EFFICIENT ALGORITHMS IN PROTEIN MODELLING

Inauguraldissertation
zur
Erlangung der Würde eines Doktors der Philosophie
vorgelegt der
Philosophisch-Naturwissenschaftlichen Fakultät
der Universität Basel

von

Gabriel Studer

aus Schüpfheim (LU)

Basel, 2017

Genehmigt von der Philosophisch-Naturwissenschaftlichen Fakultät auf Antrag von

Fakultätsverantwortlicher:
Prof. Dr. Torsten Schwede

Korreferent:
Prof. Dr. Timm Maier

Basel, 20.06.2017

Prof. Dr. Martin Spiess
Dekan

Gabriel Studer: *Efficient Algorithms in Protein Modelling*

Because it's there

— George Mallory

ABSTRACT

Proteins are key players in the complex world of living cells. No matter whether they are involved in enzymatic reactions, intercell communication or numerous other processes, knowledge of their structure is vital for a detailed understanding of their function. However, structure determination by experiment is often a laborious process that cannot keep up with the ever increasing pace of sequencing methodologies. As a consequence, the gap between proteins where we only know the sequence and the proteins where we additionally have detailed structural information is growing rapidly. Computational modelling methods that extrapolate structural information from homologous structures have established themselves as a valuable complement to experiment and help bridging this gap. This thesis addresses two key aspects in protein modelling.

(1) It investigates and improves methodologies that assign reliability estimates to protein models, so called quality estimation (QE) methods. Even a human expert cannot immediately detect errors introduced in the modelling process, thus the importance of automated methods performing this task.

(2) It assesses the available methods that perform the modelling itself, discusses solutions for current shortcomings and provides efficient implementations thereof.

When detecting errors in protein models, many knowledge based methods are biased towards the physio-chemical properties observed in soluble protein structures. This limits their applicability for the important class of membrane protein models. In an effort to improve the situation, QMEANBrane has been developed. QMEANBrane is specifically designed to detect local errors in membrane protein models by membrane specific statistical potentials of mean force that nowadays approach statistical saturation given the increase of available experimental data.

Considering the improvement of quality estimation for soluble proteins, instead of solely applying the widely used statistical potentials of mean force, QMEANDisCo incorporates the observed structural variety of experimentally determined protein structures homologous to the model being assessed. Valu-

able ensemble information can be gathered without the need of actually depending on a large ensemble of protein models, thus circumventing a main limitation of consensus QE methods.

Apart from improving QE methods, in an effort of implementing and extending state-of-the-art modelling algorithms, the lack of a free and efficient modelling engine became obvious. No available modelling engine provided an open-source codebase as a basis for novel, innovative algorithms and, at the same time, had no restrictions for usage. This contradicts our efforts to make protein modelling available to all biochemists and molecular biologists worldwide. As a consequence we implemented a new free and open modelling engine from scratch - ProMod3. ProMod3 allows to combine extremely efficient, state-of-the-art modelling algorithms in a flexible manner to solve various modelling problems.

To weaken the dogma of one template one model, basic algorithms have been explored to incorporate structural information from multiple templates into one protein model. The algorithms are built using ProMod3 and have extensively been tested in the context of the CAMEO continuous evaluation platform. The result is a highly competitive modelling pipeline that excels with extremely low runtimes and excellent performance.

CONTENTS

1 INTRODUCTION 1
- 1.1 Protein Structure 1
- 1.2 From Primary to Quaternary Structure 2
- 1.3 Experimental Methods 4
  - 1.3.1 X-ray Crystallography 5
  - 1.3.2 NMR 5
  - 1.3.3 Electron Microscopy 6
- 1.4 Sequence, Structure and Function 7
  - 1.4.1 Pairwise Sequence Alignments 7
  - 1.4.2 From Substitution Matrices to Statistical Models 9
- 1.5 Protein Modelling 10
  - 1.5.1 Homology Detection and Homology Transfer 10
  - 1.5.2 Loop Modelling 11
  - 1.5.3 Sidechain Modelling 11
  - 1.5.4 Energy Minimization 13
  - 1.5.5 Quality Estimation 14
- 1.6 Objectives 18

2 QMEANBRANE 19
- 2.1 Introduction 20
- 2.2 Materials & Methods 21
  - 2.2.1 Target Function 21
  - 2.2.2 Membrane segment definition 22
  - 2.2.3 Model quality predictors 24
  - 2.2.4 Training datasets for statitical potentials 26
  - 2.2.5 Datasets for training linear combinations 26
  - 2.2.6 Spherical smoothing for noise reduction 27
  - 2.2.7 Per amino acid weighting scheme 28
  - 2.2.8 Implementation 28
- 2.3 Results and Discussion 28
  - 2.3.1 Membrane prediction accuracy 28
  - 2.3.2 Performance on the test dataset 29
  - 2.3.3 Independent performance evaluation on models of the GPCR Dock experiments 29
  - 2.3.4 Retrospective analysis of modelling examples 30
- 2.4 Conclusion 33
- 2.5 Acknowledgments 34

2.6 Supplemental Materials 35
2.6.1 Saturation of Membrane Specific Statistical Potentials 35
2.6.2 dRMSD Definition Used in This Work 35
2.6.3 Parametrization of Statistical Potential Terms 37
2.6.4 Training data 38
2.6.5 Comparison of Different Measures of Similarity 39
2.6.6 Discrimination of Membrane and Soluble Structures 40
2.6.7 ROC Analysis on GPCRDock Test Set Using Different Measures of Similarity 40
2.6.8 Head to Head comparison with ProQM on the testset used to test/train ProQM 42
2.6.9 Performance on Our Test Set 43
2.6.10 Retrospective Modelling Analysis 44
3 QMEANDISCO 49
3.1 Introduction 50
3.2 Materials & Methods 52
3.2.1 Target Value 52
3.2.2 Training and Test set 52
3.2.3 QMEAN 53
3.2.4 DisCo 55
3.2.5 Score Combination - QMEANDisCo 57
3.2.6 Evaluation Methods 58
3.3 Results and Discussion 58
3.3.1 Overall AUC Analysis 59
3.3.2 Global Effect on Overall AUC 60
3.3.3 Per Model AUC 60
3.3.4 Analysis on High Quality Models for the CASP XI Test Set 62
3.4 Conclusions 62
3.5 Acknowledgments 63
3.6 Supplemental Materials 64
3.6.1 dRMSD as a target value for local accuracy 64
3.6.2 Parameterization of statistical potentials 65
3.6.3 Target value distribution of different test sets 65
3.6.4 Why you should not use Cα distances as a target value for local quality estimates anymore 67
3.7 Evaluation of local QE performance 73

4 PROMOD3 - A VERSATILE HOMOLOGY MODELLING TOOLBOX 87
4.1 Introduction 87
4.2 Materials & Methods 89
4.2.1 Architecture 89
4.2.2 The *Loop* Module 89
4.2.3 The *Scoring* Module 92
4.2.4 The *Sidechain* Module 93
4.2.5 The *Modelling* Module 94
4.3 Results 102
4.3.1 Fragment Detection Performance 102
4.3.2 Loop Modelling Performance 103
4.3.3 Sidechain Modelling Performance 104
4.3.4 Homology Modelling Performance 105
4.4 Discussion 106
4.5 Acknowledgments 106
4.6 Supplemental Materials 107
4.6.1 Structural Coverage in Structural Database 107
4.6.2 Relevant Loop Lengths 108
4.6.3 Fragger Scores 108
4.6.4 Scorers of the *Scoring* Module 110
4.6.5 Fragment Detection Performance 113
4.6.6 Sidechain Performance Analysis 119
4.6.7 Homology Modelling Performance - More Details 122
4.6.8 Speed Benchmarks 124
5 MULTITEMPLATE MODELLING 127
5.1 Introduction 127
5.2 Materials & Methods 128
5.2.1 The SWISS-MODEL pipeline 128
5.2.2 Rigid Blocks Algorithm 129
5.2.3 Coverage Extension Algorithm 131
5.2.4 Testing Strategies 134
5.3 Results 135
5.3.1 SWISS-MODEL Server Evaluation 135
5.3.2 Comparison to Other Multitemplate Methods 136
5.4 Discussion 138
6 ACKNOWLEDGMENTS 141

BIBLIOGRAPHY 143

1 INTRODUCTION

1.1 PROTEIN STRUCTURE

An average human cell contains several billion proteins which is in the same order of magnitude than the total number of nucleotides encoding the full human genome [118]. Proteins do not just shine with their tremendous abundance but also with their enormous functional complexity ranging from structural support and movement towards enzymatic activity as well as interaction with the outside world [4]. All this are achieved by polymers emerging from a limited alphabet of building blocks with varying properties - the amino acids. 20 amino acids are directly encoded in the genome and follow a common chemical scheme, where a $C\alpha$ carbon is linked to an amino acid specific sidechain and flanked by an amine ($-H_2N$) and carboxyl ($-COOH$) functional group. Nucleotides in RNA are also known to exhibit catalytic activity [44] but their wide range of chemical characteristics make amino acids more versatile. The desired polymer is formed by connecting amino acids through condensation reactions resulting in a peptide bond between the carboxyl carbon of amino acid at position i with the amine nitrogen at position $i+1$. The result is a continuous backbone with a repetitive pattern of the heavy atoms N, $C\alpha$ and C for each amino acid, with a carbonyl oxygen bound to C and the amino acid specific sidechain to $C\alpha$. Assuming constant bond lengths and bond angles, the overall fold of such a polymer can be characterized by a triplet of dihedral angles for every amino acid AA_i:

- **ω:** $C\alpha_{i-1}$, C_{i-1}, N_i, $C\alpha_i$
- **ϕ:** C_{i-1}, N_i, $C\alpha_i$, C_i
- **ψ:** N_i, $C\alpha_i$, C_i, N_{i+1}

The degrees of freedom are largely reduced by resonance effects that give the peptide bond a partial double bond character. As a consequence, the plane defined by ω is mostly planar and the peptide bond can either adopt *cis* or *trans* conformation with a large preference for *trans* [72]. In the case of proline, *cis* conformations are not very uncommon and occur in a noteworthy fraction [125]. Not only the ω angle is restricted, but also ϕ and ψ show clear preferences that have first been analysed statistically by Ramachandran and co-workers [130]. The visualisation of the ϕ / ψ backbone dihedral angles, the so called Ramachandran plot, looks similar for most amino acids but can in some cases have very characteristic properties (Figure 1).

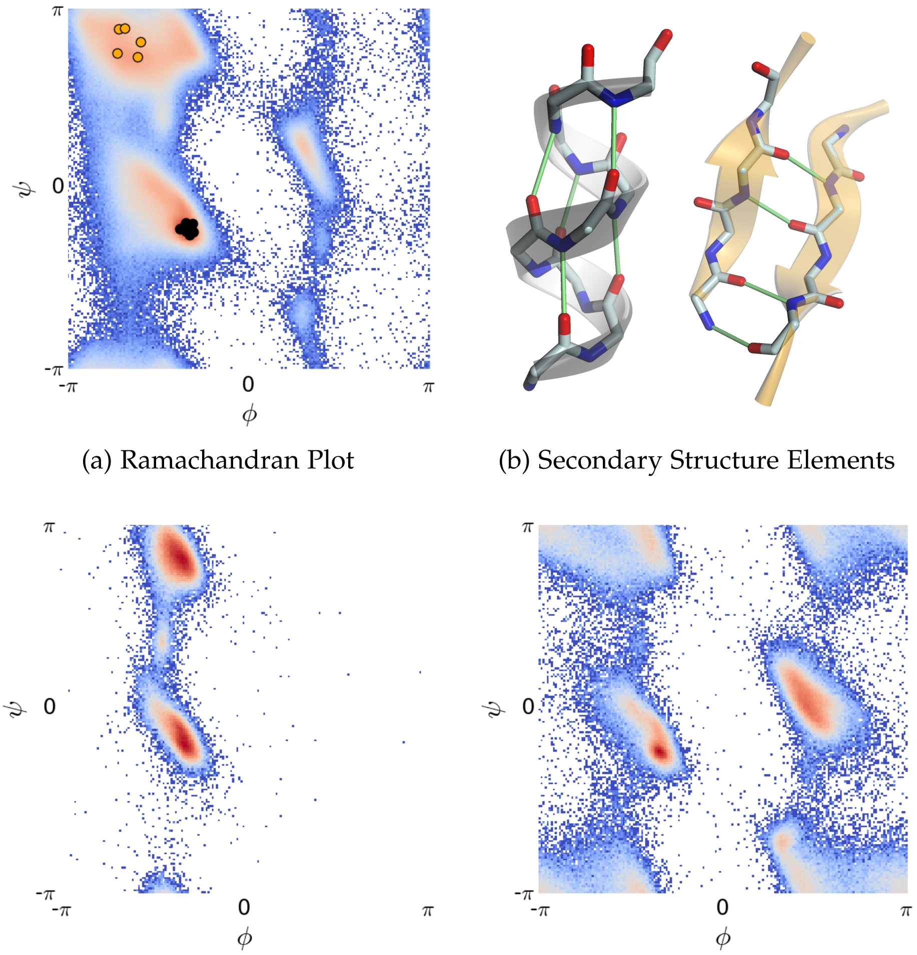
1.2 FROM PRIMARY TO QUATERNARY STRUCTURE

In Section 1.1 we discussed that a protein is a polymer of amino acids. The exact sequence of amino acids is often referred to as the protein's primary structure.

A protein's secondary structure already jumps into 3D space and describes reoccurring local arrangements of amino acids that are energetically favourable. Without having structural data for a full protein, Pauling and Corey postulated the two most common secondary structure elements, the α-helix and the β-sheet, already in 1951 [123, 124]. The key to success was the assumption of constant bond lengths / bond angles as well as a planar peptide bond. The aforementioned secondary structure elements were then a result of finding conformations given constraints that were stereochemically feasible and had a favourable hydrogen bond pattern (illustrated in Figure 1b). In detail:

- **α-helix:** CO of amino acid at position i forms a hydrogen bond with NH of amino acid at position i + 4
- **β-sheet:** Connects stretches of amino acids. The stretches are extended and form hydrogen bonds towards neighbouring stretches involving NH and CO groups. The neighbouring stretches can either run parallel or anti-parallel.

The well defined secondary structure elements prefer very characteristic pairs of ϕ/ψ backbone dihedral angles that occupy specific regions in the aforementioned Ramachandran plot (Figure 1a). The observed distributions in the Ramachandran plot are therefore not only driven by valid stereochemistry but are a direct consequence of preferred secondary structure elements. With increasing experimental data of full protein structures, the existence of α-helices and β-sheets were confirmed and new, less frequent, secondary structure elements have been



(a) Ramachandran Plot

(b) Secondary Structure Elements

(c) Ramachandran Plot - Only PRO

(d) Ramachandran Plot - Only GLY

Figure 1: Ramachandran plots showing ϕ/ψ dihedral angle pairs for all (a) or for amino acids exhibiting characteristic properties (c, d). (b): α-helix (black) and β-sheet (orange) with hydrogen bonds highlighted light green. The ϕ/ψ dihedral pairs of involved residues are plotted in (a). (c): Proline is the only proteinogenic amino acid with a secondary amine. It covalently links the Cα carbon to its N, resulting in decreased flexibility of ϕ. (d): Glycine has only a hydrogen as its sidechain, resulting in an increased structural flexibility.

observed. In an effort of standardisation, Kabsch and Sander introduced a vocabulary of secondary structures distinguishing between 8 elements with clear defined rules based on hydrogen bond patterns [79].

The full 3D arrangement of the secondary structure elements is the result of a folding process and is referred to as the tertiary structure of a protein. The main driving forces of folding are hydrophobic effects, van der Waals forces, ionic interactions and hydrogen bonds. They all oppose the reduction in entropy of the amino acid polymer that comes with folding and lead to a fold that is determined by the primary structure [7]. Exceptions to that statement include prions [129] or proteins that require the help of chaperones to fold [62]. In the case of water soluble proteins this typically leads to compact globular structures with hydrophilic sidechains pointing towards the solvent and hydrophobic sidechains buried in a hydrophobic core.

A significant fraction of the protein chains in a cellular environment have been found to be in direct contact and therefore build higher order complexes, so called oligomers [55, 76]. As an analogy to the previously discussed primary, secondary and tertiary structure, the term quaternary structures is used to describe the arrangement of protein chains into higher order complexes. If the complex is comprised by n copies of the same protein chain, we speak of a homo-oligomer. A hetero-oligomer is a combination of protein chains with different primary structure. Oligomerization allows evolution to explore an additional layer of functionality. A first example are allosteric interactions among subunits as they occur in haemoglobin [2] or GPCRs [51]. Another example demonstrating the importance of oligomerization is its role in building dynamic structures in the cytoskeleton [49] or when full viral capsids are built using dozens of protein chains as building blocks.

1.3 EXPERIMENTAL METHODS

The wold wide protein data bank (wwPDB) [15] is a database of structural information for large biological molecules, such as proteins and nucleic acids. The methods used to generate the deposited data can mostly be reduced to three: X-ray crystallography, NMR (Nuclear Magnetic Resonance) and EM (Electron Microscopy). As of May 9 2017, these methods contributed a total of 129'739 entries with almost 90% originating from X-ray crystallography. Most of the other entries come from NMR and only a tiny fraction of around 1.1% comes from EM. Nevertheless, technological advances in the field of EM have resulted in a massive increase of contributions in recent years. This section is intended to give a brief introduction to structure determination with these methods.

1.3.1 *X-ray Crystallography*

X-ray crystallography consists of growing a diffracting crystal, data acquisition, and solving the phase problem to obtain an electron density. A protein with known amino acid sequence is then fitted into the density with the help of automated computational methods.

In detail: monochromatic X-ray waves are elastically scattered upon interaction with electrons (scatterer), resulting in spherical wave fronts emerging at the locations of the scatterers. Most of these waves cancel each other out by destructive interference. However, if the scatterers are ordered in a lattice, constructive interference occurs at well defined directions depending on the lattice and the X-ray wavelength as described by Braggs law [24]. The observed diffraction pattern can be related to the underlying electron density using the Fourier transform. Since only intensities of the diffraction pattern can be measured and not the phase, the recorded information is incomplete. A problem known as the phase problem in crystallography. Direct methods to estimate phases only exist when the number of heavy atoms in the investigated protein is very low. Approaches to solve the phase problem therefore include the introduction of anomalies, solving the substructure of those anomalies using direct methods and use this information to infer the missing phases. Established techniques for this task are multiple isomorphous replacement (MIR), multi-wavelength anomalous dispersion (MAD) or single-wavelength anomalous dispersion (SAD) [3]. A widely used alternative is to guess initial phases given a protein structures that is likely to be structurally very similar (molecular replacement, [1]).

The phasing problem is indeed a key problem in the process of structure determination with X-ray crystallography. However, one should not forget that an actual protein crystal is needed in the first place. Even though many standardised protocols for crystal growing exist, this is often a laborious process [112].

1.3.2 *NMR*

In contrast to X-ray crystallography, nuclear magnetic resonance (NMR) spectroscopy is based on measurements in aqueous solution and therefore does not require crystals. While the structure determination aspect of NMR is limited to relatively small proteins, many NMR experiments exist to gain insights into the dynamics of proteins [85]. NMR exploits quantum mechanical properties of nuclei with non-zero spin, e.g. ^{1}H or ^{13}C, leading to an intrinsic magnetic moment. The application of a strong external magnetic field leads to an alignment of the magnetic moments that can be perturbed by a radio frequency pulse. As a consequence, the moments start to resonate in a measurable frequency. One could think that the resonance frequency is the same for all nuclei. This is not the case. As a result of the nucleus specific electronic environment, the magnetic field acting on each nucleus is modulated resulting in a shift (a so called chemical shift) of its resonance frequency. This effect is known as nuclear shielding. The measured signal that decays over time is the sum of all resonance frequencies and can be separated into its single frequency components using the Fourier transform. If the number of nuclei is small enough, each nucleus can be assigned a resonance frequency. Using complex patterns of radio pulses allows to exploit various types of interdependencies between nuclei from which constraints can be constructed [80]. Subsequent model building typically generates an ensemble of models satisfying the input constraints with decreasing variations as the number and the quality of input constraints increases [59].

1.3.3 *Electron Microscopy*

The main limitation of resolution in light microscopy is the wavelength of visible light (around 400-700nm). Atomic resolution would be three to four orders of magnitude smaller, hence out of reach. The wave particle dualism offers a viable alternative in the form of electrons that, according to de Broglie, have a wavelength that is inversely proportional to their velocity. Modern electron microscopes easily achieve wavelengths in the range of Å and, as an alternative to optical lenses, use magnetic lenses for magnification. While nanomaterials, e.g. metallic surfaces, can be imaged at subangstrom resolution, biological samples are much more sensitive to radiation damage [154]. Lower doses of electrons must be used instead, resulting in lower signal to noise ratios. Another challenge to overcome is the high vacuum environment in an electron microscope, requiring fixation for biological samples. A task for which Cryo-EM has established itself as a de facto standard to obtain images of biological samples [8]. Currently, the most often used approach for structure determination is single particle imaging. From every acquired image, subimages are extracted that contain exactly one protein. The proteins in a sample are randomly oriented which requires the classification of thousands of subimages into classes that represent the view onto the protein from a particular direction. Under the assumption of randomly distributed noise, class averaging increases the signal to noise ratio and allows the construction of a 3D density using the Fourier slice theorem [127]. Up to a few years ago, achieving atomic resolution with cryo-EM was limited to highly symmetric structures such as viral capsids due to advantages in class averaging. However, the introduction of direct electron detectors for image acquisition is currently revolutionising the field and starts to make such resolutions obtainable for smaller and lower symmetry proteins [98].

1.4 SEQUENCE, STRUCTURE AND FUNCTION

The primary sequence of proteins is subject to change as a direct consequence of evolution. With the assumption of tertiary structure and consequently also function being fully encoded by the primary sequence (Anfinsen Dogma, [7]), the capability to fold directly acts on evolution as selection criteria. However, protein structure has proven to be astonishingly robust towards mutation end even distantly related proteins often show high structural similarity. This fact has manifested itself as more structural information became available and has been shown by the work of Chothia & Lesk (Figure 2, [33]). This observation highlighted the importance of detecting evolutionary relationships and its value to infer structure and function. Advances in this field are among the great achievements in computational biology and current life science research could not be imagined without tools like BLAST (**B**asic **L**ocal **A**lignment **S**earch **T**ool) [5] or the various databases provided by the computational biology community. This section gives an overview over homology detection methodologies that will be of crucial importance throughout the whole thesis.

1.4.1 *Pairwise Sequence Alignments*

The goal of aligning two sequences A and B is to capture evolutionary events and to generate a residue to residue relationship. The alignment itself can be interpreted as a chain of events

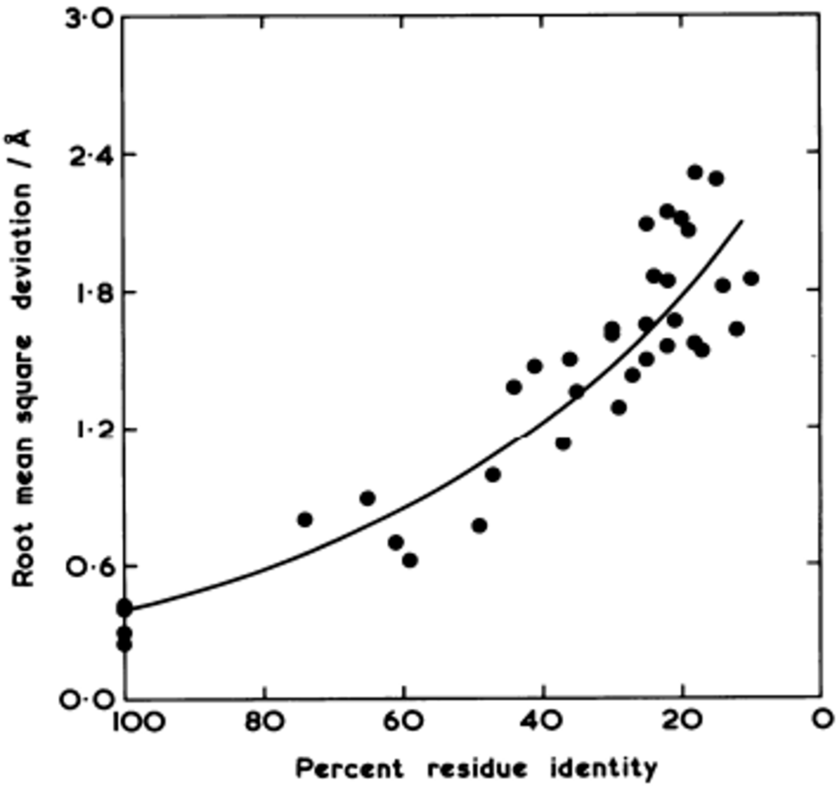


Figure 2: Relation between sequence identity and structural similarity of core residues. The results of the work from Chothia & Lesk suggest that similar sequence implies similar structure ([33], Reprinted by permission of John Wiley & Sons, Inc.).

that consists of matches (a residue in A is aligned with a residue in B), insertions in sequence A or insertions in sequence B. From a computational point of view this is an optimisation problem to find the optimal chain of events given some scoring scheme. In its simplest form, substitution matrices are used to score match events, i.e. matrices from the widely used BLOSUM (**B**lock **SU**bstitution **M**atrix) family [67]. Insertions on the other hand result in predefined penalty values. A popular algorithm to find the optimal full alignment between sequences A and B is the Needleman-Wunsch algorithm [120]. A variation thereof is the Smith-Waterman algorithm [148] that does not give a full global alignment but rather optimal local alignments. Both algorithms scale with a complexity of $\mathcal{O}(nm)$, where n and m represent the lengths of A and B. With increasing amount of available sequence data, the main application of sequence alignments shifted from actual pairwise alignments to database searches in order to identify evolutionary related sequences. The complexity of the described pairwise algorithms was the main bottleneck and led to the development of more efficient heuristics in the early 1990's which resulted in the still widely used BLAST algorithm [5]. Instead of performing full blown sequence alignments to all sequences in a database, the query sequence is split into short *words* that are used to to search for high scoring matches in the database. High scoring matches are then used as seed and extended to the left and to the right to generate *HSPs* (**H**igh **S**coring segment **P**airs). In a final step, neighbouring *HSPs* are combined to generate even longer alignments if certain scoring thresholds are fulfilled.

1.4.2 *From Substitution Matrices to Statistical Models*

BLAST in its initial form is still widely used today and good results can be obtained for closely related sequences. However, sensitivity declines quickly with evolutionary distance. The situation improved with the introduction of PSSMs (**P**osition **S**pecific **S**ubstitution **M**atrix), with PSI-BLAST (**P**osition **S**pecific **I**terative BLAST) [6] as the most prominent algorithm using them. In a first round, a sequence search is performed by standard BLAST using the classical substitution matrix approach. The identified sequences with their pairwise alignments are used to estimate amino acid frequencies for every position in the query sequence. The frequencies can be transformed to position specific substitution scores to finally obtain a PSSM. Instead of the original substitution matrix of size 20*20 (assuming 20 standard amino acids), the PSSM is a matrix of size 20*L with L being the length of the query sequence. The sequence database is searched again but now with the scores of the generated PSSM. The PSSM can iteratively be updated with newly found sequences until a maximum number of iterations is reached or no new sequences can be found. Using this approach, conservation patterns and the exact variation of amino acids at every position in the query sequence are implicitly considered. As a consequence, sensitivity for more distantly related sequences increases and the quality of the corresponding pairwise alignments is improved [6]. The drawback of this approach is that insertions are still handled heuristically by assigning constant penalty values. HMMs (**H**idden **M**arkov **M**odels) offer a way to describe insertions in a position specific manner and are able to incorporate all the advantages of a PSSM with a well understood probabilistic framework. Similar to a PSSM, HMMs are built iteratively. They finally contain position specific probabilities for all amino acids. But, additionally, also contain position specific probabilities for insertion and deletion events [88]. By not aligning HMMs to single sequences in a database anymore, but rather developing a formalism to align HMMs with other HMMs, sensitivity and alignment quality has been enhanced even further and represent the current state-of-the-art in homology detection [29]. But all this comes with the cost of increased computational complexity. Representatives of this class of algorithms are HHsearch [149] and HHblits [134].

1.5 PROTEIN MODELLING

The ever increasing improvement of DNA sequencing methodologies has lead to an explosion with regard to sequence information [65] as more and more full genomes become available. Despite the increasing number of experimentally determined protein structures, the amount of protein coding sequences without coverage of 3D structural data increases [141]. An established technique to increase this coverage and gain valuable structural insights is homology modelling [141]. Homology modelling exploits Cothia & Lesk's finding of structural similarity if two proteins are evolutionary related (Section 1.4) and has a wide range of applications including drug design [140], molecular docking [47], molecular replacement in X-ray crystallography [1], analysis of protein-protein interactions [169] and many more. Before giving a comprehensive description of all steps of a typical homology modelling procedure, we introduce the concept of a modelling engine. A modelling engine is a software package that builds 3D coordinates of a protein model given one or several homologues as template. Prominent examples are the MODELLER [158], Rosetta [151] or I-Tasser [167] software packages.

1.5.1 *Homology Detection and Homology Transfer*

Databases of known protein structure are queried for homologous entries (templates) given a desired target sequence. Methods performing this task have comprehensively been discussed in Section 1.4.1 and Section 1.4.2. For a particular template, coordinates are transferred based on the underlying pairwise sequence alignment. It's essentially copying over all residues that are aligned, even in case of a sequence mismatch. Every insertion leads to a chain break and requires attention in the subsequent loop modelling procedure. If the insertion happens in the template sequence, we speak of a deletion instead.

1.5.2 *Loop Modelling*

Every chain break is flanked by two stems that need to be connected. In case of an insertion, residues need to be added between stems that have originally been connected in the template structure. In a deletion, the stems need to be connected with no residues to do that. In any case, the stems are wrong by definition and every loop modelling algorithm must therefore handle the stems with a certain degree of flexibility. Assuming this problem as not existing, loop modelling consists of (1) generating candidate loops and (2) the selection of one of them. MODELLER generates initial loop conformations by simply placing all loop atoms on a line between the stem residues, randomising them a bit and applying a molecular mechanics minimization to obtain stereochemically feasible loop candidates. Rosetta and I-Tasser on the other hand implement sophisticated Monte-Carlo techniques to explore the available conformational space. All three engines therefore mainly rely on ab initio techniques. With the increasing amount of structural data that has experimentally been determined, database approaches gain importance and provide a viable alternative. The idea is to construct a database of observed structural conformations and query it for candidate loop conformations. Candidates with stems similar to the loop modelling problem of interest are extracted and undergo further scoring procedures. Prominent database methods include SuperLooper [68] or FREAD [32]. Whether choosing ab initio or database based methods, scoring is absolutely vital to either select a loop candidate or guide Monte Carlo procedures. Scoring can go from measures of valid stereochemistry over arbitrary energy functions towards knowledge based terms and will further be discussed in Section 1.5.5.

1.5.3 *Sidechain Modelling*

Many modelling tasks only consider backbone atoms and either completely neglect sidechains or use reduced representations thereof. As soon as sidechains are required, the already huge conformational space experiences another explosion. Due to almost constant bond lengths / angles [41], amino acid sidechains can approximately be described by dihedral angles, so called rotamers. They cluster around preferred conformations largely determined by their stereochemical properties [74]. Libraries compiled from structural analysis of high resolution X-ray structures can therefore reduce the available conformational space and serve as a starting point in the sidechain modelling problem by proposing rotamers for each amino acid. The local backbone conformation also has an influence on rotamer preference and the resulting backbone dependent rotamer libraries are considered to contain even more accurate rotamers [77]. Assuming a constant protein backbone, the sidechain modelling procedure starts with gathering a set of rotamers $R = [r_1, r_2, ..., r_n]$ for all l residues. The goal is to find $X = [x_1, x_2, ..., x_l]$ that minimizes:

$$F(X) = \sum_{i} E_{self}(R_i[x_i]) + \sum_{i} \sum_{j>i} E_{pair}(R_i[x_i], R_j[x_j]) \tag{1}$$

where E_{self} evaluates the energy of a rotamer itself and with respect to the constant environment. E_{pair} evaluates pairwise energies in between rotamers. For this formalism to hold, the underlying energy function must fulfil two properties:

1. **Pairwise Decomposable:** The energy function must allow to split between contributions towards the constant environment and contributions from pairwise interactions between rotamers.

2. **Symmetric:** $E_{pair}(R_i[x_i], R_j[x_j]) = E_{pair}(R_j[x_j], R_i[x_i])$ must hold to be independent from evaluation order.

Only pairwise energies with nonzero components have to be evaluated. An energy function with a quick convergence towards zero is therefore beneficial and leads to a problem complexity that relates roughly linear with the number of residues.

Finding the optimal solution on the other hand is more problematic. Crambin (46 residues) with 10 rotamers for every residue would already have 10^{46} possible rotamer combinations which makes a full enumeration of the solution space impossible. More sophisticated algorithms are required instead. No matter what algorithm is used in the end, the Goldstein criterion [54] has established itself as an efficient initial complexity reduction and is the basis for DEE (**D**ead **E**nd **E**limination). If there are two rotamers a and b at position i, a can be neglected if the following inequality holds:

$$\begin{aligned}&E_{self}(R_i[a]) - E_{self}(R_i[b]) + \\ &\sum_{j!=i} min_k(E_{pair}(R_i[a], R_j[k]) - E_{pair}(R_i[b], R_j[k])) > 0\end{aligned} \qquad (2)$$

In words: no matter what rotamers are set at any location $j! = i$, a is dominated by b and can therefore not be part of the overall optimum. Note, that only locations j in close proximity to i with non zero pairwise energies have to be evaluated. Despite reduced complexity, a full enumeration is still not possible and many methods rely on a graph $G = (N, V)$ to perform the final optimization, with various versions of SCWRL [27, 87] as an example. A set of rotamers at location i is represented by a node $n_i \in N$. If there is a non zero interaction between any of the rotamers in n_i and n_j, a vertex v_{ij} gets added to V. Different graph decompositions can be exploited to split the overall problem into smaller subproblems, solve them separately and merge the local solutions into the global solution [27, 166]. Alternative approaches exploit integer programming [84] or Monte Carlo techniques [69, 99]. However, the latter does not guarantee to find the overall optimum for a sidechain modelling problem at hand.

1.5.4 *Energy Minimization*

As a result of the approximative nature of many modelling algorithms, protein models contain stereochemical irregularities and clashes. Methods for their detection exist (Section 1.5.5.1), but resolving them requires more elaborate techniques. The method of choice is energy minimization using physics based molecular mechanics forcefields. A forcefield is a functional form and a set of parameters to describe the potential energy of a macromolecular system. The functional form is typically a set of functions describing covalent bonds, bond angles, dihedral angles and nonbonded terms such as Lennard-Jones and Coulomb interactions. The according parameters are derived from experiment or quantum mechanical calculations and prominent examples include the CHARMM [71] or AMBER [28] forcefields. The main application of such a forcefield is to parameterize a macromolecular system and, by solving Newtons' equations of motion, move it through time step by step to get insights into the dynamics of the system. This is not required for energy minimization. In terms of modelling, the system typically consists of the protein model in a vacuum environment and instead of solving Newtons' equations of motion, the atom positions in the system are altered to minimize the potential energy. Typical minimizers apply steepest descent or conjugate gradient approaches and are capable of resolving most of the common stereochemical irregularities.

1.5.5 *Quality Estimation*

Every protein structure, no matter whether it is a homology model or even a model constructed from high resolution X-ray data, contains errors. Quality estimation tools are developed to quantify these errors. This is particularly important, but also increasingly difficult, when more and more remote homologues are used as an underlying template in homology modelling. The range of possible applications goes from selecting the best model in a set of alternatives towards absolute quality estimates and finally the detection of local errors. The tools can largely be divided in plausibility checks, physics based, knowledge based and consensus based.

1.5.5.1 *Plausibility Checks*

Protein structures closely follow the rules of physics and chemistry that can be examined to assess the plausibility of a protein model. Various tools are routinely applied on models based on experimental data [133] and also gained importance in the field of homology modelling [81]. Examples are MolProbity [30], WHAT_CHECK [70] or PROCHECK [94]. They all provide slight variations of several stereochemistry checks. The match of of bond lengths / bond angles can be compared with reference values, e.g. from the work of Engh and Huber [41]. Also backbone dihedral angles show clear preferences and ϕ/ψ backbone dihedral pairs that violate the observations from the Ramachandran plot (Figure 1) are a strong indication for local distortions. Other checks can include the planarity of rings in amino acid sidechains or the detection of clashes, i.e. non-bonded atoms that are closer than the sum of their van der Waals radii. Despite the importance of a valid stereochemistry for structural analysis or further processing, i.e. molecular mechanics procedures, valid stereochemistry does not automatically imply an accurate protein model close to the desired native structure. The same is true for the other way around. Problems with stereochemistry can also occur in protein models that would be close to native.

1.5.5.2 *Physics Based Quality Estimation*

Given the thermodynamics hypothesis (Anfinsen dogma [7]), the native structure of a protein is determined by its free energy minimum. It should therefore be possible to detect a native structure among a set of alternatives given an accurate physics based free energy calculation. One possibility is to use a molecular mechanics forcefield such as CHARMM [71] or AMBER [28]. However, their successes in the task of quality estimation is controversial [95, 164] as the potential function can be very sensitive to small structural changes.

1.5.5.3 *Knowledge Based Quality Estimation*

Knowledge based quality assessment can include arbitrary measures of what is *known* from protein structures. This can be radius of gyration, secondary structure content, atom packing and many more [83]. One particularly important class of knowledge based quality estimation tools are statistical potentials of mean force. Their importance in the field of protein modelling makes it necessary to describe them in detail. The increasing amount of available structural data made it possible to statistically analyse certain types of interactions. In an effort to construct a scoring function for protein models, Sippl and coworkers assumed interatomic distances in a protein structure to be Boltzmann distributed [145]. The probability of a certain conformation $c_i \in C$ can be related to an energy $E(c_i)$ using the Boltzmann equation, where k_B is the Boltzmann constant and T the temperature of the system:

$$\begin{aligned} p(c_i) &= \frac{1}{Z(C)} e^{-\frac{E(c_i)}{k_B T}} \quad \text{with} \quad Z(C) = \sum_j e^{-\frac{E(c_j)}{k_B T}} && (3) \\ \Rightarrow E(c_i) &= -k_B T \ln(p(c_i)) - k_B T \ln(Z(C)) && (4) \end{aligned}$$

Instead of evaluating absolute energies, the energy difference with respect to some reference energy is more tractable. This allows to introduce the dependency on a sequence s_i that gives raise to the observed conformation c_i.

$$\begin{aligned}\Delta E(c_i|s_i) &= E(c_i|s_i) - E(c_i) \qquad (5)\\ &= -k_B T \ln\left(\frac{p(c_i|s_i)}{p(c_i)}\right) + k_B T \ln\left(\frac{Z(C)}{Z(C|S)}\right) \qquad (6)\end{aligned}$$

Assuming $Z(C|S) = Z(C)$ we get:

$$\Delta E(c_i|s_i) = -k_B T \ln\left(\frac{p(c_i|s_i)}{p(c_i)}\right) \qquad (7)$$

We can now evaluate the energy of observing a single conformation c_i given s_i, i.e. the identity of the interacting particles, with respect to the energy of observing c_i at all. In case of a full protein model with conformation C and an amino acid sequence S we can also estimate the total difference in energy by summing up all single contributions considering them as additive "microstates":

$$\Delta E(C|S) = -\sum_i k_B T \ln\left(\frac{p(c_i|s_i)}{p(c_i)}\right) \qquad (8)$$

This formalism is not limited to interatomic distances but has found to be applicable to many other structural features that can be described probabilistically. One example is the quality estimation tool QMEAN [13] that linearly combines the outcome of several different statistical potentials to evaluate the quality of a protein model. The underlying statistical potentials do not only consider interatomic distances, but also backbone dihedral angles and protein packing.

A remaining question is how exactly to derive the required probability distributions $p(c_i|s_i)$ and $p(c_i)$. The first distribution is typically extracted from experimentally determined structural information. The second distribution, the so called reference distribution, is more controversial and the literature describes different approaches to construct it. Many tools use experimental data [105, 138, 146]. Others construct reference distributions from theoretical considerations [144, 170]. There is no consensus on what works best [38] but looking at the problem from a Bayesian point of view sheds some light on what the reference distribution actually is. The probability of observing a conformation c_i given a sequence s_i is proportional to the likelihood of observing this sequence given this conformation times the prior knowledge we have about the conformation:

$$p(c_i|s_i) = \frac{p(s_i|c_i)p(c_i)}{p(s_i)} \propto p(s_i|c_i)p(c_i) \qquad (9)$$

Note, that $p(s_i)$ is considered to be constant. Under the assumption that all interactions are independent, we can express the likelihood for the sequence S of a full protein model with conformation C as:

$$p(S|C) \propto \prod_i \frac{p(c_i|s_i)}{p(c_i)} \tag{10}$$

The goal is to find C that maximises $p(S|C)$ which is equivalent to maximizing its logarithm:

$$\ln(p(S|C)) \propto \sum_i \ln\left(\frac{p(c_i|s_i)}{p(c_i)}\right) \tag{11}$$

We can immediately see the similarities to Equation 8 that is minimized by maximising the expression in Equation 11. From a Bayesian point of view, the previously discussed reference distribution is the prior knowledge we have about a conformation c_i. If applied on fairly accurate homology models, Samudrala & Moult consequently argued that this prior distribution is reasonably well approximated by experimentally determined structural information [138]. Another advantage of using a probabilistic formalism is to evade the dispute about the physical meaning of statistical potentials of mean force derived from the Boltzmann distribution [12, 155].

1.5.5.4 *Consensus Based Quality Estimation*

Consensus methods do not take into account any physical interaction. A quality estimate is predicted by assessing the consistency in an ensemble of conformations. Following Levinthals paradox [97] we can say: There is only one way of doing it right but an infinite number of ways of doing it wrong. If the conformations are not completely random, at least some of them are expected to be close to the native structure, hence similar. This makes consensus methods particularly successful in the context of the CASP experiments where a large amount of alternative conformations are available [89–91]. Another ideal application of consensus methods is when thousands of alternative ab initio conformations are generated using Monte Carlo procedures. However, in a classical homology modelling approach, the number of alternative conformations is limited. This hampers the practicability of consensus methods.

1.6 OBJECTIVES

The main objective of this thesis is to advance various aspects of protein modelling. Two main parts can be distinguished. First, tools and algorithms are presented to improve and extend the protein model quality estimation tool QMEAN. Second, the problem of protein modelling itself is discussed and ways to improve the general model accuracy in the context of SWISS-MODEL are implemented. This leads to a total of four chapters.

First, QMEAN is extended to handle the very specific case of local membrane protein model quality assessment, leading to a novel method: QMEANBrane. The original statistical potentials of mean force are retrained to faithfully reflect the large variation of molecular properties that act on a protein in a membrane environment.

Second, the power of consensus is applied to the problem of local quality assessment for soluble protein models. Instead of relying on an ensemble of models, ensemble information in form of distance constraints is extracted from the ever increasing amount of experimentally determined structures. The extracted distance constraints give a novel score for local quality assessment: DisCo. In combination with QMEAN, this gives QMEANDisCo.

Third, the requirements for modelling engines in the context of the SWISS-MODEL webserver are discussed. A lack of a free, efficient and state-of-the-art modelling engine is identified and the solution in form of ProMod3 is presented.

The fourth and last chapter is a direct continuation of the third and extends the implemented algorithms in ProMod3 to combine structural information from multiple templates to obtain one single protein model.

2 QMEANBRANE

This chapter has been published as:

G. Studer[1,2], M. Biasini[1,2], and T. Schwede[1,2]. "Assessing the local structural quality of transmembrane protein models using statistical potentials (QMEANBrane)." In: Bioinformatics 30.17 (2014), pp. i505–511.

[1] Biozentrum, University of Basel, Klingelbergstrasse 50 / 70, 4056 Basel, Switzerland; [2] SIB Swiss Institute of Bioinformatics, Basel, Switzerland

Author Contributions: GS and MB implemented the software (initials ordered by amount of contributions). GS conducted the research and wrote the manuscript, TS edited and revised the manuscript.

Motivation: Membrane proteins are an important class of biological macromolecules involved in many cellular key processes including signalling and transport. They account for one third of genes in the human genome and $> 50\%$ of current drug targets. Despite their importance, experimental structural data are sparse, resulting in high expectations for computational modelling tools to help fill this gap. However, as many empirical methods have been trained on experimental structural data, which is biased towards soluble globular proteins, their accuracy for transmembrane proteins is often limited.
Results: We developed a local model quality estimation method for membrane proteins ('QMEANBrane') by combining statistical potentials trained on membrane protein structures with a per-residue weighting scheme. The increasing number of available experimental membrane protein structures allowed us to train membrane-specific statistical potentials that approach statistical saturation. We show that reliable local quality estimation of membrane protein models is possible, thereby extending local quality estimation to these biologically relevant molecules.
Availability: Source code and datasets are available on request.

As of May 2017, QMEANBrane is available at:
https://swissmodel.expasy.org/qmean/

2.1 INTRODUCTION

Protein modelling plays a key role in exploring sequence structure relationships when experimental data are missing. Modelling techniques using evolutionary information, in particular homology/comparative modelling, developed into standardized pipelines over recent years. An indispensable ingredient of such a pipeline is the accuracy estimation of a protein model, directly providing the user with information regarding the range of its possible applications [9, 141, 142]. In this context, global model quality assessment tools are important for selecting the best model among a set of alternatives, whereas local model estimates assess the plausibility and likely accuracy of individual amino acids [13, 43]. Various techniques have been developed to address this question, with consensus methods and knowledge-based approaches showing best results in blind assessments [90]. Consensus approaches require an ensemble of models with structural variety, reflecting alternative conformations [135, 147].

in contrast, knowledge-based methods (such as statistical potentials) can be applied to single models but are in general less accurate than consensus methods and exhibit strong dependency on the structural data they have been trained on.

The unique physicochemical properties of biological membranes give rise to interactions that are energetically discouraged in soluble proteins, and vice versa [160]. However, most scoring functions using knowledge-based methods [13, 107, 132, 146, 171] have been trained on soluble proteins. Thus, they perform poorly when applied to models of membrane proteins. This specific, but highly relevant, important aspect of protein model quality assessment has received only little attention in recent years [66, 131]. With the growing amount of available high resolution membrane protein structures [52, 159] the template situation for homology modelling procedures is improving quickly and, even more important for this work, it is gradually becoming possible to adapt knowledge-based methods to this class of models.

As a result of such efforts, we present QMEANBrane, a combination of statistical potentials targeted at local quality estimation of membrane protein models in their naturally occurring

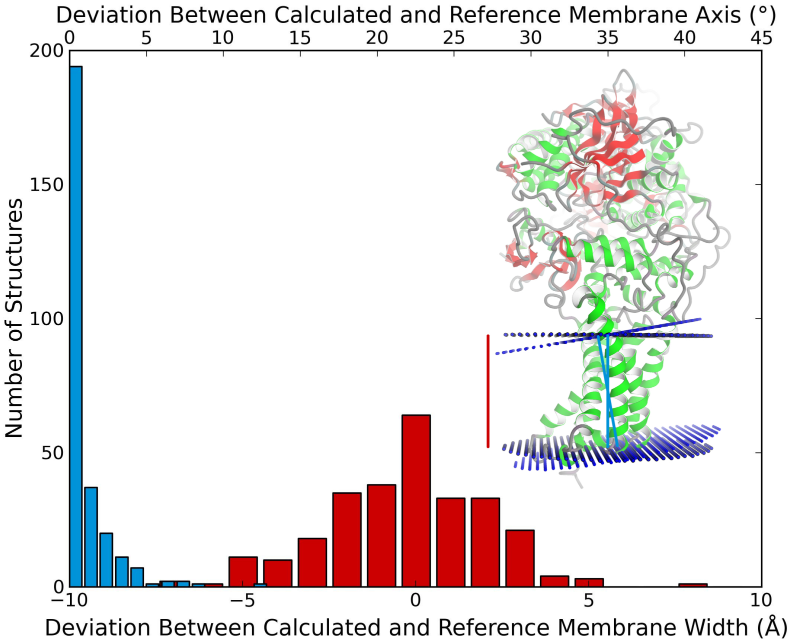


Figure 3: Difference between membrane predictions of our algorithm and the predictions of OPM on the 200 high-resolution structures used to train membrane-specific statistical potentials.

oligomeric state: after identifying the transmembrane region using an implicit solvation model, specifically trained statistical potentials get applied on the different regions of a protein model (Figure 3, Figure 4). To overcome statistical saturation problems, a novel approach for deriving statistical potentials from sparse training data has been devised. We have benchmarked the performance of the approach on a large heterogeneous test set of models and illustrate the result on the example of alignment errors in a transmembrane model.

2.2 MATERIALS & METHODS

2.2.1 *Target Function*

The similarity/difference between a model and a reference structure can be expressed in the form of distances between corresponding atoms in the model and its reference structure after performing a global superposition. However, this global superposition approach fails to give accurate results in case of domain movements. To overcome such problems, e.g. in the context of the CASP [119] experiments, the structures are manually split into so-called assessment units and evaluated separately [153]. This manual procedure is time consuming and not suitable for automate large-scale evaluation, e.g. such as performed by CAMEO [60]. Alternatively, similarity/difference between a model and reference structure can be expressed in the form of superposition-free measures such as the local Distance Difference Test (lDDT) score [109] assessing the differences in interatomic distances between model and reference structure. In this work, the lDDT inclusion radius is set to 10Å to ensure local behaviour. See Figure 9 for a comparison of different structural similarity measures (Cα-distance, dRMSD, lDDT and CAD score [121]).

2.2.2 *Membrane segment definition*

The OPM database [103] applies minimization of a free energy expression to predict the transmembrane part of a protein structure [102]. In this work, we use a similar but simplified approach, still resulting in a robust approximation of the membrane segment definition. The energy expression is defined as

$$\Delta G = \sum_{i} \sigma^{wat \to bil} f(z_i) ASA_i \tag{12}$$

with $\sigma^{wat \to bil}$ representing the transfer energy from water to decadiene for atom i per Å^2 [101], $f(z_i)$ the hydrophobicity as a function of the distance to the membrane centre z_i and ASA_i the accessible surface area of atom i in Å^2 as calculated with NACCESS (www.bioinf.manchester.ac.uk/naccess). Not all atoms facing the surface, as determined by NACCESS are in contact with the membrane, even if they fall in between the lipid bilayer, e.g. as is the case for hydrophilic pores. To determine the subset of surface atoms in direct contact with the lipid bilayer, the protein structure surface as calculated by MSMS [139] is placed onto a 3D grid, marking every cube in the grid containing surface vertices. The application of a flood fill algorithm (http://lodev.org/cgtutor/floodfill.html) on every layer along the z-axis then allows the generation of a subset of potentially membrane facing atoms.

The parameters describing the membrane (i.e. tilt angle relative to z-axis, rotation angle around z-axis, membrane width and distance of membrane centre to origin) first undergo a coarse grained sampling to identify the 10 best parameter sets for further refinement using a Levenberg–Marquardt minimizer.

This procedure is repeated several times with different initial orientations of the structure to find the set of parameters leading to the lowest total free energy.

The bilayer consists of a hydrocarbon core flanked by interface regions with a large chemical heterogeneity [161]. It is known that the properties of a membrane protein are strongly influenced by the interaction with the phospholipid bilayer, and a simple split into a membrane and soluble part would not faithfully reflect the variation of molecular properties along the membrane axis [17]. To catch these variations along the membrane axis, we split the transmembrane proteins into three parts, which are treated separately: an interface part consisting of all residues with their Cα atom positions within 5Å of the membrane defining planes, a core membrane part consisting of all residues with their Cα atom positions in between the two membrane defining planes not intersecting with the interface residues and finally, a soluble protein part consisting of all remaining residues.

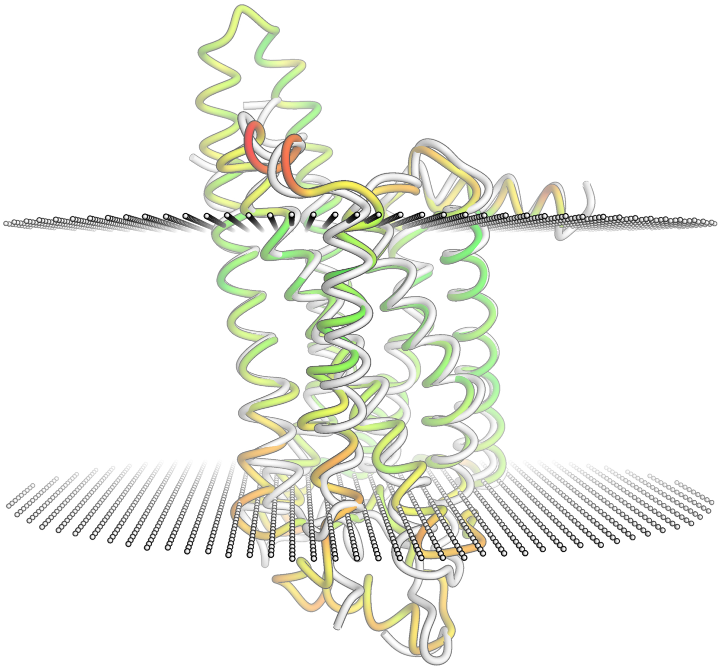


Figure 4: Local QMEANBrane scores mapped on the best performing model (mod9jk) regarding RMSD of the GPCR Dock experiment 2008. Reference structure (2.6 Å crystal structure of a human A2A ad- enosine receptor bound to ZM241385, PDB: 3eml) and membrane-defining planes are shown in white

2.2.3 *Model quality predictors*

To assess the membrane protein models quality, we mainly rely on statistical potential terms, combined with the relative solvent accessibility of each residue as calculated by DSSP [79]. The four statistical potential terms (their exact parameterizations are described in the Section 2.6.3), are the following:

1. **All-atom interaction Term:** Pairwise interactions are considered between all chemically distinguishable heavy atoms. A sequence separation threshold has been introduced to allow focusing on long-range interactions and reduce the influence of local secondary structure. Interactions originating from atoms of residues closer in sequence than this threshold are neglected.

2. **Cβ interaction Term:** This term assesses the overall fold by only considering pairwise interactions between Cβ positions of the 20 standard amino acids. In case of glycine, a representative of the Cβ position gets constructed using the backbone as anchor. The same sequence separation as in the all-atom interaction is applied.

3. **Solvation Term:** Statistics are created by counting close atoms around all chemically distinguishable heavy atoms not belonging to the assessed residue itself.

4. **Torsion Term:** The central ϕ/ψ angles of three consecutive amino acids are assessed based on the identity of the involved amino acids using a grouping scheme described by Solis and Rachovsky [150].

The torsion term trained on soluble structures is applied to the whole membrane protein model. Conversely, solvation and interaction terms are specifically trained for and applied to the soluble, membrane and interface segments with different potentials for α-helical and β-barrel transmembrane structures. A residue belonging to one of these parts 'interacts' with all atoms in the full model, and a final score is assigned by averaging all scores originating from interactions associated with this specific residue. For the solvation and torsion terms, we use a formalism closely related to the statistical potentials of mean force [145]. However, instead of referring to an energy expression, we rather look at the problem as a log odds score between the probability of observing a particular interaction between partner s with conformation c relative to some reference state:

$$S(c|s) = -\ln\left(\frac{p(c|s)}{p(c)}\right) \tag{13}$$

In case of sparse data, $p(c|s)$ cannot be expected to be saturated. Sippl and co-workers have proposed to use a combination of the extracted sequence-specific probability density function (pdf) and the reference state. The influence of the reference state vanishes at a rate determined by the newly introduced parameter σ towards large numbers of interactions (N) with sequence s:

$$p(c|s) \approx \frac{1}{1+N\sigma}p(c) + \frac{N\sigma}{1+N\sigma}p(c|s) \tag{14}$$

Using the aforementioned formalism, this leads to

$$S(c|s) \approx \ln(1+N\sigma) - \ln\left(1+N\sigma\frac{p(c|s)}{p(c)}\right) \tag{15}$$

Because of the increased abundance of structural information for soluble protein structures during the last decades, the use of the σ parameter has become largely unnecessary. However, for membrane proteins, data scarcity is still an issue and needs to be handled accordingly. In Figure 2.6.1, an analysis of the saturation behaviour of the different statistical potential terms is provided, suggesting a sufficient amount of training data for the solvation term, whereas the two interaction terms require more data to be fully saturated (Figure 8). For these cases, we introduced a treatment for sparse data by assuming that the statistics for soluble proteins are fully saturated. In other words, if there are no sufficient data available from membrane structures, we refer to the information we have from all protein structures to get a hybrid score:

$$\begin{aligned} HS(c|s) &= -\ln\left(\frac{1}{1+N\sigma}f_1 + \frac{N\sigma}{1+N\sigma}f_2\right) \\ &= \ln(1+N\sigma) - \ln(f_1 + N\sigma f_2) \end{aligned} \tag{16}$$

With f_1 representing the fraction of the probabilities of sequence-specific interactions and a reference state, where the pdfs of the specific interactions are saturated, and f_2 the fraction between the probabilities of sequence-specific interactions and a reference state, where the pdfs of the specific interactions are not necessarily saturated, as it may occur for membrane- and interface-specific cases. For regions of the pdf with zero probability as they, for example, occur at low distances in pairwise interaction terms, we applied a constant cap value to avoid infinite scores.

2.2.4 *Training datasets for statitical potentials*

The pdfs to calculate the statistical potentials for the soluble part are built using statistics extracted from a non-redundant set of high resolution X-ray structures. PISCES [157] has been used with the following parameters: sequence identity threshold 20%, resolution threshold 2 Å and R-factor threshold 0.25. Because only standard amino acids can be handled by QMEANBrane, a prior curation of the training structures is necessary. Non-standard amino acids such as phospho-serine or selenomethionine have therefore been mapped to their standard parent residues. For the selection of appropriate membrane protein structures, we rely on the OPM database [103]. As of October 2013, OPM contained 746 unique PDB IDs of structures with transmembrane segments. Applying a resolution threshold of 2.5 Å, removing all chains with <30 membrane-associated residues and considering only one chain in case of homo-oligomers results in 283 remaining chains from 200 structures. Clustering the chains based on their SEQRES sequences with kClust [63] using a sequence identity threshold of 30% resulted in 187 clusters, 140 of them from helical transmembrane structures and 47 from β-barrel structures. All entries are used in the calculation of the pdfs, where a chain originating from a cluster with n members is downweighted and contributes with a weight of $1/n$ to the final distributions. These final distributions have then been extracted by considering the corresponding chains, using the full protein structure in the oligomeric state as assigned by OPM as environment.

2.2.5 *Datasets for training linear combinations*

A set of 3745 models for soluble proteins was generated by selecting a set of non-redundant high-resolution reference structures from the PDB using PISCES (maximum 20% sequence identity, resolution better 2Å, X-ray only), extracting their amino acid sequences and building models using the automated SWISS-MODEL pipeline [82] by excluding templates with a sequence identity >90% to the target (P. Benkert, personal communication). OPM was used to identify reference structures (resolution <3.0 Å) to generate membrane protein models. Structures with <30 membrane-associated residues and hetero-oligomeric complexes were excluded. In all, 132 unique PDB IDs, which had more than one suitable template, have been selected as targets for modelling. Templates identified with HHblits [134] showing a sequence alignment coverage >50% served as input for MODELLER [137] and resulted in 3226 models with oligomeric states equivalent to the template structure. Removal of redundancy, i.e. models originating from templates with same sequence, and removal of obvious incorrect oligomeric states upon visual inspection resulted in a set of 557 models, 386 with helical transmembrane parts and 171 β-barrels.

2.2.6 *Spherical smoothing for noise reduction*

Averaging/smoothing can reduce noise introduced by quality predictors on a per-residue level, resulting in single residue scores, which more accurately reflect the local model quality. Smoothing in space tends to outperform sequential smoothing. In the proposed algorithm, every residue gets represented by its Cα position. The final quality predictor score for a residue is calculated as a weighted mean of its own value and the values associated to surrounding residues:

$$s_i = \sum_j w_j s_j \tag{17}$$

with s_i representing the final score at position i, w_j the weight of score at position j and s_j the score at position j. The weights are calculated in a Gaussian-like manner and normalized, so they sum up to one:

$$w_j = \begin{cases} \frac{1}{Z\sqrt{2\pi\sigma^2}} e^{0.5\left(\frac{d_{ij}}{\sigma}\right)^2} & \text{if } d_{ij} < 3\sigma \\ 0 & \text{else} \end{cases} \tag{18}$$

with w_j representing the weight of score at position j, d_{ij} the distance from position i to position j, σ the standard deviation of the Gaussian-like formalism to control how fast the influence of a neighbouring score vanishes as a function of the distance (5 Å turned out to be a reasonable σ) and Z as normalization factor.

2.2.7 *Per amino acid weighting scheme*

QMEANBrane uses a linear model fitted on the per-residue lDDT score to combine the single quality predictors. To remove amino acid-specific biases, such a linear model is trained for every standard amino acid:

$$s_i = \sum_j w_j s_{ij} \tag{19}$$

s_i is the combined score of residue at position i, w_j the weight of quality predictor j and s_{ij} the score of quality predictor j at position i.

2.2.8 *Implementation*

QMEANBrane is designed on a modular basis, implementing computationally expensive tasks in a C++ layer. All functionality is made fully accessible from the Python language and can directly be embedded into the computational structural biology framework OpenStructure [20, 21], allowing to assemble custom assessment pipelines to address more specific requirements.

2.3 RESULTS AND DISCUSSION

2.3.1 *Membrane prediction accuracy*

To evaluate the performance of our membrane finding algorithm, a comparison with the result obtained by OPM has been performed on the 200 structures used to train the membrane-specific statistical potentials. At this point, OPM is assumed to be the gold standard, even though it is a calculation by itself. By further considering the membrane width as the main feature of accuracy, 95% of the absolute width deviations are <4Å. In terms of translational distances, this corresponds to a 'misprediction' of 2–3 residues for helices and about 1–2 residue for sheets (Figure 3). Interestingly, using this approach, it is not only possible to automatically detect transmembrane regions but also to distinguish between transmembrane and soluble structures in general (Figure 10).

2.3.2 *Performance on the test dataset*

For a first analysis of performance on predicting local scores of membrane-associated residues in transmembrane protein models, we used the previously described model set for training the linear weights. Clusters have been built by applying kClust on the target sequences with a sequence identity threshold of 30%. The local scores for the membrane-associated residues of one cluster have then been predicted using linear models trained on all residues from models not belonging to that particular cluster (Table 1, Figure 13).

Quality Predictor	Helical structures	β-barrel structures
exposed	0.39	0.15
Torsion	0.43	0.47
Cβ interaction	0.51	0.49
Solvation	0.55	0.51
All atom interaction	0.63	0.58
All predictors combined	0.71	0.67

Table 1: Performances of single quality predictors and their combination on membrane-associated residues in our test set, measured as Pearsons' r between predicted score and actual local lDDT

2.3.3 *Independent performance evaluation on models of the GPCR Dock experiments*

Not many independent compilations of membrane protein models with known target structures exist. For a performance evaluation and comparison with other widely used quality assessment tools, we rely on the models generated during the GPCR Dock experiments 2008/2010 [92, 117] (Figure 4). A total of 491 models for three different targets, the human dopamine receptor, the human adenosine receptor and the human chemokine receptor were available. Receiver operating characteristic (ROC) analysis with the local lDDT as target value has been performed on all membrane-associated residues as defined by OPM, showing a clear superiority of QMEANBrane over other methods such as ProQ2 [132], QMEAN [13], ProQM [131], Prosa [162], Verify3D [107] or DFire [170] (Figure 5). Removing all GPCR/ Rhodopsin structures from the training data has only a minor

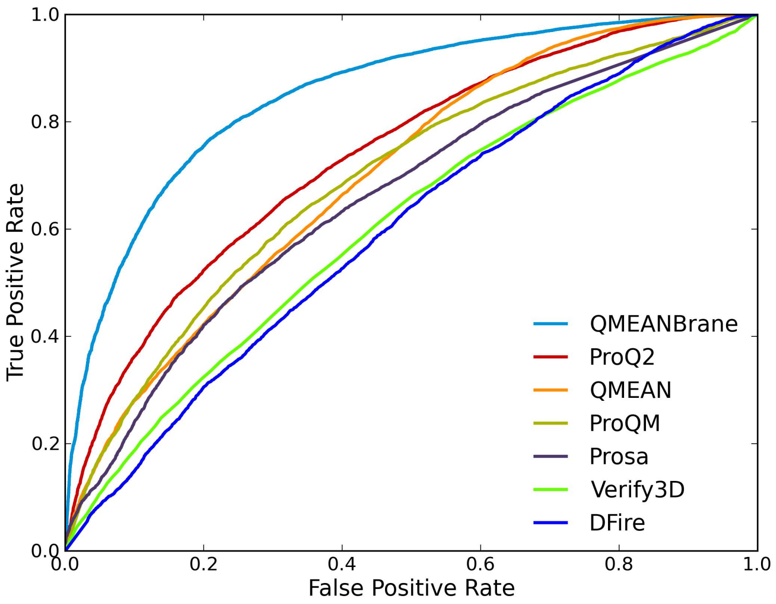


Figure 5: ROC analysis of all membrane-associated residues of the models of the GPCR Dock experiments with local lDDT as target value and a class cutoff of 0.6

effect. See Figure 11 for a more detailed performance analysis taking other measures of similarity into account. Because ProQM is the only other method specifically developed for the particular case of membrane protein model quality assessment, we also performed a direct comparison of QMEANBrane and ProQM on the dataset used to test/train ProQM in Figure 12.

2.3.4 *Retrospective analysis of modelling examples*

To illustrate the usefulness of QMEANBrane in tackling problems as they occur in real modelling cases, two targets with known structures have been selected for a more detailed analysis using the recently released SWISS-MODEL workspace [22]. The H+ translocating pyrophosphatase from *Vigna radiata* (PDB ID: 4A01) and a dopamine transporter of *Drosophila melanogaster* (PDB ID: 4M48). Models based on different target-template alignments have been compared to test QMEANBrane's capability of detecting incorrect alignments, particularly alignment shifts in transmembrane helices. (Alignments are available in the Supplementary Materials.)

The pyrophosphatase has, with the sodium translocating pyrophosphatase from *Thermotoga maritima* (PDB ID: 4AV3), a quite

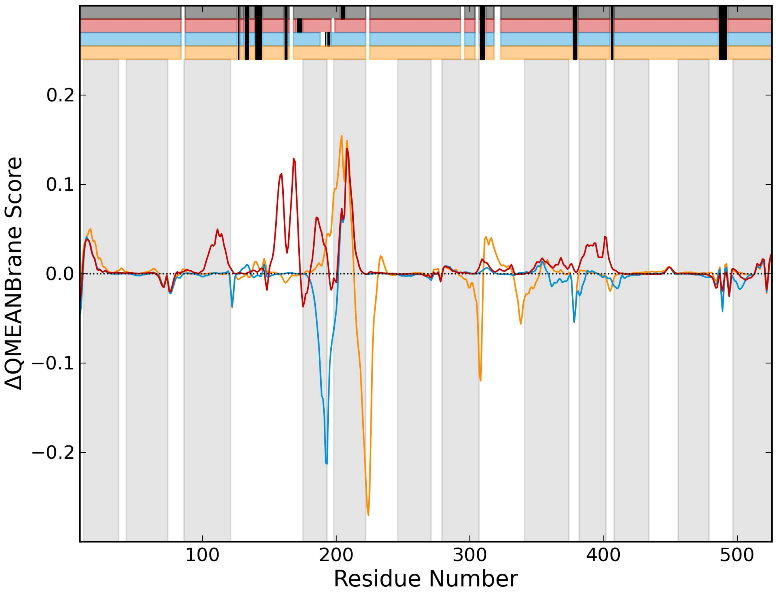


Figure 6: Difference of QMEANBrane scores of three dopamine transporter models with modified alignments versus the model built with the initial HHblits alignment, represented by the first horizontal bar. Insertions are marked black, and deletions are marked white. Second bar: shift of the insertion towards the N-terminus in front of helix 4, third bar: shift of insertion towards the N-terminus in between helices 4 and 5, fourth bar: shift of the insertion towards the C-terminus

close homologue (sequence identity >40%). Nevertheless, the alignments provided by BLAST [5] and HHblits differ significantly. Because the BLAST alignment has a lower coverage, not including the first transmembrane helix, only the part covered by both alignments is considered. Figure 14 shows a comparison of the QMEANBrane scores from the two models built with the different alignments. Two transmembrane helices contain an alignment shift of three residues, resulting in a clear local increase of the QMEANBrane scores of the model built with the HHblits alignment relative to the model built with the BLAST alignment. The higher quality of the HHblits model gets confirmed by its global lDDT of 0.63 versus 0.59 of the BLAST model.

For the dopamine transporter example, we chose an amine transporter from *Aquifex aeolicus VF5*, identified by HHblits with a sequence identity of ~24%, as the primary template. Despite the good coverage, a major problem occurs in transmembrane helix 5. The initial HHblits alignment has an insertion of three residues enforcing a helix break and an unnatural bulge within the transmembrane part. To analyse possible modifications of the initial alignment, we rely on QMEANBrane to compare the relative differences in the models with alternative alignments with the initial model (Figure 6, Figure 7).

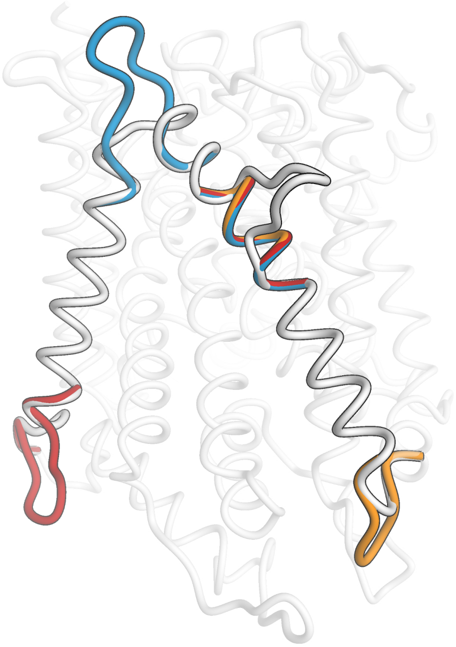


Figure 7: Structural effects of the alignment modifications shown in Figure 6. The model based on the initial HHblits alignment is coloured white; the other models are coloured according to the horizontal bar alignment representation in Figure 6

Three different alternative alignments were considered: the first is to shift the helix insertions towards the C-terminus. Despite the increase of the QMEANBrane score at the location of the alignment modification, the scores in helix 5 towards the C-terminus drop significantly, suggesting no improvement of the overall model quality. As second alternative, the insertion has been shifted into the loop connecting transmembrane helices 4 and 5. Because of their proximity, a distortion of both involved helix endings was inevitable, thus unfavourable. The third alternative, shift of the insertion towards the N-terminus before helix 4, and introducing an additional deletion in the aforementioned loop increasing the local sequence identity in helix 4, consistently increases the QMEANBrane scores in helices 4 and 5, as well as the helices close in space. These findings are confirmed by the global lDDT scores of the models built based on those alignments (initial alignment: 0.54, shift into middle: 0.54, shift towards C-terminus: 0.53, shift towards N-terminus: 0.57).

2.4 CONCLUSION

Investigating function and interactions in membrane proteins is an active field of research, with modelling techniques as an important tool to bridge the gap when structural data are missing. Comparative modelling methods automatically profit from the increased number of available experimental membrane structures, which can be used to build models for membrane proteins [50]. However, most knowledge-based approaches fail in assigning reliable local quality estimates when confronted with the unique structural features and interactions resulting from direct contact with the phospholipid bilayer.

With QMEANBrane, we present a framework that widely covers the aspects of membrane protein model quality assessment. In a first step, our membrane detection method allows to reliably locate the transmembrane part of the model. We introduce an interface region to account for the non-isotropy of protein properties along the z-axis. Statistical potential terms were trained specifically for these three regions, introducing a new hybrid potential formalism to circumvent problems arising from a lack of sufficient training data. The final local scores are then calculated using linear models trained for all 20 standard amino acids. We could show a clear improvement in accuracy over widely used quality assessment methods when considering alpha-helical transmembrane structures. It is possible to detect errors introduced in the modelling procedure such as incorrect alignments, which would facilitate the visual exploration of alternative alignments, e.g. as suggested previously in MODalign [11].

Despite similar observed overall performance for β-barrel structures, problems arise with shifted alignments, as they can occur when aligning sequences from remote homologues. The low number of pairwise atomic interactions in combination with the regular hydrophobicity pattern often observed in alignment shifts by two residues hamper the reliable detection of such errors and require further investigation.

2.5 ACKNOWLEDGMENTS

We would like to thank all members of the Torsten Schwede group at the University of Basel for helpful discussions. We are grateful to Pascal Benkert for providing the training model dataset. Special thanks go to Jürgen Haas, Lorenza Bordoli and Tjaart de Beer for valuable inputs and helping in setting up the manuscript and Alessandro Barbato for his help in performing the performance analysis using different quality assessment methods. The authors gratefully acknowledge the computational resources provided by the sciCORE/[BC]2 Basel Computational Biology Center at the University of Basel and the support by the system administration team.

Funding: This work was supported by SIB Swiss Institute of Bioinformatics. G.S. has been supported by a PhD fellowship of the SIB by the 'Swiss Foundation for Excellence and Talent in Biomedical Research'.

Conflict of interest: none declared.

2.6 supplemental materials

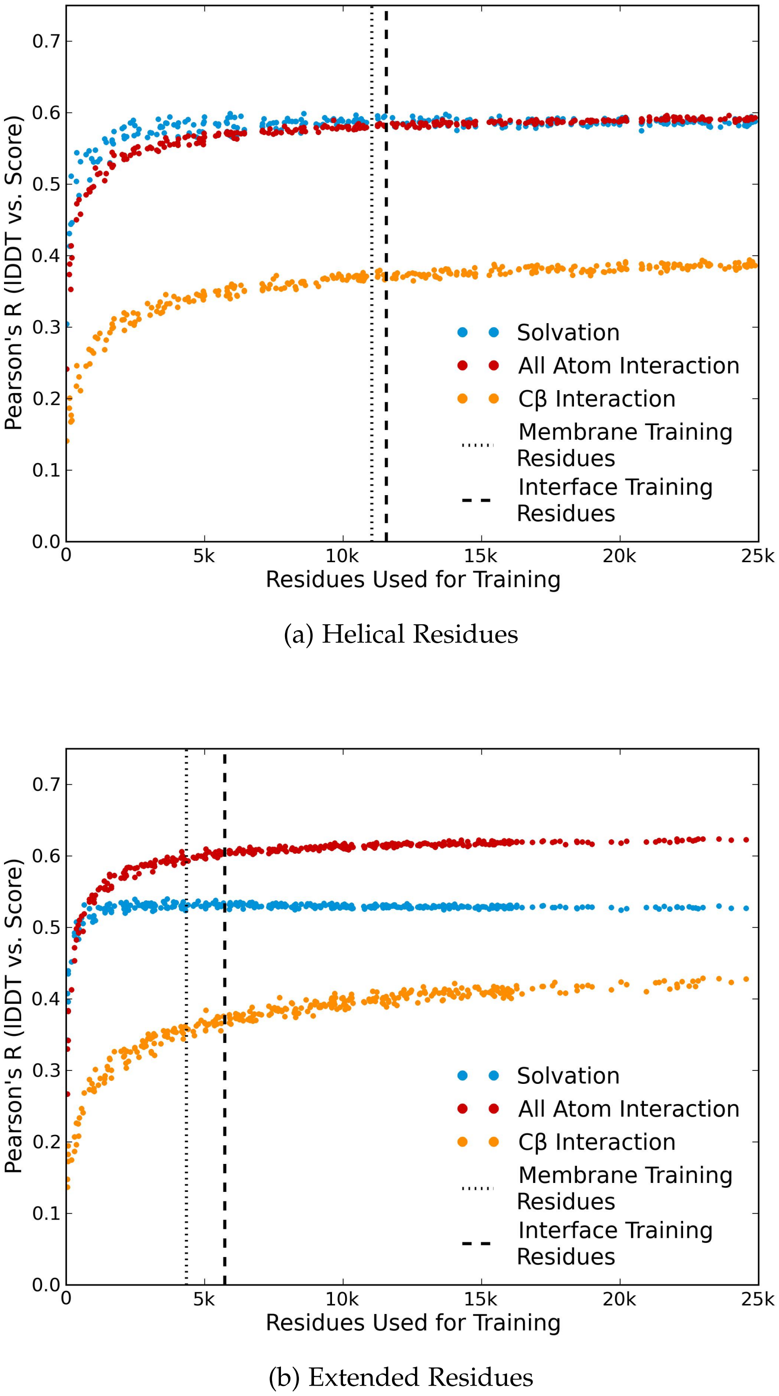
2.6.1 *Saturation of Membrane Specific Statistical Potentials*

The question at hand is, whether there is enough data to train membrane specific statistical potentials. This problem gets even more pressing upon separate treatment of membrane and interface part. Of all 3 affected statistical potential terms, a total of 360 secondary structure specific potentials have been trained with varying amount of soluble training data. Upon spherical smoothing, the local data of 2000 randomly picked structures of the soluble test set has been gathered for every single potential. To remove amino acid specific biases, one half of the data has been used to train the amino acid specific linear models on the one single feature of interest. The application of these linear models on the other half of the data then gives insights into the saturation behaviour and is further illustrated in Figure 8. In all cases the performances rapidly increase when more and more training data is provided, but the speed of asymptotic convergence is term specific. The solvation term seems to saturate fast, whereas the pairwise interaction terms need more data to be fully saturated.

2.6.2 *dRMSD Definition Used in This Work*

For some performance analysis, a local distance RMSD is used in the supplemental part of this work. In contrary to the classical RMSD approach, the dRMSD is superposition independent and represents the root mean square deviation of the difference in distance between all pairs of atoms, either on a per residue or full structure basis. To emphasize local behaviour, only distances below 10Å in the reference structure are considered in the calculation. In case of large distance differences, but also in case of missing distances, a cap value of 5Å for the difference in distance has been introduced.

$$\mathrm{dRMSD} = \sqrt{\frac{1}{N}\sum_{i,j}\left(\min(|d_{i,j,ref} - d_{i,j,model}|, cap)\right)^2} \quad (20)$$



(a) Helical Residues

(b) Extended Residues

Figure 8: Local performance of statistical potential terms when trained and applied on soluble structures and the amount of training data is varied. Every data point represents a single potential trained on a random subset of the soluble training structures.

2.6.3 *Parametrization of Statistical Potential Terms*

The pdfs of the statistical potential terms are based on histograms with following parametrization:

Soluble Potentials:

- **All Atom Interaction Term:** minimal distance: 0.0Å, maximal distance: 10.0Å, bin size: 0.5Å, sequence separation: 4 residues
- **Cβ Interaction Term:** minimal distance: 0.0Å, maximal distance: 12.0Å, bin size: 0.5Å, sequence separation: 4 residues
- **Solvation Term:** inclusion radius: 5.0Å, bin size: 1 count, max counts: 32
- **Torsion Term:** bin size: 20°

Membrane Potentials:

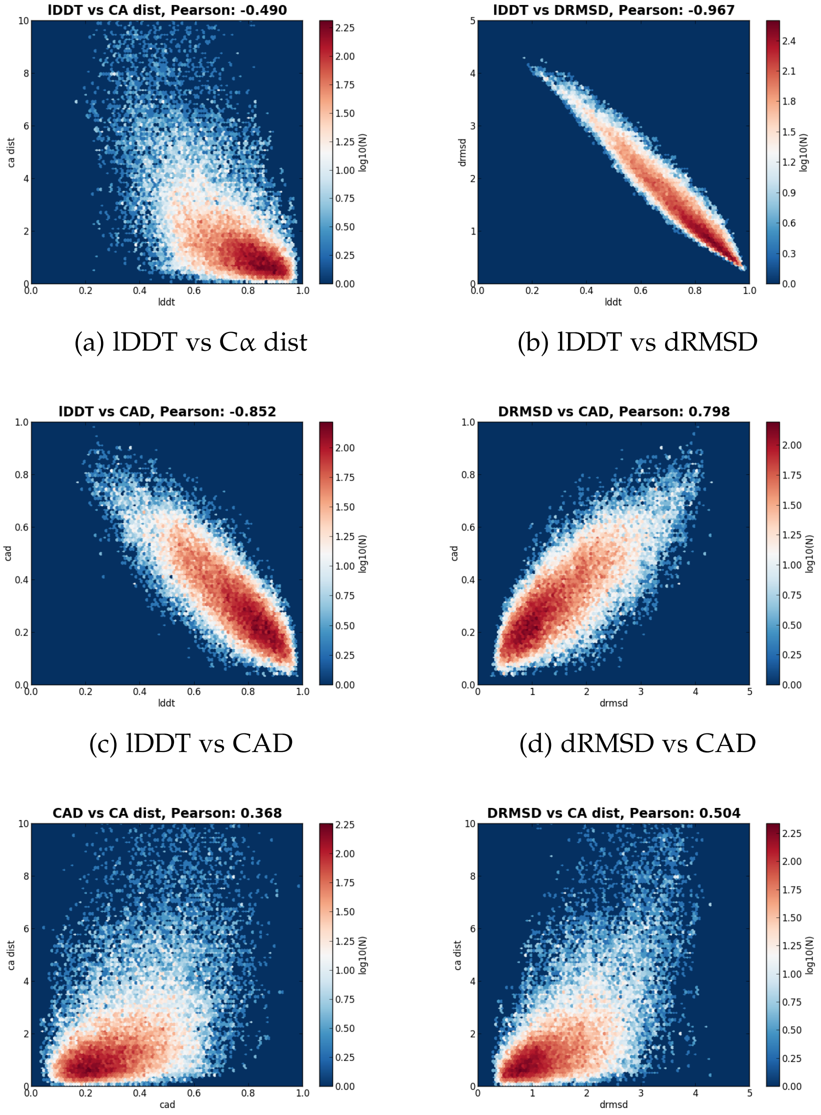
- **All Atom Interaction Term:** minimal distance: 0.0Å, maximal distance: 10.0Å, bin size: 1.0Å, sequence separation: 4 residues
- **Cβ Interaction Term:** minimal distance: 0.0Å, maximal distance: 12.0Å, bin size: 1.0Å, sequence separation: 4 residues
- **Solvation Term:** inclusion radius: 5.0Å, bin size: 1 count, max counts: 32
- **Torsion Term:** the torsion potential trained on soluble structures is applied on the transmembrane residues.

2.6.4 *Training data*

pdb id	chain	weight	pdb id	chain	weight	pdb id	chain	weight	pdb id	chain	weight	pdb id	chain	weight
4dx5	A	1.0	3b9y	A	1.0	1ymg	A	0.167	1kqf	C	1.0	2a06	G	0.25
3ar2	A	0.2	2r9r	B	1.0	3gd8	A	0.167	**4jq6**	A	1.0	4f4s	A	1.0
2agv	A	0.2	4k1c	A	1.0	3m9i	A	0.167	2fyu	E	0.2	1v55	I	1.0
3n5k	A	0.2	2j58	A	1.0	2j8d	H	0.2	1lol	E	0.2	2x2v	A	1.0
2zbd	A	0.2	**1u19**	A	1.0	1dxr	H	0.2	1pp9	E	0.2	3arc	h	1.0
1su4	A	0.2	**2z73**	A	1.0	1l9b	H	0.2	2a06	E	0.2	2yev	C	1.0
2zxe	A	1.0	4ezc	A	1.0	2j8c	H	0.2	3cx5	E	0.2	3arc	z	1.0
2yev	A	1.0	3arc	d	1.0	1eys	H	0.2	2nr9	A	0.2	1pp9	J	0.2
3ayf	A	1.0	3arc	a	1.0	1m56	B	1.0	2xow	A	0.2	2a06	W	0.2
4a01	A	1.0	3k3f	A	1.0	3kly	A	0.5	3zeb	A	0.2	2fyu	J	0.2
1jb0	A	0.5	3spc	A	0.5	3kcu	A	0.5	2xtv	A	0.2	1lol	J	0.2
1jb0	B	0.5	2qks	A	0.5	2qjy	B	1.0	2irv	A	0.2	3cx5	I	0.2
1wpg	A	1.0	1dxr	M	0.2	2bs2	C	1.0	4dve	A	1.0	1h2s	B	1.0
3s8g	A	1.0	1eys	M	0.2	1ldf	A	1.0	2qjy	C	1.0	1v55	J	1.0
1m56	A	0.5	2j8d	M	0.2	**3ddl**	A	1.0	3rlb	A	1.0	1lgh	A	1.0
1v55	A	0.5	2j8c	M	0.2	1z98	A	0.333	3s8g	B	1.0	2fyu	K	0.5
2a65	A	1.0	1l9b	M	0.2	3d9s	A	0.333	2bl2	A	1.0	1lol	K	0.5
2wsw	A	1.0	2yev	B	1.0	3cll	A	0.333	1jb0	L	1.0	1nkz	A	1.0
3arc	b	1.0	3m73	A	1.0	2vpz	C	1.0	2uuh	A	1.0	1v55	K	1.0
4iky	A	0.333	2nq2	A	1.0	4fc4	A	1.0	4alo	A	1.0	1hgz	A	1.0
4ikv	A	0.333	3zuy	A	1.0	3cx5	D	0.2	2j7a	I	1.0	1jb0	K	1.0
4ikx	A	0.333	**3odu**	A	1.0	2fyu	D	0.2	1v55	D	1.0	1v55	L	1.0
3puw	F	1.0	3v5u	A	1.0	1pp9	D	0.2	1jb0	F	1.0	1lgh	B	1.0
3arc	c	1.0	1okc	A	1.0	2a06	D	0.2	1yq3	C	0.5	1m56	D	1.0
4jkv	A	1.0	1kqf	B	1.0	1lol	D	0.2	1zoy	C	0.5	1v55	M	1.0
4mlb	C	0.5	3puw	G	1.0	2f2b	A	1.0	1ors	C	1.0	1jb0	J	1.0
3wbn	A	0.5	**3vw7**	A	1.0	3c02	A	1.0	2ahy	A	0.5	1nkz	B	1.0
2rh1	C	1.0	2j8d	L	0.2	**2ei4**	A	0.125	3ouf	A	0.5	3arc	x	1.0
3gia	A	1.0	2j8c	L	0.2	**1vgo**	A	0.125	1r3j	C	0.5	2zxe	G	1.0
2qjy	A	0.167	1l9b	L	0.2	**2zzl**	A	0.125	1s5h	C	0.5	3arc	i	1.0
3cx5	C	0.167	1eys	L	0.2	**1py6**	A	0.125	1zoy	D	0.5	3arc	J	1.0
2fyu	C	0.167	1dxr	L	0.2	**1h2s**	A	0.125	1yq3	D	0.5	1jb0	I	1.0
1lol	C	0.167	**3ug9**	A	1.0	**1mol**	A	0.125	3cx5	H	1.0	3arc	l	1.0
1pp9	C	0.167	2zxe	B	1.0	**1ap9**	A	0.125	3e86	A	1.0	3arc	k	1.0
2a06	C	0.167	1m56	C	0.5	**1h68**	A	0.125	3zk1	A	0.333	3arc	m	1.0
2qts	A	1.0	1v55	C	0.5	**4hyj**	A	1.0	1yce	L	0.333	3arc	f	1.0
3tij	A	1.0	2w2e	A	1.0	1rc2	B	0.5	2xqu	A	0.333	3s8g	C	1.0
2ns1	A	0.333	**3vvk**	A	0.333	3llq	A	0.5	1v55	G	1.0	1jb0	M	1.0
2b2f	A	0.333	**3a7k**	A	0.333	3tx3	A	1.0	3ldc	A	1.0	3arc	t	1.0
1u7g	A	0.333	**1e12**	A	0.333	1v55	B	1.0	3arc	e	1.0	3arc	y	1.0
3hd6	A	1.0	2b6p	A	0.167	1q16	C	1.0	1lol	G	0.25	1jb0	X	1.0
4kpp	A	1.0	1j4n	A	0.167	2bhw	A	1.0	2fyu	G	0.25	3bkd	E	0.5
4eiy	A	1.0	2b6o	A	0.167	**1xio**	A	1.0	1pp9	G	0.25	3lbw	A	0.5

Table 2: Structural data used to train alpha helix specific transmembrane statistical potentials. GPCR and Rhodopsin related structures are marked bold. The effect of removing them from the training data is shown in the GPCRDock performance evaluation in Figure 11

2.6.5 *Comparison of Different Measures of Similarity*



(a) lDDT vs Cα dist

(b) lDDT vs dRMSD

(c) lDDT vs CAD

(d) dRMSD vs CAD

(e) CAD vs Cα dist

(f) dRMSD vs Cα dist

Figure 9: The data of the membrane associated residues of the GPCR Dock experiments is used to show the differences in the measurements of local similarities. All atom based measures seem to agree well (b,c,d), whereas the Cα distance correlates poorly with them (a,e,f).

2.6.6 *Discrimination of Membrane and Soluble Structures*

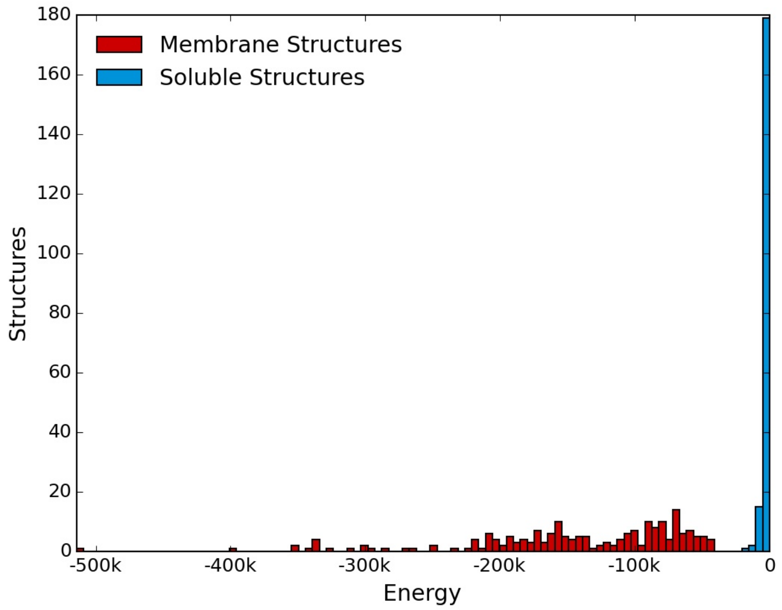
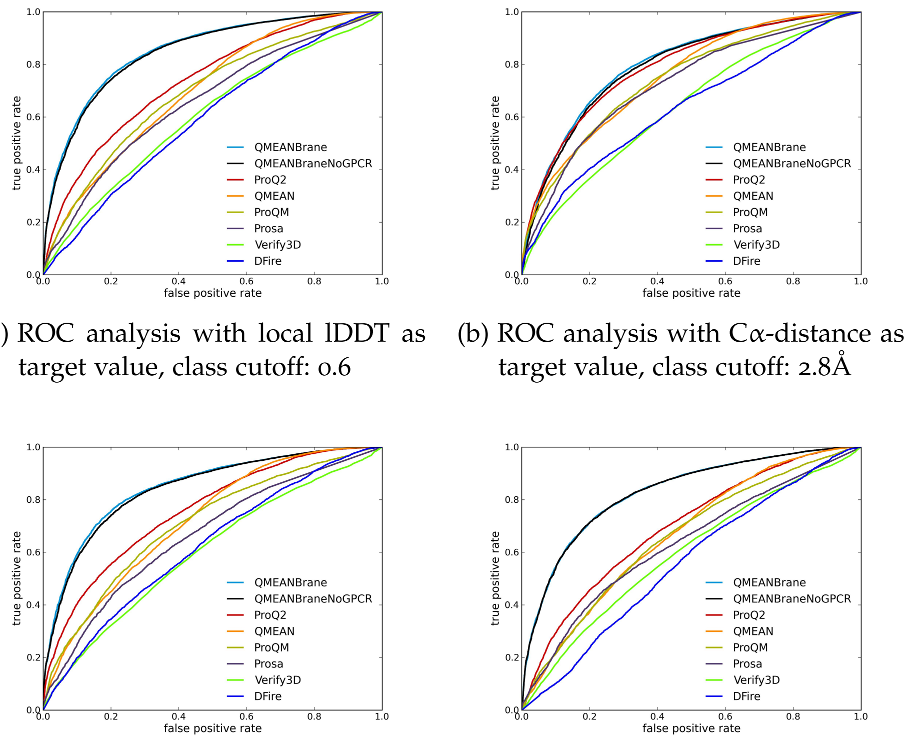


Figure 10: Calculated (pseudo) energies for 200 membrane protein structures used to generate membrane specific statistical potentials vs. the energies of 200 randomly selected soluble structures. A clear discrimination is possible.

2.6.7 *ROC Analysis on GPCRDock Test Set Using Different Measures of Similarity*

Method	AUC lDDT	AUC Cα	AUC dRMSD	AUC CAD
QMEANBrane	0.85	0.80	0.85	0.83
QMEANBraneNoGPCR	0.85	0.79	0.84	0.83
ProQ2	0.74	0.79	0.76	0.69
QMEAN	0.70	0.75	0.72	0.67
ProQM	0.69	0.74	0.71	0.66
Prosa	0.72	0.72	0.66	0.63
Verify3D	0.60	0.64	0.60	0.59
DFire	0.60	0.63	0.62	0.56

Table 3: Raw data considering all membrane associated residues as defined by OPM.



(a) ROC analysis with local lDDT as target value, class cutoff: 0.6

(b) ROC analysis with Cα-distance as target value, class cutoff: 2.8Å

(c) ROC analysis with local dRMSD as target value, class cutoff: 2.5

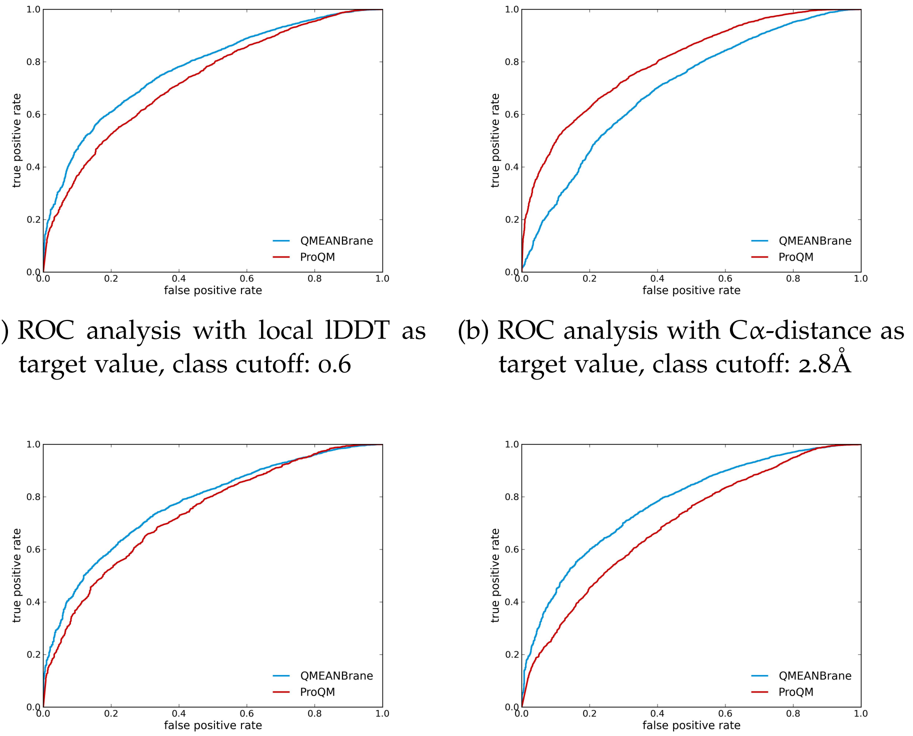
(d) ROC analysis with local CAD-score as target value, class cutoff: 0.5

Figure 11: ROC analysis considering all membrane associated residues as defined by OPM. The black curves indicate the effect of removing all GPCR/Rhodopsin related structures as defined in Table 1 from the training data.

2.6.8 *Head to Head comparison with ProQM on the testset used to test/train ProQM*

Method	AUC lDDT	AUC Cα	AUC dRMSD	AUC CAD
QMEANBrane	0.78	0.70	0.77	0.77
ProQM	0.73	0.80	0.74	0.70

Table 4: Areas under the curve for QMEANBrane and ProQM using different measures of similarity. Despite decrease in performance, QMEANBrane is superior regarding all atom measurements. In terms of Cα distances, ProQM clearly outperforms QMEANBrane.



(a) ROC analysis with local lDDT as target value, class cutoff: 0.6

(b) ROC analysis with Cα-distance as target value, class cutoff: 2.8Å

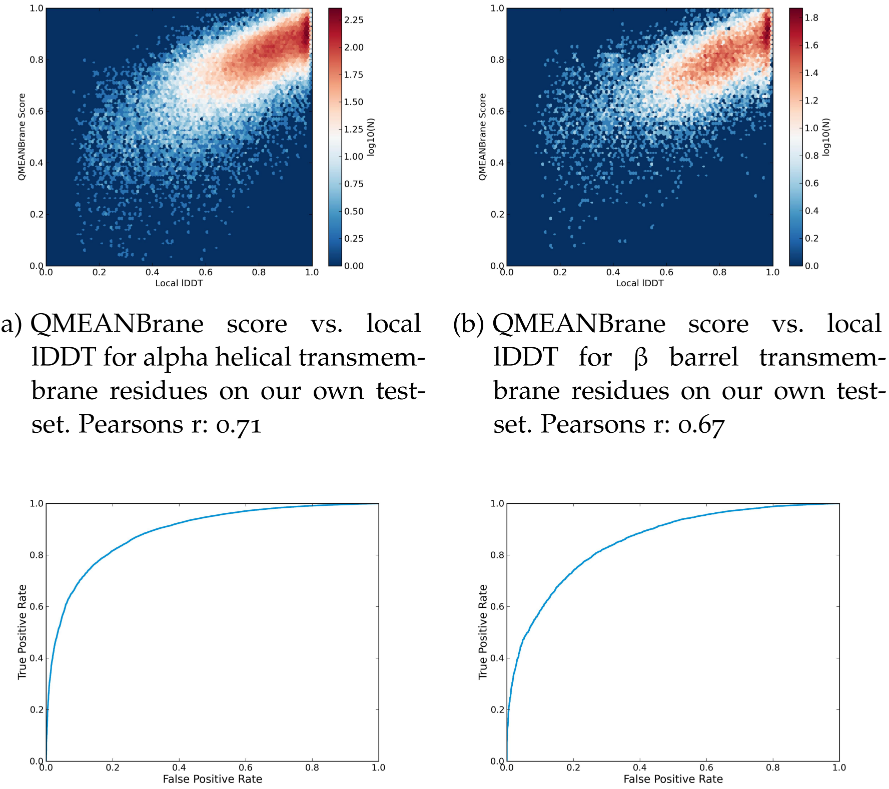
(c) ROC analysis with local dRMSD as target value, class cutoff: 2.5

(d) ROC analysis with local CAD-score as target value, class cutoff: 0.5

Figure 12: ROC analysis on membrane associated residues as defined by OPM of the testset used to test/train ProQM.

2.6.9 *Performance on Our Test Set*

Performance on our test set has been measured by a leave one out strategy. Upon clustering with a sequence identity threshold of 30%, the linear weights applied on the targets of one particular cluster are trained on all other clusters. Despite similar observed overall performance for β-barrel structures, problems arise with shifted alignments as they can occur when aligning sequences from remote homologues. The low level of pairwise interactions in combination with the regular hydrophobicity pattern often observed in alignment shifts by two residues hamper the reliable detection of such errors, and will require further investigations in the future.



(a) QMEANBrane score vs. local lDDT for alpha helical transmembrane residues on our own testset. Pearsons r: 0.71

(b) QMEANBrane score vs. local lDDT for β barrel transmembrane residues on our own testset. Pearsons r: 0.67

(c) ROC analysis on alpha helical transmembrane residues on our own testset with lDDT as target function and a class cutoff of 0.6. AUC: 0.89

(d) ROC analysis on β barrel transmembrane residues on our own testset with lDDT as target function and a class cutoff of 0.6. AUC: 0.85

Figure 13: Local performances of QMEANBrane on membrane associated residues as defined by OPM on our own testset.

2.6.10 *Retrospective Modelling Analysis*

2.6.10.1 *H^+ Translocating Pyrophosphatase*

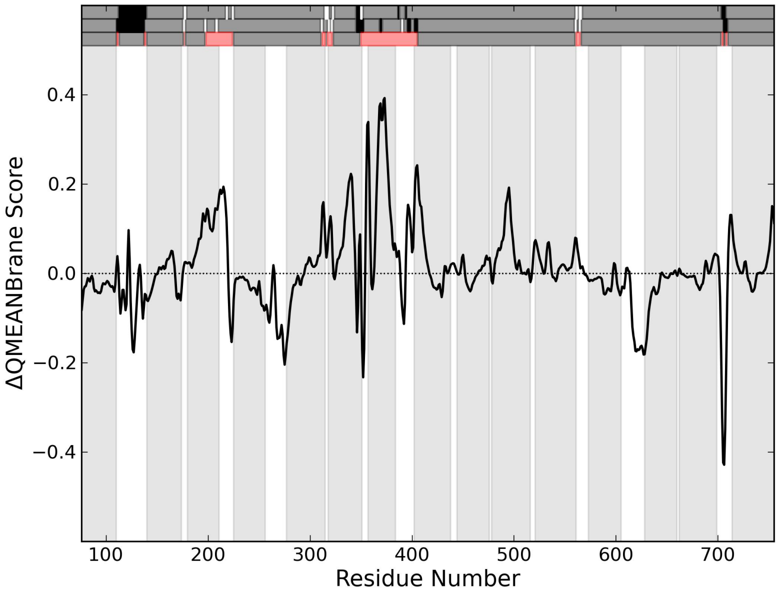


Figure 14: Comparison of QMEANBrane scores of a model built with the HHblits alignment vs. a model built with the BLAST alignment. The first horizontal bar represents the HHblits alignment with insertions marked in black and deletion marked in white, the second bar is the same for the BLAST alignment. The third bar highlights regions, where the two alignments differ.

```
Target    MGAAILPDLGTEILIPVCAVIGIAFALFQWLLVSKVKLSAVRDASPNAAAKNGYNDYLIE 60
4av3.1.A  ............................................................

EEEGINDHNVVVKCAEIQNAISEGATSFLFTEYKYVGIFMVAFAILIFLFLGSVEGFSTSPQACSYDKTK 130
...............EISSYIRSGADSFLAHETKAIFKVAIVIAILLMIF....................  35

TCKPALATAIFSTVSFLLGGVTSLVSGFLGMKIATYANARTTLEAR..KGVGKAFITAFRSGAVMGFLL. 197
.......TTWQTGVAFLLGAVMSASAGIVGMKMATRANVRVAEAARTTKKIGPALKVAYQGGSVMGLSVG  98

..AANGLLVLYIA..........INLFKIYYGDDWGGLFEAITGYGLGGSSMALFGRVGGGIYTKAADVG 255
GFALLGLVLVYLIFGKWMGQVDNLNIYTNWLGINFVPFAMTVSGYALGCSIIAMFDRVGGGVYTKAADMA 168

ADLVGKVERNIPEDDPRNPAVIADNVGDNVGDIAGMGSDLFGSYAESSCAALVVAS......ISSFGLN. 318
ADLVGKTELNLPEDDPRNPATIADNVGDNVGDVAGLGADLLESFVGAIVSSIILASYMFPIYVQKIGENL 238

.HEL......TAMLYPLIVSSVGILVCLLTTLFATDFFEIKAVKEIEPALKKQLVISTVLMTIGVAVVSF 381
VHQVPKETIQALISYPIFFALVGLGCSMLGILYV.......IVKKPSDNPQRELNIS..LWTSALLTVVL 299

VALPTSFTI..........FNFGVQKDVKSWQLFLCVAVGLWAGLIIGFVTEYYTSNAYSPVQDVADSCR 441
TAFLTYFYLKDLQGLDVLGFRFGA...ISPW...FSAIIGIFSGILIGFWAEYYTSYRYKPTQFLGKSSI 363

TGAATNVIFGLALGYKSVIIPIFAIAISIFVSFTFAAMYGIAVAALGMLSTIATGLAIDAYGPISDNAGG 511
EGTGMVISNGLSLGMKSVFPPTLTLVLGILFADYFAGLYGVAIAALGMLSFVATSVSVDSYGPIADNAGG 433

IAEMAGMSHRIRERTDALDAAGNTTAAIGKGFAIGSAALVSLALFGAFV.SRAS..........ITTVDV 570
ISEMCELDPEVRKITDHLDAVGNTTAAIGKGFAIGSAIFAALSLFASYMFSQISPSDIGKPPSLVLLLNM 503

LTPKVFIGLIVGAMLPYWFSAMTMKSVGSAALKMVEEVRRQFNTIPGLMEGTAKPDYATCVKISTDASIK 640
LDARVIAGALLGAAITYYFSGYLISAVTKAAMKMVDEIRRQAREIPGLLEGKAKPDYNRCIEITSDNALK 573

EMIPPGALVMLTPLVVGILFGVETLSGVLAGSLVSGVQIAISASNTGGAWDNAKKYIEAGASEHARSLGP 710
QMGYPAFIAILTPLVTGFLLGAEFVGGVLIGTVLSGAMLAILTANSGGAWDNAKKYLEAGNLEGY....G 639

KGSDCHKAAVIGDTIGDPLKDTSGPSLNILIKLMAVESLVFAPFFATHGGLLFKIF 766
KGSEPHKALVIGDTVGDPLKDTVGPSLDILIKIMSVVSVIAVSIF........... 684
```

Figure 15: BLAST alignment for H^+ Translocating Pyrophosphatase

```
Target    MGAAILPDLGTEILIPVCAVIGIAFALFQWLLVSKVKLSAVRDASPNAAAKNGYNDYLIE 60
4av3.1.A  ............ALFFLIPLVALGFAAANFAAVVRK.......................P 25

EEEGINDHNVVVKCAEIQNAISEGATSFLFTEYKYVGIFMVAFAILIFLFLGSVEGFSTSPQACSYDKTK 130
EG....T....ERMKEISSYIRSGADSFLAHETKAIFKVAIVIAILLMIFTT..................  69

TCKPALATAIFSTVSFLLGGVTSLVSGFLGMKIATYANARTTLEARK..GVGKAFITAFRSGAVMGFLLA 198
.........WQTGVAFLLGAVMSASAGIVGMKMATRANVRVAEAARTTKKIGPALKVAYQGGSVMGLSVG 130

ANGLLVLYIAINLFKIYYGD.....DWGGLF........EAITGYGLGGSSMALFGRVGGGIYTKAADVG 255
GFALLGLVLVYLIFGKWMGQVDNLNIYTNWLGINFVPFAMTVSGYALGCSIIAMFDRVGGGVYTKAADMA 200

ADLVGKVERNIPEDDPRNPAVIADNVGDNVGDIAGMGSDLFGSYAESSCAALVVASISSF......G.L. 317
ADLVGKTELNLPEDDPRNPATIADNVGDNVGDVAGLGADLLESFVGAIVSSIILASYMFPIYVQKIGENL 270

.....NHELT.AMLYPLIVSSVGILVCLLTTLFATDFFEI.KAVKEIEPALKKQLVISTVLMTIGVAVVS 380
VHQVPKETIQALISYPIFFALVGLGCSMLGILYV....IVKKPSDNPQRELNISLWTSALLTVVLTAFLT 336

FVALPTSFTIFNFGVQKDVKSWQLFLCVAVGLWAGLIIGFVTEYYTSNAYSPVQDVADSCRTGAATNVIF 450
YFYLKD.LQGLDVL.GFRFGAISPWFSAIIGIFSGILIGFWAEYYTSYRYKPTQFLGKSSIEGTGMVISN 404

GLALGYKSVIIPIFAIAISIFVSFTFAAMYGIAVAALGMLSTIATGLAIDAYGPISDNAGGIAEMAGMSH 520
GLSLGMKSVFPPTLTLVLGILFADYFAGLYGVAIAALGMLSFVATSVSVDSYGPIADNAGGISEMCELDP 474

RIRERTDALDAAGNTTAAIGKGFAIGSAALVSLALFGAFVS......RAS....I.TTVDVLTPKVFIGL 579
EVRKITDHLDAVGNTTAAIGKGFAIGSAIFAALSLFASYMFSQISPSDIGKPPSLVLLLNMLDARVIAGA 544

IVGAMLPYWFSAMTMKSVGSAALKMVEEVRRQFNTIPGLMEGTAKPDYATCVKISTDASIKEMIPPGALV 649
LLGAAITYYFSGYLISAVTKAAMKMVDEIRRQAREIPGLLEGKAKPDYNRCIEITSDNALKQMGYPAFIA 614

MLTPLVVGILFGVETLSGVLAGSLVSGVQIAISASNTGGAWDNAKKYIEAGASEHARSLGPKGSDCHKAA 719
ILTPLVTGFLLGAEFVGGVLIGTVLSGAMLAILTANSGGAWDNAKKYLEAGNLE....GYGKGSEPHKAL 680

VIGDTIGDPLKDTSGPSLNILIKLMAVESLVFAPFFATHGGLLFKIF 766
VIGDTVGDPLKDTVGPSLDILIKIMSVVSVIAVSIFKHVH....... 720
```

Figure 16: HHblits alignment for H^+ Translocating Pyrophosphatase

2.6.10.2 *Dopamine Transporter*

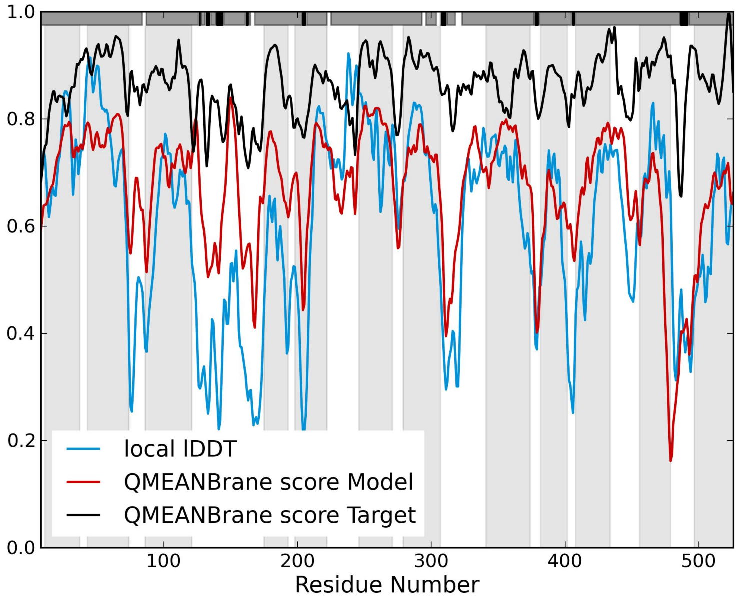
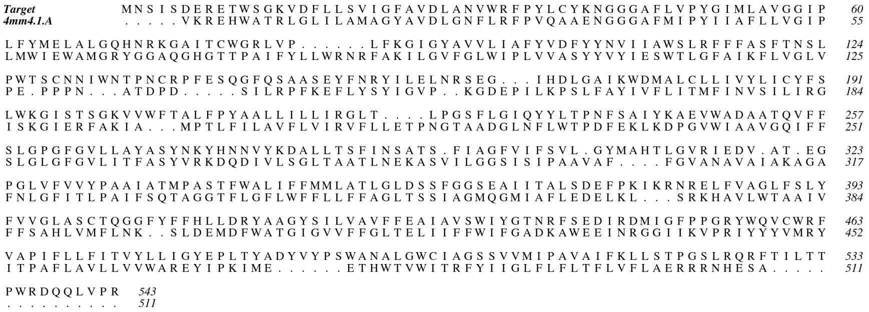


Figure 17: Local QMEANBrane scores of reference structure (black), as well as the scores from the model built with the initial HHblits alignment (red). The horizontal bar represents the alignment with insertions in black and deletions in white.



```
Target      MNSISDERETWSGKVDFLLSVIGFAVDLANVWRFPYLCYKNGGGAFLVPYGIMLAVGGIP   60
4mm4.1.A    .....VKREHWATRLGLILAMAGYAVDLGNFLRFPVQAAENGGGAFMIPYIIAFLLVGIP   55

LFYMELALGQHNRKGAITCWGRLVP......LFKGIGYAVVLIAFYVDFYYNVIIAWSLRFFFASFTNSL   124
LMWIEWAMGRYGGAQGHGTTPAIFYLLWRNRFAKILGVFGLWIPLVVASYYVYIESWTLGFAIKFLVGLV   125

PWTSCNNIWNTPNCRPFESQGFQSAASEYFNRYILELNRSEG...IHDLGAIKWDMALCLLIVYLICYFS   191
PE.PPPN...ATDPD.....SILRPFKEFLYSYIGVP..KGDEPILKPSLFAYIVFLITMFINVSILIRG   184

LWKGISTSGKVVWFTALFPYAALLILLIRGLT....LPGSFLGIQYYLTPNFSAIYKAEVWADAATQVFF   257
ISKGIERFAKIA...MPTLFILAVFLVIRVFLLETPNGTAADGLNFLWTPDFEKLKDPGVWIAAVGQIFF   251

SLGPGFGVLLAYASYNKYHNNVYKDALLTSFINSATS.FIAGFVIFSVL.GYMAHTLGVRIEDV.AT.EG   323
SLGLGFGVLITFASYVRKDQDIVLSGLTAATLNEKASVILGGSISIPAAVAF....FGVANAVAIAKAGA   317

PGLVFVVYPAAIATMPASTFWALIFFMMLATLGLDSSFGGSEAIITALSDEFPKIKRNRELFVAGLFSLY   393
FNLGFITLPAIFSQTAGGTFLGFLWFFLLFFAGLTSSIAGMQGMIAFLEDELKL...SRKHAVLWTAAIV   384

FVVGLASCTQGGFYFFHLLDRYAAGYSILVAVFFEAIAVSWIYGTNRFSEDIRDMIGFPPGRYWQVCWRF   463
FFSAHLVMFLNK..SLDEMDFWATGIGVVFFGLTELIIFFWIFGADKAWEEINRGGIIKVPRIYYYVMRY   452

VAPIFLLFITVYLLIGYEPLTYADYVYPSWANALGWCIAGSSVVMIPAVAIFKLLSTPGSLRQRFTILTT   533
ITPAFLAVLLVVWAREYIPKIME......ETHWTVWITRFYIIGLFLFLTFLVFLAERRRNHESA.....   511

PWRDQQLVPR   543
..........   511
```

Figure 18: Initial HHblits Alignment

```
Target     MNSISDERETWSGKVDFLLSVIGFAVDLANVWRFPYLCYKNGGGAFLVPYGIMLAVGGIP  60
4mm4.1.A   .....VKREHWATRLGLILAMAGYAVDLGNFLRFPVQAAENGGGAFMIPYIIAFLLVGIP  55

LFYMELALGQHNRKGAITCWGRLVP......LFKGIGYAVVLIAFYVDFYYNVIIAWSLRFFFASFTNSL  124
LMWIEWAMGRYGGAQGHGTTPAIFYLLWRNRFAKILGVFGLWIPLVVASYYVYIESWTLGFAIKFLVGLV  125

PWTSCNNIWNTPNCRPFESQGFQSAASEYFNRYILELNRSEG...IHDLGAIKWDMALCLLIVYLICYFS  191
PE.PPPN...ATDPD.....SILRPFKEFLYSYIGVP..KGDEPILKPSLFAYIVFLITMFINVSILIRG  184

LWKGISTSGKVVWFTALFPYAALLILLIRGLT.LPGSFLGIQYYLTPNFSAIYKAEVWADAATQVFFSLG  260
ISKGIERFAKIAMPTLFILAVFLVIRVFLLETPNGTAADGLNFLWTPDFEKLKDPGVWIAAVGQIFFSLG  254

PGFGVLLAYASYNKYHNNVYKDALLTSFINSATS.FIAGFVIFSVL.GYMAHTLGVRIEDV.AT.EGPGL  326
LGFGVLITFASYVRKDQDIVLSGLTAATLNEKASVILGGSISIPAAVAF....FGVANAVAIAKAGAFNL  320

VFVVYPAAIATMPASTFWALIFFMMLATLGLDSSFGGSEAIITALSDEFPKIKRNRELFVAGLFSLYFVV  396
GFITLPAIFSQTAGGTFLGFLWFFLLFFAGLTSSIAGMQGMIAFLEDELKL...SRKHAVLWTAAIVFFS  387

GLASCTQGGFYFFHLLDRYAAGYSILVAVFFEAIAVSWIYGTNRFSEDIRDMIGFPPGRYWQVCWRFVAP  466
AHLVMFLNK..SLDEMDFWATGIGVVFFGLTELIIFFWIFGADKAWEEINRGGIIKVPRIYYYVMRYITP  455

IFLLFITVYLLIGYEPLTYADYVYPSWANALGWCIAGSSVVMIPAVAIFKLLSTPGSLRQRFTILTTPWR  536
AFLAVLLVVWAREYIPKIME......ETHWTVWITRFYIIGLFLFLTFLVFLAERRRNHESA........  511

DQQLVPR  543
.......  511
```

Figure 19: Insertion Shifted Towards C-Terminus

```
Target     MNSISDERETWSGKVDFLLSVIGFAVDLANVWRFPYLCYKNGGGAFLVPYGIMLAVGGIP  60
4mm4.1.A   .....VKREHWATRLGLILAMAGYAVDLGNFLRFPVQAAENGGGAFMIPYIIAFLLVGIP  55

LFYMELALGQHNRKGAITCWGRLVP......LFKGIGYAVVLIAFYVDFYYNVIIAWSLRFFFASFTNSL  124
LMWIEWAMGRYGGAQGHGTTPAIFYLLWRNRFAKILGVFGLWIPLVVASYYVYIESWTLGFAIKFLVGLV  125

PWTSCNNIWNTPNCRPFESQGFQSAASEYFNRYILELNRSEG...IHDLGAIKWDMALCLLIVYLICY.F  190
PE.PPPN...ATDPD.....SILRPFKEFLYSYIGVP..KGDEPILKPSLFAYIVFLITMFINVSILIR.  183

.SLWKGISTSGKVVWFTALFPYAALLILLIRGLT....LPGSFLGIQYYLTPNFSAIYKAEVWADAATQV  255
G..I..SKGIERFAKIAMPTLFILAVFLVIRVFLLETPNGTAADGLNFLWTPDFEKLKDPGVWIAAVGQI  249

FFSLGPGFGVLLAYASYNKYHNNVYKDALLTSFINSATS.FIAGFVIFSVL.GYMAHTLGVRIEDV.AT.  321
FFSLGLGFGVLITFASYVRKDQDIVLSGLTAATLNEKASVILGGSISIPAAVAF....FGVANAVAIAKA  315

EGPGLVFVVYPAAIATMPASTFWALIFFMMLATLGLDSSFGGSEAIITALSDEFPKIKRNRELFVAGLFS  391
GAFNLGFITLPAIFSQTAGGTFLGFLWFFLLFFAGLTSSIAGMQGMIAFLEDELKL...SRKHAVLWTAA  382

LYFVVGLASCTQGGFYFFHLLDRYAAGYSILVAVFFEAIAVSWIYGTNRFSEDIRDMIGFPPGRYWQVCW  461
IVFFSAHLVMFLNK..SLDEMDFWATGIGVVFFGLTELIIFFWIFGADKAWEEINRGGIIKVPRIYYYVM  450

RFVAPIFLLFITVYLLIGYEPLTYADYVYPSWANALGWCIAGSSVVMIPAVAIFKLLSTPGSLRQRFTIL  531
RYITPAFLAVLLVVWAREYIPKIME......ETHWTVWITRFYIIGLFLFLTFLVFLAERRRNHESA...  511

TTPWRDQQLVPR  543
............  511
```

Figure 20: Insertion Shifted in Between Helix 4 and 5

```
Target     MNSISDERETWSGKVDFLLSVIGFAVDLANVWRFPYLCYKNGGGAFLVPYGIMLAVGGIP  60
4mm4.1.A   .....VKREHWATRLGLILAMAGYAVDLGNFLRFPVQAAENGGGAFMIPYIIAFLLVGIP  55

LFYMELALGQHNRKGAITCWGRLVP......LFKGIGYAVVLIAFYVDFYYNVIIAWSLRFFFASFTNSL  124
LMWIEWAMGRYGGAQGHGTTPAIFYLLWRNRFAKILGVFGLWIPLVVASYYVYIESWTLGFAIKFLVGLV  125

PWTSCNNIWNTPNCRPFESQGFQSAASEYFNRYILELNRSEG...IHDLGAIKWDMALCLLIVYLICYFS  191
PE.PPPN...ATDPD.....SILRPFKEFLYSYIGVP..KGDEPILKPS....LFAYIVFLITMFINVSI  180

LWKGIS.TSGKVVWFTALFPYAALLILLIRGLT....LPGSFLGIQYYLTPNFSAIYKAEVWADAATQVF  256
LIRGISKGIERFAKIAMPTLFILAVFLVIRVFLLETPNGTAADGLNFLWTPDFEKLKDPGVWIAAVGQIF  250

FSLGPGFGVLLAYASYNKYHNNVYKDALLTSFINSATS.FIAGFVIFSVL.GYMAHTLGVRIEDV.AT.E  322
FSLGLGFGVLITFASYVRKDQDIVLSGLTAATLNEKASVILGGSISIPAAVAF....FGVANAVAIAKAG  316

GPGLVFVVYPAAIATMPASTFWALIFFMMLATLGLDSSFGGSEAIITALSDEFPKIKRNRELFVAGLFSL  392
AFNLGFITLPAIFSQTAGGTFLGFLWFFLLFFAGLTSSIAGMQGMIAFLEDELKL...SRKHAVLWTAAI  383

YFVVGLASCTQGGFYFFHLLDRYAAGYSILVAVFFEAIAVSWIYGTNRFSEDIRDMIGFPPGRYWQVCWR  462
VFFSAHLVMFLNK..SLDEMDFWATGIGVVFFGLTELIIFFWIFGADKAWEEINRGGIIKVPRIYYYVMR  451

FVAPIFLLFITVYLLIGYEPLTYADYVYPSWANALGWCIAGSSVVMIPAVAIFKLLSTPGSLRQRFTILT  532
YITPAFLAVLLVVWAREYIPKIME......ETHWTVWITRFYIIGLFLFLTFLVFLAERRRNHESA....  511

TPWRDQQLVPR  543
...........  511
```

Figure 21: Insertion Shifted Towards N-Terminus in Fromt of Helix 4

3 QMEANDISCO

This chapter has been a collaborative effort between Christine Rempfer, Gabriel Studer, Andrew Waterhouse and Rafal Gumienny. Title of the Manuscript:

QMEANDisCo – Distance Constraints Applied on the Local Quality Estimation Problem

Author Contributions: CR implemented DisCo, compiled the SWISS-MODEL based test-/training sets, trained the score combination with QMEAN and helped setting up the manuscript. GS implemented QMEAN, retrained all potentials, gathered the CAMEO and CASP training data, performed the evaluation and wrote the manuscript. AW, GS, RG and CR implemented the webserver (initials ordered by amount of contributions).

Motivation: Quality estimation methods are an indispensable ingredient in any modelling pipeline. Global quality estimates give a general impression of a model's applicability or allow selecting a model in a set of alternatives. Local quality estimates on the other hand assess the reliability of individual amino acids, opening a full range of possible applications. We therefore aim to extend the local quality estimation capabilities of QMEAN by harnessing ensemble information in form of distance constraints extracted from the rapidly increasing amount of experimentally determined structural information.
Results: We improved the established quality estimation tool QMEAN and enhanced its local quality estimation capabilities with a new term based on distance constraints - DisCo. QMEAN and QMEANDisCo have been successfully tested and compared to other state of the art local quality estimation tools on a wide variety of test sets. Careful data analysis revealed that both methods particularly stand out in distinguishing wrongly from correctly modelled residues in models of reasonable overall fold.
Availability: https://swissmodel.expasy.org/qmean/

3.1 INTRODUCTION

Modelling methods, in particular homology/comparative modelling, have established themselves as a valuable complement for structural analysis when experimental data are missing [140]. While such methods have matured into pipelines that can generate models for almost any protein automatically, the quality of the generated models can be highly variable and hard to predict in the absence of experimental observables. This is a major concern as the range of applications for which the model can be used directly depends on its quality [9, 141], hence the importance of quality estimation methods. Quality estimates can be of a global nature, i.e. to pick the best model in a set of alternatives, or of a local nature. The latter allows for a more specific model selection in cases where only one particular part of the protein is of interest, for example a domain containing an active site. It can also guide the modelling process itself, such as detecting regions requiring further refinement or choosing from alternative local conformations.

Currently, two very distinct approaches exist to successfully tackle the quality estimation problem: single model methods and consensus methods. Most single model methods use knowledge based approaches such as statistical potentials of mean force, to express the expected similarity to the actual native structure with a numeric value [13, 132, 146, 170]. Such methods have the advantage of only requiring a single model as input, but they tend to be outperformed by consensus methods that base their prediction on a full ensemble of models [89–91]. Model quality is estimated from the variability of the models in the ensemble, assuming that correct structural features will tend to be more conserved [53, 110, 147]. Nevertheless, their application is somewhat limited as a set of models is not always available in many applied cases. This led to the development of the so called quasi-single model methods [110], which try to combine the predictive power of consensus with the convenience of taking a single model as input by using alternative sources of ensemble information.

One established classical single model method is QMEAN [13, 14]. QMEAN uses statistical potentials of mean force and the consistency of a model with structural features predicted from sequence to generate quality estimates on a global and local scale. A specialized version "QMEANBrane" was also de-

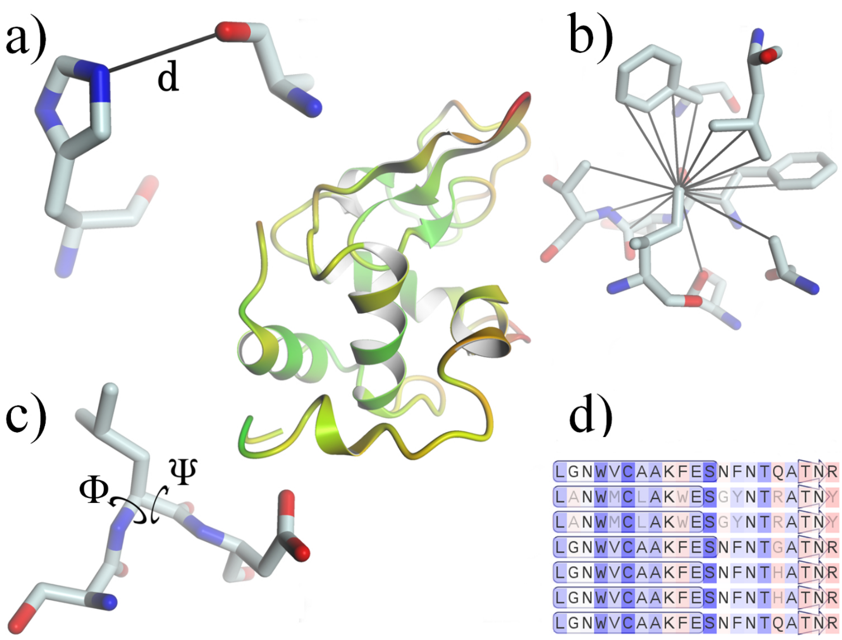


Figure 22: The central model visualizes the goal of assigning per residue quality estimates with a colour gradient. QMEAN approaches the problem with a combination of several terms: a) pairwise terms with one term considering all atoms and another term only considering Cβ positions, b) solvation term, c) backbone torsion term, d) terms assessing the consistency of the model with profile based predictions (secondary structure and solvent accessibility).

veloped, employing statistical potentials specifically trained to assess the local quality of membrane protein models [152].

In this work we improve QMEANs capabilities for providing local quality estimates (Figure 22). We introduce a new distance constraint (DisCo) score that assesses the agreement between observed pairwise distances in a model with an ensemble of constraints extracted from experimentally determined structures with sequences homologous to the model being assessed. Adding this score to QMEAN (QMEANDisCo) therefore leads to a quasi-single model method. Using the homologous structures directly when generating the set of constraints, allows us to keep the computation time low by avoiding a full blown model building process to generate a structural ensemble. We show that adding DisCo to QMEAN significantly increases the reliability in estimating local qualities on a wide variety of test sets.

3.2 MATERIALS & METHODS

In order to approach the local quality estimation problem, we first defined an appropriate target value that expresses the similarity of a model to the native structure on a local scale. We carefully compiled a set of models to train the newly developed methods towards this target value and test it towards models from various sources. The evaluation has then been targeted at the discrimination of correct from poorly modelled residues in models of reasonable overall fold.

3.2.1 *Target Value*

As a target value for local quality estimates, we use the all atom based lDDT score [109]. lDDT is a superposition free score and assesses the differences in pairwise interatomic distances between model and reference structure. Only pairwise distances up to a certain cutoff are considered, reducing the influence of domain or hinge movement events. The authors of lDDT recommend a 15Å cutoff for global full model scores. For per residue scores, this cutoff has been decreased to 10Å to emphasize local behaviour. To avoid overtraining towards a certain target value and allow for a general interpretation of the predicted local score values we repeat all evaluations of local quality estimation performance with CAD score [121] and dRMSD (Section 3.6.1). All three scores evaluate models on an all atom basis with lDDT and dRMSD additionally considering stereochemical issues and clashes. We deliberately avoid any local target value based on reduced structural representations since they do not reflect the wide variety of local interactions in great detail. One representative of this category would be the widely used Cα distances between residues in model and target after a global superposition. Section 3.6.4 gives further details about potential issues with this target value.

3.2.2 *Training and Test set*

Training and testing have been performed on a large set of models generated by the SWISS-MODEL modelling webserver [22]. Further testing and evaluation have been performed on independent test sets provided by the CAMEO continuous evaluation platform [60] and the CASP XI experiment [91].

For our own training-/test set generated with SWISS-MODEL we rely on a set of non-redundant entries from the PDB as culled by PISCES [157] to serve as targets for the model building process (sequence identity cutoff: 20%, resolution cutoff: 1.8Å, R-factor cutoff: 0.25). The returned list contained 5302 entries. On one hand we used the full list to train the statistical potential of mean force terms, on the other hand we used a randomly selected subset of 2500 items as targets to generate a large set of models for training and testing purposes. For every modelled target, a maximum of 10 models have been built by randomly selecting templates with sequence identity below 90% to the target and an alignment coverage of at least 50%. The number of models per target has further been reduced, such that no pair of models has a sequence identity above 90% considering their underlying templates. To get the desired sets of models a further split was necessary.

- Models from 625 targets (2456 models) to train linear score combinations for the local and global QMEAN scoring functions.
- Models from 1250 targets (4886 models) to train random forest regressor to combine the local QMEAN scoring function with DisCo.
- Models from 625 targets (2471 models) for testing purposes, which we will refer to as the SWISS-MODEL test set.

A CASP XI test set has been compiled by downloading publicly available models submitted for the QA2 Model 2 category [91], resulting in a total of 13,077 models.

A CAMEO test set has been compiled by downloading all the QE predictions from CAMEO in a timeframe of 3 months (2016.12.24 – 2017.03.18), resulting in a total of 2289 models.

One issue of the training / test set compilation is that the statistical potential of mean force terms are trained and tested on the same structural data. Due to the saturation behaviour of the underlying probability density functions, this is not a problem [152].

3.2.3 *QMEAN*

QMEAN is a combination of four statistical potential of mean force terms, as well as two terms comparing secondary structure-

and solvent accessibility predictions by PSIPRED [75] and ACCPRO [31] with their actual outcome in the model (Figure 22).

Compared to the 2011 version of QMEAN, which we refer to as QMEAN_OLD, QMEANBrane already introduced improvements in the statistical potential terms that are further described in Section 2.2.3. QMEANBrane employs specifically trained potentials for three different segments in a transmembrane protein model with segments being defined as membrane, interface and soluble. The potentials applied on the soluble segment are now also in use for QMEAN and have been retrained for the use as local and global quality predictors with the exact parameterization described in Section 3.6.2.

No changes have been made regarding the predicted solvent accessibility term, it is still a binary classification whether prediction and model match on a per residue basis. However, the predicted secondary structure term has been improved. Instead of simply checking for a match between prediction and outcome, QMEAN now incorporates all available information from DSSP as well as PSIPRED. A log-odds score relates the probability of observing a certain DSSP state in combination with a PSIPRED prediction with the probability of observing the two events independently of each other [149] (Equation 21). The required probabilities have been extracted from the same structural information already used to train the statistical potential terms.

$$S(d,p,c) = \log\left(\frac{p(d,p,c)}{p(d)p(p,c)}\right) \tag{21}$$

with d representing a DSSP state in [G,H,I,E,B,T,S,C], p a PSIPRED state in [H,E,C] and c a PSIPRED confidence value in [0-9].

Despite having distinct statistical potentials optimized for local and global scoring, all quality predictors are evaluated on a per residue basis. The results are further processed to obtain per residue or global scores.

3.2.3.1 *Local Quality Estimates*

Having calculated all the scores on a per residue basis, a spherical smoothing is applied to reduce noise [152]. A subsequent amino acid dependent linear combination of the scores gives per residue quality estimates [152] with linear weights trained on the specified SWISS-MODEL training set.

3.2.3.2 *Global Quality Estimates*

Despite small changes in the single quality predictors, there is no conceptual difference to QMEAN_OLD. For every quality predictor, the per residue scores are averaged in order to normalize for size. A linear combination of all four statistical potential terms gives the QMEAN4 score and additionally using the two predicted sequence feature terms gives the QMEAN6 score with linear weights trained on the specified SWISS-MODEL training set.

To relate these scores to what one would expect from high resolution X-ray structures, QMEAN still provides them as Z-scores given the corresponding score distribution from high resolution X-ray structures of similar size.

3.2.4 *DisCo*

DisCo is the successor of QMEANDist [19], a quasi-single model method that participated in the CASP IX experiment as a global quality predictor [89]. We revisited the approach of assessing the agreement of pairwise residue-residue distances with ensembles of distance constraints extracted from structures homologous to the assessed model. Instead of generating global quality estimates, DisCo aims to predict local per residue quality estimates. After extracting the target sequence of the model to be assessed, structural homologues are identified using HHblits [134]. For each homologue k, all Cα positions are extracted and mapped onto the target sequence using the HHblits alignment. Gaussian-like distance constraints $g_{ijk}(d_{ij})$ are constructed for all observed pairwise Cα-Cα distances μ_{ijk} below 15Å:

$$g_{ijk}(d_{ij}) = exp\left[-\frac{1}{2}(d_{ij} - \mu_{ijk})^2\right] \tag{22}$$

The goal is to construct a pairwise scoring function $s_{ij}(d_{ij})$, that assesses the consistency of a particular pairwise Cα-Cα distance d_{ij} from the model with all corresponding constraints $g_{ijk}(d_{ij})$. For that matter it has to be considered that HHblits may return redundant results. In order to avoid biases introduced by over-represented protein families, all found homologues are hierarchically clustered based on their normalized pairwise sequence similarity as estimated with the BLOSUM62

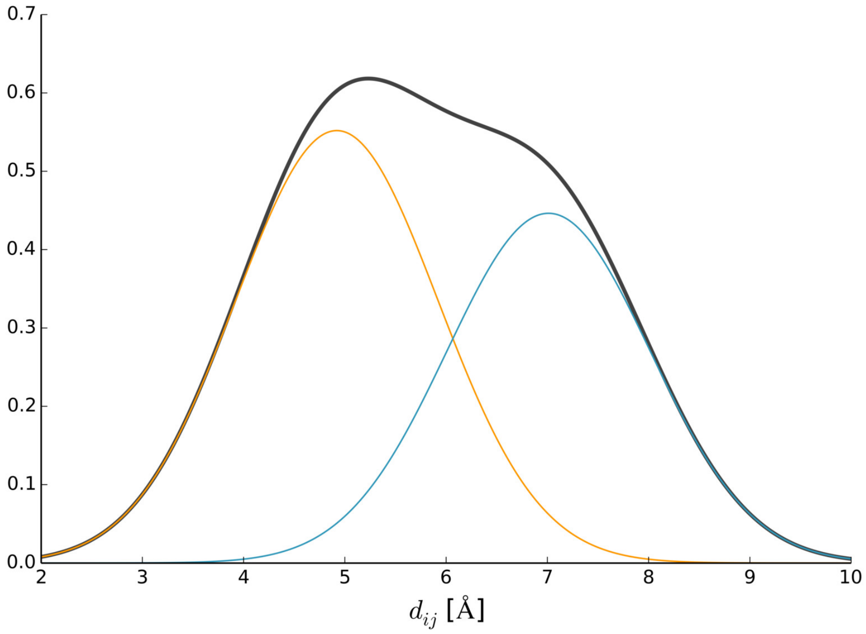


Figure 23: Example DisCo scoring function representing the Cα-Cα distance between residues i and j. The underlying templates appear in two clusters with avg. sequence similarity (SS_c) 0.276 (orange cluster) and 0.273 (blue cluster). Shown are the cluster scoring functions scaled by the cluster dependent weights $w_c h_{ijc}(d_{ij})$ (coloured lines) and the resulting DisCo scoring function $s_{ij}(d_{ij})$ (black line).

substitution matrix. For every cluster c, a cluster specific scoring function $h_{ijc}(d_{ij})$ is constructed:

$$h_{ijc}(d_{ij}) = \frac{1}{n_{ijc}} \sum_{k \in c} g_{ijk}(d_{ij}) \tag{23}$$

with n_{ijc} being the number of the underlying constraints $g_{ijk}(d_{ij})$ in that particular cluster. Note, that not every template k must necessarily contribute a constraint for a particular pair i, j. n_{ijc} can therefore potentially be zero, the full cluster is omitted in this case. To get our desired function $s_{ij}(d_{ij})$ we combine $h_{ijc}(d_{ij})$ from each cluster c in a weighted manner, such that clusters expected to be closely related to the target sequence contribute more than others:

$$s_{ij}(d_{ij}) = \sum_{c} w_c h_{ijc}(d_{ij}) \tag{24}$$

with weights w_c defined as $exp[\gamma SS_c]$ and normalized, so that the weights of all clusters in which the Cα-Cα pair is present, sum up to one. SS_c is the average normalized sequence similarity towards the target sequence of cluster c and γ is considered to be a constant that controls how fast the influence of a cluster vanishes as a function of SS_c . Best performance was observed when taking a value of 70 for γ and Figure 23 illustrates an example function. The DisCo score of a single residue of the model at position i then gets determined by averaging the outcome of all n pairwise scoring functions $s_{ij}(d_{ij})$ towards other residues j with their $C\alpha$ positions within 15Å:

$$DisCo_i = \frac{1}{n} \sum_j s_{ij}(d_i j) \tag{25}$$

3.2.5 *Score Combination - QMEANDisCo*

A simple linear model to combine QMEAN and DisCo would not faithfully reflect DisCo's dependency on the situation of found homologues. In case of many homologues with high sequence similarity to the target, the DisCo score is likely to be reliable and should have a large contribution to QMEANDisCo. However if there are only remote homologues available, QMEANDisCo should more closely, completely in the case of no homologues, rely on the statistical potentials of QMEAN. To handle these dependencies, a random forest regressor has been trained on the specified training data [126]. Besides QMEAN and DisCo, the random forest takes the following features as input to estimate the QMEANDisCo score for one particular residue i:

- Average number of cluster scoring functions towards all residues j within 15 Å
- Average of highest sequence similarities among the clusters towards all residues j within 15 Å
- Average of highest sequence identities among the clusters towards all residues j within 15 Å
- Average variance towards all residues j within 15 Å, where one variance is calculated using all observed μ_{ijk}
- Number of other residues within 15Å in the model. Only residues being covered together with residue i in at least one template are considered.
- global QMEAN4 score

3.2.6 *Evaluation Methods*

We define the local quality estimation problem we want to tackle as the ability to discriminate poorly from well modelled residues assuming an overall correct fold. For evaluation we use a receiver operation characteristic (ROC) analysis, which is common in the field [90, 91]. ROC allows one to visualize a predictors' capability of distinguishing positively from negatively classified data points and quantify the outcome with the area under the curve (AUC). Based on local lDDT as the target value, the data points of all single residues in a test set are divided into positives/negatives using a cutoff of 0.6, considering residues with local lDDT below that cutoff as positives. This reflects our definition of the quality assessment problem. In a first step we perform the ROC analysis on all described test sets by pooling all single residue predictions, which mainly analyses the capability of assigning absolute quality estimates. In a second step we perform the ROC analysis on a per model basis. To quantify the outcome, we generate a probability density function from all resulting per model AUC values of a test set using a Gaussian kernel density estimate. The expectation value of this distribution, the expected AUC when looking at one particular model, then allows a direct comparison to other predictors. For the alternative evaluations available in the supplemental materials we use a cutoff of 0.5 for CAD score and a cutoff of 2.0Å for dRMSD to classify the data points.

3.3 RESULTS AND DISCUSSION

The SWISS-MODEL test set is mainly intended to directly compare the performances of QMEAN [13] (QMEAN_OLD) to the current improved version of QMEAN and finally QMEANDisCo. To allow a comparison of the latter two to other publicly available local quality estimation tools, this analysis has been extended towards the test sets from CAMEO and CASP XI. In case of CAMEO, QMEAN and QMEANDisCo have been registered as participating servers and all predictions in the specified time range have been downloaded from the official website. A blind prediction is therefore guaranteed. In the case of CASP XI, the predictions of the CASP XI participants have been downloaded from the official data archive and complemented by QMEAN and QMEANDisCo calculated locally. No structural information published after April 2014 has been used to obtain DisCo.

3.3.1 *Overall AUC Analysis*

Applying QMEAN_OLD and QMEAN on the SWISS-MODEL test set gives significant changes of overall AUC in favor of the current version (0.80 vs 0.87). Incorporating DisCo (QMEANDisCo) increases the performance even further to 0.93 (Figure 24, Table 5). When comparing QMEAN to other methods in the CAMEO test set, QMEAN reaches an overall AUC of 0.88. The performance compared to ProQ2 [156] (0.87) and the quasi-single model methods ModFold4/ModFold6 [110] (0.89 / 0.89) differs only marginally. QMEANDisCo on the other hand clearly outperforms all other methods participating in CAMEO with an overall AUC of 0.93 (Figure 24, Table 6). While the overall AUC for QMEAN and QMEANDisCo slightly decreases when analysing the CASP XI test set (0.84 and 0.88), the previ-

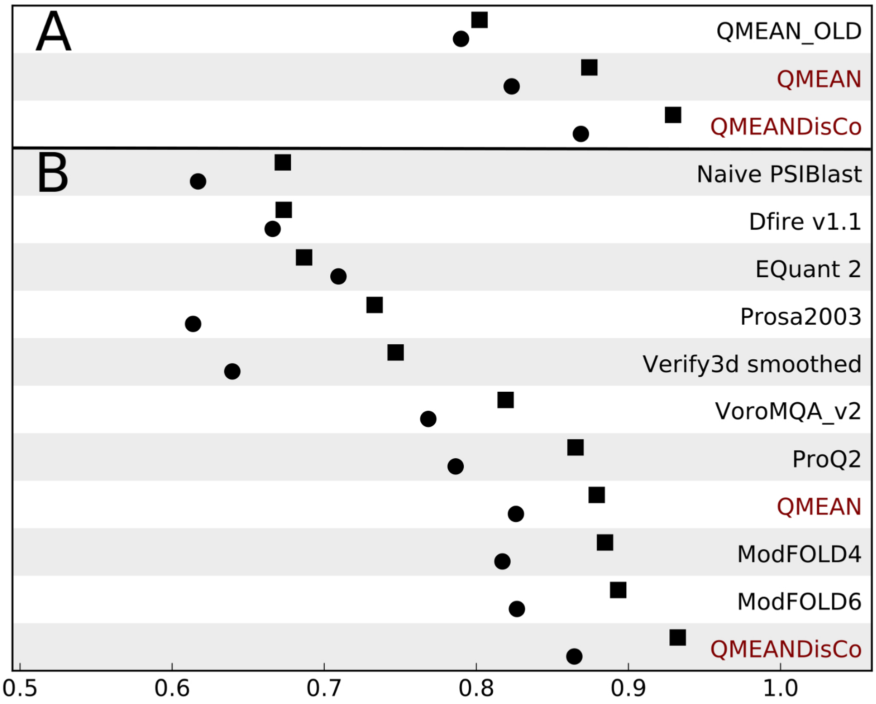


Figure 24: Evaluation of local quality estimation performance on SWISS- MODEL test set (A) and CAMEO test set (B) with local lDDT as target value. Squares represent performance in terms of overall AUC and circles the expected per model AUC. In the SWISS-MODEL test set, there is a steady increase of prediction accuracy from QMEAN_OLD, the current QMEAN and QMEANDisCo. QMEANDisCo performs best among the participants of the CAMEO continuous evaluation platform.

ously mentioned methods ProQ2 and ModFold4 (represented by its successor ModFold5_single) have an increased overall AUC value (0.86 and 0.88). The best consensus based predictors even reach overall AUC values up to 0.90 (Figure 25, Table 7).

3.3.2 *Global Effect on Overall AUC*

Because of the large number of quality estimation tools that can directly be compared, the most valuable test set used in this work clearly is the one from CASP XI. A possible problem is the substantially different single residue target value distribution compared to the other two test sets used in this work (Figure 26). The SWISS-MODEL and CAMEO test sets exhibit a unimodal distribution largely originating from models with reasonable overall fold. In contrast to that, the distribution of the CASP XI test set is bimodal. The lower quality distribution originates from a large number of random coil models that do not match with our definition of the local quality estimation problem. This gives rise to the hypothesis that most of the local quality estimation performance could already be retrieved by detecting those random coils and predicting all their residues to be of low quality. To test this hypothesis, a naive predictor has been implemented. It is based on the Davis-QAconsensus baseline predictor from the official CASP XI QE evaluation. For all residues of a particular model, the global score of the full model was assigned. Detecting random coils and scoring their residues accordingly is not necessarily a bad idea but this implementation has the obvious flaw of not being able to discriminate correctly from wrongly modelled residues in one particular model. Nevertheless, the naive predictor performs surprisingly well with an overall AUC value of 0.83 (Figure 25, Table 7). This observation clearly highlights that a good performance in terms of overall AUC is not necessarily the result of assigning meaningful per residue scores but rather a global effect. This makes a complementary per model analysis necessary and informative.

3.3.3 *Per Model AUC*

Since a ROC analysis requires data points classified as positives and negatives, particularly good or particularly bad models cannot be assessed in a per model analysis. In the case of local lDDT as target value this leaves 2421 models for the

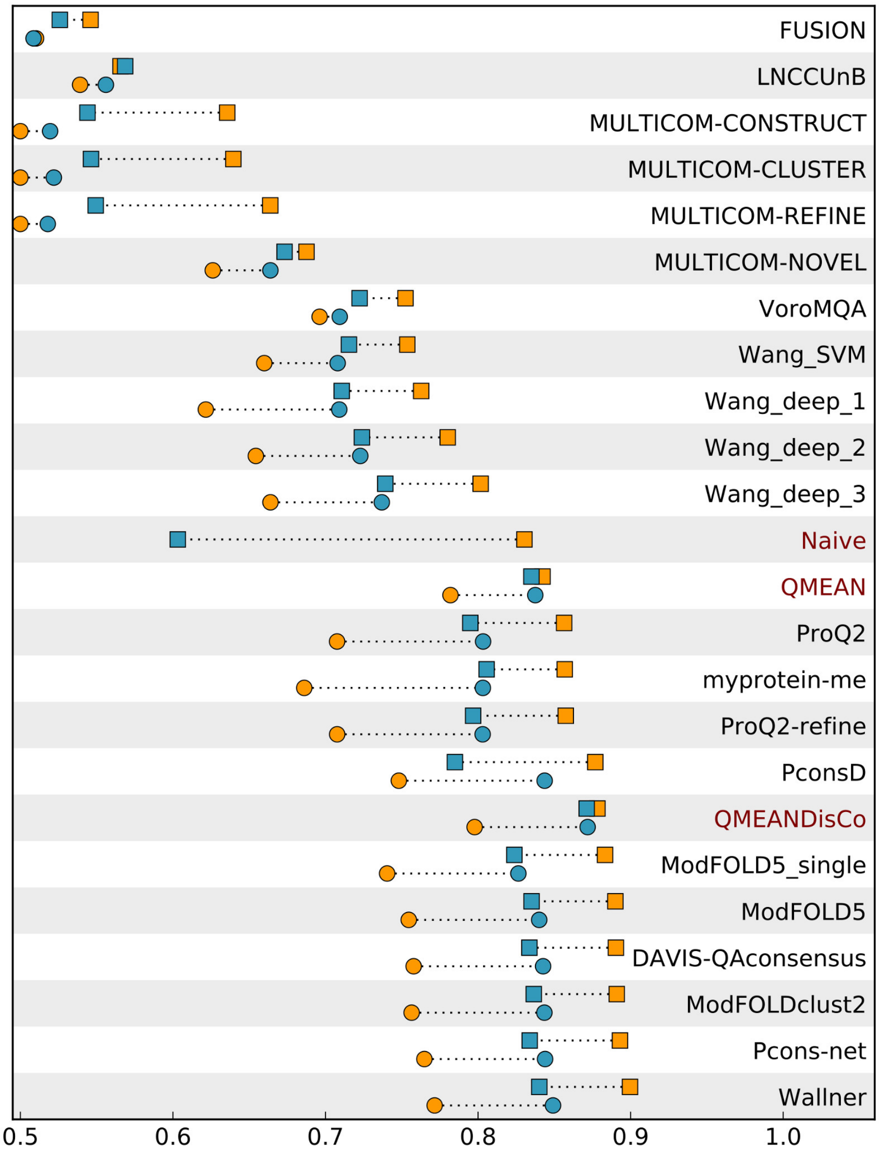


Figure 25: Evaluation on CASP XI test set (orange) and a subset of high quality models (blue) with local lDDT as target value. Squares represent performance in overall AUC and circles in expected per model AUC. For the naive predictor, the overall AUC is displayed only. The expected per model AUC tends to increase when only considering high quality models. This unveils general difficulties in discriminating correctly from wrongly modelled residues in models of low quality. Nevertheless, much of the overall AUC performance seems to originate from exactly that kind of models in many local quality estimation methods.

SWISS-MODEL test set (-2.0%), 2225 for the CAMEO test set (-2.8%) and 12506 models for the CASP XI test set (-4.4%). In the SWISS-MODEL and CAMEO test sets, the ranking regarding the expected per model AUC does not change significantly compared to the overall AUC. In both cases QMEANDisCo performs best with a general trend of the expected per model AUC being lower than the overall AUC (Figure 24; Table 5, Table 6). The situation changes for the CASP XI test set. The difference between overall AUC and expected per model AUC is rather small for QMEAN and QMEANDisCo (-0.06, -0.08 respectively) but much larger for many other methods. This hints to a possible influence of the previously described global effect. In a per model analysis, QMEANDisCo takes the lead with an expected per model AUC of 0.80. QMEAN comes second with 0.78 (Figure 25, Table 7).

3.3.4 *Analysis on High Quality Models for the CASP XI Test Set*

To simulate the situation of only having models with a reasonable overall fold, the evaluation was repeated on the CASP XI test set only considering models with at least 50% of their residues classified as correctly modelled. Only 4789 models of the original set remain (36.6%). The expected per model AUC increases for almost all evaluated methods (Figure 25, Table 10). Discriminating correctly from wrongly modelled residues in low quality models is therefore problematic in general. Another observation is the complete breakdown in performance of the naive predictor when looking at the overall AUC. Simply detecting random coils doesn't work in this setup because they're not present anymore. While QMEAN and QMEANDisCo achieve a similar overall AUC as in the overall CASP XI test set (0.84, 0.87 respectively), most other methods also significantly break down.

3.4 CONCLUSIONS

The local quality estimation problem remains an important aspect of protein structure modelling from a user perspective but also as a tool to improve the modelling process itself. We successfully evaluated the incorporation of distance constraints extracted directly from experimentally determined structures homologous to the model to be evaluated. This led to an enhanced version of QMEAN, QMEANDisCo. The overall AUC and expected per model AUC evaluations show that QMEAN and QMEANDisCo particularly stand out when assessing models with reasonable overall fold. The remaining question is whether local quality estimation on low quality models is meaningful at all. Training and testing on such data might be far from reality and introduce unnecessary biases. Detection of correct overall folds should rather be delegated to global quality estimation methods. Local quality estimation methods could then concentrate on what they're supposed to do: detecting local errors.

3.5 ACKNOWLEDGMENTS

Special thanks goes to Jürgen Haas for his effort in supervising Christine Rempfer during her master thesis and valuable inputs. We also would like to thank Tjaart de Beer and Niklaus Johner for the help in setting up the manuscript. The authors gratefully acknowledge the computational resources provided by the sciCORE / [BC]^2 Basel Computational Biology Center at the University of Basel and the support by the system administration team.

3.6 SUPPLEMENTAL MATERIALS

3.6.1 *dRMSD as a target value for local accuracy*

In contrary to the classical RMSD, the dRMSD (distance-RMSD) is superposition free and represents the root mean square deviation of the difference in distance between all pairs of atoms, either on a per residue or full structure basis. The dRMSD of a certain residue i can be calculated by first gathering the pairwise distances of all its atoms a towards other atoms b from residues $j \neq i$ below a cutoff radius in the reference structure. This gives N distances, which is then compared to the ones extracted from the model:

$$\mathrm{dRMSD}_i = \sqrt{\frac{1}{N}\sum_{a,b} \min(|d_{a,b,ref} - d_{a,b,model}|, cap)^2} \tag{26}$$

To emphasize local behaviour, the cutoff radius has been set to 10 Å. In the case of large or missing distance differences, a cap value of 5Å for the difference in distance has been introduced. The implementation in this work also considers stereochemical issues and clashes by a preprocessing step as described in the lDDT paper. If an atom of a sidechain is stereochemically problematic or involved in a clash, the entire sidechain is removed. If this is the case for a backbone atom, the entire residue is removed. This automatically leads to an increased dRMSD score due to nonexistent atoms and therefore missing distances.

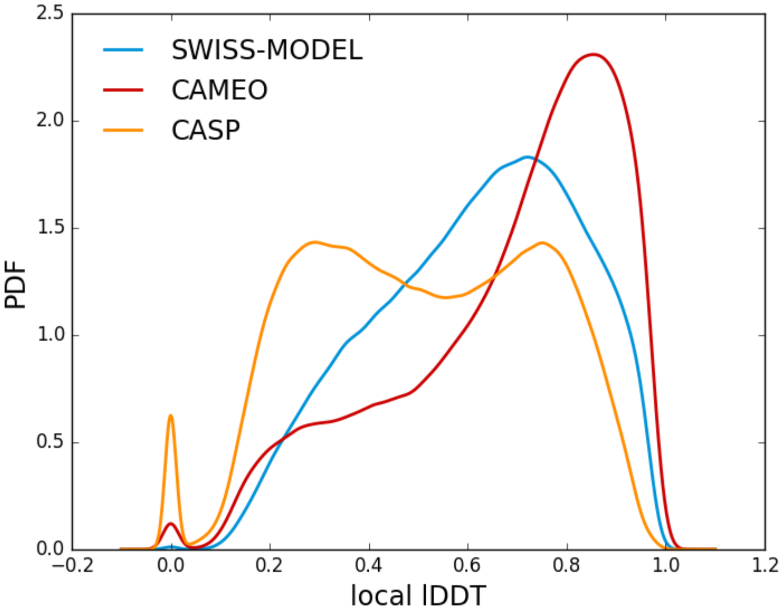
3.6.2 *Parameterization of statistical potentials*

The probability density functions of the statistical potential terms are based on histograms with the following parameterization:

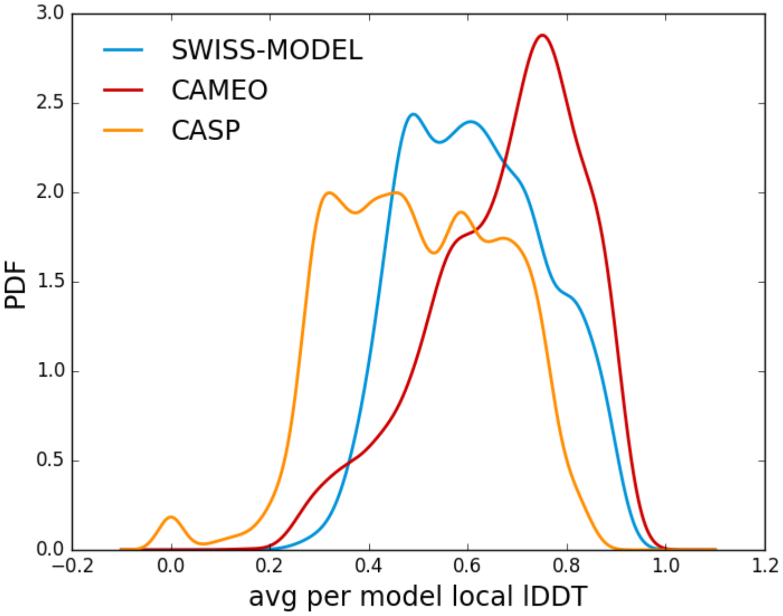
- **local All atom interaction term:** minimal distance: 0.0Å, maximal distance: 10.0Å, bin size: 0.5Å, sequence separation: 4 residues
- **local Cβ interaction term:** minimal distance: 0.0Å, maximal distance: 12.0Å, bin size: 0.5Å, sequence separation: 4 residues
- **local Solvation term:** inclusion radius: 5.0Å, max counts: 32
- **global All atom interaction term:** minimal distance: 0.0Å, maximal distance: 12.0Å, bin size: 0.5Å, sequence separation: 4 residues
- **global Cβ interaction term:** minimal distance: 0.0Å, maximal distance: 12.0Å, bin size: 0.5Å, sequence separation: 4 residues
- **global Solvation term:** inclusion radius: 4.0Å, max counts: 20
- **Torsion term (both local and global):** bin size: 20°

3.6.3 *Target value distribution of different test sets*

The test sets used in this work vary significantly regarding the quality of their underlying protein structure models. We highlight this fact by a more detailed analysis based on lDDT scores. While the SWISS-MODEL and CAMEO test sets largely consist of high quality models, the CASP test set exhibits a clear tendency towards lower quality models (Figure 26).



(a) local lDDT distribution - The peaks at 0.0 for local lDDT arise from residues with stereochemical issues / clashes with the backbone atoms involved



(b) distribution of per model averages of local lDDT values

Figure 26: Target value distribution of the different test sets

3.6.4 *Why you should not use Cα distances as a target value for local quality estimates anymore*

The Cα distance of a residue in a model and its target structure after a global superposition, is an often used measure of local model accuracy. Obvious problems include:

- The measure neglects 90% of the atoms in a protein structure. Interatomic interactions are simply neglected. This can be stereochemical issues, but also favourable interactions such as electrostatic interactions, hydrogen bonds etc.

- The measure does not account for the environment of a residue. The score of a completely buried residue does not get penalized if the entire environment actually making it a buried residue is completely missing.

- The measure is superposition dependent. This leads to two problems.

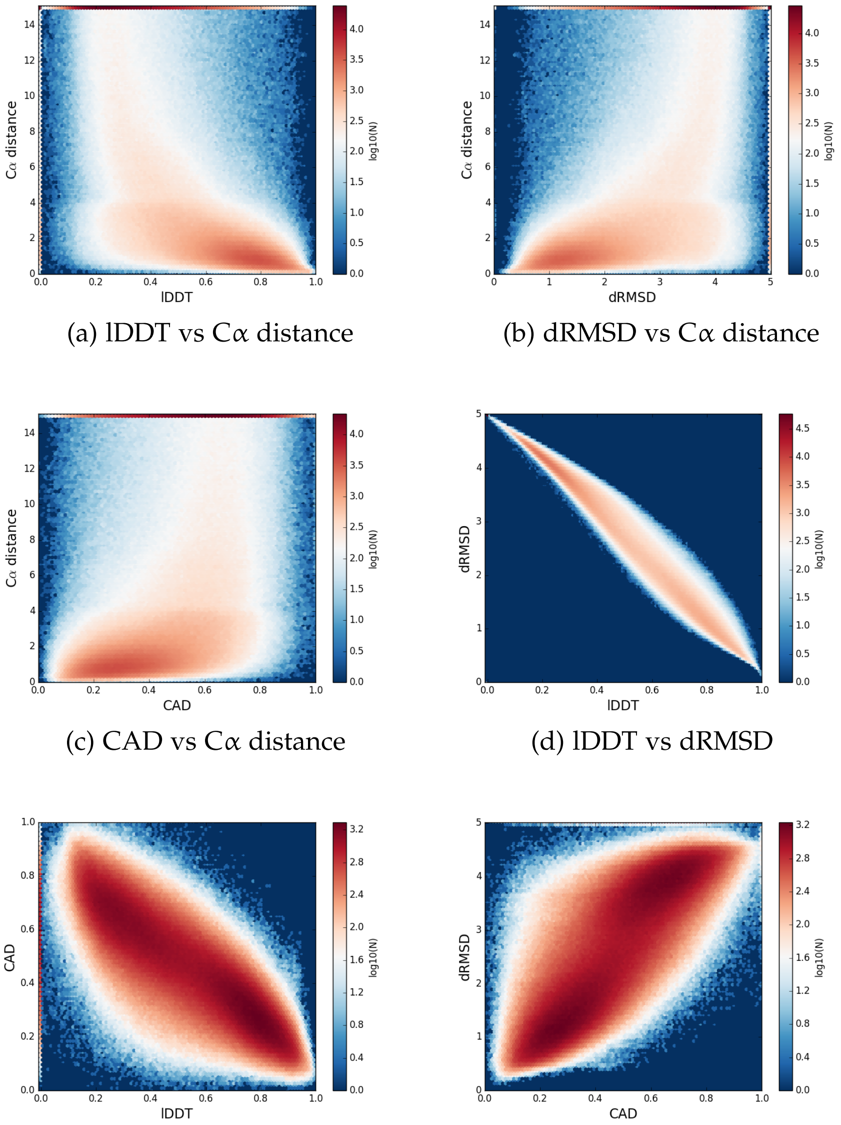
  [1] Especially in low quality models, different tools can create different global superpositions. The score is therefore only reproducible with exactly the same superposition algorithm.

  [2] Domain / hinge movements make it impossible to estimate accurate scores.

The CASP experiment typically uses Cα distances as target value for QE evaluation. We therefore take a closer look at the models of the CASP XI test set with superpositions and local score data provided by the official assessors (LGA output [168] on split domains). For a direct comparison, we also calculated the superposition free all atom scores used in this work (dRMSD, lDDT, CAD). While those three scores show a good pairwise agreement, they correlate poorly with Cα distances (Figure 27).

For further analysis we concentrate on the direct comparison between Cα distance and local lDDT score. The most disturbing observation is the large number of residues being "good" in terms of Cα distance but "bad" in terms of local lDDT. "Good" in terms of Cα distance means a Cα distance below 3.8Å (CASP terminology). "Bad" in our terminology means a local lDDT score below 0.6. It is easy to find many examples, where the Cα distance is obviously flawed and the superposition free scores show a more robust behaviour (Figure 28, Figure 29). The same is true for the opposite scenario, a large number of residues being "bad" in terms of Cα distances but "good" in terms of local lDDT. Many of those data points can be explained by the susceptibility of Cα distances to domain- and hinge movements (Figure 30, Figure 31). Again, superposition free scores show a more robust behaviour.

We therefore believe that the use of Cα distances not only introduces unnecessary noise in the evaluation of quality assessment performance, but also hinders machine learning approaches to unfold their full power when used as target value in training. Superposition free all atom scores should be used instead.



(a) lDDT vs Cα distance

(b) dRMSD vs Cα distance

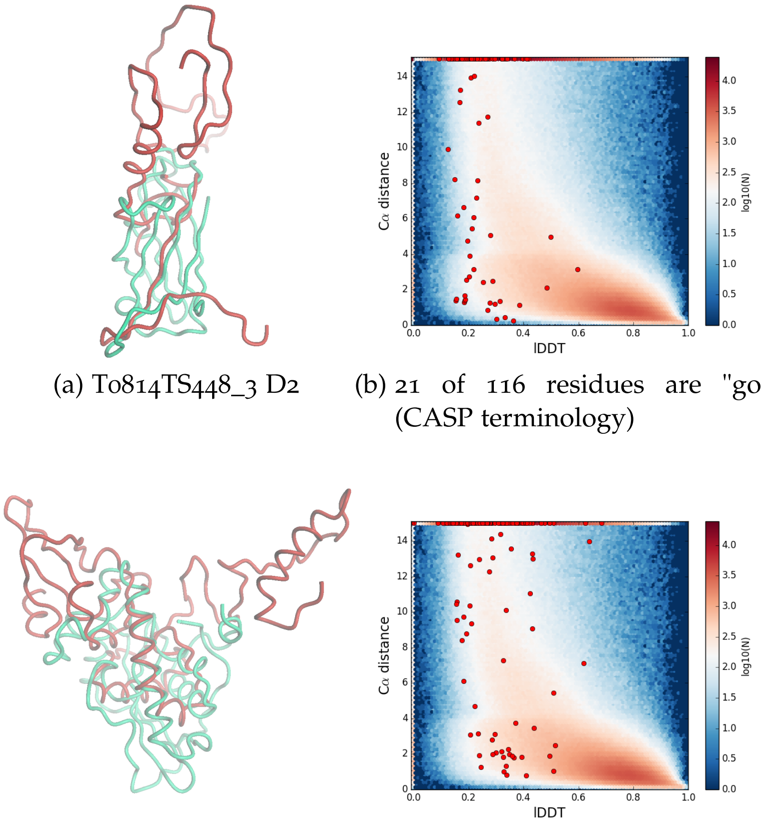
(c) CAD vs Cα distance

(d) lDDT vs dRMSD

(e) lDDT vs CAD

(f) CAD vs dRMSD

Figure 27: Data from the CASP XI (Model2) test set. Every data point compares two scores for the same residue. The Cα distances have been capped at 15 Å and correlate poorly with the superposition free all atom scores used in this work. While lDDT and dRMSD match almost perfectly, the correlation of lDDT and dRMSD towards CAD score is poorer but still good.



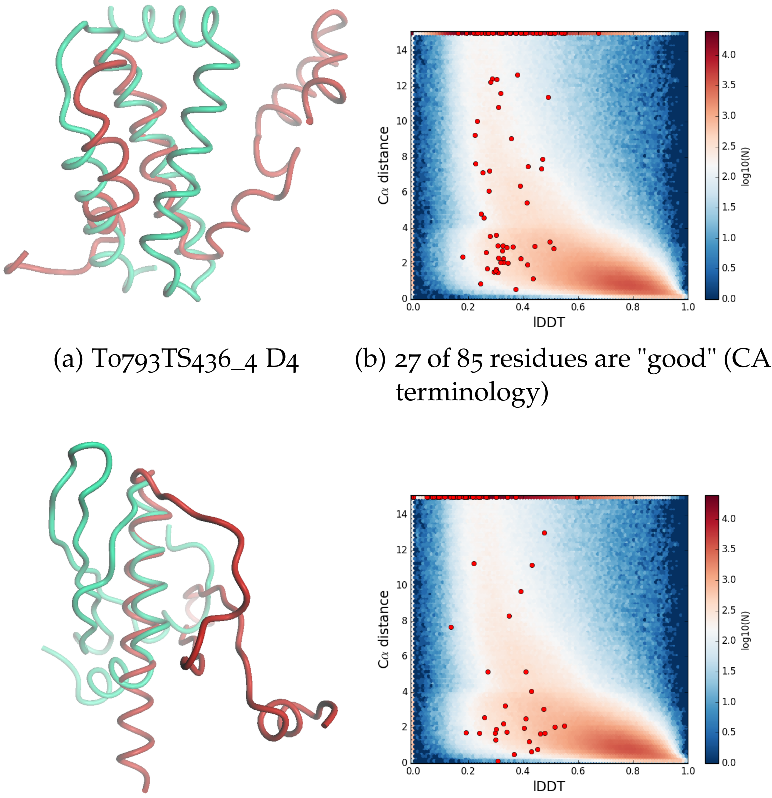
(a) T0814TS448_3 D2

(b) 21 of 116 residues are "good" (CASP terminology)

(c) T0781TS436_4 D1

(d) 23 of 200 residues are "good" (CASP terminology)

Figure 28: (a),(c): Models (red) superposed onto their corresponding targets (green) with superpositions from the official CASP assessors. (b),(d): corresponding per residue lDDT vs Cα distance plots with overall distribution as background.



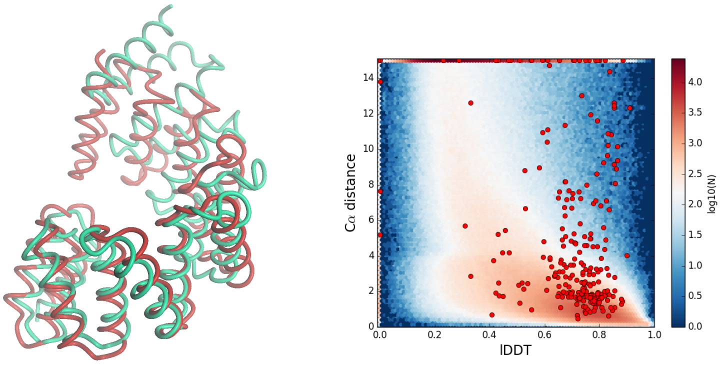
(a) T0793TS436_4 D4

(b) 27 of 85 residues are "good" (CASP terminology)

(c) T0767TS466_3 D1

(d) 21 of 76 residues are "good" (CASP terminology)

Figure 29: (a),(c): Models (red) superposed onto their corresponding targets (green) with superpositions from the official CASP assessors. (b),(d): corresponding per residue lDDT vs Cα distance plots with overall distribution as background.



(a) T0821TS454_4 D1

(b) lDDT values are mostly in the "good" range, whereas Cα distances are pretty much random

Figure 30: (a): Model (red) superposed on target (green) with superposition from the official CASP assessors. Even though the global superposition is bad, the relative orientations of the single helices in the model are largely correct. (b): corresponding per residue lDDT vs Cα distance with overall distribution as background

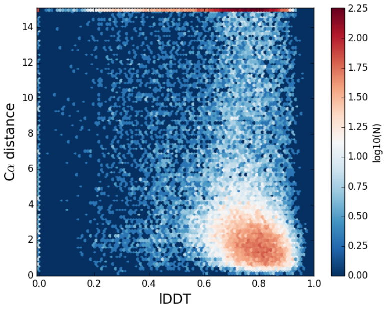


Figure 31: While Figure 30 looks at one model, this plot shows the local lDDT values vs Cα distances for all 150 models of the target T0821. Many data points agree, low Cα distances result in high local lDDT values. But many other data points show the susceptibility of Cα distances to domain-/hinge movements that are particularly pronounced in this target.

3.7 evaluation of local qe performance

Method	Target	Pearson r	Spearman r	AUC	per model AUC
QMEAN_OLD	lddt	0.569	0.590	0.802	0.790
QMEAN	lddt	0.730	0.743	0.874	0.823
QMEANDisCo	lddt	0.857	0.865	0.929	0.869
QMEAN_OLD	cad	0.445	0.450	0.746	0.709
QMEAN	cad	0.626	0.630	0.827	0.759
QMEANDisCo	cad	0.716	0.720	0.861	0.781
QMEAN_OLD	drmsd	0.551	0.567	0.776	0.747
QMEAN	drmsd	0.696	0.705	0.849	0.777
QMEANDisCo	drmsd	0.819	0.827	0.910	0.817

Table 5: Evaluation on the SWISS-MODEL test set

Method	Target	Pearson r	Spearman r	AUC	per model AUC
Verify3d smoothed	lddt	0.484	0.438	0.747	0.640
Dfire v1.1	lddt	0.264	0.285	0.673	0.666
Prosa2003	lddt	0.408	0.371	0.733	0.614
Naive PSIBlast	lddt	0.284	0.322	0.673	0.617
ModFOLD4	lddt	0.672	0.702	0.885	0.817
ProQ2	lddt	0.610	0.642	0.865	0.786
EQuant 2	lddt	0.410	0.331	0.687	0.709
VoroMQA_v2	lddt	0.573	0.562	0.819	0.768
ModFOLD6	lddt	0.678	0.712	0.893	0.827
QMEAN	lddt	0.727	0.705	0.879	0.826
QMEANDisCo	lddt	0.837	0.839	0.932	0.864
Verify3d smoothed	cad	0.411	0.364	0.724	0.611
Dfire v1.1	cad	0.183	0.199	0.645	0.609
Prosa2003	cad	0.359	0.331	0.724	0.592
Naive PSIBlast	cad	0.231	0.257	0.642	0.580
ModFOLD4	cad	0.552	0.583	0.842	0.749
ProQ2	cad	0.513	0.520	0.824	0.718
EQuant 2	cad	0.324	0.240	0.654	0.661
VoroMQA_v2	cad	0.469	0.446	0.780	0.705
ModFOLD6	cad	0.557	0.582	0.848	0.756
QMEAN	cad	0.647	0.610	0.856	0.780
QMEANDisCo	cad	0.736	0.717	0.895	0.804
Verify3d smoothed	drmsd	0.479	0.429	0.729	0.622
Dfire v1.1	drmsd	0.248	0.266	0.647	0.629
Prosa2003	drmsd	0.396	0.362	0.705	0.592
Naive PSIBlast	drmsd	0.267	0.301	0.658	0.610
ModFOLD4	drmsd	0.665	0.677	0.858	0.766
ProQ2	drmsd	0.600	0.620	0.835	0.739
EQuant 2	drmsd	0.397	0.311	0.663	0.669
VoroMQA_v2	drmsd	0.556	0.541	0.790	0.723
ModFOLD6	drmsd	0.670	0.686	0.864	0.776
QMEAN	drmsd	0.698	0.670	0.849	0.781
QMEANDisCo	drmsd	0.809	0.808	0.909	0.814

Table 6: Evaluation on the CAMEO test set

Method	Pearson r	Spearman r	AUC	per model AUC
MULTICOM-CONSTRUCT	0.143	0.257	0.636	0.491
VoroMQA	0.493	0.496	0.752	0.696
LNCCUnB	0.126	0.130	0.566	0.539
ModFOLDclust2	0.672	0.734	0.891	0.756
MULTICOM-REFINE	0.280	0.307	0.664	0.487
FUSION	0.139	0.080	0.546	0.510
Pcons-net	0.662	0.738	0.893	0.765
Wang_deep_1	0.501	0.519	0.763	0.622
MULTICOM-NOVEL	0.331	0.365	0.688	0.626
ProQ2-refine	0.523	0.675	0.857	0.708
ModFOLD5	0.679	0.733	0.890	0.755
DAVIS-QAconsensus	0.532	0.733	0.890	0.758
Wallner	0.668	0.754	0.900	0.772
MULTICOM-CLUSTER	0.195	0.262	0.640	0.494
Wang_deep_2	0.479	0.541	0.780	0.654
Wang_deep_3	0.539	0.579	0.802	0.664
myprotein-me	0.324	0.662	0.857	0.686
PconsD	0.581	0.711	0.877	0.748
ProQ2	0.523	0.673	0.856	0.708
Wang_SVM	0.452	0.497	0.754	0.660
ModFOLD5_single	0.677	0.720	0.883	0.740
QMEAN	0.639	0.664	0.842	0.782
QMEANDisCo	0.704	0.713	0.878	0.798
Naive	0.614	0.619	0.830	nan

Table 7: lDDT Evaluation on the CASP XI test set (Model2)

Method	Pearson r	Spearman r	AUC	per model AUC
MULTICOM-CONSTRUCT	0.111	0.197	0.599	0.494
VoroMQA	0.421	0.425	0.715	0.641
LNCCUnB	0.113	0.106	0.555	0.533
ModFOLDclust2	0.491	0.553	0.784	0.669
MULTICOM-REFINE	0.207	0.231	0.617	0.491
FUSION	0.105	0.072	0.533	0.523
Pcons-net	0.478	0.551	0.782	0.669
Wang_deep_1	0.376	0.391	0.692	0.582
MULTICOM-NOVEL	0.265	0.291	0.647	0.587
ProQ2-refine	0.387	0.506	0.759	0.633
ModFOLD5	0.499	0.554	0.784	0.669
DAVIS-QAconsensus	0.389	0.551	0.783	0.670
Wallner	0.492	0.570	0.792	0.676
MULTICOM-CLUSTER	0.143	0.200	0.601	0.496
Wang_deep_2	0.361	0.407	0.710	0.601
Wang_deep_3	0.410	0.442	0.727	0.608
myprotein-me	0.251	0.502	0.756	0.628
PconsD	0.440	0.538	0.776	0.665
ProQ2	0.388	0.505	0.758	0.632
Wang_SVM	0.354	0.380	0.694	0.609
ModFOLD5_single	0.501	0.547	0.779	0.662
QMEAN	0.526	0.561	0.779	0.714
QMEANDisCo	0.584	0.603	0.800	0.725
Naive	0.457	0.467	0.739	nan

Table 8: CAD Evaluation on the CASP XI test set (Model2)

Method	Pearson r	Spearman r	AUC	per model AUC
MULTICOM-CONSTRUCT	0.142	0.258	0.638	0.485
VoroMQA	0.470	0.471	0.741	0.676
LNCCUnB	0.128	0.135	0.566	0.535
ModFOLDclust2	0.683	0.737	0.888	0.726
MULTICOM-REFINE	0.284	0.310	0.666	0.479
FUSION	0.133	0.088	0.554	0.518
Pcons-net	0.665	0.735	0.889	0.732
Wang_deep_1	0.497	0.508	0.758	0.607
MULTICOM-NOVEL	0.323	0.353	0.680	0.609
ProQ2-refine	0.520	0.668	0.855	0.681
ModFOLD5	0.690	0.735	0.887	0.722
DAVIS-QAconsensus	0.542	0.736	0.888	0.727
Wallner	0.671	0.751	0.895	0.738
MULTICOM-CLUSTER	0.197	0.265	0.642	0.487
Wang_deep_2	0.467	0.524	0.770	0.627
Wang_deep_3	0.528	0.565	0.793	0.639
myprotein-me	0.331	0.664	0.856	0.666
PconsD	0.595	0.724	0.881	0.718
ProQ2	0.521	0.667	0.854	0.681
Wang_SVM	0.436	0.477	0.741	0.635
ModFOLD5_single	0.688	0.721	0.881	0.709
QMEAN	0.605	0.626	0.830	0.748
QMEANDisCo	0.673	0.677	0.872	0.768
Naive	0.643	0.642	0.841	nan

Table 9: dRMSD Evaluation on the CASP XI test set (Model2)

Method	Pearson r	Spearman r	AUC	per model AUC
MULTICOM-CONSTRUCT	0.032	0.070	0.544	0.520
VoroMQA	0.411	0.387	0.722	0.709
LNCCUnB	0.154	0.085	0.569	0.556
ModFOLDclust2	0.544	0.565	0.837	0.843
MULTICOM-REFINE	0.087	0.077	0.549	0.518
FUSION	0.107	0.074	0.526	0.509
Pcons-net	0.568	0.572	0.834	0.844
Wang_deep_1	0.325	0.354	0.711	0.709
MULTICOM-NOVEL	0.320	0.284	0.673	0.664
ProQ2-refine	0.487	0.503	0.797	0.803
ModFOLD5	0.545	0.573	0.835	0.840
DAVIS-QAconsensus	0.431	0.557	0.834	0.843
Wallner	0.573	0.583	0.840	0.849
MULTICOM-CLUSTER	0.091	0.073	0.546	0.522
Wang_deep_2	0.377	0.361	0.724	0.723
Wang_deep_3	0.410	0.386	0.739	0.737
myprotein-me	0.307	0.524	0.806	0.803
PconsD	0.421	0.446	0.785	0.844
ProQ2	0.487	0.502	0.795	0.803
Wang_SVM	0.394	0.361	0.715	0.708
ModFOLD5_single	0.539	0.568	0.824	0.826
QMEAN	0.632	0.633	0.835	0.838
QMEANDisCo	0.716	0.728	0.871	0.872
Naive	0.165	0.174	0.603	nan

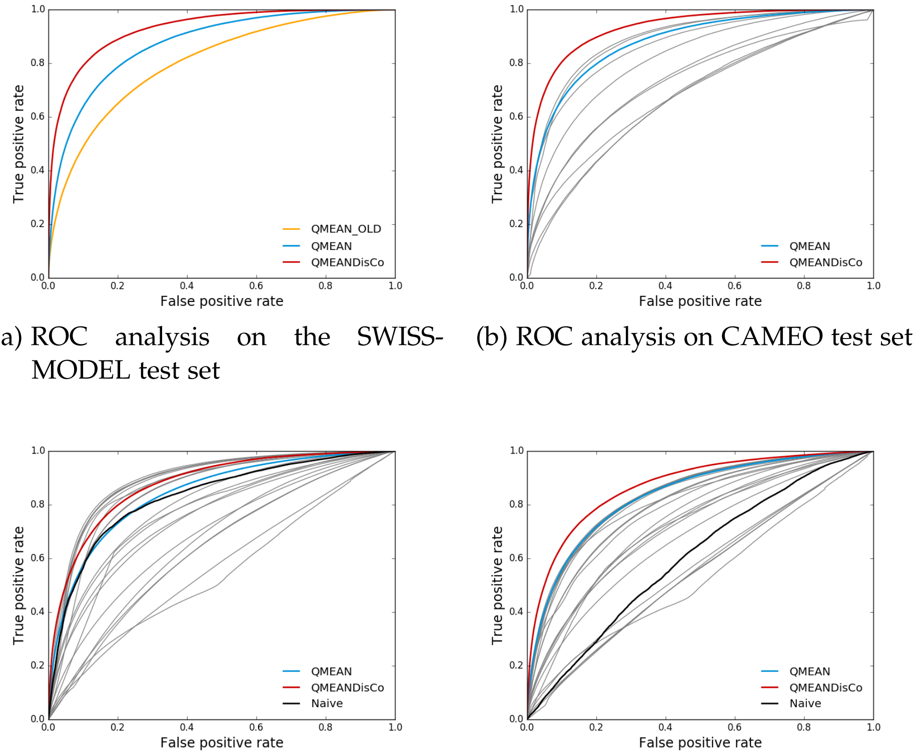
Table 10: lDDT Evaluation on the CASP XI test set (Model2) - only considering models with 50% of their residues having a local lDDT > 0.6

Method	Pearson r	Spearman r	AUC	per model AUC
MULTICOM-CONSTRUCT	0.020	0.044	0.529	0.507
VoroMQA	0.342	0.312	0.701	0.674
LNCCUnB	0.113	0.058	0.563	0.557
ModFOLDclust2	0.390	0.401	0.778	0.772
MULTICOM-REFINE	0.056	0.049	0.534	0.507
FUSION	0.093	0.057	0.527	0.521
Pcons-net	0.405	0.408	0.773	0.769
Wang_deep_1	0.236	0.240	0.673	0.663
MULTICOM-NOVEL	0.243	0.213	0.648	0.632
ProQ2-refine	0.362	0.355	0.746	0.735
ModFOLD5	0.393	0.408	0.778	0.770
DAVIS-QAconsensus	0.302	0.395	0.775	0.771
Wallner	0.413	0.414	0.780	0.774
MULTICOM-CLUSTER	0.057	0.046	0.531	0.508
Wang_deep_2	0.273	0.246	0.683	0.673
Wang_deep_3	0.304	0.273	0.698	0.684
myprotein-me	0.229	0.389	0.756	0.739
PconsD	0.301	0.331	0.739	0.774
ProQ2	0.364	0.357	0.746	0.734
Wang_SVM	0.292	0.247	0.674	0.658
ModFOLD5_single	0.392	0.407	0.770	0.760
QMEAN	0.534	0.509	0.806	0.789
QMEANDisCo	0.595	0.572	0.834	0.815
Naive	0.129	0.131	0.587	nan

Table 11: CAD Evaluation on the CASP XI test set (Model2) - only considering models with 50% of their residues having a local lDDT > 0.6

Method	Pearson r	Spearman r	AUC	per model AUC
MULTICOM-CONSTRUCT	0.027	0.065	0.538	0.516
VoroMQA	0.391	0.367	0.700	0.694
LNCCUnB	0.161	0.087	0.563	0.554
ModFOLDclust2	0.548	0.535	0.800	0.803
MULTICOM-REFINE	0.080	0.069	0.543	0.511
FUSION	0.097	0.069	0.527	0.508
Pcons-net	0.564	0.539	0.797	0.803
Wang_deep_1	0.311	0.327	0.685	0.683
MULTICOM-NOVEL	0.309	0.269	0.651	0.641
ProQ2-refine	0.472	0.471	0.766	0.769
ModFOLD5	0.547	0.542	0.799	0.799
DAVIS-QAconsensus	0.442	0.528	0.797	0.802
Wallner	0.568	0.547	0.803	0.809
MULTICOM-CLUSTER	0.085	0.068	0.540	0.518
Wang_deep_2	0.360	0.332	0.693	0.692
Wang_deep_3	0.390	0.354	0.708	0.707
myprotein-me	0.311	0.486	0.769	0.765
PconsD	0.430	0.435	0.757	0.804
ProQ2	0.472	0.469	0.764	0.770
Wang_SVM	0.374	0.333	0.685	0.680
ModFOLD5_single	0.539	0.536	0.788	0.786
QMEAN	0.591	0.579	0.802	0.805
QMEANDisCo	0.672	0.678	0.841	0.837
Naive	0.181	0.184	0.605	nan

Table 12: dRMSD Evaluation on the CASP XI test set (Model2) - only considering models with 50% of their residues having a local lDDT > 0.6



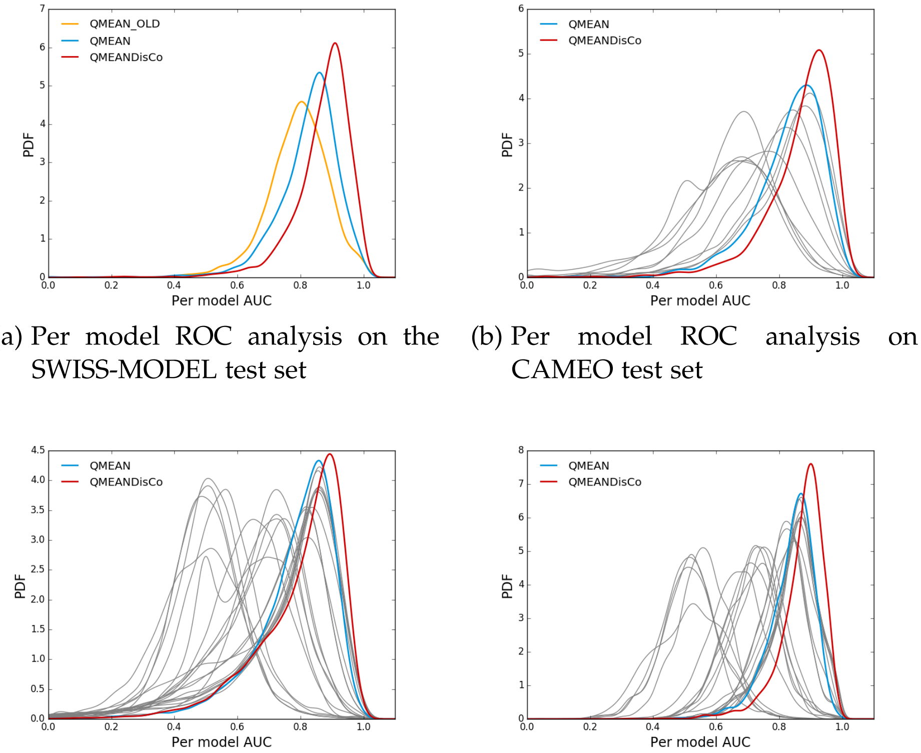
(a) ROC analysis on the SWISS-MODEL test set

(b) ROC analysis on CAMEO test set

(c) ROC analysis on the CASP XI test set

(d) ROC analysis on CASP XI test set - high quality

Figure 32: Raw curves for overall ROC analysis with lDDT as target value



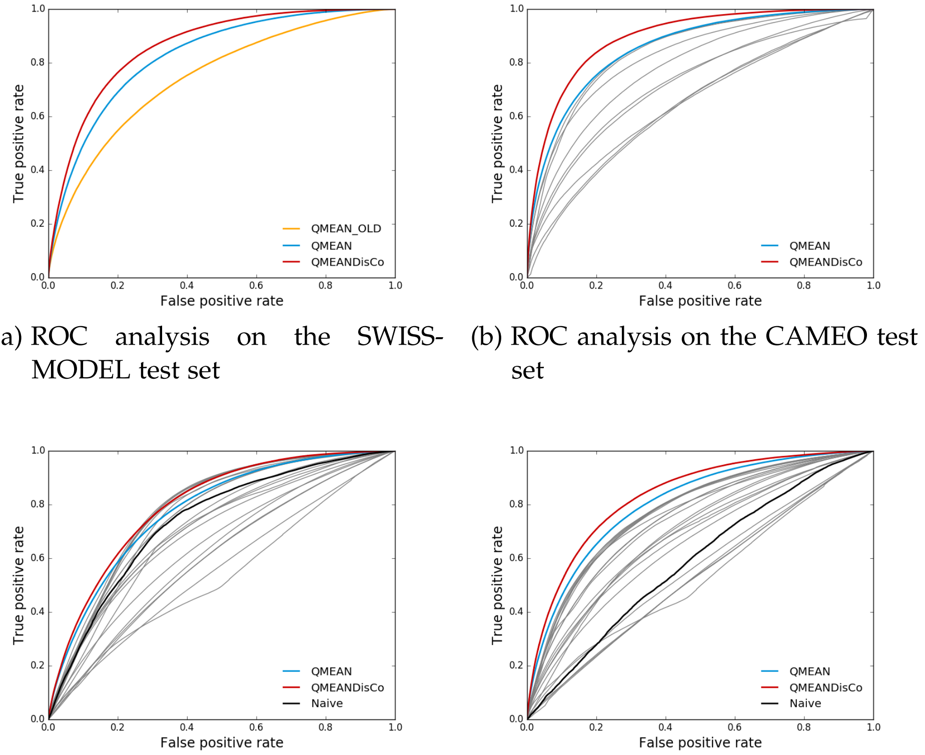
(a) Per model ROC analysis on the SWISS-MODEL test set

(b) Per model ROC analysis on CAMEO test set

(c) Per model ROC analysis on the CASP XI test set

(d) Per model ROC analysis on CASP XI test set - high quality

Figure 33: Raw curves for per model ROC analysis with lDDT as target value



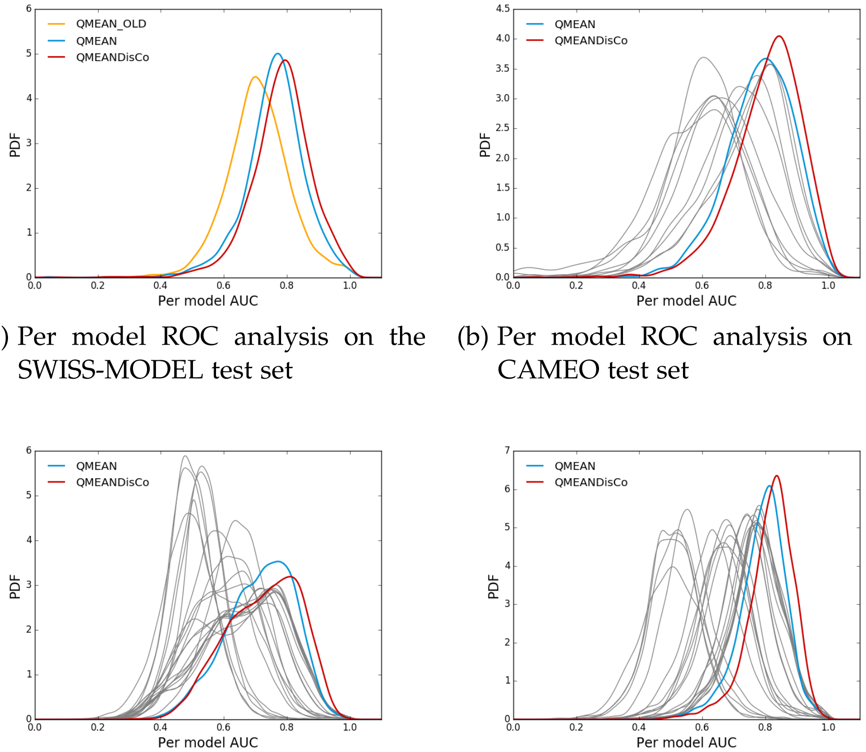
(a) ROC analysis on the SWISS-MODEL test set

(b) ROC analysis on the CAMEO test set

(c) ROC analysis on the CASP XI test set

(d) ROC analysis on the CASP XI test set - high quality

Figure 34: Raw curves for overall ROC analysis with CAD as target value



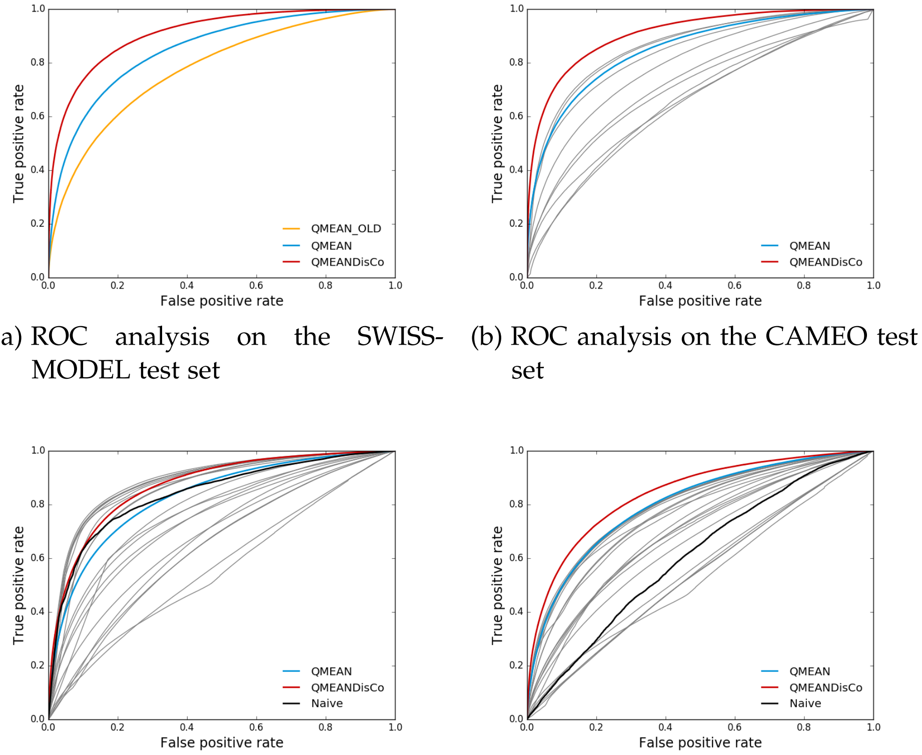
(a) Per model ROC analysis on the SWISS-MODEL test set

(b) Per model ROC analysis on the CAMEO test set

(c) Per model ROC analysis on the CASP XI test set

(d) Per model ROC analysis on the CASP XI test set - high quality

Figure 35: Raw curves for per model ROC analysis with CAD as target value



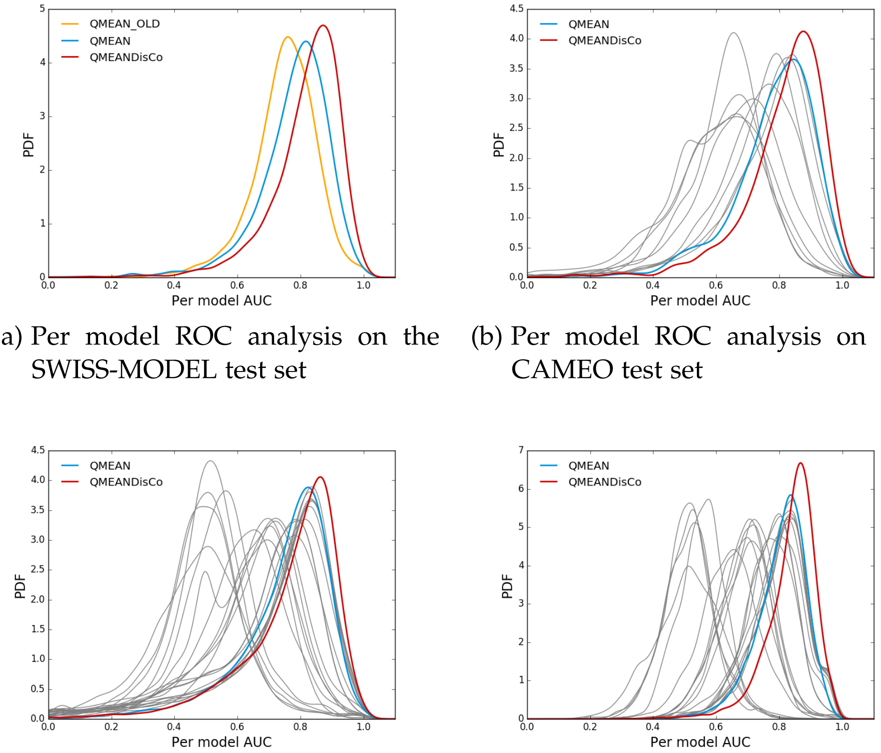
(a) ROC analysis on the SWISS-MODEL test set

(b) ROC analysis on the CAMEO test set

(c) ROC analysis on the CASP XI test set

(d) ROC analysis on the CASP XI test set - high quality

Figure 36: Raw curves for overall ROC analysis with dRMSD as target value



(a) Per model ROC analysis on the SWISS-MODEL test set

(b) Per model ROC analysis on the CAMEO test set

(c) Per model ROC analysis on the CASP XI test set

(d) Per model ROC analysis on the CASP XI test set - high quality

Figure 37: Raw curves for per model ROC analysis with dRMSD as target value

4

PROMOD3 - A VERSATILE HOMOLOGY MODELLING TOOLBOX

This chapter has been a collaborative effort between Gabriel Studer, Gerardo Tauriello, Stefan Bienert and Niklaus Johner.

Author Contributions: GS, GT and SB implemented the software (initials ordered by amount of contributions). GS made major contributions to the modelling pipeline, performed the research related to sidechain modelling, loop / fragment detection, performed all evaluations and wrote the manuscript. GT made major contributions to the modelling pipeline and trained the linear weights for the linear score combination in loop modelling. NJ trained a first set of linear weights for loop scoring.

Motivation: Protein models that extrapolate structural information from evolutionary related proteins are an attractive alternative when experimental data are missing. The underlying computational methods have therefore received great attention in the last decades. Nevertheless, the field is lacking a free, efficient, state-of-the-art modelling engine.
Results: A complete modelling engine has been developed that can perform all steps required to generate a protein model by homology - ProMod3. Its modular design aims at implementing flexible modelling pipelines and fast prototyping of novel algorithms. All modelling tasks, such as loop modelling, sidechain modelling or generating a full protein model by homology have extensively been tested and compared to state-of-the-art methods. In all aspects, ProMod3 has proven to be highly accurate while being extremely performant with respect to computation time.
Availability: ProMod3 is available through the SWISS-MODEL webserver: https://swissmodel.expasy.org

4.1 INTRODUCTION

The Schwede lab provides services such as SWISS-MODEL [22] and the associated SWISS-MODEL repository [23] with the purpose of making protein modelling accessible to all biochemists and molecular biologists worldwide. This leads to several thousand requests to model a tertiary / quaternary protein structures per day with the goal of providing feedback to the user within a few minutes. Given all this information we can define specifications that the SWISS-MODEL modelling pipeline must fulfill:

1. The pipeline must provide state of the art algorithms that are flexible and extensible to fulfil evolving requirements

2. Building the models must be computationally efficient (returning a result within a few minutes)

3. Given the limited runtime, the generated models must be as accurate as possible

4. The models must be free for everybody, no licencing constraints

The first step concerning homology detection can be achieved by using the extremely efficient tools BLAST and HHblits [6] [134]. These tools have proven themselves to be fast and accurate and thus will not be changed in the scope of this project. The second step of generating an actual model has been performed by the ProMod2 modelling engine [58] until recently. ProMod2 is implemented in efficient but hard to maintain C code. Improving the modelling algorithms by including the latest developments in the field has proven itself to be difficult. ProMod2 furthermore failed on certain loop modelling problems, not returning any modelling result in those cases. The first and most important point in the specifications is therefore clearly not fulfilled and we needed a replacement. One obvious choice would have been the widely used modelling engine MODELLER [158]. However, the licencing is restrictive and the source code is not fully available. A clear contradiction to point one and four of our specifications. Another option that has been evaluated was the use of the homology modelling capabilities of the Rosetta [151] or I-Tasser [167] software packages but they both failed to fulfil our requirements regarding point two and four. We therefore decided to implement a modelling engine from scratch - ProMod3. As we demonstrate in this work, the modelling engine is capable of generating highly accurate models of protein structures using limited computational resources. It provides efficient data structures that can be manipulated with state-of-the-art algorithms and allows the implementation of flexible pipelines to solve modelling problems at hand. After obtaining excellent results in an extensive testing phase in the CAMEO continuous evaluation platform [60], the engine has been deployed as default modelling engine in the SWISS-MODEL pipeline as of June 2016 and has ever since generated thousands of homology models for the scientific community worldwide.

4.2 MATERIALS & METHODS

4.2.1 *Architecture*

ProMod3 can considered to be an extension to the OpenStructure software framework [21], which is specifically tailored to homology modelling. Its modular design aims at implementing flexible modelling pipelines and fast prototyping of novel algorithms. The *loop* module provides algorithms and data structures designed to generate and manipulate short peptide segments to model target regions without direct template information. To generate all atom representations of peptide segments, *sidechain* can be used. *scoring* is concerned with the selection of alternative conformations and measuring model reliability in general. Specific modelling tasks that use the aforementioned modules, are gathered in *modelling*. Molecular mechanics tasks to regularize structures or segments thereof are not directly implemented in ProMod3 but the functionality of OpenStructure has been extended to provide wrappers around the OpenMM molecular mechanics library [40]. To ensure efficiency, most tasks and algorithms have been implemented in C++ and made available to the Python scripting language. This allows for rapid prototyping of novel algorithms in Python with the option to easily port them to C++ if execution speed matters. The following sections summarize the implementation details of the individual ProMod3 modules.

4.2.2 *The Loop Module*

OpenStructure provides a flexible object for representing and editing structural information called the EntityHandle. It's flexibility comes at the cost of efficiency in terms of structural manipulations or memory usage. This can be problematic in case of expensive sampling approaches or when large numbers of conformations have to be processed. The *loop* module therefore provides optimized data structures representing peptide segments differentiating between backbone only and all atom representations. They can be created efficiently from an OpenStructure EntityHandle or merged back in. Alternatively, they can be created from scratch. Several objects / algorithms are available for this task. There is a database providing structural information from non-redundant high-resolution X-ray structures (Section 4.2.2.1) and two objects to actually access it: the fragment database that extracts fragments based on geometric criteria (Section 4.2.2.2) and the Fragger that extracts fragments based on sequence derived scores (Section 4.2.2.3). The structural database and fragment database have their origin in the Fragra method [19] but both have massively been refactored to increase accuracy and information content.

4.2.2.1 *Structural Database*

To generate meaningful structural fragments, an efficient approach is to rely on the fact that the available conformational space for short fragments is largely covered by high resolution structures [45, 46] (Section 4.6.1). The *loop* module contains a database to serve as a source for structural information that can be extracted by arbitrary accessors. The information stored in the database is similar to the Rosetta Vall database [56] and is optimized for fast access speed and low memory usage. For every added protein chain we store the amino acid sequence, coordinates of the backbone atoms (N, Cα, C, O), the DSSP [79] secondary structure assignments including the matching DSSP solvent accessibilities, the ϕ/ψ backbone dihedral angles, the sequence profiles derived from HHblits [134] and sequence profiles derived from structural data [171]. A linear memory layout guarantees fast access to the stored information and makes it possible to generate keys to uniquely identify any fragment from the database by only 3 integer values: an entry index, the offset from the start of that entry and the fragment length. Arbitrary accessor objects can be built on top of the structural database that relate fragment keys to arbitrary criteria. The default database shipped with ProMod3 contains a non-redundant set of protein chains as generated with the PISCES webserver [157] using a sequence identity threshold of 90% and a resolution threshold of 2.5Å. This gives ~24 000 chains with >5 000 000 residues and requires about 550MB of memory.

4.2.2.2 *Fragment Database*

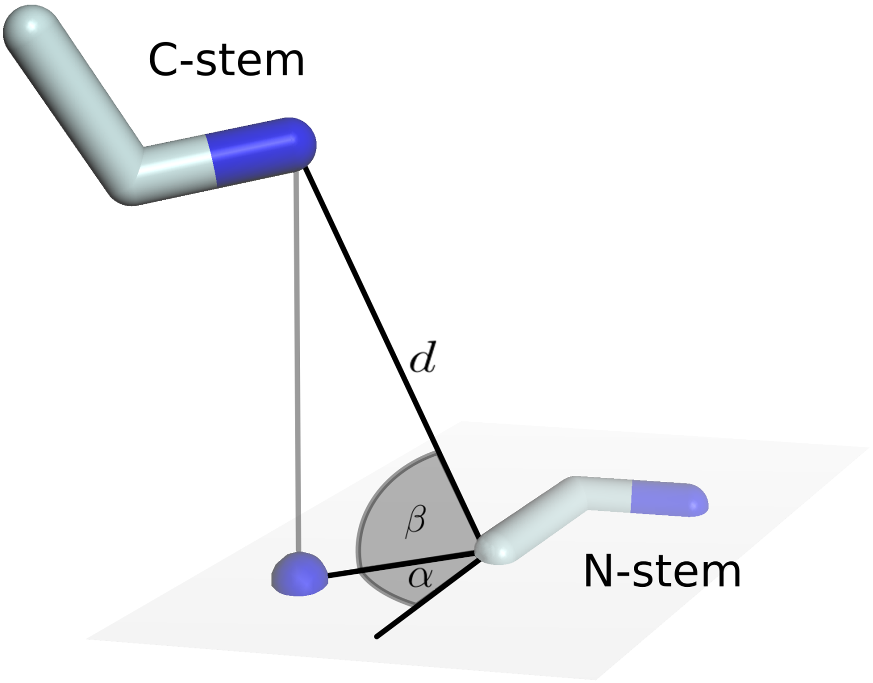


Figure 38: The FragDB accesses an underlying StructureDB using a description of the relative orientation of two stem residues, represented by their N, Cα and C atoms. Descriptors are: number of residues in between (*l*, not shown), distance from N-stem C to C-stem N (*d*) and 4 angles. α and β describe the position of the C-stem N relative to N-Stem N, Cα, C. Accordingly, γ and δ define the N-stem C relative to C-stem N, Cα, C (not shown).

A typical loop modelling problem involves providing structural candidates for a loop that are geometrically constrained by two stem residues [32, 115, 116]. ProMod3 implements a fragment database that accesses the previously described structural database based on geometric criteria. The fragment database reduces the relative orientation of two stem residues to 6 numerical descriptors as visualised in Figure 38. Given a structural database, one can build a fragment database by grouping all possible fragments from the structural database with similar stem geometry. To avoid redundancy, a Cα-RMSD threshold can be enabled. Lightning fast access is guaranteed by the simple stem geometry description and the organization of the fragment database as a hash map. Given 2 stem residues as input query, the description of their relative orientation generates a key in the hash map and gives instant access to all structural fragments with similar stem geometries. The default fragment database shipped with ProMod3 is based on the underlying default structural database and contains information for ~21 000 000 fragments of length 1-12 with ~3 400 000 different stem geometries and requires about 260MB of memory.

4.2.2.3 *Fragger*

The Fragger object provides another way to access the structural database and makes use of the fact that sequence based properties exhibit preferences for local structural conformations [25, 163]. This allows one to massively reduce the immense conformational space to locally preferred regions by querying the structural database for fragments of matching properties. This fact is widely used to propose conformations to sample loop regions or complete peptides without template information [136, 165]. The Fragger object provides a variety of scoring functions to evaluate all possible fragments of a certain length in the structural database for their match towards a query sequence. They are further discussed in Section 4.6.3. Due to the linear memory layout of the structural database, a sliding window approach allows for an extremely efficient search for a list of fragments optimizing a certain score or a linear combination thereof.

4.2.3 *The Scoring Module*

Accurate scoring capabilities are absolutely crucial in many modelling tasks such as the selection of alternative local conformations, guidance of sampling procedures or measuring the general local or global reliability of a protein model. Scorer objects in ProMod3 range from stereochemistry related scorers such as clash scorers [26] to knowledge based scorers implementing statistical potentials of mean force [145]. Additional scorers allow the user to correlate local structural segments to density information or allow to incorporate arbitrary constraint functions between residue pairs. All available scorers are optimized to efficiently assess local structural stretches given a constant environment and are further discussed in Section 4.6.4. From a design point of view, ProMod3 separates between scorer and environment. In order to get access to model-specific data, every scorer requires to be attached to an environment object that can be updated as the modelling proceeds. While scoring of backbone-only segments is straightforward, ProMod3 reconstructs all sidechains of the segment to be scored as well as the residues being close in the scoring environment if a certain scorer requires all heavy atoms to be present.

4.2.4 *The Sidechain Module*

ProMod3 comes with state-of-the-art sidechain modelling algorithms to generate full atom representations of protein models. They are inspired by SCWRL4 [87] but the design of the module allows for interference with the sidechain modelling process at several stages. ProMod3 provides rotamer libraries with access to rotamers with or without dependency on the backbone. It is possible to build custom libraries, although the backbone independent Penultimate [104] and the backbone dependent Dunbrack 2010 [143] libraries are directly provided to the user in binary format.

The rotamers can be represented by their heavy atoms and polar hydrogens as rigid rotamers (Rigid Rotamer Model → RRM). An alternative are flexible rotamers. The same set of atoms builds the basis for an ensemble of conformations, so called sub-rotamers, exhibiting small variations around the sidechain χ dihedral angles to better express the flexibility of sidechains (Flexible Rotamer Model → FRM) [114]. In both cases, RRM and FRM, ProMod3 employs the SCWRL4 energy function to estimate the pairwise energies between rotamers and towards parts of the protein model that are kept rigid. In case of RRM, this is simply summing up all pairwise energies, whereas FRM exploits a thermodynamics based formalism [114].

Having rotamers and all required energies, the optimal combination of rotamers minimizing Equation 1 has to be found. This is extremely complex and a full enumeration of the solution space is computationally not feasible. Preprocessing steps in the form of dead end elimination [54] or edge decomposition [87] are implemented to reduce the problem size so it can finally be decomposed and solved by the graph based TreePack algorithm [166].

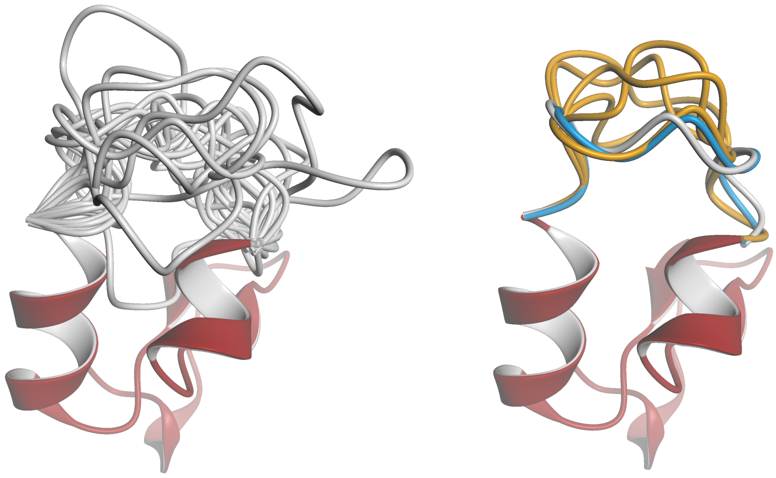
As an alternative, ProMod3 also allows searching for a set of suboptimal solutions regarding the energy function using the A* algorithm [96]. The rational behind considering suboptimal solutions is based on the fact, that the energy difference to the optimal solution might well be within the accuracy limit of the applied energy function.

4.2.5 *The Modelling Module*

Efficient development of pipelines and new functionality is impossible when reoccuring tasks repeatedly have to be reimplemented from scratch. The *modelling* module overcomes this drawback by providing higher level functionality that operates on the data structures from the previously described modules. This can be the adaptation of loop structures on stem residues with CCD [26] / KIC [108], interfaces to the molecular mechanics functionality in OpenStructure for relaxation / minimization, or pipelines to perform full modelling tasks. These pipelines include detection of fragments based on sequence features, loop modelling, sidechain modelling or a full homology modelling pipeline.

4.2.5.1 *Loop Closing*

The task of loop closing is to fit a loop conformation onto the target stem residues that need to be connected. In the context of ProMod3, this is necessary to ensure valid stereochemistry after extraction of loop conformations from the fragment database (Figure 39a) or in the process of Monte Carlo sampling. A first possibility is to run an energy minimization on the loop conformation and enforce matching stems by adding harmonic position constraints accordingly. A drawback of this approach are the computational costs if thousands of loop conformations have to be closed as it is often the case in a typical homology modelling scenario. For computational efficiency, two algorithms were implemented that are both inspired by the field of robotics. The CCD (**C**yclic **C**oordinate **D**escent) [26] and KIC (**Ki**nematic **C**losure) [108] algorithms. The idea of CCD is to first superpose the N-stem of the loop conformation onto the target N-stem and then iteratively alter ϕ/ψ backbone dihedral angles to minimize the RMSD of the loop C-stem and the target C-stem until convergence is reached. This is achieved by describing the RMSD between C-stem and target C-stem as a



(a) Loop candidates extracted from the structural database using the fragment database

(b) Randomly selected loop candidate from Figure 39a (white) closed with CCD (blue) and all possible solutions found by KIC (orange)

Figure 39: Illustration of loop candidates and their fit onto the desired stem residues using CCD / KIC

function of the dihedral angle to be altered. The optimal angle can then analytically be derived. Optionally one can activate a filter to avoid movements resulting in unfavourable backbone dihedral angles. Given a probability density function for the involved amino acid A one can estimate $p_1 = p(\phi, \psi|A)$ and $p_2 = p(\phi_{new}, \psi_{new}|A)$ from which an acceptance probability $p_{accept} = min(1, p_2/p_1)$ can be derived. If a random number r in range $[0, 1]$ fulfils $r < p_{accept}$, the suggested change of the dihedral angles is applied, the next dihedral angles are subject to change otherwise. Despite increased runtime and a lower probability of convergence, ProMod3 uses this filter by default.

In contrary to CCD, the second loop closing algorithm, KIC, is not an iterative approach. It requires to specify 3 pivot Cα atoms that lead to a fragmentation of the loop conformation into 4 rigid pieces. KIC estimates solutions for the ϕ/ψ angles flanking the pivot atoms that lead to loop closure under the constraint that the bond lengths and angles of the pivot atoms remain constant. The constraints can be formulated as three polynomials where common zeros define valid solutions that can be found with the approach of polynomial resultants [35]. This gives up to 16 valid solutions for a certain loop closing problem and requires a selection procedure if only one solution is requested. To summarize: CCD returns exactly one solution per loop closing problem, where changes of backbone dihedral angles are distributed over the whole loop, whereas KIC returns up to 16 solutions but only 3 ϕ/ψ dihedral angle pairs are touched. As a direct consequence, CCD usually only introduces small changes to a loop conformation if the stems already approximately match as it is the case in our structural database / fragment database approach, whereas KIC can alter the complete loop orientation (Figure 39b). CCD should therefore be preferred in the context of the database approaches in ProMod3 to stay as close to the original database loop as possible. For ab initio sampling approaches, KIC should be favoured as it has been shown to improve the exploration of the conformational space available to a given loop modelling problem [108].

4.2.5.2 *Fragment Detection Pipeline*

Given a secondary structure prediction and a sequence profile, the goal is to find *n* structural fragments of length *l* as close to the native structure as possible for every location in the query sequence. The default fragment detection pipeline in ProMod3 uses a Fragger object (Section 4.2.2.3) parametrized with the SSAgreement, TorsionProbability, SeqProfile and StructProfile terms (Section 4.6.3). For optimal performance, a heuristic described in Section 4.6.5.1 is applied on the TorsionProbability term.

To obtain an optimal linear combination of all involved terms, test and training sets have been generated on the basis of 500 randomly selected protein chains present in the default structural database. For each fragment length $l \in (5, 7, 9, 11, 15)$, a set of 1000 fragments has been created by random selection of structural fragments in those 500 protein chains. For each set, 40% are considered for training and 60% for testing. To avoid redundancy, all subsequent training / testing has been performed by querying a new structural database with no structural chains having a sequence identity above 90% to any of the 500 chains selected before.

For every target fragment, n fragments can be extracted from the structural database. A possible measure of success is to evaluate the Cα-RMSDs with respect to the target fragment and estimate the fraction being below a certain threshold. A continuous curve is the result of varying this threshold (in this case 0-3Å) and gives the precision characteristics of a fragment detection method [18, 122]. The area under the curve (AUC) breaks this characteristic down to one number and the average AUC across a whole training set has proven to be a good measure of accuracy to optimize for the desired linear weights.

A problem with this measure is the fact that fragments with common local conformations as native structures, e.g. α-helices, tend to exhibit better performance. This is simply because of the higher probability to select a similar fragment by pure chance. For every target fragment, one can estimate a performance curve resulting from selecting m fragments randomly in the database with $m >> n$ in order to reduce noise. From the original performance curve, this random curve can be subtracted to correct for conformational bias and measure the performance relative to random. For every fragment length we started with initial guesses of the linear weights and performed a conjugate gradient optimization with the average AUC on the corresponding training sets as target value. The test sets have then been used to assure consistency in performance and a direct comparison to the widely used Rosetta fragment picker [56] (Rosetta 3.7) with the underlying Vall jul19.2011 database.

4.2.5.3 *Loop Modelling Pipeline*

The *loop* module provides the capabilities for generating loop candidates based on the structural / fragment databases. Given the observed structural coverage in the structural database Table 4.6.1, high quality candidates can be expected up to a loop length of around 12 residues with the fragment database approach. This satisfies most needs in realistic homology modelling scenarios (Figure 43) and therefore received our main attention for solving loop modelling problems. The goal of the loop modelling pipeline is to (1) propose structural candidates for a certain loop modelling problem and (2) select one of them by employing the *scoring* module. Only in the rare case of longer loops, a Monte Carlo sampling procedure is used as fallback. The parameterization of the sampling is chosen to solve the loop modelling problem in a matter of seconds with the main attention on providing a stereo-chemically valid loop. This does not automatically imply accuracy.

GENERATING LOOP CANDIDATES As a consequence of homology transfer based on a sequence alignment, it is impossible to directly propose structural candidates bridging a gap between stem residues enclosing insertion / deletion events. In case of an insertion, the stems have been connected by a peptide bond in the template structure and one or more residues need to be modelled in this non existent gap. A deletion on the other hand results in a gap with no residues to fill it. In any case, template information must be omitted to allow an extensions of the original gap by shifting the stem residues in order to find stereochemically viable structural candidates. Given the initial stems, a simple elongation schema can be defined according to algorithm 1. Structural candidates can now be found by iterating over the resulting gap extensions and query the internal databases. The iteration stops, as soon as enough candidates have been found, introducing a direct dependency of the found structural candidates on the elongation schema. A variation has therefore been introduced by first generating all possible gap extensions and subsequently apply a scoring based reordering of the found gap extensions before any structural candidates are extracted. Every residue in the target structure gets assigned a penalty if it is part of the gap extension and therefore omitted. The score of a gap extension is then the sum of the penalties from omitted residues plus an additional penalty per elongation. The underlying idea is to give a high penalty for residues that are likely to be structurally conserved in order to first process gap extensions that affect more variable regions. Currently, a penalty of 1.0 is assigned if a residue has been involved in a secondary structure element in the template structure, 0.0 otherwise. Given these values, an elongation penalty of 0.8 per elongation step has been found to give optimal modelling performance (data not shown). Having reordered the extensions, loop candidates are extracted from the databases and fitted onto the target structure using the CCD loop closing algorithm. The processed loop candidates are subject to scoring to decide for one final candidate.

SCORING AND CANDIDATE SELECTION All structural candidates undergo a scoring procedure that linearly combines different scores available from the *scoring* module. To avoid any size effects, the maximal observed extension of the stems towards their corresponding termini is estimated to determine the full scoring range. The scores for every candidate are calculated by merging the candidate into the target structure and calculating all scores for the full scoring range, even though certain candidates might not fully cover it.

To obtain a set of linear weights for any possible combination of scores, a large training set has been generated. From 4000 randomly selected chains in the default structural database, 5000 target loops have randomly been selected for all loop lengths in the range [3, 12] with the only requirements being: no terminal loops and more than 50% of the residues assigned as coil by DSSP. The 4000 chains have then been removed from the default structural database and a new fragment database has been built to query for non redundant candidates for all target loops. A total of 8'185'642 candidates have been found and fitted onto their corresponding stem residues with CCD. In order to find optimal linear weights for any combination of scores, we have defined the target function as follows: the integral (area under the curve) of the cumulative distribution for the probability of the selected candidates for all the target loops being below xÅ in a range of 0-3Å. The resulting optimization problem turned out to be rather complex and a simple conjugate gradient approach showed poor convergence behaviour, CMA [34] as an alternative optimization strategy significantly improved the situation and allowed to estimate weights for different score combinations. As a compromise between speed and accuracy, the default for the database approach is to use backbone only scores (CBPackingScore, CBetaScore, ClashScore, HBondScore, ReducedScore, TorsionScore), complemented by database specific scores that compare the sequence / structure profiles from the loop candidate with the target sequence profile as well as the stem RMSD before applying CCD. The application of all

Algorithm 1: Gap Elongation Procedure

```
n_stem_orig and c_stem_orig are the initial stems;
for elongation ← 1 to max_elongation do
    n_stem ← n_stem_orig - elongation;
    c_stem ← c_stem_orig;
    for shift ← 0 to elongation do
        ProcessGap(n_stem, c_stem);
        n_stem ← n_stem+1;
        c_stem ← c_stem+1;
    end
end
```

atom scores (AllAtomInteractionScore, AllAtomPackingScore, AllAtomClashScore) can be enabled but significantly increases runtime, as the sidechains for every candidate as well as the sidechains in close proximity have to be constructed. Benchmarks have shown, that this further improves loop modelling performance when loops in experimentally determined structures are remodelled. However, the effects for the case of homology modelling have been found to be marginal (data not shown).

Performance of the full loop candidate selection procedure in combination with the subsequent default scoring has been tested on a benchmark used to evaluate FREAD [32], a database driven loop modelling method. For every loop length within four and twenty residues, there are 30 loops to model in experimentally determined structures. To avoid redundancy, all chains in the structural database with sequence identity above 90% to any of the chains in the loop test set have been removed. The data extracted from the FREAD publication allows for direct comparison to other widely used methods such as MODELLER [48], Rapper [10, 36], PLOP [73] and FREAD itself.

4.2.5.4 *Sidechain Modelling Pipeline*

ProMod3 provides a default sidechain modelling pipeline that follows exactly the same steps as SCWRL4 but adds a post-processing step in the end. It takes an input structure, uses rotamers with sub-rotamers (FRM) and solves the sidechain modelling problem using the described graph algorithms. Note, that every rotamer in the FRM gets represented by an ensemble of sub-rotamers to express the variability around the χ dihedral angles. By default, one central sub-rotamer is considered to be the representative. That is the one being enabled when the rotamer is set in the target structure. Instead, as a post-processing step, every rotamer that is part of the solution is transformed to a set of rigid rotamers representing all of its sub-rotamers. Those sets re-enter the pairwise energy calculation and graph solving to decide on the optimal sub-rotamer of each set. The resulting optimal sub-rotamers are considered as a final solution and inserted in the input structure.

To ensure state of the art performance, the set of structures used to test SCWRL4 has been used. It consists of 379 experimentally determined structures. All sidechains of the full asymetric units have been removed and reconstructed. In case of several chains with the same sequence, only the first is considered for evaluation. The main criteria for reconstruction performance is the fraction of χ_1 angles being within 20° of the reference value given by the crystal structure. This is a widely used measurement in the field [27, 87, 99, 106].

4.2.5.5 *Homology Modelling Pipeline*

The default homology modelling pipeline is fully customizable and is intended to serve as a starting point for custom versions. Given an alignment and a template structure, all conserved structural information is transferred to an initial model of the target sequence. In the first step, small deletions are processed by relaxing neighbouring residues and closed if a non-clashing solution can be found. Non-closed deletions from now on get treated as normal insertions and enter the loop modelling procedure. Once the model has a continuous backbone, sidechains are reconstructed using the default reconstruction pipeline. Please note that conserved sidechains that have never been touched in the modelling process are kept rigid. Energy minimization resolves stereochemical irregularities and clashes introduced in the modelling process. Short steepest descent and conjugate gradient minimization runs are iteratively applied on the model until all stereochemical problems are resolved or an upper bound of iterations is reached. Once this is completed, the final model is ready to be used.

To test the performance, a realistic homology modelling scenario has been created by selecting 226 target sequences from 3 months of the CAMEO continuous evaluation platform [60]. The templates with the corresponding alignments exhibiting the best HHblits [134] e-value have been searched in the SWISS-MODEL template library [22] at the day of the CAMEO submission before the according target structures got released to the public. The extracted information allows ProMod3 to run in parallel to the widely used MODELLER tool given the exactly same HHblits profile, template structure and the target-template sequence alignment as input. The goal was to build models as close to native as possible, measured by the superposition free all atom based lDDT score [109]. The MolProbity overall score [30] evaluates stereochemistry as an additional but equally important measure. To avoid unrealistic terminal tail conformations, all models have been trimmed to the region actually covered by the provided template structure.

4.3 RESULTS

4.3.1 *Fragment Detection Performance*

Figure 40 shows similar overall performance in fragment detection between ProMod3 and the Rosetta fragment picker with default parameterization. While being worse when considering fragments of helical target structure, ProMod3 does better at fragments with extended target structures. Conceptually equal terms to the four that are used in ProMod3 are also in use in the Rosetta fragment picker, the two terms considering sequence profiles even share the exact same mathematical formalism. They both outperform their Rosetta counterparts when in use as single terms (Figure 47). Assuming similar sequence profile quality (ProMod3 used HHblits, Rosetta uses PsiBlast [6]), the difference in performance must come from the increased amount of data in the structural database in ProMod3. Consequently, ProMod3 also outperforms the Rosetta fragment picker when only the four equivalent scores get used(Figure 46).

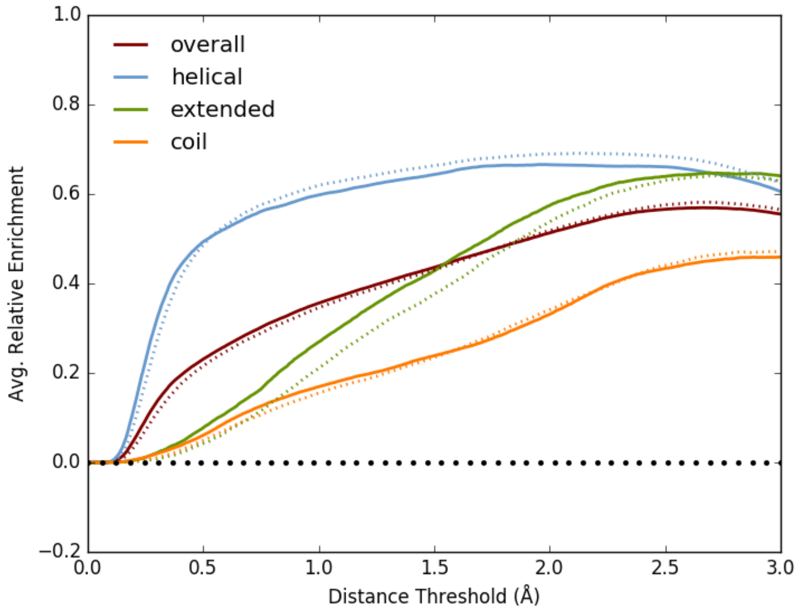


Figure 40: Fragment detection performance on fragments of length 9 relative to random. Solid lines: ProMod3, dotted lines: Rosetta fragment picking protocol. Secondary structure definitions are based on DSSP assignments on native structure. >50% helical residues→helical, >50% extended residues→extended, coil otherwise.

Rosetta can only catch up because of incorporating additional terms including solvent accessibility and backbone dihedral angle predictions [42]. This leads to the conclusion that the ProMod3 performance could further be improved by adding additional terms and Rosetta could profit from a larger Vall database. Besides prediction accuracy, computation time is of great interest for many applications. A speed benchmark further specified in Section 4.6.8.2 gives a speedup of ~6.0x in favour of ProMod3.

4.3.2 *Loop Modelling Performance*

In case of the FREAD benchmark, ProMod3 clearly outperforms all methods it has been compared against (Table 13). This is especially true for shorter loops that predominantly need to be modelled in realistic homology modelling scenarios (Figure 43). The improvements described in the cited FREAD article could unfortunately not be considered. The reason is that

Length	MODELLER	Rapper	PLOP	Original FREAD	ProMod3
4	1.73	1.10	1.79	1.29	0.60
5	2.30	1.23	2.76	2.19	0.66
6	2.38	1.92	3.25	1.79	0.98
7	3.44	2.60	3.73	2.53	1.43
8	4.25	2.88	4.34	2.88	1.89
9	4.31	3.03	5.58	3.08	1.51
10	5.69	3.90	6.41	4.25	2.15
11	5.34	4.63	6.52	4.55	1.93
12	7.18	5.10	6.86	3.99	3.65
13	6.96	5.72	7.86	5.54	2.56
14	7.24	6.02	8.37	6.07	5.21
15	7.93	6.41	9.60	6.41	3.99
16	8.65	7.29	9.86	7.50	4.50
17	9.61	7.35	9.00	7.84	6.28
18	7.64	7.56	10.54	5.48	6.61
19	10.52	9.10	11.51	7.67	7.04
20	10.49	10.64	11.14	7.64	9.23

Table 13: Loop Modelling Performance Analysis - Comparing average backbone RMSD (N, CA, C, O) for each loop length in the FREAD benchmark set with data extracted from the FREAD publication

the described method did not return results for all loops in the test set and the reported RMSD values for each loop length are based only on a subset (typically around 60%). FREAD is therefore represented by the Original FREAD column [37]. As the benchmark only considers loops in high resolution X-ray structures, it doesn't reflect a realistic homology modelling scenario. Consider this missing aspect to be covered in the overall homology modelling performance analysis.

4.3.3 *Sidechain Modelling Performance*

The similarity of the algorithms in ProMod3 and SCWRL4 leads to comparable sidechain reconstruction performance when using rotamers with sub-rotamers (average fraction of correct χ1: 82.30% (ProMod3), 82.18% respectively; Table 16, Table 18). This fraction slightly increases for ProMod3 when applying the described post-processing of selecting the optimal sub-rotamers (82.52%; Figure 41, Table 15). When analyzing the performance increase on a per amino acid basis, bulky sidechains profit most (PHE 92.09%→93.85%, TRP 87.13%→89.38, TYR 90.66%→92.33%;

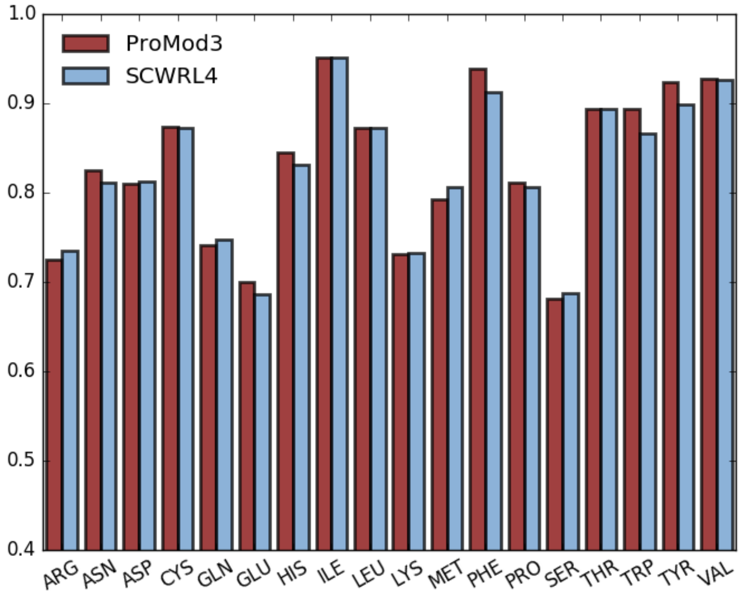


Figure 41: Comparison of sidechain modelling performance with SCWRL4 by measuring the fraction of χ1 angles being within 20° of the reference angles observed in the SCWRL4 test set.

Table 15, Table 16). Even with post-processing, a speedup of ~2.8x (4.5x when not using sub-rotamers (RRM)) compared to SCWRL4 can be observed on a speed benchmark further described in Section 4.6.8.3.

4.3.4 *Homology Modelling Performance*

Given the analysis on the generated models, ProMod3 shows an average lDDT score increase of 1.68 (Figure 42). Also regarding MolProbity overall scores, ProMod3 produces significantly better results by an average decrease of 1.27 (Figure 42). Note that the MolProbity overall score is intended to relate with X-ray resolution, lower is therefore better. The decomposition of the MolProbity overall score into its single components (Clashscore, Ramachandran not favored and percentage bad sidechain rotamers; Figure 50) reveals MODELLERs inability to resolve clashes when default settings are used. The situation for MODELLER models only slightly improves with increased re-

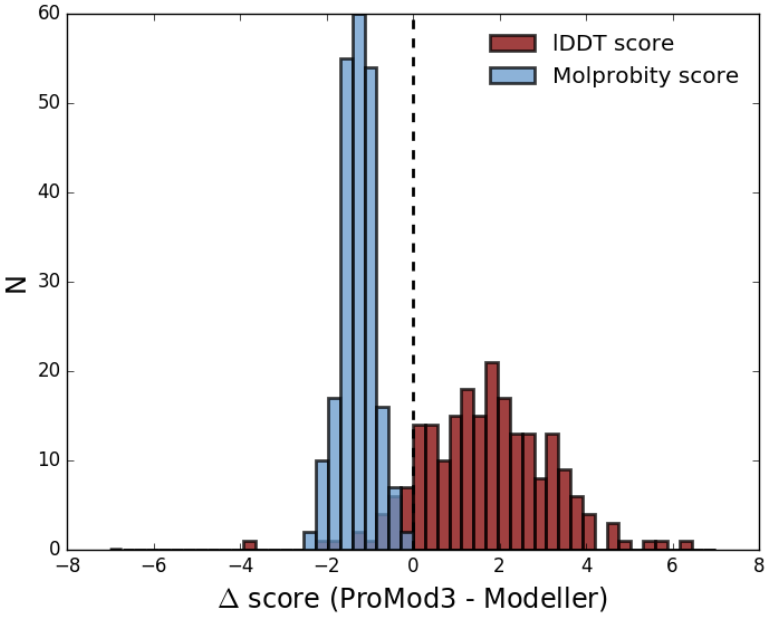


Figure 42: Comparison of overall homology modelling performance with MODELLER. Exactly the same alignment and template serve as input to model a total of 226 targets. The similarity to the native structure is measured by the lDDT score (higher is better) and stereo-chemistry by the MolProbity score (lower is better).

finement level (automodel.md_level set to refine.very_slow) at high cost of computation time (Figure 51). When using the default settings in MODELLER, ProMod3 has a moderate increase of modelling speed of a factor 1.4x in a benchmark further specified in Section 4.6.8.4. With the increased md refinement, this factor increases to 7.2x.

4.4 DISCUSSION

ProMod3 has matured to a competitive homology modelling engine implementing state of the art algorithms with tricks and tweaks to make them as efficient as possible. Accurate protein loops can be modelled within seconds given the presented database approach in combination with a wide variety of scoring functions. If one is willing to spend more computational time, structural fragments open up the space for fully customizable sampling procedures. All this gets complemented by highly accurate and fast sidechain modelling algorithms that not only allow the construction of all atom protein models but also extend the loop modelling capabilities by feeding back the all atom information into the available all atom scorers. In combination with the OpenMM wrappers provided by OpenStructure, ProMod3 also gains highly customizable molecular mechanics capabilities and can therefore considered to be a complete modelling engine. The combination of state of the art algorithms with the flexibility that comes with exporting all functionality to the Python programming language helps implementing and testing of new approaches. ProMod3 can therefore considered to be the basis and a promise for the future.

4.5 ACKNOWLEDGMENTS

Special thanks goes to the whole CAMEO team for countless resubmissions when something was awry with one of the test servers. Also the whole Schwede group was extraordinary helpful for input during endless group meetings, coffee breaks or simply while drinking a beer.

4.6 supplemental materials

4.6.1 *Structural Coverage in Structural Database*

Fragment based approaches rely on the fact that the structural space for fragments up to a certain length is fully covered. This section evaluates the current state of structural coverage in the default structural database in ProMod3. To represent the observed conformational space for every fragment length between 3 and 15, empty sets have been created. For every set, all possible fragments of according length in the structural database have been iterated and added, if no similar fragment was already present (no other fragment with Cα-RMSD below 3Å). In a second step, every fragment in all of the sets has been checked for similarity to any fragment of same length in a different entry of the structural database (with a stringent definition of similarity of Cα-RMSD below 1Å).

Length	Unique Fragments	Covered	Fraction Covered
3	1	1	1.00
4	1	1	1.00
5	2	2	1.00
6	3	3	1.00
7	6	6	1.00
8	15	15	1.00
9	27	26	0.96
10	53	50	0.94
11	114	102	0.89
12	237	182	0.77
13	495	350	0.71
14	1016	650	0.64
15	2124	1292	0.61

Table 14: Structural Coverage in Structural Database

4.6.2 *Relevant Loop Lengths*

Loop modelling accuracy decreases with increasing loop lengths, as can be seen in the loop modelling performance analysis. This is a result of the increasing conformational space. In case of database approaches, this space is less and less covered and in case of de novo or sampling approaches it gets increasingly expensive to explore it. To get an idea of the general difficulty of the problem, it is worth analyzing loop lengths as they occur in typical homology modelling scenarios. For that we analyzed all inserted loops when running the default pipeline in ProMod3 on the full homology modelling benchmark. Modelling all 226 targets required to model a total of 1018 loops. Figure 43 clearly shows, that short loops dominate. 978 loops (96.1%) are of length 12 or less.

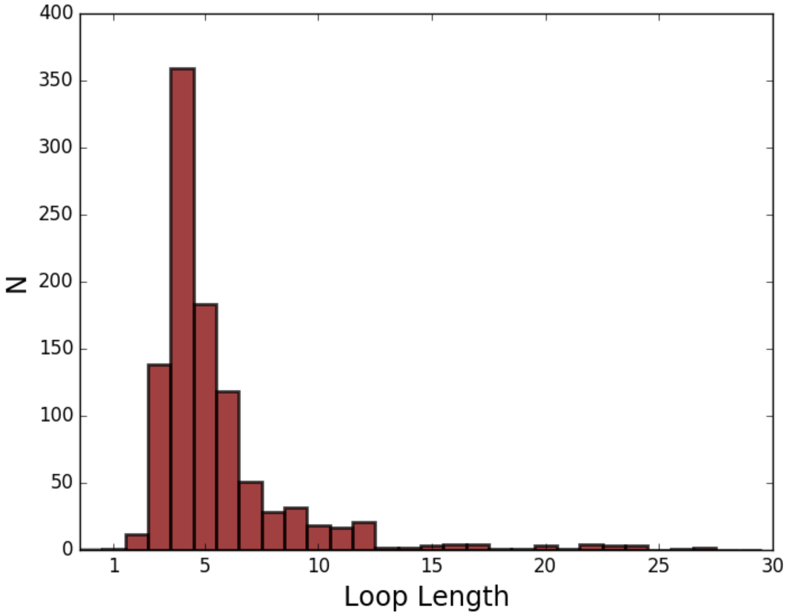


Figure 43: Observed lengths of loops inserted in the full homology modelling benchmark using the default ProMod3 pipeline. From a total of 1018 loops, 978 are of length 12 or shorter. 6 loops longer than 30 residues are not shown in the plot.

4.6.3 *Fragger Scores*

All scores implemented in the Fragger object (Section 4.2.2.3) are listed here with a detailed description:

- **SeqID:** Fraction of identical amino acids to the query sequence.
- **SeqSim:** Avg. substitution score to the query sequence as estimated by any substitution matrix available in OpenStructure, e.g. BLOSUM62.
- **TorsionProbability:** Avg. probabilities of ϕ/ψ dihedral angles in the structural database given the input sequence. The probabilities also consider the identity of the flanking residues. Instead of generating distributions for all possible combinations of flanking residues, the flanking residues can be grouped arbitrarily. The default grouping scheme follows Solis & Rachovsky [150]. Default distributions are available to the user, but custom distributions can be generated with custom grouping schemes and training data.
- **SSAgreement:** Avg. secondary structure agreement score given a PSIPRED [75] prediction and the observed secondary structure in the structural database as estimated with DSSP. The used formalism is probabilistic [149]:

  $$S(d,p,c) = \log\left(\frac{p(d,p,c)}{p(d)p(p,c)}\right) \tag{27}$$

  where *d* is the secondary structure assignment by DSSP, *p* the secondary structure prediction from PSIPRED and *c* the according PSIPRED confidence. The exact same distributions are also in use in the QMEAN scoring function.
- **SeqProfile:** Avg. L1 distance of profile columns in query sequence profile and the sequence profiles present in the structural database. The same formalism is used in the Rosetta fragment picking protocol [56]:

  $$S(p,q) = \sum_{i=1}^{20} |p(i) - q(i)| \tag{28}$$

  where $p(i)$ represent the probabilities of the 20 standard amino acids in the query sequence profile and $q(i)$ the same in the target sequence profile. Another formalism as it is in use in HHblits would be:

$$S(p,q) = \log\left(\sum_{i=1}^{20} \frac{p(i)q(i)}{f(i)}\right) \tag{29}$$

where *f* additionally represents a reference distributions. This is computationally more expensive but did not improve performance in fragment detection (data not shown).

- **StructProfile:** Same as SeqProfile, but the query profile is compared to the structural profiles in the structural database.

4.6.4 *Scorers of the Scoring Module*

Many of the scorers in the *scoring* module are based on statistical potentials of mean force. They are only containers and internally operate on lookup tables. These need to be filled upon parameterization. For all of them, ProMod3 provides default versions that can be loaded from disk. Available scorers are:

- CBPackingScore: Statistical potential that evaluates the number of other Cβ positions within a certain cutoff radius of the Cβ position of the residue to be evaluated.

  The calculated scores are summed and normalized by the number of scored residues.

- CBetaScore: Statistical potential that evaluates pairwise interactions between Cβ atoms which are located within a cutoff and that are at least seq_sep residues apart. A score is assigned to each distance using equally sized bins and distinguishing all possible pairs of amino acids.

  Every pairwise interaction within the loop and towards the environment is evaluated, summed up and finally normalized by the number of evaluated interactions.

- ReducedScore: Statistical potential that evaluates pairwise interactions between the reduced representation of residues with Cα distance $<$ cutoff and that are at least seq_sep residues apart. Every residue gets represented by its Cα position p and a directional component $v = norm(ca\_pos -$

$n\_pos) + norm(ca\_pos - c\_pos)$. For interacting residues $r1$ and $r2$, we can define a line l between $p1$ and $p2$. The statistical potential then considers:

  - dist: distance between $p1$ and $p2$
  - α: angle between $v1$ and l
  - β: angle between $v2$ and l
  - γ: dihedral between $(p1 + v1, p1, p2, p2 + v2)$

  A score is assigned to each combination of parameters using equally sized bins and distinguishing all possible pairs of amino acids.

  Every pairwise interaction within the loop and towards the environment is evaluated, summed up and finally normalized by the number of evaluated interactions.

- HBondScore: Statistical potential that evaluates hydrogen bonds similar to the one defined in the Rosetta energy function [86]. It considers the Cα, C and O positions from backbone hydrogen bond acceptors in interaction with the N and H positions from the backbone hydrogen bond donors. Four Parameters describe their relative orientation:

  - dist: H-O distance
  - α: O-H-N angle
  - β: C-N-H angle
  - γ: Cα-C-O-H dihedral angle

  A scoring function with equally sized bins for all combinations of these parameters for three different states is generated. State 1 for helical residues, state 2 for extended residues and state 0 for all other residues. If the state of two interacting residues is the same, that is the one from which the score is extracted. In all other cases, the energy is extracted from the 0 state.

  Every pairwise interaction within the loop and towards the environment is evaluated, summed up and finally normalized by the number of residues in the loop.

- TorsionScore: Statistical potential that evaluates ϕ/ψ backbone dihedral angles taking into account the identity of the scored residue, but also its flanking residues. Instead of generating a scoring function with equally sized ϕ and ψ bins for all possible combinations of flanking residues, the flanking residues can be grouped arbitrarily. The default grouping scheme follows Solis & Rachovsky [150].

All evaluated scores are summed and normalized by the number of residues in the loop. The first ϕ and last ψ angle of the loop are determined with the help of the scoring environment if set.

- AllAtomInteractionScore: Statistical Potential that evaluates pairwise interactions between all atoms that are located within a cutoff and that are at least seq_sep residues apart. A score is assigned to each distance using equally sized bins and distinguishing all possible pairs of chemically distinguishable atoms.

  Every pairwise interaction within the loop and towards the environment is evaluated, summed up and finally normalized by the number of evaluated interactions.

- AllAtomPackingScore: Statistical potential that evaluates the number other heavy atoms within a certain cutoff radius around all heavy atoms of a residue not belonging to the assessed residue itself.

  The calculated per atom scores are summed and normalized by the number of atoms being assessed.

- ClashScore: Calculates a simple clash score between all pairs of atoms among the evaluated residues and towards the set environment. There is no need to define any parameters here as all interaction energies are fixed [27].

  All calculated scores are summed and normalized by the number of residues in the loop.

- AllAtomClashScore: Same as ClashScore but considering all atoms.

  The calculated score is normalized by the number of atoms being assessed.

- DensityScore: Given an input structure, the scorer generates a density map of the loop to be scored [39] and estimates the normalized cross correlation to a user defined target map.

- SSAgreementScore: Evaluates a secondary structure agreement score as it is already defined as Fragger score. In every score evaluation, the secondary structure of the loop is estimated by searching for hydrogen bonds leading to a secondary structure as defined by DSSP. The hydrogen bonds are searched internally in the loop as well as towards the environment.

  The final per residue scores are summed and normalized by the number of residues in the loop.

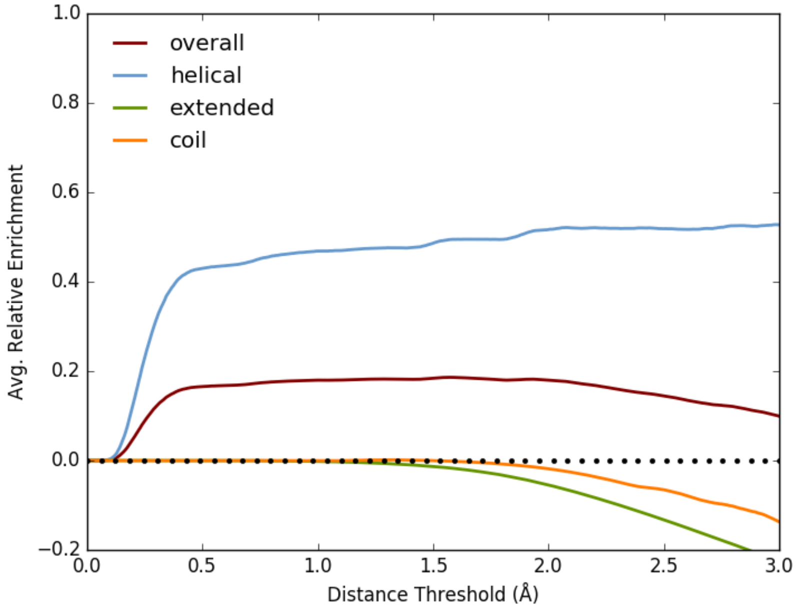
- PairwiseScore: Evaluates a list of generic pairwise functions. They are user defined and can either be simple contact functions (evaluate to x if $dist < max\_dist$, 0.0 otherwise) or arbitary lookup tables.

  When evaluating a loop, the scores of all pairwise functions that involve a residue in the loop are summed up (the other residue can be either in the loop or in the scoring environment) and normalized by the number of residues in the loop.

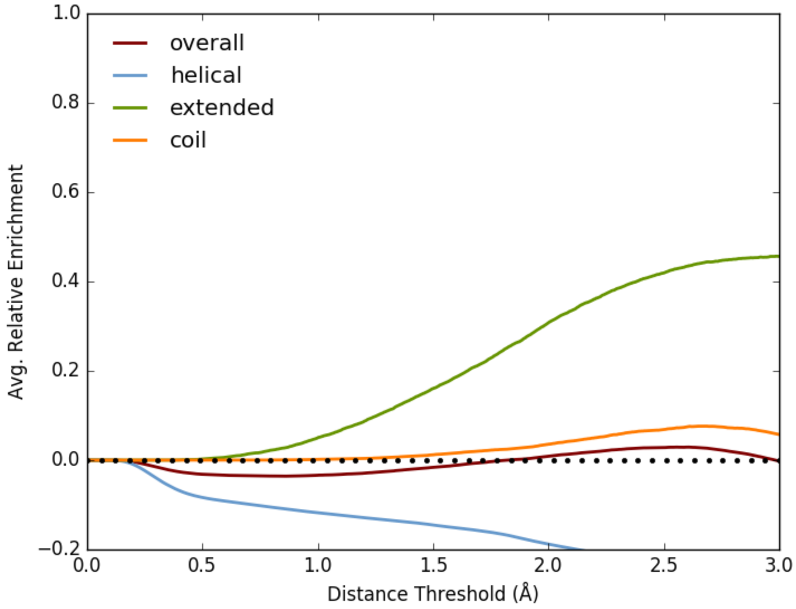
4.6.5 *Fragment Detection Performance*

4.6.5.1 *Torsion Score Heuristic*

The outcome of the torsion score in fragment detection is highly dependent on the secondary structure of the target fragment and the used torsion probability distributions. The top scoring fragments typically represent the secondary structure that dominates the data those distributions have been trained on (Figure 44a, Figure 44b, Figure 45a). To reduce this effect, ProMod3 incorporates the secondary structure prediction from PSIPRED in the default torsion score term and assigns torsion probability distributions on a per residue basis. If the PSIPRED prediction for a certain residue is helical or extended with a confidence of $>= 6$, the applied torsion probability distributions are trained on residues with the according secondary structure. In all other cases, the distributions are trained on coiled residues. Using this heuristic, the secondary structure dependency of the the torsion score is reduced (Figure 45b).

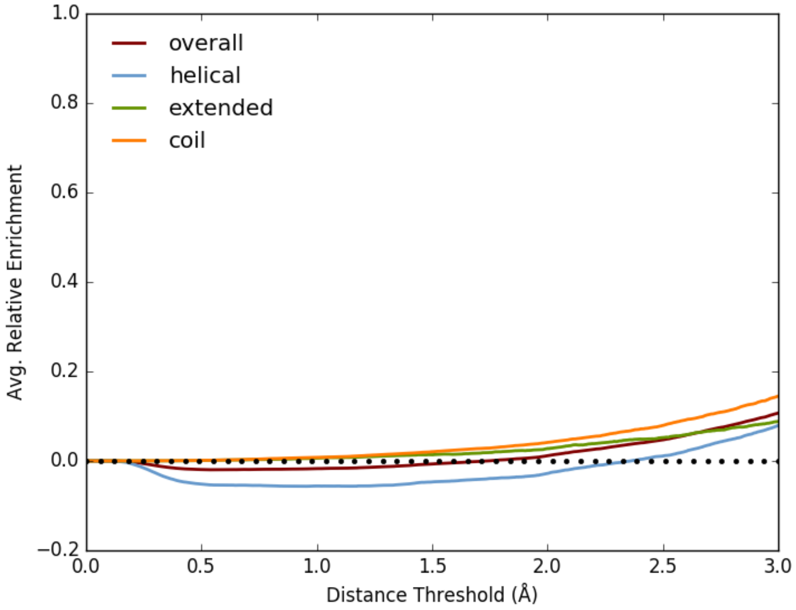


(a) Torsion probability distributions trained on helical residues

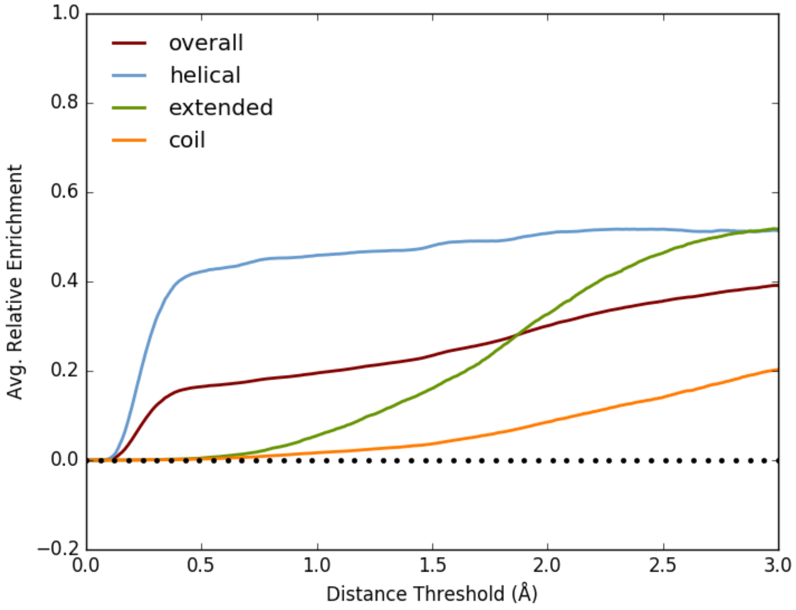


(b) Torsion probability distributions trained on extended residues

Figure 44: Comparison of fragment detection performance of TorsionProbability term on fragments of length 9 relative to random. The secondary structure specific performance is highly dependent on the underlying probability distributions.



(a) Torsion probability distributions trained on coiled residues



(b) Per residue torsion probability distributions selected according described heuristic

Figure 45: Continuation of Figure 44 and the result of the described torsion heuristic.

4.6.5.2 *Comparison to Rosetta*

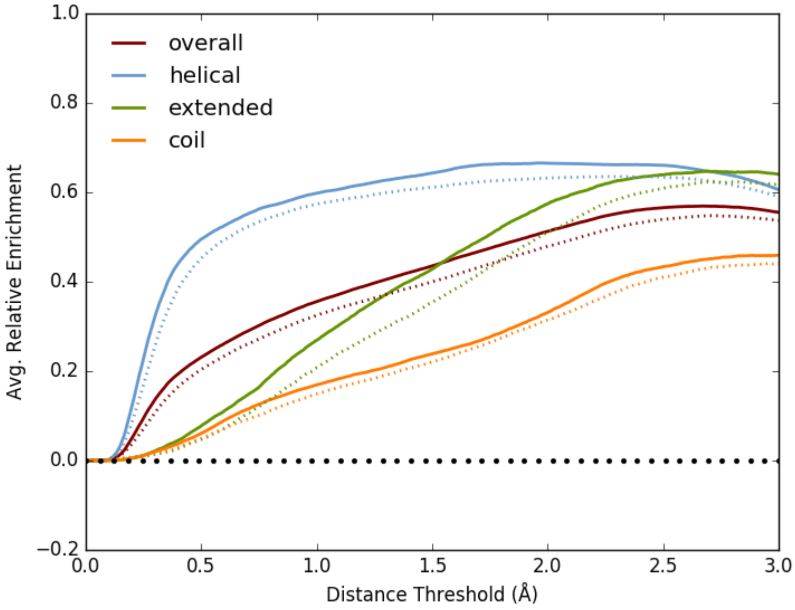
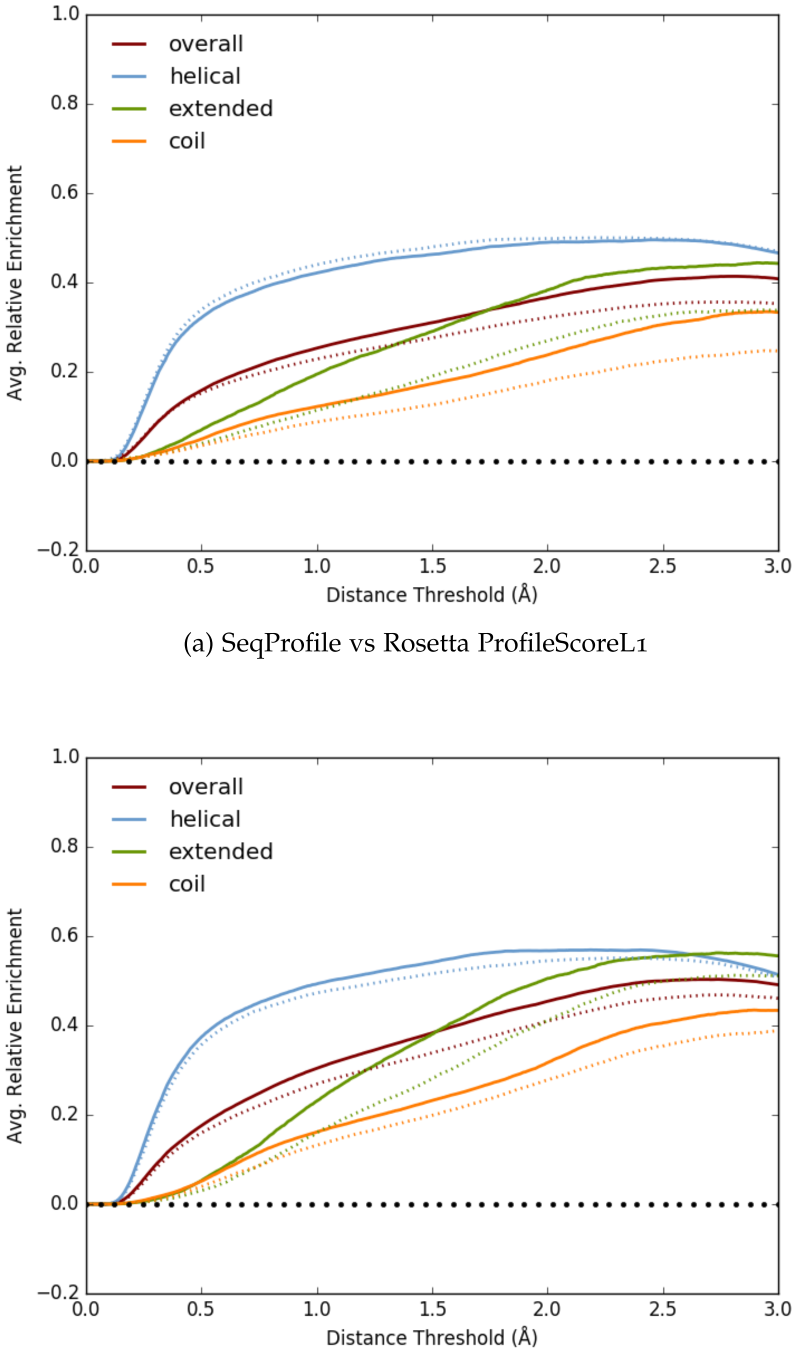


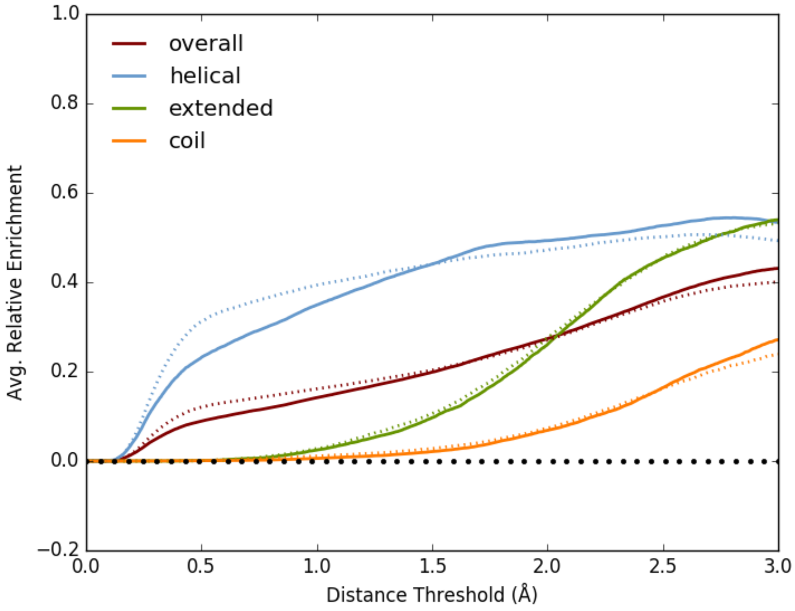
Figure 46: Comparison of fragment detection performance on fragments of length 9 relative to random. The performance of the default ProMod3 pipeline is compared to the Rosetta fragment picking protocol when only the subset of the four equivalent scores is used (ProfileScoreL1, ProfileScoreStructL1, SecondarySimilarity, RamaScore).



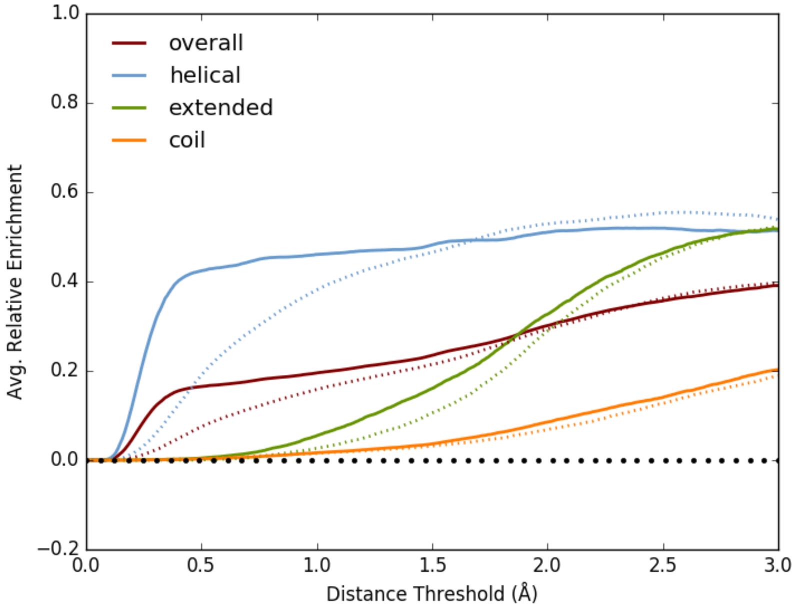
(a) SeqProfile vs Rosetta ProfileScoreL1

(b) StructProfile vs Rosetta ProfileScoreStructL1

Figure 47: Comparison of the fragment detection performance on fragments of length 9 relative to random. The performance of the two profile dependent terms used in the default fragment detection pipeline of ProMod3 (solid lines) are compared to their equivalend in the Rosetta fragment picking protocol (ProfileScoreL1, ProfileScoreStructL1; dotted lines). ProMod3 and Rosetta use exactly the same underlying mathematical formalism for scoring.



(a) SSAgreement vs Rosetta SecondarySimilarity



(b) TorsionProbability vs Rosetta RamaScore

Figure 48: Continuation of Figure 47. The performance of the SSAgreement and TorsionProbability terms used in the default fragment detection pipeline of ProMod3 (solid lines) are compared to their equivalend in the Rosetta fragment picking protocol (SecondarySimilarity, RamaScore; dotted lines)

4.6.6 *Sidechain Performance Analysis*

This section provides further details regarding sidechain modelling performance in ProMod3 and SCWRL4 with different settings. Additionally to only provide the fraction of correct χ1 angles, further analysis is performed regarding χ2 angles and average RMSD calculated on the heavy atoms of the sidechains (not including Cβ atoms).

AA	num	χ1 correct (%)	χ2 correct (%)	χ2 correct given χ1 (%)	avg RMSD
ARG	3638	72.54	67.92	74.04	1.97
ASN	2883	82.48	45.47	52.44	0.64
ASP	4019	81.04	58.87	68.35	0.81
CYS	1001	87.31			0.19
GLN	2512	74.12	61.86	70.73	1.15
GLU	4644	70.00	63.95	70.84	1.50
HIS	1543	84.51	48.48	52.38	0.90
ILE	3968	95.16	84.48	86.52	0.28
LEU	6558	87.28	84.54	94.55	0.35
LYS	3901	73.14	73.98	77.95	1.16
MET	1410	79.22	70.78	79.59	0.77
PHE	2717	93.85	85.94	88.43	0.62
PRO	3233	81.13	79.77	98.25	0.13
SER	4107	68.10			0.33
THR	3790	89.34			0.21
TRP	979	89.38	76.71	83.89	1.00
TYR	2346	92.33	83.33	86.47	0.74
VAL	5019	92.69			0.25

Table 15: SCWRL4 test set - ProMod3 flexible rotamer model (FRM) with optimal subrotamer selection (default).

AA	num	χ1 correct (%)	χ2 correct (%)	χ2 correct given χ1 (%)	avg RMSD
ARG	3638	72.51	70.09	75.89	1.98
ASN	2883	82.34	45.33	52.32	0.65
ASP	4019	81.04	58.90	68.13	0.81
CYS	1001	87.11			0.19
GLN	2512	74.44	61.86	70.37	1.15
GLU	4644	70.00	65.40	72.29	1.50
HIS	1543	84.25	48.48	52.08	0.91
ILE	3968	95.04	84.45	86.53	0.28
LEU	6558	87.13	84.58	94.52	0.35
LYS	3901	73.21	74.03	77.94	1.16
MET	1410	79.57	71.28	79.50	0.78
PHE	2717	92.09	85.17	87.65	0.67
PRO	3233	81.13	79.77	98.25	0.13
SER	4107	68.01			0.33
THR	3790	89.26			0.21
TRP	979	87.13	75.49	82.53	1.08
TYR	2346	90.66	83.21	86.27	0.80
VAL	5019	92.67			0.25

Table 16: SCWRL4 test set - ProMod3 flexible rotamer model (FRM, no subrotamer selection).

AA	num	χ1 correct (%)	χ2 correct (%)	χ2 correct given χ1 (%)	avg RMSD
ARG	3638	70.62	68.55	74.58	2.05
ASN	2883	81.27	43.36	50.49	0.67
ASP	4019	79.07	56.21	66.39	0.88
CYS	1001	87.41			0.19
GLN	2512	74.28	61.46	69.45	1.16
GLU	4644	68.11	64.06	71.07	1.56
HIS	1543	83.15	44.13	47.62	0.96
ILE	3968	94.53	83.92	85.98	0.30
LEU	6558	86.29	83.42	94.06	0.37
LYS	3901	72.16	73.39	77.16	1.18
MET	1410	79.57	70.21	78.79	0.81
PHE	2717	90.76	81.82	85.48	0.74
PRO	3233	80.30	78.94	98.23	0.14
SER	4107	67.71			0.33
THR	3790	89.10			0.22
TRP	979	86.01	72.32	79.57	1.18
TYR	2346	89.05	79.54	84.20	0.90
VAL	5019	92.13			0.26

Table 17: SCWRL4 test set - ProMod3 rigid rotamer model (no subrotamers, RRM).

AA	num	χ1 correct (%)	χ2 correct (%)	χ2 correct given χ1 (%)	avg RMSD
ARG	3638	73.53	70.40	76.34	1.96
ASN	2883	81.10	47.35	55.22	0.68
ASP	4019	81.19	60.34	69.94	0.83
CYS	1001	87.21			0.20
GLN	2512	74.80	62.78	71.26	1.15
GLU	4644	68.60	65.35	73.20	1.51
HIS	1543	83.15	45.37	49.26	0.96
ILE	3968	95.11	84.45	86.38	0.27
LEU	6558	87.19	84.93	94.84	0.36
LYS	3901	73.29	73.88	77.51	1.17
MET	1410	80.57	71.42	78.79	0.79
PHE	2717	91.24	84.95	87.78	0.76
PRO	3233	80.58	79.18	98.20	0.14
SER	4107	68.79			0.33
THR	3790	89.31			0.21
TRP	979	86.62	75.18	81.84	1.13
TYR	2346	89.86	82.48	86.34	0.89
VAL	5019	92.61			0.26

Table 18: SCWRL4 test set - SCWRL4 flexible rotamer model (FRM).

AA	num	χ1 correct (%)	χ2 correct (%)	χ2 correct given χ1 (%)	avg RMSD
ARG	3638	71.94	68.00	73.67	2.02
ASN	2883	80.71	46.72	54.83	0.68
ASP	4019	79.37	57.55	67.90	0.88
CYS	1001	87.41			0.20
GLN	2512	74.04	62.98	71.13	1.16
GLU	4644	67.38	64.15	71.94	1.55
HIS	1543	83.41	42.51	46.15	0.97
ILE	3968	94.76	83.74	85.93	0.28
LEU	6558	86.38	83.93	94.39	0.38
LYS	3901	72.19	72.88	76.70	1.19
MET	1410	79.08	70.64	78.83	0.83
PHE	2717	90.84	82.08	85.33	0.77
PRO	3233	79.71	78.26	98.10	0.15
SER	4107	67.86			0.34
THR	3790	88.76			0.22
TRP	979	85.90	68.74	76.10	1.23
TYR	2346	88.53	80.18	85.03	0.97
VAL	5019	92.09			0.27

Table 19: SCWRL4 test set - SCWRL4 rigid rotamer model (no subrotamers, RRM).

4.6.7 *Homology Modelling Performance - More Details*

When using default settings, ProMod3 is clearly betther than MODELLER in both, the lDDT score as well as MolProbity overall score (Figure 42). The MolProbity overall score is a composition score based on three contributors representing clashes, Ramachandran- and rotamer outliers. An analysis on the single scores suggests that MODELLER has issues in resolving clashes when running with default settings (Figure 49; avg. MolProbity clash scores: ProMod3: 6.9, MODELLER: 88.9). Also regarding rotamer outliers, a significant difference can be observed (Figure 50a; avg. Molprobity rotamer outlier scores: ProMod3: 1.8, MODELLER: 3.5), whereas the Ramachandran outliers are very similar (Figure 50b; avg. MolProbity rotamer outliers: ProMod3: 2.1, MODELLER: 1.8). The obvious approach to improve stereochemistry in MODELLER is to increase the refinement. This has been achieved by remodelling the full test set with the md_level variable in the automodel class set to refine.very_slow. Despite small improvements, the stereochemical issues persist (Figure 51, MODELLER averages: clash score: 81.7, rotamer outlier score: 2.6, Ramachandran outlier score: 1.6).

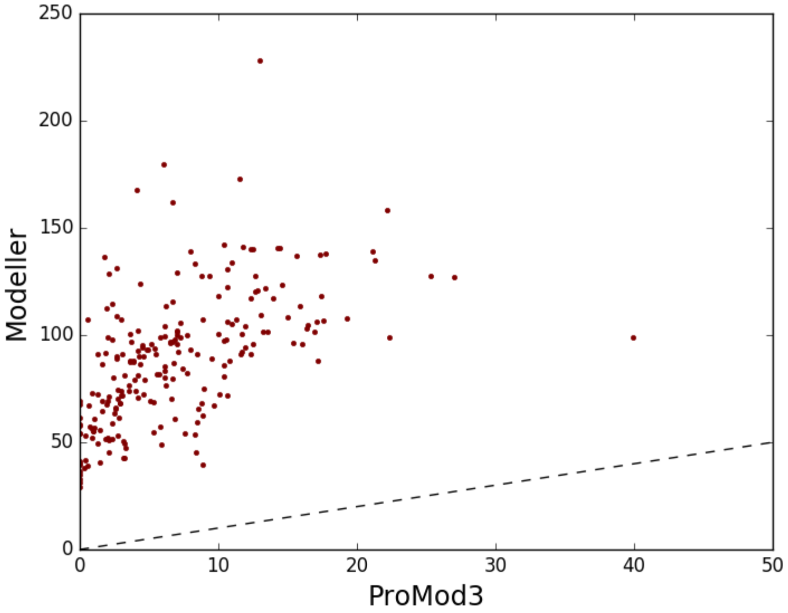
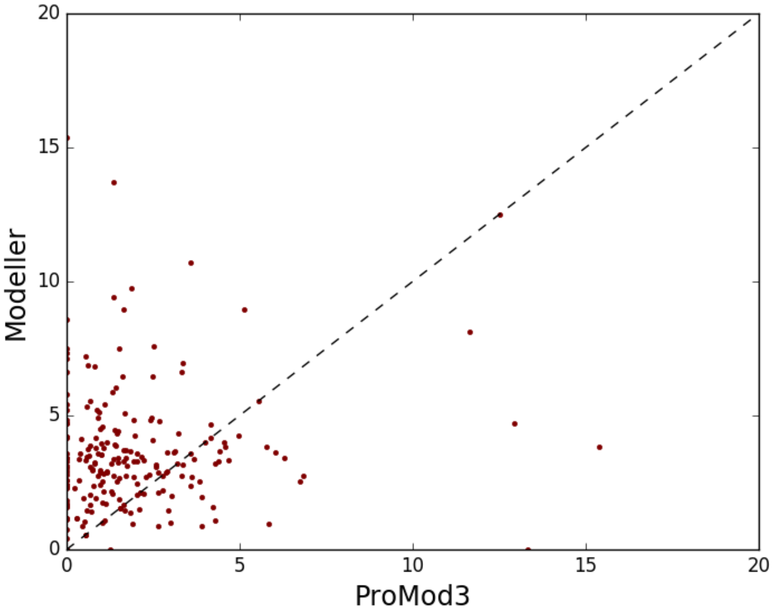
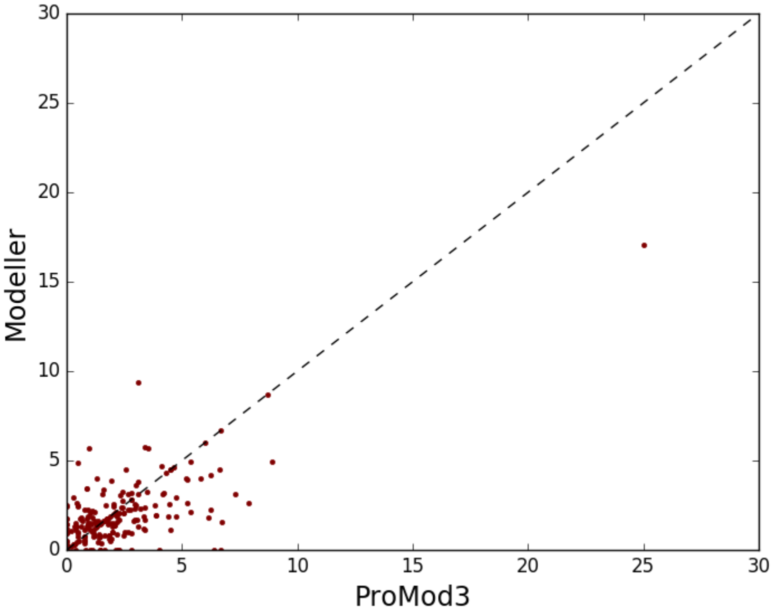


Figure 49: Comparison of MolProbity clash scores on models built with ProMod3 / MODELLER (default settings) using the same input. Every dot represents two models of the same target. If it lies on the dashed line, the scores are equal.



(a) Rotamer Outlier Scores



(b) Ramachandran Outlier Scores

Figure 50: Continuation from Figure 49, comparing the MolProbity rotamer / ramachandran outlier scores.

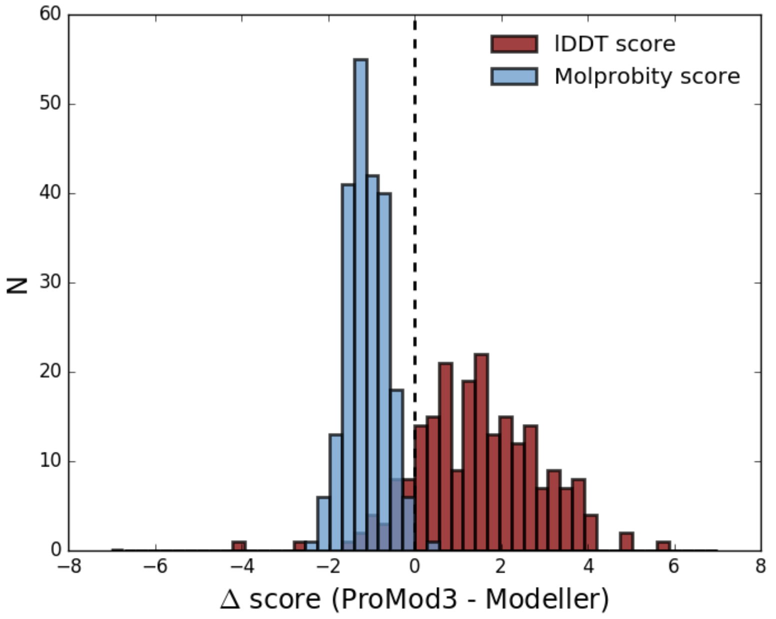


Figure 51: Comparison of overall modelling performance on models built with ProMod3 vs MODELLER (with md_level set to refine.very_slow in the automodel class) using the same input data.

4.6.8 *Speed Benchmarks*

4.6.8.1 *Setup*

All speed benchmarks have been performed on the exactly same hardware:

- CPU: Intel i7-6600U 2.60GHz
- Memory: 16GB DDR4 2133 MHz

For all benchmarks, ProMod3 has been built with GCC 5.4 and all optimizations turned on. (according cmake Flags: -DOPTIMIZE=1, -DENABLE_SSE=1)

4.6.8.2 *Fragment Detection Speed Benchmark*

The runtime is compared with Rosetta 3.7 compiled with GCC 5.4 and all optimizations turned on. The used test set contains five chains of the test used to measure the fragment detection performance. The pdb id, chain id and number of fragments are:

- 3QWW, A, 425 fragments of length 9
- 4HDH, A, 631 fragments of length 9
- 1WG8, A, 277 fragments of length 9
- 1DEV, B, 33 fragments of length 9
- 2ODI, A, 230 fragments of length 9

This gives a total of 1596 fragments of length 9. In one run, the top 100 fragments of length 9 are searched for all possible positions in the target chains. To only consider the raw search performance, all required input (profiles, PSIPRED predictions, etc.) is provided.

The average timings over three independent runs on the same hardware are:

- ProMod3: 1751 s
- Rosetta 3.7: 10446 s

This gives the reported speedup of ~6.0x in favour of ProMod3.

4.6.8.3 *Sidechain Modelling Speed Benchmark*

ProMod3 is compared to the distributed binary of SCWRL4. Computation time is measured as the average over three independent runs on the full test set used to measure sidechain modelling accuracy. The observed timings are:

- ProMod3 with subrotamers (FRM) and post-processing: 555 s
- ProMod3 no subrotamers (RRM): 109 s
- SCWRL4 with subrotamers (FRM): 1563 s
- SCWRL4 no subrotamers (RRM): 495 s

This gives the reported speedups of 2.8x in case of FRM and 4.5x in case of RRM.

4.6.8.4 *Homology Modelling Speed Benchmark*

ProMod3 is compared to MODELLER 9.17 from the officially distributed Linux (Debian/Ubuntu) package. Computation time is measured as the average over three independent runs on the full test set used to measure homology modelling accuracy. The observed timings are:

- ProMod3: 1605 s
- MODELLER default settings: 2196 s
- MODELLER increased refinement: 11548 s

This gives the reported speedups of 1.4x and 7.2x.

5 MULTITEMPLATE MODELLING

Motivation: Template detection is the first step in a homology modelling procedure. Especially for well characterized protein families, hundreds of found templates is no exception. While most homology modelling methods select one single template for modelling, multiple templates might contain complementary information. If combined with clever strategies, model accuracy could be improved.

Results: The default homology modelling pipeline in ProMod3 has been extended to take several templates as input. The proposed algorithms are mainly targeted at increasing the structural coverage of the target sequence and have extensively been tested and compared to other state-of-the-art homology modelling methods.

5.1 INTRODUCTION

The first step of basic homology modelling typically consists of finding experimentally determined structures that are homologous to the sequence to be modelled. One of them is selected to serve as template for the subsequent modelling steps. Despite often being redundant, structural information from multiple templates can, in some cases, be complementary and the modelling procedure potentially benefits from the added information from alternative templates [93]. Examples include adding structural information towards N-/C-terminus, i.e. to increase coverage, or to exploit alternative local conformations.

One form of multitemplate modelling is to represent template information as internal constraints, as it is implemented in the MODELLER [158] software. MODELLER then uses a maximum likelihood approach to generate 3D-coordinates maximizing the agreement with the input constraints. Such internal constraints can also be combined to include information from several sources and then be fed into exactly the same maximum likelihood workflow to obtain a final model. However, contradicting template information has been found to be a major problem. A possible solution is to start with a single template, the "seed", and only add additional templates if their structural information is consistent. Two publicly available methods following this approach are HHPred [113] and RaptorX [128].

Another form of multitemplate modelling use fragments extracted from various templates. This gives two benefits. First, a model with high coverage can be generated by fragment assembly. Second, local structural variations can be explored by fragment replacement. Two prominent representatives of this approach are ITasser [167] and Robetta [151].

The aim of this chapter is to develop an approach to increase the model coverage by exploiting structural information from various templates with ProMod3. To avoid detrimental effects from inconsistent internal constraints, fragment assembly will be used instead. The performance increase compared to the default SWISS-MODEL homology modelling pipeline in ProMod3 will be analysed and discussed in the context of the CAMEO continuous evaluation platform [60].

5.2 MATERIALS & METHODS

5.2.1 *The SWISS-MODEL pipeline*

1. **Template Search:** The SWISS-MODEL pipeline runs both BLAST [6] and HHblits [134] against the SWISS-MODEL template library [22] (SMTL) in order to detect protein structures homologous to the target sequence. The SMTL is a curated copy of the protein data bank [16], annotated for the needs of homology modelling with SWISS-MODEL.

2. **Template Ranking:** The identified templates are ranked based on properties extracted from their alignment to the target sequence. The relevant score used is GMQE (**G**lobal **M**odel **Q**uality **E**stimate) [19].

3. **Template Filtering:** In some cases, the number of potential templates reaches several thousand. The goal of template filtering is to remove redundancy and only use a subset for further processing.

4. **Template Selection:** The filtered templates are presented to the user for manual selection. In the case of fully automated modelling, the list of already filtered templates is further reduced. The goal is to cover as much as possible

of the target sequence and to capture structural diversity [23].

5. **Model Building:** Build a model for every selected template using ProMod3.
6. **Model Selection:** The GMQE used in the template selection step gets updated using the QMEAN4 score (Section 3.2.3) from the built models. All built models are ranked according the updated GMQE and presented to the user. If the modelling task has been submitted by CAMEO, the finally returned model is selected according to an additional scoring step incorporating consensus information from all built models [19].

5.2.2 *Rigid Blocks Algorithm*

The motivation of the the rigid blocks algorithm is to identify structurally consistent regions in two protein structures - rigid blocks. The first and most important usage of rigid blocks is to define viable anchor points to connect structural fragments in assembly procedures. they furthermore allow to define alternative local conformations of structural stretches not being part of any rigid block, in case both of their stems are part of the same rigid block. In our case there is no need to define biologically relevant domains as is the goal in published methods such as DynDom [64]. Another viable option to achieve this task would be to use the graph based Domain-Find algorithm [19] that aims to detect and cluster residues with matching structural environments. Despite its elegance, it introduces unnecessary complexity to the problem, hence the usage of a simple and fast alternative.

A protein structure is represented by its atoms that have a defined position in 3D space. If there are two protein structures with known correspondence for every atom in the first structure towards the second, their similarity can be expressed as the root mean square deviation (RMSD):.

$$\mathrm{RMSD}(A, B) = \sqrt{\frac{1}{N}\sum_{i} d(a_i, b_i)^2} \tag{30}$$

with $d(a_i, b_i)$ being the Cartesian distance between atom i in the two proteins A and B. For obvious reasons this formalism is only meaningful when the two proteins are optimally

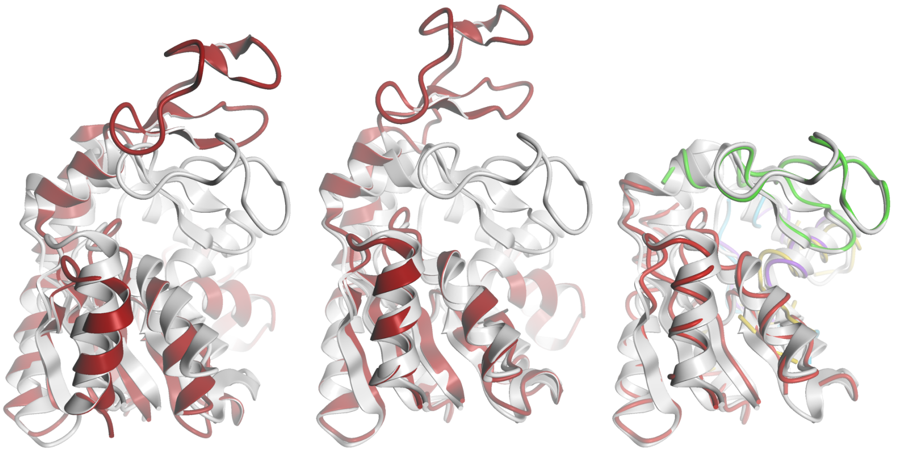


Figure 52: Adenylate kinase in its open form (pdb id: 4ake) superposed onto inhibitor bound form (pdb id: 1ake; white) using three different algorithms discussed in this chapter. From left to right: Kabsch, iterative Kabsch and rigid blocks.

superposed. This can be done by first representing the atoms of the two proteins as lists of positions and directly performing a translation to move their geometric centre to the origin of the coordinate system. In a second step, the Kabsch algorithm returns the optimal rotation matrix leading to the minimal possible RMSD [78]. Following this procedure therefore not only results in the minimal possible RMSD between two proteins, but also the corresponding transformation in space.

For many applications, the transformation resulting from the Kabsch algorithm is sufficient, but it is highly susceptible to outlier positions and cannot deal with hinge-/domain movements (Figure 52). One possibility to reduce this effect is to perform an initial superposition and introduce a distance threshold. From the two position lists originating from the input structures, two subsets can be created by gathering all positions with pairwise distances below that threshold. The two subsets are then used as input for another round of the Kabsch algorithm and two new subsets are generated based on the distance threshold. This procedure is iteratively applied until the subsets converge or a maximum number of iterations is reached. The iterative superposition largely diminishes the effect of outlier positions but in case of a hinge-/domain movement, it behaves unpredictably. Despite being deterministic, the converged solution depends on the initial superposition, which is not optimal. Typically it just converges towards the largest subsets which are structurally consistent (Figure 52). This is exactly what is required in this work. But instead of only having one consistent subset (from now on called a rigid block), we want several of them to identify regions of consistent structure. That is where the rigid blocks algorithm comes in.

The algorithm requires two protein structures and a sequence alignment for the residue-residue mapping as input. Two lists of positions with size L are generated by only extracting the $C\alpha$ positions of the L aligned residue pairs. Instead of using the full lists as starting point of the iterative superposition, a sliding window of length l is defined. This gives $L - l + 1$ consecutive slices that can be extracted from the initial position lists. They all serve as starting point for an iterative superposition and all unique rigid blocks emerging from this procedure are stored. Note, that they are potentially overlapping or very similar. The algorithm thus takes a threshold parameter as input. Two rigid blocks are merged, if their fraction of matching elements is above that threshold.

The underlying calculations are extremely efficient by not using the original Kabsch algorithm for minimum RMSD superposition anymore but the more efficient quaternion characteristic polynomial method [100]. The most expensive computation reduces to calculating a covariance matrix between the position lists [61]. Using Eigen matrices [57] for this task allows to exploit sophisticated vectorization techniques of modern C++ compilers and reduces the runtime to only a few milliseconds for a medium sized problem.

5.2.3 *Coverage Extension Algorithm*

The multitemplate modelling algorithm presented in this section replaces step 5 in the SWISS-MODEL pipeline and relies on extending coverage using two different strategies:

1. overlapping extension strategy
2. linker sampling extension strategy

For every template a model is built. The template serves as seed and all other templates are considered to be alternative templates. Starting from the seed, the algorithm tries to iteratively extend every chain by 1. and then 2. using the alternative templates until nothing happens anymore.

5.2.3.1 *Overlapping Extension Strategy*

Every alternative template of the current chain in the seed is checked for increased coverage towards the N- or C-terminus and whether there is a structural overlap. The rigid blocks algorithm identifies structurally consistent regions in that overlap and uses the rigid block with the shortest distance d to the current terminus as a superposition anchor (if d is smaller than 30 residues). Every superposable extension is subject to the default ProMod3 loop closing procedure (Section 4.2.5.3) to ensure a continuous amino acid chain and is finally checked for clashes with the seed. Only if potential clashes are close to the joining region, will a Monte Carlo procedure try to resolve them. All non-clashing extensions are now subject to a final scoring using the default loop scoring procedure in ProMod3 (algorithm 4.2.5.3) and the best extension for each terminus is added to the current chain in the seed.

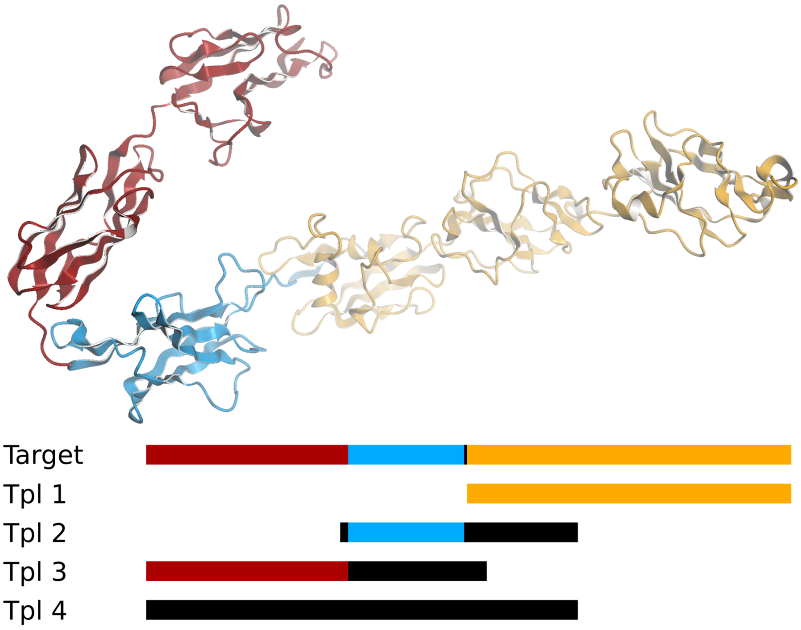


Figure 53: Server 49 model of CAMEO target 5ftx. Horizontal bars indicate the target coverage for each of the four input templates and the colouring their actual contribution to the final model. The relative orientation is based on the overlap strategy.

5.2.3.2 *Linker Sampling Extension Strategy*

Every alternative template of the current chain in the seed is checked to see whether it increases coverage towards the N- or C-terminus. The found extensions are extracted and undergo the default ProMod3 loop closing procedure (Section 4.2.5.3) to ensure a continuous amino acid chain. The optimal relative orientation is then determined by fragment replacement. A total of 1000 fragments are searched using the default ProMod3 fragment detection pipeline (Section 4.2.5.2) for the range not covered (gap between extension and terminus). If there is no gap between the extension and the associated terminus in the current chain, the extension is shortened to have a fragment length of at least 3 residues. The relative orientation of each extension is determined by inserting each of the found fragments and scoring the full extension in its resulting relative orientation using the default loop scoring procedure in ProMod3 (algorithm 4.2.5.3). Having done that for all extensions, a final scoring step determines and directly adds the optimal extension for each terminus in the current chain of the seed.

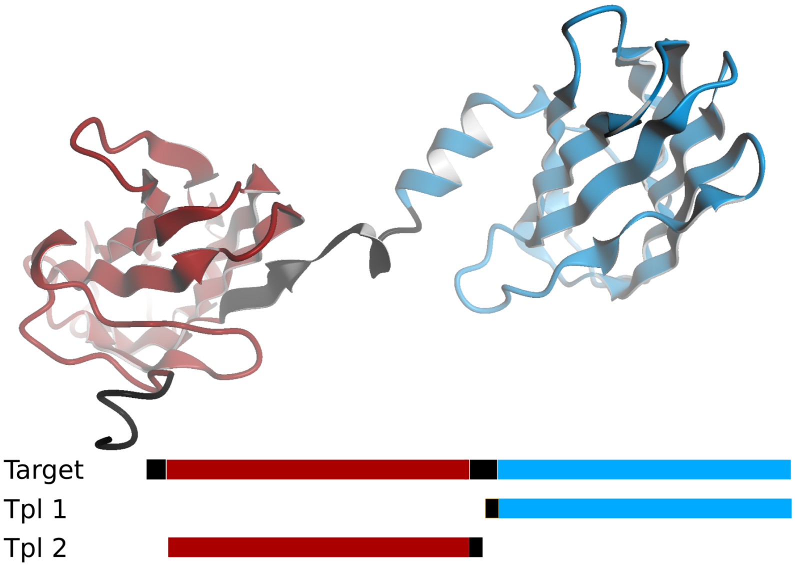


Figure 54: Server 49 model of CAMEO target 5iku. Horizontal bars indicate the target coverage for the two input templates and the colouring their actual contribution to the final model. The relative orientation is based on the linker sampling strategy.

5.2.4 *Testing Strategies*

To test the impact on modelling performance when using the simple coverage extension, we registered three testing servers to CAMEO.

- **server 54:** Baseline to evaluate the modelling performance when using the default SWISS-MODEL pipeline based on the ProMod3 modelling engine as described in Section 5.2.1.

- **server 55:** Exactly the same as server 54 but enforcing a 100% target sequence coverage. Each terminus not covered by the used template is modelled by performing 20 fragment based Monte Carlo runs with 5000 Monte Carlo steps each. The termini are likely to be of low quality but allow to evaluate the effect of increased coverage on coverage dependent target scores. The reason for that extension is the observation that many servers registered to CAMEO always return models with 100% target sequence coverage.

- **server 49:** Same setup as for the other two servers. But the coverage extension algorithm described in section Section 5.2.3 is used for model building. For a direct comparison to server 55, the same termini modelling strategy is applied to each model to enforce 100% target sequence coverage.

The performance of all three servers is evaluated on 80 models submitted by CAMEO in the period of one month (2017.03.11 - 2017.04.01), using the all atom based superposition independent lDDT score [109] as target value. As an additional evaluation, the models of server 49 are directly compared to other widely used modelling services employing multitemplate modelling techniques (Robetta [151], HHPred [113], RaptorX [128] and IntFold4-TS [111]).

5.3 RESULTS

5.3.1 *SWISS-MODEL Server Evaluation*

A direct comparison reveals that server 55 always produces equally good or better results than server 54 in terms of lDDT score (Figure 55). This is purely a result of adding low quality termini if they're not covered by the underlying template as illustrated in Figure 56. This is expected as the lDDT score estimates the fraction of differences in interatomic distances below a certain threshold. This fraction can only go up as there is no punishment for badly modelled interatomic interactions. The authors of lDDT already discussed this effect by estimating a nonzero baseline for random structures [109]. Despite not necessarily being useful, the (close to) random termini are necessary for a direct comparison to the top performing predictors registered to CAMEO since they almost always return models with full coverage (Table 20). In some cases, the models of server 49 experience a significant improvement of lDDT score relative to server 55.

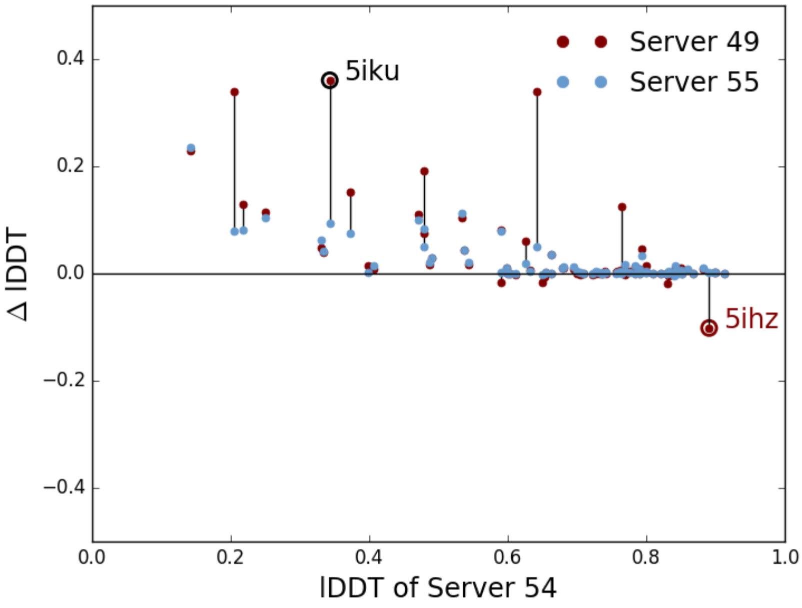


Figure 55: Per model comparison with server 54 as reference. CAMEO target 5iku will further be discussed in Figure 56. CAMEO target 5ihz represents a fail of server 49 due to its template library exceptionally not being up to date.

This can largely be attributed to adding full domains and confirms the observations of Larsson et al. [93]. They state that coverage extension is the main source of improvements in multitemplate modelling. That is exactly what we do in server 49 with the remaining challenge being to improve relative domain orientations, especially when no overlap is present in the underlying templates. One case of lower lDDT compared to server 55, CAMEO target 5ihz, is marked in Figure 55. The reason was that the template library of server 49 was exceptionally not up to date, which caused an obvious template to be missed.

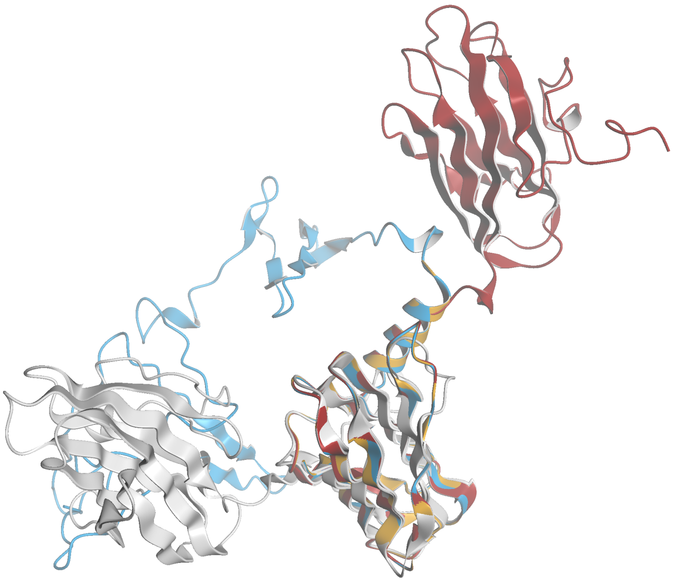


Figure 56: Alternative models of CAMEO target 5iku. White: target structure, orange: server54, blue: server 55, red: server 49

5.3.2 *Comparison to Other Multitemplate Methods*

Server 49 is highly competitive compared to state-of-the-art multitemplate modelling methods (Table 20). Even though all SWISS-MODEL related servers return models in a fraction of the runtime of all other servers, server 49 is only significantly outperformed by Robetta in terms of lDDT. This is not true anymore when considering binding sites. The heavy refinement strategies of Robetta seem to be detrimental to these, often more conserved parts of a protein model. For a more detailed

Method	Returned Models	Runtime (h:min:s)	Coverage	lDDT	lDDT Binding Site
server 54	80	00:09:33	0.87	0.6737	0.7588
server 55	80	00:19:39	1.0	0.6942	0.7842
server 49	80	00:16:41	1.0	0.7076	0.7823
Robetta	79	34:04:36	0.99	0.7306	0.7612
HHPred	79	24:21:09	0.97	0.6679	0.7413
IntFold4-TS	79	17:34:26	1.0	0.7079	0.8011
RaptorX	80	09:35:36	0.95	0.6938	0.7429

Table 20: Per server averages on modelled targets. 32 of the 80 targets contain a ligand classified as relevant. They contribute to lDDT binding site and have been modelled by all methods.

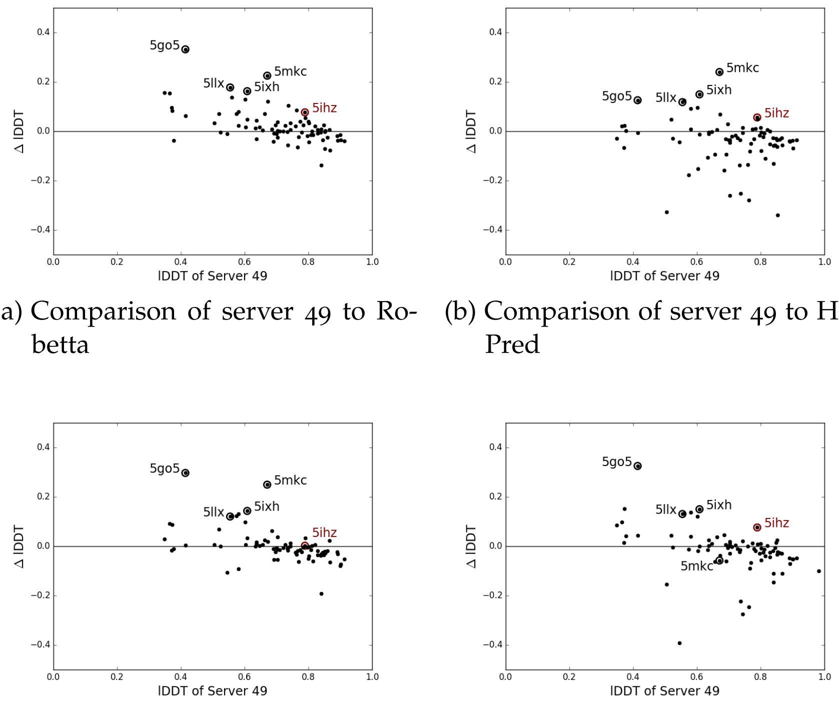
analysis, the comparison is extended on a per model basis (Figure 57). If server 49 generates a high quality model (lDDT > 0.70), it rarely gets outperformed by any of the other methods. The situation changes for the lower quality range, where especially Robetta is capable to consistently outperform server 49 (Figure 57a). Since we ruled out the influence of coverage to overall performance, the cases where Robetta performs better than server 49 must either come from successful improvement over the best available single templates, or flaws in the server 49 modelling pipeline. In order to identify cases of the latter option, we analyze the 10 targets where Robetta shows the largest improvements compared to server 49 manually. Additionally to the already discussed template library issue (CAMEO target 5ihz), 4 of the 10 cases are a result of flaws in the current SWISS-MODEL pipeline and not the multitemplate modelling algorithm itself. The corresponding models consistently stick out in all comparisons. It's therefore not something Robetta does particularly well, but rather something server 49 does wrong:

1. **5g05:** Known issue in final consensus based scoring step to select model returned to CAMEO. A faulty normalization tends to prefer oligomers. In this case a low quality dimer instead of a monomer with much higher quality (lDDT score 0.414 vs 0.7487) has been selected. Correct stoichiometry would have been monomeric.

2. **5mkc:** The obvious template (sequence identity > 90%) contained an insertion comprising a full domain with ~350 residues. The SWISS-MODEL pipeline identified both, the part towards C-terminus and N-terminus around that insertion as single templates.

3. **5llx:** A small domain with an obvious template has been neglected due to the template filtering step in the SWISS-MODEL pipeline.

4. **5ixh:** Same issue as 5g05. A low quality dimer instead of a monomer with much higher quality (lDDT score 0.6081 vs 0.698) has been selected in the final consensus based scoring step. The correct stoichiometry would have been monomeric.



(a) Comparison of server 49 to Robetta

(b) Comparison of server 49 to HH-Pred

(c) Comparison of server 49 to IntFold4-TS

(d) Comparison of server 49 to RaptorX

Figure 57: Direct comparison of server 49 with 4 servers employing multitemplate modelling. In addition to the model marked red (see Figure 55), 4 more models are specifically marked and will further be discussed in Section 5.3.2.

5.4 DISCUSSION

This chapter uses the current SWISS-MODEL pipeline and introduces relatively small changes to incorporate information from multiple templates into one model. The biggest success comes from adding full additional domains to a seed template, with the remaining challenge being to improve the relative orientation. This largely agrees with the findings of Larsson et al. [93]. They stated that most of the improvements from combining multiple templates can be expected by extending the coverage of the target sequence. Another finding is the fact that even extending coverage by low quality termini increases performance in terms of coverage dependent scores such as lDDT. Even though this does not add any valuable information to a protein model, this effect must be considered when comparing towards methods that always return full coverage models. A final analysis using publicly available data from CAMEO, showed that the presented multitemplate pipeline can generate competitive models at a fraction of the runtime of all other methods. Room for improvement has mainly been identified in the overall SWISS-MODEL pipeline and not the multitemplate modelling algorithm itself. We are therefore confident to further improve modelling performance by tackling the discussed problems in the SWISS-MODEL pipeline. This would further strengthen our position of generating high quality models in very little runtime.

6 ACKNOWLEDGMENTS

First of all I want to acknowledge the generous support I received through a PhD fellowship of the SIB by the 'Swiss Foundation for Excellence and Talent in Biomedical Research'. Furthermore I would like to thank my supervisor Torsten Schwede who runs a lab of modelling enthusiasts and seems to have an answer for everything. Also a great thanks goes to my PhD committee consisting of Timm Maier and Matteo dal Peraro. The whole Schwede group was a great help during my struggle as a PhD student. It is difficult to thank all the people as they deserve it, so I do that in globo: THANKS! Well, OK, some people really need to be mentioned separately since they had a direct impact on the writing of this thesis. Jürgen Haas for his efforts to keep me on track, correction works and trafficking of delicious Italian specialities against even more delicious chocolate from central Switzerland. Tjaart de Beer for nit-picking but super useful corrections to get many details of this thesis right. What would a student be without a family behind! A big thank you also goes to Schwyz. It was always a warm welcome, lots of support and healthy food to recover from long periods of thai and pizza. And no, I did not forget you. Thank you so much Lara. This thesis would not be the same without you!

BIBLIOGRAPHY

[1] C. Abergel. "Molecular replacement: tricks and treats." In: *Acta Crystallogr. D Biol. Crystallogr.* 69.Pt 11 (2013), pp. 2167–2173.

[2] G. K. Ackers and J. M. Holt. "Asymmetric cooperativity in a symmetric tetramer: human hemoglobin." In: *J. Biol. Chem.* 281.17 (2006), pp. 11441–11443.

[3] P. D. Adams, P. V. Afonine, R. W. Grosse-Kunstleve, R. J. Read, J. S. Richardson, D. C. Richardson, and T. C. Terwilliger. "Recent developments in phasing and structure refinement for macromolecular crystallography." In: *Curr. Opin. Struct. Biol.* 19.5 (2009), pp. 566–572.

[4] B. Alberts. *Molecular Biology of the Cell: Reference edition*. Molecular Biology of the Cell: Reference Edition Bd. 1. Garland Science, 2008. ISBN: 9780815341116.

[5] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. "Basic local alignment search tool." In: *J. Mol. Biol.* 215.3 (1990), pp. 403–410.

[6] Stephen F. Altschul, Thomas L. Madden, Alejandro A. Schäffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman. "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs." In: *Nucleic Acids Research* 25.17 (1997), p. 3389.

[7] C. B. Anfinsen. "Principles that govern the folding of protein chains." In: *Science* 181.4096 (1973), pp. 223–230.

[8] X. C. Bai, G. McMullan, and S. H. Scheres. "How cryo-EM is revolutionizing structural biology." In: *Trends Biochem. Sci.* 40.1 (2015), pp. 49–57.

[9] D. Baker and A. Sali. "Protein structure prediction and structural genomics." In: *Science* 294.5540 (2001), pp. 93–96.

[10] P. I. de Bakker, M. A. DePristo, D. F. Burke, and T. L. Blundell. "Ab initio construction of polypeptide fragments: Accuracy of loop decoy discrimination by an all-atom statistical potential and the AMBER force field with the Generalized Born solvation model." In: *Proteins* 51.1 (2003), pp. 21–40.

[11] A. Barbato, P. Benkert, T. Schwede, A. Tramontano, and J. Kosinski. "Improving your target-template alignment with MODalign." In: *Bioinformatics* 28.7 (2012), pp. 1038–1039.

[12] A BenNaim. "Statistical potentials extracted from protein structures: Are these meaningful potentials?" In: *JOURNAL OF CHEMICAL PHYSICS* 107.9 (1997), pp. 3698–3706.

[13] P. Benkert, M. Biasini, and T. Schwede. "Toward the estimation of the absolute quality of individual protein structure models." In: *Bioinformatics* 27.3 (2011), pp. 343–350.

[14] P. Benkert, S. C. Tosatto, and D. Schomburg. “QMEAN: A comprehensive scoring function for model quality assessment.” In: *Proteins* 71.1 (2008), pp. 261–277.

[15] H. M. Berman. “The Protein Data Bank: a historical perspective.” In: *Acta Crystallogr., A, Found. Crystallogr.* 64.Pt 1 (2008), pp. 88–95.

[16] H. Berman, K. Henrick, and H. Nakamura. “Announcing the worldwide Protein Data Bank.” In: *Nat. Struct. Biol.* 10.12 (2003), p. 980.

[17] A. Bernsel, H. Viklund, J. Falk, E. Lindahl, G. von Heijne, and A. Elofsson. “Prediction of membrane-protein topology from first principles.” In: *Proc. Natl. Acad. Sci. U.S.A.* 105.20 (2008), pp. 7177–7181.

[18] D. Bhattacharya, B. Adhikari, J. Li, and J. Cheng. “FRAGSION: ultra-fast protein fragment library generation by IOHMM sampling.” In: *Bioinformatics* 32.13 (2016), pp. 2059–2061.

[19] M. Biasini. “Accurate Modeling of Protein Structures by Homology.” PhD thesis. "Biozentrum - University of Basel", 2013.

[20] M. Biasini, V. Mariani, J. Haas, S. Scheuber, A. D. Schenk, T. Schwede, and A. Philippsen. “OpenStructure: a flexible software framework for computational structural biology.” In: *Bioinformatics* 26.20 (2010), pp. 2626–2628.

[21] M. Biasini, T. Schmidt, S. Bienert, V. Mariani, G. Studer, J. Haas, N. Johner, A. D. Schenk, A. Philippsen, and T. Schwede. “OpenStructure: an integrated software framework for computational structural biology.” In: *Acta Crystallogr. D Biol. Crystallogr.* 69.Pt 5 (2013), pp. 701–709.

[22] M. Biasini et al. “SWISS-MODEL: modelling protein tertiary and quaternary structure using evolutionary information.” In: *Nucleic Acids Res.* 42.Web Server issue (2014), W252–258.

[23] S. Bienert, A. Waterhouse, T. A. de Beer, G. Tauriello, G. Studer, L. Bordoli, and T. Schwede. “The SWISS-MODEL Repository-new features and functionality.” In: *Nucleic Acids Res.* 45.D1 (2017), pp. D313–D319.

[24] W. H. Bragg and W. L. Bragg. “The Reflection of X-rays by Crystals.” In: *Proceedings of the Royal Society of London Series A* 88 (July 1913), pp. 428–438.

[25] C. Bystroff, K. T. Simons, K. F. Han, and D. Baker. “Local sequence-structure correlations in proteins.” In: *Curr. Opin. Biotechnol.* 7.4 (1996), pp. 417–421.

[26] A. A. Canutescu and R. L. Dunbrack. “Cyclic coordinate descent: A robotics algorithm for protein loop closure.” In: *Protein Sci.* 12.5 (2003), pp. 963–972.

[27] A. A. Canutescu, A. A. Shelenkov, and R. L. Dunbrack. “A graph-theory algorithm for rapid protein side-chain prediction.” In: *Protein Sci.* 12.9 (2003), pp. 2001–2014.

[28] D. A. Case, T. E. Cheatham, T. Darden, H. Gohlke, R. Luo, K. M. Merz, A. Onufriev, C. Simmerling, B. Wang, and R. J. Woods. “The Amber biomolecular simulation programs.” In: *J Comput Chem* 26.16 (2005), pp. 1668–1688.

[29] J. Chen, M. Guo, X. Wang, and B. Liu. "A comprehensive review and comparison of different computational methods for protein remote homology detection." In: *Brief. Bioinformatics* (2016).

[30] V. B. Chen, W. B. Arendall, J. J. Headd, D. A. Keedy, R. M. Immormino, G. J. Kapral, L. W. Murray, J. S. Richardson, and D. C. Richardson. "MolProbity: all-atom structure validation for macromolecular crystallography." In: *Acta Crystallogr. D Biol. Crystallogr.* 66.Pt 1 (2010), pp. 12–21.

[31] J. Cheng, A. Z. Randall, M. J. Sweredoski, and P. Baldi. "SCRATCH: a protein structure and structural feature prediction server." In: *Nucleic Acids Res.* 33.Web Server issue (2005), W72–76.

[32] Y. Choi and C. M. Deane. "FREAD revisited: Accurate loop structure prediction using a database search algorithm." In: *Proteins* 78.6 (2010), pp. 1431–1440.

[33] C. Chothia and A. M. Lesk. "The relation between the divergence of sequence and structure in proteins." In: *EMBO J.* 5.4 (1986), pp. 823–826.

[34] Y. Collette, N. Hansen, G. Pujol, D. Salazar Aponte, and R. Le Riche. "On Object-Oriented Programming of Optimizers – Examples in Scilab." In: *Multidisciplinary Design Optimization in Computational Mechanics*. Ed. by P. Breitkopf and R. F. Coelho. in print. Wiley, 2010. Chap. 14, pp. 527–565.

[35] Evangelos A. Coutsias, Chaok Seok, Michael J. Wester, and Ken A. Dill. "Resultants and loop closure." In: *International Journal of Quantum Chemistry* 106.1 (2006), pp. 176–189.

[36] M. A. DePristo, P. I. de Bakker, S. C. Lovell, and T. L. Blundell. "Ab initio construction of polypeptide fragments: efficient generation of accurate, representative ensembles." In: *Proteins* 51.1 (2003), pp. 41–55.

[37] C. M. Deane and T. L. Blundell. "CODA: a combined algorithm for predicting the structurally variable regions of protein models." In: *Protein Sci.* 10.3 (2001), pp. 599–612.

[38] H. Deng, Y. Jia, Y. Wei, and Y. Zhang. "What is the best reference state for designing statistical atomic potentials in protein structure prediction?" In: *Proteins* 80.9 (2012), pp. 2311–2322.

[39] F. DiMaio, M. D. Tyka, M. L. Baker, W. Chiu, and D. Baker. "Refinement of protein structures into low-resolution density maps using rosetta." In: *J. Mol. Biol.* 392.1 (2009), pp. 181–190.

[40] P. Eastman et al. "OpenMM 4: A Reusable, Extensible, Hardware Independent Library for High Performance Molecular Simulation." In: *J Chem Theory Comput* 9.1 (2013), pp. 461–469.

[41] R. A. Engh and R. Huber. "Accurate bond and angle parameters for X-ray protein structure refinement." In: *Acta Crystallographica Section A* 47.4 (1991), pp. 392–400.

[42] 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." In: *Structure* 17.11 (2009), pp. 1515–1527.

[43] M. Fasnacht, J. Zhu, and B. Honig. "Local quality assessment in homology models using statistical potentials and support vector machines." In: *Protein Sci.* 16.8 (2007), pp. 1557–1568.

[44] M. J. Fedor and J. R. Williamson. "The catalytic diversity of RNAs." In: *Nat. Rev. Mol. Cell Biol.* 6.5 (2005), pp. 399–412.

[45] N. Fernandez-Fuentes and A. Fiser. "Saturating representation of loop conformational fragments in structure databanks." In: *BMC Struct. Biol.* 6 (2006), p. 15.

[46] Narcis Fernandez-Fuentes, Joseph M. Dybas, and Andras Fiser. "Structural Characteristics of Novel Protein Folds." In: *PLOS Computational Biology* 6.4 (Apr. 2010), pp. 1–11.

[47] L. G. Ferreira, R. N. Dos Santos, G. Oliva, and A. D. Andricopulo. "Molecular docking and structure-based drug design strategies." In: *Molecules* 20.7 (2015), pp. 13384–13421.

[48] A. Fiser, R. K. Do, and A. Sali. "Modeling of loops in protein structures." In: *Protein Sci.* 9.9 (2000), pp. 1753–1773.

[49] D. A. Fletcher and R. D. Mullins. "Cell mechanics and the cytoskeleton." In: *Nature* 463.7280 (2010), pp. 485–492.

[50] L. R. Forrest, C. L. Tang, and B. Honig. "On the accuracy of homology modeling and sequence alignment methods applied to membrane proteins." In: *Biophys. J.* 91.2 (2006), pp. 508–517.

[51] K. Fuxe et al. "GPCR heteromers and their allosteric receptor-receptor interactions." In: *Curr. Med. Chem.* 19.3 (2012), pp. 356–363.

[52] E. F. Garman. "Developments in x-ray crystallographic structure determination of biological macromolecules." In: *Science* 343.6175 (2014), pp. 1102–1108.

[53] K. Ginalski, A. Elofsson, D. Fischer, and L. Rychlewski. "3D-Jury: a simple approach to improve protein structure predictions." In: *Bioinformatics* 19.8 (2003), pp. 1015–1018.

[54] R. F. Goldstein. "Efficient rotamer elimination applied to protein side-chains and related spin glasses." In: *Biophys. J.* 66.5 (1994), pp. 1335–1340.

[55] D. S. Goodsell and A. J. Olson. "Structural symmetry and protein function." In: *Annu Rev Biophys Biomol Struct* 29 (2000), pp. 105–153.

[56] D. Gront, D. W. Kulp, R. M. Vernon, C. E. Strauss, and D. Baker. "Generalized fragment picking in Rosetta: design, protocols and applications." In: *PLoS ONE* 6.8 (2011), e23294.

[57] G. Guennebaud, B. Jacob, et al. http://eigen.tuxfamily.org.

[58] N. Guex and M. C. Peitsch. "SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling." In: *Electrophoresis* 18.15 (1997), pp. 2714–2723.

[59] Peter Güntert. "Automated NMR Structure Calculation With CYANA." In: *Protein NMR Techniques*. Ed. by A. Kristina Downing. Totowa, NJ: Humana Press, 2004, pp. 353–378. ISBN: 978-1-59259-809-0.

[60] J. Haas, S. Roth, K. Arnold, F. Kiefer, T. Schmidt, L. Bordoli, and T. Schwede. "The Protein Model Portal–a comprehensive resource for protein structure and model information." In: *Database (Oxford)* 2013 (2013), bat031.

[61] Imran S Haque, Kyle A Beauchamp, and Vijay S Pande. "A Fast 3 x N Matrix Multiply Routine for Calculation of Protein RMSD." In: *bioRxiv* (2014).

[62] F. U. Hartl, A. Bracher, and M. Hayer-Hartl. "Molecular chaperones in protein folding and proteostasis." In: *Nature* 475.7356 (2011), pp. 324–332.

[63] M. Hauser, C. E. Mayer, and J. Soding. "kClust: fast and sensitive clustering of large protein sequence databases." In: *BMC Bioinformatics* 14 (2013), p. 248.

[64] Steven Hayward and Richard A Lee. "Improvements in the analysis of domain motions in proteins from conformational change: DynDom version 1.50." In: *Journal of Molecular Graphics and Modelling* 21.3 (2002), pp. 181 –183.

[65] J. M. Heather and B. Chain. "The sequence of sequencers: The history of sequencing DNA." In: *Genomics* 107.1 (2016), pp. 1–8.

[66] A. J. Heim and Z. Li. "Developing a high-quality scoring function for membrane protein structures based on specific inter-residue interactions." In: *J. Comput. Aided Mol. Des.* 26.3 (2012), pp. 301–309.

[67] S. Henikoff and J. G. Henikoff. "Amino acid substitution matrices from protein blocks." In: *Proc. Natl. Acad. Sci. U.S.A.* 89.22 (1992), pp. 10915–10919.

[68] P. W. Hildebrand, A. Goede, R. A. Bauer, B. Gruening, J. Ismer, E. Michalsky, and R. Preissner. "SuperLooper–a prediction server for the modeling of loops in globular and membrane proteins." In: *Nucleic Acids Res.* 37.Web Server issue (2009), W571–574.

[69] L. Holm and C. Sander. "Fast and simple Monte Carlo algorithm for side chain optimization in proteins: application to model building by homology." In: *Proteins* 14.2 (1992), pp. 213–223.

[70] R. W. Hooft, G. Vriend, C. Sander, and E. E. Abola. "Errors in protein structures." In: *Nature* 381.6580 (1996), p. 272.

[71] J. Huang, S. Rauscher, G. Nawrocki, T. Ran, M. Feig, B. L. de Groot, H. Grubmuller, and A. D. MacKerell. "CHARMM36m: an improved force field for folded and intrinsically disordered proteins." In: *Nat. Methods* 14.1 (2017), pp. 71–73.

[72] Andreas Jabs, Manfred S Weiss, and Rolf Hilgenfeld. "Non-proline Cis peptide bonds in proteins 1." In: *Journal of Molecular Biology* 286.1 (1999), pp. 291 –304.

[73] M. P. Jacobson, D. L. Pincus, C. S. Rapp, T. J. Day, B. Honig, D. E. Shaw, and R. A. Friesner. "A hierarchical approach to all-atom protein loop prediction." In: *Proteins* 55.2 (2004), pp. 351–367.

[74] Joël Janin, Shoshanna Wodak, Michael Levitt, and Bernard Maigret. "Conformation of amino acid side-chains in proteins." In: *Journal of Molecular Biology* 125.3 (1978), pp. 357 –386.

[75] D. T. Jones. "Protein secondary structure prediction based on position-specific scoring matrices." In: *J. Mol. Biol.* 292.2 (1999), pp. 195–202.

[76] S. Jones and J. M. Thornton. "Principles of protein-protein interactions." In: *Proc. Natl. Acad. Sci. U.S.A.* 93.1 (1996), pp. 13–20.

[77] Roland L Dunbrack Jr. "Rotamer Libraries in the 21st Century." In: *Current Opinion in Structural Biology* 12.4 (2002), pp. 431 –440.

[78] W. Kabsch. "A solution for the best rotation to relate two sets of vectors." In: *Acta Crystallographica Section A* 32.5 (1976), pp. 922–923.

[79] W. Kabsch and C. Sander. "Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features." In: *Biopolymers* 22.12 (1983), pp. 2577–2637.

[80] Voula Kanelis, Julie D. Forman-Kay, and Lewis E. Kay. "Multidimensional NMR Methods for Protein Structure Determination." In: *IUBMB Life* 52.6 (2001), pp. 291–302.

[81] D. A. Keedy et al. "The other 90% of the protein: assessment beyond the Calphas for CASP8 template-based and high-accuracy models." In: *Proteins* 77 Suppl 9 (2009), pp. 29–49.

[82] F. Kiefer, K. Arnold, M. Kunzli, L. Bordoli, and T. Schwede. "The SWISS-MODEL Repository and associated resources." In: *Nucleic Acids Res.* 37.Database issue (2009), pp. D387–392.

[83] D. Kihara, H. Chen, and Y. D. Yang. "Quality assessment of protein structure models." In: *Curr. Protein Pept. Sci.* 10.3 (2009), pp. 216–228.

[84] C. L. Kingsford, B. Chazelle, and M. Singh. "Solving and analyzing side-chain positioning problems using linear and integer programming." In: *Bioinformatics* 21.7 (2005), pp. 1028–1036.

[85] I. R. Kleckner and M. P. Foster. "An introduction to NMR-based approaches for measuring protein dynamics." In: *Biochim. Biophys. Acta* 1814.8 (2011), pp. 942–968.

[86] T. Kortemme, A. V. Morozov, and D. Baker. "An orientation-dependent hydrogen bonding potential improves prediction of specificity and structure for proteins and protein-protein complexes." In: *J. Mol. Biol.* 326.4 (2003), pp. 1239–1259.

[87] G. G. Krivov, M. V. Shapovalov, and R. L. Dunbrack. "Improved prediction of protein side-chain conformations with SCWRL4." In: *Proteins* 77.4 (2009), pp. 778–795.

[88] A. Krogh, M. Brown, I. S. Mian, K. Sjolander, and D. Haussler. "Hidden Markov models in computational biology. Applications to protein modeling." In: *J. Mol. Biol.* 235.5 (1994), pp. 1501–1531.

[89] A. Kryshtafovych, K. Fidelis, and A. Tramontano. "Evaluation of model quality predictions in CASP9." In: *Proteins* 79 Suppl 10 (2011), pp. 91–106.

[90] A. Kryshtafovych, A. Barbato, K. Fidelis, B. Monastyrskyy, T. Schwede, and A. Tramontano. "Assessment of the assessment: evaluation of the model quality estimates in CASP10." In: *Proteins* 82 Suppl 2 (2014), pp. 112–126.

[91] A. Kryshtafovych, A. Barbato, B. Monastyrskyy, K. Fidelis, T. Schwede, and A. Tramontano. "Methods of model accuracy estimation can help selecting the best models from decoy sets: Assessment of model accuracy estimations in CASP11." In: *Proteins* (2015).

[92] I. Kufareva, M. Rueda, V. Katritch, R. C. Stevens, and R. Abagyan. "Status of GPCR modeling and docking as reflected by community-wide GPCR Dock 2010 assessment." In: *Structure* 19.8 (2011), pp. 1108–1126.

[93] P. Larsson, B. Wallner, E. Lindahl, and A. Elofsson. "Using multiple templates to improve quality of homology models in automated homology modeling." In: *Protein Sci.* 17.6 (2008), pp. 990–1002.

[94] R. A. Laskowski, M. W. MacArthur, D. S. Moss, and J. M. Thornton. "*PROCHECK*: a program to check the stereochemical quality of protein structures." In: *Journal of Applied Crystallography* 26.2 (1993), pp. 283–291.

[95] R. J. Law, C. Capener, M. Baaden, P. J. Bond, J. Campbell, G. Patargias, Y. Arinaminpathy, and M. S. Sansom. "Membrane protein structure quality in molecular dynamics simulation." In: *J. Mol. Graph. Model.* 24.2 (2005), pp. 157–165.

[96] A. R. Leach and A. P. Lemon. "Exploring the conformational space of protein side chains using dead-end elimination and the A* algorithm." In: *Proteins* 33.2 (1998), pp. 227–239.

[97] C. Levinthal. "Are there pathways for protein folding?" In: *J.Chim.Phys.* 65 (1968), pp. 44–45.

[98] X. Li, P. Mooney, S. Zheng, C. R. Booth, M. B. Braunfeld, S. Gubbens, D. A. Agard, and Y. Cheng. "Electron counting and beam-induced motion correction enable near-atomic-resolution single-particle cryo-EM." In: *Nat. Methods* 10.6 (2013), pp. 584–590.

[99] S. Liang and N. V. Grishin. "Side-chain modeling with an optimized scoring function." In: *Protein Sci.* 11.2 (2002), pp. 322–331.

[100] Pu Liu, Dimitris K. Agrafiotis, and Douglas L. Theobald. "Fast determination of the optimal rotational matrix for macromolecular superpositions." In: *Journal of Computational Chemistry* 31.7 (2010), pp. 1561–1563.

[101] A. L. Lomize, I. D. Pogozheva, and H. I. Mosberg. "Quantification of helix-helix binding affinities in micelles and lipid bilayers." In: *Protein Sci.* 13.10 (2004), pp. 2600–2612.

[102] A. L. Lomize, I. D. Pogozheva, M. A. Lomize, and H. I. Mosberg. "Positioning of proteins in membranes: a computational approach." In: *Protein Sci.* 15.6 (2006), pp. 1318–1333.

[103] M. A. Lomize, A. L. Lomize, I. D. Pogozheva, and H. I. Mosberg. "OPM: orientations of proteins in membranes database." In: *Bioinformatics* 22.5 (2006), pp. 623–625.

[104] S. C. Lovell, J. M. Word, J. S. Richardson, and D. C. Richardson. "The penultimate rotamer library." In: *Proteins* 40.3 (2000), pp. 389–408.

[105] H. Lu and J. Skolnick. "A distance-dependent atomic knowledge-based potential for improved protein structure selection." In: *Proteins* 44.3 (2001), pp. 223–232.

[106] M. Lu, A. D. Dousis, and J. Ma. "OPUS-Rota: a fast and accurate method for side-chain modeling." In: *Protein Sci.* 17.9 (2008), pp. 1576–1585.

[107] R. Luthy, J. U. Bowie, and D. Eisenberg. "Assessment of protein models with three-dimensional profiles." In: *Nature* 356.6364 (1992), pp. 83–85.

[108] D. J. Mandell, E. A. Coutsias, and T. Kortemme. "Sub-angstrom accuracy in protein loop reconstruction by robotics-inspired conformational sampling." In: *Nat. Methods* 6.8 (2009), pp. 551–552.

[109] V. Mariani, M. Biasini, A. Barbato, and T. Schwede. "lDDT: a local superposition-free score for comparing protein structures and models using distance difference tests." In: *Bioinformatics* 29.21 (2013), pp. 2722–2728.

[110] L. J. McGuffin, M. T. Buenavista, and D. B. Roche. "The ModFOLD4 server for the quality assessment of 3D protein models." In: *Nucleic Acids Res.* 41.Web Server issue (2013), W368–372.

[111] L. J. McGuffin, J. D. Atkins, B. R. Salehe, A. N. Shuid, and D. B. Roche. "IntFOLD: an integrated server for modelling protein structures and functions from amino acid sequences." In: *Nucleic Acids Res.* 43.W1 (2015), W169–173.

[112] A. McPherson. "Introduction to protein crystallization." In: *Methods* 34.3 (2004), pp. 254–265.

[113] A. Meier and J. Soding. "Automatic Prediction of Protein 3D Structures by Probabilistic Multi-template Homology Modeling." In: *PLoS Comput. Biol.* 11.10 (2015), e1004343.

[114] J. Mendes, A. M. Baptista, M. A. Carrondo, and C. M. Soares. "Improved modeling of side-chains in proteins with rotamer-based methods: a flexible rotamer model." In: *Proteins* 37.4 (1999), pp. 530–543.

[115] M. A. Messih, R. Lepore, and A. Tramontano. "LoopIng: a template-based tool for predicting the structure of protein loops." In: *Bioinformatics* 31.23 (2015), pp. 3767–3772.

[116] E. Michalsky, A. Goede, and R. Preissner. "Loops In Proteins (LIP)–a comprehensive loop database for homology modelling." In: *Protein Eng.* 16.12 (2003), pp. 979–985.

[117] M. Michino, E. Abola, C. L. Brooks, J. S. Dixon, J. Moult, and R. C. Stevens. "Community-wide assessment of GPCR structure modelling and ligand docking: GPCR Dock 2008." In: *Nat Rev Drug Discov* 8.6 (June 2009), pp. 455–463.

[118] R. Milo. "What is the total number of protein molecules per cell volume? A call to rethink some published values." In: *Bioessays* 35.12 (2013), pp. 1050–1055.

[119] J. Moult, K. Fidelis, A. Kryshtafovych, T. Schwede, and A. Tramontano. "Critical assessment of methods of protein structure prediction (CASP)–round x." In: *Proteins* 82 Suppl 2 (2014), pp. 1–6.

[120] S. B. Needleman and C. D. Wunsch. "A general method applicable to the search for similarities in the amino acid sequence of two proteins." In: *J. Mol. Biol.* 48.3 (1970), pp. 443–453.

[121] K. Olechnovic, E. Kulberkyte, and C. Venclovas. "CAD-score: a new contact area difference-based function for evaluation of protein structural models." In: *Proteins* 81.1 (2013), pp. 149–162.

[122] S. H. de Oliveira, J. Shi, and C. M. Deane. "Building a better fragment library for de novo protein structure prediction." In: *PLoS ONE* 10.4 (2015), e0123998.

[123] L. PAULING and R. B. COREY. "The pleated sheet, a new layer configuration of polypeptide chains." In: *Proc. Natl. Acad. Sci. U.S.A.* 37.5 (1951), pp. 251–256.

[124] L. PAULING, R. B. COREY, and H. R. BRANSON. "The structure of proteins; two hydrogen-bonded helical configurations of the polypeptide chain." In: *Proc. Natl. Acad. Sci. U.S.A.* 37.4 (1951), pp. 205–211.

[125] D. Pal and P. Chakrabarti. "Cis peptide bonds in proteins: residues involved, their conformations, interactions and locations." In: *J. Mol. Biol.* 294.1 (1999), pp. 271–288.

[126] F. Pedregosa et al. "Scikit-learn: Machine Learning in Python." In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[127] P. A. Penczek. "Fundamentals of three-dimensional reconstruction from projections." In: *Meth. Enzymol.* 482 (2010), pp. 1–33.

[128] J. Peng and J. Xu. "A multiple-template approach to protein threading." In: *Proteins* 79.6 (2011), pp. 1930–1939.

[129] S. B. Prusiner. "Novel proteinaceous infectious particles cause scrapie." In: *Science* 216.4542 (1982), pp. 136–144.

[130] G. N. RAMACHANDRAN, C. RAMAKRISHNAN, and V. SASISEKHARAN. "Stereochemistry of polypeptide chain configurations." In: *J. Mol. Biol.* 7 (1963), pp. 95–99.

[131] A. Ray, E. Lindahl, and B. Wallner. "Model quality assessment for membrane proteins." In: *Bioinformatics* 26.24 (2010), pp. 3067–3074.

[132] A. Ray, E. Lindahl, and B. Wallner. "Improved model quality assessment using ProQ2." In: *BMC Bioinformatics* 13 (2012), p. 224.

[133] R. J. Read et al. "A new generation of crystallographic validation tools for the protein data bank." In: *Structure* 19.10 (2011), pp. 1395–1412.

[134] M. Remmert, A. Biegert, A. Hauser, and J. Soding. "HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment." In: *Nat. Methods* 9.2 (2011), pp. 173–175.

[135] D. B. Roche, M. T. Buenavista, and L. J. McGuffin. "Assessing the quality of modelled 3D protein structures using the ModFOLD server." In: *Methods Mol. Biol.* 1137 (2014), pp. 83–103.

[136] C. A. Rohl, C. E. Strauss, K. M. Misura, and D. Baker. "Protein structure prediction using Rosetta." In: *Meth. Enzymol.* 383 (2004), pp. 66–93.

[137] A. Sali and T. L. Blundell. "Comparative protein modelling by satisfaction of spatial restraints." In: *J. Mol. Biol.* 234.3 (1993), pp. 779–815.

[138] R. Samudrala and J. Moult. "An all-atom distance-dependent conditional probability discriminatory function for protein structure prediction." In: *J. Mol. Biol.* 275.5 (1998), pp. 895–916.

[139] M. F. Sanner, A. J. Olson, and J. C. Spehner. "Reduced surface: an efficient way to compute molecular surfaces." In: *Biopolymers* 38.3 (1996), pp. 305–320.

[140] T. Schmidt, A. Bergner, and T. Schwede. "Modelling three-dimensional protein structures for applications in drug design." In: *Drug Discov. Today* 19.7 (2014), pp. 890–897.

[141] T. Schwede. "Protein modeling: what happened to the "protein structure gap"?" In: *Structure* 21.9 (2013), pp. 1531–1540.

[142] T. Schwede et al. "Outcome of a workshop on applications of protein models in biomedical research." In: *Structure* 17.2 (2009), pp. 151–159.

[143] M. V. Shapovalov and R. L. Dunbrack. "A smoothed backbone-dependent rotamer library for proteins derived from adaptive kernel density estimates and regressions." In: *Structure* 19.6 (2011), pp. 844–858.

[144] M. Y. Shen and A. Sali. "Statistical potential for assessment and prediction of protein structures." In: *Protein Sci.* 15.11 (2006), pp. 2507–2524.

[145] M. J. Sippl. "Calculation of conformational ensembles from potentials of mean force. An approach to the knowledge-based prediction of local structures in globular proteins." In: *J. Mol. Biol.* 213.4 (1990), pp. 859–883.

[146] M. J. Sippl. "Recognition of errors in three-dimensional structures of proteins." In: *Proteins* 17.4 (1993), pp. 355–362.

[147] M. J. Skwark and A. Elofsson. "PconsD: ultra rapid, accurate model quality assessment for protein structure prediction." In: *Bioinformatics* 29.14 (2013), pp. 1817–1818.

[148] T. F. Smith and M. S. Waterman. "Identification of common molecular subsequences." In: *J. Mol. Biol.* 147.1 (1981), pp. 195–197.

[149] J. Soding. "Protein homology detection by HMM-HMM comparison." In: *Bioinformatics* 21.7 (2005), pp. 951–960.

[150] A. D. Solis and S. Rackovsky. "Improvement of statistical potentials and threading score functions using information maximization." In: *Proteins* 62.4 (2006), pp. 892–908.

[151] Y. Song, F. DiMaio, R. Y. Wang, D. Kim, C. Miles, T. Brunette, J. Thompson, and D. Baker. "High-resolution comparative modeling with RosettaCM." In: *Structure* 21.10 (2013), pp. 1735–1742.

[152] G. Studer, M. Biasini, and T. Schwede. "Assessing the local structural quality of transmembrane protein models using statistical potentials (QMEANBrane)." In: *Bioinformatics* 30.17 (2014), pp. i505–511.

[153] T. J. Taylor, C. H. Tai, Y. J. Huang, J. Block, H. Bai, A. Kryshtafovych, G. T. Montelione, and B. Lee. "Definition and classification of evaluation units for CASP10." In: *Proteins* 82 Suppl 2 (2014), pp. 14–25.

[154] Robert E. Thach and Sigrid S. Thach. "Damage to Biological Samples Caused by the Electron Beam during Electron Microscopy." In: *Biophysical Journal* 11.2 (1971), pp. 204 –210.

[155] P. D. Thomas and K. A. Dill. "Statistical potentials extracted from protein structures: how accurate are they?" In: *J. Mol. Biol.* 257.2 (1996), pp. 457–469.

[156] K. Uziela and B. Wallner. "ProQ2: estimation of model accuracy implemented in Rosetta." In: *Bioinformatics* 32.9 (2016), pp. 1411–1413.

[157] G. Wang and R. L. Dunbrack. "PISCES: a protein sequence culling server." In: *Bioinformatics* 19.12 (2003), pp. 1589–1591.

[158] B. Webb and A. Sali. "Comparative Protein Structure Modeling Using MODELLER." In: *Curr Protoc Bioinformatics* 54 (2016), pp. 1–5.

[159] S. H. White. "The progress of membrane protein structure determination." In: *Protein Sci.* 13.7 (2004), pp. 1948–1949.

[160] S. H. White. "Biophysical dissection of membrane proteins." In: *Nature* 459.7245 (2009), pp. 344–346.

[161] S. H. White, A. S. Ladokhin, S. Jayasinghe, and K. Hristova. "How membranes shape protein structure." In: *J. Biol. Chem.* 276.35 (2001), pp. 32395–32398.

[162] M. Wiederstein and M. J. Sippl. "ProSA-web: interactive web service for the recognition of errors in three-dimensional structures of proteins." In: *Nucleic Acids Res.* 35.Web Server issue (2007), W407–410.

[163] P. E. Wright, H. J. Dyson, and R. A. Lerner. "Conformation of peptide fragments of proteins in aqueous solution: implications for initiation of protein folding." In: *Biochemistry* 27.19 (1988), pp. 7167–7175.

[164] L. Wroblewska and J. Skolnick. "Can a physics-based, all-atom potential find a protein's native structure among misfolded structures? I. Large scale AMBER benchmarking." In: *J Comput Chem* 28.12 (2007), pp. 2059–2066.

[165] D. Xu and Y. Zhang. "Ab initio protein structure assembly using continuous structure fragments and optimized knowledge-based force field." In: *Proteins* 80.7 (2012), pp. 1715–1735.

[166] Jinbo Xu. "Rapid Protein Side-Chain Packing via Tree Decomposition." In: *Research in Computational Molecular Biology: 9th Annual International Conference, RECOMB 2005, Cambridge, MA, USA, May 14-18, 2005. Proceedings*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2005, pp. 423–439.

[167] J. Yang, R. Yan, A. Roy, D. Xu, J. Poisson, and Y. Zhang. "The I-TASSER Suite: protein structure and function prediction." In: *Nat. Methods* 12.1 (2015), pp. 7–8.

[168] A. Zemla. "LGA: A method for finding 3D similarities in protein structures." In: *Nucleic Acids Res.* 31.13 (2003), pp. 3370–3374.

[169] Q. C. Zhang, D. Petrey, J. I. Garzon, L. Deng, and B. Honig. "PrePPI: a structure-informed database of protein-protein interactions." In: *Nucleic Acids Res.* 41.Database issue (2013), pp. D828–833.

[170] H. Zhou and Y. Zhou. "Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction." In: *Protein Sci.* 11.11 (2002), pp. 2714–2726.

[171] H. Zhou and Y. Zhou. "Fold recognition by combining sequence profiles derived from evolution and from depth-dependent structural alignment of fragments." In: *Proteins* 58.2 (2005), pp. 321–328.

COLOPHON

This document was typeset using the typographical look-and-feel `classicthesis` developed by André Miede. The style was inspired by Robert Bringhurst's seminal book on typography "*The Elements of Typographic Style*". `classicthesis` is available for both LaTeX and LyX:

`https://bitbucket.org/amiede/classicthesis/`

Happy users of `classicthesis` usually send a real postcard to the author, a collection of postcards received so far is featured here:

`http://postcards.miede.de/`

Final Version as of September 4, 2017 (`classicthesis` version 4.2).