Contents

List of Figures

List of Tables

List of Abbreviations

ACL Average Chain Length.

AMPLE Ab initio Modelling of Proteins for moLEcular replacement.

CC Correlation Coefficient.

FP False Positive.

LLG Log-Likelihood Gain.

M$_{eff}$ Number of Effective Sequences.

MR Molecular Replacement.

MSA Multiple Sequence Alignment.

MX Macromolecular Crystallography.

PDB Protein Data Bank.

PDBTM Protein Data Bank of Transmembrane Proteins.

RIO Residue-Independent Overlap.

RMSD Root Mean Square Deviation.

TFZ Translation Function Z-score.

TM-score Template-Modelling score.

TP True Positive.

CHAPTER 1

Introduction

1.1 Macromolecular X-ray crystallography

The discovery of X-ray diffraction by crystals by Max van Laue [1, 2] marked the origins of modern crystallography. However, it was not until the work of William Lawrence Bragg and William Henry Bragg that X-ray scattering could be translated into atomic positions [3–5]. Since then, X-ray crystallography and the determination of atomic positions in organic and inorganic molecules of has come a long way and shaped the path for many 21$^{\text{st}}$ century discoveries. Amongst those ground-breaking discoveries are the earliest structural models of biological molecules including DNA [6], vitamin B12 [7], and the first protein structures [8–11]. These structure elucidations hallmarked the dawn of a new era in biological and biomedical research. At the time of writing, 124,551 structural models were determined by X-ray diffraction studies [12], and thus X-ray crystallography is a key method in biological research.

1.1.1 X-ray scattering

X-rays are high energy photons part of the electromagnetic spectrum with a wavelength 0.1-100Å [13]. X-rays can be described as packets of travelling electromagnetic waves, whose electric field vector interacts with the charged electrons of matter [13]. Such interaction, typically termed scattering, results in the diffraction of the incoming wave, which X-ray crystallography relies on.

In its simplest form, scattering of X-ray radiation can be explained in the scenario of exposure to a single free electron. The resulting scattering can be classed as elastic (Thomson scattering) or inelastic (Crompton scattering) [13]. The latter — scattering that results in a loss of energy of the emitting photon due to energy transfer onto the electron — does not contribute to discrete scattering, the type of scattering X-ray diffraction relies on. In comparison, Thomson scattering does not result in a loss of energy of the emitting photon. This has significant effects, the incoming photon emits with the same frequency causing the electron to oscillate identically further enhancing the signal.

If we expand the example to include all electrons in an atom and expose the atom to X-ray radiation, our theory needs to be slightly expanded. Given that one or more electrons in an atom are not free but orbit around the atom's nucleus in a stable and defined manner, the distribution of these electrons around the nucleus determines the scattering of the incoming X-ray photons. The distribution of scattered photon waves is thus an overall representation of the probability distributions of each electron in the atom and is refered to as electron density $\rho(\boldsymbol{r})$. In X-ray scattering, it suffices to approximate the shape of the electron density to a sphere. If we now consider the emitting wave $\boldsymbol{s_1}$ of an X-ray photon scattered by any position $\boldsymbol{r}$ in the electron density of an atom, then the phase difference $\Delta\varphi$ to the incoming wave $\boldsymbol{s_0}$ can be described by Eq. (1.1) [13].

$$\Delta\varphi = 2\pi\left(\boldsymbol{s_1} - \boldsymbol{s_0}\right)\boldsymbol{r} = 2\pi \cdot \boldsymbol{Sr} \tag{1.1}$$

If more than one electron in an atom's electron density scatter the incoming X-ray wave, then the emitting partial waves can be described by the atomic scattering function f_s (Eq. (1.2)), which describes the interference of all scattered waves [13]. The total scattering power of an atom is proportional to the number of electrons and element-specific with heavier atoms scattering more strongly. Given the approximation of a centrosymmetric electron density, the atomic scattering function is also symmetric.

$$f_s = \int_{\boldsymbol{r}}^{V(atoms)} \rho\left(\boldsymbol{r}\right) \cdot e^{2\pi i \boldsymbol{Sr}} \cdot d\boldsymbol{r} \tag{1.2}$$

With an enhanced understanding of X-ray scattering of electrons orbiting a single atom, it is important to consider X-ray scattering of adjacent atoms, such as it is typically found in molecules. If the electromagnetic wave of a X-ray photon excites all electrons of adjacent atoms, then the resulting partial waves result in constructive or destructive interference. Maximal interference can be obtained when all partial waves are in phase, and maximal destructive interference when out-of-phase. This leads to varying intensities of the emitting X-ray photon at different points in space. To obtain the overall scattering power F_s of all contributing atoms, Eq. (1.2) needs to be modified to include the sum over all atoms j as described in Eq. (1.3).

$$F_s = \sum_{j=1}^{atoms} f_{s,j}^0 \cdot e^{2\pi i \boldsymbol{S} \boldsymbol{r}_j} \tag{1.3}$$

If we now translate our hypothetical experiment into a crystal lattice then our understanding described in Eq. (1.3) needs to be expanded from a 1-dimensional distance vector $\boldsymbol{r}$ to the three dimensional lattice translation vectors $\boldsymbol{a}$, $\boldsymbol{b}$ and $\boldsymbol{c}$. The Laue equations (Eq. (1.4)) do exactly that and ultimately determine the positions of the diffraction peaks in 3-dimensional space.

$$\boldsymbol{S} \cdot \boldsymbol{a} = n_1, \qquad \boldsymbol{S} \cdot \boldsymbol{b} = n_2, \qquad \boldsymbol{S} \cdot \boldsymbol{c} = n_3 \tag{1.4}$$

$$n\lambda = 2d_{hkl} sin\theta \tag{1.5}$$

Such determination is possible through the findings made by Bragg and Bragg [3], who identified the relationship between the scattering vector $\boldsymbol{S}$ and the planes in the crystal lattice. Today, this relationship is defined by the Bragg equation (Eq. (1.5)) [3], which allows us to interpret X-ray diffraction as reflections on discrete lattice planes, which relates the diffraction angle θ to the lattice spacing d_{hkl} (Fig. 1.1) [13]. For maximum diffraction n needs to be integer multiples to result in maximum constructive interference of wavelength λ.

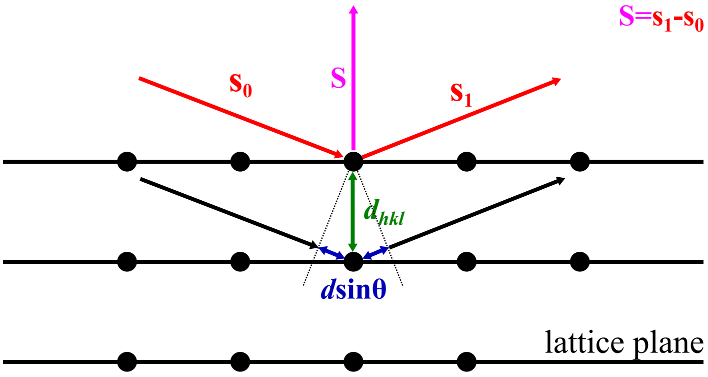

Figure 1.1: Schematic of Bragg scattering.

Lastly, if the hypothetical model is expanded to molecular crystals, then the total scattering from the unit cell is merely a summation of all molecular unit cell scattering contributions in the crystal. Mathematically, this results in Eq. (1.3) being generalised to Eq. (1.6) through the application of the Laue equations (Eq. (1.4)) to express the scattering vector $\boldsymbol{Sr}_j$ as Miller indices of the reflection planes $\boldsymbol{hx}_j$.

$$F_h = \sum_{j=1}^{atoms} f_{s,j}^{0} \cdot e^{2\pi i \boldsymbol{hx}_j} \tag{1.6}$$

The structure factor equation defines the scattering power from a crystal in a given reciprocal lattice direction $\boldsymbol{h}$. The scattering is enhanced by the number of repeating units of lattice translation vectors $\boldsymbol{a}$, $\boldsymbol{b}$ and $\boldsymbol{c}$, and thus the overall scattering power is proportional to the number of unit cells in the crystal.

It should be noted that Eq. (1.6) is a simplification of the problem at hand. In reality, instrument and experimental corrections need to be applied to the structure factor equation. A correction factor for each experiment-dependent parameter needs to be applied to the structure factor equation. However, in the scope of this work the details of such correction factors do not need to be discussed.

$$\rho(x,y,z) = \frac{1}{V} \sum_{h=0}^{+\infty} \sum_{k=-\infty}^{+\infty} \sum_{l=-\infty}^{+\infty} \boldsymbol{F}(hkl) \cdot e^{-2\pi i(hx+ky+lz)} \tag{1.7}$$

Since complex structure factors describe the molecular structure in the reciprocal space domain, the conversion to the real space domain in form of electron density is required. This can be conveniently done through the bijective Fourier transform, which allows to convert complex structure factors to electron density and vice versa without the loss of any information [13]. Thus, electron density can be obtained from the complex structure factors using Eq. (1.7). The normalisation factor $1/V$ provides the correct units for the electron density $\rho(x,y,z)$.

1.1.2 From crystal to structure

In X-ray crystallographic experiments, X-ray radiation is measured using light detectors. However, the measurement taken is incomplete. Light detectors only capture the intensity of the scattered X-ray photons but crucially lose the phase information. The latter is essential for atomic reconstruction of the molecule in the crystal, and thus needs to be obtained. In Macromolecular Crystallography (MX), experimentalists have a number of alternative techniques to compensate for the lost phase information.

Prior to the big advances in computing power and the successful elucidation of many protein structures, MX crystallographers primarily recovered the lost phase information through Direct Methods or Experimental Phasing [13]. Today, the most popular method to recovering the lost phase information is Molecular Replacement (MR) [14, 15]. In a MR search, a known structure ('search model') similar to the unknown is relocated in the unit cell until the solution with the best fit between calculated and observed diffraction data is obtained [13]. A 6-dimensional search, i.e. a simultaneous rotation and translation search, is possible [16–18], however computationally very expensive and less suitable for challenging cases. In comparison, most modern crystalloraphic applications opt for two distinct sub-searches, the rotation search to orient the search model within the unit cell followed by the translation search to locate it [13]. The benefits over a combined search include search-specific target functions that enable increased sensitivity and additional terms to compensate for imperfect data.

The most successful MR algorithms perform the rotation and translation searches using Patterson methods or Maximum Likelihood functions. Patterson methods — originally developed by Rossmann and Blow [19] — rely on the use of a map of vectors between the scattering atoms, which can be determined for the calculated and observed structure factor amplitudes. Patterson vectors can be sub-classed as intra- and inter-molecular vectors. A distinct separation of the observed vectors is impossible. However, inter-molecular vectors appear further away from the central peak of self-vector (vector from atom to itself) in the Patterson map [13]. The calculated Patterson vectors for the search model allow for a clearer distinction between the intra- and inter-molecular vectors. If the search model is placed in a large unit cell, then inter-molecular vectors must scale with the unit cell dimension [13]. Ultimately, using the intra-molecular Patterson vectors, the search probe can be oriented against the experimentally determined Patterson vectors. Similarly, the inter-molecular vectors can be used to identify the correct translation of the search probe. Patterson methods are very sensitive to small orientation errors of the search probe [13]. Thus, orientations with the highest vector peak overlaps are trialed in the subsequence translation search.

In comparison to the Patterson methods, Maximum Likelihood methods do not rely on inter-atomic vectors in Patterson maps. Instead, Maximum Likelihood methods make use of Bayes' theorem [20] to compare calculated structure factors and observed structure factor amplitudes directly [18]. Bayes' theorem in crystallographic Maximum Likelihood methods is applied to compute the likelihood that an experimental value is observed given the current search model. The maximal likelihood indicates the best search model given the observed experimental data. Since the search model likelihood term is the product of many individual probabilities, which are difficult to represent computationally due to floating point representations, the log of the likelihood is commonly used [13]. The major advantage of Maximum likelihood methods over Patterson methods centres on the more realist target functions, which consider errors and incompleteness of the search model, applies bulk solvent correction and conducts multi-model searches [18]. The latter is of particularl relevance since the Maximum likelihood rotation function can thus consider already placed search model probes in a fixed position whilst trialling additional ones [21], which proves to be a major advantage over Patterson methods. Furthermore, likelihood target functions consider the structural variance of multiple superposed models in an ensemble search model, which is used to weight structure factors at the various positions to improve the overall likelihood term [18].

The initial electron density map after MR is almost always inaccurate because of the search model-based phases. Inaccuracies arise from experimental errors, model incompleteness, low signal-to-noise or model bias. Thus, approaches for improving the phases used to calculated the initial electron density map have been developed and are routinely applied in MX. Density modification describes a set of methods that improve the obtained electron density typically by applying statistical corrections to electron density distributions. These corrections are based on prior knowledge or assumptions of the physical properties of macromolecular structures [13]. This process can transform initially poor or uninterpretable initial electron density maps to high quality ones. Three pre-dominant density modification approaches exist: solvent flattening, histogram matching and the "sphere-of-influence" method. Solvent flattening is an approach first proposed by Wang [22], which exploits the fact that solvent regions in protein crystals are disordered, and thus differ in electron density volume from macromolecule-containing regions. If solvent electron density is set to a constant, then it is essentially flattened which will result in improved structure factors with improved phases and thus improved electron density. Histogram matching [23] exploits the defined characteristics of an electron density distribution determined from sets of proteins at the same resolution, irrespectivce of individual structural details. The electron density distribution for noisy maps are Gaussian-shaped. In contrast, the electron density distribution of a feature-defined map is positively skewed. The "sphere-of-influence" method was introduced by Sheldrick [24] and classifies solvent and protein electron density by observing its variance across the shell surface of a 2.42Å sphere (dominant 1-3 atom distance in macromolecular structures). If the sphere is positioned in the disordered solvent region typically found in intermolecular channels, the density variance will be low. Thus, this approach allows to smoothen solvent-containing regions of the electron density [24]. Independent of the density modification strategy applied, it is important to understand that improvements to the electron density map anywhere lead to improvements everywhere by transferral of information from one part of the map to another [25].

A second approach to improving the initial electron density is termed Refinement. Iteratively, the placed search model is optimised to better describe the experimentally observed data. This optimisation problem is typically broken down into three main steps: the definition of the model parameters, the scoring function and the optimisation method. The model parameters describe the crystal and its content and can be subdivided into atomic and non-atomic model parameters [26]. These parameters combined are used to score the current model. The scoring function relates the experimental data to the model parameters. The scoring function contains two primary terms, the refinement data target and an *a priori* knowledge term. The former defines a target function that assesses the similarity between calculated and experimental structure factors. The target function is commonly a Maximum Likelihood-based function that considers missing or incomplete data [26, 27]. The *a priori* knowledge term in the scoring function defines the properties of a good model by including stereochemical property terms. Lastly, optimisation methods provide tools to vary the model parameters to better fit the experimental data. Different optimisation techniques can be used depending on the severity of model parameter alteration, which generally depend on the entrapment of states in local energy minima. The three steps combined form a macrocycle that iteratively modifies the model to optimise its fit to the experimental data. This ultimately improves both the electron density map interpretability and model quality. MX refinement can be performed in structure-factor-based reciprocal space and electron-density-based real space [26]. A combination allows global and local refinement strategies and enables grid-like searches to optimise the model parameters until convergence.

Once initial phase information is improved through refinement and density modification, attempts can be made to build atomic model coordinates into the electron density map. This process is typically coupled with refinement or density modification to iteratively improve the quality of the partially built model and the electron density map [13]. A small number of distinct algorithms are currently used to automatically build atomic coordinates into electron density: main-chain autotracing [28], fitting pseudo-atoms into electron density [29], or fitting reference coordinates with similar electron density maps [30, 31]. In essence, all algorithms attempt to maximise the number of correctly identified and placed atomic coordinates into available electron density. Whilst autotracing solely builds main-chain peptides, the other two approaches rely on sequence information to also build side-chains. Independent of the complexity of the model building task, the higher the resolution and the more complete the initial starting model, the less ambiguous and challenging this overall task becomes [13].

1.1.3 Unconventional Molecular Replacement

The process of macromolecular structure determination via conventional MR has been outlined previously. Search models are typically derived from structural homologs identified by sequence similarity to the crystallised target [13]. However, with decreasing sequence similarity between homologs, it becomes more challenging to identify structural templates suitable for MR. Furthermore, experimental phasing approaches to circumvent the absence of MR templates can be expensive, unsuccessful and very challenging for certain protein targets, and thus remain unfeasable to pursue at times. Under such circumstances, alternative approaches are required, which are referred to as “unconventional” MR approaches from here onwards. The unconventional MR approach most relevant to the work presented in this thesis utilises the 3-dimensional structure prediction of a protein target starting from its sequence [32–34].

1.2 *Ab initio* protein structure prediction

The folding of protein structures is commonly described by the folding funnel hypothesis [35]. It assumes that the native state of a protein fold corresponds to its global minimum free energy state along its energy surface (Fig. 1.2) [36]. *In silico* protein folding experiments attempt to find this lowest-free-energy state of the protein fold; however, to unambiguously identify it sampling of all polypeptide chain conformations is necessary. In theory, sampling of all conformations for a 100-residue protein takes in the order of approximately 10^{52} years (10^7 configurations with 10^{-11} seconds per configuration), yet in practice an equivalent polypeptide chain would fold in milliseconds to seconds [37, 38]. This paradox — termed the Levinthal paradox [37] — created the basis for the folding funnel hypothesis.

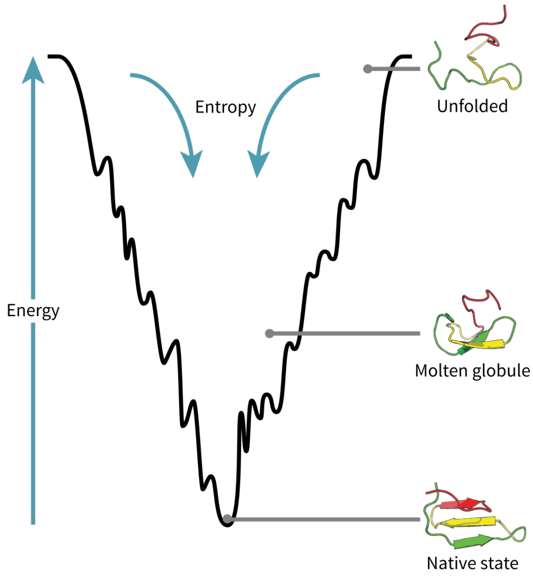

Figure 1.2: Schematic of the folding funnel hypothesis [35]. Diagram produced by Wikipedia [39] contributors.

In *ab initio* protein structure prediction, the tertiary structure of a protein is predicted using its primary structure alone. This problem is in its nature identical to finding the lowest-energy state along the protein's energy landscape. However, in an attempt to avoid the Levinthal paradox, different knowledge- and physics-based energy functions coupled with a variety of conformational-search sampling algorithms are employed [40].

Physics-based energy functions use physiochemical force fields typically coupled with Molecular Dynamics simulations to sample the folding trajectory of a protein sequence (true physics-based approaches are computationally intractable because quantum mechanics models would need to be used). Force fields describe parameter sets used to calculate energy potentials for a system of atoms in a simulation run, and include potentials such as van der Waals and electrostatic interactions [40]. In the context of *ab initio* protein structure prediction, pure physics-based approaches are often less favourable, because the computational complexity to find the lowest free-energy state of a large protein structure remains intractable without the use of supercomputers.

Knowledge-based energy functions rely on empirical energy terms derived from statistics and regularities of experimentally determined structures [40]. These energy terms can be subdivided into two types, the generic or sequence-independent terms and amino-acid or sequence-dependent terms [41]. The former include terms to describe the backbone hydrogen-bonds and local backbone stiffness of a polypeptide chain. The latter describes terms such as pairwise residue contact potential, distance-dependent atomic contact potential, and secondary structure propensities. However, predicting local or global tertiary structure of a protein sequence using empirical energy terms alone is very difficult. Subtle differences in the local and global environment of a primary structure alongside the subtle differences in primary structures leading to common secondary structure features are very difficult to reproduce in a modelling scenario. Thus, knowledge-based energy functions are often coupled with the assembly of fragments extracted from other protein structures to predict the unknown tertiary structure of the target sequence [40].

The most successful *ab initio* structure prediction protocols use knowledge-based and physics-based energy functions combined with fragment-assembly-based conformational searches to find the lowest free-energy state [42–46]. Structural fragments of varying lengths (typically 3-20 residues) are extracted from existing protein structures [47–54]. These fragments are used in a Monte-Carlo simulation to search the conformational space of the polypeptide chain to search for low free-enery states [55]. The insertion of overlapping fragments results in the replacement of torsion angles either at random positions or sequentially from pre-defined starting position (such as N- or C-termini), and each move is scored against the Metropolis criterion [55] consisting of knowledge-based and physics-based terms. If a fragment passed the Metropolis criterion, its torsion angles are accepted and integrated in the polypeptide chain for the next fragment-insertion iteration. This process is repeated until convergence of the decoy, i.e. no lower free-energy state can be found. In all routines, these steps are independently repeated thousands of times to create a pool of decoys.

In order to identify the correct fold amongst the thousands of generated decoys, clustering approaches are commonly in combination with *ab initio* protocols. Shortle et al. [56] identified that the most-similar decoy to the native structure is most often the centroid (decoy with most neighbours in the cluster) of the largest cluster. Further studies showed that the selection of those centroid decoys helps to identify the most native-like folds amongst the many thousands generated [57–59]. Some protocols use clustering as an intermediate or final step to identify decoys for which it will perform more computationally demanding all-atom refinement [58] or other decoy hybridisation [43, 60, 61] approaches to further approach the native-like fold [62].

1.3 Residue-residue contact prediction

CHAPTER 2

General methodology

2.1 Dataset creation

2.1.1 FLUME dataset

A test set of 21 globular protein targets was manually selected to include a range of chain lengths, fold architectures, X-ray diffraction data resolutions and Multiple Sequence Alignment (MSA) depths for contact prediction (Table 2.1). The test set covered the three fold classes (α-helical, mixed α-β and β-sheet) and targets were grouped using their DSSP [63] secondary-structure assignment. Target chain lengths fell in the range of [62, 221] residues. Each crystal structure contained one molecule per asymmetric unit and the resolutions of the experimental data was in range from 1.0 to 2.3Å.

2.1.2 KEENO dataset

An unbiased selection of 27 non-redundant protein targets was selected using the following protocol (Table 2.2).

The Pfam v29.0 [84] database was filtered for all protein families with at least one representative structure in the RCSB PDB [12] database. Each representative had to have monomeric protein stoichiometry and its fold classified in the SCOPe v2.05 database [85]. Targets with fold assignments other than ”a” (all-α), ”b” (all-β), ”c” (mixed α+β) or ”d” (mixed α/β) were excluded to exclusively focus on regular globular protein folds. Each resulting protein target was screened against the RESTful API of the RCSB PDB (`www.rcsb.org`) webserver to identify targets meeting the following criteria: experimental technique is X-ray crystallography; chain length is $\geq$ 100 residues and $\leq$ 250 residues; resolution is between 1.3 and 2.3Å; structure factor amplitudes are deposited in the Protein Data Bank [12] database; and there is only a single molecule in the asymmetric unit. The resulting protein structures were cross-validated against the Protein Data Bank of Transmembrane Proteins (PDBTM) [86] to exclude any possible matches. Subsequently, one representative entry was randomly selected for each Pfam family.

The final set of 27 non-redundant targets was determined using further target characterisation and grouping of Pfam families. All targets were grouped using three criteria:

Table 2.1: Summary of the FLUME dataset.

PDB ID	Molecule	Resolution (Å)	Space Group	Chain ID	Chain Length	Molecules per ASU	Matthew's Coefficient	Solvent Content (%)	Fold	Citation
1a6m	Oxy-myoglobin	1.00	$P2_1$	A	151	1	1.90	36.00	all-α	[64]
1aba	T4 glutaredoxin	1.45	$P2_12_12_1$	A	87	1	2.22	44.62	mixed α/β	[65]
1bdo	Biotinyl domain of acetyl-coenzyme A carboxylase	1.80	$P2_12_12$	A	80	1	2.48	49.00	all-β	[66]
1bkr	Calponin Homology (CH) domain from β-spectrin	1.10	$P2_1$	A	109	1	2.04	39.80	all-α	[67]
1chd	CheB methylesterase domain	1.75	$P3_221$	A	203	1	2.35	47.65	mixed α/β	[68]
1e0s	G-protein Arf6-GDP	2.28	$P6_122$	A	174	1	2.18	37.00	mixed α/β	[69]
1eaz	Phosphoinositol (3,4)-bisphosphate PH domain	1.40	$C222_1$	A	125	1	2.48	48.00	mixed α+β	[70]
1hh8	N-terminal region of P67Phox	1.80	$P3_1$	A	213	1	2.71	45.00	all-α	[71]
1kjl	Galectin-3 domain	1.40	$P2_12_12_1$	A	146	1	2.15	42.68	all-β	[72]
1kw4	Polyhomeotic SAM domain	1.75	$P6_5$	A	89	1	2.25	45.27	all-α	[73]
1lo7	4-hydroxybenzoyl CoA thioesterase	1.50	I222	A	141	1	2.06	40.22	mixed α+β	[74]
1npu	Extracellular domain of murine PD-1	2.00	$P2_12_12_1$	A	117	1	1.67	25.80	all-β	[75]
1pnc	Poplar plastocyanin	1.60	$P2_12_12_1$	A	99	1	1.82	32.48	all-β	[76]
1tjx	Synaptotagmin I C2B domain	1.04	$P3_221$	A	159	1	2.40	48.00	mixed α+β	[77]
1tlv	LicT PRD	1.95	$P3_221$	A	221	1	2.80	50.00	all-α	[78]
2nuz	α-spectrin SH3 domain	1.85	$P2_12_12_1$	A	62	1	2.57	52.16	all-β	
2qyj	Ankyrin	2.05	$P6_1$	A	166	1	2.28	45.99	all-α	[79]
3w56	C2 domain	1.60	I2	A	131	1	2.05	40.10	all-β	[80]
4cl9	N-terminal bromodomain of Brd4	1.40	$P2_12_12_1$	A	127	1	2.21	44.37	all-α	[81]
4u3h	FN3con	1.98	$P4_132$	A	100	1	2.47	50.27	all-β	[82]
4w97	KstR2	1.60	C2	A	200	1	2.75	55.25	all-α	[83]

domain fold, target chain length and alignment depth. The former consisted of the three fold classes all-α, all-β, and mixed α-β (α+β and α/β) and targets were group using the SCOPe assignment. The target chain lengths were obtained from the deposited information via the RESTful API of the RCSB PDB web server and split into three bins, using 150 and 200 residues as bin edges. Furthermore, the alignment depth was calculated for the sequence alignment of each Pfam family and three bins established with bin edges of 100 and 200 sequences. Thus, all targets were classed in three bins for each of the three features.

The final selection of the 27 targets was performed by randomly selecting one target for each feature combination. To ensure even sampling across the three different fold categories, a target function was employed to identify roughly even target characteristics in each group. The alignment depth and chain length were used as metrics, and had to be within ±15 units to the values of the other fold classes. This created two conditions that had to be met for a randomly chosen sample to be accepted.

2.1.3 ETHERWOOD dataset

The selection of this dataset was done by [110]. In summary, 14 non-redundant transmembrane protein targets were selected from the PDBTM [86], with a chain length of < 250 residues and resolution of < 2.5Å. The final selection is summarised in Table 2.3.

2.2 Enhancement of β-sheet restraints

Structure prediction of β-strand containing protein targets *ab initio* is a notoriously challenging task. β-strands, potentially far in sequence space, form a β-sheet in 3-dimensions. Since fragment-assembly algorithms work on the basis of randomly inserting one fragment at the time, the probability of β-strand formation is much lower compared to α-helices.

Recent advances in *ab initio* structure prediction have seen great improvements in structure prediction quality through the use of predicted residue-residue contacts as distance restraints (see Section 1.3). However, only a single approach specifically focused

Table 2.2: Summary of the KEENO dataset.

PDB ID	Molecule	Resolution (Å)	Space Group	Chain ID	Chain Length	Molecules per ASU	Matthew's Coefficient	Solvent Content (%)	Fold	Citation
1fcy	Retinoic acid nuclear receptor HRAR	1.30	$P4_12_12$	A	236	1	2.25	45.50	all-α	[87]
1fvg	Peptide methionine sulfoxide reductase	1.60	C121	A	199	1	2.10	41.55	mixed α+β	[88]
1gm4	Cytochrome C3	2.05	$P6_122$	A	107	1	2.48	50.43	all-α	[89]
1gv8	N-II domain of ovotransferrin	1.95	$P3_1$	A	159	1	2.24	45.00	mixed α/β	[90]
1k40	FAT domain of focal adhesion kinase	2.25	C121	A	126	1	2.21	44.40	all-α	[91]
1oee	Hypothetical protein YodA	2.10	C121	A	193	1	2.30	46.20	all-β	[92]
1oz9	Hypothetical protein AQ_1354	1.89	$P4_32_12$	A	150	1	2.76	55.07	mixed α+β	[93]
1q8c	Hypothetical protein MG027	2.00	$P4_1$	A	151	1	2.42	49.25	all-α	[94]
1rlh	Conserved hypothetical protein	1.80	$P6_3$	A	173	1	2.12	41.98	mixed α+β	
1s2x	Cag-Z	1.90	$P2_12_12_1$	A	206	1	2.74	54.70	all-α	[95]
1u61	Putative Ribonuclease III	2.15	$I4_132$	A	138	1	6.50	80.80	all-α	
1zxu	At5g01750 protein	1.70	$P2_12_12_1$	A	217	1	2.50	50.20	mixed α+β	
2eum	Glycolipid transfer protein	2.30	C121	A	209	1	2.25	45.39	all-α	[96]
2ol8	Outer surface protein A	1.90	$P12_11$	O	249	1	2.19	43.87	all-β	[97]
2oqz	Sortase B	1.60	$P12_11$	A	223	1	2.07	40.71	all-β	[98]
2x6u	T-Box transcription factor TBX5	1.90	$P2_12_12_1$	A	203	1	2.20	44.21	all-β	[99]
2y64	Xylanase	1.40	$P2_12_12_1$	A	167	1	2.15	43.00	all-β	[100]
2yjm	TtrD	1.84	C121	A	176	1	2.08	40.80	all-α	[101]
2yq9	2, 3-cyclic-nucleotide 3-phosphodiesterase	1.90	$P2_12_12_1$	A	221	1	2.10	41.70	mixed α+β	[102]
3dju	Protein BTG2	2.26	$P2_12_12_1$	B	122	1	1.98	37.73	mixed α+β	[103]
3g0m	Cysteine desulfuration protein sufE	1.76	$P12_11$	A	141	1	1.88	34.58	mixed α+β	
3qzl	Iron-regulated surface determinant protein A	1.30	$P2_12_12$	A	127	1	2.42	49.12	all-β	[104]
4aaj	N-(5-phosphoribosyl)anthranilate isomerase	1.75	$P6_1$	A	228	1	2.38	48.30	mixed α/β	[105]
4dbb	Amyloid-β A4 precursor protein-binding family A1	1.90	$P4_12_12$	A	162	1	3.25	62.10	all-β	[106]
4e9e	Methyl-CpG-binding domain protein 4	1.90	H3	A	161	1	2.42	49.23	all-α	[107]
4lbj	Galectin-3	1.80	$P2_12_12_1$	A	138	1	2.09	41.01	all-β	[108]
4pgo	Hypothetical protein PF0907	2.30	$P6_522$	A	116	1	3.25	62.10	all-β	[109]

Table 2.3: Summary of the ETHERWOOD dataset.

PDB ID	**Molecule**	**Resolution (Å)**	**Space Group**	**Chain ID**	**Chain Length**	**Molecules per ASU**	**Matthew's Coefficient**	**Solvent Content (%)**	**Fold**	**Citation**
1gu8	Sensory rhodopsin II	2.27	$C222_1$	A	239	1	2.75	53.00	all-α	[111]
2bhw	Chlorophyll A-B binding protein AB80	2.50	C121	A	232	3	4.10	69.00	all-α	[112]
2evu	Aquaporin aqpM	2.30	I4	A	246	1	3.38	63.57	all-α	[113]
2o9g	Aquaporin Z	1.90	I4	A	234	1	3.34	63.19	all-α	[114]
2wie	ATP synthase C chain	2.13	$P6_322$	A	82	5	3.41	68.00	all-α	[115]
2xov	Rhomboid protease GLPG	1.65	H32	A	181	1	3.50	64.92	all-α	[116]
3gd8	Aquaporin 4	1.80	$P42_12$	A	223	1	2.73	54.97	all-α	[117]
3hap	Bacteriorhodopsin	1.60	$C222_1$	A	249	1	2.73	54.99	all-α	[118]
3ldc	Calcium-gated potassium channel mthK	1.45	$P42_12$	A	82	1	2.48	50.44	all-α	[119]
3ouf	Potassium channel protein	1.55	I2	A	97	2	2.40	48.76	all-α	[120]
3pcv	Leukotriene C4 synthase	1.90	F23	A	156	1	4.91	74.77	all-α	[121]
3rlb	ThiT	2.00	C121	A	192	2	3.89	68.39	all-α	[122]
3u2f	ATP synthase subunit C	2.00	$P4_222$	K	76	5	2.32	46.92	all-α	[123]
4dve	Biotin transporter BioY	2.09	C121	A	198	3	3.27	62.40	all-α	[124]

on improvements to the structure prediction of β-sheet formation [125]. To enhance the probability of β-sheet formation in *ab initio* structure prediction, part of this thesis focused on a more general model to enrich restraints between β-strands to attempt better super-secondary quality in the final decoys.

A more general approach, compared to [125] focusing on β-barrel proteins, was developed combining a starting set of contact pairs with a specifically-prepared set obtained from BBCONTACTS [126]. A HHBLITS [127] MSA was constructed using two sequence-search iterations with an E-value cutoff of 10^{-3} against the UniProt20 database [128]. Redundant sequences were removed from the MSA to 90% sequence identity using HHFILTER [127]. Subsequently, the MSA was subjected to CCMPRED [129] for co-evolution based contact prediction. The BBCONTACTS algorithms also requires a secondary-structure prediction, which was obtained using the ADDSS.PL script [127] distributed with the HHSUITE [130]. Both input files were subjected to BBCONTACTS to obtain a final set of β-strand specific contact pairs.

The BBCONTACTS contact pairs were added to a base set of contact pairs usually obtained from a separate (meta-)predictor. The combination of the two sets of contact pairs was done by simple union of the lists; however, if a contact pair was in the intersection, a contact-pair related weight was doubled to allow subsequent modifications of the energy term in distance restraint creation. Furthermore, additional contact pairs were inferred if not present in the base set of contact pairs. The inference worked on the basis that any neighbouring contacts (i.e. $i, j \pm 1$; $i, j \pm 2$; $i \pm 1, j$; $i \pm 2, j$) to contact i, j must be present, and thus any missing were automatically added to the final list. Again, any already present contact pair was assigned double the weight compared to the rest.

2.3 Evaluation of data

This section defines and describes concepts used throughout this thesis to assess and/or validate various data.

2.3.1 Sequence alignment data

2.3.1.1 Seqeuence alignment depth

Co-evolution based residue-residue contact prediction is dependent on an input MSA ideally containing all homolohous sequences found in the queried database. However, the MSA needs a certain level of sequence diversity amongst the homologs to accurately capture the co-evolution signal. The alignment depth — often also referred to as Number of Effective Sequences (M$_{eff}$) — captures this diversity by computing the number of non-redundant sequences in the MSA.

$$M_{eff} = \sum_i \frac{1}{\sum_j S_{i,j}} \tag{2.1}$$

Various approaches exist for computing M$_{eff}$ [131–133] yielding similar results [134]. In this thesis, the approach defined by Morcos et al. [131] is used. Morcos et al. [131] first described the approach by which sequence weights are computed by means of Hamming distances between all possible sequence combinations in the MSA (Eq. (2.1)). If a Hamming distance was < 0.2 (sequence identity of 80%), the binary value $S_{i,j}$ was assigned 1 and otherwise a 0. The sum of fractional weights of the similarity of each sequence compared to all others ultimately describes the alignment depth.

2.3.2 Contact prediction data

2.3.2.1 Contact map coverage

The fraction of residues covered by a set of contact pairs (N$_{\text{map}}$) out of the total number of residues in the target sequence (N$_{\text{sequence}}$) (Eq. (2.2)).

$$Cov = \frac{N_{map}}{N_{sequence}} \tag{2.2}$$

2.3.2.2 Contact map precision

The precision of a set of contact pairs is equivalent to the the proportion of True Positive (TP) contact pairs in the overall set (Eq. (2.3)). A contact pair was defined as TP if the equivalent Cβ (Cα in case of Gly) atoms in the native crystal structure were < 8Å apart. The precision value is in range [0, 1], whereby a value of 1 means all contact pairs are TPs.

$$Prec = \frac{TP}{TP - FP} \tag{2.3}$$

If contacts were unmatched between the target sequence and reference structure, they were not taken into account in the calculation of the precision score.

2.3.2.3 Contact map Jaccard index

The Jaccard index quantifies the similarity between two sets of contact pairs. It describes the proportion of contact pairs in the intersection compared to the union between the two sets [135] (Eq. (2.4)).

$$J_{x,y} = \frac{|x \cap y|}{|x \cup y|} \tag{2.4}$$

The variables x and y are two sets of contact pairs. $|x \cap y|$ is the number of elements in the intersection of x and y, and the $|x \cup y|$ represents the number of elements in the union of x and y. The Jaccard index falls in the range [0,1], with a value of 1 corresponding to identical sets of contact pairs and 0 to non-identical ones. It is worth noting that only exact matches are considered and the neighbourhood of a single contact ignored.

2.3.2.4 Contact map singleton content

Almost all sliced sets of residue-residue contact pairs contain a fraction of contact pairs not co-localising with others. These contact pairs — referred to as singleton contact pairs from here onwards — typically show a high False Positive (FP) rate and could be considered noise (although sometimes they encode TP contacts in an oligomeric interface). To quantify this fraction, a distance-based clustering analysis was defined to identify singleton contact pairs, and thus describe the level of noise in the prediction, or alternatively how well contact pairs co-localise typically between secondary structure features.

To identify singleton contact pairs in a set of contacts, the neighbourhood of each pair was searched for the presence of other contacts. The search radius was defined by ±2 residues in a 2D-representation of the contact map. If no other contact pair was identified under such constraint, the contact pair was classified as singleton.

2.3.3 Structure prediction data

2.3.3.1 Root Mean Squared Deviatio

The Root Mean Square Deviation (RMSD) is a measure to quantify the average atomic distance between two protein structures (Eq. (2.5)). The RMSD is sequence-independent, and measures the distance between Cα atoms.

$$RMSD = \sqrt{\frac{1}{n}\sum_{i,j}(x_i - x_j)^2 + (y_i - y_j)^2 + (z_i - z_j)^2} \tag{2.5}$$

2.3.3.2 Template-Modelling score

The Template-Modelling score (TM-score) is a more accurate measure of structure similarity between two protein structures than the RMSD [136]. Unlike the RMSD, the TM-score score assigns a lenght-dependent weight to the distances between atoms, with shorter distances getting assigned stronger weights [136]. The TM-score has widely been accepted as a standard for assessing the similarity between two structures, particularly in the field of *ab initio* structure prediction.

$$TMscore = max\left[\frac{1}{L_{target}}\sum_{i}^{L_{aligned}}\frac{1}{1+\left(\frac{d_i}{d_0}\right)^2}\right] \tag{2.6}$$

d_i describes the distance between the ith pair of residues. The distance scale d_0 to normalise the distances is defined by the equation $1.24\sqrt[3]{L_{target} - 15} - 1.8$. The TM-score value falls in the range (0, 1]. A TM-score value of < 0.2 indicates two random unrelated structures, and a value > 0.5 roughly the same fold [137]

2.3.3.3 Long-range contact precision

The long-range contact precision score is computed identically to the precision of sets of contact pairs (Section 2.3.2.2). However, the precision score is computed solely for long-range contacts ($>$ 23 residues sequence separation).

2.3.4 Molecular Replacement data

2.3.4.1 Register-Independent Overlap

The Residue-Independent Overlap (RIO) score [138] is a measure of structural similarity between two protein structures considering the total number of atoms within < 1.5Å. The RIO can be separated into the in- ($\mathrm{RIO_{in}}$) and out-of-register ($\mathrm{RIO_{out}}$) score considering the sequence register between the model and the target. The RIO score is primarily a measure for post-MR search models to assess the placement of search model atoms with respect to the previously solved crystal structure. To avoid the addition of single atoms place correctly purely by chance, the RIO metric requires at least three consecutive Cα atoms to be within th 1.5Å threshold.

2.3.4.2 Structure solution

MR structure solutions were assessed throughout all works presented in this thesis by the Correlation Coefficient (CC) [139] and Average Chain Length (ACL) scores computed by SHELXE. SHELXE performs density modification and main-chain tracing of the refined MR solution [140]. Thorn and Sheldrick [140] highlighted in their work that a CC of $\geq 25\%$ indicates a successful structure solution. Additionally, previous research with Ab initio Modelling of Proteins for moLEcular replacement (AMPLE) [138] has shown that an ACL of the trace needs to be ≥ 10 residues.

In most studies in this thesis, additionally to the SHELXE metrics the post-SHELXE auto-built structures needed R values of ≤ 0.45. The R values had to be acquired by at least one of the Buccaneer [31] or ARP/wARP [141] solutions.

Lastly, the PHASER Translation Function Z-score (TFZ) and Log-Likelihood Gain (LLG) metrics were also considered when automatically judging a MR solution. Values of > 8 and > 120 were required, respectively. However, the PHASER metrics do not always indicate a structure solution — particularly for smaller fragments — and thus was not considered an essential metric to pass to be considered a successful solution.

Bibliography

[1] W Friedrich, P Knipping, M Laue, *Ann. Phys.* **1913**, *346*, 971–988.

[2] M Laue, *Ann. Phys.* **1913**, *346*, 989–1002.

[3] W. H. Bragg, W. L. Bragg, *Proceedings of the Royal Society A: Mathematical Physical and Engineering Sciences* **July 1913**, *88*, 428–438.

[4] W. L. Bragg, *Scientia* **1929**, *23*, 153.

[5] W. L. Bragg, *Nature* **Dec. 1912**, *90*, 410.

[6] J. D. Watson, F. H. C. Crick, Others, *Nature* **1953**, *171*, 737–738.

[7] D. C. Hodgkin, J Kamper, M Mackay, J Pickworth, K. N. Trueblood, J. G. White, en, *Nature* **July 1956**, *178*, 64–66.

[8] T. L. Blundell, J. F. Cutfield, S. M. Cutfield, E. J. Dodson, G. G. Dodson, D. C. Hodgkin, D. A. Mercola, M Vijayan, en, *Nature* **June 1971**, *231*, 506–511.

[9] C. C. Blake, D. F. Koenig, G. A. Mair, A. C. North, D. C. Phillips, V. R. Sarma, en, *Nature* **May 1965**, *206*, 757–761.

[10] M. F. Perutz, M. G. Rossmann, A. F. Cullis, H Muirhead, G Will, A. C. North, en, *Nature* **Feb. 1960**, *185*, 416–422.

[11] J. C. Kendrew, G Bodo, H. M. Dintzis, R. G. Parrish, H Wyckoff, D. C. Phillips, en, *Nature* **Mar. 1958**, *181*, 662–666.

[12] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, P. E. Bourne, *Nucleic Acids Res.* **Jan. 2000**, *28*, 235–242.

[13] B. Rupp, *Biomolecular crystallography : principles, practice, and application to structural biology*, English, Garland Science, New York, **2010**.

[14] M. G. Rossmann, *Acta Crystallogr. D Biol. Crystallogr.* **2001**, *57*, 1360–1366.

[15] M. G. Rossmann, en, *Acta Crystallogr. A* **Feb. 1990**, *46 (Pt 2)*, 73–82.

[16] C. R. Kissinger, D. K. Gehlhaar, D. B. Fogel, en, *Acta Crystallogr. D Biol. Crystallogr.* **Feb. 1999**, *55*, 484–491.

[17] N. M. Glykos, M Kokkinidis, en, *Acta Crystallogr. D Biol. Crystallogr.* **Feb. 2000**, *56*, 169–174.

[18] R. J. Read, en, *Acta Crystallogr. D Biol. Crystallogr.* **Oct. 2001**, *57*, 1373–1382.

[19] M. G. Rossmann, D. M. Blow, *Acta Crystallogr.* **Jan. 1962**, *15*, 24–31.

[20] M. Bayes, M. Price, *Philosophical Transactions of the Royal Society of London* **Jan. 1763**, *53*, 370–418.

[21] L. C. Storoni, A. J. McCoy, R. J. Read, en, *Acta Crystallogr. D Biol. Crystallogr.* **Mar. 2004**, *60*, 432–438.

[22] B. C. Wang, en, *Methods Enzymol.* **1985**, *115*, 90–112.

[23] V. Y. Lunin, *Acta Crystallogr. A* **Mar. 1988**, *44*, 144–150.

[24] G. M. Sheldrick, *Zeitschrift für Kristallographie - Crystalline Materials* **Jan. 2002**, *217*, 371.

[25] T. C. Terwilliger, en, *Acta Crystallogr. D Biol. Crystallogr.* **Aug. 2000**, *56*, 965–972.

[26] P. V. Afonine, R. W. Grosse-Kunstleve, N. Echols, J. J. Headd, N. W. Moriarty, M. Mustyakimov, T. C. Terwilliger, A. Urzhumtsev, P. H. Zwart, P. D. Adams, en, *Acta Crystallogr. D Biol. Crystallogr.* **Apr. 2012**, *68*, 352–367.

[27] G. N. Murshudov, P. Skubák, A. A. Lebedev, N. S. Pannu, R. A. Steiner, R. A. Nicholls, M. D. Winn, F. Long, A. A. Vagin, en, *Acta Crystallogr. D Biol. Crystallogr.* **Apr. 2011**, *67*, 355–367.

[28] G. M. Sheldrick, en, *Acta Crystallogr. D Biol. Crystallogr.* **Apr. 2010**, *66*, 479–485.

[29] V. S. Lamzin, A Perrakis, K. S. Wilson, *International Tables for Crystallography* **2001**, 720–722.

[30] T. Terwilliger, en, *J. Synchrotron Radiat.* **Jan. 2004**, *11*, 49–52.

[31] K. Cowtan, en, *Acta Crystallogr. D Biol. Crystallogr.* **Sept. 2006**, *62*, 1002–1011.

[32] B. Qian, S. Raman, R. Das, P. Bradley, A. J. McCoy, R. J. Read, D. Baker, en, *Nature* **Nov. 2007**, *450*, 259–264.

[33] D. J. Rigden, R. M. Keegan, M. D. Winn, en, *Acta Crystallogr. D Biol. Crystallogr.* **Dec. 2008**, *64*, 1288–1291.

[34] R. Das, D. Baker, en, *Acta Crystallogr. D Biol. Crystallogr.* **Feb. 2009**, *65*, 169–175.

[35] P. E. Leopold, M Montal, J. N. Onuchic, en, *Proc. Natl. Acad. Sci. U. S. A.* **Sept. 1992**, *89*, 8721–8725.

[36] C. B. Anfinsen, en, *Science* **July 1973**, *181*, 223–230.

[37] C. Levinthal, *Mossbauer spectroscopy in biological systems* **1969**, *67*, 22–24.

[38] M. Karplus, en, *Nat. Chem. Biol.* **June 2011**, *7*, 401–404.

[39] Wikipedia, Folding Funnel — Wikipedia, The Free Encyclopedia, [Online; accessed 09-April-2018], **2004**.

[40] J. Lee, P. L. Freddolino, Y. Zhang in *From Protein Structure to Function with Bioinformatics (2nd Ed.) Vol. 69*, (Ed.: D. J. Rigden), Springer Netherlands, Dordrecht, **2017**, pp. 3–35.

[41] J. Skolnick, en, *Curr. Opin. Struct. Biol.* **Apr. 2006**, *16*, 166–171.

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

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

[44] M. Blaszczyk, M. Jamroz, S. Kmiecik, A. Kolinski, en, *Nucleic Acids Res.* **July 2013**, *41*, W406–11.

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

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

[47] J. Abbass, J.-C. Nebel, en, *BMC Bioinformatics* **Apr. 2015**, *16*, 136.

[48] Y. Shen, G. Picord, F. Guyon, P. Tuffery, en, *PLoS One* **Nov. 2013**, *8*, e80493.

[49] S. C. Li, D. Bu, X. Gao, J. Xu, M. Li, en, *Bioinformatics* **July 2008**, *24*, i182–9.

[50] I. Kalev, M. Habeck, en, *Bioinformatics* **Nov. 2011**, *27*, 3110–3116.

[51] D. Bhattacharya, B. Adhikari, J. Li, J. Cheng, en, *Bioinformatics* **July 2016**, *32*, 2059–2061.

[52] T. Wang, Y. Yang, Y. Zhou, H. Gong, en, *Bioinformatics* **Mar. 2017**, *33*, 677–684.

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

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

[55] N Metropolis, S Ulam, en, *J. Am. Stat. Assoc.* **Sept. 1949**, *44*, 335–341.

[56] D Shortle, K. T. Simons, D Baker, en, *Proc. Natl. Acad. Sci. U. S. A.* **Sept. 1998**, *95*, 11158–11162.

[57] Y. Zhang, J. Skolnick, *J. Comput. Chem.* **2004**, *25*, 865–871.

[58] P. Bradley, K. M. S. Misura, D. Baker, en, *Science* **Sept. 2005**, *309*, 1868–1871.

[59] S Ołdziej, C Czaplewski, A Liwo, M Chinchio, M Nanias, J. A. Vila, M Khalili, Y. A. Arnautova, A Jagielska, M Makowski, H. D. Schafroth, R Kaźmierkiewicz, D. R. Ripoll, J Pillardy, J. A. Saunders, Y. K. Kang, K. D. Gibson, H. A. Scheraga, en, *Proc. Natl. Acad. Sci. U. S. A.* **May 2005**, *102*, 7547–7552.

[60] Y. Zhang, J. Skolnick, en, *Proc. Natl. Acad. Sci. U. S. A.* **May 2004**, *101*, 7594–7599.

[61] J. Yang, R. Yan, A. Roy, D. Xu, J. Poisson, Y. Zhang, en, *Nat. Methods* **Jan. 2015**, *12*, 7–8.

[62] A. Kryshtafovych, A. Barbato, B. Monastyrskyy, K. Fidelis, T. Schwede, A. Tramontano, en, *Proteins* **Sept. 2016**, *84 Suppl 1*, 349–369.

[63] W. Kabsch, C. Sander, en, *Biopolymers* **Dec. 1983**, *22*, 2577–2637.

[64] J. Vojtěchovský, K. Chu, J. Berendzen, R. M. Sweet, I. Schlichting, en, *Biophys. J.* **Oct. 1999**, *77*, 2153–2174.

[65] H. Eklund, M. Ingelman, B. O. Söderberg, T. Uhlin, P. Nordlund, M. Nikkola, U. Sonnerstam, T. Joelson, K. Petratos, en, *J. Mol. Biol.* **Nov. 1992**, *228*, 596–618.

[66] F. K. Athappilly, W. A. Hendrickson, en, *Structure* **Dec. 1995**, *3*, 1407–1419.

[67] S. Bañuelos, M. Saraste, K. D. Carugo, en, *Structure* **Nov. 1998**, *6*, 1419–1431.

[68] A. H. West, E. Martinez-Hackert, A. M. Stock, en, *J. Mol. Biol.* **July 1995**, *250*, 276–290.

[69] J. Ménétrey, E. Macia, S. Pasqualato, M. Franco, J. Cherfils, en, *Nat. Struct. Biol.* **June 2000**, *7*, 466–469.

[70] C. C. Thomas, S Dowler, M Deak, D. R. Alessi, D. M. van Aalten, en, *Biochem. J.* **Sept. 2001**, *358*, 287–294.

[71] S. Grizot, F. Fieschi, M. C. Dagher, E. Pebay-Peyroula, en, *J. Biol. Chem.* **June 2001**, *276*, 21627–21631.

[72] P. Sörme, P. Arnoux, B. Kahl-Knutsson, H. Leffler, J. M. Rini, U. J. Nilsson, en, *J. Am. Chem. Soc.* **Feb. 2005**, *127*, 1737–1743.

[73] C. A. Kim, M. Gingery, R. M. Pilpa, J. U. Bowie, en, *Nat. Struct. Biol.* **June 2002**, *9*, 453–457.

[74] J. B. Thoden, H. M. Holden, Z. Zhuang, D. Dunaway-Mariano, en, *J. Biol. Chem.* **July 2002**, *277*, 27468–27476.

[75] X. Zhang, J.-C. D. Schwartz, X. Guo, S. Bhatia, E. Cao, M. Lorenz, M. Cammer, L. Chen, Z.-Y. Zhang, M. A. Edidin, S. G. Nathenson, S. C. Almo, en, *Immunity* **Mar. 2004**, *20*, 337–347.

[76] B. A. Fields, H. H. Bartsch, H. D. Bartunik, F Cordes, J. M. Guss, H. C. Freeman, en, *Acta Crystallogr. D Biol. Crystallogr.* **Sept. 1994**, *50*, 709–730.

[77] Y. Cheng, S. M. Sequeira, L. Malinina, V. Tereshko, T. H. Söllner, D. J. Patel, en, *Protein Sci.* **Oct. 2004**, *13*, 2665–2672.

[78] M. Graille, C. Z. Zhou, V. Receveur-Bréchot, B. Collinet, N. Declerck, H. Van Tilbeurgh, en, *J. Biol. Chem.* **Apr. 2005**, *280*, 14780–14789.

[79] T. Merz, S. K. Wetzel, S. Firbank, A. Plückthun, M. G. Grütter, P. R. E. Mittl, en, *J. Mol. Biol.* **Feb. 2008**, *376*, 232–240.

[80] D. A. K. Traore, A. J. Brennan, R. H. P. Law, C. Dogovski, M. A. Perugini, N. Lukoyanova, E. W. W. Leung, R. S. Norton, J. A. Lopez, K. A. Browne, H. Yagita, G. J. Lloyd, A. Ciccone, S. Verschoor, J. A. Trapani, J. C. Whisstock, I. Voskoboinik, en, *Biochem. J.* **Dec. 2013**, *456*, 323–335.

[81] S. J. Atkinson, P. E. Soden, D. C. Angell, M. Bantscheff, C.-W. Chung, K. A. Giblin, N. Smithers, R. C. Furze, L. Gordon, G. Drewes, I. Rioja, J. Witherington, N. J. Parr, R. K. Prinjha, en, *Med. Chem. Commun.* **Feb. 2014**, *5*, 342–351.

[82] B. T. Porebski, A. A. Nickson, D. E. Hoke, M. R. Hunter, L. Zhu, S. McGowan, G. I. Webb, A. M. Buckle, en, *Protein Eng. Des. Sel.* **Mar. 2015**, *28*, 67–78.

[83] A. M. Crowe, P. J. Stogios, I. Casabon, E. Evdokimova, A. Savchenko, L. D. Eltis, en, *J. Biol. Chem.* **Jan. 2015**, *290*, 872–882.

[84] R. D. Finn, P. Coggill, R. Y. Eberhardt, S. R. Eddy, J. Mistry, A. L. Mitchell, S. C. Potter, M. Punta, M. Qureshi, A. Sangrador-Vegas, G. A. Salazar, J. Tate, A. Bateman, en, *Nucleic Acids Res.* **Jan. 2016**, *44*, D279–D285.

[85] J. M. Chandonia, N. K. Fox, S. E. Brenner, en, *J. Mol. Biol.* **Feb. 2017**, *429*, 348–355.

[86] G. E. Tusnády, Z. Dosztányi, I. Simon, en, *Nucleic Acids Res.* **Jan. 2005**, *33*, D275–8.

[87] B. P. Klaholz, A. Mitschler, D. Moras, en, *J. Mol. Biol.* **Sept. 2000**, *302*, 155–170.

[88] W. T. Lowther, N Brot, H Weissbach, B. W. Matthews, en, *Biochemistry* **Nov. 2000**, *39*, 13307–13312.

[89] R. O. Louro, I. Bento, P. M. Matias, T. Catarino, A. M. Baptista, C. M. Soares, M. A. Carrondo, D. L. Turner, A. V. Xavier, en, *J. Biol. Chem.* **Nov. 2001**, *276*, 44044–44051.

[90] P. Kuser, D. R. Hall, L. H. Mei, M. Neu, R. W. Evans, P. F. Lindley, en, *Acta Crystallogr. D Biol. Crystallogr.* **May 2002**, *58*, 777–783.

[91] I. Hayashi, K. Vuori, R. C. Liddington, en, *Nat. Struct. Biol.* **Feb. 2002**, *9*, 101–106.

[92] G. David, K. Blondeau, M. Schiltz, S. Penel, A. Lewit-Bentley, en, *J. Biol. Chem.* **Oct. 2003**, *278*, 43728–43735.

[93] V. Oganesyan, D. Busso, J. Brandsen, S. Chen, J. Jancarik, R. Kim, S. H. Kim, en, *Acta Crystallographica - Section D Biological Crystallography* **July 2003**, *59*, 1219–1223.

[94] J. Liu, H. Yokota, R. Kim, S. H. Kim, en, *Proteins: Structure Function and Genetics* **June 2004**, *55*, 1082–1086.

[95] L. Cendron, A. Seydel, A. Angelini, R. Battistutta, G. Zanotti, en, *J. Mol. Biol.* **July 2004**, *340*, 881–889.

[96] L. Malinina, M. L. Malakhova, A. T. Kanack, M. Lu, R. Abagyan, R. E. Brown, D. J. Patel, en, *PLoS Biol.* **Nov. 2006**, *4*, 1996–2011.

[97] K. Makabe, S. Yan, V. Tereshko, G. Gawlak, S. Koide, en, *J. Am. Chem. Soc.* **Nov. 2007**, *129*, 14661–14669.

[98] A. W. Maresso, R. Wu, J. W. Kern, R. Zhang, D. Janik, D. M. Missiakas, M. E. Duban, A. Joachimiak, O. Schneewind, en, *J. Biol. Chem.* **Aug. 2007**, *282*, 23129–23139.

[99] C. U. Stirnimann, D. Ptchelkine, C. Grimm, C. W. Müller, en, *J. Mol. Biol.* **July 2010**, *400*, 71–81.

[100] L. Von Schantz, M. Håkansson, D. T. Logan, B. Walse, J. Osterlin, E. Nordberg-Karlsson, M. Ohlin, M. Hkansson, D. T. Logan, B. Walse, J. Österlin, E. Nordberg-Karlsson, M. Ohlin, en, *Glycobiology* **July 2012**, *22*, 948–961.

[101] S. J. Coulthurst, A. Dawson, W. N. Hunter, F. Sargent, en, *Biochemistry* **Feb. 2012**, *51*, 1678–1686.

[102] M. Myllykoski, A. Raasakka, M. Lehtimäki, H. Han, I. Kursula, P. Kursula, en, *J. Mol. Biol.* **Nov. 2013**, *425*, 4307–4322.

[103] X. Yang, M. Morita, H. Wang, T. Suzuki, W. Yang, Y. Luo, C. Zhao, Y. Yu, M. Bartlam, T. Yamamoto, Z. Rao, en, *Nucleic Acids Res.* **Dec. 2008**, *36*, 6872–6881.

[104] J. C. Grigg, C. X. Mao, M. E. P. Murphy, en, *J. Mol. Biol.* **Oct. 2011**, *413*, 684–698.

[105] H. Repo, J. S. Oeemig, J. Djupsjöbacka, H. Iwaï, P. Heikinheimo, en, *Acta Crystallogr. D Biol. Crystallogr.* **Nov. 2012**, *68*, 1479–1487.

[106] M. F. Matos, Y. Xu, I. Dulubova, Z. Otwinowski, J. M. Richardson, D. R. Tomchick, J. Rizo, A. Ho, en, *Proc. Natl. Acad. Sci. U. S. A.* **Mar. 2012**, *109*, 3802–3807.

[107] S. Moréra, I. Grin, A. Vigouroux, S. Couvé, V. Henriot, M. Saparbaev, A. A. Ishchenko, en, *Nucleic Acids Res.* **Oct. 2012**, *40*, 9917–9926.

[108] P. M. Collins, K. Bum-Erdene, X. Yu, H. Blanchard, en, *J. Mol. Biol.* **Apr. 2014**, *426*, 1439–1451.

[109] T. Weinert, V. Olieric, S. Waltersperger, E. Panepucci, L. Chen, H. Zhang, D. Zhou, J. Rose, A. Ebihara, S. Kuramitsu, D. Li, N. Howe, G. Schnapp, A. Pautsch, K. Bargsten, A. E. Prota, P. Surana, J. Kottur, D. T. Nair, F. Basilico, V. Cecatiello, S. Pasqualato, A. Boland, O. Weichenrieder, B. C. Wang, M. O. Steinmetz, M. Caffrey, M. Wang, en, *Nat. Methods* **Feb. 2015**, *12*, 131–133.

[110] J. M. H. Thomas, F. Simkovic, R. M. Keegan, O. Mayans, Y. Zhang, D. J. Rigden, C. Zhang, Y. Zhang, D. J. Rigden, *Acta Crystallographica Section D Structural Biology* **Dec. 2017**, *73*, 985–996.

[111] K. Edman, A. Royant, P. Nollert, C. A. Maxwell, E. Pebay-Peyroula, J. Navarro, R. Neutze, E. M. Landau, en, *Structure* **Apr. 2002**, *10*, 473–482.

[112] J. Standfuss, A. C. T. Van Scheltinga, M. Lamborghini, W. Kühlbrandt, en, *EMBO J.* **Mar. 2005**, *24*, 919–928.

[113] J. K. Lee, D. Kozono, J. Remis, Y. Kitagawa, P. Agre, R. M. Stroud, en, *Proc. Natl. Acad. Sci. U. S. A.* **Dec. 2005**, *102*, 18932–18937.

[114] D. F. Savage, R. M. Stroud, en, *J. Mol. Biol.* **May 2007**, *368*, 607–617.

[115] D. Pogoryelov, Ö. Yildiz, J. D. Faraldo-Gómez, T. Meier, en, *Nat. Struct. Mol. Biol.* **Oct. 2009**, *16*, 1068–1073.

[116] K. R. Vinothkumar, K. Strisovsky, A. Andreeva, Y. Christova, S. Verhelst, M. Freeman, en, *EMBO J.* **Nov. 2010**, *29*, 3797–3809.

[117] J. D. Ho, R. Yeh, A. Sandstrom, I. Chorny, W. E. C. Harries, R. A. Robbins, L. J. W. Miercke, R. M. Stroud, en, *Proceedings of the National Academy of Sciences* **May 2009**, *106*, 7437–7442.

[118] N. H. Joh, A. Oberai, D. Yang, J. P. Whitelegge, J. U. Bowie, en, *J. Am. Chem. Soc.* **Aug. 2009**, *131*, 10846–10847.

[119] S. Ye, Y. Li, Y. Jiang, en, *Nat. Struct. Mol. Biol.* **Aug. 2010**, *17*, 1019–1023.

[120] M. G. Derebe, D. B. Sauer, W. Zeng, A. Alam, N. Shi, Y. Jiang, en, *Proceedings of the National Academy of Sciences* **Jan. 2011**, *108*, 598–602.

[121] H. Saino, Y. Ukita, H. Ago, D. Irikura, A. Nisawa, G. Ueno, M. Yamamoto, Y. Kanaoka, B. K. Lam, K. F. Austen, M. Miyano, en, *J. Biol. Chem.* **May 2011**, *286*, 16392–16401.

[122] G. B. Erkens, R. P. A. Berntsson, F. Fulyani, M. Majsnerowska, A. Vujiĉić-Žagar, J. Ter Beek, B. Poolman, D. J. Slotboom, en, *Nat. Struct. Mol. Biol.* **June 2011**, *18*, 755–760.

[123] J. Symersky, V. Pagadala, D. Osowski, A. Krah, T. Meier, J. D. Faraldo-Gómez, D. M. Mueller, en, *Nat. Struct. Mol. Biol.* **Apr. 2012**, *19*, 485–91, S1.

[124] R. P.-A. Berntsson, J. ter Beek, M. Majsnerowska, R. H. Duurkens, P. Puri, B. Poolman, D.-J. Slotboom, en, *Proceedings of the National Academy of Sciences* **Aug. 2012**, *109*, 13990–13995.

[125] S. Hayat, C. Sander, D. S. Marks, A. Elofsson, en, *Proc. Natl. Acad. Sci. U. S. A.* **Apr. 2015**, *112*, 5413–5418.

[126] J. Andreani, J. Söding, en, *Bioinformatics* **June 2015**, *31*, 1729–1737.

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

[128] A. Bateman, M. J. Martin, C. O'Donovan, M. Magrane, E. Alpi, R. Antunes, B. Bely, M. Bingley, C. Bonilla, R. Britto, B. Bursteinas, H. Bye-AJee, A. Cowley, A. Da Silva, M. De Giorgi, T. Dogan, F. Fazzini, L. G. Castro, L. Figueira, P. Garmiri, G. Georghiou, D. Gonzalez, E. Hatton-Ellis, W. Li, W. Liu, R. Lopez, J. Luo, Y. Lussi, A. MacDougall, A. Nightingale, B. Palka, K. Pichler, D. Poggioli, S. Pundir, L. Pureza, G. Qi, S. Rosanoff, R. Saidi, T. Sawford, A. Shypitsyna, E. Speretta, E. Turner, N. Tyagi, V. Volynkin, T. Wardell, K. Warner, X. Watkins, R. Zaru, H. Zellner, I. Xenarios, L. Bougueleret, A. Bridge, S. Poux, N. Redaschi, L. Aimo, G. ArgoudPuy, A. Auchincloss, K. Axelsen, P. Bansal, D. Baratin, M. C.

Blatter, B. Boeckmann, J. Bolleman, E. Boutet, L. Breuza, C. Casal-Casas, E. De Castro, E. Coudert, B. Cuche, M. Doche, D. Dornevil, S. Duvaud, A. Estreicher, L. Famiglietti, M. Feuermann, E. Gasteiger, S. Gehant, V. Gerritsen, A. Gos, N. Gruaz-Gumowski, U. Hinz, C. Hulo, F. Jungo, G. Keller, V. Lara, P. Lemercier, D. Lieberherr, T. Lombardot, X. Martin, P. Masson, A. Morgat, T. Neto, N. Nouspikel, S. Paesano, I. Pedruzzi, S. Pilbout, M. Pozzato, M. Pruess, C. Rivoire, B. Roechert, M. Schneider, C. Sigrist, K. Sonesson, S. Staehli, A. Stutz, S. Sundaram, M. Tognolli, L. Verbregue, A. L. Veuthey, C. H. Wu, C. N. Arighi, L. Arminski, C. Chen, Y. Chen, J. S. Garavelli, H. Huang, K. Laiho, P. McGarvey, D. A. Natale, K. Ross, C. R. Vinayaka, Q. Wang, Y. Wang, L. S. Yeh, J. Zhang, en, *Nucleic Acids Res.* **Jan. 2017**, *45*, D158–D169.

[129] S. Seemayer, M. Gruber, J. S??ding, en, *Bioinformatics* **Nov. 2014**, *30*, 3128–3130.

[130] J. Söding, en, *Bioinformatics* **Apr. 2005**, *21*, 951–960.

[131] F. Morcos, A. Pagnani, B. Lunt, A. Bertolino, D. S. Marks, C. Sander, R. Zecchina, J. N. Onuchic, T. Hwa, M. Weigt, en, *Proceedings of the National Academy of Sciences* **Dec. 2011**, *108*, E1293–E1301.

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

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

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

[135] Q. Wuyun, W. Zheng, Z. Peng, J. Yang, *Brief. Bioinform.* **Oct. 2016**, bbw106.

[136] Y. Zhang, J. Skolnick, en, *Proteins: Structure Function and Genetics* **Dec. 2004**, *57*, 702–710.

[137] J. Xu, Y. Zhang, en, *Bioinformatics* **Apr. 2010**, *26*, 889–895.

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

[139] M Fujinaga, R. J. Read, *J. Appl. Crystallogr.* **Dec. 1987**, *20*, 517–521.

[140] A. Thorn, G. M. Sheldrick, en, *Acta Crystallogr. D Biol. Crystallogr.* **Nov. 2013**, *69*, 2251–2256.

[141] S. X. Cohen, M. Ben Jelloul, F. Long, A. Vagin, P. Knipscheer, J. Lebbink, T. K. Sixma, V. S. Lamzin, G. N. Murshudov, A. Perrakis, en, *Acta Crystallogr. D Biol. Crystallogr.* **Jan. 2007**, *64*, 49–60.