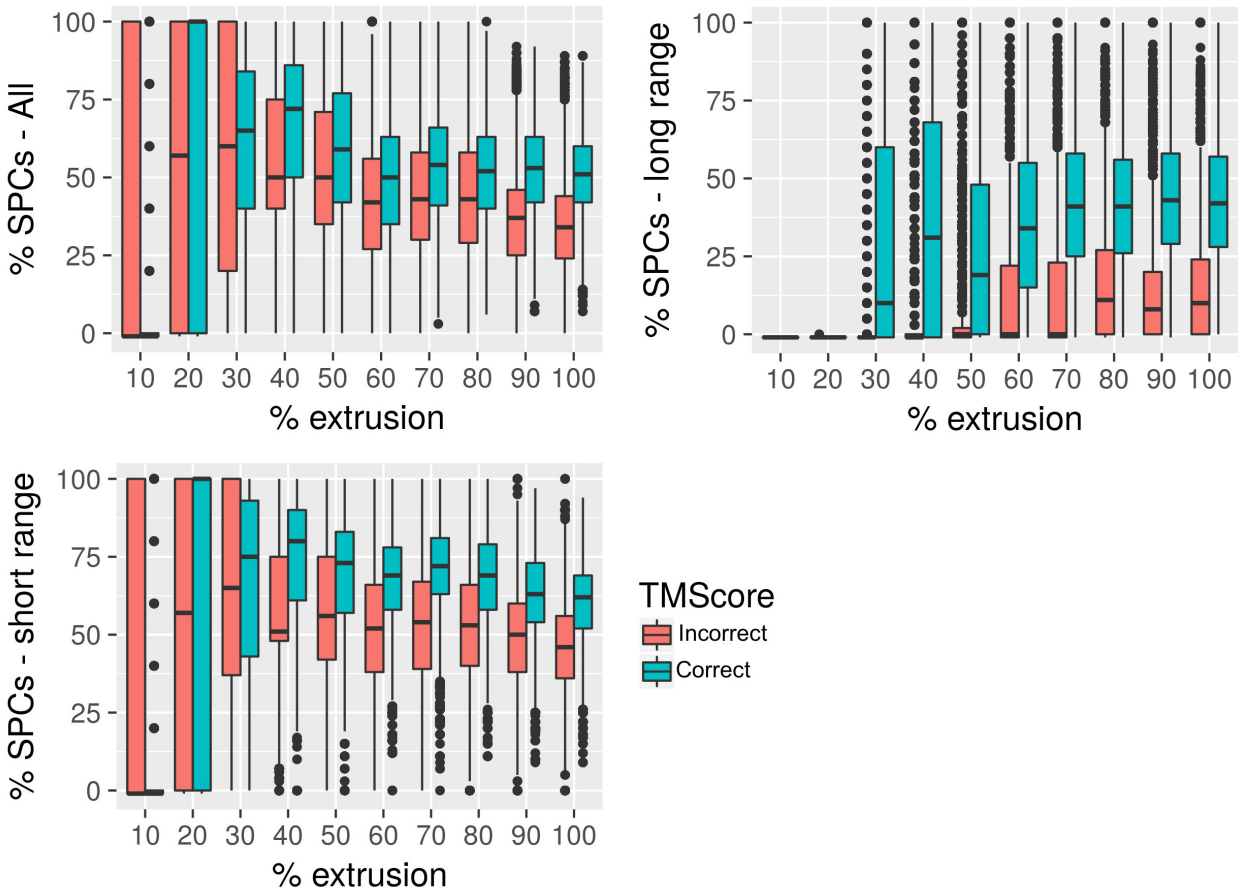



**Figure 3.2:** Satisfied Predicted Contacts at each percentage length for 10,000 decoys for the Targets with Correct Models data set, split into correct (blue) and incorrect (red) decoys. (top left) Total predicted contacts satisfied. (top right) LR predicted contacts only. (bottom left) SR contacts only.

Looking at the soluble proteins in the Targets with Correct Models data set, we see separation in SPC levels between correct and incorrect models. Correct models have higher SPCs and short-range SPCs at all lengths of extrusion, however looking exclusively at long-range SPCs, this distinction between correct and incorrect models is larger, but only observed from 30% extrusion. Below this extrusion level many proteins have few or zero predicted LR contacts; the shortest protein in the data set contains 54 residues overall, and therefore only 10 residues at 20% extrusion. LR contacts are between residues greater than 23 residues apart and thus are impossible at this length. Interestingly, at 50% extrusion there is a drop

in long-range SPCs; this may be attributed to an increase in predicted contacts available at that extrusion length. For example, 1NAT has 16 long-range contacts available at 40%, but this almost doubles, with 31 long-range contacts available at 50% extrusion.

We also analysed SPCs in the membrane proteins from the Targets with Correct Models data set (Table 2.1). For these simulations we generated 10,000 models per protein.

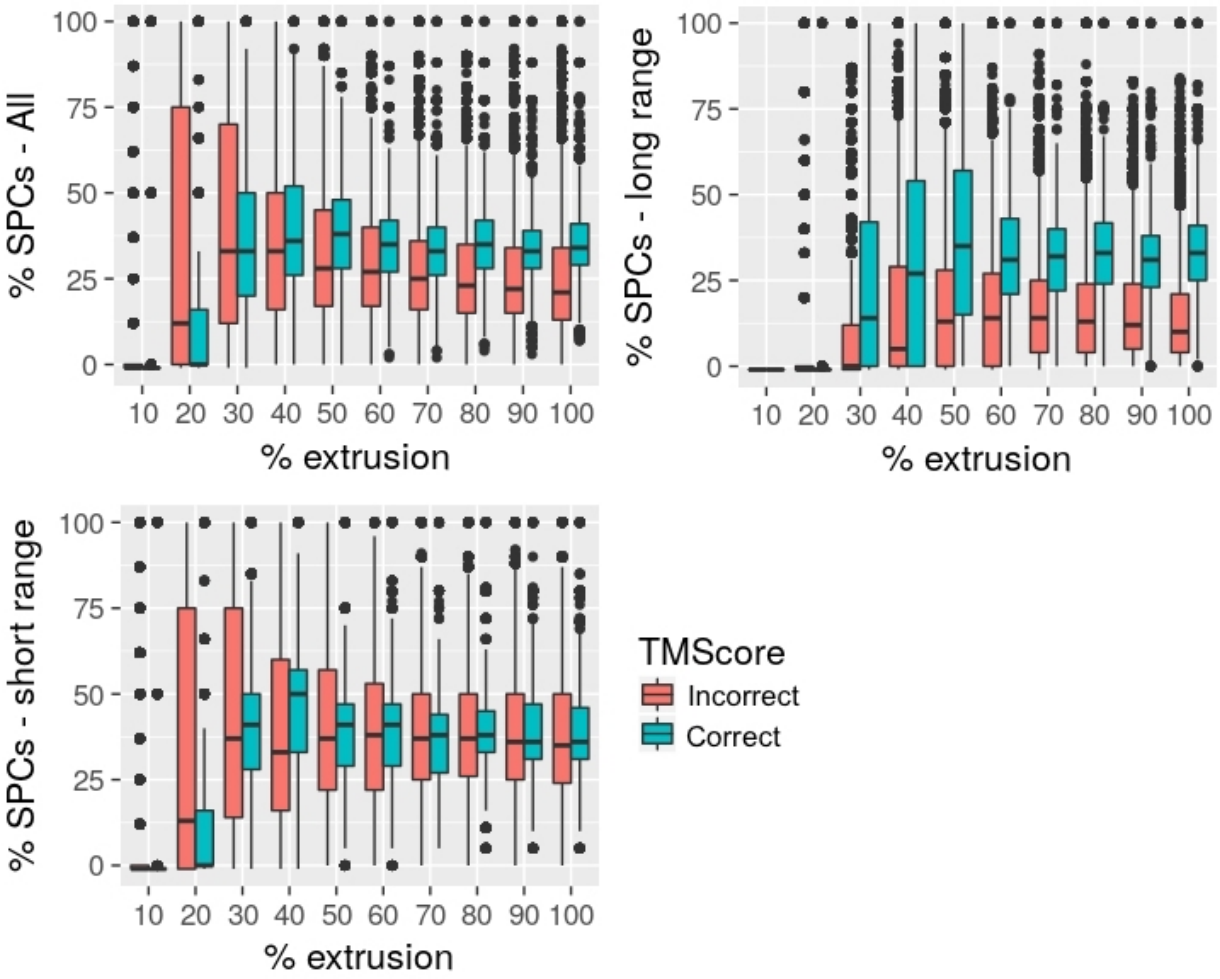



**Figure 3.3:** Satisfied Predicted Contacts at each % length for 10,000 decoys for the all membrane proteins in the data set, split into correct (blue) and incorrect (red) decoys. (top left) Total predicted contacts satisfied. (top right) LR predicted contacts only. (bottom left) SR contacts only.

Evaluating SPCs during membrane protein model extrusion (Figure 3.3) shows that there is little difference in short-range SPCs between correct and incorrect final models. This

is perhaps not surprising, given that this set is comprised of $\alpha$-helical bundles and most short-range contacts are captured by secondary structure. Comparing the soluble cases to membrane cases, SPCs are generally higher in soluble models; this is particularly true for short-range SPCs. We see a distinction between long-range SPCs in correct and incorrect models, especially from 50% extrusion; correct models generally have higher long-range SPCs than incorrect models. This suggests that evaluating the long-range SPCs of models at 50% extrusion may allow us to predict whether the final model will be correct.

## 3.3 Design of an SPC cut-off

### 3.3.1 Identification of suitable cut-off value for SPCs

The results shown above suggest that at 50% extrusion, we can predict whether the complete model will be correct based on the LR-SPCs. To explore this feature in structure prediction, we assessed whether there is a suitable cut-off value where models with less than a given percentage of LR-SPCs are likely to be incorrect. We looked at the density distributions of correct and incorrect decoys with varying levels of LR-SPCs at 50% extrusion.

Figure 3.4 shows that it is possible to identify a cut-off value that would lead to the separation of the majority of incorrect decoys from correct decoys. For LR-SPCs this cut-off value is around 30%.

The proteins in our set cover a range of lengths, numbers of predicted contacts and precision of predicted contacts. Therefore it is possible that the optimal cut-off value would change on a per case basis. To evaluate whether the cut-off value of 30% LR-SPCs would be generally applicable to a number of proteins or whether there was a lot of variation in the optimal cut-off, we evaluated the density of correct and incorrect models at 50% extrusion in a run of standard SAINT2, on the N-terminal contacts data set. Figure 3.5 shows that the

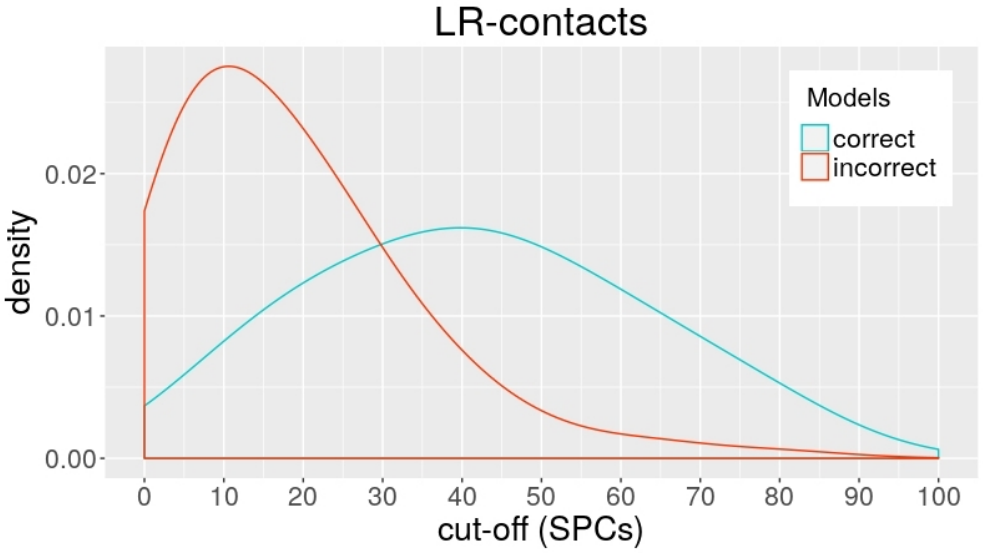



**Figure 3.4:** Density of LR-SPCs for correct (blue) and incorrect (red) models in a run of 10,000 models for proteins from the N-Terminal Contacts Data Set (Table S2).

ideal cut-off value for each protein is not consistent, and ranges from 10% to 30% for the cases shown. We also observe that some cases have better separation in the densities of LR-SPCs in correct and incorrect models, and thus the cut-off would be more effective these cases.

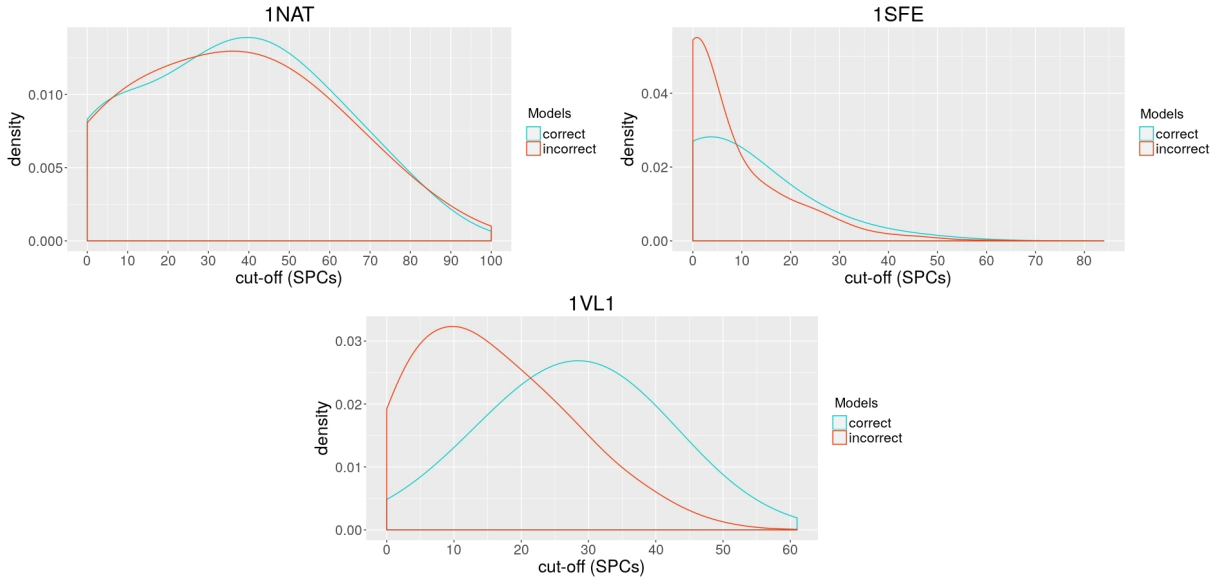



**Figure 3.5:** Density of LR-SPCs for correct (blue) and incorrect (red) models from a run of 10,000 models for the soluble proteins 1NAT, 1SFE and 1VL1.

## 3.4 Testing SPC Cut-off Implementations

All versions of SAINT2 tested are described in Table 2.3. We investigated whether implementing a LR-SPC cut-off led to enrichment of correct models in the final ensemble. We also assessed whether we could produce this enrichment when running time was limited to the equivalent computational time required by standard SAINT2. We tested several versions of a cut-off protocol on a small number of soluble proteins to save computational time(the Reduced Correct data set -see methods). For the same reason we applied a time limit of 24 hours for versions 1-3 of the cut-off; in version 3.2 we implemented a separate limit based on a time regression.

### 3.4.1 Version 1: Strict Cut-off

Implementing a strict cut-off of 30% LR-SPCs at 50% extrusion (version 1) produced enrichment of correct models of between 1 and 1.5 times for the cases tested (Figure 3.6). While the proportion of correct models increased, there did not seem to be any clear effect on the quality of the best models produced; the TM-score of the best model produced for each case was within 0.05 of the best model produced using standard SAINT2.

There was a slight increase in the running time used to generate 10,000 complete models with this cut-off compared to standard SAINT2; extra time is required in version 1 to generate the first half of the decoys that are produced and eventually labelled as incorrect and discarded. Generation of the next model then begins from scratch. 1NAT and 206L both complete runs of 10,000 complete models within the time limit of 24 hours, however 1VIN, 1SL8 and 1SFE all reach the time limit without producing 10,000 models.

Version 1 of our cut-off assumed that the optimal cut-off value of 30% LR-SPCs would be applicable to all proteins. However, based on the observation in Figure 3.5, this is not true and different proteins have different optimal cut-off values. Therefore we attempted to

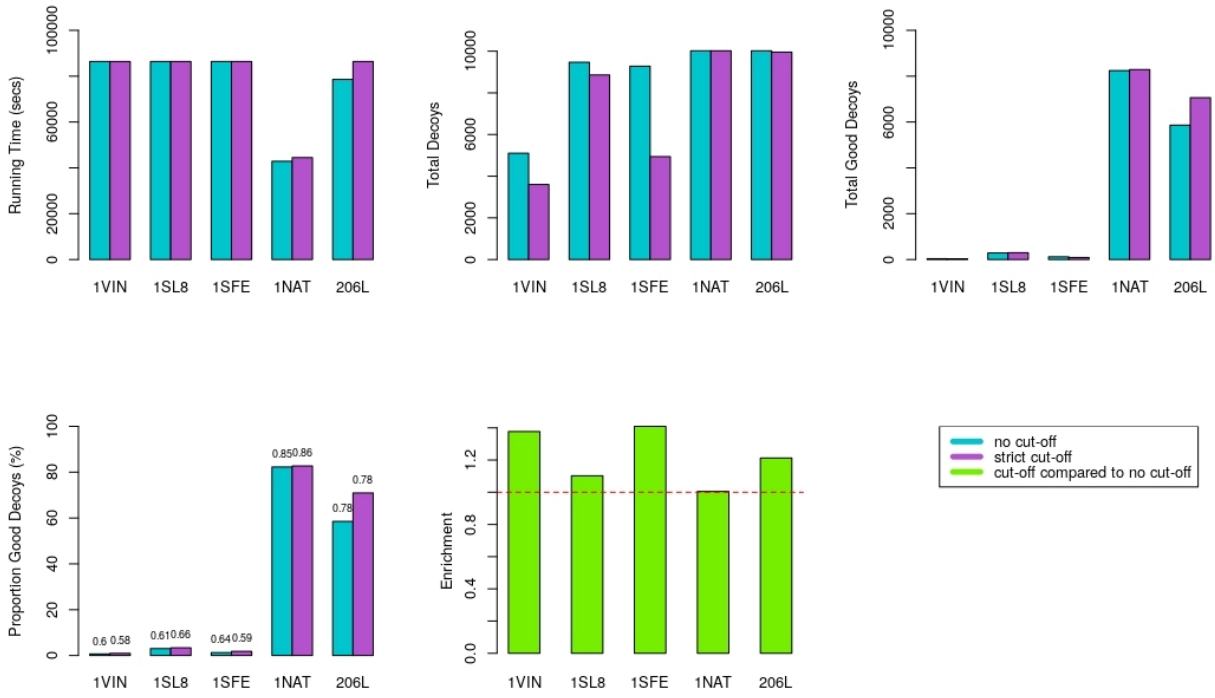



**Figure 3.6:** A comparison of structure prediction and generation of decoys with no cut-off (blue), and the strict cut-off protocol (purple). Decoys were generated as detailed in Table 2.3. The values compared were running time (limited to a maximum of 24 hours) (top left); Total number of complete decoys produced (top left); Total number of correct decoys produced for each protein (top right); the proportion of decoys produced that are correct for each protein, with TM-score of top scoring decoy labelled (bottom left); the enrichment of correct decoys using the cut-off (bottom middle).

improve upon version 1 by introducing a flexible cut-off value that started at 100% LR-SPCs and incrementally stepped down for each run of decoys produced.

### 3.4.2 Version 2: Stepped Cut-off

We wanted to improve on the efficiency of version 1 at producing correct models; in version 2 we implemented a stepping down approach with the LR-SPC cut-off value. This potentially would allow us to generate models at the optimal cut-off for every protein. The cut-off started at 100% LR-SPCs, and decreased by increments of 10% for every run of 100 attempts at building models.

Figure 3.7 shows that version 2 (stepped) produced enrichment of correct models for every protein tested. However, this was at the expense of an increased running time. Figure 3.8 shows that each of the proteins tested has a different cut-off value where the most partial decoys pass the cut-off and are completed to produce full models. For example, for 1VIN this is with a cut-off value of 10% and for 1NAT this is 40%. In general, the lower the cut-off value, the more models are produced as more can pass the cut-off - the exception to this occurs when most of the 1,000 models have been completed at higher cut-off levels.

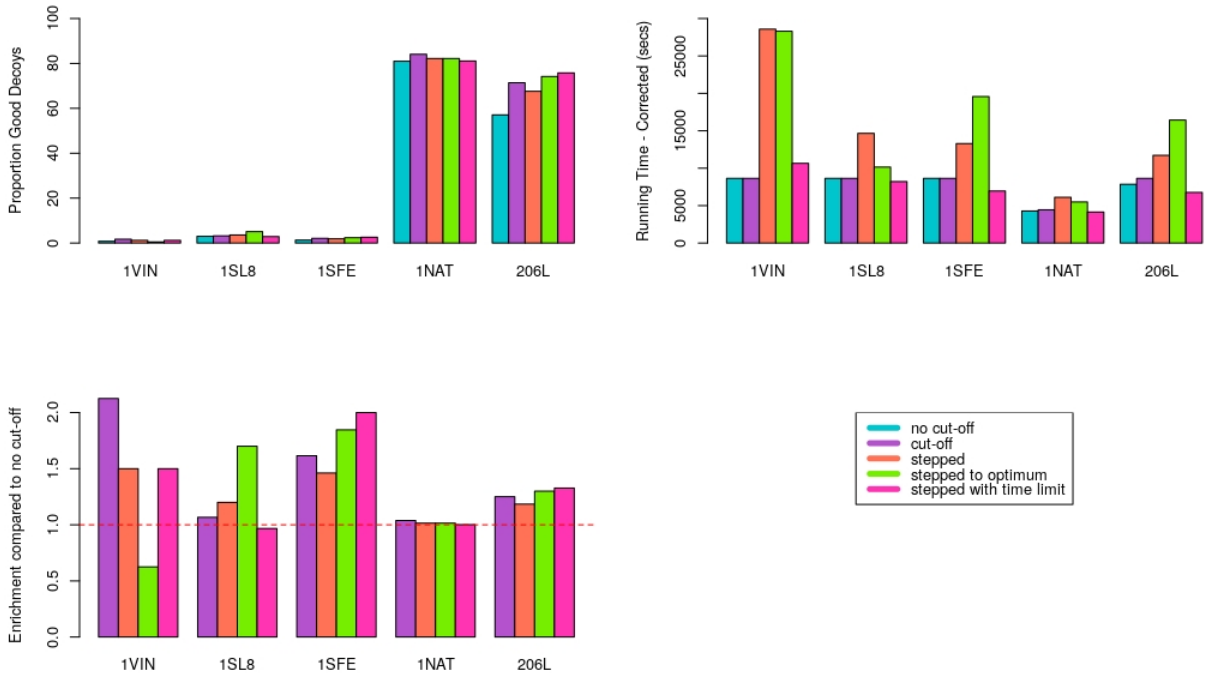



**Figure 3.7:** Comparing runs of 1,000 models for standard SAINT2 (blue), to the strict cut-off (purple), stepped cut-off (orange), stepped cut-off to optimal value (green) and stepped cut-off to optimal value with time restriction (pink) versions (as described in Table 2.3). Proportion of correct models produced (top left), the corrected running time in seconds (top right), enrichment of correct decoys produced for each method as compared to standard SAINT2 (represented by red line) (bottom left).

For the majority of proteins we see that a higher proportion of correct models are produced at higher cut-off levels. For example, 1SFE, while producing relatively few complete models

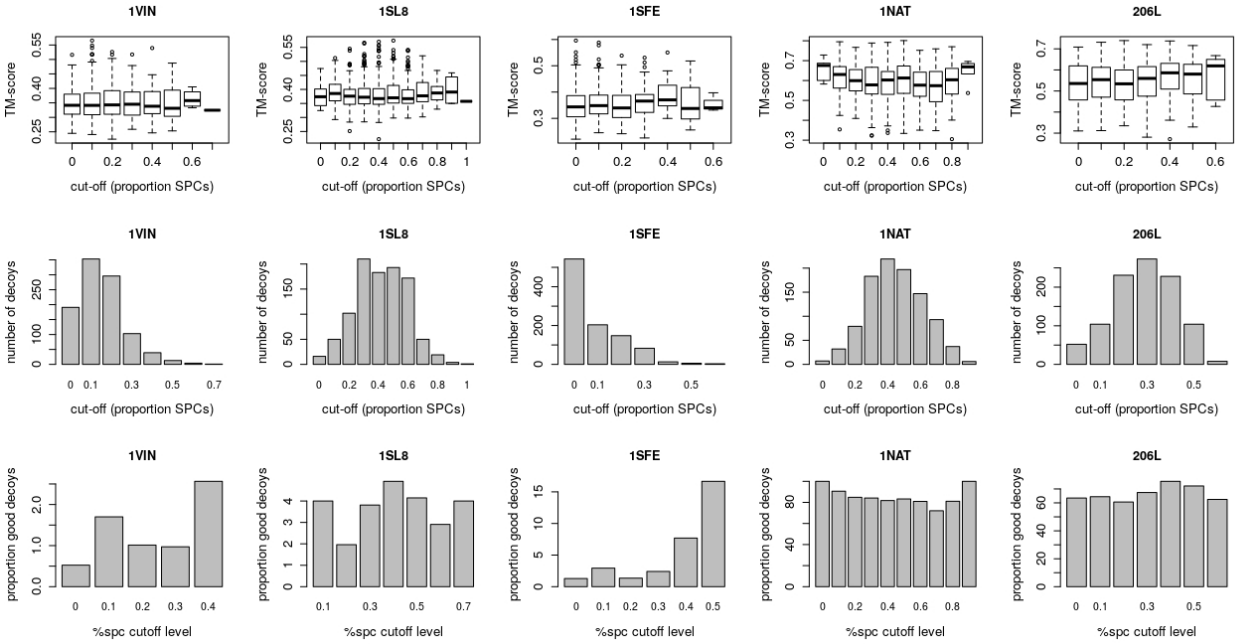



**Figure 3.8:** Looking at the decoys generated for each protein from the test set described in Table 2.3, using version 2. The TM-scores of decoys produced at each step-down of cut-off (top), the total number of complete decoys produced at each step-down in cut-off (middle), the proportion of correct decoys produced at each step-down in cut-off (bottom).

at a cut-off value of 50%, produces the highest proportion of correct models here. The cut-off values that produce a higher proportion of correct models also generally produce better quality models, with higher TM-score. We may be able to improve the efficiency and enrichment of correct models produced by identifying this optimal cut-off for each individual protein (where there is a high proportion of correct models), and producing the maximum number of models at this cut-off.

### 3.4.3 Version 3: Stepped to Optimum Cut-off

We created version 3 to improve enrichment of correct models produced by identifying the optimal cut-off for each protein and producing the majority of models at this cut-off. We wanted to identify the cut-off value for each protein that is high enough to produce

better enrichment of correct models and also better quality models. The cut-off should be low enough that an adequate number of models are accepted and completed, or it could lead to prohibitively high running times.

We decided to build upon the stepping down concept introduced in version 2, although instead of stepping down by 10% LR-SPCs every run, we will only step down when the previous run results in prohibitively few complete models. We set the level of complete models required as 10 out of each run of 1,000 models.

Figure 3.7 shows that for three of the five proteins tested, version 3 (stepped to optimum) provides increased enrichment of correct models over versions 1 and 2. For one protein, 1VIN, version 3 actually produces a smaller proportion of correct models than the standard SAINT2. The running time, similar to version 2, is increased compared to version 1 and standard SAINT2. Version 3 identifies different optimal cut-off points for different proteins; at these optimal cut-off values the highest proportion of correct models are produced, as well as the highest quality model.

While versions 2 and 3 provide enrichment over standard SAINT2 for the majority of the cases tested, the running time is higher. To improve efficiency of these cut-off protocols, we altered some of the parameters in version 3, to create version 3.2. We also here introduced a time limit, based on a regression to estimate the time that standard SAINT2 would take to generate models for a protein of equivalent length.

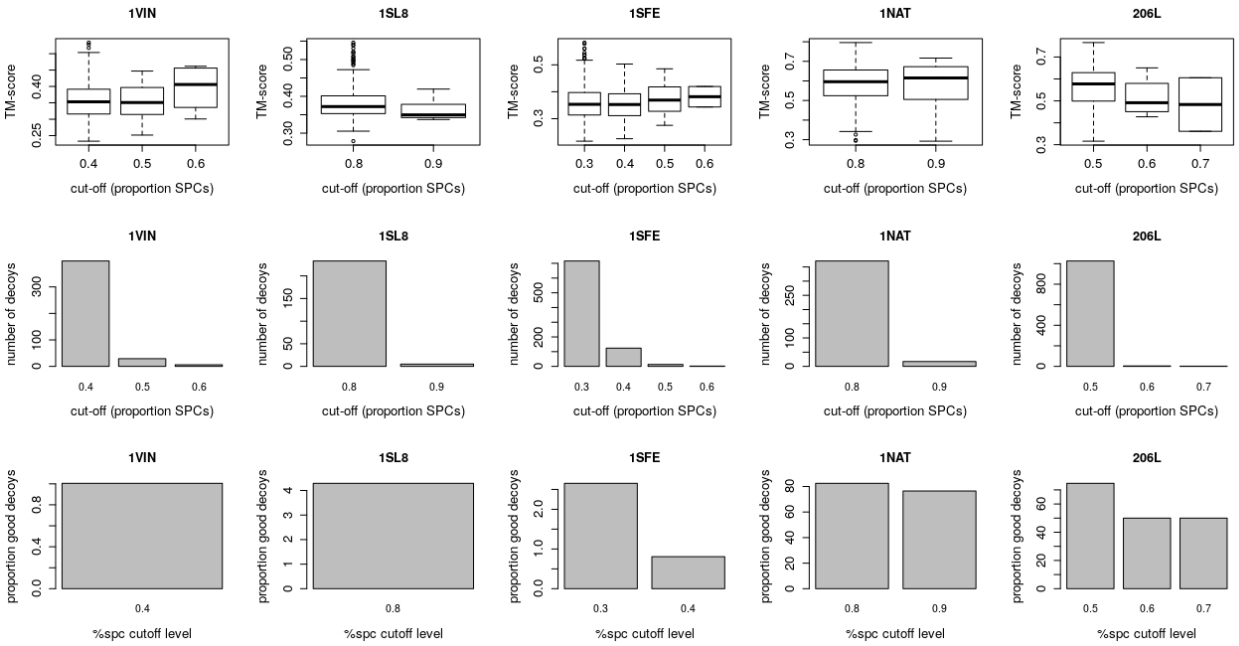



**Figure 3.9:** Looking at the decoys generated for each protein from the test set described in Table 2.3, using version 3. The TM-scores of decoys produced at each step-down of cut-off (top), the total number of complete decoys produced at each step-down in cut-off (middle), the proportion of correct decoys produced at each step-down in cut-off (bottom).

### 3.4.4 Time Regression

In order to more directly compare the enrichment of the final version of SAINT2 to the standard version, we implemented a time limit to constrain the runtime of SAINT2 version 3.2 to that of the standard SAINT2 when generating the same number of models for a protein of a given length. Equation 3.1 and Figure 3.10 show the linear regression model calculated with the R "linear model" function ("lm"), using the square of the length of the protein ($L^2$) as the input terms, and the times taken to produce 60 decoys, on 60 cores as the response, for each of the soluble proteins listed in Table 3.2.

| PDB code | Time (secs) | Length |
|---|---|---|
| 1CSP | 28 | 67 |
| 1CEW | 66 | 108 |
| 1SFE | 135 | 165 |
| 1FTG | 145 | 168 |
| 1SQW | 159 | 176 |
| 1SL8 | 167 | 181 |
| 1SDI | 227 | 212 |
| 1VL1 | 232 | 217 |
| 1W66 | 221 | 217 |
| 1VIN | 306 | 252 |
| 1RL0 | 297 | 255 |
| 1XWY | 323 | 260 |
| 2HVM | 352 | 273 |
| 1JKS | 347 | 277 |
| 1MSK | 496 | 327 |
| 1AYE | 705 | 398 |
| 1SMD | 1094 | 495 |

**Table 3.2:** Length and time taken in seconds for 60 models to be generated on 60 cores for each of the listed soluble proteins.

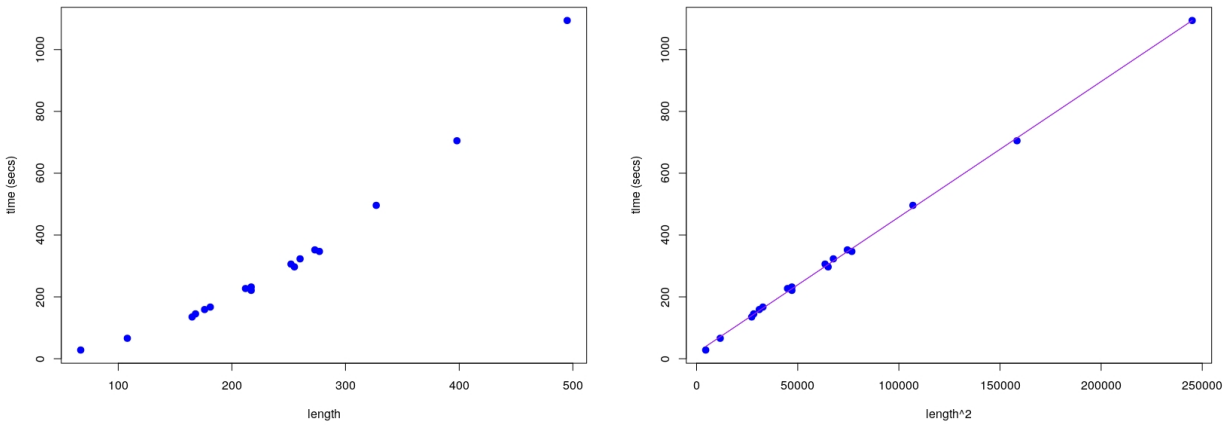



**Figure 3.10:** Length of the proteins in Table 3.2 and the time in seconds to generate 60 models on 60 cores (left). Time regression from linear model to describe the time in seconds to generate a model for a protein using the square of its length (right)

$$t_{runtime} = \frac{L^2}{228.1} + 19.77 \tag{3.1}$$

### 3.4.5 Version 3.2: Stepped to Optimum Cut-off with Time Limit

To create version 3.2 we modified some of the parameters from version 3, to increase efficiency and enrichment. Firstly we started the cut-off at 80% LR-SPCs rather than 100% as we observed that none of the cases tested used an optimum cut-off value above this, and it was rare that any correct models were produced at levels this high. We also raised the threshold of complete models required to be produced per run to maintain the cut-off level and avoid stepping down to 0.4% (40 models in each 1,000 decoy run). We also implemented a time limit, based on Equation 3.1 that would limit version 3.2 to the equivalent running time of standard SAINT2.

Version 3.2 (stepped with time limit) produced increased enrichment of correct models compared to versions 2 and 3 for three out of the five cases tested. For one case the enrichment was roughly the same, and for one case (1SL8) there was a lower proportion of correct models produced than the other versions. However, the running time for version 3.2 was reduced compared to versions 2 and 3, with roughly the same running time as standard SAINT2.

All of the versions of modified SAINT2 produced enrichment of correct models in the final decoy pool compared to standard SAINT2 for the majority of cases (Figure 3.7). Versions 1 and 2 produced enrichment for all proteins tested, while versions 3 and 3.2 produced enrichment for four out of the five proteins tested. Versions 2 and 3 had increased run time as compared to standard SAINT2, whereas versions 1 and 3.2 had comparable run time to standard SAINT2. We selected version 3.2 to carry forward for further testing as it represented a compromise in enrichment and runtime compared to versions 2 and 3. We selected version 3.2 over version 1 despite enrichment appearing slightly worse, as it was expected to be more

effective with proteins that SAINT2 produces few correct structures for (not included in the Reduced Correct data set), due to its flexibility with identifying an optimal cut-off value.

### 3.4.6 Further Testing of Version 3.2

We tested version 3.2 on a slightly extended data set (Figure 3.11). We generated 10,000 models for three of the proteins from the Reduced Correct data set, as well as a further seven proteins from our original data set of 41 soluble proteins (1FTG, 1SDI, 1SQW, 1VL1, 1W66, 1XWY, 2HVM). We disregarded 1NAT and 206L from the Reduced Correct data set, as these proteins were already predicted well by standard SAINT2, and from the preliminary tests of versions 1-3.2. The additional seven proteins were included in our test to represent cases that SAINT2 could predict less well, or not predict correctly at all as this is where we would like to increase efficiency and enrichment most.

Within the time restriction applied to limit the runtime to that of an equivalent standard SAINT2 run, 8 of the 10 cases tested completed producing 10,000 complete models (Figure 3.11). For 5 of the 10 cases (1FTG, 1SFE, 1VIN, 1VL1 and 1XWY) there is enrichment of the correct models in the final decoy pool. 1SL8 was the only protein with a decrease in the proportion of correct models produced after implementation of the cut-off. Four proteins (1SDI, 1SQW, 1W66 and 2HVM) produced no correct models; these proteins also had no correct models using standard SAINT2.

Using version 3.2, the best decoy was produced at the cut-off level identified as optimum (Figure S1), even though the highest proportion of correct models were not necessarily produced at this level (Figure S3), suggesting that the parameter changes implemented from version 3 are effective. For example, 1SFE and 1VIN both generate their best model with a 30% SPC cut-off, but have a higher proportion of correct models produced with a 40% SPC cut-off.

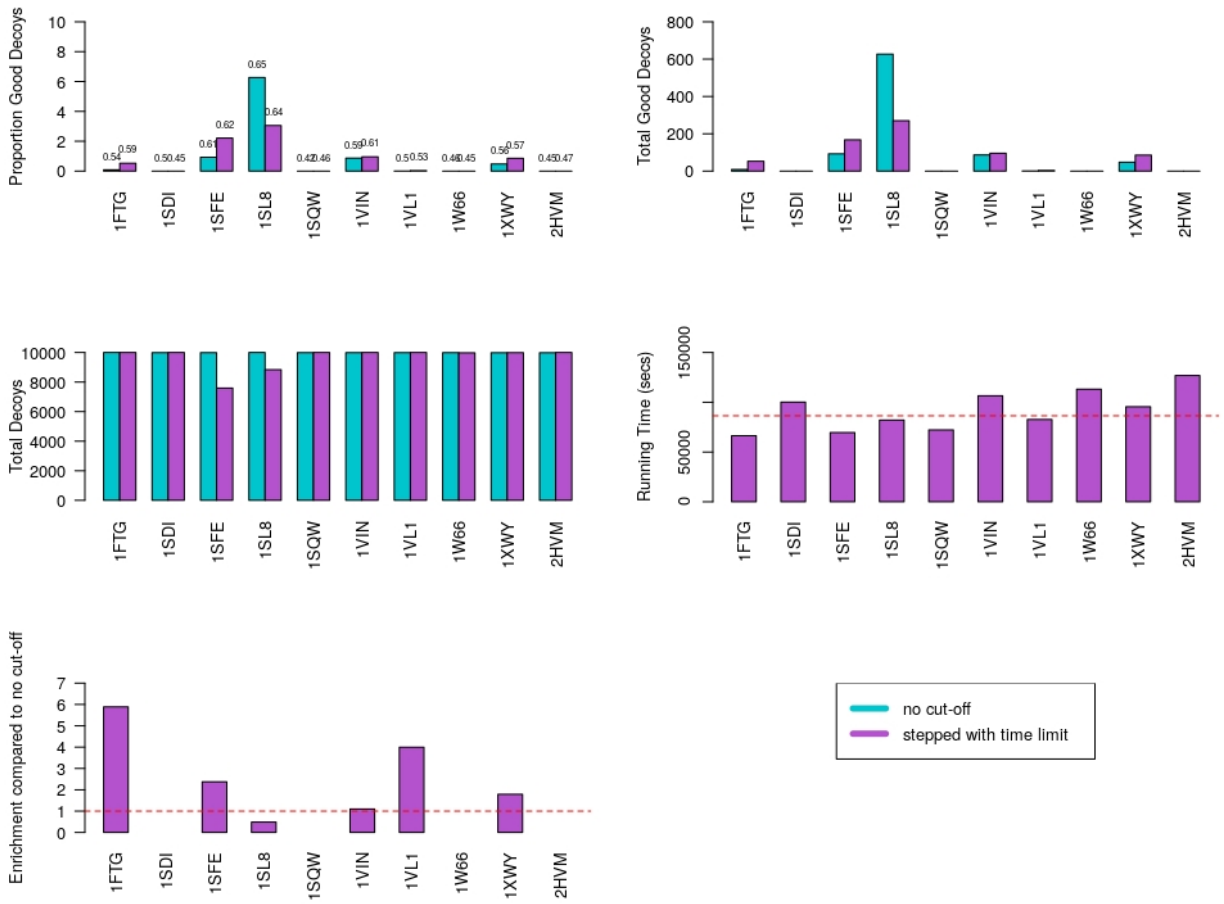



**Figure 3.11:** Comparing the stepped cut-off to optimal value protocol with an added time restriction (purple) to no cut-off (blue), as described in Table 2.3. Proportion of correct decoys produced, with top TM-score of decoys labelled (top left); number of decoys produced that were correct (top right); the total number of decoys produced for each protein (middle left); running time in seconds (middle right); Enrichment of correct decoys for each protein (bottom left).

#### 3.4.6.1 Number of decoys needed for correct model

The number of decoys per run required for standard SAINT2 to predict a structure correctly was previously estimated at 10,000 [13]. In order to investigate whether the increase in efficiency brought about by implementing a cut-off into SAINT2 could lead to any meaningful decrease in computational requirement, we evaluated the sample sizes required for the generation of at

least one correct model, or for the generation of at least five correct models, in a run of version 3.2 of the modified SAINT2 on the proteins tested in the previous section. The analysis was carried out in a smilar way to that described in [13]. We took 20 samples of each size tested, and looked at the minimum sample size where there was one correct model, and five correct models in at least 19 of these samples. Figure 3.12 shows the minimum sample size required for at least one correct model, and at least five correct models, using version 3.2 and standard SAINT2.

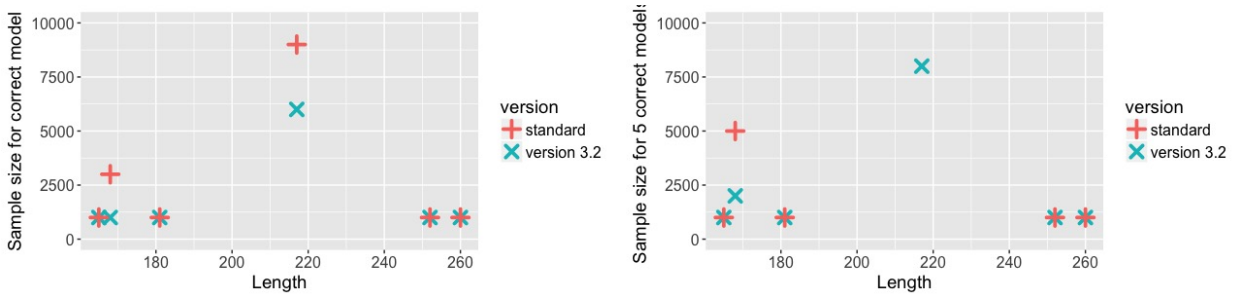



**Figure 3.12:** Minimum sample size for proteins of varying lengths so that best model is correct (left) or at least five models are correct (right), using version 3.2 (blue) and SAINT2 (red).

Of the 10 proteins tested on version 3.2 in the previous section, we could predict correct structures for six: 1FTG, 1SFE, 1SL8, 1VIN, 1VL1, 1XWY). For four of these cases, version 3.2 of modified SAINT2 is able to predict at least 5 correct structures in a sample size of less than 1,000 models. For the protein 1FTG, version 3.2 produced one correct model within 1,000 models, but a sample size of 2,000 was required to produce 5 correct models. There is also one case (1VL1, a 217 residue protein), where a run of 6,000 models is needed to predict one correct structure and 8,000 needed for at least five correct structures.

We wanted to compare our results for the required sample size to generate a correct model using version 3.2, to those needed using standard SAINT2. For the six cases tested, SAINT2 could produce a correct model for four of the proteins in a run of 1,000 models. A run of 3,000 models was required to give a correct structure prediction for 1FTG, and a run of 9,000 was needed for 1VL1. Our version 3.2 decreases the sample size needed to produce a correct model for both 1FTG and 1VL1 (Figure 3.12).

Investigating the run size needed to generate at least five correct models for each case using standard SAINT2 produced similar results; we could predict correct structures for four proteins within a run of 1,000 models. A run size of 5,000 models was required for 1FTG, and standard SAINT2 did not produce five correct models for 1VIN. Again, our version 3.2 required less models to predict five correct models for 1FTG, and could produce at least five correct models for 1VL1 within a run size of 7,000 decoys (Figure 3.12).

The results shown on this small test set show that for proteins where a correct model was less frequently predicted, our new version of SAINT2 reduced the number of models needed for a correct prediction. It may be applicable to reduce the number of models produced by SAINT2 for some proteins - but there would need to be further investigation into any properties that may separate these 'more easily predictable' proteins from those that require larger run sizes. With a run size of 10,000 models, our new version of SAINT2 appears to increase the frequency of predicting a correct model for 'harder' proteins when compared to standard SAINT2.

## 3.5 SAINT2 Diagnostics

There are still a number of cases for which we could not predict correct structures. In order to establish why SAINT2 and SAINT2 version 3.2 were unable to correctly predict the structures of some proteins, we assessed the quality of the data that was input (Table S1). We used 14 soluble proteins, selected from the original data set of 44 soluble proteins; this was a combination of cases that standard SAINT2 could predict correct structures for, and could not, so that we could identify any differences in input data quality between these cases.

Using an $N_{eff}$ value of 32 or above as an indicator of a sufficient quality MSA to produce predicted contacts that allow fold-level accuracy [6], we observe that all of the good cases (that SAINT2 could produce correct models for) had $N_{eff}$ values greater than 32. Of the bad cases, 1SDI, 1SQW and 1MSK all have $N_{eff}$ values lower than 32, indicating insufficient sequence

information in their MSAs; this could explain the failure of SAINT2 to produce correct decoys. A lack of sequence information (number and diversity of sequences) will result in low precision of predicted contacts, which in turn will lead to poor quality structure prediction. 1W66, 2HVM, 1BMG and 1NEP all have $N_{eff}$ values greater than 32 and for these proteins, the failure to produce correct models cannot be attributed to the quality of the MSA.

We see no significant difference between the precision and coverage of the fragment libraries, the average precision and coverage was 0.57 and 0.95 for cases SAINT2 could predict, and 0.54 and 0.99 for cases SAINT2 could not predict. Similarly the average precision of secondary structure prediction is 65% for cases SAINT2 could predict and 66% for cases SAINT2 could not predict. The mean absolute error (MAE) for torsion angles was higher for those proteins where SAINT2 could produce a correct model. The accuracy of predicted contacts at 50% extrusion (LR and SR, where the cut-off is enforced) was lower for the proteins SAINT2 could predict correctly. From this analysis, we cannot attribute SAINT2 failure to input quality alone.

## 3.6 Testing Protocol on Larger Datasets

### 3.6.1 Dataset 1

We aimed to further test the observations found when running version 3.2 of SAINT2, that we may be able to enrich the number of correct models produced in a run of SAINT2, especially in cases where standard SAINT2 could only predict a few correct structures. Therefore we tested two protocols on all of the soluble cases given in Table S2. First the SAINT2 with no modifications was used to generate 10,000 models for each protein. The second 'version 3.2' was used without a time regression to produce a set number of models for each protein.

For each of the 44 proteins in Table S2, we generated 1,000 models using the new protocol. We also took 20 samples of 1,000 models from runs of standard SAINT2 that had generated

10,000 models for each of the 44 proteins. We averaged the number of correct models from these samples and compared them with those produced using the new protocol. Error bars were not included, as they were generally very small with the highest standard deviation of 14 where the number of correct models was 617. More typically the standard deviation for the number of correct models was between 0 and 3.

I then used the new protocol to generate a further 9,000 models for each of the 19 cases from Table S2 where standard SAINT2 had produced at least one correct model in its run of 10,000 models. This set of 10,000 models generated using the new protocol were then compared with the 10,000 models generated using standard SAINT2.

We group the cases tested into three different categories. The first category contains cases that standard SAINT2 is able to predict well, generating a large number (greater than 100 of the 10,000 model run) of correct models for. The second category contains cases where only a few correct models are generated (between 0 and 99 correct models). The final category contains the cases where standard SAINT2 did not generate any correct models.

Looking initially at 1,000 models per protein, we observe a different behaviour for each of these categories when comparing the number of correct models produced by the new protocol, to standard SAINT2. There is little to no enrichment for those cases SAINT2 already predicted well. For those cases in the third category, where standard SAINT2 predicted no correct models, the new protocol also produced no correct models.

There were 12 proteins in the second category, where standard SAINT2 produced at least one correct model, but less than 100. Of these 12 proteins, using the new protocol produced enrichment for five cases. There was no change in enrichment for two cases. There were two cases where the new protocol produced fewer correct models than standard SAINT2; for 1CSP the new protocol produced one correct model, rather than two produced with standard SAINT2; for 1EKG the new protocol produced no correct models, where standard SAINT2 produced one. There were three cases where sampling 1,000 models from those produced

using standard SAINT2 gave no correct models (although correct models were present in the full 10,000 models) and the new protocol did produce a small number of correct models.

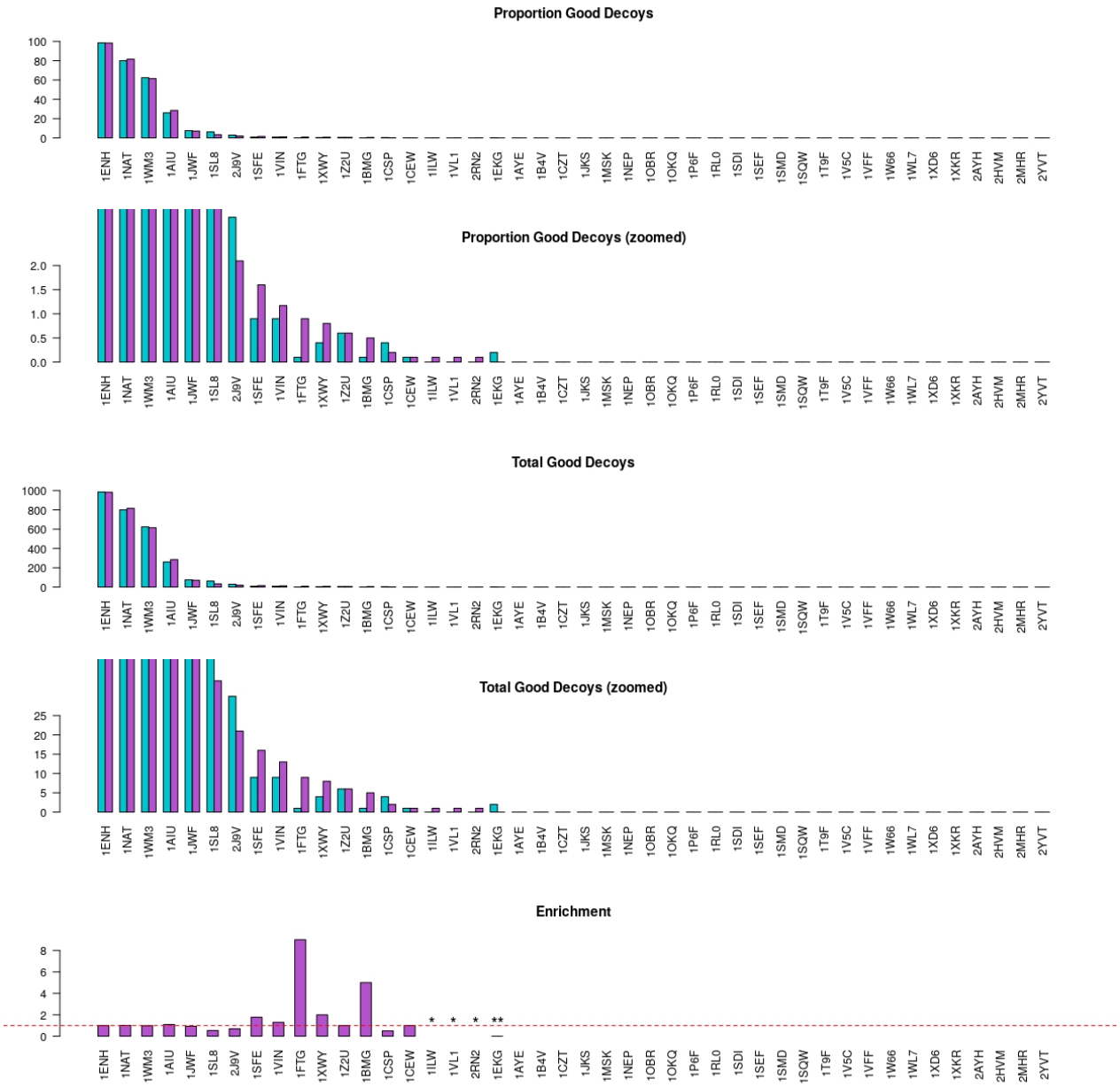



**Figure 3.13:** Comparison of 1,000 models generated with the new protocol (purple) with a sample of 1,000 models generated with standard SAINT2 (blue). Proportion of models produced that were correct (top), and zoomed in (second from top), the total number of correct models produced (middle), and zoomed in (second from bottom), and the enrichment of correct models produced using the improved protocol over standard SAINT2, for the 44 soluble proteins in the data set. Cases where standard SAINT2 produces no correct models, but the new protocol produces one or more correct models are marked with *. Cases where standard SAINT2 produces one or more correct models and the new protocol produce no correct models are marked with **.

The results in Figure 3.13 suggest that using the new protocol may provide enrichment for cases where standard SAINT2 only produces a small number correct models. However, there is some element of randomness related to the stochastic process of SAINT2 that may influence the production of correct models, especially when only very small numbers are produced. Therefore in order to investigate this fully, more models would need to be produced/ more runs of SAINT2 carried out.

The results in Figure 3.14, comparing runs of 10,000 models, show a similar pattern to those produced in Figure 3.13 when generating 1,000 models. This time we only include proteins in two categories: those standard SAINT2 predicts well, and those standard SAINT2 can predict only a few correct models for. In general, there was again little to no enrichment when using the new protocol for the seven cases standard SAINT2 produced a good number of correct models for. Of the remaining 12 proteins, using the new protocol produced enrichment for seven cases. There were four cases where using the new protocol produced fewer correct models, and one case where standard SAINT2 predicted a correct model, but the new protocol produced no correct models.

The new version of SAINT2 - when compared to standard SAINT2 - enriches the correct models produced in the 'harder' cases, where less correct predictions are made. If we can identify properties, such as protein length, the variation of sequences included in the MSA or the number of output predicted contacts for a case, that would identify it may fall in this group of harder cases - we could selectively use our new protocol for those cases it would more likely produce enrichment for.




**Figure 3.14:** Comparison of 10,000 models generated with the new protocol (purple) with 10,000 models generated with standard SAINT2 (blue). Proportion of models produced that were correct (top), and zoomed in (second from top), the total number of correct models produced (middle), and zoomed in (second from bottom), and the enrichment of correct models produced using the improved protocol over standard SAINT2, for the 44 soluble proteins in the data set. Cases where standard SAINT2 produces one or more correct models and the new protocol produce no correct models are marked with **.

### 3.6.2 Dataset 2

The results from the larger data set suggested that using our protocol enriched the production of correct models in some cases. We next tested our protocol on an even larger data set comprised of 245 soluble proteins (Table S3), generating 1,000 models for each case and comparing these results to a run of 1,000 models produced using standard SAINT2.

Firstly, we wanted to investigate whether using our protocol, and restricting search space based on satisfied predicted contacts, produced any increase in the quality of models produced. Figure 3.15 shows the density of TM-scores for all models produced for all cases, using standard SAINT2 and our protocol. Here we see that there is very little difference in the distributions of model TM-scores using either version of SAINT2.

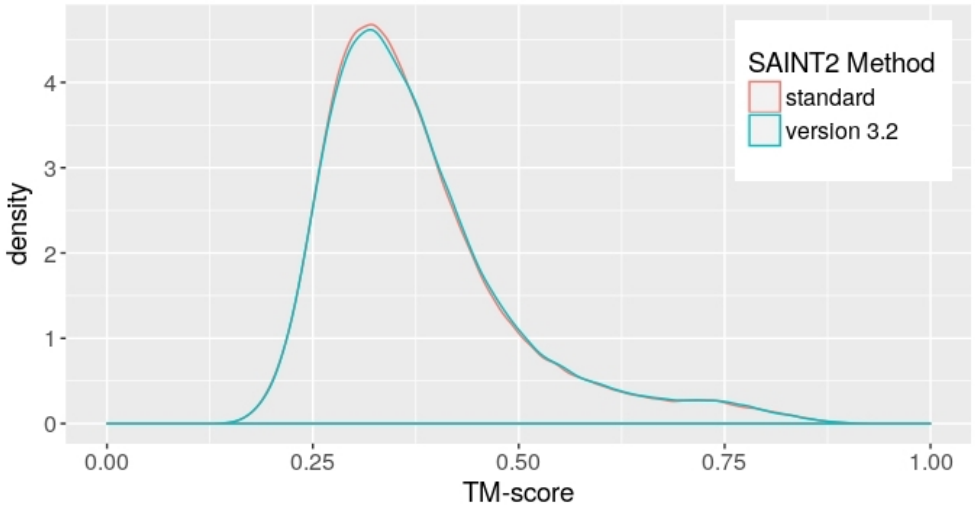



**Figure 3.15:** The density of TM-scores for 1,000 models produced for all of the 245 proteins in the larger data set, produced using standard SAINT2 (red) and the new protocol (blue). Generated using the default kernel density estimation method in R.

Figure 3.16 shows the distribution of the TM-scores of the best models produced for each of the 245 cases tested, using standard SAINT2 compared to the new protocol. Again, there is very little difference in the distribution of best TM-scores produced using each method.

We next investigated whether using our new protocol caused enrichment of correct models.

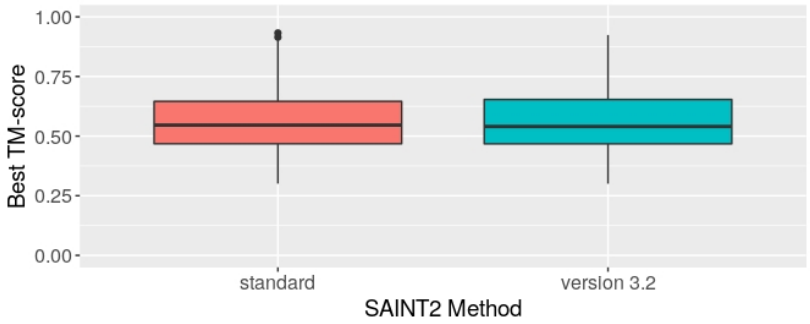



**Figure 3.16:** The distribution of TM-scores of the best scoring model for each of the 245 proteins in the larger data set produced using standard SAINT2 compared to the new protocol.

In Figure 3.17, we see that out of the 245 cases tested, using the new protocol enriched the number of correct models produced in 89 cases. For 90 cases there was no difference in the number of correct models produced, and for 66 cases, using the new protocol produced less correct models than using standard SAINT2.

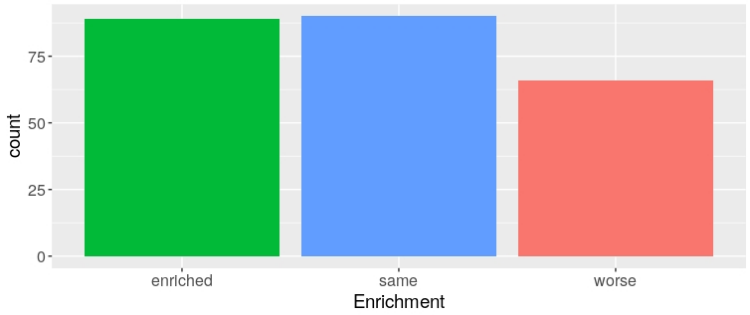



**Figure 3.17:** The number of cases from the larger data set of 245 proteins where using the new protocol enriched the number of correct models produced (green), produced the same number of correct models (blue) and produced less correct models (red).

We next wanted to investigate whether there were certain cases that our new protocol would be better at enriching. We analysed the length, $N_{eff}$ (number of effective sequences in the MSA [61]), and number of predicted contacts, for the cases we could enrich, produce the same number of correct models, and produced fewer correct models for (Figure 3.18).

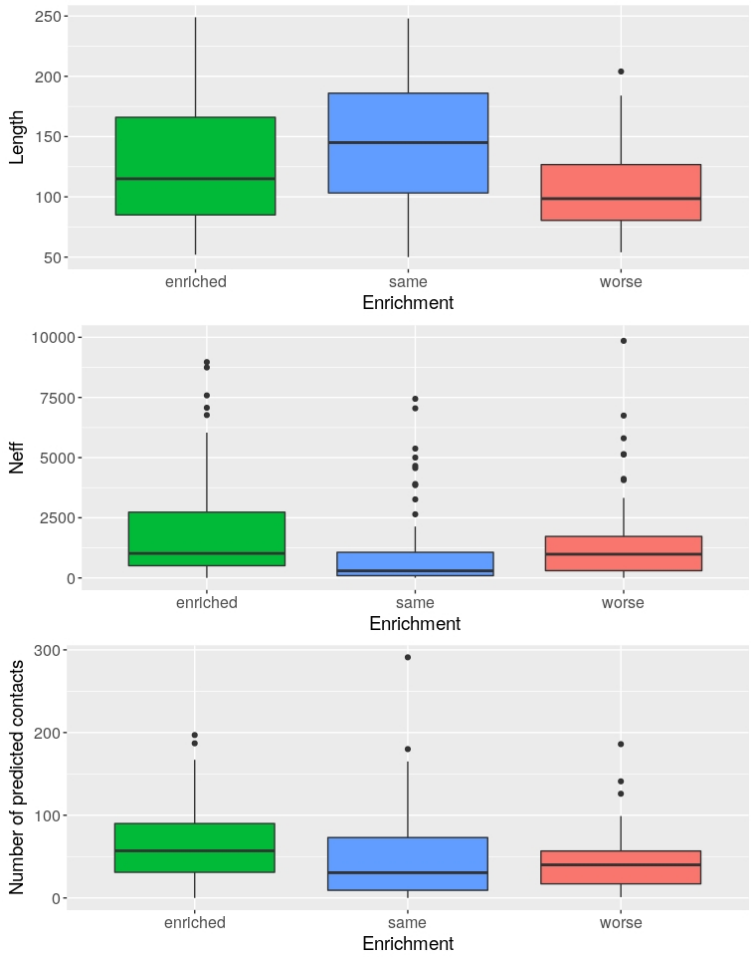



**Figure 3.18:** The distribution of lengths (top), $N_{eff}$ values (middle), and number of predicted contacts (bottom) for those cases the new protocol can enrich (green), produce the same number of correct models for (blue) and produces fewer correct models for (red).

Figure 3.18 shows that in general, the shortest proteins in the data set are those that we produce fewer correct models for. Proteins that we cannot enrich using the new protocol tend to be the longest in length, with enrichment occuring mostly on medium-length proteins. There is little difference in the average $N_{eff}$ value for those cases we can enrich, compared to those the new protocol produces fewer correct models for, however there is a lower average $N_{eff}$ value for those cases we perform the same for. When comparing the number of predicted contacts in the cases we can enrich compared to those we couldn't, the proteins

our new protocol enriched had a higher average number of predicted contacts than those we produced the same, or fewer correct models for.

We also wanted to investigate whether there was any difference in the length, $N_{eff}$ value, and number of predicted contacts for those cases that we could get correct (using the new protocol), and those we couldn't. Figure 3.19 shows the results obtained when comparing these variables. For those cases we could predict at least one correct model for, there was a shorter average length, higher average $N_{eff}$ value, and a slightly higher number of predicted contacts.

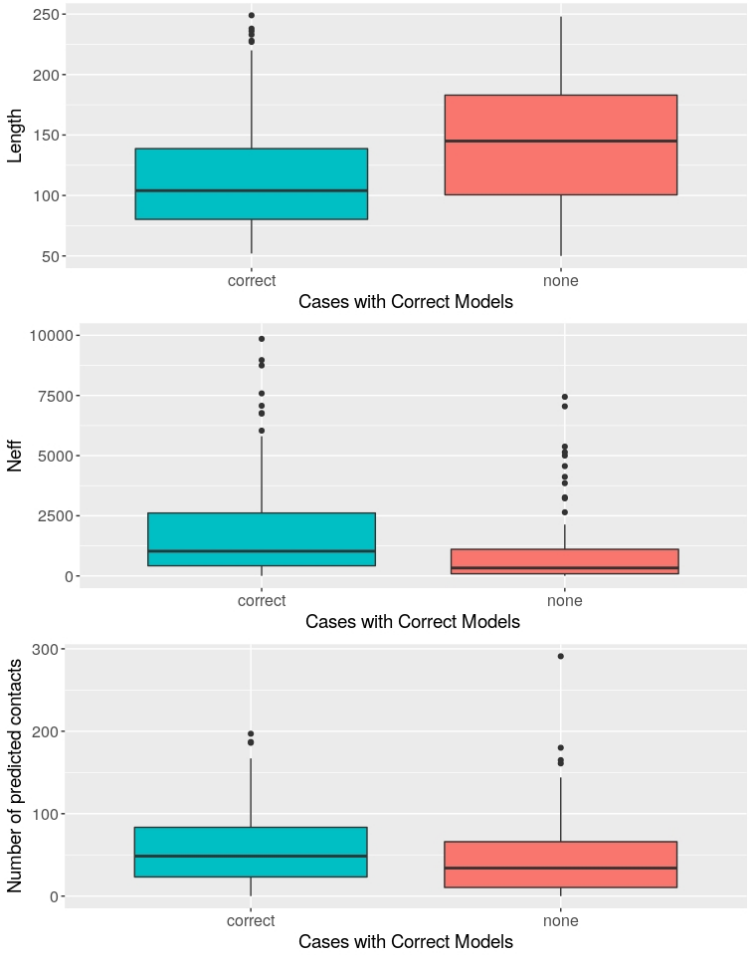



**Figure 3.19:** The distribution of lengths (top), $N_{eff}$ values (middle), and number of predicted contacts (bottom) for those cases the new protocol can predict at least one correct model for (red), and cannot predict correctly (blue).

If we can identify certain properties of proteins that our new protocol can enrich, then it may be possible to selectively implement it for cases with these properties, that it will be more likely to produce enrichment for. These results suggest that our protocol may be more useful on proteins with more predicted contacts.

Previous results suggested that using our new protocol may be able to produce enrichment in correct models produced for cases, at smaller run sizes. Therefore we analysed the number of cases we could generate at least one correct model for, at incrementally increasing run sizes, from 1 to 1,000 (Figure 3.20). At all run sizes, there were similar numbers of cases with correct models when using standard SAINT2 and the new protocol. With the final run size of 1,000 models, there were 157 cases with correct models using standard SAINT2, and 158 cases with correct models using the new protocol. However, this analysis was only completed up to run sizes of 1,000 - if we could perform it on larger sample sizes (up to 10,000 models), we may begin to see more cases with correct models using our cut-off.

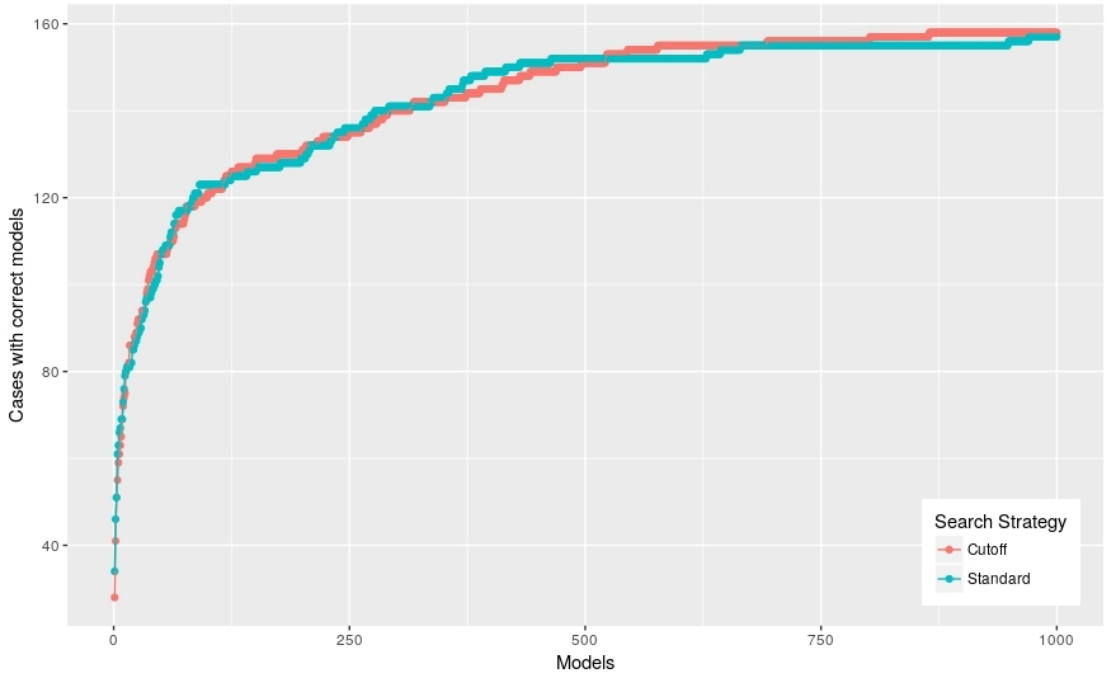



**Figure 3.20:** The number of cases the new SAINT2 can produce at lease one correct model for, at incrementally increasing run sizes, from 1 to 1,000 models.

## 3.7 Discussion

My work so far has focused on how predicted contacts are satisfied during the extrusion of a model, and whether we can use this information to improve protein structure prediction. Initial implementation of a cut-off based on SPCs into SAINT2 provided enrichment of correct models for cases where SAINT2 produces only a few correct models. For proteins that SAINT2 did not produce a correct prediction, the new versions of SAINT2 still failed.

Assessment of the quality of the SAINT2 input information (including sequence information provided by the MSA, predicted secondary structure, predicted torsion angles, fragment library and predicted contacts) showed that for some proteins an insufficient MSA $N_{eff}$ value may explain our inability to produce a correct model. However for other proteins there was no clear difference in input information quality when compared to cases we could correctly predict. We found no difference between predicted secondary structure precision, predicted torsion angle precision, and fragment library quality, between good cases and bad cases. Therefore, one explanation for the inability of SAINT2 to predict correct structures for some proteins may be that the scoring function is ineffective. For example, we apply a relatively heavy weight to the predicted contacts potential; however this may encourage entrapment of models in incorrect conformations; further testing is required to corroborate this claim.

Further testing of the new protocol on larger data sets showed that in general there is no increase in model quality, or the number of cases that SAINT2 can correctly predict when using this alternative strategy to search conformational space, however there may be some cases (with regards to protein length, $N_{eff}$, and number of predicted contacts) where our protocol may be more useful in enriching the correct models produced. More specifically it seems as though proteins with more predicted contacts are more likely to be enriched by the new protocol. If this is the case, it would be interesting to explore why this is happening and how individual contacts have an effect upon structure prediction. It may also suggest that

improvements in contact prediction methods may lead to further enrichment in this way.

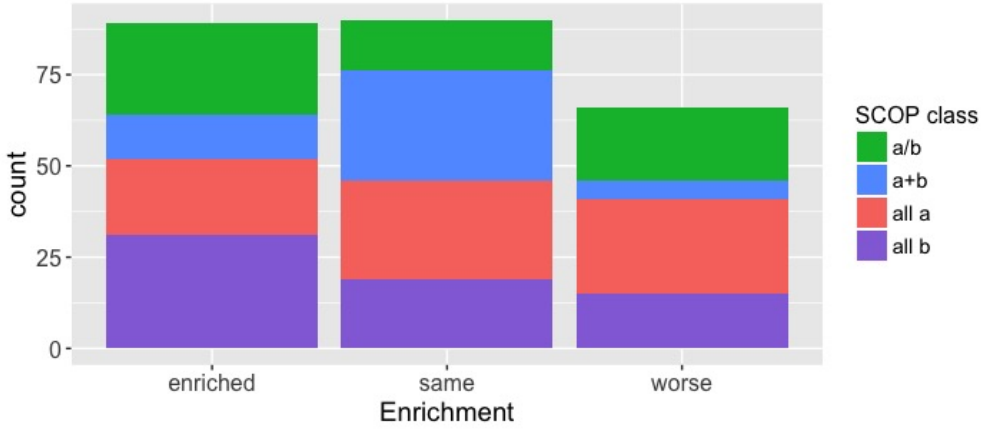



**Figure 3.21:** The SCOP classes - $\alpha/\beta$ (green), $\alpha/\beta$ (blue), all $\beta$ (red), and all $\alpha$ (purple) for the cases from the larger dataset of 245 proteins where the new protocol enriched the number of correct models produced, produced the same number of correct predictions or produced less correct predictions.

While there has not been an extensive analysis in this project regarding the correlation between the ability of SAINT2 to correctly predict the structure of a protein and its SCOP class - for the final analysis using the larger dataset, Figure 3.21 shows the distribution of SCOP classes for the cases tested. In general, the new protocol performed better on $\alpha/\beta$ and all $\beta$ cases, while $\alpha + \beta$ cases were largely not affected and all $\alpha$ cases were worse. In the case of all $\alpha$ proteins, the new protocol may cause a decrease in correct structures as these are generally well predicted by standard SAINT2 anyway.

It would be interesting to see whether we could selectively use the new protocol on cases with more predicted contacts produced (or whether producing more predicted contacts for other cases increases the impact of the new protocol on other cases), or specifically $\alpha/\beta$ and all $\beta$ cases, to produce enrichment without decreasing the correct predictions for some cases.

If we can increase this enrichment, or understand more about what is causing it, we may be able to produce correct models for cases that we are currently unable to predict.

# 4

# Investigating the Behaviour of Predicted Contacts

The results of the previous chapter show that there are cases with good precision of predicted contacts where SAINT2 can not predict any correct models. In this chapter we investigate how SAINT2 is utilising predicted contacts during extrusion and whether there is any difference between cases where we can correctly predict structures, and those where we can't. We also wanted to investigate whether there was any difference in how predicted contacts were satisfied and broken during the generation of correct and incorrect models. This could help us identify and select correct models within a run of SAINT2. Our unique insight into the behaviour of predicted contacts during model generation may allow us to detect patterns

that could not be seen when looking just at the complete chain.

We first investigated the specific contacts that were being satisfied in correct and incorrect models, to see if there was any difference between these cases. We looked for any contacts that were only being satisfied by correct models, and conversely whether there were any contacts that were only satisfied by incorrect models and if so whether these contacts were correct or incorrect predictions. We measured this behaviour throughout extrusion to see whether contacts were satisfied and broken differently in correct and incorrect models. If we can identify contacts that are advantageous to the prediction of correct models, we could up-weight these contacts during extrusion and increase the chances of producing a correct structure prediction. Conversely, if there are contacts more likely to be associated with incorrect models, we can down-weight them to avoid producing incorrect models.

Predicted contacts are satisfied differently in different models; this may be due to a number of factors including the probablility of the predicted contact being correct (PPV score as assigned by metaPSICOV), whether it is a true contact or not, its position in the protein sequence and the separation of the residues involved in the contact within the sequence. Therefore, we also analysed what the satisfaction of the predicted contacts in a run of models could tell us about whether the predicted contact was correct and whether there could be any way to identify incorrect predicted contacts using information from a SAINT2 run.

## 4.1 Methods for Investigating the Behaviour of Predicted Contacts

### 4.1.1 Analysis of Contacts Present in Correct and Incorrect Models

A run of 10,000 models for the soluble protein 1AIU was used to evaluate the presence of individual contacts in correct and incorrect models. The presence of each contact was recorded at each 10% extrusion length and recorded as a heatmap, scaled to represent the difference in satisfaction of each predicted contact in correct and incorrect models.

The difference in contact satisfaction between correct and incorrect models was calculated by taking the proportion of correct models minus the proportion of incorrect models containing each contact, at each point of model extrusion.

### 4.1.2 Tracking Forming and Breaking of Contacts

We randomly selected 10 models generated by standard SAINT2 for the soluble protein 1TZT: five correct and five incorrect models. We recorded the levels of All, LR-, and SR-SPCs at each 10% length of extrusion for each model.

We then calculated the lost contacts, as described in Equation 4.1 at every 10% extrusion for each of the 10 models to evaluate how contacts were being broken during extrusion.

$$\text{Lost Contacts} = \frac{BrokenContacts_{(n-1)\rightarrow n}}{SatisfiedContacts_{(n-1)}} \tag{4.1}$$

### 4.1.3 Analysing the Satisfaction of Individual Predicted Contacts

We used the Large Dataset, comprised of 245 proteins, as described in Table S3 and described previously in chapter 2. We generated 1,000 models for each protein and recorded which predicted contacts were satisfied in the final models.

A ROC (receiver operating characteristic) curve was generated using this set of 1,000 models. The threshold that was varied was the satisfaction rank of the predicted contact, which was used to classify a contact as correct or incorrect. The satisfaction rank is the rank of a predicted contact for a protein as given by its percentage satisfaction in the set of 1,000 models. This value is normalised by dividing by the number of total predicted contacts for that protein.

## 4.2 Satisfaction of Predicted Contacts in Correct and Incorrect Models

We first wanted to identify whether there were particular predicted contacts that were more likely to be satisfied in correct models, or incorrect models. If this was the case, the aim would be to attempt to identify whether a model was correct based on the specific predicted contacts satisfied. These contacts could be identified with properties such as position in sequence or PPV. Alternatively it would be useful to investigate whether there was any reason that some predicted contacts were not being satisfied, and whether adjusting the weighting of these contacts could increase the chances of producing a correct model. If there were contacts that were only ever satisfied in correct (or incorrect) models, this may provide interesting perspective on how SAINT2 is producing correct models and allow us to skew the process more towards producing correct predictions.

We show here the predicted contacts for the soluble protein 1AIU, ordered first by the position in sequence (Figure 4.1) and then by PPV (Figure 4.2). The heatmap is scaled

based on the percentage of correct models with each contact satisfied minus the percentage of incorrect models with each contact satisfied, thus giving us the difference in how often each contact is satisfied in correct and incorrect models. We used the protein 1AIU as it had 75 predicted contacts, a good compromise between having enough predicted contacts to be able to view how they were being satisfied during extrusion, but not too many that it would be hard to see what was happening on an individual level. Our SAINT2 run of 10,000 models produced 2,706 correct models - this was enough correct models to be able to compare how predicted contacts were being satisfied with those in incorrect models.

As expected, the majority of predicted contacts are more often satisfied in correct models - the scale of Figures 4.1 and 4.2 represents this as the asymmetrical range spans roughly 4% more satisfied in incorrect models to 30% more satisfied in correct models.

In general, those contacts available earlier in extrusion (during the first 50%) and in the final 10-20% are satisfied to similar levels in correct and incorrect models. Contacts available in the mid-lengths of extrusion (from around 60-80%) are mostly more satisfied in correct models. However, there are a couple of contacts within this group that are either more satisfied in incorrect models, or are satisfied to similar levels in all models. It would be interesting to investigate the properties of these contacts to establish why they appear to be associated more with incorrect models. There does not seem to be any pattern between the PPV of a predicted contact, and the difference in satisfaction between correct and incorrect models.

If it is the case that there are certain predicted contacts that are more important for correct structure prediction than others, it could be possible to bias our model generation algorithm towards satisfying those contacts that are more correlated with correct models. Alternatively we could down-weight contacts that we believe would be more likely to produce incorrect models. It would be interesting to see how predicted contacts are satisfied in individual models, to establish whether correct models are forming contacts early on in extrusion and keeping these contacts in order to maintain correct substructures, or whether

they are breaking them in order to sample a variety of conformations.

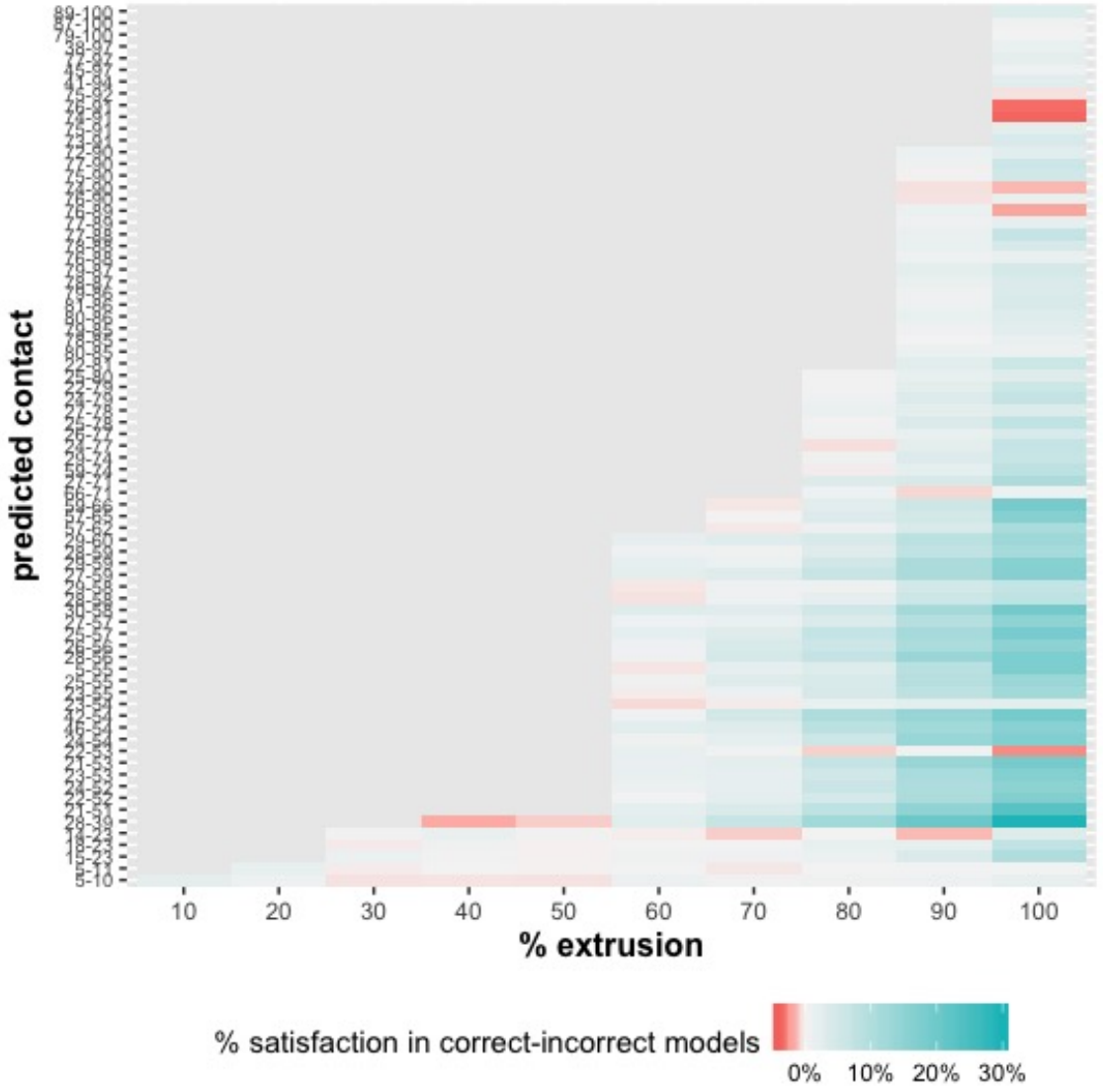



**Figure 4.1:** Heatmap displaying the predicted contacts for 1AIU, ordered by position in the sequence, scaled based on the percentage of correct models with each contact satisfied, minus the percentage satisfaction in incorrect models for each extrusion length from a run of 10,000 decoys. Grey areas on the figure represent extrusion lengths where a contact is not yet available.

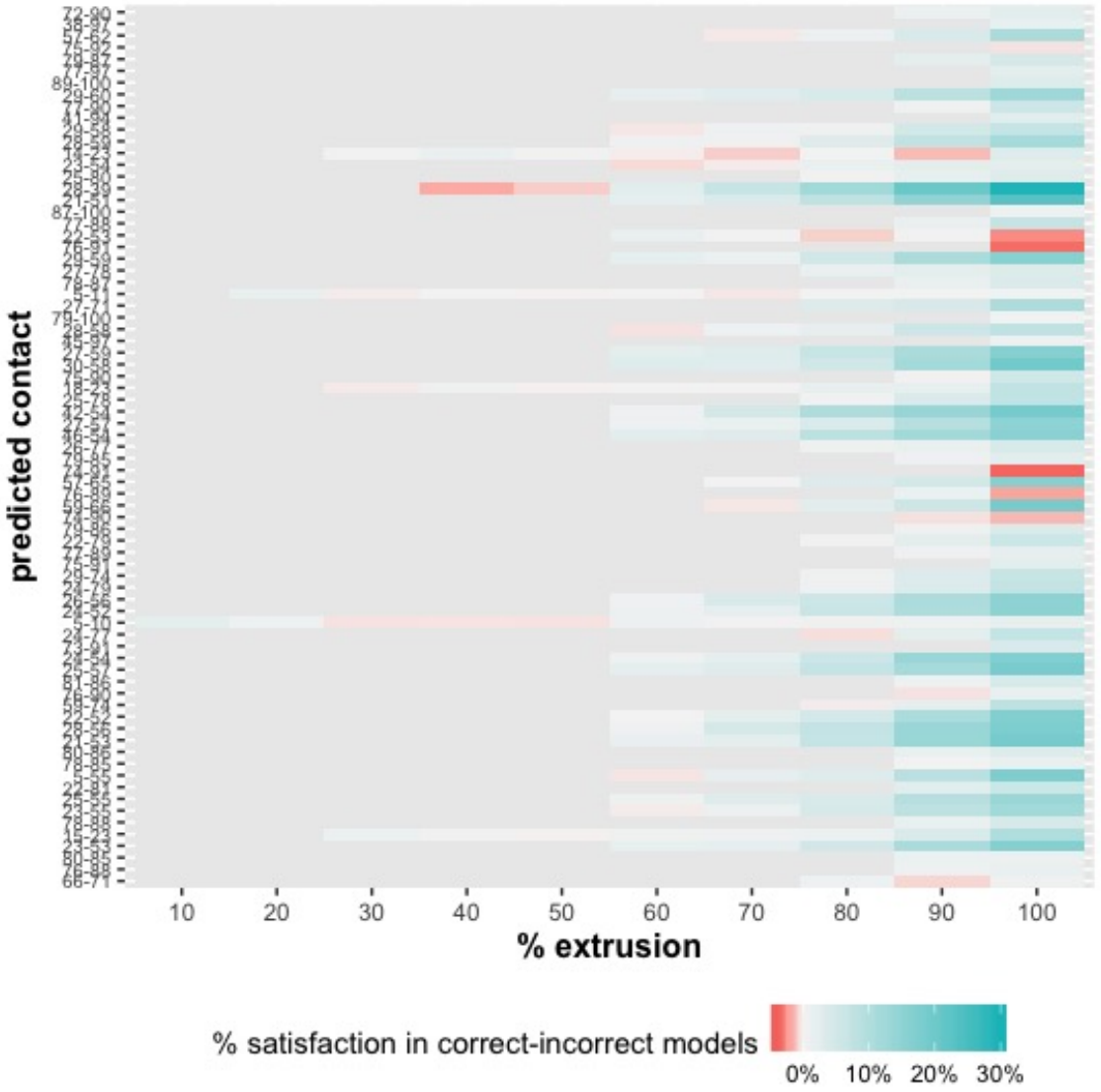



**Figure 4.2:** Heatmap displaying the predicted contacts for 1AIU, ordered by PPV (highest PPV at the bottom), scaled based on the percentage of correct models with each contact satisfied, minus the percentage satisfaction in incorrect models for each extrusion length from a run of 10,000 decoys. Grey areas on the figure represent extrusion lengths where a contact is not yet available.

## 4.3 Decoy Tracking and Lost Contacts

In order to further investigate the behaviour shown in Figures 4.1 and 4.2 that some contacts are satisfied differently in correct and incorrect models, we looked at the SPCs in individual models during generation. This may show us whether models form contacts early on in generation and maintain these contacts throughout extrusion, or whether they are constantly breaking and reforming contacts.

A random sample of 10 models with a good TM-Score (greater than 0.5), and 10 models with a poor TM-Score were selected from our set of decoys for the protein 1TZT. We selected this protein as it was a mid-size protein (126 residues), with 59 predicted contacts. The SPCs for each snapshot were recorded separately for each model.

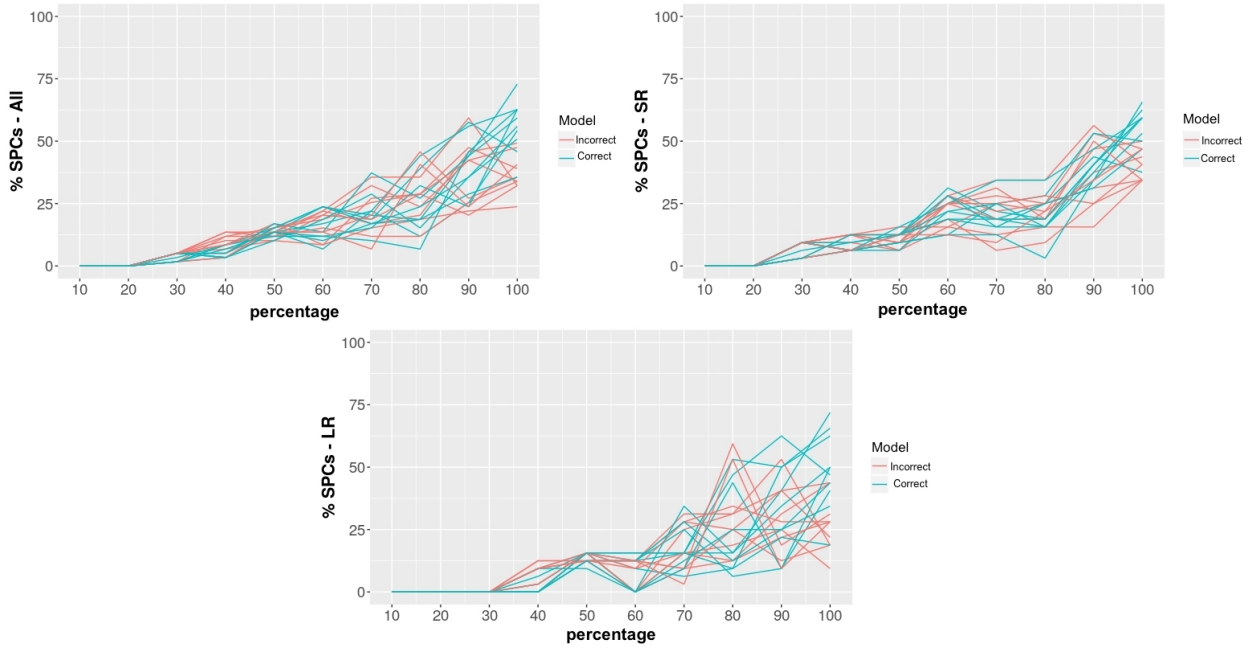



**Figure 4.3:** Proportion of predicted contacts available that were satisfied at each 10% length of decoy generated. Data from 20 Decoys, 10 correct (red) and 10 incorrect (blue), for the soluble protein 1TZT; (top left) All contacts; (top right) Short-range contacts and (bottom) long-range contacts.

Here we considered how predicted contacts are satisfied during the production of a particular model. 1TZT has 59 predicted contacts, of which 51 were correct predictions. Of these predicted contacts, 32 are LR and 27 are SR. We again assessed the behaviour of 10 models (five correct and five incorrect), shown in Figure 4.3.

While it is a very small sample, we observe little difference in the behaviour of SPCs between correct models and incorrect models for this case. Consistent with previous results, in the final 10-20% extrusion, correct models have more SPCs than incorrect models. For LR-SPCs only, there is a lot more variation in the levels of satisfaction as extrusion occurs. For the final 30% of extrusion it does seem that LR contacts are satisfied to higher levels in correct models. This variability suggests that there may be more breaking and forming of LR contacts as models are generated.

To try and answer this question, we considered another method of tracking individual decoys. We monitored the predicted contacts that are broken at a certain percentage length extrusion. Equation 4.1, was used to calculate the 'lost contacts' at each snapshot of decoy generation for each of the 20 decoys for the 1TZT protein that were used for decoy tracking.

In general, Figure 4.4 shows correct models breaking more contacts during the middle stages of model generation. Towards the final extrusion length, correct models break less contacts - this may be the result of being in a more stable conformation where the breaking of contacts would be more likely to have a negative effect on the scoring function. When looking specifically at SR contacts, incorrect models appear to break contacts far more often than correct models. This occurs throughout the entire length of extrusion. This may suggest that correct models are forming more stable sub-structures involving SR contacts, whereas incorrect models are not.

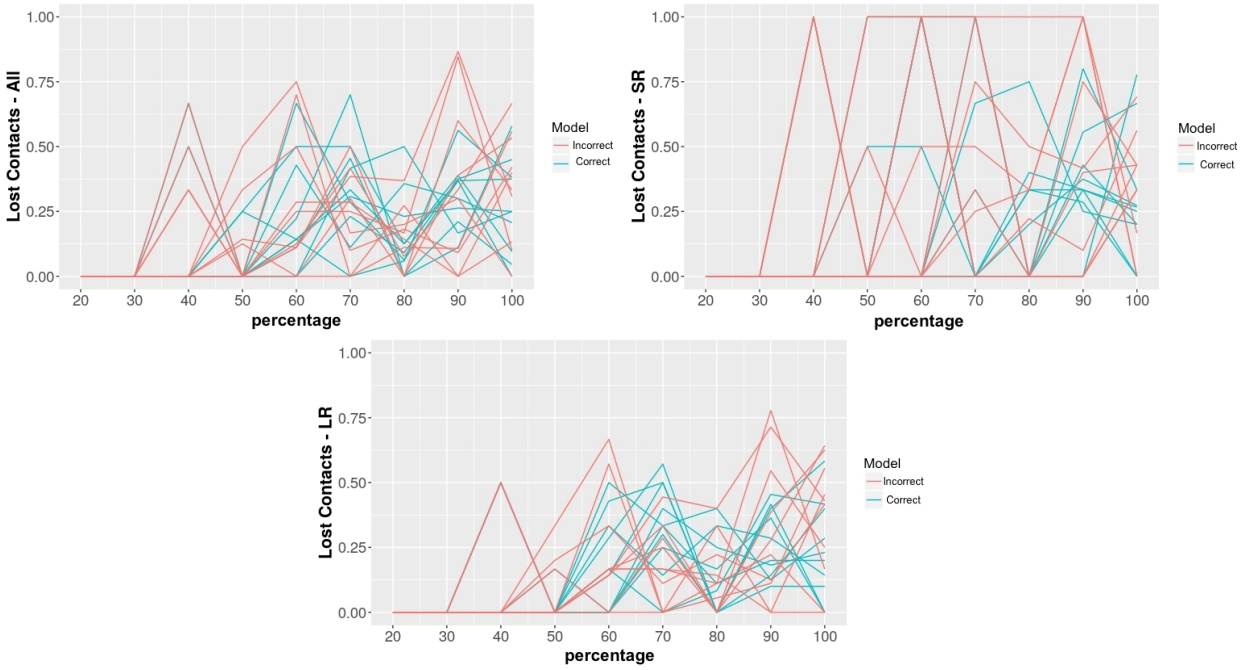



**Figure 4.4:** Lost contacts (Equation 4.1) at each 10% length of decoy generated. Data from 20 Decoys, 10 correct (red) and 10 incorrect(blue) for the soluble protein 1TZT. All contacts (top left); Short-range contacts (top right) and long-range contacts (bottom).

## 4.4 Satisfaction of Individual Predicted Contacts

We wanted to further investigate how predicted contacts were being satisfied in models generated by SAINT2. If there was any calculable difference to the predicted contacts present in models, it may be possible to use this to either identify correct models, or bias the production of models to those that include any 'beneficial' contacts.

We wanted to explore whether correct models satisfied different predicted contacts to incorrect models. We also wanted to investigate whether it would be possible to identify whether a predicted contact was incorrect based on information about its PPV, position in protein sequence, range between the two residues involved in the contact, or level of satisfaction in a sample of models. To do this, we recorded the presence of each predicted contact in 1000 models, for each protein in the Larger Dataset of 245 proteins, described

in Table S3. For each case, we generated a heatmap, indicating whether each contact was satisfied. The incorrect contacts were marked with $*$ and the predicted contacts were sorted first by PPV, then position in the sequence, range of the contact, and the level of satisfaction.

Figures 4.5-4.8 show the relationship between PPV, position, range and level of satisfaction, and whether a contact is correct, for the protein 1TZT. This protein was selected as it displays a relatively typical pattern compared to the other proteins in the Larger Dataset. In agreement with earlier work in chapter 3, correct models have more contacts satisfied than incorrect models. However, there does not appear to be any obvious pattern for specific contacts that this applies to - for example, contacts that would be more integral to producing a correct structure prediction.

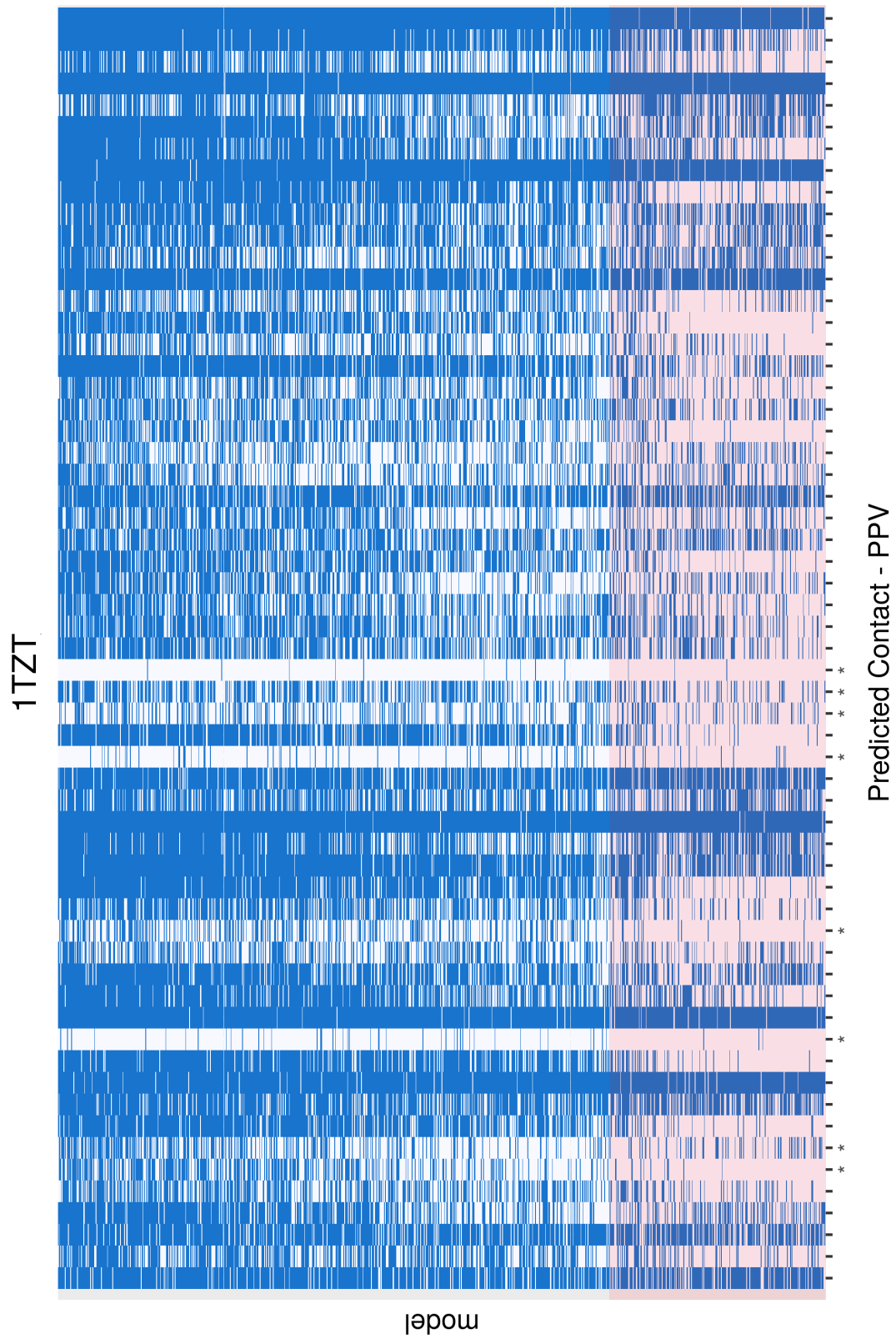



**Figure 4.5:** Heatmap displaying the predicted contacts for 1TZT, ordered from low to high PPV for 1,000 models, ordered from low to high TM-score. Predicted contacts marked in blue indicate satisfaction, white indicated unsatisfied. Models within the red box are incorrect and those outside the red box are correct. Predicted contacts marked with $*$ are incorrect.

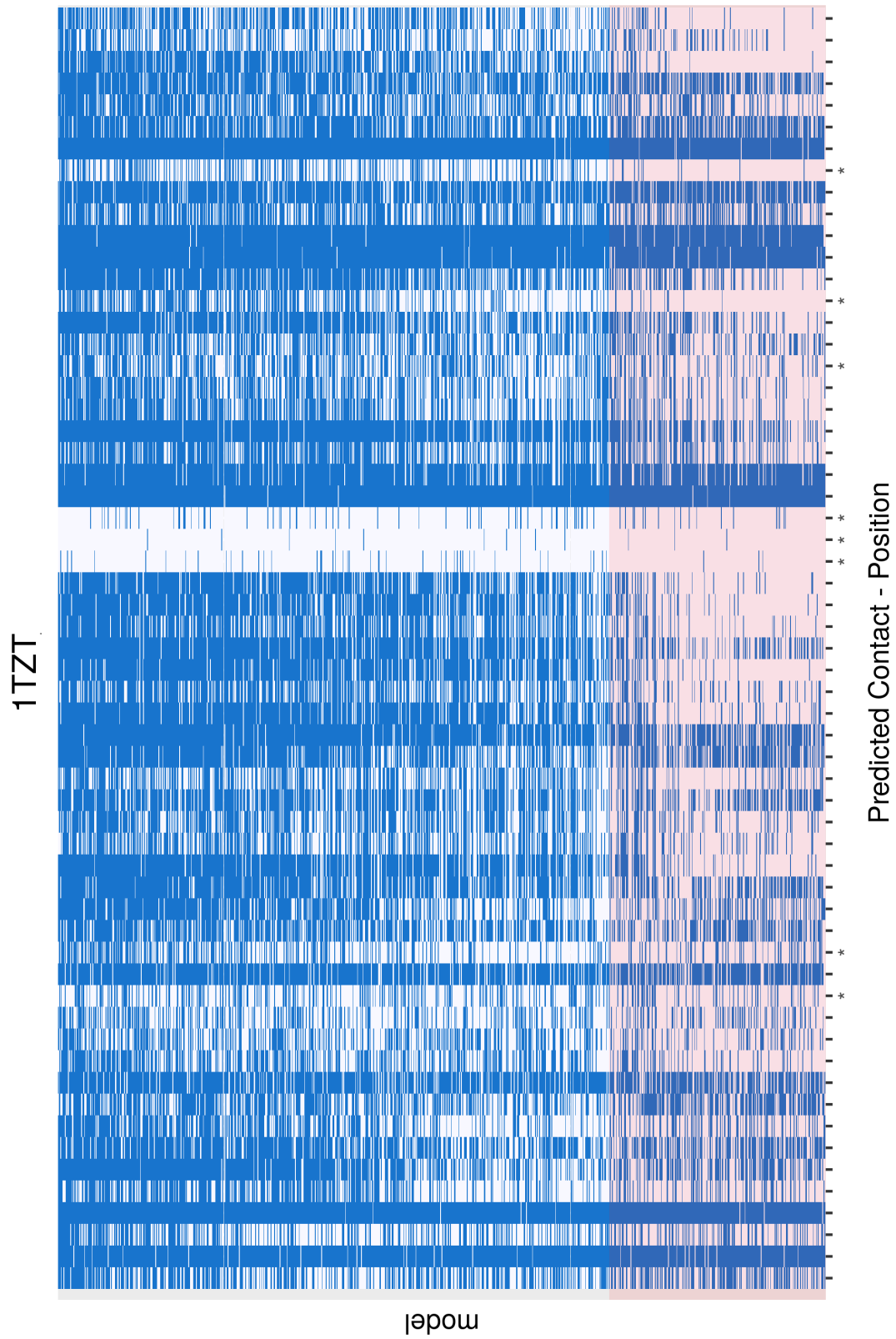



**Figure 4.6:** Heatmap displaying the predicted contacts for 1TZT, ordered from early to late position in protein sequence for 1,000 models, ordered from low to high TM-score. Predicted contacts marked in blue indicate satisfaction, white indicated unsatisfied. Models within the red box are incorrect and those outside the red box are correct. Predicted contacts marked with $*$ are incorrect.

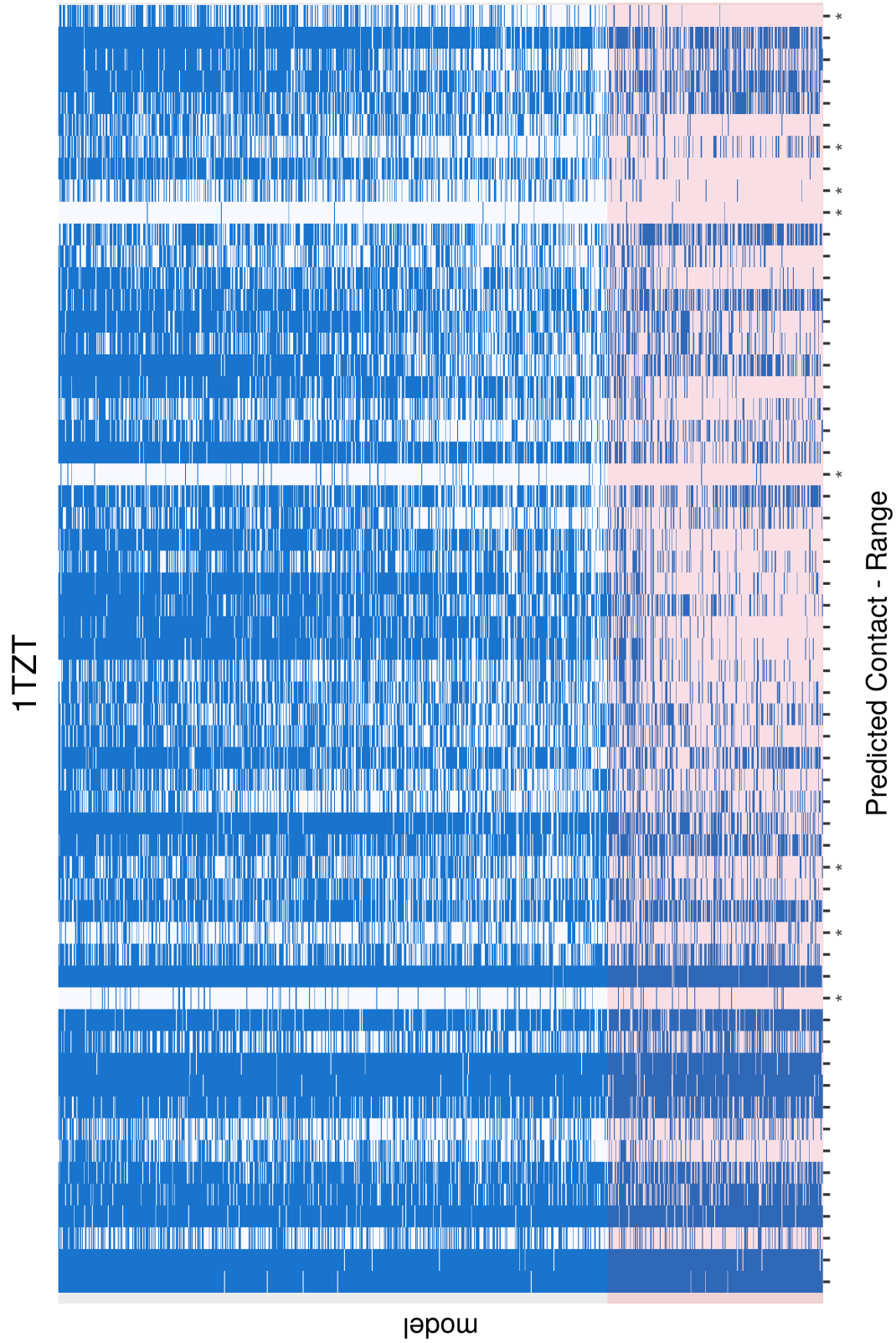



**Figure 4.7:** Heatmap displaying the predicted contacts for 1TZT, ordered from low to high residue range for 1,000 models, ordered from low to high TM-score. Predicted contacts marked in blue indicate satisfaction, white indicated unsatisfied. Models within the red box are incorrect and those outside the red box are correct. Predicted contacts marked with $*$ are incorrect.

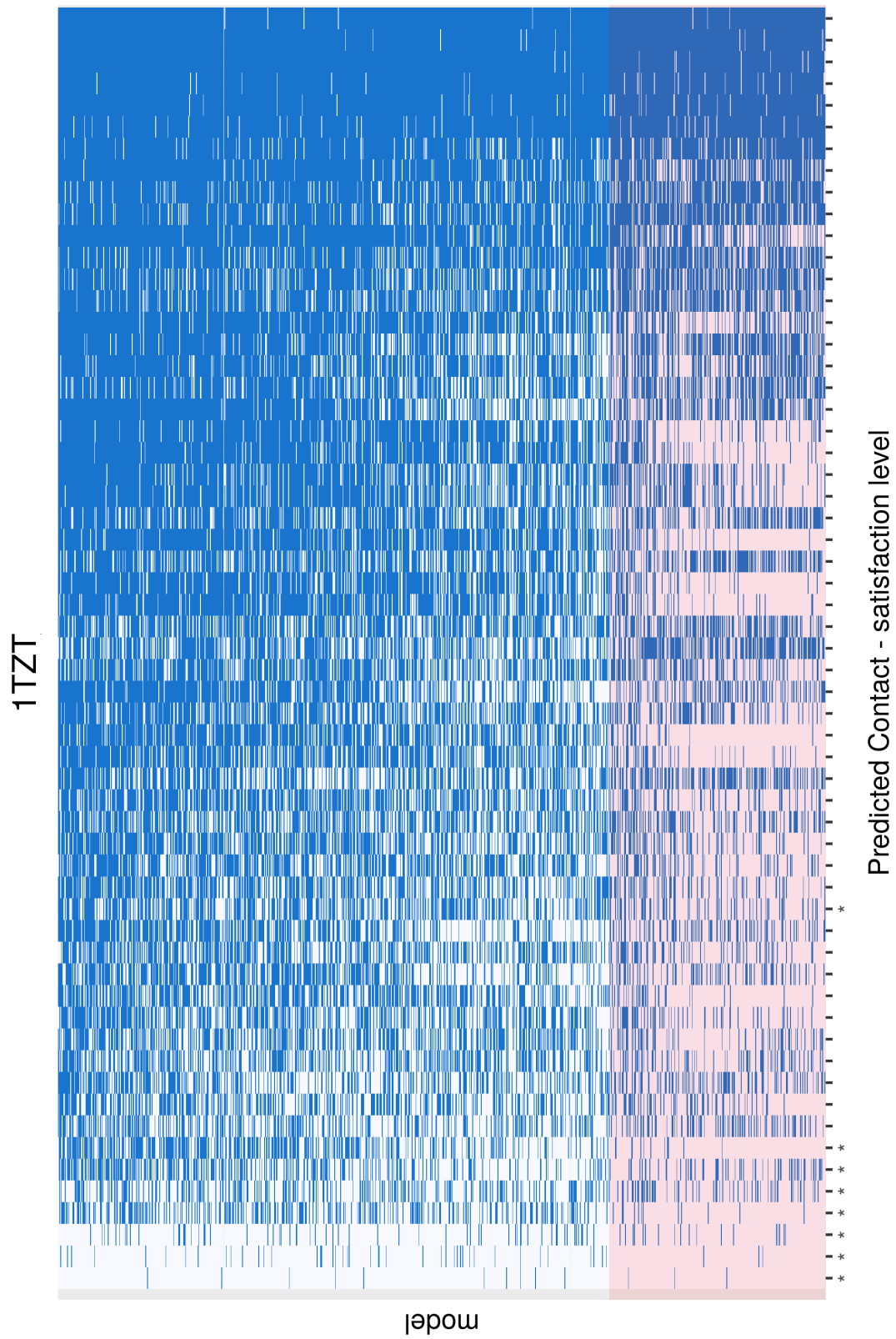



**Figure 4.8:** Heatmap displaying the predicted contacts for 1TZT, ordered from low to high satisfaction for 1,000 models, ordered from low to high TM-score. Predicted contacts marked in blue indicate satisfaction, white indicated unsatisfied. Models within the red box are incorrect and those outside the red box are correct. Predicted contacts marked with $*$ are incorrect.

The distribution of incorrect predicted contacts appears to be relatively random with regards to position in protein sequence, or range of the contact. While there is also a similarly random appearance to the distribution of incorrect contacts, there are generally more incorrect contacts with lower PPV scores. When sorting predicted contacts by the level of satisfaction, the majority of incorrect predicted contacts are towards the lower levels of satisfaction, as displayed in Figure 4.8.

Table 4.1 shows that as expected, incorrect predicted contacts are generally satisfied to lower levels in correct and incorrect models. While the levels of satisfaction are similar for incorrect contacts for all models, correct predicted contacts are satisfied to higher levels in correct models compared to incorrect models.

| **Contacts** | **Correct Models** | **Incorrect Models** |
|---|---|---|
| **Correct** | 54% | 46% |
| **Incorrect** | 27% | 26% |

**Table 4.1:** Average satisfaction of correct and incorrect predicted contacts in correct and incorrect models for the larger dataset of 245 proteins.

In order to investigate whether it would be possible to identify predicted contacts likely to be incorrect based on levels of satisfaction in a run of models, we plotted a ROC curve, varying the threshold of satisfaction of a contact required for us to label it as correct or incorrect (Figure 4.9). Rather than using the raw value of satisfaction of the predicted contact, we used the rank of the contact when sorting by the level of satisfaction over 1000 models in a protein, normalised using the number of predicted contacts for that protein. This would allow us to compare the values over all proteins.

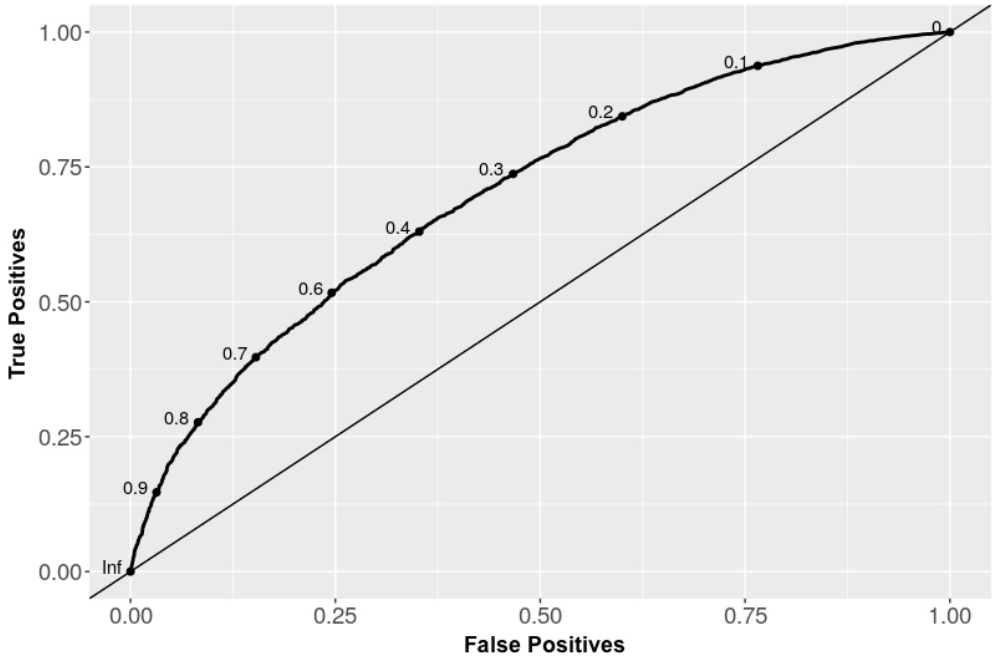



**Figure 4.9:** ROC curve varying the threshold of satisfaction rank of predicted contacts for all 245 proteins in the Large Dataset.

Figure 4.9 shows that especially at higher threshold values for predicted contact rank, there are higher levels of true positives than false positives. For example, using the cut-off of 0.6 (classifying those predicted contacts ranked over 60%, or in the top 40%, as correct and those under that as incorrect), gives roughly 50% true positives, and 25% false positives. This would mean that we would correctly identify around half of the true predicted contacts, and around 25% of the incorrect predicted contacts as being true contacts. Using a less stringent cut-off of 0.3 (classifying the top 70% ranked predicted contacts as correct), we identify 75% of the correct predicted contacts and under 50% of the incorrect predicted contacts.

We continued by investigating the pattern seen above in Figure 4.8, that incorrect predicted contacts generally had lower levels of satisfaction than correct predicted contacts. We compared the satisfaction rank of predicted contacts, and whether they were correct (Figure 4.10). We also used a filtered list of the models, produced using a model quality assessment protocol based on contact map alignment, described in [62].

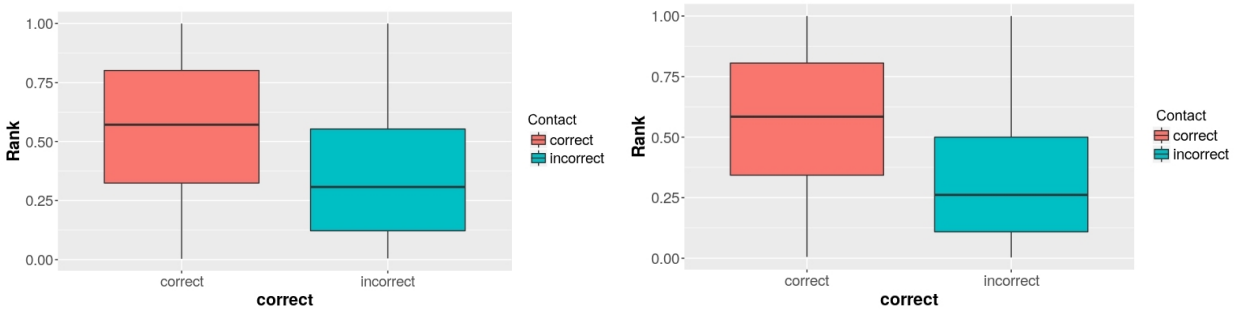



**Figure 4.10:** Boxplots describing the satisfaction ranks for correct predicted contacts (red) and incorrect predicted contacts (blue) for 1,000 models (left) and a filtered subset of models (right) for 245 proteins in the Large Dataset.

Figure 4.10 shows that incorrect predicted contacts are generally ranked lower than correct predicted contacts. After filtering to a reduced set of models more likely to be correct, there is a larger difference in the satisfaction rank of the contacts in correct and incorrect predicted contacts, suggesting that models produced by SAINT2 are satisfying correct predicted contacts more often.

## 4.5 Discussion for Investigating the Behaviour of Predicted Contacts

### 4.5.1 Behaviour of Predicted Contacts during Model Extrusion

We had previously observed in Chapter 3 that there were cases that SAINT2 could predict no good structures for, despite having accurate contact predictions. We therefore wanted to investigate how contacts were being satisfied and broken during model extrusion.

We observed in Figure 4.1 that as models are extruded, consistent with our previous observations, most predicted contacts are satisfied more in correct models than incorrect models. This was more likely to be the case with predicted contacts that became available in the middle regions of the sequence (from around 60-80% extrusion). With these predicted

contacts, as extrusion increased, there was more of a difference between how often the contacts were satisfied in correct models compared to incorrect models. It may be the case that the breaking of other contacts that have trapped the model in an incorrect formation is required in order to allow these contacts to form.

There are some contacts that are satisfied to similar levels in correct and incorrect models. In general, these are either earlier (during the first 50% extrusion), or in the final 10-20% extrusion. With the earlier contacts, this is likely because they are SR contacts that occur within sections of ordered secondary structure, such as within $\alpha$-helices, and thus are easier to satisfy. For the contacts that occur later in extrusion, we may not be giving enough computational time, or moves to SAINT2 during these later steps in order to sample the conformational space and satisfy these contacts.

There are a small subset of predicted contacts that are satisfied more often in incorrect models. These contacts are all correct, and the majority involve residues in a similar area of the protein chain, in the last loop section. It may be the case that incorrect models may be in a conformation that allows more flexibility with satisfying these contacts, whereas correct models may not being given enough computational time to accurately sample these final lengths of extrusion.

Our observations from Figure 4.2 do not show any correlation between the PPV of a predicted contact, and whether it is more likely to be satisfied in correct or incorrect models. It may be the case that looking at other cases, with more contacts that are satisfied to higher levels in incorrect models may provide more information between any link with PPV.

In order to further investigate whether there was any difference in how contacts were being formed in correct and incorrect models, we looked how often satisfied contacts formed (Figure 4.3) and were being broken (Figure 4.4) in individual models.

Our results show that while there is a general upward trend with each model as to how many contacts were satisfied as extrusion occurred, there are some lengths where there is a

decrease in the percentage of SPCs. This could be due to contact breakage, or new contacts becoming available at that extrusion length. Again, it appears as though satisfying LR contacts during extrusion is indicative of a final decoy with a correct TM-score. This is particularly the case during the middle lengths of extrusion, when there is a greater distinction between the contacts satisfied by correct and incorrect models. However, it may be a good idea to track more decoys, to see whether this changes. We provide two explanations for this finding; either we are losing contacts that were previously satisfied towards the end of extrusion, or contacts formed near the C-terminus are not being given enough moves to be satisfied.

We looked at the proportion of contacts that were lost at each stage of extrusion in individual models; we observed that correct models where breaking less contacts as extrusion occured, while incorrect models seemed to be breaking contacts fairly consistently throughout extrusion. This may suggest that correct models are forming more stable conformations. Particularly, looking at breaking SR contacts suggests that correct models break these far less often, and therefore may be forming more stable sub-structures that incorrect models are not. However, conclusions are hard to draw given our sample size. It would be useful to look at more models, more proteins, and to look at the correct contacts separately from the incorrect ones.

### 4.5.2 Satisfaction of Individual Predicted Contacts

In general, the majority of predicted contacts are more often satisfied in correct models than incorrect (Figures 4.5-4.8). This is not surprising as previous results in this report have shown a similar pattern. It may be interesting to compare the difference in density of the satisfaction values from the heatmaps for each predicted contact, between the correct and incorrect models. This could provide further information about the predicted contacts that behave differently in correct and incorrect models, as well as those that behave in the same way.

When ordering the predicted contacts based on PPV, position, range and satisfaction

level, there was a correlation between predicted contacts that were satisfied the least, and the predicted contact being incorrect. When investigating this further, for all of the 245 proteins in the Dataset, there were 21 predicted contacts that were never satisfied in any model. Of these, 16 were incorrect and 5 were correct. When using SAINT2 to predict a structure, we do no know which predicted contacts will be correct or incorrect. If we can identify incorrect predicted contacts using some metric and discard them, we may be able to increase model quality.

It could be possible to implement a cut-off based on the level of satisfaction of a predicted contact within a run of SAINT2. If a predicted contact is not satisfied at all, or only at a low level, then it may be useful to discard the predicted contact for subsequent runs of SAINT2 for the same protein, which would hopefully lead to an increase in model quality. Figure 4.9 shows that at a number of cut-off values, we retain a higher fraction of true positive contacts, than false positive. However, there are generally more correct predicted contacts than incorrect (9756 correct, 3084 incorrect), and therefore this may actually lead to more correct predicted contacts being discarded than incorrect.

Ranking the predicted contacts based on the level of satisfaction within a run of models, we see that incorrect predicted contacts are generally ranked lower than correct predicted contacts. We may be able to use this satisfaction rank when creating a cut-off, as described above. For example, as shown in Figure 4.10, we could dispose of over 50% of the incorrect predicted contacts while keeping over 75% of the correct contacts. It would be interesting to investigate whether we could increase the quality of models produced, by using a cut-off to refine the list (by changing the weights) of predicted contacts used as input into SAINT2 for further runs.

During this Chapter, we have only considered contacts as a binary classification - if two residues are closer than 8Å, we say a contact is present. It may be useful instead, to analyse the distances between two residues predicted to be in contact and compare differences with these values in correct and incorrect models.

# 5

# Conclusions and Future Work

Satisfied predicted contacts can be used as a measure of the quality of a final structure prediction, and the results of this thesis show that this could also be the case for partial models. Discarding models that fail to satisfy enough predicted contacts leads to enrichment of correct models in some cases. The question remains as to whether there are specific properties that would be correlated with a higher rate of success and enrichment when using our new SAINT2 protocol.

Preliminary results suggest that proteins with more predicted contacts (output by metaPSICOV), or certain SCOP folds (all $\beta$ or $\alpha/\beta$) are more enriched with the new protocol. It would be interesting to investigate whether the new SAINT2 protocol could be used selectively with proteins with these properties. In the case of number of predicted contacts, it would be interesting to see whether increasing the number of predicted contacts

- perhaps by including metaPSICOV predictions with a lower PPV - could increase the enrichment produced in more cases.

Looking at specific predicted contacts, it may be possible to identify incorrect predicted contacts based on how often they are satisfied during a run of SAINT2. If these less-satisfied contacts are discarded, it may lead to more accurate predicted contacts, and consequently better quality models.

However, much of this work is dependent on the binary definition of a contact between residues 8Å apart. It may be the case that changing how a contact is defined, either by using a different threshold (perhaps 5Å between side chains), or a linear score based on the distance between two residues, affects how the protocol enriches correct models.

It may also be the case that we are not allowing SAINT2 enough moves at longer extrusion steps to sufficiently sample the conformational space. One way to investigate this would be to observe the quality of the models at different extrusion steps and see whether this changes if we allow more moves. Other further work could involve the inclusion of more features into SAINT2 - for example there are currently no refinement or model selection stages.

Satisfied predicted contacts provide much potential for indicating the quality of a model during extrusion - saving computational time by discarding incorrect models early on, and guiding the generation of models to more correct structures. There is much more work to be done in implementing this, and finding the cases where using satisfied contacts would be most applicable, but the results in this thesis provide many different interesting questions to be answered!

# References


[1] L. Slabinski, L. Jaroszewski, P. C. Rodrigues, L. Rychlewski, I. Wilson, S. Lesley, and A. Godzik, "The challenge of protein structure determination - lessons from structural genomics," *Protein Science*, vol. 16, no. 11, pp. 2472–2482, 2007.

[2] D. A. Benson, M. Cavanaugh, K. Clark, I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, and E. W. Sayers, "GenBank," *Nucleic Acids Research*, vol. 41, no. D1, 2013.

[3] P. W. Rose, A. Prlić, A. Altunkaya, C. Bi, A. R. Bradley, C. H. Christie, L. D. Costanzo, J. M. Duarte, S. Dutta, Z. Feng, R. K. Green, D. S. Goodsell, B. Hudson, T. Kalro, R. Lowe, E. Peisach, C. Randle, A. S. Rose, C. Shao, Y.-P. Tao, Y. Valasatava, M. Voigt, J. D. Westbrook, J. Woo, H. Yang, J. Y. Young, C. Zardecki, H. M. Berman, and S. K. Burley, "The RCSB protein data bank: integrative view of protein, gene and 3D structural information.," *Nucleic Acids Research*, vol. 45, no. D1, pp. 271–281, 2016.

[4] J. Moult, K. Fidelis, A. Kryshtafovych, T. Schwede, and A. Tramontano, "Critical assessment of methods of protein structure prediction: Progress and new directions in round XI," *Proteins: Structure, Function and Bioinformatics*, vol. 84, pp. 4–14, 2016.

[5] R. D. Finn, P. Coggill, R. Y. Eberhardt, S. R. Eddy, J. Mistry, A. L. Mitchell, S. C. Potter, M. Punta, M. Qureshi, A. Sangrador-Vegas, G. A. Salazar, J. Tate, and A. Bateman, "The Pfam protein families database: Towards a more sustainable future," *Nucleic Acids Research*, vol. 44, no. D1, pp. D279–D285, 2016.

[6] S. Ovchinnikov, H. Park, N. Varghese, P.-S. Huang, G. A. Pavlopoulos, D. E. Kim, H. Kamisetty, N. C. Kyrpides, and D. Baker, "Protein structure determination using metagenome sequence data," *Science*, vol. 355, no. 6322, pp. 294–298, 2017.

[7] J. C. Norvell and J. M. Berg, "Update on the Protein Structure Initiative," *Structure*, vol. 15, no. 12, pp. 1519–1522, 2007.

[8] S. H. P. de Oliveira, J. Shi, and C. M. Deane, "Comparing co-evolution methods and their application to template-free protein structure prediction," *Bioinformatics*, vol. 33, no. 3, pp. 373–381, 2016.

[9] D. S. Marks, L. J. Colwell, R. Sheridan, T. A. Hopf, A. Pagnani, R. Zecchina, and C. Sander, "Protein 3D structure computed from evolutionary sequence variation," *PLoS ONE*, vol. 6, no. 12, p. E28766, 2011.

[10] D. S. Marks, T. A. Hopf, and C. Sander, "Protein structure prediction from sequence variation," *Nature Biotechnology*, vol. 30, no. 11, pp. 1072–1080, 2012.

[11] D. T. Jones and L. J. McGuffin, "Assembling Novel Protein Folds From Super-secondary Structural Fragments," *Proteins: Structure, Function and Genetics*, vol. 53, no. SUPPL. 6, pp. 480–485, 2003.

[12] S. Raman, R. Vernon, J. Thompson, M. Tyka, R. Sadreyev, J. Pei, D. Kim, E. Kellogg, F. Dimaio, O. Lange, L. Kinch, W. Sheffler, B. H. Kim, R. Das, N. V. Grishin, and D. Baker, "Structure prediction for CASP8 with all-atom refinement using Rosetta," *Proteins: Structure, Function and Bioinformatics*, vol. 77, no. SUPPL. 9, pp. 89–99, 2009.

[13] S. H. De Oliveira, E. C. Law, J. Shi, and C. M. Deane, "Sequential search leads to faster,

more efficient fragment-based de novo protein structure prediction," *Bioinformatics*, vol. 34, no. 7, pp. 1132–1140, 2018.

[14] D. E. Kim, F. Dimaio, R. Yu-Ruei Wang, Y. Song, and D. Baker, "One contact for every twelve residues allows robust and accurate topology-level protein structure modeling," *Proteins: Structure, Function and Bioinformatics*, vol. 82, no. SUPPL.2, pp. 208–218, 2014.

[15] B. Adhikari and J. Cheng, "Protein residue contacts and prediction methods," *Methods in Molecular Biology*, vol. 1415, pp. 463–476, 2016.

[16] S. Miyazawa and R. L. Jernigan, "Residue – Residue Potentials with a Favorable Contact Pair Term and an Unfavorable High Packing Density Term, for Simulation and Threading," *Journal of Molecular Biology*, vol. 256, no. 3, pp. 623–644, 1996.

[17] B. Adhikari, D. Bhattacharya, R. Cao, and J. Cheng, "CONFOLD: Residue-residue contact-guided ab initio protein folding," *Proteins: Structure, Function and Bioinformatics*, vol. 83, no. 8, pp. 1436–1449, 2015.

[18] A. Kryshtafovych, K. Fidelis, and J. Moult, "CASP10 results compared to those of previous CASP experiments," *Proteins: Structure, Function and Bioinformatics*, vol. 82, no. SUPPL.2, pp. 164–174, 2014.

[19] R. Zwanzig, A. Szabo, and B. Bagchi, "Levinthal's paradox.," *Proceedings of the National Academy of Sciences*, vol. 89, no. 1, pp. 20–22, 1992.

[20] S. H. P. De Oliveira, J. Shi, and C. M. Deane, "Building a better fragment library for de novo protein structure prediction," *PLoS ONE*, vol. 10, no. 4, pp. 1–20, 2015.

[21] D. T. Jones, "Predicting novel protein folds by using FRAGFOLD," *Proteins: Structure, Function and Genetics*, vol. 45, pp. 127–132, 2001.

[22] L. N. Kinch, W. Li, B. Monastyrskyy, A. Kryshtafovych, and N. V. Grishin, "Evaluation of free modeling targets in CASP11 and ROLL," *Proteins: Structure, Function and Bioinformatics*, vol. 84, no. Suppl 1, pp. 51–66, 2016.

[23] K. T. Simons, C. Kooperberg, E. Huang, and D. Baker, "Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and bayesian scoring functions," *Journal of Molecular Biology*, vol. 268, no. 1, pp. 209–225, 1997.

[24] D. Gront, D. W. Kulp, R. M. Vernon, C. E. M. Strauss, and D. Baker, "Generalized fragment picking in rosetta: Design, protocols and applications," *PLoS ONE*, vol. 6, no. 8, p. E23294, 2011.

[25] I. Kalev and M. Habeck, "HHfrag: HMM-based fragment detection using HHpred," *Bioinformatics*, vol. 27, no. 22, pp. 3110–3116, 2011.

[26] S. C. Li, D. Bu, X. Gao, J. Xu, and M. Li, "Designing succinct structural alphabets," *Bioinformatics*, vol. 24, no. 13, pp. i182–i189, 2008.

[27] C. A. Rohl, C. E. M. Strauss, K. M. S. Misura, and D. Baker, "Protein Structure Prediction Using Rosetta," *Methods in Enzymology*, vol. 383, no. 2003, pp. 66–93, 2004.

[28] J. E. Jones, "On the Determination of Molecular Fields. I. From the Variation of the Viscosity of a Gas with Temperature," *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 106, no. 738, pp. 441–462, 1924.

[29] R. Samudrala and J. Moult, "An all-atom distance-dependent conditional probability discriminatory function for protein structure prediction," *Journal of Molecular Biology*, vol. 275, no. 5, pp. 895–916, 1998.

[30] T. Lazaridis and M. Karplus, "Effective energy functions for protein structure prediction.," *Current Opinion in Structural Biology*, vol. 10, no. 2, pp. 139–45, 2000.

[31] S. C. Ngan, M. T. Inouye, and R. Samudrala, "A knowledge-based scoring function based on residue triplets for protein structure prediction," *Protein Engineering, Design and Selection*, vol. 19, no. 5, pp. 187–193, 2006.

[32] M. J. Skwark, A. Abdel-Rehim, and A. Elofsson, "PconsC: Combination of direct information methods and alignments improves contact prediction," *Bioinformatics*, vol. 29, no. 14, pp. 1815–1816, 2013.

[33] H. Park, J. Ko, K. Joo, J. Lee, C. Seok, and J. Lee, "Refinement of protein termini in template-based modeling using conformational space annealing," *Proteins: Structure, Function and Bioinformatics*, vol. 79, no. 9, pp. 2725–2734, 2011.

[34] L. Hovan, V. Oleinikovas, H. Yalinca, A. Kryshtafovych, G. Saladino, and F. Gervasio, "No Title," *Proteins: Structure, Function, and Bioinformatics*, pp. 1–16, 2017.

[35] U. Gobel, C. Sander, R. Schneider, and A. Valencia, "Correlated mutations and residue contacts in proteins," *Proteins: Structure, Function, and Bioinformatics*, vol. 18, no. 4, pp. 309–317, 1994.

[36] F. Morcos, A. Pagnani, B. Lunt, A. Bertolino, D. S. Marks, C. Sander, R. Zecchina, J. N. Onuchic, T. Hwa, and M. Weigt, "Direct-coupling analysis of residue coevolution captures native contacts across many protein families," *Proceedings of the National Academy of Sciences*, vol. 108, no. 49, pp. E1293–E1301, 2011.

[37] D. T. Jones, D. W. A. Buchan, D. Cozzetto, and M. Pontil, "PSICOV: Precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments," *Bioinformatics*, vol. 28, no. 2, pp. 184–190, 2012.

[38] S. Seemayer, M. Gruber, and J. Soding, "CCMpred - Fast and precise prediction of protein residue-residue contacts from correlated mutations," *Bioinformatics*, vol. 30, no. 21, pp. 3128–3130, 2014.

[39] B. Monastyrskyy, D. D'Andrea, K. Fidelis, A. Tramontano, and A. Kryshtafovych, "New encouraging developments in contact prediction: Assessment of the CASP11 results," *Proteins: Structure, Function and Bioinformatics*, vol. 84, no. Suppl. 1, pp. 131–144, 2015.

[40] D. T. Jones, T. Singh, T. Kosciolek, and S. Tetchner, "MetaPSICOV: Combining coevolution methods for accurate prediction of contacts and long range hydrogen bonding in proteins," *Bioinformatics*, vol. 31, no. 7, pp. 999–1006, 2015.

[41] M. J. Skwark, A. Abdel-Rehim, and A. Elofsson, "PconsC: Combination of direct information methods and alignments improves contact prediction," *Bioinformatics*, vol. 29, no. 14, pp. 1815–1816, 2013.

[42] M. J. Skwark, D. Raimondi, M. Michel, and A. Elofsson, "Improved Contact Predictions Using the Recognition of Protein Like Contact Patterns," *PLoS Computational Biology*, vol. 10, no. 11, p. e1003889, 2014.

[43] M. Michel, D. Menéndez Hurtado, K. Uziela, and A. Elofsson, "Large-scale structure prediction by improved contact predictions and model quality assessment," *Bioinformatics*, vol. 33, no. 14, pp. i23–i29, 2017.

[44] S. Wang, S. Sun, Z. Li, R. Zhang, and J. Xu, "Accurate De Novo Prediction of Protein Contact Map by Ultra-Deep Learning Model," *PLoS Computational Biology*, vol. 13, no. 1, p. e1005324, 2017.

[45] D. W. Buchan and D. T. Jones, "Improved protein contact predictions with the MetaPSICOV2 server in CASP12," *Proteins: Structure, Function and Bioinformatics*, vol. 86, pp. 78–83, 2018.

[46] S. Ovchinnikov, L. Kinch, H. Park, Y. Liao, J. Pei, D. E. Kim, H. Kamisetty, N. V. Grishin, and D. Baker, "Large-scale determination of previously unsolved protein structures using evolutionary information," *eLife*, vol. 4, p. e09248, 2015.

[47] D. C. Phillips, "The hen egg-white lysozyme molecule," *Proceedings of the National Academy of Sciences*, vol. 57, no. 3, pp. 484–95, 1967.

[48] W. R. Taylor, "Topological accessibility shows a distinct asymmetry in the folds of beta-alpha proteins," *FEBS Letters*, vol. 580, no. 22, pp. 5263–5267, 2006.

[49] Y. E. Kim, M. S. Hipp, A. Bracher, M. Hayer-Hartl, and F. Ulrich Hartl, "Molecular Chaperone Functions in Protein Folding and Proteostasis," *Annual Review of Biochemistry*, vol. 82, no. 1, pp. 323–355, 2013.

[50] E. Wang, J. Wang, C. Chen, and Y. Xiao, "Computational evidence that fast translation speed can increase the probability of cotranslational protein folding," *Scientific Reports*, vol. 5, p. 15316, 2015.

[51] A. N. Fedorov, "Cotranslational Protein Folding - Minireview," *Journal of Biological Chemistry*, vol. 272, no. 52, pp. 32715–32718, 1997.

[52] C. H. Yu, Y. Dang, Z. Zhou, C. Wu, F. Zhao, M. S. Sachs, and Y. Liu, "Codon Usage Influences the Local Rate of Translation Elongation to Regulate Co-translational Protein Folding," *Molecular Cell*, vol. 59, no. 5, pp. 744–754, 2015.

[53] Y. Zhang and J. Skolnick, "Scoring function for automated assessment of protein structure template quality.," *Proteins*, vol. 57, no. 4, pp. 702–10, 2004.

[54] J. Xu and Y. Zhang, "How significant is a protein structure similarity with TM-score = 0.5?," *Bioinformatics*, vol. 26, no. 7, pp. 889–895, 2010.

[55] A. G. Murzin, S. E. Brenner, T. Hubbard, and C. Chothia, "SCOP: A structural classification of proteins database for the investigation of sequences and structures," *Journal of Molecular Biology*, vol. 247, no. 4, pp. 536–540, 1995.

[56] M. A. Lomize, A. L. Lomize, I. D. Pogozheva, and H. I. Mosberg, "OPM: Orientations of proteins in membranes database," *Bioinformatics*, vol. 22, no. 5, pp. 623–625, 2006.

[57] D. T. Jones, "Protein secondary structure prediction based on position-specific scoring matrices," *Journal of Molecular Biology*, vol. 292, no. 2, pp. 195–202, 1999.

[58] E. Faraggi, Y. Yang, S. Zhang, and Y. Zhou, "Predicting continuous local structure and the effect of its substitution for secondary structure in fragment-free protein structure prediction.," *Structure (London, England : 1993)*, vol. 17, no. 11, pp. 1515–27, 2009.

[59] S. F. Altschul, T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman, "Gapped BLAST and PSI-BLAST: A new generation of protein database search programs," *Nucleic Acids Research*, vol. 25, no. 17, pp. 3389–3402, 1997.

[60] L. Kaján, T. A. Hopf, M. Kalaš, D. S. Marks, and B. Rost, "FreeContact: fast and free software for protein contact prediction from residue co-evolution," *BMC Bioinformatics*, vol. 15, no. 1, p. 85, 2014.

[61] H. Kamisetty, S. Ovchinnikov, and D. Baker, "Assessing the utility of coevolution-based residue-residue contact predictions in a sequence- and structure-rich era," *Proceedings of the National Academy of Sciences*, vol. 110, no. 39, pp. 15674–15679, 2013.

[62] S. de Oliveira, C. E. West, and C. M. Deane, "Model quality assessment using alignments of predicted contact maps - Unpublished.".

# A

# SI

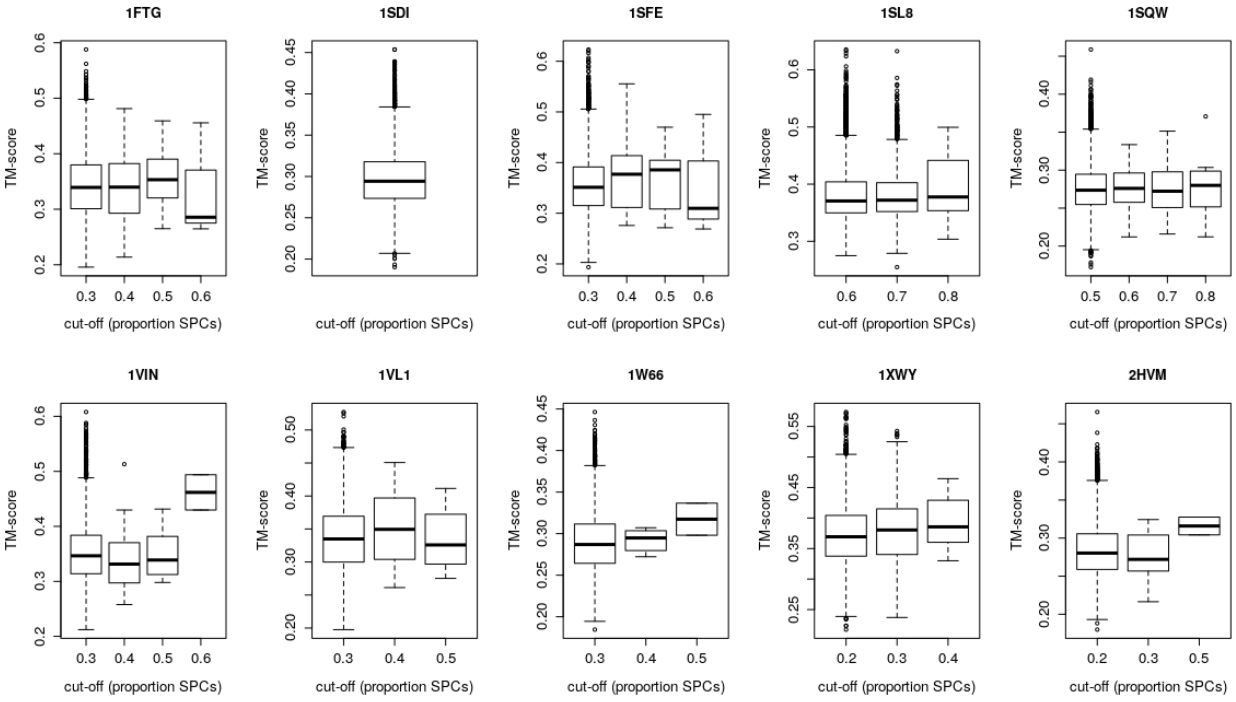



**Figure S1:** TM scores of the decoys produced at each step of cut-off for 11 soluble proteins, using the stepped cut-off to optimal value protocol with an added time regression

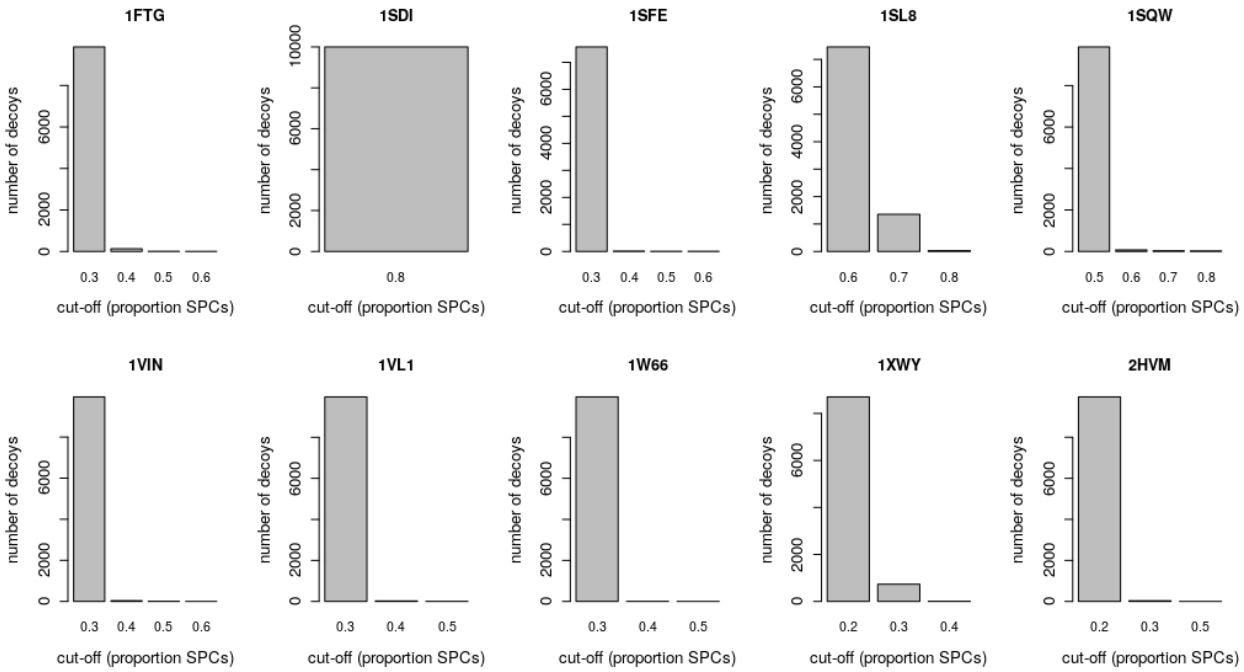



**Figure S2:** Total number of complete decoys produced at each step of cut-off for 11 soluble proteins, using the stepped cut-off to optimal value protocol with an added time regression

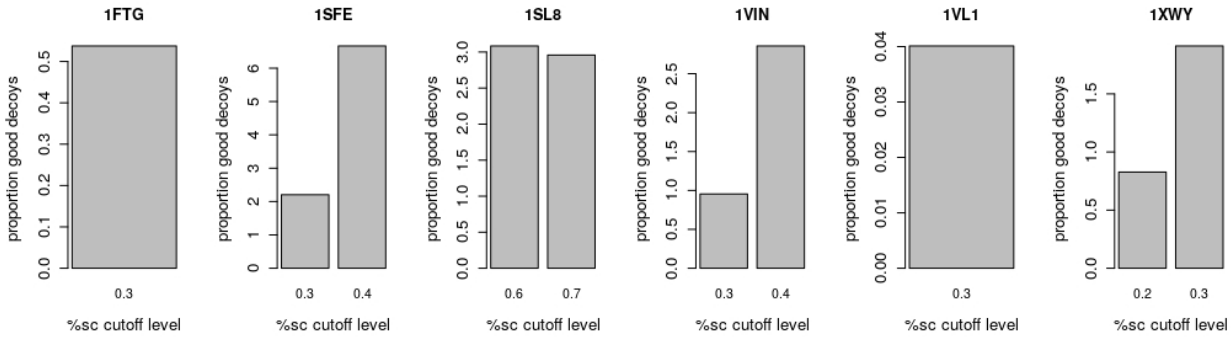



**Figure S3:** Proportion of correct models produced at each step of cut-off for 11 soluble proteins, using the stepped cut-off to optimal value protocol with an added time regression

| | PDB code | $N_{eff}$ | Precision @ 50% | | | Fragment Library | | Precision | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | All | LR | SR | Precision | Coverage | H | B | O | L |
| **Good** | 1VIN | 113 | 0.00 | 0.00 | 0.00 | 0.70 | 0.99 | 0.89 | NA | 0.61 | 0.37 |
| | 1SFE | 178 | 7.69 | 8.82 | 8.51 | 0.60 | 0.99 | 0.95 | 0.46 | 0.39 | 0.33 |
| | 1SL8 | 468 | 0.00 | 0.00 | 0.00 | 0.74 | 0.93 | 0.87 | NA | 0.61 | 0.43 |
| | 1VL1 | 111 | 9.52 | 4.35 | 7.69 | 0.51 | 0.99 | 0.88 | 0.56 | 0.40 | 0.28 |
| | 1FTG | 463 | 81.63 | 81.82 | 81.69 | 0.53 | 0.92 | 0.86 | 0.68 | 0.40 | 0.36 |
| | 1XWY | 317 | 45.00 | 0.00 | 29.67 | 0.64 | 1.00 | 0.94 | 0.41 | 0.56 | 0.35 |
| | 1JKS | 1745 | 100.00 | 93.55 | 94.59 | 0.25 | 0.86 | NA | 0.37 | 0.18 | 0.15 |
| **Bad** | 1SDI | 10 | 33.33 | 75.00 | 57.14 | 0.75 | 0.98 | 0.90 | NA | 0.45 | 0.31 |
| | 1SQW | 19 | 100.00 | 100.00 | 100.00 | 0.48 | 0.99 | 0.70 | 0.48 | 0.47 | 0.30 |
| | 1W66 | 128 | 88.89 | 53.12 | 66.00 | 0.50 | 1.00 | 0.96 | 0.36 | 0.36 | 0.28 |
| | 2HVM | 91 | 85.71 | 85.71 | 85.71 | 0.52 | 0.99 | 0.89 | 0.53 | 0.54 | 0.28 |
| | 1MSK | 29 | 0.00 | 2.78 | 2.13 | 0.52 | 0.96 | 0.89 | 0.28 | 0.47 | 0.29 |
| | 1BMG | 345 | NA | 88.89 | 88.89 | 0.45 | 1.00 | NA | 0.68 | 0.34 | 0.31 |
| | 1NEP | 47 | NA | 100.00 | 100.00 | 0.54 | 1.00 | NA | 0.76 | 0.46 | 0.32 |

| | PDB code | Coverage | | | Secondary Structure | | Torsion Angles | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | H | B | L | Precision | 3-state prec. | $\phi$ | $\psi$ | $\phi$ loop | $\psi$ loop |
| **Good** | 1VIN | 0.99 | NA | 0.99 | 69% | 84% | 27.40 | 66.39 | 42.02 | 88.92 |
| | 1SFE | 1.00 | 1.00 | 0.99 | 64% | 82% | 40.16 | 93.27 | 45.34 | 91.91 |
| | 1SL8 | 0.41 | 0.75 | 0.31 | 74% | 89% | 40.91 | 77.18 | 61.39 | 95.52 |
| | 1VL1 | 0.99 | 0.96 | 0.98 | 60% | 81% | 43.60 | 101.88 | 40.36 | 92.83 |
| | 1FTG | 0.43 | 0.69 | 0.50 | 61% | 82% | 21.89 | 34.00 | 30.27 | 49.90 |
| | 1XWY | 1.00 | 1.00 | 1.00 | 67% | 87% | 34.03 | 61.99 | 44.25 | 85.29 |
| | 1JKS | 0.00 | 0.51 | 0.20 | 61% | 83% | 24.80 | 43.53 | 34.10 | 53.93 |
| **Bad** | 1SDI | 0.99 | NA | 0.93 | 75% | 87% | 12.69 | 27.59 | 24.80 | 56.45 |
| | 1SQW | 1.00 | 1.00 | 0.99 | 58% | 72% | 23.07 | 39.62 | 30.42 | 44.90 |
| | 1W66 | 1.00 | 1.00 | 0.97 | 65% | 87% | 21.29 | 29.14 | 27.65 | 42.69 |
| | 2HVM | 1.00 | 1.00 | 0.99 | 60% | 86% | 20.55 | 27.99 | 29.21 | 40.80 |
| | 1MSK | 1.00 | 0.88 | 0.96 | 62% | 80% | 35.79 | 70.46 | 47.53 | 91.95 |
| | 1BMG | NA | 1.00 | 1.00 | 71% | 92% | 23.54 | 24.61 | 24.27 | 31.46 |
| | 1NEP | NA | 1.00 | 1.00 | 68% | 95% | 17.28 | 20.38 | 16.81 | 26.67 |

**Table S1:** SAINT2 Diagnostics run. “Good proteins" are ones SAINT2 can produce at least one correct decoy for. “Bad proteins", SAINT2 cannot produce any correct decoys for. Precision @ 50% refers to the precision of predicted contacts available at 50% extrusion; Fragment library precision and coverage are measured overall, as well as specifically by predominant predicted secondary structure. In the case of Precision and Coverage; H includes fragments that are predominantly $\alpha$-helical; B mostly $\beta$-sheet; L predominantly loop; O refers to other (no predominant predicted secondary structure). The precision and 3-state precision of secondary structre predictions are measured, as well as the MAE of torsion angles.

| SCOP Class $\alpha/\beta$ | | | | | |
|---|---|---|---|---|---|
| PDB ID | Name | Pfam Family ID | Length | Resolution | R-value |
| 1AIU | Human thioredoxin | PF00085 | 105 | 2.00 Å | 0.179 |
| 1NAT | Spo0F | PF00072 | 124 | 2.45 Å | 0.181 |
| 2RN2 | Ribonuclease H | PF00075 | 155 | 1.48 Å | 0.196 |
| 1FTG | Apoflavodoxin | PF00258 | 168 | 2.00 Å | 0.182 |
| 1ILW | Pyrazinamidase | PF00857 | 180 | 2.05 Å | 0.181 |
| 1VL1 | 6-phosphogluconolactonase | PF01182 | 232 | 1.55 Å | 0.165 |
| 1XWY | tatD deoxyribonuclease | PF01026 | 264 | 2.00 Å | 0.171 |
| 2HVM | Hevamine | PF00704 | 273 | 1.80 Å | 0.149 |
| 1OBR | Carboxypeptidase t | PF00246 | 326 | 2.3 Å | 0.152 |
| 1VFF | Beta-glycosidase | PF00232 | 423 | 2.50 Å | 0.176 |
| 1SMD | Human salivary amylase | PF00128 | 496 | 1.6 Å | 0.184 |
| SCOP Class $\alpha + \beta$ | | | | | |
| 1WM3 | Human SUMO-2 protein | PF00240 | 72 | 1.2 Å | 0.119 |
| 1CEW | Chicken egg white cystatin | PF00031 | 107 | 2.0 Å | 0.198 |
| 1EKG | Mature human frataxin | PF01491 | 127 | 1.8 Å | 0.197 |
| 1Z2U | Ubiquitin-conjugating enzyme | PF00179 | 150 | 1.1 Å | 0.131 |
| 206L | Phage Lysozyme | P00720 | 164 | 1.75 Å | 0.170 |
| 1SQW | KD93 | PF03657 | 188 | 1.90 Å | 0.225 |
| 1W66 | Lipoate-protein ligase b | PF03099 | 232 | 1.08 Å | 0.116 |
| 1RL0 | Dianthin 30 | PF00161 | 255 | 1.4 Å | 0.184 |
| 2YVT | aq_1956 | PF12850 | 260 | 1.60 Å | 0.170 |
| 1JKS | Protein Kinase | PF00069 | 294 | 1.5 Å | 0.183 |
| 1MSK | Methionine synthase | PF02965 | 331 | 1.8 Å | 0.198 |
| 1AYE | Human procarboxypeptidase A2 | PF02244 | 401 | 1.8 Å | 0.178 |
| 1B4V | Cholesterol oxidase | PF00732 | 504 | 1.5 Å | 0.181 |
| SCOP Class All $\beta$ | | | | | |
| 1CSP | Major cold-shock protein | PF00313 | 67 | 2.45 Å | 0.195 |
| 1BMG | Beta 2-microglobulin | PF13895 | 98 | 2.5 Å | 0.191 |
| 1XD6 | Gastrodianin | PF01453 | 112 | 2.0 Å | 0.199 |
| 1NEP | Niemann-Pick C2 protein | PF02221 | 130 | 1.7 Å | 0.194 |
| 1CZT | Human coagulation factor V | PF00754 | 160 | 1.87 Å | 0.201 |
| 1T9F | NA | PF02815 | 187 | 2.00 Å | 0.217 |
| 2AYH | H(A16-M) | PF00722 | 214 | 1.6 Å | 0.143 |
| 1P6F | NKp46 | PF13895 | 241 | 2.2 Å | 0.207 |
| 1SEF | EF2996 | PF07883 | 274 | 2.05 Å | 0.208 |
| 1WL7 | ABN-TS | PF04616 | 312 | 1.9 Å | 0.183 |
| 1OKQ | Laminin alpha 2 chain LG4-5 | PF00054 | 394 | 2.80 Å | 0.220 |

| PDB ID | Name | Pfam Family ID | Length | Resolution | R-value |
|---|---|---|---|---|---|
| SCOP Class All $\alpha$ | | | | | |
| 1ENH | Engrailed homodomain | PF00046 | 54 | 2.1 Å | 0.197 |
| 2J9V | ESCRT-I | PF03997 | 99 | 2.0 Å | 0.228 |
| 2MHR | Myohemerythrin | 118 | PF01814 | 1.3 Å | 0.158 |
| 1JWF | GGA-1 | PF00790 | 147 | 2.10 Å | 0.222 |
| 1SFE | ADA DNA methyltransferase | PF01035 | 180 | 2.10 Å | 0.219 |
| 1SL8 | Calcium-regulated photoprotein | PF07164 | 191 | 1.7 Å | 0.216 |
| 1SDI | YcfC | PF04356 | 213 | 1.65 Å | 0.150 |
| 1VIN | Cyclin A3 | PF00134 | 268 | 2.0 Å | 0.216 |
| 1V5C | Chitosanase | PF01270 | 386 | 2.0 Å | 0.184 |
| Transmembrane $\alpha$-helical bundles | | | | | |
| 1ORS_C | KvAP potassium channel | No annotation | 132 | 1.9 Å | 0.231 |
| 3CZ5_C | LuxR | PF00072 | 153 | 2.7 Å | 0.196 |
| 4HUQ_S | Folate energy-coupling factor | PF12822 | 174 | 3.0 Å | 0.215 |
| 2XOW_A | Rhomboid protease | PF01694 | 179 | 2.09 Å | 0.198 |
| 3RLB_A | Thiamin S-component | PF09515 | 192 | 2.0 Å | 0.206 |
| 4A2N_B | Methyltransferase-ICMT | PF04140 | 194 | 3.4 Å | 0.238 |
| 4O6Y_A | Cytochrome b561 | PF03188 | 230 | 1.7 Å | 0.201 |
| 1KQF_C | Formate dehydrogenase N | PF00033 | 217 | 1.6 Å | 0.177 |
| 3B4R_B | Site-2 protease | PF02163 | 224 | 3.3 Å | 0.251 |
| 4B4A_A | Tatc core | PF00902 | 249 | 3.5 Å | 0.252 |
| 1E12_A | Halorhodopsin | PF01036 | 253 | 1.8 Å | 0.237 |
| 2VPZ_C | Polysulfide reductase | PF14589 | 253 | 2.4 Å | 0.249 |
| 2DYR_C | Cytochrome C oxidase | PF00510 | 261 | 1.8 Å | 0.202 |
| 2W2E_A | Aquaporin | PF00230 | 279 | 1.15 Å | 0.140 |
| 3KLY_A | Pentameric formate channel | PF01226 | 280 | 2.1 Å | 0.179 |
| 1OKC_A | Mitochondrial ADP/ATP carrier | PF00153 | 297 | 2.2 Å | 0.220 |
| 4OD5_A | Ubiquinone homologue | PF01040 | 303 | 3.56 Å | 0.279 |
| 4N7W_A | Bile acid transporter protein | PF01758 | 307 | 1.95 Å | 0.186 |
| 3M73_A | SLAC1 homologue | PF03595 | 314 | 1.15 Å | 0.141 |
| 2QI9_A | ABC-transporter BtuCD | PF06609 | 326 | 2.6 Å | 0.262 |
| 4EZC_A | UT-B Urea transporter | PF03253 | 384 | 2.36 Å | 0.196 |
| 1U7G_A | Ammonia channel AmtB | PF00909 | 385 | 1.4 Å | 0.133 |
| 1ZCD_A | Na+/H+ antiporter NhaA | PF06965 | 388 | 3.45 Å | 0.316 |
| 4BWZ_A | Sodium antiporter NapA | PF00999 | 394 | 2.98 Å | 0.222 |

**Table S2:** The 44 proteins comprising our PDB-Representative soluble validation data set separated by SCOP classes and the 24 proteins comprising our transmembrane $\alpha$-helical bundle validation data set. Soluble proteins are single-domain, single chain, and belong to distinct PFam families.

| PDB ID | Length | Pfam Family ID | Resolution | SCOP |
|---|---|---|---|---|
| 3RT3 | 93 | PF02942 | 2.01 | $\alpha/\beta$ |
| 2EFV | 73 | PF10802 | 1.9 | $\alpha/\beta$ |
| 1N81 | 173 | PF09216 | 2.1 | $\alpha/\beta$ |
| 2OX6 | 161 | PF14590 | 1.7 | $\alpha/\beta$ |
| 1ZKE | 78 | PF10398 | 1.6 | $\alpha/\beta$ |
| 4JJF | 96 | PF03201 | 2.2 | $\alpha/\beta$ |
| 1ZS4 | 78 | PF05269 | 1.7 | $\alpha/\beta$ |
| 1D2Z | 145 | PF14786 | 2.0 | $\alpha/\beta$ |
| 1V74 | 84 | PF09204 | 2.0 | $\alpha/\beta$ |
| 1XG7 | 231 | PF09171 | 1.88 | $\alpha/\beta$ |
| 1SED | 106 | PF08963 | 2.1 | $\alpha/\beta$ |
| 2X9C | 60 | PF09392 | 2.45 | $\alpha/\beta$ |
| 1MN8 | 94 | PF01140 | 1.0 | $\alpha/\beta$ |
| 3WU2 | 92 | PF06514 | 1.9 | $\alpha/\beta$ |
| 1EEX | 132 | PF02287 | 1.7 | $\alpha/\beta$ |
| 4GKF | 139 | PF09701 | 2.1 | $\alpha/\beta$ |
| 2J6Y | 74 | PF08673 | 1.85 | $\alpha/\beta$ |
| 4FZO | 67 | PF04010 | 1.3 | $\alpha/\beta$ |
| 1BOU | 86 | PF07746 | 2.2 | $\alpha/\beta$ |
| 1UZ3 | 69 | PF03735 | 1.1 | $\alpha/\beta$ |
| 1LJP | 93 | PF00964 | 1.8 | $\alpha/\beta$ |
| 2NN4 | 59 | PF06014 | 2.1 | $\alpha/\beta$ |
| 2IB0 | 123 | PF14530 | 2.0 | $\alpha/\beta$ |
| 2J9U | 93 | PF03997 | 2.0 | $\alpha/\beta$ |
| 4DHX | 82 | PF10163 | 2.1 | $\alpha/\beta$ |
| 2IV1 | 69 | PF02560 | 1.88 | $\alpha/\beta$ |
| 1RR7 | 93 | PF08765 | 2.2 | $\alpha/\beta$ |
| 1K78 | 121 | PF00292 | 2.25 | $\alpha/\beta$ |
| 2HUO | 248 | PF05153 | 2.0 | $\alpha/\beta$ |
| 4W4K | 76 | PF00934 | 1.95 | $\alpha/\beta$ |
| 1TJF | 185 | PF01213 | 2.21 | $\alpha/\beta$ |
| 1H0A | 123 | PF01417 | 1.7 | $\alpha/\beta$ |
| 4W4L | 156 | PF00823 | 2.45 | $\alpha/\beta$ |
| 1UB9 | 78 | PF13601 | 2.05 | $\alpha/\beta$ |
| 2FI0 | 58 | PF08984 | 2.1 | $\alpha/\beta$ |
| 2BV7 | 149 | PF08718 | 1.79 | $\alpha/\beta$ |
| 1FPO | 75 | PF07743 | 1.8 | $\alpha/\beta$ |
| 1HFE | 55 | PF02256 | 1.6 | $\alpha/\beta$ |
| 1RP3 | 50 | PF04316 | 2.3 | $\alpha/\beta$ |

| PDB ID | Length | Pfam Family ID | Resolution | SCOP |
|---|---|---|---|---|
| 2GA1 | 55 | PF04255 | 2.0 | $\alpha/\beta$ |
| 2O70 | 156 | PF09349 | 1.8 | $\alpha/\beta$ |
| 1FXK | 116 | PF02996 | 2.3 | $\alpha/\beta$ |
| 2ION | 113 | PF02847 | 1.57 | $\alpha/\beta$ |
| 4J41 | 80 | PF06013 | 1.52 | $\alpha/\beta$ |
| 3LTL | 182 | PF01369 | 2.2 | $\alpha/\beta$ |
| 3H8M | 65 | PF07647 | 2.1 | $\alpha/\beta$ |
| 2H5N | 121 | PF05099 | 2.01 | $\alpha/\beta$ |
| 2GSC | 105 | PF05635 | 2.45 | $\alpha/\beta$ |
| 1ZK8 | 75 | PF13305 | 2.15 | $\alpha/\beta$ |
| 1NG6 | 143 | PF09424 | 1.4 | $\alpha/\beta$ |
| 1B79 | 101 | PF00772 | 2.3 | $\alpha/\beta$ |
| 4PC6 | 192 | PF00889 | 2.2 | $\alpha/\beta$ |
| 3X0F | 87 | PF00335 | 1.47 | $\alpha/\beta$ |
| 1TZT | 126 | PF01029 | 1.55 | $\alpha/\beta$ |
| 1Z4A | 137 | PF00210 | 2.3 | $\alpha/\beta$ |
| 1VZ0 | 89 | PF02195 | 2.3 | $\alpha/\beta$ |
| 3IU5 | 85 | PF00439 | 1.63 | $\alpha/\beta$ |
| 3F31 | 103 | PF00435 | 2.3 | $\alpha/\beta$ |
| 1FSE | 56 | PF00196 | 2.05 | $\alpha/\beta$ |
| 1S7M | 52 | PF15403 | 2.1 | $\alpha+\beta$ |
| 2OJ6 | 160 | PF01664 | 1.85 | $\alpha+\beta$ |
| 4YG3 | 159 | PF00426 | 2.29 | $\alpha+\beta$ |
| 1VIF | 59 | PF06442 | 1.8 | $\alpha+\beta$ |
| 4L5F | 95 | PF02832 | 2.45 | $\alpha+\beta$ |
| 1GPC | 88 | PF08804 | 2.2 | $\alpha+\beta$ |
| 1OSY | 110 | PF09259 | 1.7 | $\alpha+\beta$ |
| 2GN5 | 81 | PF02303 | 2.3 | $\alpha+\beta$ |
| 1SR4 | 144 | PF03498 | 2.0 | $\alpha+\beta$ |
| 1Y2X | 138 | PF07367 | 2.36 | $\alpha+\beta$ |
| 1OA8 | 113 | PF08517 | 1.7 | $\alpha+\beta$ |
| 1EAO | 122 | PF00853 | 1.4 | $\alpha+\beta$ |
| 3MH8 | 190 | PF07161 | 2.0 | $\alpha+\beta$ |
| 4UU3 | 157 | PF05870 | 1.15 | $\alpha+\beta$ |
| 1TD0 | 100 | PF02924 | 1.95 | $\alpha+\beta$ |
| 3T5I | 140 | PF05351 | 2.1 | $\alpha+\beta$ |
| 1J3Q | 160 | PF06560 | 1.85 | $\alpha+\beta$ |
| 2CAY | 88 | PF11605 | 1.9 | $\alpha+\beta$ |
| 2F5T | 228 | PF11495 | 1.45 | $\alpha+\beta$ |
| 2NQD | 92 | PF09394 | 1.75 | $\alpha+\beta$ |

| PDB ID | Length | Pfam Family ID | Resolution | SCOP |
|---|---|---|---|---|
| 2FMM | 52 | PF01393 | 1.8 | $\alpha + \beta$ |
| 3HSB | 63 | PF17209 | 2.2 | $\alpha + \beta$ |
| 1DU3 | 128 | PF00229 | 2.2 | $\alpha + \beta$ |
| 2QW7 | 82 | PF03319 | 2.4 | $\alpha + \beta$ |
| 4IJH | 99 | PF04057 | 1.5 | $\alpha + \beta$ |
| 3HKS | 69 | PF01287 | 2.3 | $\alpha + \beta$ |
| 4C9Z | 183 | PF03145 | 1.95 | $\alpha + \beta$ |
| 3QZB | 115 | PF01880 | 1.1 | $\alpha + \beta$ |
| 2P12 | 67 | PF04167 | 1.63 | $\alpha + \beta$ |
| 1HF2 | 96 | PF03775 | 2.2 | $\alpha + \beta$ |
| 3I4O | 63 | PF01176 | 1.47 | $\alpha + \beta$ |
| 2F15 | 84 | PF16561 | 2.0 | $\alpha + \beta$ |
| 2AR1 | 145 | PF01878 | 1.6 | $\alpha + \beta$ |
| 1PZ2 | 200 | PF06964 | 2.0 | $\alpha + \beta$ |
| 3I98 | 152 | PF00719 | 1.85 | $\alpha + \beta$ |
| 2PHC | 198 | PF02682 | 2.29 | $\alpha + \beta$ |
| 4Q97 | 86 | PF07654 | 2.4 | $\alpha + \beta$ |
| 3L9E | 141 | PF00080 | 2.05 | $\alpha + \beta$ |
| 4LHN | 158 | PF07691 | 2.12 | $\alpha + \beta$ |
| 1Y0G | 162 | PF04264 | 2.2 | $\alpha + \beta$ |
| 2BM6 | 75 | PF13599 | 2.2 | $\alpha + \beta$ |
| 2ANE | 107 | PF02190 | 2.03 | $\alpha + \beta$ |
| 1RU4 | 181 | PF13229 | 1.6 | $\alpha + \beta$ |
| 2QDL | 135 | PF01584 | 2.2 | $\alpha + \beta$ |
| 1G6G | 67 | PF00498 | 1.6 | $\alpha + \beta$ |
| 1QWY | 101 | PF01551 | 1.3 | $\alpha + \beta$ |
| 4OFF | 83 | PF00027 | 1.89 | $\alpha + \beta$ |
| 2CZR | 98 | PF15517 | 2.3 | All $\beta$ |
| 1EYU | 153 | PF09225 | 1.78 | All $\beta$ |
| 1ESG | 159 | PF02923 | 1.9 | All $\beta$ |
| 2I9I | 220 | PF05211 | 1.8 | All $\beta$ |
| 1YTL | 154 | PF02552 | 1.8 | All $\beta$ |
| 1LU9 | 80 | PF09176 | 1.9 | All $\beta$ |
| 2YY8 | 118 | PF01994 | 2.48 | All $\beta$ |
| 2WCW | 92 | PF01870 | 1.58 | All $\beta$ |
| 1IWP | 114 | PF02288 | 2.1 | All $\beta$ |
| 1YQE | 203 | PF04414 | 1.83 | All $\beta$ |
| 1SYX | 132 | PF02966 | 2.35 | All $\beta$ |
| 2D1P | 85 | PF04077 | 2.15 | All $\beta$ |
| 1Z6N | 130 | PF14595 | 1.5 | All $\beta$ |

| PDB ID | Length | Pfam Family ID | Resolution | SCOP |
|---|---|---|---|---|
| 2PJU | 163 | PF06506 | 2.1 | All $\beta$ |
| 1U7P | 157 | PF12689 | 1.9 | All $\beta$ |
| 3WYZ | 204 | PF01876 | 2.21 | All $\beta$ |
| 1U2Z | 205 | PF08123 | 2.2 | All $\beta$ |
| 2ZBT | 206 | PF01680 | 1.65 | All $\beta$ |
| 4BLV | 245 | PF04378 | 2.0 | All $\beta$ |
| 3SLP | 145 | PF09588 | 2.3 | All $\beta$ |
| 1CFZ | 128 | PF01750 | 2.2 | All $\beta$ |
| 3E7D | 196 | PF02570 | 1.8 | All $\beta$ |
| 1IXN | 233 | PF03740 | 2.3 | All $\beta$ |
| 3JS3 | 227 | PF01487 | 2.2 | All $\beta$ |
| 2NZX | 131 | PF00852 | 1.9 | All $\beta$ |
| 1G64 | 168 | PF02572 | 2.1 | All $\beta$ |
| 3IIK | 120 | PF13238 | 1.95 | All $\beta$ |
| 1O73 | 94 | PF13905 | 2.28 | All $\beta$ |
| 1LK5 | 171 | PF06026 | 1.75 | All $\beta$ |
| 1YR7 | 237 | PF03029 | 2.08 | All $\beta$ |
| 1UQR | 137 | PF01220 | 1.7 | All $\beta$ |
| 4W8G | 101 | PF13676 | 1.95 | All $\beta$ |
| 2XWQ | 109 | PF01903 | 2.01 | All $\beta$ |
| 1RVV | 139 | PF00885 | 2.4 | All $\beta$ |
| 2WQK | 183 | PF01975 | 1.5 | All $\beta$ |
| 1UPF | 205 | PF14681 | 2.3 | All $\beta$ |
| 3HEE | 137 | PF02502 | 2.0 | All $\beta$ |
| 2B3J | 148 | PF14437 | 2.0 | All $\beta$ |
| 1KDP | 214 | PF02224 | 2.3 | All $\beta$ |
| 1HTW | 132 | PF02367 | 1.7 | All $\beta$ |
| 1R5Y | 238 | PF01702 | 1.2 | All $\beta$ |
| 2FCA | 169 | PF02390 | 2.1 | All $\beta$ |
| 1JR2 | 236 | PF02602 | 1.84 | All $\beta$ |
| 3RQ5 | 240 | PF01256 | 1.7 | All $\beta$ |
| 4QT4 | 184 | PF01195 | 2.19 | All $\beta$ |
| 2PBR | 180 | PF02223 | 1.96 | All $\beta$ |
| 3U3G | 127 | PF13456 | 1.4 | All $\beta$ |
| 5DLE | 91 | PF02302 | 1.55 | All $\beta$ |
| 1X07 | 216 | PF01255 | 2.2 | All $\beta$ |
| 1QYR | 249 | PF00398 | 2.1 | All $\beta$ |
| 4F4J | 184 | PF00625 | 2.45 | All $\beta$ |
| 1KA9 | 228 | PF00977 | 2.3 | All $\beta$ |
| 1JFL | 215 | PF01177 | 1.9 | All $\beta$ |

| PDB ID | Length | Pfam Family ID | Resolution | SCOP |
|---|---|---|---|---|
| 4F82 | 160 | PF08534 | 1.85 | All $\beta$ |
| 1WDJ | 148 | PF05685 | 2.0 | All $\beta$ |
| 4P92 | 216 | PF01738 | 1.65 | All $\beta$ |
| 3ZYW | 64 | PF00462 | 1.84 | All $\beta$ |
| 1JLJ | 146 | PF00994 | 1.6 | All $\beta$ |
| 5A1J | 233 | PF01497 | 1.6 | All $\beta$ |
| 3KL2 | 187 | PF00857 | 2.3 | All $\beta$ |
| 4PK9 | 196 | PF01734 | 1.96 | All $\beta$ |
| 1J54 | 166 | PF00929 | 1.7 | All $\beta$ |
| 2C1G | 118 | PF01522 | 1.75 | All $\beta$ |
| 2NQH | 220 | PF01261 | 1.1 | All $\beta$ |
| 1H5R | 238 | PF00483 | 1.9 | All $\beta$ |
| 1KAF | 102 | PF09158 | 1.6 | All $\alpha$ |
| 4NBP | 93 | PF02217 | 1.31 | All $\alpha$ |
| 1EL6 | 207 | PF08677 | 2.0 | All $\alpha$ |
| 3O9S | 120 | PF00600 | 2.48 | All $\alpha$ |
| 1HUF | 120 | PF09013 | 2.0 | All $\alpha$ |
| 1RKI | 88 | PF11424 | 1.6 | All $\alpha$ |
| 4GH9 | 121 | PF02097 | 1.65 | All $\alpha$ |
| 1NLN | 179 | PF00770 | 1.6 | All $\alpha$ |
| 2D4P | 114 | PF09390 | 1.7 | All $\alpha$ |
| 1TFO | 87 | PF11429 | 2.3 | All $\alpha$ |
| 4JET | 178 | PF06438 | 2.2 | All $\alpha$ |
| 1F3V | 110 | PF09034 | 2.0 | All $\alpha$ |
| 2FZP | 122 | PF08941 | 1.87 | All $\alpha$ |
| 1UKF | 183 | PF03543 | 1.35 | All $\alpha$ |
| 2FK3 | 55 | PF12924 | 2.4 | All $\alpha$ |
| 1WNH | 214 | PF06907 | 1.83 | All $\alpha$ |
| 1MWP | 92 | PF02177 | 1.8 | All $\alpha$ |
| 1SKO | 115 | PF08923 | 2.0 | All $\alpha$ |
| 1K8K | 165 | PF05856 | 2.0 | All $\alpha$ |
| 4OK9 | 138 | PF09021 | 1.91 | All $\alpha$ |
| 3EX7 | 141 | PF02792 | 2.3 | All $\alpha$ |
| 4ILY | 209 | PF01374 | 1.83 | All $\alpha$ |
| 2EKM | 153 | PF04008 | 2.06 | All $\alpha$ |
| 2OHW | 116 | PF07997 | 1.4 | All $\alpha$ |
| 1B33 | 54 | PF01383 | 2.3 | All $\alpha$ |
| 1VQN | 187 | PF00827 | 2.4 | All $\alpha$ |
| 2P0W | 162 | PF10394 | 1.9 | All $\alpha$ |
| 1UW4 | 90 | PF03467 | 1.95 | All $\alpha$ |

| PDB ID | Length | Pfam Family ID | Resolution | SCOP |
|---|---|---|---|---|
| 3MJZ | 80 | PF14552 | 2.4 | All $\alpha$ |
| 1ZD0 | 129 | PF08617 | 1.7 | All $\alpha$ |
| 2OSD | 61 | PF02830 | 2.3 | All $\alpha$ |
| 1J27 | 85 | PF04456 | 1.7 | All $\alpha$ |
| 2FV5 | 230 | PF13688 | 2.1 | All $\alpha$ |
| 3ZXW | 92 | PF00101 | 2.1 | All $\alpha$ |
| 1NFJ | 63 | PF01918 | 2.0 | All $\alpha$ |
| 2J3W | 127 | PF04628 | 2.1 | All $\alpha$ |
| 1U55 | 158 | PF07700 | 1.77 | All $\alpha$ |
| 3T3T | 108 | PF01491 | 1.38 | All $\alpha$ |
| 1MB1 | 80 | PF04383 | 2.1 | All $\alpha$ |
| 2CM0 | 78 | PF09409 | 1.9 | All $\alpha$ |
| 2O0Q | 89 | PF06108 | 1.8 | All $\alpha$ |
| 1IAH | 196 | PF02816 | 2.4 | All $\alpha$ |
| 1YOC | 130 | PF14539 | 1.7 | All $\alpha$ |
| 1JYH | 148 | PF14526 | 1.8 | All $\alpha$ |
| 2F9Z | 99 | PF03975 | 2.4 | All $\alpha$ |
| 1R44 | 201 | PF01427 | 2.25 | All $\alpha$ |
| 2JGB | 159 | PF01652 | 1.7 | All $\alpha$ |
| 3DVH | 83 | PF01221 | 2.0 | All $\alpha$ |
| 1GYB | 113 | PF02136 | 1.9 | All $\alpha$ |
| 3L7H | 90 | PF03259 | 1.95 | All $\alpha$ |
| 1WTF | 62 | PF13370 | 1.6 | All $\alpha$ |
| 4QR1 | 78 | PF09827 | 2.19 | All $\alpha$ |
| 2QSW | 72 | PF09383 | 1.5 | All $\alpha$ |
| 2A1V | 96 | PF04237 | 2.15 | All $\alpha$ |
| 3KVT | 90 | PF02214 | 2.0 | All $\alpha$ |
| 1VMH | 112 | PF01894 | 1.31 | All $\alpha$ |
| 1TIF | 69 | PF05198 | 1.8 | All $\alpha$ |
| 2JLJ | 111 | PF01312 | 1.3 | All $\alpha$ |
| 2J5A | 90 | PF01250 | 2.3 | All $\alpha$ |
| 1ZHV | 63 | PF13840 | 1.5 | All $\alpha$ |
| 3RMU | 111 | PF13669 | 1.8 | All $\alpha$ |
| 4HEI | 88 | PF02482 | 1.6 | All $\alpha$ |
| 1MKB | 130 | PF07977 | 2.0 | All $\alpha$ |
| 1OZ9 | 130 | PF02130 | 1.89 | All $\alpha$ |
| 1ZAX | 100 | PF00466 | 2.1 | All $\alpha$ |
| 3LNW | 81 | PF00381 | 2.3 | All $\alpha$ |
| 1AYA | 75 | PF00017 | 2.05 | All $\alpha$ |
| 4NRN | 86 | PF05016 | 1.8 | All $\alpha$ |

| PDB ID | Length | Pfam Family ID | Resolution | SCOP |
|---|---|---|---|---|
| 1S7I | 115 | PF03795 | 1.8 | All $\alpha$ |
| 1TUV | 86 | PF03992 | 1.7 | All $\alpha$ |
| 1LN0 | 88 | PF01541 | 2.0 | All $\alpha$ |
| 3B2G | 75 | PF00111 | 1.76 | All $\alpha$ |
| 2GJ3 | 111 | PF00989 | 1.04 | All $\alpha$ |
| 1LKD | 118 | PF00903 | 1.7 | All $\alpha$ |

**Table S3:** The 245 proteins comprising our PDB-Representative soluble validation data set. Soluble proteins are single-domain, single chain, and belong to distinct PFam families.