universität wien

DIPLOMARBEIT / DIPLOMA THESIS

Titel der Diplomarbeit / Title of the Diploma Thesis

„Developing a Workflow for the Validation and Correction of PDB Macromolecular and Ligand Structure Data“

verfasst von / submitted by

Stefan Comployer

angestrebter akademischer Grad / in partial fulfilment of the requirements for the degree of

Magister der Pharmazie (Mag.pharm.)

Wien, 2016 / Vienna, 2016

Studienkennzahl lt. Studienblatt / degree programme code as it appears on the student record sheet:	A 449
Studienrichtung lt. Studienblatt / degree programme as it appears on the student record sheet:	Diplomstudium Pharmazie
Betreut von / Supervisor:	Univ.-Prof. Dr. Thierry Langer

Abstract

Three-dimensional models of macromolecular structures are frequently used as the basis for computational studies. Currently x-ray crystallography is considered to be the most accurate method of structure analysis. However, due to the inherent limitations of these experiments, models are often incomplete or contain errors resulting from misinterpretation of experimental data. The aim of this thesis was to investigate the most common errors found in such structures and to provide a guideline for analyzing and preparing PDB files for further use in simulations. In order to achieve this, structures were examined for deviations from a variety of validation parameters, fit of binding sites and ligands to their electron density map, coordinates missing from the model, incorrectly assigned rotamers and protonation states of ionizable entities. If possible the identified errors were corrected immediately. This was done using available open source software and web services. The results showed that the majority of PDB files would benefit from refinement prior to their usage in computational experiments. Based on the results of this work we propose a workflow consisting of 6 different checks that should guide novices and people not familiar with this field through the initial steps of PDB file preparation.

Zusammenfassung

3D Modelle von Makromolekülen dienen als Grundlage für eine Vielzahl von computerbasierten Studien. Durch das hohe Ausmaß an Genauigkeit ist die Röntgenstrukturanalyse derzeit meist das Verfahren der Wahl um die Struktur von Biomolekülen aufzuklären. Trotz aller Vorteile, sind aber auch die Möglichkeiten dieser Methode limitiert. So sind Modelle oft unvollständig oder enthalten Fehler, die aus Missinterpretationen experimenteller Daten resultieren. Das Ziel dieser Arbeit war es, die am häufigsten auftretenden Fehler in solchen Strukturen zu identifizieren und einen Leitfaden für deren Analyse und Korrektur zu erstellen. Um das zu erreichen, wurden diverse Strukturen auf Abweichungen von verschiedenen Validierungsparametern, die Übereinstimmung von Bindungsstellen und Liganden der Modelle mit ihren Elektronendichtekarten, fehlende Modellkoordinaten, falsch zugeordnete Konformationen von Rotameren und Protonierungszustände ionisierbarer Komponenten untersucht. Dabei gefundene Fehler wurden, soweit die Möglichkeit bestand, sofort mit verfügbarer lizenzfreier Software oder Web-Diensten korrigiert. Die Ergebnisse dieser Arbeit legen nahe, dass die meisten PDB Dateien von einer Optimierung vor der Nutzung in computerbasierten Experimenten profitieren würden. Außerdem wurde, basierend auf den Ergebnissen dieser Arbeit, ein Workflow mit 6 Schritten erstellt der als Leitfaden für die Vorbereitung von PDB Dateien für Anfänger beziehungsweise mit diesem Feld weniger vertraute Personen fungieren soll.

Contents

Introduction....1
1.1 The Protein Data Bank....2
1.2 X-ray Crystallography....3
1.2.1 Crystallization....4
1.2.2 Obtaining Diffraction Pattern and Electron Density Map....5
1.2.3 Model Building....6
1.2.4 Resulting Problems....9
1.3 Quality Metrics of the 3D Structure Model....10
1.3.1 Global Quality Parameters....10
1.3.2 Local Quality Parameters....15
1.4 Missing Residues and Missing Atoms....21
1.5 Flipped Sidechains....26
1.6 Protonation State....29
Methods....33
2.1 Validation of Selected Input Structures....33
2.2 Evaluation of the Ligands Fit to the Electron Density Map....33
2.3 Modelling Missing Coordinates....34
2.4 Correction of Wrongly Assigned ASN and GLN Residues....35
2.5 Calculating pK_a Values of Ionizable Residues and Ligands....35
2.6 Applying Charges to Ligands....36
Results and Discussion....37
3.1 Model Validation....39
3.2 Evaluation of the Ligands Fit to the Electron Density Map....40
3.3 Assessing and Modelling Missing Coordinates....44
2.4 Flipped ASN and GLN Rotamers....65
2.5 pK_a Calculations....82
2.6 Applying Charges to Ligands....90
Conclusion....94
Bibliography....97

List of Figures

1.1 Amount of structures currently present in the Protein Data Bank....2
1.2 Steps involved in solving a protein structure via x-ray crystallography....3
1.3 Format of a PDB file....6
1.4 Myoglobin structure (PDB entry 1mbi) solved at 2.0 Å resolution....7
1.5 Myoglobin structure (PDB entry 1a6m) solved at atomic resolution (1.0 Å)....9
1.6 Appearance of electron density at different resolutions of experimental data shown on the example of the N-terminal fragment of lysozyme....11
1.7 Scale of values for resolution achieved by x-ray experiments....12
1.8 Scale of R-factor values....13
1.9 Scale of difference between Rfree and R values....14

1.10 Behavior of R and Rfree during the refinement of human immunoglobulin (IgG) and the C2 domain of protein G at resolutions ~ 3.5 Å 14
1.11 Per-residue plot of RSR values for retinoic acid binding proteins I&II (PDB entry 1CBS) as provided by the EDS 16
1.12 Schematic representation of torsion angles of the protein backbone 19
1.13 Examples of a Ramachandran Plot 20
1.14 Distribution of all missing residues in PDB protein structures 21
1.15 Remarks for missing coordinates in PDB files 22
1.16 Structure of α map kinase inhibitor (PDB entry 2QD9) before and after modelling the sections of missing coordinates 23
1.17 Illustration of a “flipped” asparagine sidechain in the structure of abl kinase domain (PDB entry 1HZI) 26
1.18 Six possible rotameric and protonation states of histidine with marked formal charges on nitrogen in doubly protonated HIP states 28
1.19 Characteristics of proteins with ionizable sidechains 30
1.20 Factors influencing the pK_a of ionizable groups in proteins 31

2.1 Process of classification as implemented in VHELIBS for evaluating a models fit to its electron density 34

3.1 Workflow for validation and preparation of PDB files 38
3.2 Example for “Bad” /”Good” fit to electron density as labeled by VHELIBS 40
3.3 Distribution of ligands and their binding sites for default- (PDB) and custom settings 43
3.4 Occurrence of missing strings of certain length 45
3.5 Example for alignment file in the PIR format 46
3.6 Distribution of residues with unexpected pK_a values calculated by PROPKA 86
3.7 Steps of ligand preparation for MD input 90

List of Tables

3.1 Evaluation of ligands and binding sites using VHELIBS 42
3.2 Missing atoms and missing residues of PDB files within the test set 63
3.3 Incorrectly assigned ASN & GLN sidechain rotamers 81
3.4 Residues of PDB files with unexpected pK_a values calculated by PROPKA 85
3.5 Residues of PDB files with unexpected pK_a values calculated by H++ 89
3.6 Ligands of PDB structures with ionizable functional groups 93

Chapter 1

Introduction

Since the introduction of the Protein Data Bank (PDB) and the first published protein structure of myoglobin, structural biology has come a long way in the determination of macromolecular structures. With the improvement of X-ray techniques and the introduction of NMR structure determination numbers of submitted structures increased constantly and the growth continues to date. While initially limited to experts in structural analysis, nowadays depositors and users of the PDB belong to a large variety of research fields, including biology, chemistry, computer sciences and many more, with different levels of experience (Berman et al. 2000). Most of the people using structures provided by the PDB and other scientific platforms consider the data as correct and error free. However, as the structural information given by X-ray experiments is subject to the interpretation of electron density maps by crystallographers or automated software, they can contain errors or misinterpretations (Brown and Ramaswamy 2007; Davis, St-Gallay, and Kleywegt 2008). Since the accuracy of computational follow up studies like simulations or molecular modelling critically depends on the quality of the input structure, it is important to particularly pay attention to possible contained errors or misinterpretations (Dauter et al. 2014).

The aim of this thesis is to provide an overview of frequent errors in protein structures solved by x-ray experiments. A workflow for the preparation of PDB files, containing protein-ligand complexes, for further use in molecular dynamic (MD) simulations and other computational studies is presented. All steps involved within the provided workflow are carried out by license-free software or web services, to make it suitable for everyone working with macromolecular structures, regardless of access to expensive software packages.

1.1 The Protein Data Bank

The World Wide Protein Databank (wwPDB) is the internationally recognized archive for three dimensional structure data of proteins and nucleic acids. It was established in 1971 at Brookhaven National Laboratories (BNL) and started up with only seven macromolecular structures. Due to improvements of crystallographic techniques and the addition of structures determined by nuclear magnetic resonance (NMR), in the 1980s the number of deposited structures rocketed (Berman et al. 2000). Since then the number of PDB entries, most of them resolved by x-ray crystallography or NMR spectroscopy, steadily increased, currently exceeding 120.000 structures (September 2016).

Exp.Method	Proteins	Nucleic Acids	Protein/NA Complexes	Other	Total
X-RAY	102166	1764	5228	4	109162
NMR	10124	1171	236	8	11539
ELECTRON MICROSCOPY	816	30	279	0	1125
HYBRID	92	3	2	1	98
other	174	4	6	13	197
Total	113372	2972	5751	26	122121

Figure 1.1: Amount of structures currently present in the Protein Data Bank (September 2016). The vast majority of structures are solved by x-ray crystallography and NMR.[1]

Each structure can be identified by its PDB ID, which consists of four alphanumeric characters (e.g “1oj9”). Files provided by the PDB contain xyz coordinates as well as important information about the chemistry of the macromolecule, ligands, structural descriptors and details of the data collection and refinement (Berman 2007). In addition to PDB files the Protein Data Bank provides detailed validation reports and the experimental observations (e.g structure factors) that were used to determine the corresponding coordinates (Rose et al. 2015).

[1] http://www.rcsb.org/pdb/statistics

1.2 X-ray Crystallography

For the purpose of gathering structural information of proteins, x-ray diffraction often is the method of choice as in many cases it is the most advanced tool available for structure analysis. Thus it is no surprise that by far the most of today's published macromolecular structures were determined by x-ray crystallography (Acharya and Lloyd 2005).

Solving a structure via x-ray crystallography involves 4 Steps:

- Crystallization
- Obtaining the diffraction pattern
- Calculation of the electron density map
- Model building

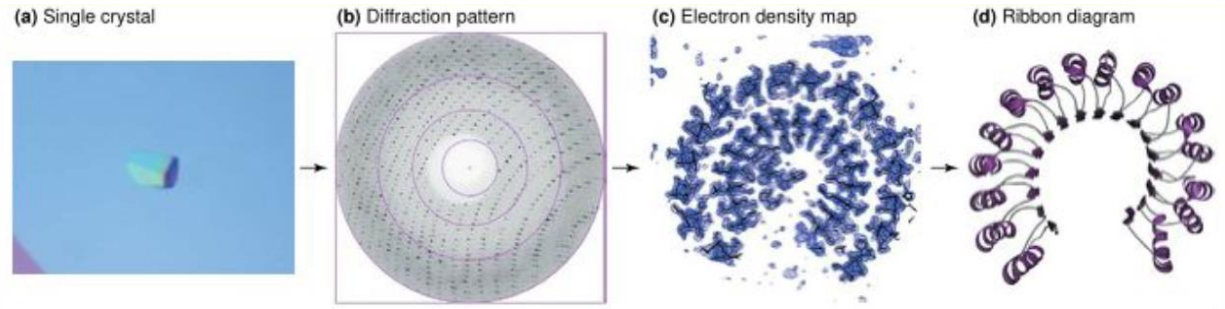

Figure 1.2: Steps involved in solving a protein structure via x ray crystallography. a) a single protein crystal b) the diffraction pattern with data extending to 2.0 Å resolution c) a part of the electron density map with superimposed atomic positions d) the corresponding molecular model of a hRI (human placental ribonuclease inhibitor) molecule, displayed as a ribbon diagramm (Acharya and Lloyd 2005).

1.2.1 Crystallization

Diffraction of x-ray beams from one molecule alone would be too weak to measure. Therefore x-ray experiments require multiple copies of the molecule stacked in a highly ordered three dimensional array (a crystal lattice) to magnify the signal[2].
Protein crystallization is considered the rate limiting and least understood step in protein crystallographic work. It requires a very high purity and relatively large amounts (usually a few milligrams) of the protein. Although the principle of crystallization is the same for proteins and simple salts, the amount of factors influencing the likelihood of proteins forming a crystal, like concentration of the protein and precipitant, pH, buffer, temperature, crystallization technique and the inclusion of additives and many more, make protein crystallization a rather sophisticated procedure. Also protein crystals for x-ray experiments are required to have a size of at least 0.1 mm to provide sufficient volume of crystal lattice that can be exposed to the x-ray beam (Smyth and Martin 2000).

Problems:

- Many flexible, less conformationally constrained or highly glycosylated proteins cannot be resolved by x-ray crystallography due to the inability to form crystals from them.
- Crystals are highly packed structures. Regions, interacting with neighboring molecules, of the crystal structure are fixed in certain conformations, providing only one snapshot of many possible conformations.
- Because of the non-physiological environment in crystals, protein structures may differ from their native conformation in solution.

[2] https://www.jic.ac.uk/staff/david-lawson/xtallog/summary.htm

1.2.2 Obtaining Diffraction Pattern and Electron Density Map

Once crystallization was successful the protein crystal is mounted and rotated while being subjected to x-ray beams. For high resolution experiments x-ray beams are usually generated by synchrotrons which produce highly focused x-rays with a very high intensity. This allows for a higher signal to noise ratio of the diffraction images and shorter exposure time (Helliwell 1992). When hitting the molecules of the crystal, the x-rays are diffracted by the structure´s electrons. As a result of constructive and destructive interferences, characteristic diffraction patterns can be observed on the detector. By rotating the crystal it is possible to obtain diffraction spots for every atom of the investigated molecule (Acharya and Lloyd 2005). Each spot is defined by the parameters wavelength, amplitude and phase. Calculation of the positions of each atom requires knowledge of all three parameters. Wavelengths are selected by the x-ray source, amplitudes can be calculated from the intensities of the diffraction spots but the information about the phase is lost during the experiment. However, it is possible to estimate phases indirectly by different methods. Once solved the “phase problem” and having access to all parameters needed, electron densities for each point of the diffraction map can be obtained by calculating the structure factors (Smyth and Martin 2000).

Problems:

- Resolution of diffraction data varies throughout different experiments and is strongly correlated with the degree of order within the protein crystal. Local flexibility or motion across the molecules of the crystal will result in lower resolution and less detailed and accurate electron density maps.
- Differences of single electrons usually cannot be observed by x-ray diffraction experiments. Hence it is not possible to differentiate between carbon, oxygen and nitrogen atoms from electron density maps.

- Another consequence of the inability to resolve single electrons in most x-ray experiments, is that positions of hydrogen atoms usually cannot be determined experimentally (R. W. W. Hooft, Sander, and Vriend 1996).

1.2.3 Model Building

The final step of x-ray crystallography is to fit the known amino acid sequence into the observed electron density and build a three dimensional structural model of the investigated protein. After the initial process of model building, structures are usually refined iteratively by potential energy minimization operations to improve the fit of the structure to the observed electron density. One has to keep in mind that, as diffraction occurs simultaneously for all molecules in the crystal, the model is a time and averaged picture of the whole crystal lattice. The final model is then output as a PDB file (file format of the protein data bank) which contains the coordinates as well as occupancy and temperature factors for every atom of the structure.

ATOM	1	N	PRO	A	1	28.993	7.932	-2.761	1.00	26.01
ATOM	2	CA	PRO	A	1	28.136	7.033	-3.526	1.00	24.42
ATOM	3	C	PRO	A	1	26.695	7.150	-3.028	1.00	19.43
ATOM	4	O	PRO	A	1	26.440	7.751	-1.994	1.00	17.54
ATOM	5	CB	PRO	A	1	28.703	5.651	-3.205	1.00	28.18

Figure 1.3: Format of a PDB file. The data consists of 11 columns containing following information: The first column indicates the section of the PDB file (in the illustrated case lines belong to the ATOM records of the PDB file); the second column is the atom number; the third informs about the atom type; the fourth shows the residue type in a three letter code; the fifth is the chain identifier in the case of more subunits; the sixth, the residue number; seventh, eighth and ninth are the x,y and z coordinates, the tenth displays the occupancy and the eleventh the B-factor.

Temperature factors

Beside coordinates, for every atom of a macromolecular crystal structure, temperature factors (also called “displacement factors” or “B-factors”) are given. B-factors express the mean atomic displacement in units of $Å^2$ and thus reflect the uncertainty of atomic positions (Radivojac et al. 2004). Because proteins are not rigid objects, their backbones and sidechains are constantly moving due to thermal motion and their kinetic energy. B-factors describe these movements around an atoms averaged position (Yuan, Bailey, and Teasdale 2005).

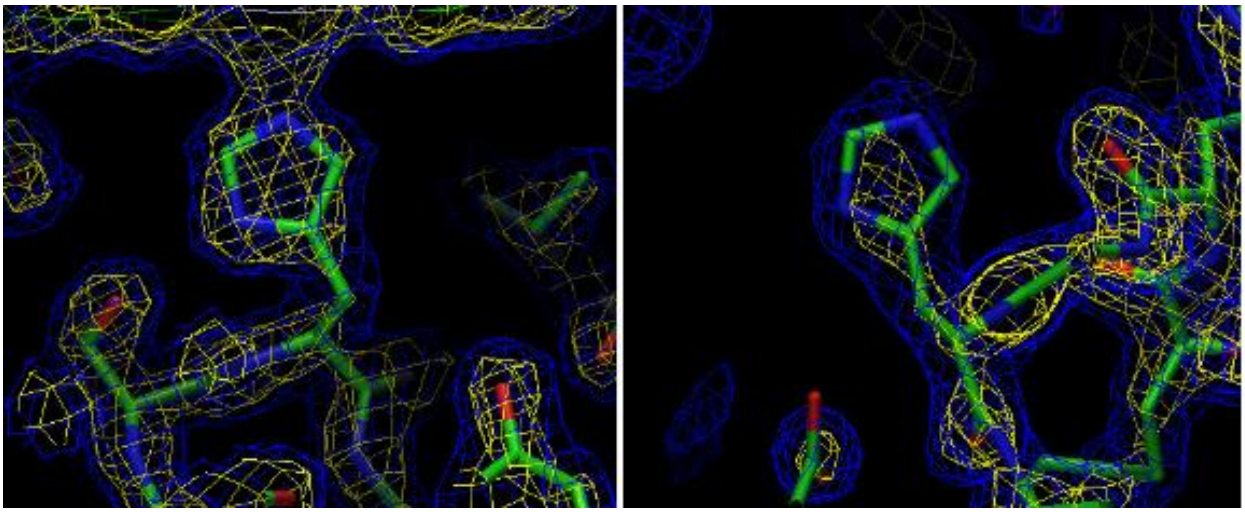

Figure 1.4: Myoglobin structure (PDB entry 1mbi) solved at 2.0 Å resolution. The left picture shows HIS93 with lower B-factors in the range of 15-20, as recognizable from the sharp electron density (yellow grid) surrounding the whole amino acid. On the right picture of HIS81 with high B-factors in the range of 22-74 the electron density is smeared out, resulting in a smaller region of high electron density that only covers parts of the residue[3].

[3] https://pdb101.rcsb.org/

More flexible parts of a protein therefore show a higher mean square displacement, reflecting disorder (Kleywegt 2000). For more information on disorder see section 1.4 “Missing residues and missing atoms”. As a result regions with higher B-factors (>50 Å^2) usually have little or no observable electron densities, which means their coordinates are less accurate. These regions can often be found on the surface of proteins where they are exposed to water and have a higher freedom to move. Low B-factors (<10 Å^2) correspond to well-ordered parts of the protein that show little movement and have nearly the same position in all the molecules of the crystal.[4]

Occupancy

Even though molecules are highly ordered in protein crystals, slight differences between single atoms occur throughout the crystal lattice. A residue sidechain, for example, can show various conformations within molecules of the same crystal. When electron density maps allow to distinguish between multiple conformations, occupancies for the different orientations can be assigned by their relative occurrence. If an atom is identical within all molecules of the crystal, an occupancy of 1.00 (1.00 corresponds to 100% occupancy) will be assigned for this atom. Alternatively, when distinguishing between two definite orientations, an occupancy of 0.6 would mean that 60% of the crystals atoms show the same distinct conformation, while the others adapt a conformation found in the other 40% of the molecules. However occupancies always sum up to a total of 1[5].

[4] http://pdb101.rcsb.org
[5] http://www.proteinstructures.com/Structure

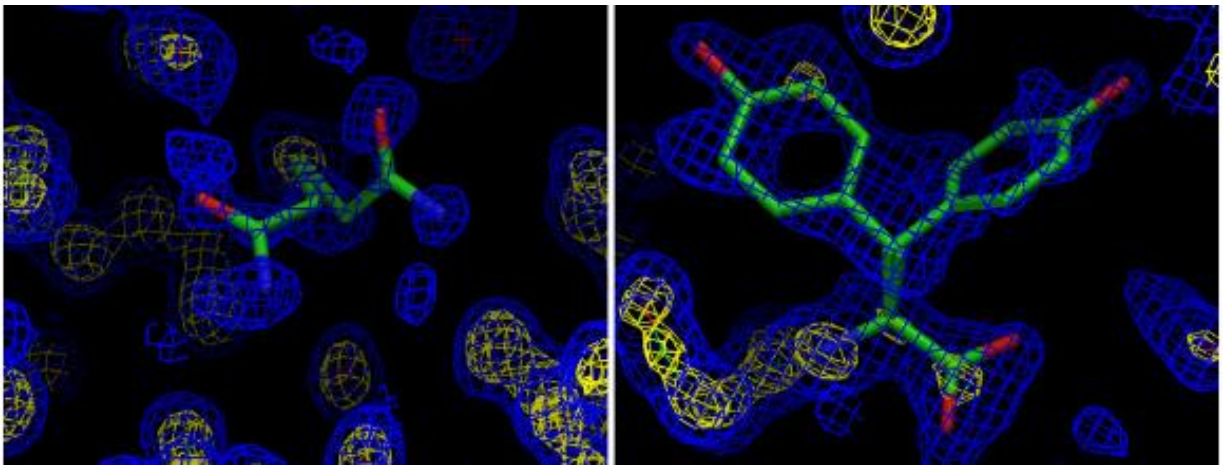

Figure 1.5: Myoglobin structure (PDB entry 1a6m) solved at atomic resolution (1.0 Å). The left picture shows glutamine (GLN) 8 with assigned occupancies of 0.53 and 0.47 for two distinct conformations. The right picture shows tyrosine (TYR) 151 with occupancies of 0.5 for each conformation[6].

1.2.4 Resulting Problems

- Quality of the determined protein structure
- Missing residues and missing atoms
- Side chain flips
- Correct protonation state

[6] https://pdb101.rcsb.org/

1.3 Quality Metrics of the 3D Structure Model

With the constant growing number of deposited structures, careful validation of models used for computational or biological studies becomes more and more important. Overall, protein structures provided by the PDB are of good quality, especially those automatically determined by high-throughput methods (Deller and Rupp 2015). Nevertheless, model building depends on the subjective interpretation of experimental data and as a consequence is prone to errors. Investigators often consider structures, provided by scientific platforms, as error free (Brown and Ramaswamy 2007). The use of models with poor quality can lead to their propagation or even worse to a derivation of rules from incorrectly interpreted data (Cooper 2012). Quality indices can help to avoid conclusions based on potentially erroneous data (Barker and Clevers 2000). Generally quality assessment can be subdivided into global and local quality parameters.

1.3.1 Global Quality Parameters

Global quality parameters like R values and resolution are closely coupled to each other and allow an estimation of the models accuracy. Models, showing poor global quality parameters tend to have a higher amount of residual errors as well (Deller and Rupp 2015).

Resolution of diffraction data

One of the most important global quality parameters in assessing the quality of macromolecular structures is the resolution. The resolution of crystallographic data is usually expressed in Å (Ångström, 10^{-10} m, 0.1 nm), where lower numbers indicate higher resolution. With a higher the resolution more experimental data can be collected, thus leading to more reliable models. At high resolutions, about 95% of the resulting model is a consequence of the collected data and therefore misinterpretations are reduced. Vice versa, the lower the resolution, the higher the chance of incomplete modelling and rate of errors (Davis, Teague, and Kleywegt 2003).

In general resolution can be described as “the minimum spacing (d) of crystal lattice planes that still provide measurable diffraction of X-rays”. Hence, it reflects the “level of detail, or the minimum distance between structural features that can be distinguished in the electron-density maps”.

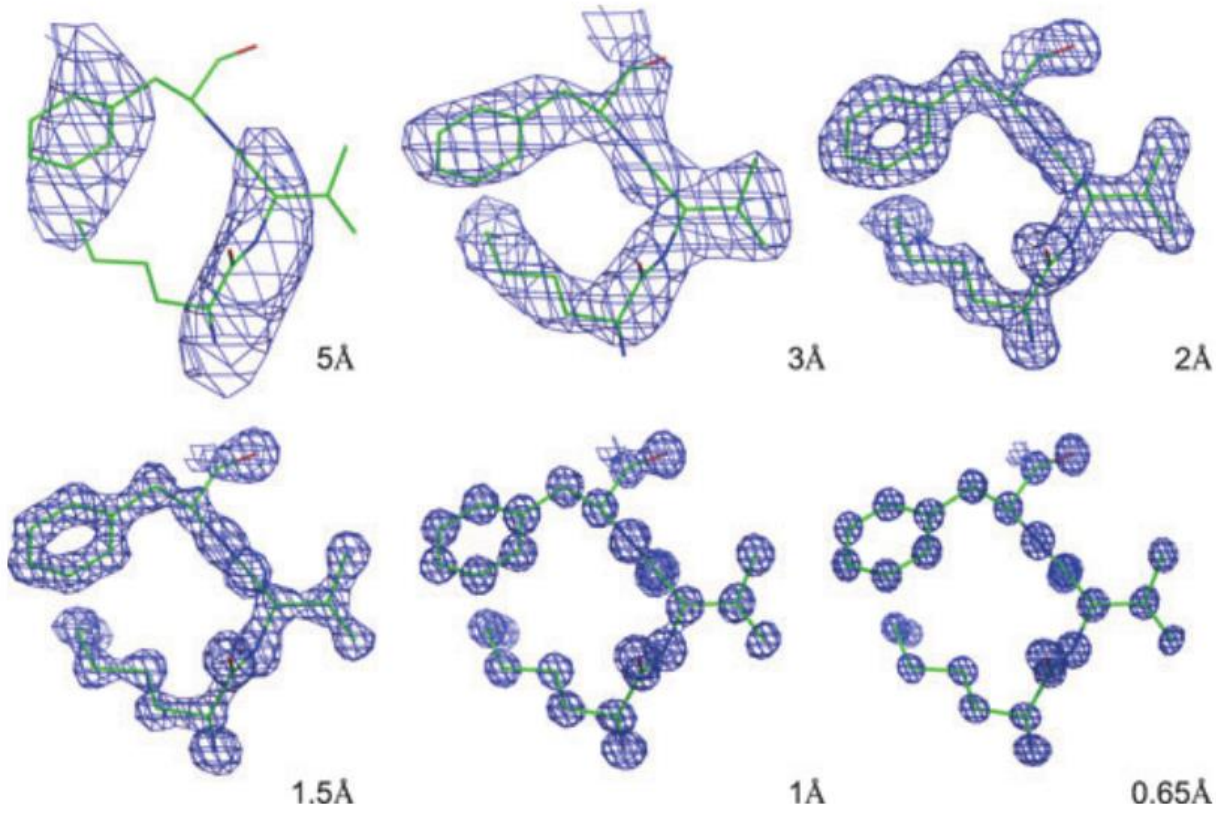

Figure 1.6: Appearance of electron density at different resolutions of experimental data shown on the example of the N-terminal fragment of lysozyme (PDB entry 2vb1). While 184 676 reflections were used for the highest resolution of 0.65 Å, map calculations for the 5 Å resolution only include 415 reflections (Wlodawer et al. 2008).

Typical covalent carbon-carbon bonds in macromolecules have a bond length of 1.5 Å. At 1.2 Å, atomic resolution is reached, which is equal to the shortest covalent bond distance (C=O) observed in Proteins, hydrogens excluded (R. J. Morris and Bricogne 2003). Currently, most of the structures deposited in the PDB show a resolution of 1.5 Å – 2.5 Å, which is considered to be medium to high quality[7].

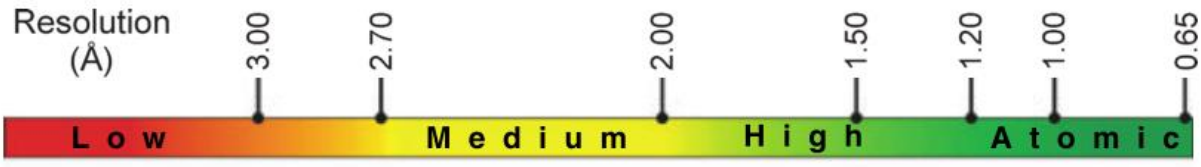

Figure 1.7: Scale of values for resolution achieved by x-ray experiments. Lower values indicate better quality. Structures resolved at 3.00 Å or below are considered to be of poor quality, structures between 2.70 Å and 2.00 Å indicate medium quality and structures solved beneath 2 Å are of high quality. Atomic resolution is reached at 1.2 Å (Wlodawer et al. 2008).

R Values

R-Factor

Another global quality parameter is the “reliability” factor, also called R-factor, which is used to measure the refinement of the model against the diffraction data of the x-ray experiment (Acharya and Lloyd 2005). It is one of the most used global quality indicators, as its value not only provides information about the quality of the model but also the data. The R value is closely coupled to the resolution of the data and the completeness of the model. Therefore more complete models or models with higher resolution data will normally result in a lower R-factors (Brown and Ramaswamy 2007).

[7] http://www.rcsb.org/pdb/statistics

The R-Factor is defined by:

$$\sum ||F_o|-|F_c|| / \sum |F_o|$$

It measures the discrepancy between F_o, representing the experimentally observed structure factor amplitudes, and F_c, the structure factor amplitudes calculated from the model (Kleywegt and Jones 1995). Nevertheless this parameter should be examined critically, since R-values can be made arbitrarily low by adding more parameters to the model than there are experimental observations (so called “overfitting”). Thus a low R-value alone does not necessarily mean that a structure is of good quality.

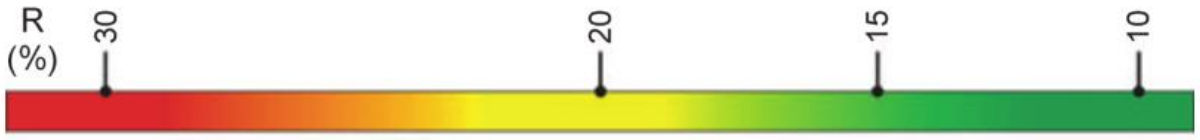

Figure 1.8: Scale of R-factor values. Lower values indicate better agreement between diffraction data and the refined model. Reliable models of macromolecular structures should have R values below 20%. R values approaching 30 % should be treated with extreme caution, as they may contain at least some incorrect parts (Wlodawer et al. 2008).

R_{free}

To overcome the problem of overfitting the use of a second reliability index was suggested by Brünger in 1992. The so called “free R value” or R_{free} is calculated analogous to the R-factor, but uses two randomly split data sets. Consisting of 90 - 95% of the diffraction data, the first set (working set) is used for model refinement. For evaluation purposes the second set (test set), with the rest of the data, is left out from refinement and later used to see how good the model predicts the experimental observations of the test set. Models with R_{free} values >40% should be treated with caution as they might contain serious errors (Axel T. Brünger 1992; Kleywegt and Brünger 1996).

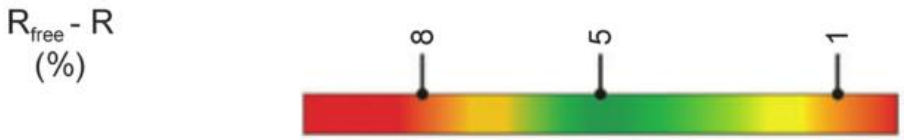

Figure 1.9: Scale of difference between Rfree and R values. Different to R and R_{free}, strong deviations in both directions should be alarming. Optimal values are around 5%. While values exceeding 7% can be a sign for over interpretation of experimental data, extraordinary low values (<2%) suggest that the model is not "truly" free for some reason (Wlodawer et al. 2008).

.

By introducing new parameters to the model the R value may be decreased, but the R_{free} will remain the same or increase. Thus good models should have similar R_{free} and R values. Comparing these two values can give important information about the model quality. The difference of R_{free} – R should be as small as possible (usually R_{free} values are a bit higher). Differences over 7% can indicate overfitting or model errors and should be alarming. Very low differences however could signal that e.g. the test set is not "truly" free (Wlodawer et al. 2008).

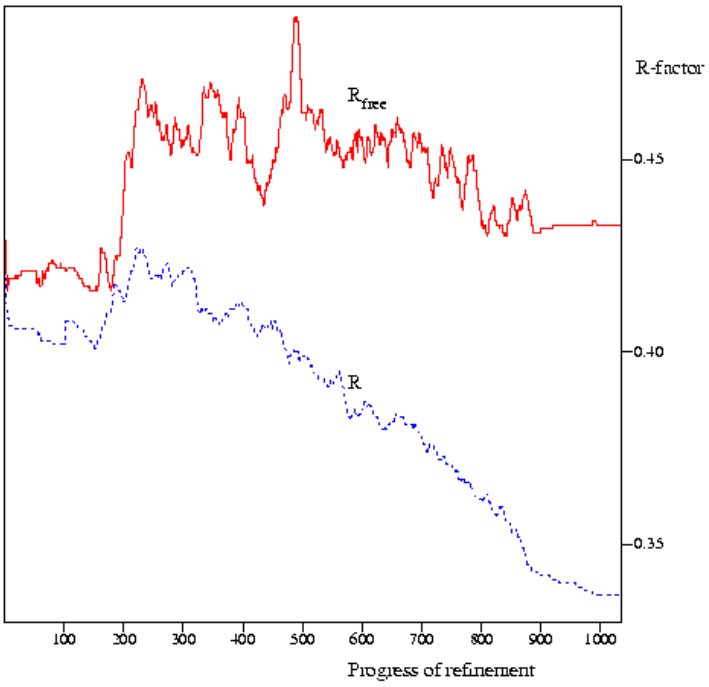

Figure 1.10: Behavior of R and Rfree during the refinement of human immunoglobulin (IgG) and the C2 domain of protein G at resolutions ~ 3.5 Å. The solid line shows the behavior of R_{free}*, changes of R values are shown by the dotted line. While R values drop by 0.1, the* R_{free} *even increases (Kleywegt and Jones 1997).*

1.3.2 Local Quality Parameters

In some fields of research not only global quality, but rather the correctness of a specific part of a model is of particular interest. In structure based drug design (SBDD) for instance, accurately modeled binding sites and their bound ligands are much more important than the overall quality of the model (Cereto-Massagué et al. 2013).
Local quality metrics (also called real space quality metrics), always relate to the model and its electron density map rather than directly to the diffraction data. Although, in contrast to global quality indicators, they focus on providing information about certain parts of the protein (e.g single residues or ligands), local quality parameters can also be averaged and used to measure regional or global quality of a model (Deller and Rupp 2015).

Important parameters for assessing local quality:

- Real space R-Factor
- Real space correlation coefficient
- Occupancy weight B-Factor
- Ramachandran outliers

The Real space R-Factor

Originally introduced by Jones et al. (1991), the Real space R-Factor (RSR) is defined as

$$RSR = \sum |\rho_{obs} - \rho_{calc}| / \sum |\rho_{obs} + \rho_{calc}|$$

where ρ_{obs} is the electron density obtained from the experiment and ρ_{calc} the models calculated electron density. Before the actual calculation a certain radius around the entities atoms (entity refers to the part the RSR is calculated for i.e. a ligand or residue) is chosen. The observed and calculated electron densities are then compared for every

point in this area (Read et al. 2011). The size of this radius depends on the resolution of the data used. RSR calculations offered by the EDS (Electron density server) or within PDB Validation reports are analyzed residue by residue, resulting in per residue plots. RSR values range from 0 to 1, low values indicating a good fit of the examined part of the model and the corresponding experimental data.

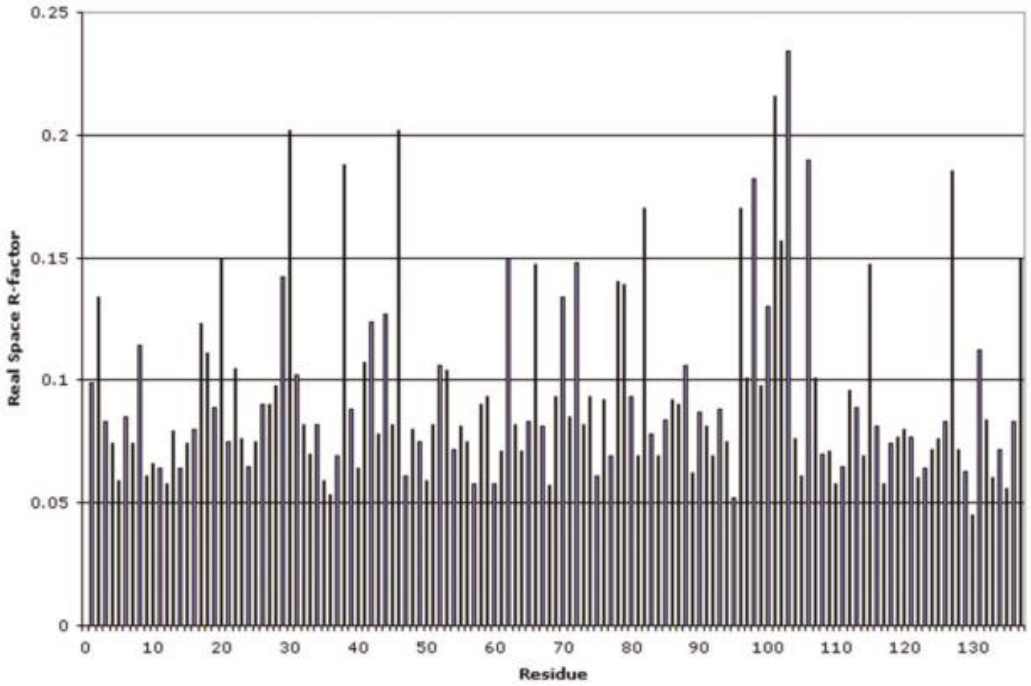

Figure 1.11: Per-residue plot of RSR values for retinoic acid binding proteins I&II (PDB entry 1CBS) as provided by the EDS. Every amino acids RSR value is displayed by a single bar according to its residue number (Kleywegt et al. 2004).

For standard residues RSRZ scores (normalized RSR scores) can be calculated that compare the calculated RSR value to the average RSR value of all residues of the same type (e.g. Lysine) at similar resolution (Kleywegt et al. 2004). This can be particularly interesting for assessing the local fit of certain residues, as typical RSR values for a certain residue are not only resolution dependent but also vary within different amino acids (e.g. glutamates tend to have higher average RSR values as they often appear on proteins surfaces). Residues with RSRZ scores >2 are considered to be outliers.

The percentage of all residues with RSRZ >2 within the model can also give information about the global quality (Read et al. 2011). Unfortunately as identical ligands only rarely appear in the PDB it is not possible to calculate their RSRZ with statistical significance (Gore, Velankar, and Kleywegt 2012).

The Real space correlation coefficient

Another very important and often used parameter for assessing local quality is the Real space correlation coefficient (RSCC). Unlike the RSR, for calculating the RSCC there is no need to scale observed and calculated electron densities together, which is considered one of the weaknesses of the RSR (Kleywegt et al. 2004). Defined as

$$RSCC = \sum (\rho_{obs} - \rho_{obs}) * (\rho_{calc} - \rho_{calc}) / [\sum (\rho_{obs} - \rho_{obs})^2 * \sum (\rho_{calc} - \rho_{calc})^2]^{½}$$

the RSCC is a standard linear sample correlation coefficient, where ρ_{obs} again stands for the observed and ρ_{calc} for the calculated electron density. Values can range from 0, meaning that the looked at element is actually not in the suggested position, to 1, demonstrating a perfect fit to the experimental data. Generally RSCC values >0.9 show that e.g. the ligand fits the electron density very well, while values <0.8 indicate poor- or over-modelled structures where significant fragments of the model are not covered by the electron density (Deller and Rupp 2015). A downside of the RSCC parameter is that it is not sensitive to varying intensities of electron densities, meaning that even a very weak, but spherical densities would correlate very well with the model, if for instance a water molecule was placed there (Read et al. 2011).

The Occupancy weighted average B-Factor

A local quality parameter also to consider is the Occupancy weighted average B-Factor (OWAB). It can be used to check quality of a models residue or ligand and is defined by (Kleywegt et al. 2004):

$$OWAB = (\sum B * Q) / (\sum Q)$$

It is calculated by summing up the B-Factors (B) of all atoms in the residue or ligand multiplied by their occupancy (Q) and then dividing by the sum of occupancies of the atoms [Weichenberger et al., 2013]. OWAB values can range from 0 to infinity, lower values indicating better model quality for the group of atoms it is calculated for.
As for RSR and RSCC, when the calculated value is poor, it is not possible to distinguish between a bad model (accuracy), pool diffraction data (precision) and a high amount of motion by this metric alone. However, by combining them it is possible to retrieve more information. If, for example, the OWAB is low but the RSR is high the model is of bad quality (Warren et al. 2012).

Ramachandran outliers

The backbone conformation of every, except the terminal, residue in a Protein is determined by three torsion angles (figure 1.12). Rotations around the N-Cα bond of a residues backbone (C_{i-1}-**N_i-C^{α}_i**-C_i) are defined by φ, rotations around Cα-C bonds of backbone atoms (N_i-**C^{α}_i-C_i**-N_{i+1}) by ψ and rotations around C and N peptide bonds (C^{α}_i-**C_i-N_{i+1}**-C^{α}_{i+1}) by ω (Zhou, O'Hern, and Regan 2011). Due to the partial double bond character of peptide bonds, ω angles are strictly limited to values near 180° for trans peptides and 0° for cis peptides. Although ω angles only have little importance regarding validation purposes, values between +/- 20 and +/-160 are unusual and should be treated carefully.

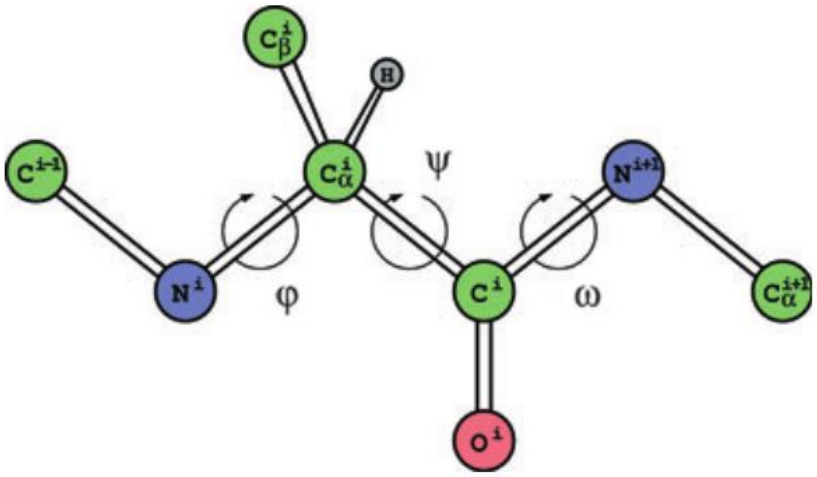

Figure 1.12: Schematic representation of torsion angles of the protein backbone. Torsion angles are defined by φ, ψ and ω for the i^{th} residue. In the eclipsed conformation their reference value is 0°. Torsion angles shown in this figure are all equal 180° (Wlodawer et al. 2008).

In contrast φ (phi) and (psi) ψ torsion angles are much less restricted. Because of steric hindrance, resulting from the amino acids sidechains, there are preferred combinations of φ and ψ values. This holds true even for the majority of residues forming elements without typical secondary structure (Kleywegt 2000). Deviations from these preferred torsion angle conformations can indicate errors in the model and should be subject to further inspection. However outliers are not necessarily wrong and evaluation with the experimental information given is recommended as they may also show interesting features of protein function (Kleywegt and Jones 1996; R. W. Hooft, Sander, and Vriend 1997). An extremely useful tool for assessing backbone torsion angles is the so called "Ramachandran" plot, which maps the φ and ψ torsion angles pairs of every residue against their predicted distribution (Wlodawer et al. 2008). In this two dimensional plot the horizontal axis shows φ values, while the vertical axis shows ψ values. Values on both axes scale from -180° to 180°[8]. Ramachandran plots illustrate "core" regions, which contain the most favorable combinations of torsion angles, "allowed" regions, either around the core regions or not associated with them, and "disallowed" regions (A.

[8] http://proteopedia.org/

L. Morris et al. 1992). The two large of the three core regions represent preferred conformations found in alpha helices and beta strands, while the third smaller core region corresponds to backbone conformations in left handed alpha helices [Hollingsworth et al., 2011]. Glycine (GLY) and proline (PRO) residues however, are exceptions to the regularity of typical torsion angle conformations. GLY residues lack sidechains, are therefore less sterically hindered, and can be found in regions not allowed for alanine (ALA) like residues (i.e. residues that are neither GLY nor PRO). Proline on the other hand, due to its sidechain covalently bonding to the backbones nitrogen, can adapt only very limited torsion angle conformations. As a consequence GLY and PRO outliers are typically not included in overall summary validation measures (Lovell et al. 2003). Φ and ψ angles are usually not restraint in refinement processes (A. T. Brünger et al. 1998). Hence the percentage of Ramachandran outliers is an excellent parameter for assessing local quality and for validating a model (Read et al. 2011).

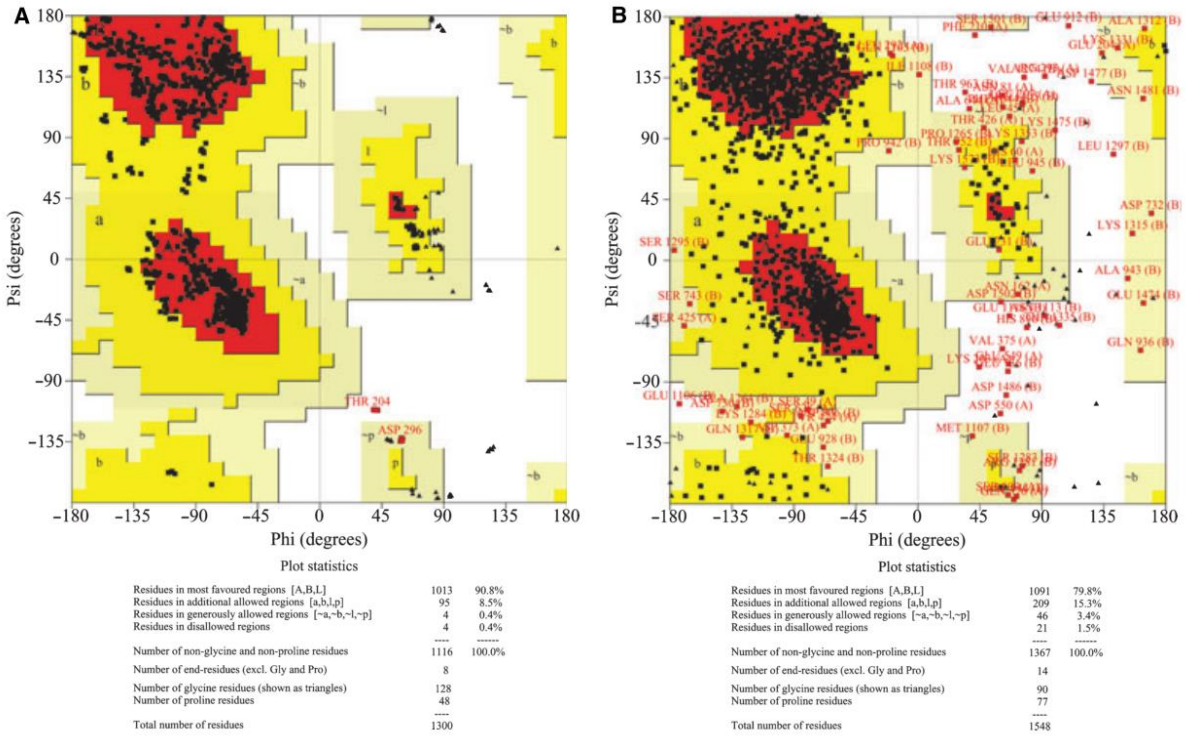

Figure 1.13: Examples of a Ramachandran Plot. (A) Left shows the Plot for Erwinia chrysanthemi L-asparaginase (PDB entry 1o7j) at atomic resolution. (B) The right Plot shows the C3b complement protein (PDB centry 2hr0) at 2.26 Å, characterized by a large number of dihedral angles in disallowed regions (Wlodawer et al. 2008).

1.4 Missing Residues and Missing Atoms

Each PDB file contains two sections giving important information about the structure of the investigated Protein. In the “SEQRES” records all amino acids (residues) of the crystal are listed, irrespective of whether or not the corresponding coordinates are present in the file. The “ATOM” records provide the positional information of each atom, including atom type, matching residue, chain ID, its x, y and z coordinates, the occupancy- and the B-Factor.[9] However not every amino acid displayed in the SEQRES records automatically comes with its coordinates provided in the ATOM records. A large number of protein structures are partially incomplete. They contain missing residues or missing atoms, meaning that these parts of the protein have no experimentally determined position in space. Roughly 64% of all the proteins deposited in the PDB contain at least 1 missing residue while 26% miss at least ten residues. In most cases these missing regions are located at the beginning or the end of the structure (Brandt, Heringa, and Leunissen 2008).

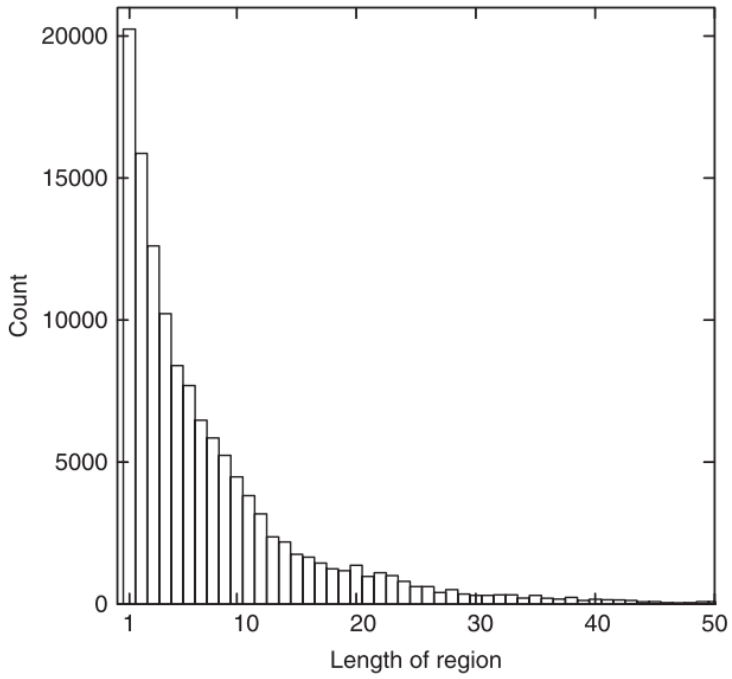

Figure 1.14: Distribution of all missing residues in PDB protein structures. The count of missing residues is larger than the actual number of chains as a single chain can contain more than one missing region. Missing regions are predominantly found at the beginning or end of a protein sequence (Brandt, Heringa, and Leunissen 2008).

[9] http://pdb101.rcsb.org/

In PDB files header, missing coordinates are displayed in the Remark section as either REMARK 465 (for missing residues) or REMARK 470 (for missing atoms).

```
REMARK 465 MISSING RESIDUES
REMARK 465 THE FOLLOWING RESIDUES WERE NOT LOCATED IN THE
REMARK 465 EXPERIMENT. (M=MODEL NUMBER; RES=RESIDUE NAME; C=CHAIN
REMARK 465 IDENTIFIER; SSSEQ=SEQUENCE NUMBER; I=INSERTION CODE.)
REMARK 465
REMARK 465   M RES C SSSEQI
REMARK 465     GLY A   248
REMARK 465     HIS A   249
```

```
REMARK 470 MISSING ATOM
REMARK 470 THE FOLLOWING RESIDUES HAVE MISSING ATOMS (M=MODEL NUMBER;
REMARK 470 RES=RESIDUE NAME; C=CHAIN IDENTIFIER; SSEQ=SEQUENCE NUMBER;
REMARK 470 I=INSERTION CODE):
REMARK 470   M RES CSSEQI  ATOMS
REMARK 470     CYS A 277    SG
REMARK 470     PHE A 278    CG   CD1  CD2  CE1  CE2  CZ
REMARK 470     PHE A 424    CG   CD1  CD2  CE1  CE2  CZ
```

Figure 1.15: Remarks for missing coordinates in PDB files. The illustration shows REMARK 465 as displayed in PDB files for residues absent from the coordinate section but present in the SEQRES records. Missing non-hydrogen atoms of standard residues are displayed under REMARK 470. Missing residues or atoms are listed with their residue type, chain identifier and the residue number.

Missing residues and atoms in PDB files derive from missing electron density for this regions, making them invisible for the interpreter. Except from technical limitations, the most common cause for this circumstance is the movement of flexible parts in the protein (Felli and Pierattelli 2015). The phenomenon of these movements is called structural “disorder” and can be divided into two classes, both effects contributing to very high B-values in these regions.

Static disorder

Static disorder occurs, when a fragment of a protein (e.g. a residue side chain), shows distinct variations of its conformations across different unit cells. At low resolutions, due to the multiple conformations, the resulting electron density can be smeared out and therefore be hidden in the noise. Unless the experiment is repeated at higher resolutions these fragments remain uninterpretable. In the case of static disorder, at higher resolutions, distinguishable electron densities may be observed for its alternating conformations. In this case occupancies for the fragment can be assigned.

Dynamic disorder

In contrast to static disorder, dynamic disorder results from thermal vibrations or increased mobility of a fragment within every unit cell. Due to the constant fluctuation of the disordered region between a large number of states (with a time scale shorter than the duration of the experiment) and because the distribution of electrons is averaged throughout all unit cells, the electron density appears smeared out or even worse is completely lacking (Wlodawer et al. 2008).

Sometimes the degree of disorder is so high that it is not possible to assign any atomic positions. Although this can also be observed with structured elements, the majority of missing electron densities is represented by disordered regions that lack fixed secondary structure (also called “random coils”) (Radivojac et al. 2004). Thus regions of the model with missing residues are often loop elements or flexible termini, especially when they are located at the protein surface and therefore are exposed to solvent. While the length of these regions can span from one to hundreds of residues, most of them are <30 residues long (Dosztányi, Mészáros, and Simon 2009).

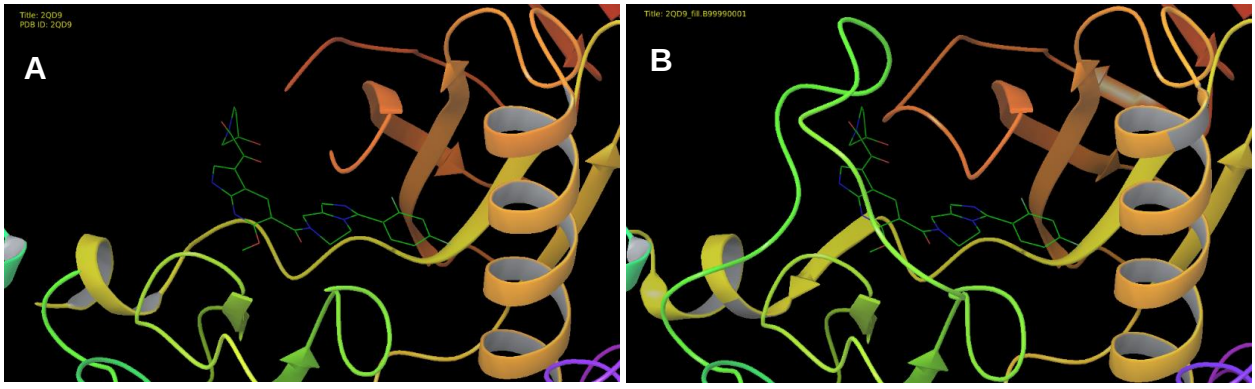

Figure 1.16: Structure of α map kinase inhibitor (PDB entry 2QD9) before and after modelling the sections of missing coordinates. (A) In the original PDB structure residues 1-10; 37-38; 124-126; 179-191; 359-366 are missing from the coordinate section. (B) PDB structure after modelling the missing loops with MODELLER (Sali and Blundell 1993b).

Since loops serve as linker between two elements of secondary structure and often contribute to the biological function being part of active- or binding sites that determine the proteins specificity, their absence from the model can be a significant problem for research purposes (Fiser and Sali 2003). For this reason and because many software packages decline incomplete coordinate files, missing residues and atoms should be added to the model to prevent biased simulations (Djinovic-Carugo and Carugo 2015).

There are two different approaches for rebuilding missing residues:

- Homology modelling
- Ab initio prediction

Homology modelling

Homology modelling, also referred as comparative modelling, is a databased approach for adding missing residues to a model that involves three steps. At first a database search for known structures, with sequence identity to the stem regions of the missing loop, is performed. This procedure is usually carried out by standard tools such as psi-blast. In many cases more than one segment, fitting the stem regions, can be found. Having found matching structures, the sequences of template(s) and target are aligned and, in the case of multiple results, sorted by sequence similarity or geometric criteria. As sequence identity decreases, alignments become increasingly ambiguous. Therefore homology modeling heavily relies on accurate structure alignment. During the second step, fragment(s) of the template(s) are then inserted in between the stem regions of the incomplete target structure to fill the missing gaps. In the simplest scenario of having only one template the coordinates are merely copied to the target structure according to the alignment (Jamroz and Kolinski 2010). However some software packages like MODELLER also allow using multiple templates to build a consensus scaffold (Sali and Blundell 1993). In the third and last step, the entire completed structure is refined by energy minimization and repacking sidechains.

Ab initio prediction

In contrast to homology modelling, where parts of known protein structures with sequence similarity are used to complete incomplete structure models, ab initio loop prediction is used to model missing parts of protein models from scratch without any use of templates. Although approaches and algorithms vary throughout different software, ab initio loop prediction is generally based on the same procedure. First, a set of different loop conformers is generated. The built loop conformers are then inserted into the missing region of the protein structure, followed by refinement of the completed models using energy functions and selection of the most appropriate model by a scoring function. Compared to the template based approaches, ab initio loop prediction is much more difficult as the number of possible conformations increases exponentially with increasing loop length. Additionally the choice of the loop conformer to pick is often a non-trivial procedure, as the lowest energy conformer does not necessarily represents the structure with the lowest RMSD value (Galaktionov, Nikiforovich, and Marshall 2001). On the other hand template based approaches are limited by the relatively small number of known protein structures and the fact that homologous proteins often lack structural information in loop regions (Park et al. 2014).

1.5 Flipped Sidechains

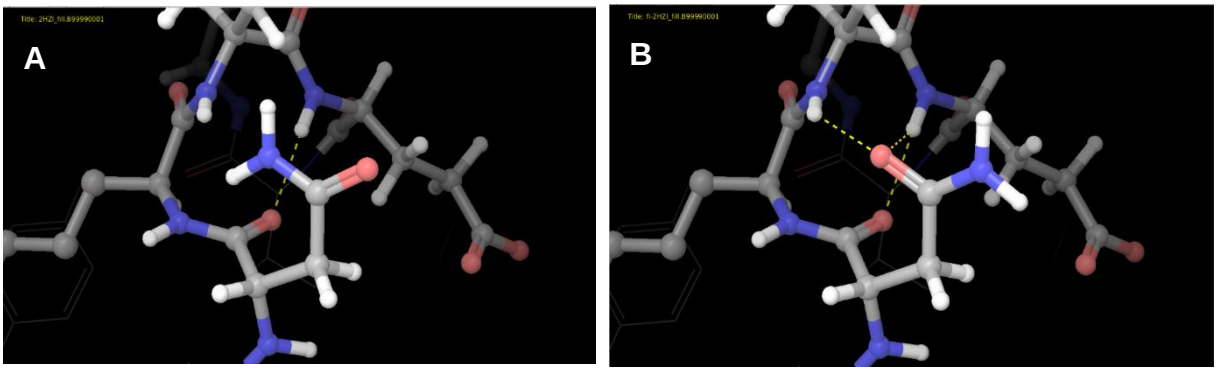

Figure 1.17: Illustration of a "flipped" asparagine sidechain in the structure of abl kinase domain (PDB entry 1HZI). (A) Original orientation of the sidechain without any interaction of its functional group to its environment. (B) Sidechain after a 180° rotation. The amide is now capable of building two hydrogen bonds with neighboring atoms.

Asparagine and glutamine

Incorrect assignment of asparagine (ASN) and glutamine (GLN) sidechain orientation is a common mistake made in x-ray analysis. Analyses suggest that across all the protein crystal structures provided by the PDB the error-rate of incorrect assigned ASN and GLN rotamers is around 20%. This also applies for high quality structures with a resolution of 1.5 Å or better. Compared to random assignments, in which the rate is expected to be 50%, this is a fairly high number (Weichenberger, Byzia, and Sippl 2008). Both ASN and GLN sidechains contain a terminal amide. This functional group can concurrently act as hydrogen bond donor and acceptor and has the ability to form up to 4 hydrogen bonds (two accepted by the oxygen and two donated by the nitrogen) (Weichenberger and Sippl 2007). Thus it often plays an important role in building up hydrogen bond networks, protein-protein interactions, ligand docking, substrate binding or catalysis. Except with extremely high resolutions, it is not possible to distinguish between O, C and N atoms based on their electron density. Therefore, by interpreting electron density maps obtained by x-ray experiments, the exact position of the amides nitrogen and oxygen atoms can be determined precisely, while their identities remain

unknown. Wrong assignment of the rotamers can lead to unfavorable interactions with surrounding atoms (Weichenberger and Sippl 2006). By analyzing their environment sidechain orientations can be validated and corrected if evidently flipped. Very accurate assignments are possible if obligate donors (e.g. peptide NH groups) or acceptors (e.g. carboxyl groups) are found in close distance to the rotamer. Conversely, in cases where surrounding atoms are ambiguous donors or acceptors the entire local hydrogen bond network needs to be analyzed for assignment (Word et al. 1999).

Histidine

Similar to asparagine and glutamine sidechains, it is not possible to determine the correct orientation of histidine (HIS) sidechains directly by x-ray experiments. Just like sidechains of ASP and GLU residues, both HIS rotamers fit the electron density equally well. Compared to the simpler chemistry of ASP and GLU sidechains, HIS sidechains are much more complex. The imidazole ring of HIS has a pK_a (~6.5) close to physiological pH values (~7.4) and two nitrogen groups which are capable of accepting a proton. Hence, depending on the pH and its environment, HIS residues can adapt a neutral form, with either δ-nitrogen (HID) or ε-nitrogen (HIE) protonated, or a doubly protonated single cationic form (HIP). When flipped, sidechains can adopt additional three states (flipped HID, HIE or HIP), altogether summing up to six different states (Kim et al. 2013). Because all protonation and tautomerisation states are stable, HIS residues commonly participate in catalysis and can often be found in active sites (McDonald and Thornton 1995). In the case of HIS residues, a flipped sidechain changes position of C and N atoms within the imidazole ring. Compared to ASN and GLN, for which the procedure of determining the right conformation is often straight forward, the different protonation states HIS can adapt, make the assignment of correct sidechain orientations much more subtle (Rupp 2009).

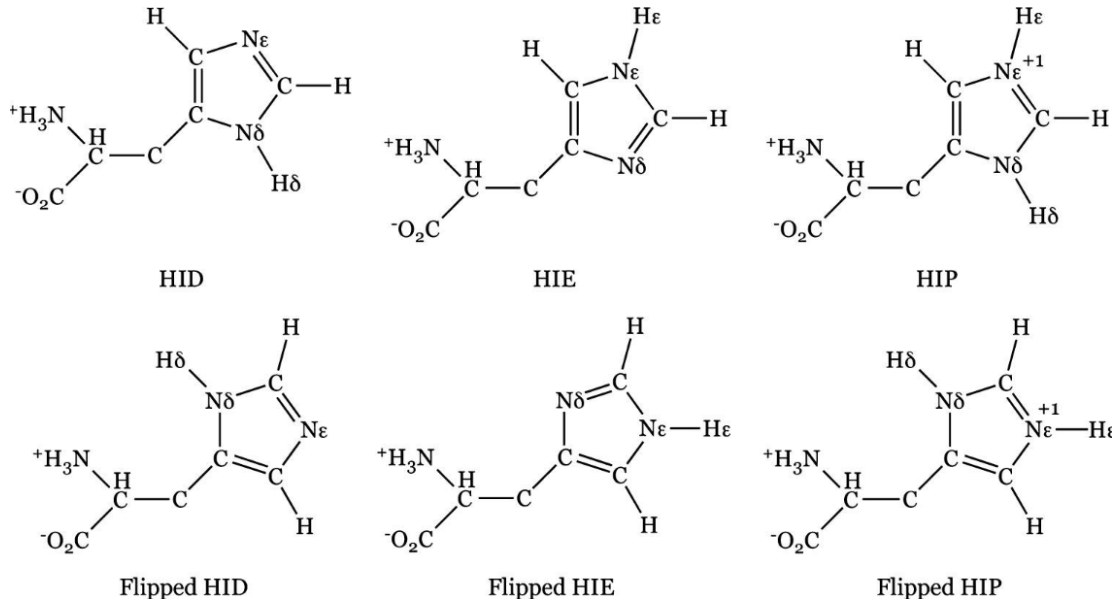

Figure 1.18: Six possible rotameric and protonation states of histidine with marked formal charges on nitrogen in doubly protonated HIP states (Kim et al. 2013).

1.6 Protonation State

PDB coordinate files usually do not contain hydrogen atoms. The reason why they are missing, although representing ~ 50%[10] of all atoms within a protein and being responsible for ~ 75% of all atomic contacts, simply is that most models deposited in the PDB were determined by x-ray crystallography which, except for experiments with resolutions <1 Å, is unable to resolve hydrogen atoms. Fortunately the position of most missing hydrogen atoms can be estimated with high accuracy, due to knowing the conformations of the atoms on which they reside (Rhodes 2006). Whereas this holds true for most of the hydrogen atoms, in some cases, such as in functional groups that can be (de-) protonated (e.g. amino or carboxyl groups), the situation is more complicated. The protonation state of residues, containing titratable groups, depend on complex electrostatic interactions between themselves and their surrounding environment, which makes the accurate prediction of their protonation state a non-trivial task (Anandakrishnan, Aguilar, and Onufriev 2012).

Among the 20 proteinogenic amino acids, seven carry a functional group in their sidechain, ionizable between pH values of 1 and 14. On average they make up 29% of the amino acids in proteins. Acidic amino acids, namely aspartic acid (ASP), glutamic acid (GLU), cysteine (CYS) and tyrosine (TYR), are negatively charged above their pK_a and uncharged below their pK_a values. Vice versa, basic amino acids, namely arginine (ARG), lysine (LYS) and histidine (HIS), are positively charged below their pK_a and uncharged above their pK_a (Pace, Grimsley, and Scholtz 2009). Every ionizable amino acid has an experimental determined intrinsic pK_a value ($pK_{a\,int}$). These values refer to the pK_a of the corresponding amino acid in solution without any interactions except the influence of the neighboring peptide bonds. They were determined by building pentapeptides, consisting only of alanine (ALA) and the desired amino acid (ALA-ALA-X-ALA-ALA), and should reflect the unperturbed pK_a values of amino acid sidechains (Thurlkill et al. 2006).

Essentially three effects are responsible for a shift apart from intrinsic pK_a values:

[10] http://proteopedia.org/

Group	Content[a]	Buried[b]	p*K* value in alanine pentapeptides (pK_{int})[c]	Average p*K* value	Low p*K* value	High p*K* value
	%	%				
Asp	5.2	56	3.9	3.5 ± 1.2	0.5	9.2
Glu	6.5	48	4.3	4.2 ± 0.9	2.1	8.8
His	2.2	72	6.5	6.6 ± 1.0	2.4	9.2
Cys	1.2	90	8.6	6.8 ± 2.7	2.5	11.1
Tyr	3.2	67	9.8	10.3 ± 1.2	6.1	12.1
Lys	5.9	34	10.4	10.5 ± 1.1	5.7	12.1
Arg	5.1	56	12.3[d]			
C terminus			3.7	3.3 ± 0.8	2.4	5.9
N terminus			8.0	7.7 ± 0.5	6.8	9.1

Figure 1.19: Characteristics of proteins with ionizable sidechains. (a) Average current content (%) of amino acids from all lifeforms. (b) Percentage of buried ionizable groups of amino acid sidechains. (c) Intrinsic pK_a values of amino acids in alanine pentapeptides (Pace, Grimsley, and Scholtz 2009).

1) Dehydration (Born effect)

Residues with ionizable sidechains within the hydrophobic interior of proteins or at interfaces between molecules and the protein (“buried” residues) often show significantly shifted pK_a values from their $pK_{a\,int}$. This is a consequence of the absence of water (dehydration) and interactions with polar elements in these regions. Because it is energetically unfavorable for ionized groups to be surrounded by environments with a low dielectric constant, pK_a values are shifted towards their neutral form. pK_a values of buried acidic residues therefore rise, while pK_a values of buried basic residues drop compared to pK_a values of the same residues exposed to water (Karp et al. 2007).

2) Charge – Charge interactions

When encountering charges in their environment, ionizable sidechains also show a shift in their pK_a values. These interactions highly correlate with the distance and whether the charge is positive or negative. A Positive charge near the ionizable group for instance will generally lower pK_a values, favoring the neutral state of basic residues or the deprotonated state of acidic residues. Negative charges on the other hand will generally raise pK_a values of ionizable groups nearby. Charge – charge interactions are responsible for the majority of pK_a shifts on the protein surface, which is where most of the ionizable residues are located (Pace, Grimsley, and Scholtz 2009).

3) Charge – Dipole interactions

Partial charges or hydrogen bonds are also able to interact with ionizable residues. Changes in pK_a values depend on whether hydrogen bonding is tighter to the protonated or deprotonated form of a residue. In the case of a more favorable protonated form pK_a values will increase, while in case of a more favorable deprotonated form pK_a values will decrease. It is also possible for ionizable groups to form more hydrogen bonds at once, in which case every hydrogen bond contributes to the pK_a shift (Grimsley, Scholtz, and Pace 2009).

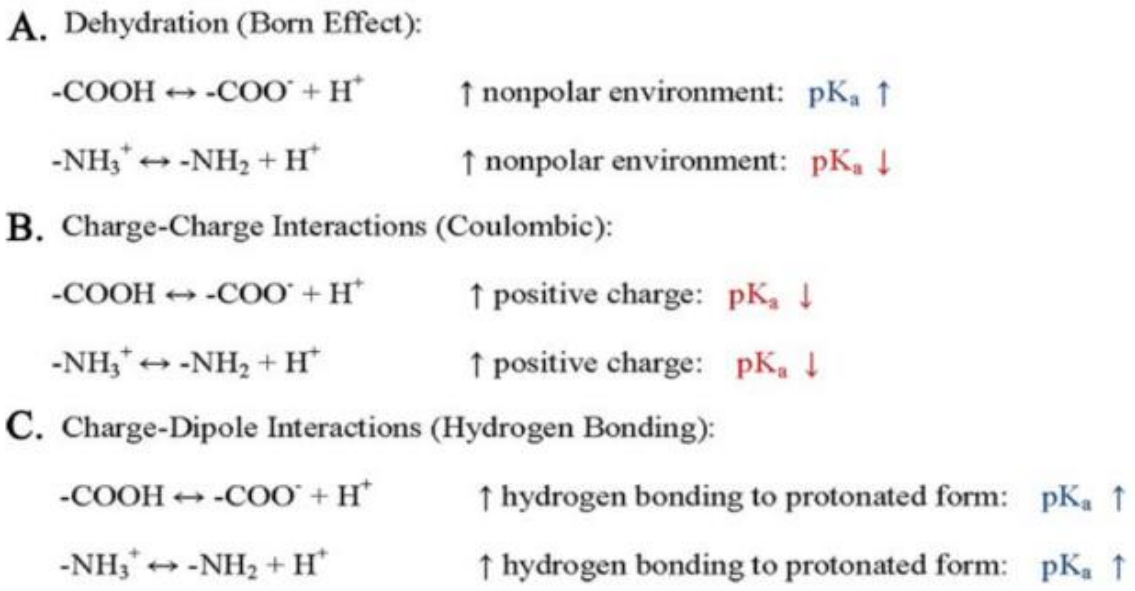

Figure 1.20: Factors influencing the pK_a of ionizable groups in proteins. (A) Changes of pK_a values due to dehydration (Born effect) result from ionizable groups buried in the interior of proteins favoring neutral states because of the low dielectric constant. (B) pK_a values of ionizable groups rise when near to a negatively charged environment and decline in the near environment of positive charges. (C) Depending on whether hydrogen bonds are tighter to the protonated or deprotonated form, pK_a values increase (favored protonated form) or decrease (favored deprotonated form) (Pace, Grimsley, and Scholtz 2009).

Knowledge of the protonation state of ionizable residues is crucial as they contribute to important properties of proteins like protein stability, solubility, catalytic activity and binding affinity. Hence, for researchers working with protein models, it is of big importance to check for ionizable groups and their protonation state, considering their local environment (Rostkowski et al. 2011).

Chapter 2

Methods

2.1 Validation of Selected Input Structures

Prior to any modification applied to PDB files a set of structure validation tests, implemented in the WHAT IF web service (Hekkelman et al. 2010), were run. A set of numerous validation parameters were tested using the “Protein Model check” option that returns a WHAT IF check report after uploading files. Beside administrative checks, coordinate parameters (B-factors, absence of sidechain atoms, occupancy), nomenclature problems, geometric parameters (e.g. bond length, bond angles), torsion angles, atomic clashes, packing environment and hydrogen bond related parameters, are evaluated within the model check. Reports to corresponding PDB structures were inspected for devastating errors and saved.

2.2 Evaluation of the Ligands Fit to the Electron Density Map

For the inspection of ligands and corresponding binding sites, PDB structures were inspected by the software tool VHELIBS (Cereto-Massagué et al. 2013). All Residues within a radius of 4.5 Å distance to the ligand were considered to be part of the binding site. Quality parameters taken into account during inspection included RSR values (Threshold for “good” RSR: 0.24; Upper cap: 0.4), RSCC (Threshold for lowest “good” RSCC: 0.9), R – Rfree (Threshold: 0.05 ≙ 5%), minimum average Occupancy (1.0), OWAB (Threshold: 50), Resolution (Limit: 3.5 Å), and Rfree values (Threshold: 1%). Assessment of mentioned quality indices was followed by a visual inspection of the fit of residues and ligands to their electron density map.

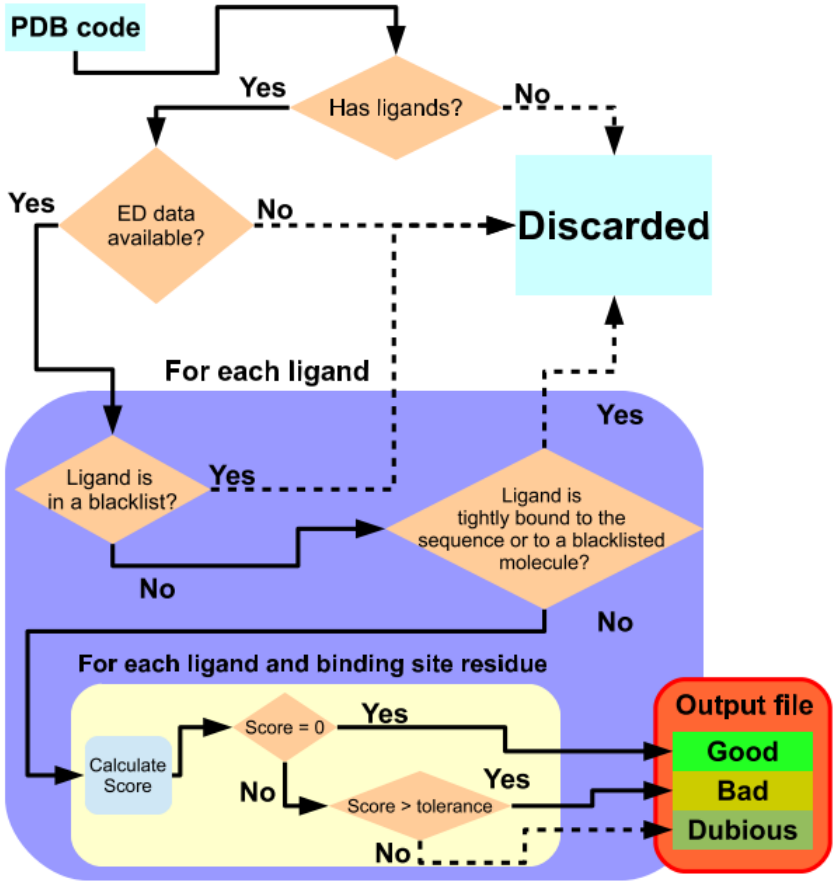

Figure 2.1: Process of classification as implemented in VHELIBS for evaluating a models fit to its electron density (Cereto-Massagué et al. 2013).

2.3 Modelling Missing Coordinates

The MODELLER package (Sali and Blundell 1993) was used to add missing coordinates and write the model out into a new PDB file. In the simpler case of just being confronted with missing atoms the “complete_pdb” routine from MODELLERs scripts module was used to fill in atoms absent from the coordinate file. For modelling missing residues and loops, first an alignment file in the PIR database format was prepared. The sequence of the protein of interest was read out from the PDB records and aligned to the corresponding FASTA sequence. Residues missing were marked with gap characters “-“ at the correct position. For each ligand in the template structure a “.” character was set at the end of the two sequences to include them into model building. Then a

subclass of MODELLERs “automodel” class, with a select_atoms routine to exclusively select residues of the model originally missing from the coordinate file, was created to stop MODELLER from refining and moving all atoms of the newly generated model. The env.io.hetatm parameter was set to “True” to enable reading ligand coordinates from the alignment file. As for the last step a script for ab initio modelling, using the created sub-class, was used to model and refine the incomplete parts of the structure.

2.4 Correction of Wrongly Assigned ASN and GLN Residues

For identifying the correct orientation of ASN and GLN sidechain rotamers the “NQ-flipper” web service (Weichenberger and Sippl 2007) was used. Previously completed PDB files were uploaded and analyzed. Residues where energy differences Δz (calculated as difference between z(PDB) and z(flipped)) reached values over 1, indicating a clearly flipped sidechain, were marked for a 180° rotation. For reasons of prudence and to prevent over interpretation, flip recommendations with Δz between 0.3 and 1 were skipped and original sidechain orientations have been retained. Files with the corrected rotamers were downloaded from the webserver.

2.5 Calculating pK_a Values of Ionizable Residues and Ligands

pK_a calculations were carried out by the software PROPKA 3.1 (Søndergaard et al. 2011; Olsson et al. 2011) and the web server H++ (accessible under: http://biophysics.cs.vt.edu) (Anandakrishnan, Aguilar, and Onufriev 2012). For previously processed PDB files with corrected rotamers, an empirical prediction of titration states for each ionizable residue and ligand was performed by PROPKA 3.1. A smaller subset of PDB structures was uploaded to the H++ server to calculate pK_a values using the Poisson–Boltzmann equation. Residues with unusual pK_a values, referring to shifts far enough from their intrinsic pK_a values to result in a different protonation state at pH levels of 7.4 (LYS, ARG, CYS, TYR: $pK_a \leq 7.4$; ASN, GLN, HIS: $pK_a \geq 7.4$), and suggested protonation states of ligands were listed for each individual PDB file.

2.6 Applying Charges to Ligands

For ligands carrying formal charges on functional groups at pH levels of 7.4, as suggested by the PROPKA 3.1 software, a manual assignment of their protonation state was carried out using the MAESTRO software suite. Ligand coordinates were isolated from the PDB file and inserted into a newly created entry. Bond orders were assigned using the corresponding tool implemented in MAESTRO, followed by a visual inspection to secure the correctness of the automated procedure. Next, formal charges were applied to atoms of de- or protonated functional groups if not recognized by MAESTRO or different to the automatically computed charges. Having assigned and checked all formal charges, hydrogen atoms were added by the “add Hydrogens” function. Ligand entries with attached Hydrogen atoms and edited protonation states were then saved as mol2 files for the further use in MD simulations.

Chapter 3

Results and Discussion

PDB files used in this thesis originate from the target list of the DUD-E database (Mysinger et al. 2012) and were downloaded directly from the RCSB Protein Data Bank. For the conducted experiments a pool of 77 structures was chosen. A workflow for evaluation and preparation of PDB files, targeting frequent problems and errors that can be found in most models of macromolecular structures, was created (Figure 3.1). This workflow can be sectioned into a structure evaluation part (Step 1, 2) and a part for analysis and preparation (Step 3, 4, 5, 6). It includes the following procedures:

Step 1 – PDB files are uploaded to the WHAT IF server to determine various validation parameters. Received WHAT IF reports are then checked for gross errors.

Step 2 – Structures are analyzed by the software VHELIBS. Binding sites and ligands are evaluated and divided into the three categories: “good”, “dubious” or “bad”. “Dubious” binding sites and ligands undergo a visual inspection to decide on the further use of the file. Structures containing “bad” segments are discarded.

Step 3 – Missing atoms and residues within PDB files are modelled with the software MODELLER, using its ab initio modelling function.

Step 4 – Completed PDB files are analyzed for incorrectly assigned ASN and GLN sidechains via the NQ-flipper web service. Unambiguously flipped rotamer orientations are corrected.

Step 5 – pK_a values of ionizable residues and ligands are calculated by PROPKA 3.1 to detect amino acids with unexpected protonation states at physiological pH values.

Step 6 – Hydrogen atoms are assigned to ligands by the MAESTRO suite according to protonation states of functional groups as determined by PROPKA 3.1.

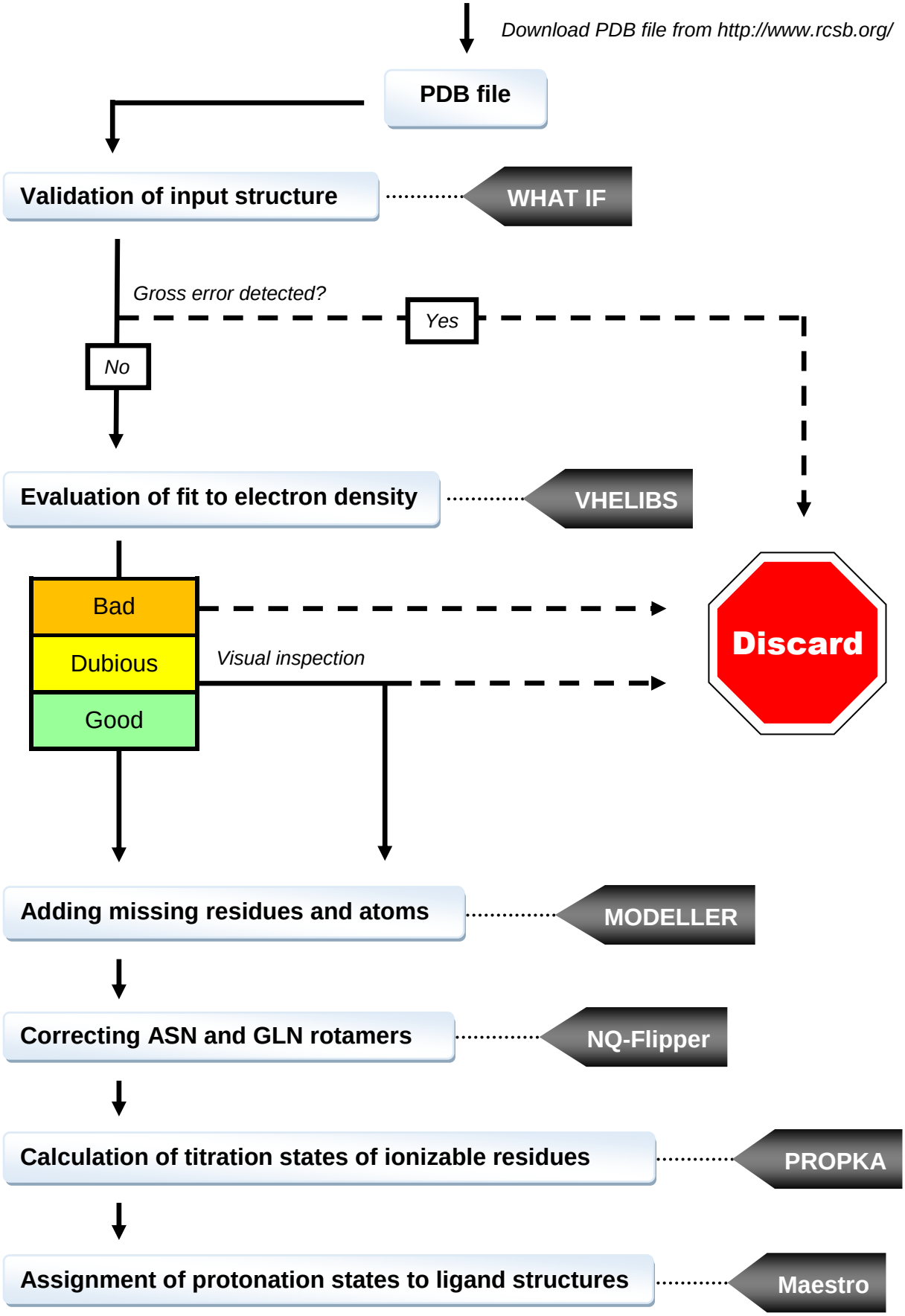

Fig 3.1 *Workflow for validation and preparation of PDB files.*

3.1 Model Validation

As a first step all PDB files were uploaded to the WHAT IF webserver and model checks, evaluating a large number of validation parameters, were carried out. Although the web service offers a good way to get an impression of the overall model quality, errors occurred in all investigated structures and do not necessarily indicate bad model quality. Especially tests for tau angle deviations, side chain planarity, bumps (short interatomic distances) and chi-1/chi-2 rotamers frequently give rise to errors even with structures of good resolution. Also errors indicating sidechain flips can be ignored as they are analyzed and corrected within step 4 of the workflow. However it is recommended to use what check reports to identify structures of very poor quality or gross errors like backwards tracing of chains.

Removing atomic clashes

The WHAT IF website also offers a tool for removing bumps from structures by rotating concerned sidechains. However an evaluation of structures on which this script was run showed that in some cases bumps fall into worse bins or have their numbers even increased after the process. Additionally, by also rotating ASN and GLN sidechains, analysis of wrongly assigned rotamers become impossible and already flipped rotamers could rotate towards unfavorable orientations. Taking these uncertainties into account it was decided to exclude this step from the presented workflow.

3.2 Evaluation of the Ligands Fit to the Electron Density Map

In order to evaluate the fit of ligands and binding sites to their experimental determined electron density maps, VHELIBS was run on a subset of 50 PDB structures (containing 73 binding sites and corresponding ligands). Each structure was evaluated individually using two different sets of parameters (as shown in Table 3.1). During the first run the default settings of VHELIBS for PDB files, including the radius around ligands considered as binding sites, caps for RSR values and "good" RSR values, minimum RSCC values, the maximum R_{free} value and a preset score tolerance were taken into account. As for the second run using custom settings, the parameters OWAB, a Resolution limit and a maximum R-R_{free} value were added to the default settings. A total of 7 PDB structures were lacking ligands or available reflection data, contained blacklisted ligands or, in less severe cases, were missing information on a set parameter (e.g. R_{free} of 1LRU) and thus were discarded and substituted. Remains from cell lysis and artefacts from purification or crystallization labelled as ligands by VHELIBS were excluded from the evaluation.

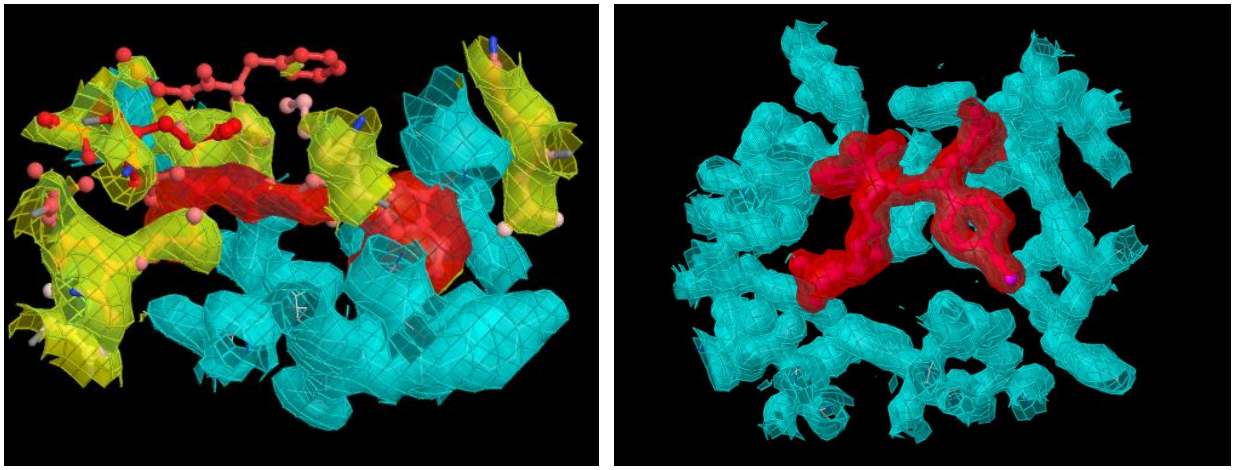

Figure 3.2: Example for "Bad" /"Good" fit to electron density as labeled by VHELIBS. The left picture shows the structure of c-Src (PDB code 3el8), which was found to have a "bad" fit for both binding site and ligand to their electron density. On the right, the structure of farnesyltransferase (PDB code 3e37) with a "good" fit of binding site and ligand is illustrated.

Using the default parameters, 18 binding sites (25%) were labelled as “good”, 46 (63%) as “dubious” and 9 (12%) as “bad”. Of the corresponding ligands 47 (64%) show a “good” fit, 22 (30%) are marked as “dubious” and only 4 (6%) as “bad”. After adding the additional parameters VHELIBS considered 8 binding sites (11%) as “good”, 50 (68%) as “dubious” and 15 (21%) as “bad”. Analogous 31 ligands (42%) were considered to have a “good” fit, 34 (47%) to be “dubious” and 8 (11%) were marked as “bad”.

PDB code	(A) Default parameters (PDB)						(B) Custom set parameters					
	Binding site			*Ligand*			*Binding site*			*Ligand*		
	Good	Dubious	Bad	Good	Dubious	Bad	Good	Dubious	Bad	Good	Dubious	Bad
2OJ9			1		1				1		1	
1XL2		1			1			1			1	
2HZI		2		2				2		2		
3L3M		1			1			1			1	
1UYG		1			1				1		1	
3EL8			1		1				1			1
2GTK		1		1				1			1	
2P54		1		1				1		1		
3BQD	1			1				1			1	
2AZR		1			1			1			1	
2QD9		1			1			1			1	
3KBA		2			2			2			2	
2OJG		1		1				1		1		
2ZDT		1		1				1		1		
1L2S		3		2	1			3		2	1	
3EML		1		1				1			1	
3BKL		1		1				1		1		
1E66	1			1			1			1		
2E1W		1		1				1		1		
2VT4	4			4				4			4	
3NY8		1			1			1				1
2HV5		1		1					1		1	
2AM9		1		1				1		1		
3L5D		2			2			2			2	
3D4Q	2			2				2		2		
1H00			2			2			2			2
3BWM	1			1				1			1	

1R9O		1			1			1			1	
3NXU		2		1	1			2		1	1	
3KRJ	1			1				1			1	
3ODU		1	1	2				1	1	1	1	
3FRJ			1	1					1	1		
2I78	1			1				1			1	
3PBL		2		2				2			2	
3NXO		1		1				1		1		
2RGP		1		1				1		1		
1SJ0		1		1					1		1	
2FSZ	1	3		1	3		1	3		1	3	
3KL6		1		1					1	1		
1W7X		1		1				1		1		
2NNQ		1		1				1		1		
3BZ3		1		1					1		1	
3E37	2			2			2			2		
1ZW5			2		2				2			2
2V3F	2			2			2			2		
3KGC		1		1				1		1		
1VSO		1		1				1		1		
3MAX	2	1		3			2	1		3		
3F07		1	1			2			2			2
3NF7		2			2			2			2	

Table 3.1: Evaluation of ligands and binding sites using VHELIBS. (A) Results for structures of the test set using default (PDB) parameters. (B) Results after adding three additional quality parameters to the analysis. Segments with a "good" fit of the model to the corresponding electron density are highlighted in green, "dubious" fits are shown in yellow and "bad" fits are marked with orange.

Analysis of the investigated models, using the preset default parameters of VHELIBS for PDB files, show that the majority of ligands and binding sites essentially fit their electron density. Switching from default to the custom set parameters the amount of ligands rated “good” drops by 22%. At the same time numbers for ligand structures rated “dubious” increase by 17%. The rate of ligands marked with “bad” slightly rises by 5% using the more restricted settings. Similar results are obtained for the fit of binding sites to their electron density. While the amount of binding sites rated as “good” drops by 14 % and “dubious” ratings increase by 5%, the number of binding sites categorized as “bad” rises by 9% (see figure 3.2).

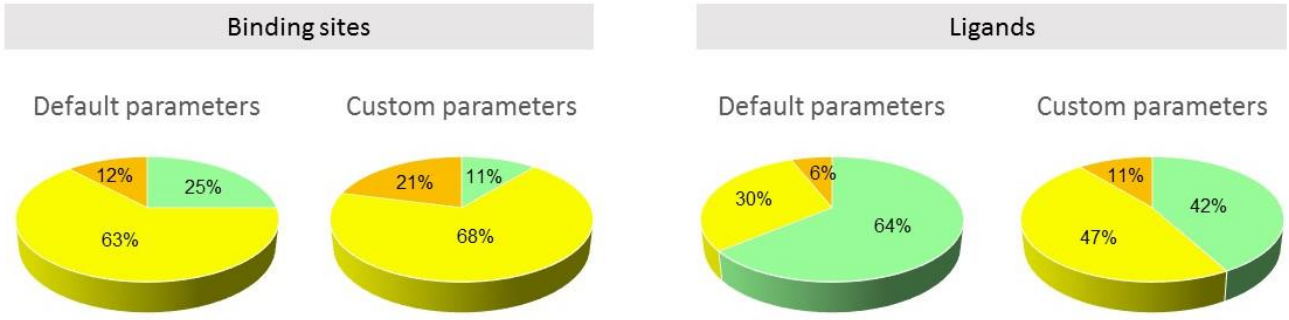

Figure 3.3 Distribution of ligands and their binding sites for default- (PDB) and custom settings. “Good” fits are displayed in green, “dubious” fits in yellow and “bad” fits in orange.

In both cases “bad” fits double after adding OWAB, R-Rfree and the resolution cut-off, showing the strong validation power of these parameters. Considering importance of accuracy, especially for binding sites and ligands, for most computational studies in the field of SBDD, analysis of these regions help to select good models and prevent serious misinterpretations by using structures otherwise discarded prior to simulation. Generally, compared to ligands, binding sites tend to easier fall into worse bins. This could be a consequence of their bigger size, making it easier to exceed VHELIBS tolerance scores. Nevertheless every structure containing binding sites or ligands rated “dubious”, should be examined carefully, whereas discarding “bad” structures for further use is recommended.

3.3 Assessing and Modelling Missing Coordinates

For the purpose of assessing amount and positions of missing coordinates 77 structures were examined for REMARK 470 (missing atoms) and REMARK 465 (missing residues) within the PDB files remark section. Only 10 (13%) of all inspected models were already complete and therefore didn't need further treatment. Missing coordinates for sidechain atoms were found in a total of 28 (36%) of all investigated structures. While the vast majority of them also lacked coordinates for at least one complete residue, two structures exclusively missed several sidechain atoms. A detailed list of missing strings and atoms of each PDB file is showed in Table 3.2.

Adding missing atoms

For mentioned PDB files containing REMARK 470 and therefore missing atoms, although providing a complete set of coordinates for the protein backbone, a short script using MODELLERs "complete_pdb" routine was run to rebuild absent sidechain atoms. Absent atom coordinates in structures containing both REMARK 470 and REMARK 465 were concurrently added within the process of modelling missing strings of residues.

Script used for adding only missing atoms to otherwise complete structure:

```
1  from modeller import *
2  from modeller.scripts import complete_pdb
3  import sys
4
5  arg = sys.argv[1]
6
7  env = environ()
8  env.libs.topology.read(file='$(LIB)/top_heav.lib')
9  env.libs.parameters.read(file='$(LIB)/par.lib')
10
11 m = complete_pdb(env, '/path_to_pdb/'+arg+'.pdb')
12 m.write(file='/path_to_save_location/complete_'+arg+'.pdb')
```

Adding missing residues

Considering the amount of atoms within a macromolecular protein structure and the limitations of x-ray crystallography, it comes as no surprise that 65 structures (84%) of the investigated test set are missing coordinates for at least one residue. As a single file often contains more than one missing string, the total amount of missing strings found in the test set was considered. Of the 242 found missing strings 154 are located at C or N termini while 88 are located somewhere within the polypeptide chain (hereinafter referred as internal string). Figure 3.4 illustrates the occurrence of missing strings with varying length.

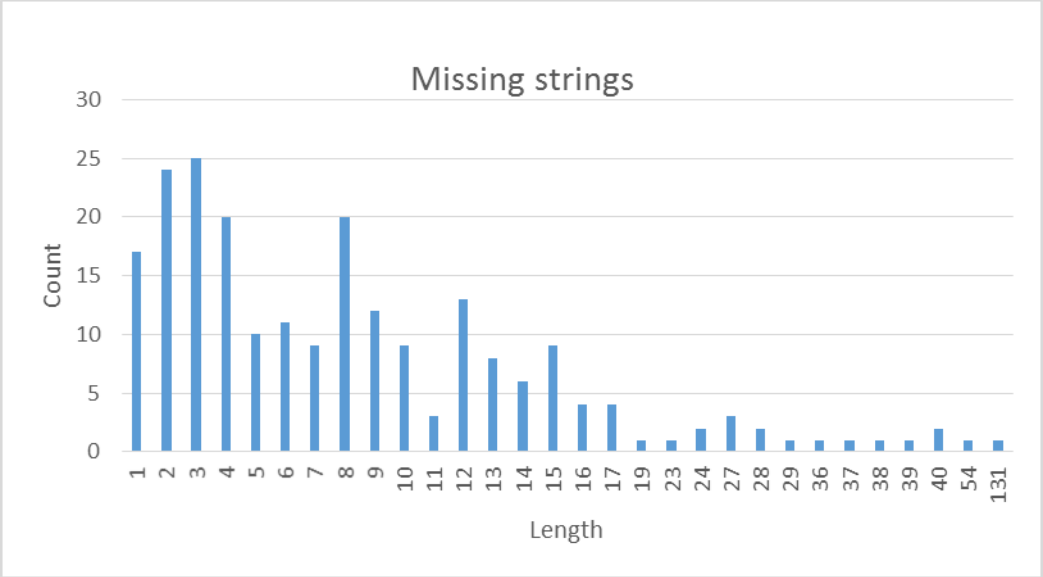

Figure 3.4: Occurrence of missing strings of certain length. The amount of missing strings decreases with increasing length. The horizontal axis shows the length of a missing string and the vertical axis shows the number of missing strings found for this specific length.

In order to model the missing strings alignment files, providing information about the PDB sequence, location of missing residues, chain breaks, the amount of HETATM records (ligands) to be included and the full sequence of the primary structure, were generated (see Figure 3.5).

```
>P1;2HZI
structureX:2HZI::A::B::::
---------DKWEMERTDITMKHKLGGGQYGEVYEGVWKKYSLTVAVKTLKEDTMEVEEFLKEAAVMKEIKHPNLVQLLG
VCTREPPFYIITEFMTYGNLLDYLRECNRQEVNAVVLLYMATQISSAMEYLEKKNFIHRDLAARNCLVGENHLVKVADFG
LSRLMTGDTYTAHAGAKFPIKWTAPESLAYNKFSIKSDVWAFGVLLWEIATYGMSPYPGIDLSQVYELLEKDYRMERPEG
CPEKVYELMRACWQWNPSDRPSFAEIHQAFETMFQES/---------DKWEMERTDITMKHKLGGGQYGEVYEGVWKKYS
LTVAVKTL----MEVEEFLKEAAVMKEIKHPNLVQLLGVCTREPPFYIITEFMTYGNLLDYLRECNRQEVNAVVLLYMAT
QISSAMEYLEKKNFIHRDLAARNCLVGENHLVKVADFGLSRLMTGDTYTAHAGAKFPIKWTAPESLAYNKFSIKSDVWAF
GVLLWEIATYGMSPYPGIDLSQVYELLEKDYRMERPEGCPEKVYELMRACWQWNPSDRPSFAEIHQAFETMFQES..*

>P1;2HZI_fill
sequence:2HZI::::::::
GAMDPSPNYDKWEMERTDITMKHKLGGGQYGEVYEGVWKKYSLTVAVKTLKEDTMEVEEFLKEAAVMKEIKHPNLVQLLG
VCTREPPFYIITEFMTYGNLLDYLRECNRQEVNAVVLLYMATQISSAMEYLEKKNFIHRDLAARNCLVGENHLVKVADFG
LSRLMTGDTYTAHAGAKFPIKWTAPESLAYNKFSIKSDVWAFGVLLWEIATYGMSPYPGIDLSQVYELLEKDYRMERPEG
CPEKVYELMRACWQWNPSDRPSFAEIHQAFETMFQES/GAMDPSPNYDKWEMERTDITMKHKLGGGQYGEVYEGVWKKYS
LTVAVKTLKEDTMEVEEFLKEAAVMKEIKHPNLVQLLGVCTREPPFYIITEFMTYGNLLDYLRECNRQEVNAVVLLYMAT
QISSAMEYLEKKNFIHRDLAARNCLVGENHLVKVADFGLSRLMTGDTYTAHAGAKFPIKWTAPESLAYNKFSIKSDVWAF
GVLLWEIATYGMSPYPGIDLSQVYELLEKDYRMERPEGCPEKVYELMRACWQWNPSDRPSFAEIHQAFETMFQES..*
```

Figure 3.5: Example for alignment file in the PIR format. The first passage contains information about the experiment used for structure determination (e.g. structureX for x-ray), the PDB code from which structural information is obtained, chain identifiers, the PDB sequence, location of missing residues (marked with "-"), chain breaks (marked with "/") and the amount of ligands to include (marked with "."). The second passage provides the complete sequence of the primary structure.

Script for writing out sequence of PDB:

```
1  import sys
2  from modeller import *
3
4  arg = sys.argv[1]
5
6  e = environ()
7  m = model(e, file='/path_to_pdb/'+arg+'.pdb')
8  aln = alignment(e)
9  aln.append_model(m, align_codes=arg1)
10 aln.write(file='/path_to_save_location/'+arg+'.seq'
```

After careful inspection of alignment files, missing residues and atoms were added using MODELLERS ab initio modelling function. In order to preserve the experimentally determined coordinates of the models as far as possible, refinement for coordinates already provided within PDB files was disabled. This step ensured that only newly modelled strings were refined and moved during the modelling process to minimize biases in further refinement steps and analyses of the structure. A total of 5 PDB files contained modified amino acids, such as phosphorylated residues, and were excluded from the modelling process as MODELLER was not able to process them correctly.

Example script used for modelling missing strings and atoms:

```
1  from modeller import *
2  from modeller.automodel import *
3
4  log.verbose()
5
6  # Select regions for refinement
7  class MyModel(automodel):
8          def select_atoms(self):
9                  return selection(self.residue_range('1:A, 9:A'),
   self.residue_range('278:B, 286:B'), self.residue_range('328:B, 331:B'))
10
11 env = environ()
12
13 # Directories for input atom files
14 env.io.atom_files_directory = ['/path_to_pdb/2HZI.pdb']
15
16 # Enable ligands
17 env.io.hetatm = True
18
19 # Directory for the alignment file
20 a = MyModel(env, alnfile = '/path_to_alignmentfile/2HZI_alignment.ali',
   knowns = '2HZI', sequence = '2HZI_fill')
21
22 # Number of models created
23 a.starting_model= 1
24 a.ending_model  = 2
25
26 a.make()
```

PDB code	(A) Missing residues					(B) Missing atoms			
	Chain ID	*Missing strings*	*Amount*	*Terminal*	*Internal*	*Residue type*	*Chain ID*	*SSEQI*	*Atoms*
2OJ9	A	1 - 4	4	x		GLU	A	1067	CG CD OE1 OE2
	A	119 - 126	8		x				
1JH4									
1XL2									
2HZI	A	1 - 9	9	x		LYS	A	234	NZ
	B	278 - 286	9	x		LYS	A	247	NZ
	B	328 - 331	4		x	ARG	A	307	CD NE CZ NH1 NH2
						GLU	A	308	CG CD OE1 OE2
						LYS	A	356	CE NZ
						LYS	A	400	CD CE NZ
						GLU	A	466	CD OE1 OE2
						LYS	A	467	NZ
						LYS	B	234	NZ
						GLU	B	238	CD OE1 OE2
						ARG	B	239	NE CZ NH1 NH2
						LYS	B	245	CE NZ
						LYS	B	247	NZ
						LYS	B	262	NZ
						LYS	B	263	CE NZ
						GLU	B	279	OE1 OE2
						GLU	B	281	CG CD OE1 OE2
						GLU	B	282	CD OE1 OE2
						ARG	B	307	NE CZ NH1 NH2
						GLU	B	308	CG CD OE1 OE2
						LYS	B	400	CG CD CE NZ
						GLU	B	462	CG CD OE1 OE2
						GLU	B	466	CD OE1 OE2

							LYS	B	467	NZ
3L3M	A	1 - 1	1	x						
	A	350 - 350	1	x						
1UYG	A	1 - 15	15	x			GLU	A	163	CG CD
	A	225 - 236	12	x			LYS	A	224	CA C O CB CG CD CE NZ
3EL8	A	1 - 10	10	x			CYS	A	277	SG
	A	164 - 176	13		x		PHE	A	278	CG CD1 CD2 CE1 CE2 CZ
	B	287 - 294	8	x			PHE	A	424	CG CD1 CD2 CE1 CE2 CZ
	B	337 - 342	6		x		PHE	B	278	CG CD1 CD2 CE1 CE2 CZ
	B	450 - 462	13		x		ARG	B	409	CG CD NE CZ NH1 NH2
							PHE	B	424	CG CD1 CD2 CE1 CE2 CZ
2GTK	A	56 - 68	13		x					
2P54							ASP	A	202	CG OD1 OD2
							ASN	A	235	CG OD1 ND2
							ASN	A	261	CG OD1 ND2
							ASN	A	393	CG OD1 ND2
							TYR	A	468	CG CD1 CD2 CE1 CE2 CZ OH
							ARG	B	686	CG CD NE CZ NH1 NH2
							LYS	B	688	CG CD CE NZ
3BQD	A	1 - 2	2	x			GLN	B	741	CG CD OE1 NE2
	B	256 - 257	2	x						
2AZR	A	1 - 1	1	x						
	A	299 - 299	1	x						
2QD9	A	1 - 10	10	x						
	A	37 - 38	2		x					
	A	124 - 126	3		x					
	A	179 - 191	13		x					
	A	359 - 366	8	x						
3KBA	A	1 - 1	1	x			GLN	A	682	CG CD OE1 NE2
	A	24 - 28	5		x		LEU	A	683	CG CD1 CD2
	A	112 - 113	2		x		ILE	A	684	CG1 CG2 CD1

	A	180 - 180	1		x
	B	254 - 254	1	x	
	B	505 - 506	2	x	

2OJG	A	1 - 38	38	x	
	A	352 - 353	2	x	x
	A	378 - 380	3	x	
2ZDT	A	1 - 6	6	x	
	A	174 - 178	5	x	x
	A	336 - 344	9	x	x
	A	363 - 364	2	x	
1L2S	A	287 - 289	3		x

ASP	A	697	CG OD1 OD2
ASP	A	709	CG OD1 OD2
GLU	A	786	OE1
GLN	A	787	CD OE1 NE2
GLU	A	791	CG CD OE1 OE2
GLN	A	803	CD OE1 NE2
ARG	A	836	NE CZ NH1 NH2
LYS	A	861	CG CD CE NZ
ARG	A	899	CG CD NE CZ NH1 NH2
LYS	A	932	CG CD CE NZ
LYS	A	933	CA C O CB CG CD CE
LYS	A	933	NZ
GLN	B	682	CG CD OE1 NE2
LEU	B	683	CG CD1 CD2
GLN	B	720	OE1 NE2
GLN	B	812	NE2
GLN	B	868	CG CD OE1 NE2

GLU	A	21	CG CD OE1 OE2
GLN	A	22	CG CD OE1 NE2
LYS	A	99	CG CD CE NZ
GLU	A	124	CG CD OE1 OE2
LYS	A	126	CG CD CE NZ
GLU	A	196	CG CD OE1 OE2
LYS	A	207	CG CD CE NZ

3EML	A	1 - 17	17	x		
	A	164 - 170	7		x	
	A	473 - 487	15	x		
3BKL	A	1 - 3	3	x		
	A	261 - 261	1		x	
	A	399 - 400	2		x	
	A	589 - 591	3	x		
1E66	A	1 - 3	3	x		
	A	536 - 543	8	x		

LYS	A	246	CG CD CE NZ
ASP	A	264	CG OD1 OD2
ARG	A	296	CG CD NE CZ NH1 NH2
GLN	B	7	CG CD OE1 NE2
GLN	B	52	CG CD OE1 NE2
GLN	B	57	CG CD OE1 NE2
ASP	B	123	CG OD1 OD2
LYS	B	126	CG CD CE NZ
GLU	B	205	CG CD OE1 OE2
LYS	B	207	CG CD CE NZ
LYS	B	246	CG CD CE NZ
LYS	B	290	CG CD CE NZ
LYS	B	299	CG CD CE NZ

LEU	A	7	CG CD1 CD2
ASN	A	42	CG OD1 ND2
ARG	A	46	CD NE CZ NH1 NH2
ARG	A	47	CD NE CZ NH1 NH2
LYS	A	52	CD CE NZ
GLU	A	89	CD OE1 OE2
GLN	A	162	CD OE1 NE2
LYS	A	192	CD CE NZ
ASN	A	253	CG OD1 ND2
ASN	A	257	CG OD1 ND2
GLU	A	260	CD OE1 OE2

2E1W	A	1 - 2	2	x	
	A	352 - 356	5	x	
2OI0					
2VT4	A	1 - 9	9	x	
	A	210 - 224	15		x
	A	299 - 313	15	x	
	B	314 - 321	8	x	
	B	523 - 537	15		x
	B	613 - 626	14	x	
	C	627 - 634	8	x	
	C	836 - 850	15		x
	C	913 - 939	27	x	
	D	940 - 949	10	x	
	D	1149 - 1163	15		x

GLU	A	268	CG CD OE1 OE2
LYS	A	270	CE NZ
GLU	A	299	CD OE1 OE2
GLU	A	350	CD OE1 OE2
ASP	A	365	OD1 OD2
ASN	A	382	OD1 ND2
LYS	A	413	CD CE NZ
GLU	A	434	CD OE1 OE2
LYS	A	454	CD CE NZ
GLU	A	455	CG CD OE1 OE2
LYS	A	478	CG CD CE NZ
HIS	A	486	CG ND1 CD2 CE1 NE2
GLN	A	488	CD OE1 NE2
GLU	A	489	CD OE1 OE2
SER	A	490	OG
LYS	A	511	CD CE NZ
GLN	A	526	CD OE1 NE2

	D	1237 - 1252	16	x	
3NY8	A	1 - 39	39	x	
	A	479 - 490	12	x	
3CQW	A	1 - 5	5	x	
	A	310 - 325	16		x
	A	341 - 342	2	x	
3D0E	A	310 - 320	11		x
	B	645 - 655	11		x
2HV5	A	1 - 2	2	x	
2AM9	A	1 - 16	16	x	

LYS	A	836	CG CD CE NZ
LYS	A	847	CG CD CE NZ
ASN	A	848	CG OD1
GLU	A	893	CG CD OE1 OE2
ILE	A	501	CG1 CG2 CD1

1S3B	A	1 - 2	2	x	
	A	502 - 520	19	x	
	B	521 - 522	2	x	
	B	1017 - 1040	24	x	
3L5D	A	1 - 17	17	x	
	A	92 - 95	4		x
	A	408 - 414	7	x	
	B	415 - 431	17	x	
	B	821 - 828	8	x	
3D4Q	A	1 - 28	28	x	
	A	182 - 193	12		x
	A	305 - 307	3	x	
	B	308 - 335	28	x	
	B	489 - 500	12		x
	B	612 - 614	3	x	
2CNK	A	175 - 175	1	x	
1BCD	A	1 - 1	1	x	
1H00	A	13 - 14	2		x

LYS	A	9	CG CD CE NZ

	A	36 - 43	8		x
	A	152 - 161	10		x

GLU	A	12	CG CD OE1 OE2
TYR	A	15	CG CD1 CD2 CE1 CE2 CZ OH
LEU	A	25	CG CD1 CD2
LYS	A	34	CG CD CE NZ
ARG	A	50	CG CD NE CZ NH1 NH2
GLU	A	51	CG CD OE1 OE2
GLU	A	73	CG CD OE1 OE2
ASN	A	74	CG OD1 ND2
LYS	A	75	CG CD CE NZ
LEU	A	96	CG CD1 CD2
ARG	A	150	CG CD NE CZ NH1 NH2
GLU	A	162	CG CD OE1 OE2
VAL	A	164	CG1 CG2
LYS	A	178	CG CD CE NZ
LYS	A	278	CG CD CE NZ
ARG	A	297	CG CD NE CZ NH1 NH2

3BWM

1R9O	A	1 - 8	8	x	
	A	21 - 25	5		x
	A	197 - 203	7		x
	A	476 - 477	2	x	
3NXU	A	1 - 6	6	x	
	A	243 - 245	3		x
	A	259 - 266	8		x
	A	475 - 485	11	x	
	B	486 - 491	6	x	
	B	728 - 730	3		x
	B	744 - 751	8		x
	B	960 - 970	11	x	
3KRJ	A	1 - 9	9	x	
	A	149 - 161	13		x

PRO	A	544	CG CD
LYS	A	545	CG CD CE NZ

	A	227 - 227	1		x
	A	328 - 335	8	x	
3ODU	A	1 - 36	36	x	
	B	503 - 539	37	x	
	B	996 - 1004	9	x	
1LRU	A	165 - 168	4	x	
	B	333 - 336	4	x	
	C	498 - 504	7	x	
3FRJ	A	1 - 14	14	x	
	A	224 - 225	2		x
	A	279 - 286	8	x	
	B	287 - 299	13	x	
	B	565 - 572	8	x	

MET	B	63	CE
LYS	B	68	CG CD CE NZ
SER	B	319	O

2I78

3PBL	A	1 - 40	40	x	
	A	473 - 481	9	x	
	B	482 - 521	40	x	
	B	626 - 634	9		x
	B	954 - 962	9	x	

GLN	A	144	CG CD OE1 NE2
SER	A	145	OG
GLU	A	1108	CG CD OE1 OE2
THR	A	357	OG1 CG2
SER	B	145	OG
ARG	B	149	CG CD NE CZ NH1 NH2
ASN	B	1040	CG OD1 ND2
ASN	B	1053	CG OD1 ND2
ASN	B	1055	CG OD1 ND2
ARG	B	1080	CG CD NE CZ NH1 NH2

3NXO

2RGP	A	33 - 36	4		x
	A	48 - 52	5		x
	A	167 - 174	8		x
	A	287 - 294	8		x
	A	303 - 308	6		x
1SJ0	A	225 - 226	2		x

GLU	A	1015	CG CD OE1 OE2

	A	246 - 248	3	x	
2FSZ	A	1 - 7	7	x	
	A	155 - 166	12		x
	A	224 - 226	3		x
	A	246 - 246	1	x	
	B	247 - 249	3	x	
	B	401 - 412	12		x
	B	470 - 472	3		x
	B	492 - 492	1	x	
3KL6	A	233 - 241	9	x	
	B	242 - 245	4	x	
	B	296 - 298	3	x	
1W7X	A	164 - 167	4		x
	B	271 - 273	3		x

LYS	H	60	CE NZ
ARG	H	62	NE CZ NH1 NH2
GLN	H	170	CG CD OE1 NE2
LYS	L	143	CG CD CE NZ
ARG	L	144	CG CD NE CZ NH1 NH2
GLU	A	54	CB CG CD OE1 OE2
LYS	A	79	CD CE NZ

2NNQ

3BZ3	A	32 - 35	4		x
	A	158 - 170	13		x
3C4F	A	117 - 128	12		x
	B	419 - 430	12		x

1J4H

3E37	A	1 - 54	54	x	
	A	370 - 379	10	x	
	B	380 - 393	14	x	
	B	804 - 816	13	x	
1ZW5	A	1 - 14	14		x
	A	35 - 36	2	x	

LYS	A	28	CE NZ
GLN	A	29	CG CD OE1 NE2
GLN	A	37	CG CD OE1 NE2

2V3F

A	32 - 34	3		x
A	501 - 505	5	x	
B	506 - 507	2	x	
B	534 - 539	6		x
B	1005 - 1010	6	x	

ARG	A	40	CD NE CZ NH1 NH2
ASP	A	45	CG OD1 OD2
LYS	A	90	CE NZ
LYS	A	135	NZ
GLN	A	194	CG CD OE1 NE2
LYS	A	205	CD CE NZ
LYS	A	212	NZ
GLU	A	236	CD OE1 OE2
LYS	A	237	CG CD CE NZ
GLU	A	295	CG CD OE1 OE2
LYS	A	301	NZ
GLU	A	302	CG CD OE1 OE2
LYS	A	307	CD CE NZ
LYS	A	361	CE NZ
ARG	A	366	NE CZ NH1 NH2
LYS	A	367	CE NZ
GLU	A	72	CD OE1 OE2
ASN	A	146	OD1 ND2
LYS	A	155	CE NZ
ARG	A	211	CD NE CZ NH1 NH2
GLU	A	222	CD OE1 OE2
LYS	A	224	CE NZ
LYS	A	408	CE NZ
LYS	A	441	CG CD CE NZ
LYS	A	466	CG CD CE NZ
LEU	A	498	CA C O CB CG CD1 CD2
GLU	B	111	CD OE1 OE2
LYS	B	155	CD CE NZ
ARG	B	211	NE CZ NH1 NH2
LYS	B	441	CD CE NZ
GLN	B	497	CA C O CB CG CD OE1 NE2

3KGC	A	1 - 4	4	x	
	A	262 - 263	2	x	
	B	264 - 266	3	x	
	B	526 - 526	1	x	
1VSO	A	1 - 4	4	x	
	A	65 - 69	5		x
	A	164 - 166	3		x
	A	256 - 257	2	x	
3MAX	B	368 - 370	3	x	
	C	735 - 735	1	x	
3F07	A	1 - 12	12	x	
	A	379 - 388	10	x	
	B	389 - 400	12	x	
	B	769 - 776	8	x	
	C	777 - 788	12	x	
	C	860 - 883	24		x
	C	1153 - 1164	12	x	
3NF7	A	1 - 27	27	x	
	A	161 - 164	4		x
	A	181 - 183	3	x	
	B	184 - 210	27	x	
	B	344 - 347	4		x
	B	364 - 366	3	x	
3LAN	A	554 - 560	7	x	
	B	561 - 564	4	x	
	B	626 - 627	2		x
	B	776 - 791	16		x
	B	917 - 920	4		x
	B	990 - 1120	131	x	
3CCW	A	1 - 6	6	x	
	A	428 - 441	14	x	

LYS	C	13	CG CD CE NZ

HIS	B	361	CG ND1 CD2 CE1 NE2

	B	442 - 447	6	x		
	B	869 - 882	14	x		
	C	883 - 891	9	x		
	C	903 - 905	3		x	
	C	1309 - 1323	15	x		
	D	1324 - 1346	23	x		
	D	1365 - 1372	8		x	
	D	1749 - 1764	16	x		
3F9M	A	1 - 8	8	x		
	A	99 - 102	4		x	
	A	464 - 470	7	x		
4TRJ	A	1 - 1	1	x		
2ICA	A	1 - 3	3	x		
	A	183 - 183	1	x		

3LPB	A	1 - 4	4	x		
	A	82 - 86	5		x	
	A	230 - 234	5		x	
	A	295 - 295	1	x		
	B	296 - 296	1	x		
	B	377 - 381	5		x	
	B	471 - 473	3		x	

MET	A	128	SD CE
ASN	A	163	OD1 ND2
ASP	A	229	OD1 OD2
LYS	A	268	CD CE NZ
GLU	A	269	CD OE1 OE2
ASP	A	290	OD1 OD2
GLU	A	293	CD OE1 OE2
GLU	A	301	CD OE1 OE2
GLN	A	843	CG CD OE1 NE2
ARG	A	847	CZ NH1 NH2
LYS	A	857	CE NZ
LYS	A	883	CE NZ
HIS	A	886	CG ND1 CD2 CE1 NE2
SER	A	887	OG
GLU	A	889	CG CD OE1 OE2
GLU	A	890	CG CD OE1 OE2
ARG	A	893	CD NE CZ NH1 NH2
ARG	A	897	CD NE CZ NH1 NH2
GLN	A	906	OE1 NE2

ASN	A	924	CG OD1 ND2
LYS	A	926	CE NZ
LYS	A	943	CD CE NZ
LYS	A	945	NZ
ARG	A	947	CG CD NE CZ NH1 NH2
GLN	A1	3	CG CD OE1 NE2
GLU	A1	15	CG CD OE1 OE2
LYS	A1	53	CE NZ
ILE	A1	65	CG1 CG2 CD1
GLN	A1	72	CG CD OE1 NE2
MET	A1	73	CG SD CE
ILE	A1	74	CG1 CG2 CD1
LYS	A1	83	CG CD CE NZ
SER	B	839	OG
ASP	B	840	CG OD1 OD2
GLU	B	845	CG CD OE1 OE2
ARG	B	847	CZ NH1 NH2
LYS	B	850	CG CD CE NZ
LYS	B	857	CD CE NZ
ASN	B	859	CG OD1 ND2
ASN	B	874	CG OD1 ND2
GLU	B	877	CG CD OE1 OE2
GLU	B	890	CG CD OE1 OE2
GLU	B	896	CD OE1 OE2
ARG	B	897	CZ NH1 NH2
LYS	B	912	CE NZ
ASN	B	924	CG OD1 ND2
LYS	B	945	CE NZ
GLN	B1	3	CD OE1 NE2
LYS	B1	5	CE NZ
LYS	B1	11	CG CD CE NZ

3G0E	A	333	336	4	x	
2B8T	A	1	10	10	x	
	A	217	223	7	x	
	B	224	233	10	x	
	B	274	288	15		x
	B	438	446	9	x	
	C	447	456	10	x	
	C	498	505	8		x
	C	664	669	6	x	
	D	670	680	11	x	
	D	719	722	4		x
	D	889	892	4	x	
2I0E	A	1	18	18	x	
	A	305	306	2		x
	A	350	353	4	x	
	B	354	371	18	x	
	B	655	683	29		x
	B	703	706	4	x	
2OF2						
3CHP	A	1	2	2	x	
3M2W	A	1	2	2	x	
	A	201	207	7		x

GLU	B1	15	CG CD OE1 OE2
SER	B1	29	OG
LYS	B1	53	CG CD CE NZ
GLN	B1	70	CG CD OE1 NE2
GLN	B1	72	CG CD OE1 NE2

LYS	D	217	CG CD CE NZ
LYS	D	218	CG CD CE NZ
ARG	D	219	CG CD NE CZ NH1 NH2

HIS	A	47	CG ND1 CD2 CE1 NE2
LYS	A	55	NZ
LYS	A	64	CD CE NZ
THR	A	66	OG1 CG2
GLN	A	68	CG CD OE1 NE2
VAL	A	69	CG1 CG2

2AA2	A	1 - 17	17	x	
	A	200 - 204	5		x
	A	275 - 275	1	x	

LEU	A	70	CG CD1 CD2
LEU	A	72	CG CD1 CD2
ILE	A	74	CG1 CG2 CD1
LYS	A	77	CG CD CE NZ
LEU	A	79	CG CD1 CD2
GLN	A	80	CG CD OE1 NE2
ARG	A	85	CZ NH1 NH2
GLN	A	87	CD OE1 NE2
LYS	A	89	CD CE NZ
GLN	A	96	CD OE1 NE2
GLN	A	113	CD OE1 NE2
ARG	A	131	CD NE CZ NH1 NH2
ASP	A	155	CG OD1 OD2
GLN	A	156	CG CD OE1 NE2
ASN	A	200	CG OD1 ND2
GLU	A	238	OE1 OE2
SER	A	265	OG
ARG	A	280	NE CZ NH1 NH2
LYS	A	330	CD CE NZ
LYS	A	346	CE NZ
GLU	A	347	CD OE1 OE2
GLU	A	354	CD OE1 OE2
LEU	A	727	CG CD1 CD2
ARG	A	732	CG CD NE CZ NH1 NH2
GLU	A	763	CG CD OE1 OE2
GLU	A	847	CG CD OE1 OE2
ARG	A	861	CG CD NE CZ NH1 NH2
LYS	A	887	CG CD CE NZ
SER	A	888	OG
LYS	A	905	CG CD CE NZ
SER	A	914	OG

							GLN	A	916	CG CD OE1 NE2
3LQ8	A	1 - 4	4	x					
	A	68 - 70	3		x				
	A	100 - 103	4		x				
	A	181 - 182	2		x				
	A	189 - 194	6		x				
	A	298 - 302	5	x					
3EQH	A	1 - 5	5	x			CYS	A	277	C O CB SG
	A	245 - 273	29		x				
	A	349 - 360	12	x					
1QW6									
1B9V									

Table 3.2: Missing atoms and missing residues of PDB files within the test set. Models without any missing coordinates are marked green. Structures with modified residues declined by MODELLER are marked red. (A) Missing strings of residues are listed for every model including their chain identifier, length, residue numbers and information about their location within chains (terminal or internal). Residues listed were renumbered from 1 prior to modelling. (B) The second column lists missing atoms of the investigated structures with corresponding amino acids, chain identifier and residue number as listed PDB files.

Overall, results obtained from the analyses of missing coordinates lie within previously assumed values. Although the found numbers exceed the percentage of PDB files containing missing residues published in previous studies, they match the study of Dijnovic-Carugo & Carugo in 2015 which found that more than 80% of all PDB structures with resolutions of 1.75 Å and below contain at least one missing residue. As structures were chosen randomly, the test set covers a wide spectrum of resolutions, ranging from 1.35 Å to 3.30 Å. However the size of the test set used is too small to divide it into different resolution bins. Considering that most of the structures examined were refined with resolutions below 1.75 Å though, values around 80% are not surprising. Observing occurrences of missing strings of certain lengths, it was found that the amount of detected missing strings decreases with increasing length. Looking at the location of these strings, the results show that 64% are located at C or N termini which most likely is due to the enhanced freedom of movement of terminal segments. Missing coordinates at termini are most often considered to be minor problems. Nevertheless not taking them into account could lead to biased results in some cases (e.g. statistical analysis). On the contrary, internally located missing strings are far more troublesome for end users as they lead to unintended chain breaks during molecular dynamic simulations, might lead to misinterpretations of electrostatic potentials at protein surfaces and could cause biased results in docking simulations due to their absence. For the purpose of modelling ab initio prediction was chosen over comparative modelling, because very often absent strings are located in regions also missing in homolog structures and the majority of missing strings in PDB structures are of short length. When encountering modified residues, structures have to be discarded or substituted with equivalent amino acids to allow MODELLER to read input sequences properly. In summary it can be stated, that due to the extremely high frequency of missing residues, rebuilding missing coordinates is of great importance to prevent possible consequential errors.

2.4 Flipped ASN and GLN Rotamers

For the next step an analysis of ASN and GLN sidechains was carried out to detect energetically unfavorable rotamer orientations. Both of these residues are polar and therefore most often are located in surface regions exposed to solvent, which increases the probability of being absent from experimentally determined coordinates due to their higher flexibility. Residues not included in original PDB files could then alter the overall percentage of likely flipped ASN and GLN compared to the total occurrence of the two amino acids. Therefore, as suggested in the presented workflow, only structures with previously modelled strings and models already providing a complete set of coordinates were uploaded to the NQ-flipper website. Excluding the experimentally determined parts of structures from refinement during the modelling process ensured the remaining of sidechains in their original state. By taking all structures into account that passed the modelling step, a total of 72 PDB files were checked for energy differences of their ASN and GLN rotamer conformations. Table 3.3 shows residues marked for a flip, z-scores for the two corresponding rotamers and the flip indicator as suggested by NQ-flipper. While PDB files within the test set on average contain 23% incorrectly assigned sidechains, evaluation of the examined files suggest that models can hold up to 55% flipped residues. Altogether 682 (356 GLN, 326 ASN) sidechains were marked for a 180° flip by the web service. However, only those with energy differences (Δz) ≥ 1 (marked with the flip indicator “F”), which made up 82% of all suggested flips, were corrected for further steps of the workflow. Sidechains with a $\Delta z \leq 1$ (marked with the flip indicator “f”) were excluded from the flipping process in order to prevent over interpretation, as flips for sidechains with smaller energy differences may be recommended but are not clear indicators for such.

PDB code	%	Chain ID	SSEQI	Residue type	z(PDB)	z(Flipped)	Δz	Flip?
2OJ9	25%	A	28	GLN	5.9	4.1	1.8	F
		A	57	ASN	11.7	-0.7	12.4	F
		A	121	ASN	2.8	0.9	1.9	F
		A	150	ASN	1.0	-1.5	2.5	F
		A	235	GLN	2.5	1.7	0.8	f

1XL2	5%	B	18	GLN	0.3	-0.4	0.7	f
2HZI	13%	A	8	ASN	3.1	0.3	2.8	F
		A	29	GLN	0.3	-1.4	1.7	F
		A	74	ASN	0.1	-1.0	1.1	F
		A	108	ASN	8.9	-1.2	10.1	F
		B	468	ASN	1.9	-0.8	2.7	F
3L3M	26%	A	33	GLN	0.5	-1.1	1.6	F
		A	92	ASN	2.5	-0.8	3.3	F
		A	159	ASN	2.3	0.9	1.4	F
		A	166	ASN	6.7	2.1	4.6	F
		A	192	GLN	1.3	0.4	0.9	f
		A	300	ASN	4.3	0.1	4.2	F
		A	319	ASN	1.5	0.1	1.4	F
		A	337	ASN	2.7	-0.5	3.2	F
1UYG	15%	A	40	ASN	0.3	-0.7	1.0	F
		A	79	ASN	2.4	0.3	2.1	F
		A	194	GLN	2.6	0.0	2.6	F
3EL8	21%	A	40	ASN	2.9	-0.3	3.2	F
		A	77	GLN	1.7	1.0	0.7	f
		A	279	GLN	5.0	2.9	2.1	F
		A	285	ASN	1.1	-1.3	2.4	F
		B	290	GLN	2.4	-1.7	4.1	F
		B	314	GLN	5.8	2.1	3.7	F
		B	326	ASN	2.2	-1.4	3.6	F
		B	513	GLN	0.3	-0.6	0.9	f
		B	536	GLN	1.0	-0.8	1.8	F
2GTK	20%	A	139	GLN	0.9	-1.8	2.7	F
		A	196	ASN	1.1	0.7	0.4	f
		A	206	ASN	2.6	2.0	0.6	f
		A	245	GLN	1.4	0.0	1.4	F
		A	248	GLN	1.8	-0.7	2.5	F
2P54	41%	A	34	ASN	5.9	1.8	4.1	F
		A	60	ASN	7.0	1.0	6.0	F
		A	63	GLN	4.5	3.6	0.9	f
		A	76	GLN	5.9	-0.1	6.0	F
		A	98	ASN	2.0	-1.3	3.3	F
		A	135	ASN	1.2	-0.8	2.0	F
		A	165	ASN	2.5	-1.4	3.9	F
		A	192	ASN	78.0	73.4	4.6	F
		A	200	GLN	2.4	0.9	1.5	F
		A	241	GLN	1.7	-1.2	2.9	F
3BQD	34%	A	64	ASN	4.1	0.4	3.7	F
		A	93	GLN	3.5	0.9	2.6	F

		A	108	ASN	4.7	2.3	2.4	F
		A	185	ASN	6.3	2.9	3.4	F
		A	188	GLN	1.9	0.8	1.1	F
		A	189	ASN	1.3	-0.5	1.8	F
		A	191	GLN	1.7	-0.4	2.1	F
		A	216	GLN	1.1	-0.6	1.7	F
		A	238	GLN	4.4	3.9	0.5	f
2AZR	12%	A	123	GLN	3.8	-0.4	4.2	F
		A	139	ASN	2.6	0.6	2.0	F
		A	166	GLN	1.0	-1.0	2.0	F
2QD9	20%	A	31	GLN	2.7	-0.6	3.3	F
		A	120	ASN	2.8	-1.0	3.8	F
		A	121	ASN	2.8	1.0	1.8	F
		A	202	ASN	2.1	1.0	1.1	F
		A	208	GLN	3.3	0.5	2.8	F
		A	361	GLN	1.6	-1.0	2.6	F
3KBA	29%	A	2	GLN	2.0	1.5	0.5	f
		A	40	GLN	0.4	-0.2	0.6	f
		A	107	GLN	2.1	-1.2	3.3	F
		A	123	GLN	2.1	-0.2	2.3	F
		A	132	GLN	1.6	-1.5	3.1	F
		A	135	GLN	1.2	-0.5	1.7	F
		A	217	GLN	1.4	-0.9	2.3	F
		B	255	GLN	2.4	-1.0	3.4	F
		B	293	GLN	1.0	-1.3	2.3	F
		B	314	ASN	1.4	-1.0	2.4	F
		B	376	GLN	1.8	-0.7	2.5	F
		B	385	GLN	7.5	5.6	1.9	F
		B	388	GLN	3.4	0.5	2.9	F
		B	441	GLN	2.2	-0.6	2.8	F
		B	459	GLN	0.8	0.0	0.8	f
		B	470	GLN	0.9	-1.2	2.1	F
2OJG	16%	A	107	ASN	1.4	0.6	0.8	f
		A	174	ASN	2.9	0.1	2.8	F
		A	221	ASN	3.3	-0.7	4.0	F
		A	273	ASN	0.0	-0.5	0.5	f
		A	282	ASN	1.7	-1.6	3.3	F
2ZDT	24%	A	9	GLN	2.2	-1.4	3.6	F
		A	28	ASN	2.5	-0.3	2.8	F
		A	84	ASN	0.4	0.0	0.4	f
		A	120	GLN	1.3	0.8	0.5	f
		A	253	GLN	1.0	-2.0	3.0	F
		A	258	ASN	2.2	1.4	0.8	f

		A	317	GLN	1.8	-0.5	2.3	F
		A	362	ASN	2.3	-0.5	2.8	F
1L2S	19%	A	4	GLN	1.6	-0.8	2.4	F
		A	19	GLN	2.1	-1.5	3.6	F
		A	20	GLN	1.9	-0.6	2.5	F
		A	32	GLN	2.9	-1.9	4.8	F
		A	134	ASN	2.9	-1.8	4.7	F
		A	169	GLN	0.8	0.0	0.8	f
		A	172	GLN	2.5	2.1	0.4	f
		A	247	GLN	3.0	0.8	2.2	F
		A	358	GLN	3.3	0.2	3.1	F
		B	362	GLN	3.0	0.9	2.1	F
		B	377	GLN	2.9	-0.4	3.3	F
		B	378	GLN	2.6	-0.8	3.4	F
		B	407	GLN	3.7	1.3	2.4	F
		B	412	GLN	2.0	0.6	1.4	F
		B	492	ASN	2.0	-1.9	3.9	F
		B	716	GLN	2.6	1.0	1.6	F
3EML	4%	A	362	ASN	2.0	0.8	1.2	F
		A	442	ASN	3.4	0.2	3.2	F
3BKL	9%	A	20	GLN	2.7	-2.0	4.7	F
		A	54	GLN	1.4	-1.0	2.4	F
		A	94	GLN	0.4	-0.4	0.8	f
		A	159	GLN	1.5	-1.4	2.9	F
		A	205	GLN	0.3	-0.7	1.0	f
		A	518	GLN	1.2	-1.1	2.3	F
1E66	21%	A	9	ASN	7.2	2.3	4.9	F
		A	42	ASN	8.3	0.4	7.9	F
		A	68	GLN	1.1	0.0	1.1	F
		A	162	GLN	5.2	-1.8	7.0	F
		A	253	ASN	6.1	0.4	5.7	F
		A	257	ASN	5.2	1.3	3.9	F
		A	310	ASN	4.9	-2.2	7.1	F
		A	374	GLN	3.4	2.1	1.3	F
		A	382	ASN	1.6	-0.3	1.9	F
		A	488	GLN	3.7	-1.8	5.5	F
		A	526	GLN	1.9	-1.7	3.6	F
2E1W	39%	A	2	GLN	2.5	0.2	2.3	F
		A	137	GLN	1.2	0.1	1.1	F
		A	157	GLN	0.7	-0.7	1.4	F
		A	174	GLN	2.4	0.5	1.9	F
		A	198	GLN	0.9	-0.5	1.4	F
		A	221	ASN	1.2	0.1	1.1	F

		A	255	ASN	1.9	-1.6	3.5	F
		A	286	GLN	5.1	2.0	3.1	F
		A	308	GLN	3.6	0.6	3.0	F
2OI0	36%	A	249	ASN	1.6	-1.1	2.7	F
		A	264	ASN	2.3	-0.5	2.8	F
		A	377	ASN	3.9	-1.4	5.3	F
		A	389	ASN	2.7	0.1	2.6	F
		A	410	ASN	0.5	0.1	0.4	f
		A	429	GLN	2.2	0.9	1.3	F
		A	456	GLN	0.8	-1.6	2.4	F
		A	467	GLN	4.6	2.7	1.9	F
		A	471	GLN	1.8	0.0	1.8	F
2VT4	13%	A	158	GLN	1.4	-0.9	2.3	F
		B	321	GLN	8.3	7.9	0.4	f
		B	582	ASN	1.1	0.6	0.5	f
		B	592	ASN	16.2	14.9	1.3	F
		C	895	ASN	1.2	0.6	0.6	f
		C	905	ASN	8.9	6.6	2.3	F
		D	948	GLN	2.5	-0.7	3.2	F
		D	1146	GLN	-0.5	-1.3	0.8	f
		D	1218	ASN	16.7	15.7	1.0	f
3NY8	14%	A	23	ASN	4.7	0.2	4.5	F
		A	204	ASN	5.2	3.7	1.5	F
		A	290	ASN	1.1	0.0	1.1	F
		A	377	ASN	3.0	-0.2	3.2	F
		A	448	ASN	1.4	0.9	0.5	f
		A	454	ASN	2.6	0.2	2.4	F
2HV5	35%	A	8	ASN	4.2	-0.4	4.6	F
		A	27	GLN	0.8	-1.0	1.8	F
		A	50	GLN	0.6	-1.2	1.8	F
		A	51	ASN	3.2	-1.4	4.6	F
		A	94	GLN	0.9	-0.9	1.8	F
		A	137	ASN	4.6	0.6	4.0	F
		A	183	ASN	4.2	-1.0	5.2	F
		A	201	GLN	0.8	-2.0	2.8	F
		A	242	ASN	1.6	-0.6	2.2	F
		A	284	GLN	3.8	-1.8	5.6	F
2AM9	25%	A	22	ASN	7.9	4.8	3.1	F
		A	39	ASN	0.7	0.2	0.5	f
		A	58	GLN	1.3	0.1	1.2	F
		A	74	ASN	1.2	-1.8	3.0	F
		A	80	GLN	2.1	-1.1	3.2	F
		A	195	ASN	16.8	0.0	16.8	F

		A	222	GLN	1.1	-0.9	2.0	F
1S3B	13%	A	117	ASN	1.8	0.2	1.6	F
		A	170	ASN	1.0	-0.8	1.8	F
		A	251	ASN	4.3	0.8	3.5	F
		A	416	GLN	0.5	-0.6	1.1	F
		B	637	ASN	1.8	0.1	1.7	F
		B	690	ASN	1.1	-0.8	1.9	F
		B	771	ASN	4.2	0.8	3.4	F
		B	936	GLN	0.5	-0.6	1.1	f
3L5D	12%	A	132	ASN	5.1	-0.5	5.6	F
		A	315	GLN	1.3	-0.5	1.8	F
		A	409	GLN	2.5	1.5	1.0	F
		B	463	ASN	1.3	-0.1	1.4	F
		B	508	GLN	2.8	-1.0	3.8	F
		B	546	ASN	4.0	0.0	4.0	F
		B	706	GLN	1.4	-1.5	2.9	F
		B	729	GLN	-0.6	-1.0	0.4	f
3D4Q	31%	A	37	GLN	5.7	4.6	1.1	F
		A	42	GLN	0.6	-1.1	1.7	F
		A	74	GLN	1.7	0.2	1.5	F
		A	75	GLN	4.2	1.7	2.5	F
		A	162	ASN	6.1	5.4	0.7	f
		A	190	GLN	1.3	-1.8	3.1	F
		A	241	ASN	2.8	1.9	0.9	f
		A	265	ASN	7.2	5.5	1.7	F
		B	349	GLN	1.7	-0.4	2.1	F
		B	382	GLN	6.1	3.2	2.9	F
		B	450	GLN	1.7	1.3	0.4	f
		B	516	GLN	2.3	-0.5	2.8	F
		B	519	ASN	4.2	3.2	1.0	F
		B	541	GLN	1.7	1.1	0.6	f
		B	546	ASN	7.1	0.7	6.4	F
		B	572	ASN	9.1	7.1	2.0	F
		B	597	GLN	2.9	-0.1	3.0	F
1BCD	19%	A	66	ASN	2.5	-0.6	3.1	F
		A	135	GLN	2.0	-0.1	2.1	F
		A	228	ASN	1.3	0.7	0.6	f
		A	251	ASN	2.8	-0.5	3.3	F
1H00	17%	A	6	GLN	0.8	-0.5	1.3	F
		A	63	ASN	0.0	-0.7	0.7	f
		A	75	ASN	16.7	8.0	8.7	F
3BWM	13%	A	41	ASN	-0.6	-1.4	0.8	f
		A	100	GLN	1.8	-1.4	3.2	F

1R9O	10%	A	200	ASN	3.2	-0.6	3.8	F
		A	339	GLN	-0.7	-1.1	0.4	f
		A	381	ASN	1.8	1.2	0.6	f
		A	467	GLN	2.0	0.3	1.7	F
3NXU	21%	A	243	GLN	3.7	1.7	2.0	F
		A	330	GLN	1.2	-0.9	2.1	F
		A	362	ASN	3.9	2.3	1.6	F
		A	404	ASN	1.6	-0.6	2.2	F
		A	429	ASN	1.8	-0.8	2.6	F
		A	450	GLN	1.7	-1.1	2.8	F
		B	660	ASN	2.3	-0.4	2.7	F
		B	728	GLN	1.7	-1.5	3.2	F
		B	742	GLN	3.4	0.8	2.6	F
		B	792	GLN	1.6	-1.5	3.1	F
		B	815	GLN	1.5	-0.8	2.3	F
		B	847	ASN	3.1	1.4	1.7	F
		B	914	ASN	1.3	-0.2	1.5	F
		B	935	GLN	5.5	4.6	0.9	f
3KRJ	20%	A	8	GLN	1.7	1.0	0.7	f
		A	108	GLN	2.0	-0.4	2.4	F
		A	177	GLN	1.0	-0.2	1.2	F
		A	201	ASN	0.7	-0.7	1.4	F
		A	248	GLN	0.7	0.0	0.7	f
		A	267	ASN	29.0	24.3	4.7	F
		A	328	GLN	1.0	-0.8	1.8	F
3ODU	18%	A	45	ASN	14.3	5.1	9.2	F
		A	129	ASN	3.2	2.0	1.2	F
		A	155	GLN	0.7	-0.9	1.6	F
		A	212	GLN	0.6	-0.8	1.4	F
		A	446	GLN	1.6	-0.6	2.2	F
		A	452	ASN	1.6	0.7	0.9	f
		B	523	ASN	2.9	-0.1	3.0	F
		B	549	ASN	0.8	-0.9	1.7	F
		B	578	GLN	1.4	-0.7	2.1	F
		B	655	ASN	3.7	0.6	3.1	F
		B	714	GLN	0.7	-0.6	1.3	F
		B	882	ASN	0.9	0.1	0.8	f
		B	883	GLN	1.3	-1.3	2.6	F
		B	948	GLN	2.6	-0.2	2.8	F
1LRU	11%	A	55	GLN	1.5	-1.3	2.8	F
		A	96	GLN	1.4	1.1	0.3	f
		C	391	GLN	0.5	-0.9	1.4	F
		C	488	GLN	1.6	-1.8	3.4	F

3FRJ	31%	A	10	GLN	17.6	13.9	3.7	F
		A	15	GLN	15.5	14.0	1.5	F
		A	27	GLN	0.6	-1.3	1.9	F
		A	121	ASN	1.4	-0.2	1.6	F
		A	154	GLN	1.7	-1.0	2.7	F
		A	201	ASN	2.3	0.0	2.3	F
		A	264	ASN	6.7	-1.0	7.7	F
		A	285	ASN	6.8	0.2	6.6	F
		B	294	GLN	2.5	-1.3	3.8	F
		B	352	GLN	2.2	0.0	2.2	F
		B	399	ASN	2.0	0.6	1.4	F
		B	440	GLN	1.4	0.1	1.3	F
		B	487	ASN	5.8	1.5	4.3	F
		B	514	GLN	0.4	-0.3	0.7	f
		B	550	ASN	8.1	-1.8	9.9	F
2I78	31%	A	51	ASN	2.5	-0.1	2.6	F
		A	72	GLN	2.7	-0.6	3.3	F
		A	74	ASN	3.6	0.7	2.9	F
		A	75	ASN	1.5	0.6	0.9	f
		A	103	ASN	2.4	0.0	2.4	F
		A	141	GLN	0.7	-1.0	1.7	F
		A	169	ASN	2.7	-1.3	4.0	F
		A	170	ASN	2.9	-0.4	3.3	F
		A	247	GLN	2.8	-1.0	3.8	F
		A	272	ASN	5.5	-1.9	7.4	F
		A	314	GLN	2.6	-0.6	3.2	F
		A	321	ASN	0.8	-0.2	1.0	f
		A	338	ASN	0.4	0.0	0.4	f
		A	369	ASN	4.2	-1.4	5.6	F
		A	435	GLN	0.3	0.0	0.3	f
		A	505	GLN	1.9	0.9	1.0	f
		A	520	ASN	1.6	0.7	0.9	f
		A	572	ASN	2.0	0.9	1.1	F
		A	586	GLN	0.0	-0.4	0.4	f
		A	595	ASN	1.7	0.9	0.8	f
		A	612	GLN	0.8	-1.2	2.0	F
		A	679	ASN	2.4	-1.8	4.2	F
		A	694	ASN	3.0	-0.6	3.6	F
		A	697	GLN	0.8	0.1	0.7	f
		A	718	GLN	3.2	-0.8	4.0	F
		A	761	GLN	4.7	1.2	3.5	F
		B	74	ASN	2.1	1.5	0.6	f
		B	92	ASN	3.4	0.3	3.1	F

B	103	ASN	2.3	0.0	2.3	F
B	169	ASN	3.2	0.0	3.2	F
B	170	ASN	3.2	-0.5	3.7	F
B	247	GLN	2.3	-1.3	3.6	F
B	272	ASN	7.6	-2.2	9.8	F
B	314	GLN	2.2	0.4	1.8	F
B	338	ASN	0.4	-0.9	1.3	F
B	344	GLN	-0.3	-1.4	1.1	F
B	369	ASN	2.3	-1.5	3.8	F
B	388	GLN	0.5	-1.0	1.5	F
B	435	GLN	1.1	0.3	0.8	f
B	505	GLN	1.8	0.9	0.9	f
B	506	ASN	1.0	-0.8	1.8	F
B	572	ASN	0.9	-0.9	1.8	F
B	586	GLN	0.3	-0.5	0.8	f
B	612	GLN	1.5	0.1	1.4	F
B	679	ASN	3.1	-1.8	4.9	F
B	694	ASN	2.6	-0.2	2.8	F
B	718	GLN	3.6	-0.4	4.0	F
B	731	GLN	0.0	-0.8	0.8	f
B	761	GLN	4.2	2.1	2.1	F
C	72	GLN	1.8	-0.5	2.3	F
C	74	ASN	1.6	-0.3	1.9	F
C	92	ASN	3.6	2.1	1.5	F
C	103	ASN	1.8	0.2	1.6	F
C	169	ASN	10.4	5.9	4.5	F
C	170	ASN	2.7	-0.5	3.2	F
C	247	GLN	3.7	-0.2	3.9	F
C	272	ASN	4.2	-0.2	4.4	F
C	314	GLN	3.0	-0.5	3.5	F
C	338	ASN	1.2	0.5	0.7	f
C	369	ASN	2.4	1.1	1.3	F
C	435	GLN	1.0	0.4	0.6	f
C	487	ASN	2.5	2.1	0.4	f
C	505	GLN	0.9	0.1	0.8	f
C	506	ASN	2.7	-0.8	3.5	F
C	572	ASN	1.2	-1.1	2.3	F
C	595	ASN	1.4	0.8	0.6	f
C	612	GLN	1.7	-0.5	2.2	F
C	679	ASN	2.6	-0.4	3.0	F
C	694	ASN	2.0	-0.9	2.9	F
C	718	GLN	4.2	-0.2	4.4	F
D	51	ASN	2.5	0.6	1.9	F

		D	103	ASN	2.1	0.4	1.7	F
		D	169	ASN	1.8	-1.3	3.1	F
		D	170	ASN	1.7	-1.3	3.0	F
		D	247	GLN	4.3	-0.3	4.6	F
		D	272	ASN	5.8	-1.8	7.6	F
		D	314	GLN	2.9	-0.9	3.8	F
		D	321	ASN	0.9	0.3	0.6	f
		D	369	ASN	1.6	-1.5	3.1	F
		D	435	GLN	1.4	-0.2	1.6	F
		D	487	ASN	3.3	2.8	0.5	f
		D	505	GLN	1.1	-0.2	1.3	F
		D	572	ASN	1.0	0.5	0.5	f
		D	595	ASN	3.2	0.3	2.9	F
		D	612	GLN	1.9	-0.6	2.5	F
		D	679	ASN	0.9	-1.2	2.1	F
		D	694	ASN	1.6	-0.9	2.5	F
		D	718	GLN	4.1	-0.2	4.3	F
3PBL	29%	A	15	GLN	3.5	0.7	2.8	F
		A	21	ASN	10.1	9.2	0.9	f
		A	34	GLN	1.8	0.6	1.2	F
		A	56	ASN	23.1	20.2	2.9	F
		A	148	GLN	1.4	-0.1	1.5	F
		A	153	GLN	2.4	-1.3	3.7	F
		A	226	GLN	1.4	1.1	0.3	f
		A	297	ASN	-0.2	-2.3	2.1	F
		A	369	ASN	1.2	-1.2	2.4	F
		A	373	ASN	1.5	-0.4	1.9	F
		A	428	GLN	8.8	2.9	5.9	F
		B	515	GLN	0.3	-1.1	1.4	F
		B	537	ASN	18.5	16.4	2.1	F
		B	551	GLN	2.5	-0.2	2.7	F
		B	587	ASN	-0.3	-0.7	0.4	f
		B	634	GLN	4.4	0.1	4.3	F
		B	750	ASN	5.1	1.9	3.2	F
		B	763	ASN	5.3	1.3	4.0	F
		B	765	ASN	6.7	1.7	5.0	F
		B	850	ASN	0.6	-2.0	2.6	F
		B	928	ASN	11.7	10.7	1.0	f
		B	940	ASN	6.8	2.3	4.5	F
3NXO	29%	A	5	ASN	2.3	1.2	1.1	F
		A	12	GLN	4.6	3.8	0.8	f
		A	35	GLN	1.7	-2.5	4.2	F
		A	107	ASN	0.1	-1.2	1.3	F

		A	185	ASN	5.6	-0.5	6.1	F
2RGP	12%	A	107	ASN	8.6	2.6	6.0	F
		A	275	GLN	4.5	1.4	3.1	F
1SJ0	17%	A	107	ASN	1.4	-1.4	2.8	F
		A	108	GLN	4.3	1.0	3.3	F
		A	226	ASN	3.8	-1.3	5.1	F
2FSZ	28%	A	11	GLN	2.5	-0.6	3.1	F
		A	162	GLN	1.7	-0.1	1.8	F
		A	201	ASN	3.5	-0.6	4.1	F
		A	227	ASN	5.1	0.9	4.2	F
		B	257	GLN	3.2	0.0	3.2	F
		B	447	ASN	1.3	-1.1	2.4	F
		B	460	ASN	1.2	0.8	0.4	f
		B	473	ASN	7.4	3.1	4.3	F
3KL6	31%	A	5	GLN	0.3	-0.9	1.2	F
		A	15	GLN	2.9	-0.4	3.3	F
		A	46	GLN	2.7	-1.3	4.0	F
		A	122	GLN	1.8	-0.5	2.3	F
		A	139	GLN	1.8	0.1	1.7	F
		A	154	ASN	1.4	-0.6	2.0	F
		A	166	GLN	0.8	0.5	0.3	f
1W7X	18%	A	88	ASN	2.6	-1.6	4.2	F
		A	149	ASN	4.4	-0.2	4.6	F
		A	160	GLN	2.8	-0.4	3.2	F
		A	161	GLN	0.0	-1.2	1.2	F
2NNQ	50%	A	15	ASN	0.3	-0.5	0.8	f
		A	45	ASN	6.5	0.1	6.4	F
		A	59	ASN	5.6	-1.2	6.8	F
3BZ3	28%	A	25	GLN	0.7	0.1	0.6	f
		A	45	ASN	1.3	-1.1	2.4	F
		A	57	GLN	0.6	-1.6	2.2	F
		A	64	GLN	2.1	0.7	1.4	F
		A	182	ASN	1.8	-2.0	3.8	F
		A	216	ASN	0.0	-1.7	1.7	F
3C4F	21%	A	80	ASN	0.2	-0.9	1.1	F
		A	127	ASN	3.6	2.1	1.5	F
		A	196	ASN	3.4	1.6	1.8	F
		B	382	ASN	-0.1	-0.8	0.7	f
		B	425	ASN	6.4	5.8	0.6	f
		B	498	ASN	4.7	2.4	2.3	F
		B	519	GLN	2.4	-0.8	3.2	F
		B	588	GLN	2.1	-1.0	3.1	F
		B	602	ASN	1.8	0.3	1.5	F

1J4H	33%	A	3	GLN	0.1	-0.5	0.6	f
		A	53	GLN	1.3	0.2	1.1	F
3E37	17%	A	21	GLN	1.3	-1.1	2.4	F
		A	32	GLN	3.2	-1.7	4.9	F
		A	108	GLN	-0.7	-2.3	1.6	F
		A	221	GLN	1.5	-1.5	3.0	F
		A	246	ASN	1.7	-1.9	3.6	F
		A	303	GLN	1.7	-0.1	1.8	F
		A	329	ASN	1.3	-1.0	2.3	F
		A	335	ASN	2.8	1.2	1.6	F
		A	364	GLN	0.8	-1.1	1.9	F
		A	379	GLN	2.0	-0.5	2.5	F
		B	415	GLN	0.2	-1.3	1.5	F
		B	435	GLN	1.7	0.2	1.5	F
		B	467	GLN	-0.2	-0.5	0.3	f
		B	513	GLN	2.6	0.7	1.9	F
		B	796	GLN	1.8	-1.4	3.2	F
1ZW5	16%	A	7	GLN	1.4	-1.0	2.4	F
		A	8	ASN	0.6	-0.1	0.7	f
		A	17	GLN	1.5	-1.2	2.7	F
		A	25	GLN	2.0	-0.3	2.3	F
		A	182	GLN	3.0	-1.4	4.4	F
		A	184	ASN	5.2	1.9	3.3	F
2V3F	14%	A	61	ASN	0.3	-0.3	0.6	f
		A	148	ASN	2.0	-1.5	3.5	F
		A	202	GLN	0.6	-1.1	1.7	F
		B	564	GLN	2.1	0.6	1.5	F
		B	566	ASN	0.6	-0.1	0.7	f
		B	650	GLN	0.3	-1.0	1.3	F
		B	653	ASN	1.3	-1.6	2.9	F
		B	707	GLN	0.7	-1.0	1.7	F
		B	733	GLN	1.0	-2.0	3.0	F
		B	947	GLN	1.0	-0.2	1.2	F
		B	1004	GLN	3.0	-0.6	3.6	F
3KGC	20%	A	3	ASN	2.3	0.7	1.6	F
		A	252	ASN	2.9	-1.1	4.0	F
		B	266	ASN	2.8	2.0	0.8	f
		B	465	GLN	7.2	1.5	5.7	F
		B	495	ASN	1.7	1.3	0.4	f
1VSO	11%	A	3	ASN	3.1	0.4	2.7	F
		A	45	ASN	0.5	-0.7	1.2	F
3MAX	16%	A	121	GLN	1.9	-2.0	3.9	F
		A	162	GLN	1.9	0.1	1.8	F

		A	247	GLN	0.2	-0.9	1.1	F
		A	343	ASN	4.5	-0.7	5.2	F
		B	488	GLN	1.8	-1.6	3.4	F
		B	600	GLN	3.0	-1.6	4.6	F
		B	614	GLN	0.5	-2.3	2.8	F
		C	772	ASN	1.9	1.4	0.5	f
		C	855	GLN	2.3	-1.6	3.9	F
		C	896	GLN	2.0	0.5	1.5	F
		C	967	GLN	2.6	-1.2	3.8	F
		C	981	GLN	0.6	-1.1	1.7	F
		C	1077	ASN	3.4	0.3	3.1	F
		C	1081	GLN	2.4	0.1	2.3	F
		C	1092	GLN	1.4	-1.2	2.6	F
3F07	24%	A	80	GLN	3.8	1.0	2.8	F
		A	136	ASN	2.1	1.7	0.4	f
		A	236	GLN	-0.3	-1.7	1.4	F
		A	253	GLN	1.9	-0.3	2.2	F
		A	256	ASN	3.4	0.1	3.3	F
		A	372	ASN	3.0	1.4	1.6	F
		B	468	GLN	3.8	1.3	2.5	F
		B	472	GLN	2.8	0.6	2.2	F
		B	524	ASN	1.8	1.5	0.3	f
		B	624	GLN	0.1	-1.5	1.6	F
		B	641	GLN	2.0	-0.3	2.3	F
		B	760	ASN	3.0	1.5	1.5	F
		C	856	GLN	3.6	1.0	2.6	F
		C	860	GLN	7.8	4.4	3.4	F
		C	1012	GLN	-0.3	-1.6	1.3	F
		C	1029	GLN	2.2	-0.2	2.4	F
		C	1032	ASN	3.6	0.1	3.5	F
		C	1148	ASN	2.9	1.5	1.4	F
3NF7	13%	A	66	GLN	1.8	-1.2	3.0	F
		A	108	GLN	3.8	0.8	3.0	F
		B	291	GLN	3.0	0.8	2.2	F
		B	300	GLN	3.2	-0.9	4.1	F
3LAN	30%	A	23	GLN	4.5	4.1	0.4	f
		A	91	GLN	4.9	2.2	2.7	F
		A	137	ASN	2.9	1.7	1.2	F
		A	145	GLN	1.5	-0.3	1.8	F
		A	255	ASN	1.8	0.5	1.3	F
		A	332	GLN	3.1	1.1	2.0	F
		A	348	ASN	1.8	-0.9	2.7	F
		A	407	GLN	4.8	3.4	1.4	F

		A	474	ASN	2.9	1.8	1.1	F
		A	500	GLN	2.9	1.4	1.5	F
		A	509	GLN	-0.2	-0.6	0.4	f
		A	512	GLN	0.7	-0.2	0.9	f
		B	645	GLN	2.9	-1.8	4.7	F
		B	734	GLN	1.8	-1.3	3.1	F
		B	735	ASN	5.3	3.2	2.1	F
		B	815	ASN	3.0	0.8	2.2	F
		B	818	GLN	2.3	-0.2	2.5	F
		B	825	ASN	1.3	0.7	0.6	f
		B	829	GLN	0.5	-1.0	1.5	F
		B	866	ASN	6.0	0.3	5.7	F
		B	894	GLN	2.7	-1.7	4.4	F
		B	896	GLN	0.9	0.2	0.7	f
		B	908	ASN	2.8	-0.7	3.5	F
		B	927	GLN	1.6	0.0	1.6	F
		B	954	GLN	2.2	-0.6	2.8	F
		B	967	GLN	0.5	-1.0	1.5	F
		B	978	ASN	-0.4	-1.0	0.6	f
		B	988	GLN	5.5	3.6	1.9	F
		B	1020	ASN	5.6	-0.6	6.2	F
		B	1047	GLN	0.4	-1.3	1.7	F
		B	1054	ASN	3.5	0.3	3.2	F
		B	1080	GLN	0.7	-1.0	1.7	F
		B	1084	GLN	4.2	-1.3	5.5	F
		B	1107	GLN	3.0	-1.7	4.7	F
3CCW	17%	A	16	GLN	1.3	-1.3	2.6	F
		A	76	GLN	0.8	-0.9	1.7	F
		A	84	ASN	3.2	1.0	2.2	F
		A	133	ASN	3.6	0.6	3.0	F
		A	198	GLN	2.3	-0.5	2.8	F
		A	354	ASN	1.6	0.1	1.5	F
		A	390	GLN	2.9	2.3	0.6	f
		B	479	ASN	1.9	-1.3	3.2	F
		B	517	GLN	1.2	-1.6	2.8	F
		B	536	ASN	0.1	-0.6	0.7	f
		B	639	GLN	2.2	-1.0	3.2	F
		B	877	ASN	4.0	0.4	3.6	F
		C	917	GLN	3.0	1.1	1.9	F
		C	920	ASN	1.7	-1.4	3.1	F
		C	945	GLN	2.0	-1.9	3.9	F
		C	1080	GLN	1.9	0.4	1.5	F
		C	1127	GLN	1.6	-0.1	1.7	F

		C	1272	GLN	3.1	0.4	2.7	F
		D	1334	ASN	4.0	-0.8	4.8	F
		D	1361	ASN	1.3	-1.3	2.6	F
		D	1386	GLN	1.3	-1.7	3.0	F
		D	1407	ASN	5.9	1.9	4.0	F
		D	1521	GLN	1.6	-0.2	1.8	F
		D	1531	ASN	5.3	0.4	4.9	F
		D	1677	ASN	1.3	-0.1	1.4	F
		D	1713	GLN	2.4	1.7	0.7	f
3F9M	17%	A	111	GLN	3.2	1.9	1.3	F
		A	171	ASN	1.1	0.4	0.7	f
		A	318	ASN	1.1	-0.2	1.3	F
		A	328	GLN	4.4	1.1	3.3	F
		A	342	GLN	1.1	-0.2	1.3	F
4TRJ	5%	A	48	GLN	0.8	-1.5	2.3	F
2ICA	16%	A	5	ASN	1.5	-1.1	2.6	F
		A	39	ASN	6.1	1.4	4.7	F
3G0E	50%	A	15	GLN	2.2	0.1	2.1	F
		A	23	ASN	2.0	1.4	0.6	f
		A	25	ASN	1.4	-1.6	3.0	F
		A	26	ASN	6.6	-0.3	6.9	F
		A	111	ASN	-0.2	-1.1	0.9	f
		A	139	ASN	1.5	-0.6	2.1	F
		A	198	ASN	5.4	-1.0	6.4	F
		A	220	ASN	1.9	-1.1	3.0	F
		A	223	ASN	1.4	-1.3	2.7	F
		A	320	GLN	4.8	-1.6	6.4	F
		A	328	GLN	1.2	-0.7	1.9	F
2B8T	18%	A	115	ASN	1.1	0.5	0.6	f
		A	154	ASN	2.3	0.4	1.9	F
		A	175	ASN	1.2	-1.3	2.5	F
		A	216	ASN	1.5	-0.6	2.1	F
		B	228	ASN	0.4	-1.3	1.7	F
		B	282	GLN	0.7	-0.6	1.3	F
		B	338	ASN	0.3	-0.2	0.5	f
		B	398	ASN	1.6	-1.0	2.6	F
		B	407	GLN	14.9	9.6	5.3	F
		C	561	ASN	0.4	-1.0	1.4	F
		C	621	ASN	1.1	-0.6	1.7	F
		D	779	ASN	4.6	-0.2	4.8	F
		D	784	ASN	1.7	-1.0	2.7	F
		D	844	ASN	0.4	-1.1	1.5	F
2OF2	27%	A	255	GLN	1.5	-1.5	3.0	F

		A	265	ASN	2.8	-1.2	4.0	F
		A	277	GLN	2.1	0.0	2.1	F
		A	339	ASN	5.0	3.0	2.0	F
		A	413	ASN	2.3	-0.9	3.2	F
3CHP	38%	A	45	GLN	1.0	-0.4	1.4	F
		A	69	GLN	2.1	0.8	1.3	F
		A	136	GLN	2.2	0.4	1.8	F
		A	226	GLN	0.8	-0.9	1.7	F
		A	272	ASN	5.5	3.5	2.0	F
		A	341	ASN	1.8	-1.2	3.0	F
		A	350	GLN	1.3	-0.9	2.2	F
		A	440	ASN	1.4	-0.7	2.1	F
		A	441	GLN	2.1	-0.2	2.3	F
		A	445	ASN	1.8	0.8	1.0	f
		A	466	ASN	1.7	-0.5	2.2	F
		A	484	ASN	2.9	0.6	2.3	F
		A	521	GLN	5.0	2.3	2.7	F
		A	525	ASN	2.4	1.1	1.3	F
		A	527	ASN	0.9	-0.9	1.8	F
		A	544	GLN	2.0	-2.3	4.3	F
		A	561	GLN	6.3	2.2	4.1	F
		A	589	GLN	-0.3	-0.9	0.6	f
3M2W	37%	A	9	GLN	1.6	-0.1	1.7	F
		A	24	GLN	15.0	7.5	7.5	F
		A	31	ASN	0.6	0.0	0.6	f
		A	36	GLN	22.0	14.7	7.3	F
		A	43	GLN	4.9	-1.6	6.5	F
		A	69	GLN	6.5	-0.6	7.1	F
		A	156	ASN	8.0	1.3	6.7	F
		A	201	ASN	3.2	0.5	2.7	F
		A	268	GLN	0.0	-0.8	0.8	f
2AA2	55%	A	11	ASN	1.0	-0.7	1.7	F
		A	55	ASN	2.5	-0.9	3.4	F
		A	70	GLN	1.1	-0.6	1.7	F
		A	89	GLN	4.3	0.7	3.6	F
		A	114	ASN	-0.1	-0.6	0.5	f
		A	116	GLN	1.8	-1.3	3.1	F
		A	133	GLN	1.4	0.6	0.8	f
		A	145	GLN	1.3	-0.1	1.4	F
		A	149	GLN	0.9	-1.0	1.9	F
		A	154	GLN	1.4	0.3	1.1	F
		A	189	ASN	1.4	-1.2	2.6	F
		A	207	GLN	2.8	1.1	1.7	F

		A	210	GLN	8.0	5.6	2.4	F
		A	214	GLN	1.0	-0.4	1.4	F
		A	258	GLN	0.0	-1.7	1.7	F
3LQ8	21%	A	75	GLN	1.3	0.6	0.7	f
		A	90	ASN	-0.7	-1.1	0.4	f
		A	127	ASN	1.9	1.0	0.9	f
		A	191	ASN	5.4	4.9	0.5	f
3EQH	23%	A	25	GLN	2.6	-0.9	3.5	F
		A	45	ASN	7.7	7.2	0.5	f
		A	89	ASN	2.6	1.4	1.2	F
		A	181	GLN	3.0	2.1	0.9	f
		A	245	GLN	12.2	9.6	2.6	F
		A	349	ASN	6.4	0.0	6.4	F
1QW6	33%	A	353	GLN	5.5	2.2	3.3	F
		A	364	GLN	0.7	-0.9	1.6	F
		A	451	ASN	1.7	-0.7	2.4	F
		A	478	GLN	-0.4	-0.8	0.4	f
		A	498	ASN	4.2	1.2	3.0	F
		A	507	GLN	1.2	0.3	0.9	f
		A	508	GLN	1.9	1.2	0.7	f
		A	569	ASN	2.5	1.6	0.9	f
		A	628	GLN	3.6	-1.0	4.6	F
		A	634	ASN	4.5	1.6	2.9	F
		A	707	GLN	3.4	-0.2	3.6	F
		A	712	ASN	2.2	-0.9	3.1	F
1B9V	38%	A	88	GLN	1.9	1.4	0.5	f
		A	93	GLN	2.2	0.7	1.5	F
		A	144	ASN	4.3	0.4	3.9	F
		A	169	ASN	2.2	-0.7	2.9	F
		A	220	ASN	4.5	2.7	1.8	F
		A	340	ASN	3.5	0.5	3.0	F
		A	373	ASN	7.3	3.1	4.2	F

Table 3.3: Incorrectly assigned ASN & GLN sidechain rotamers. Flipped residues are listed by PDB code, chain identifier and sequence ID. Percentages of ASN and GLN residues recommended to flip are listed for each file. Sidechains with Δz (z(PDB)-z(flipped)) > 1 are marked with the flip indicator "F" and sidechains with Δz between 0.3 and 1 "f".

2.5 pK_a Calculations

Tools available for protein pK_a prediction essentially can be divided into two different groups: software using the Poisson–Boltzmann equation and empirical methods. Approaches using the Poisson–Boltzmann equation (e.g. H++) usually require a higher computational effort. Thus calculations can easily take several minutes to hours depending on the size of the protein. Empirical methods (e.g. PROPKA) on the other hand may be less accurate in some cases, but can perform pK_a calculations within a few seconds, making them a good choice when analyzing greater amounts of data. Preparation of PDB files as an initial step for further experiments should be kept as simple and straight forward as possible. Keeping that in mind, the PROPKA 3.1 software was used to calculate pK_a values for all PDB files that managed passing the previous step of rotamer analysis and correction. Residues with protonation states other than expected at pH levels of 7.4 were extracted from PROPKA reports and are listed in Table 3.4.

PDB code	Chain ID	SSEQI	Residue type	pK_a	$pK_{a\ int}$
2P54	A	210	HIS	7.5	6.5
1L2S	A	64	LYS	6.29	10.5
	B	422	LYS	6.23	10.5
	B	541	HIS	7.48	6.5
2HZI	A	58	GLU	10.42	4.5
	A	158	ASP	8.83	3.8
	B	435	ASP	8.86	3.8
	B	544	HIS	7.69	6.5
2GTK	A	161	LYS	7.39	10.5
2I78	A	206	GLU	11.61	4.5
	A	230	ASP	11.48	3.8
	A	258	LYS	6.03	10.5
	A	740	HIS	8.04	6.5
	B	162	HIS	7.71	6.5
	B	230	ASP	12.09	3.8
	B	258	LYS	6.19	10.5
	B	663	ASP	10.69	3.8
	B	709	ASP	8.16	3.8
	B	754	HIS	7.48	6.5
	C	230	ASP	11.73	3.8

	C	258	LYS	5.99	10.5
	C	663	ASP	11.51	3.8
	C	709	ASP	7.5	3.8
	C	740	HIS	7.93	6.5
	D	230	ASP	12.12	3.8
	D	258	LYS	6.01	10.5
	D	663	ASP	11.67	3.8
	D	740	HIS	7.9	6.5
1QW6	A	326	CYS	7.19	9
	A	592	GLU	9.56	4.5
	A	597	ASP	8.49	3.8
3MAX	A	134	HIS	9.3	6.5
	A	172	HIS	9.92	6.5
	A	173	HIS	7.88	6.5
	A	275	HIS	8.1	6.5
	A	355	LYS	7.04	10.5
	B	370	LYS	5.19	10.5
	B	418	HIS	7.53	6.5
	B	501	HIS	9.53	6.5
	B	539	HIS	10.32	6.5
	B	540	HIS	7.75	6.5
	B	676	GLU	9.4	4.5
	C	868	HIS	9.3	6.5
	C	906	HIS	10.01	6.5
	C	907	HIS	7.91	6.5
3CCW	B	695	CYS	7.27	9
	B	707	GLU	9.9	4.5
3D4Q	A	14	GLU	10.28	4.5
	A	15	ASP	11.65	3.8
	A	18	ARG	6.04	12.5
	A	20	LYS	6.44	10.5
	A	29	ASP	11.22	3.8
	A	157	ASP	13.49	3.8
	A	172	LYS	7.38	10.5
	A	219	ASP	11.41	3.8
	B	321	GLU	13.17	4.5
	B	322	ASP	9.22	3.8
	B	327	LYS	5.42	10.5
	B	336	ASP	7.97	3.8
	B	397	ARG	5.53	12.5
1LRU	A	132	HIS	11.23	6.5
	B	300	HIS	10.36	6.5
	C	426	CYS	4.86	9
	C	468	HIS	11.01	6.5
3FRJ	A	181	LYS	7.29	10.5

	B	288	LYS	6.4	10.5
	B	467	LYS	7.27	10.5
	B	567	ASP	11.85	3.8
2B8T	A	148	LYS	7.28	10.5
	A	153	CYS	3.43	9
	B	274	ASP	8.39	3.8
	B	349	LYS	7.19	10.5
	B	376	CYS	3.91	9
	C	594	LYS	7.38	10.5
	C	599	CYS	3.4	9
	D	817	LYS	7.09	10.5
1S3B	A	271	LYS	6.99	10.5
	A	347	HIS	8.66	6.5
	A	437	GLU	7.63	4.5
	B	670	GLU	8.7	4.5
	B	791	LYS	6.76	10.5
	B	867	HIS	8.68	6.5
	B	957	GLU	7.61	4.5
3PBL	A	4	ASP	8.78	3.8
	A	84	ASP	12.59	3.8
	B	565	ASP	12.1	3.8
2VT4	A	57	ASP	10.09	3.8
	A	100	GLU	9.42	4.5
	A	212	ASP	7.73	3.8
	A	216	LYS	5.39	10.5
	B	370	ASP	8.99	3.8
	B	413	GLU	8.27	4.5
	B	529	LYS	3.28	10.5
	B	531	LYS	6.66	10.5
	C	726	GLU	8.2	4.5
	C	842	LYS	5.89	10.5
	C	844	LYS	6.93	10.5
	D	996	ASP	8.02	3.8
3C4F	A	68	GLU	7.62	4.5
3E37	A	320	GLU	8.41	4.5
	B	664	GLU	7.69	4.5
	B	673	LYS	6.51	10.5
	B	741	HIS	12.89	6.5
3ODU	A	20	ASP	14.75	3.8
	A	32	ASP	12.53	3.8
	A	36	GLU	8.6	4.5
	A	107	ASP	9.03	3.8
	A	197	ASP	7.72	3.8
	A	462	GLU	8.47	4.5
	B	773	HIS	7.41	6.5

	B	957	HIS	7.78	6.5
	B	964	GLU	7.51	4.5
1B9V	A	226	GLU	8.34	4.5
	A	349	LYS	7.16	10.5
3LAN	A	443	ASP	8.77	3.8
	B	746	ASP	8.27	3.8
	B	823	LYS	7.07	10.5
3NXU	A	83	ARG	7.28	12.5
	A	298	GLU	7.75	4.5
	B	787	HIS	7.89	6.5
3NF7	A	17	ARG	5.5	12.5
	A	87	ASP	10.65	3.8
	B	200	ARG	5.69	12.5
	B	270	ASP	9.98	3.8
2V3F	A	235	GLU	11.38	4.5
	A	313	HIS	10.17	6.5
	A	382	ASP	8.17	3.8
	B	740	GLU	11.32	4.5
	B	818	HIS	10.27	6.5
	B	887	ASP	8.22	3.8
3F07	A	142	HIS	10.07	6.5
	A	180	HIS	9.33	6.5
	A	181	HIS	8.35	6.5
	A	216	ASP	8.15	3.8
	A	267	ASP	7.46	3.8
	B	478	HIS	7.49	6.5
	B	530	HIS	10.15	6.5
	B	568	HIS	9.29	6.5
	B	569	HIS	8.63	6.5
	B	604	ASP	8.54	3.8
	B	655	ASP	7.65	3.8
	C	871	GLU	10.48	4.5
	C	877	ASP	9.37	3.8
	C	882	GLU	11.17	4.5
	C	918	HIS	8.95	6.5
	C	956	HIS	10.36	6.5
	C	957	HIS	8.45	6.5
	C	992	ASP	8.41	3.8
3BWM	A	144	LYS	5.08	10.5
3NXO	A	30	GLU	8.51	4.5

Table 3.4: Residues of PDB files with unexpected pK_a values calculated by PROPKA. The table lists ionizable residues with protonation states different to their protonation states without any interactions at pH 7.4.

For 32% of the 72 structures inspected, PROPKA calculates pK_a shifts far enough from $pK_{a\ int}$ values of residues to result in a change of normally assumed protonation states at physiological pH levels. Figure 3.6 illustrates the relative occurrence of ionizable amino acids with unusual pK_a values of these PDB files.

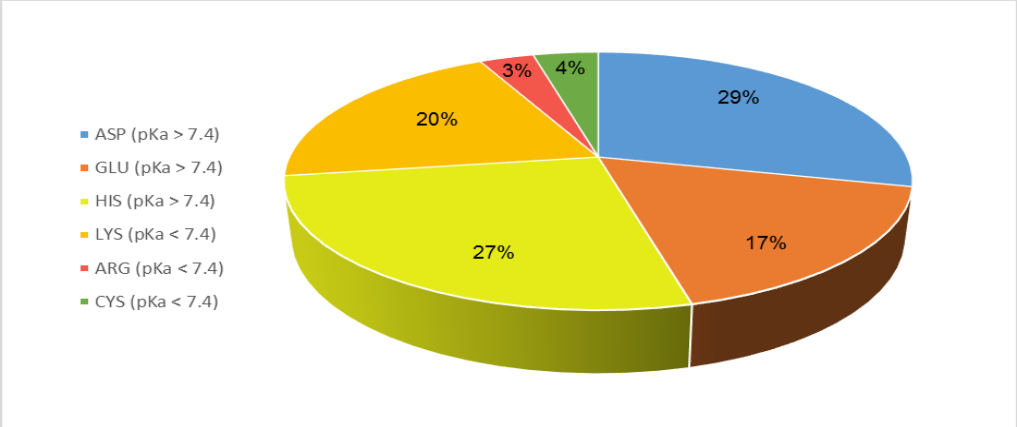

Figure 3.6: Distribution of residues with unexpected pK_a values calculated by PROPKA.

The distribution found suggests that there is similar likelihood for ASP, GLU, HIS and LYS residues to show a change in their titration states, whereas CYS or ARG residues are less effected. The relatively low amount of ARG and CYS residues found could be a result of ARG being by far the most basic amino acid and CYS being protonated even at higher pH levels, so that burial won't result in changes of their protonation states under physiological conditions. A smaller subset of 10 randomly chosen PDB structures containing residues with discussed properties were uploaded to the H++ web service for a comparison of results obtained by different approaches for pK_a calculations. Results for these PDB files are listed in Table 3.6. Although the sum of residues with unusual pK_a values is nearly the same over the 10 inspected structures (PROPKA: 104; H++: 101), only 38 (37%) of the residues found with PROPKA were identical to the residues found by H++. Given the differences in results it is not possible to make a general statement about the correctness of these calculations. Nevertheless, considering that over on third of all PDB structures contain amino acids with protonation states other than one would usually suspect or assign, calculating pK_a values for ionizable residues should be considered an integral part of PDB files preparation as electrostatic interactions often determine important properties of biomolecules.

PDB code	Chain ID	SSEQI	Residue Type	pK_a value	$pK_{a\ int}$
1L2S	A	147	TYR	6.679	14.075
	A	207	HIS	7.819	6.434
	B	505	TYR	5.091	13.204
	B	541	HIS	8.5	5.415
	B	565	HIS	7.749	6.386
1S3B	A	178	HIS	7.465	3.85
	A	252	HIS	9.041	6.568
	A	347	HIS	9.547	6.05
	A	397	CYS	5.324	9.051
	B	610	HIS	7.856	5.541
	B	772	HIS	8.764	6.544
	B	867	HIS	8.986	5.933
	B	917	CYS	5.849	9.484
2I78	A	162	HIS	8.896	5.741
	A	230	ASP	10.805	4.584
	A	663	ASP	9.25	4.515
	A	740	HIS	10.582	3.425
	A	754	HIS	7.722	4.719
	B	162	HIS	8.825	5.184
	B	230	ASP	11.264	4.842
	B	363	HIS	8.82	6.628
	B	663	ASP	>12.000	5.256
	B	740	HIS	11.517	4.17
	B	754	HIS	9.583	5.216
	C	162	HIS	8.908	6.009
	C	200	ASP	11.338	3.335
	C	206	GLU	9.539	5.425
	C	230	ASP	8.127	4.197
	C	363	HIS	7.431	5.699
	C	740	HIS	9.464	3.524
	C	754	HIS	8.397	4.935
	D	162	HIS	7.49	5.552
	D	200	ASP	>12.000	3.831
	D	206	GLU	>12.000	5.654
	D	740	HIS	9.231	3.638
	D	754	HIS	9.012	5.26
2P54	A	41	HIS	7.507	5.592
	A	195	HIS	9.234	8.575
	A	215	HIS	9.222	5.306
	A	267	TYR	3.394	3.988
2VT4	A	57	ASP	7.894	8.595
	A	216	LYS	0.735	4.325
	A	227	LYS	6.317	8.974

	A	291	LYS	6.087	9.561
	B	537	ARG	6.652	8.378
	C	844	LYS	7.175	6.491
	C	853	LYS	6.53	9.011
	C	909	TYR	5.392	13.212
	D	1157	LYS	5.936	5.606
	D	1230	LYS	6.386	9.18
3D4Q	A	157	ASP	5.799	8.923
	B	327	LYS	7.46	6.711
	B	578	LYS	8.011	6.878
3F07	A	90	HIS	10.523	4.527
	A	142	HIS	>12.000	3.421
	A	180	HIS	10.643	4.653
	A	186	GLU	9.745	6.462
	A	201	HIS	9.794	6.419
	B	530	HIS	>12.000	3.58
	B	568	HIS	11.328	4.796
	B	574	GLU	10.097	6.857
	B	589	HIS	10.294	6.585
	C	861	GLU	7.442	4.835
	C	866	HIS	8.981	5.879
	C	877	ASP	10.99	5.276
	C	882	GLU	>12.000	8.192
	C	918	HIS	>12.000	3.396
	C	956	HIS	9.85	4.689
	C	962	GLU	10.926	6.544
	C	977	HIS	10.155	6.272
3MAX	A	172	HIS	10.205	4.143
	A	173	HIS	10.713	3.479
	A	275	HIS	>12.000	4.268
	B	369	LYS	6.075	8.375
	B	539	HIS	10.747	4.31
	B	540	HIS	9.37	3.102
	B	642	HIS	>12.000	3.628
	C	868	HIS	>12.000	2.917
	C	906	HIS	10.715	4.311
	C	907	HIS	10.689	3.428
	C	1009	HIS	>12.000	4.557
2B8T	A	145	LYS	7.112	10.427
	A	191	CYS	<0.000	10.122
	B	376	CYS	5.355	7.556
	C	637	CYS	<0.000	9.623
	D	814	LYS	7.34	10.476
	D	822	CYS	6.671	7.397
	D	852	CYS	7.351	8.525

3ODU	A	20	ASP	10.847	7.684
	A	32	ASP	>12.000	9.011
	A	36	GLU	7.85	6.272
	A	123	HIS	10.297	5.54
	A	271	HIS	8.901	6.353
	A	408	LYS	7.286	9.435
	A	436	ASP	7.767	6.356
	A	468	HIS	9.878	5.61
	B	625	HIS	8.347	5.974
	B	773	HIS	8.458	6.235
	B	910	LYS	4.425	7.692
	B	957	HIS	10.896	6.235
	B	970	HIS	7.852	4.801

Table 3.5: Residues of PDB files with unexpected pKa values calculated by H++. Amino acids marked green were also found in the analysis carried out by PROPKA.

2.6 Applying Charges to Ligands

For the final step PROPKA reports of all previously analyzed PDB files were searched for ligands carrying titratable compounds. Results for the analysis are shown in Table 3.6. Among the 72 inspected structures, 29 contain a ligand with at least one ionizable compound. Since not every atom capable of carrying a charge actually is charged at physiological pH levels, PROPKAs model pK_a values were compared to the calculated pK_a values. It was found that 22 (31%) of the investigated PDB files include ligands with at least one functional group ionized at pH 7.4. The remaining 6 structures hold aromatic nitrogen, which under normal circumstances is considered neutral by the software.

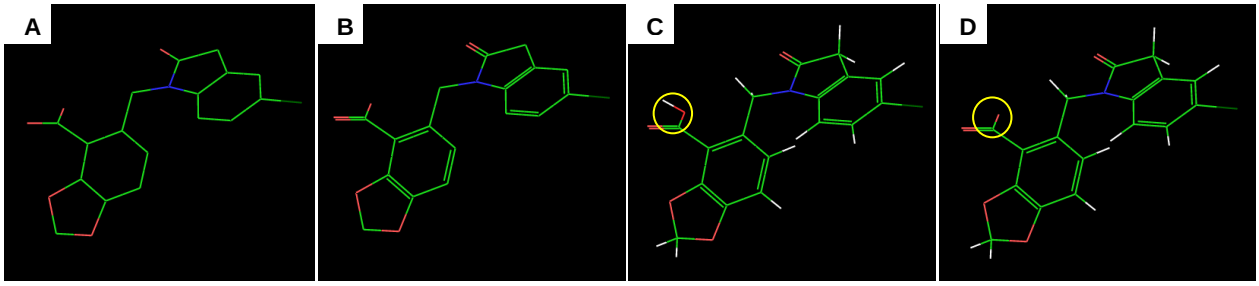

Figure 3.6: Steps of ligand preparation for MD input. The figure shows the ligand of HIV integrase (PDB code "3NF7") during different steps of assigning hydrogens and the correct protonation state. (A) Ligand as extracted from PDB file (B) Ligand with bond assigned bond orders (C) Ligand structure when assigning hydrogens without considering electrostatics (D) Ligand with added hydrogens after assigning correct formal charges.

With over 30% of all PDB files containing charged ligands, it becomes clear that correct assignment of protonation states prior to simulation is a crucial step for achieving reliable results of interactions within binding pockets. Thus special attention has to be paid when generating topology- and parameter files of ligands for MD simulations. Structure Data Files (SDF) files of ligands downloaded from RCSB contain information about atoms, coordinates and bonds of ligands, but lack data for hydrogen atoms. Hydrogens can be added automatically when creating topology- and parameter files, but are allo-

cated based on bond orders provided in SDF files. Electrostatics are not considered, though. Functional groups typically charged at physiological pH values, such as acids (e.g. carboxylic acids) or bases (e.g. amines), appear in their neutral states. In order to build ligand files with accurate protonation states the MAESTRO software suite was used to extract ligands from PDB files, assign bond orders, formal charges and finally add hydrogen atoms. In general it is also possible to skip the first step of extraction and edit formal charges and hydrogens from downloaded SDF files. Extraction from original models has the immediate advantage of retaining the coordinates of the original ligand model. In three cases (3BQD, 1H00, 2RGP) PROPKA calculated pK_a values for amines slightly beneath 7.4. However it was decided to also apply charges to these atoms, as values near the actual pH would mean an approximate 50:50 distribution of protonated and deprotonated states. In summary it can be said, that many ligand structures need treatment before using them for any kind of simulation. Not taking into account charges could have a large effect on calculated binding energies and thus could bias results of any experiment investigating interactions between ligands and binding sites.

PDB code	ID	Number	Compound	Position	pK_a	Model pK_a
2OJ9	BMI	1	Amine	N	7.47	10
		1	Aromatic Nitrogen	N1	-2.06	5
		1	Aromatic Nitrogen	N2	-0.16	5
		1	Aromatic Nitrogen	N4	4.32	5
		1	Aromatic Nitrogen	N6	2.11	5
1L2S	STC	1	Amine	N1	9.11	10
		1	Carboxylic acid	C21	0.04	4.5
		2	Amine	N1	9.14	10
		2	Carboxylic acid	C21	-0.2	4.5
		3	Amine	N1	9.65	10
		3	Carboxylic acid	C21	4.6	4.5
2HZI	JIN	1	Amine	N08	8.34	10
		1	Aromatic Nitrogen	N03	2.02	5
		2	Amine	N08	8.45	10
		2	Aromatic Nitrogen	N03	2.07	5
3BQD	DAY	1	Amine	N2	6.75	10
2I78	KIQ	1	Amine	N	12.15	10
		1	Aromatic Nitrogen	N14	3.52	5

1QW6	HEM	1	Aromatic Nitrogen	NA	-1.62	5
		1	Aromatic Nitrogen	NB	2.98	5
		1	Aromatic Nitrogen	NC	-3.2	5
		1	Aromatic Nitrogen	ND	-3.14	5
		1	Carboxylic acid	CGA	2.56	4.5
		1	Carboxylic acid	CGD	6.42	4.5
1BCD	FMS	1	Amine	N	8.35	10
3BZ3	YAM	1	Amine	N10	7.91	10
		1	Amine	N9	9.15	10
		1	Aromatic Nitrogen	N1	2.24	5
		1	Aromatic Nitrogen	N4	1.62	5
		1	Aromatic Nitrogen	N25	3.63	5
1H00	FAP	1	Amine	N7	8.64	10
		1	Amine	N16	7.27	10
		1	Amine	N27	9.65	10
		1	Aromatic Nitrogen	N6	-0.11	5
3BKL	KAW	1	Aromatic Nitrogen	NE1	2.98	5
		1	Carboxylic acid	C	-0.97	4.5
1ZW5	ZOL	1	Aromatic Nitrogen	N17	1.22	5
		1	Phosphonic acid	O16	17.35	6
		1	Phosphonic acid	O15	13.9	6
		1	Phosphonic acid	O10	4.15	6
		1	Phosphonic acid	O11	4.36	6
2HV5	ZST	1	Aromatic Nitrogen	N2	3.39	5
		1	Aromatic Nitrogen	N3	0.94	5
		1	Carboxylic acid	C18	-0.66	4.5
3NY8	JRZ	1	Amine	N1	9.86	10
2B8T	THM	1	Aromatic Nitrogen	N3	2.15	5
		2	Aromatic Nitrogen	N3	0.87	5
		3	Aromatic Nitrogen	N3	2.58	5
		4	Aromatic Nitrogen	N3	2.79	5
3EML	ZMA	1	Amine	N10	8.01	10
		1	Aromatic Nitrogen	N13	0.8	5
3PBL	ETQ	1	Amine	N2	10.3	10
		2	Amine	N2	10.04	10
1J4H	SUB	1	Amine	N2	8.64	10
3F9M	MRK	1	Aromatic Nitrogen	N4	2.86	5
		1	Aromatic Nitrogen	N19	3.59	5
3KRJ	KRJ	1	Amine	N26	9.99	10
		1	Aromatic Nitrogen	N05	2.43	5
3ODU	ITD	1	Aromatic Nitrogen	N1	-0.1	5
		2	Aromatic Nitrogen	N1	0.99	5
3L5D	BDV	1	Aromatic Nitrogen	N1	3.64	5

		2	Aromatic Nitrogen	N1	3.69	5
3LAN	KBT	1	Aromatic Nitrogen	N24	2.51	5
3NXU	RIT	1	Aromatic Nitrogen	N5	2.84	5
		1	Aromatic Nitrogen	N83	2.08	5
		2	Aromatic Nitrogen	N5	2.88	5
		2	Aromatic Nitrogen	N83	2.27	5
2FSZ	OHT	1	Amine	N24	10.36	10
		2	Amine	N24	9.3	10
		3	Amine	N24	10.33	10
		4	Amine	N24	9.29	10
3NF7	CIW	1	Carboxylic acid	C18	4.89	4.5
		2	Carboxylic acid	C18	6.08	4.5
1R9O	FLP	1	Carboxylic acid	C14	3.74	4.5
2RGP	HYZ	1	Amine	N3	7.35	10
		1	Amine	N8	9.26	10
		1	Aromatic Nitrogen	N2	-0.12	5
		1	Aromatic Nitrogen	N4	-0.18	5
3BWM	SAM	1	Amine	N	13.01	10
		1	Aromatic Nitrogen	N3	3.36	5
		1	Carboxylic acid	C	5.12	4.5
3NXO	D2B	1	Aromatic Nitrogen	N1	1.54	5

Table 3.6: Ligands of PDB structures with ionizable functional groups.

Chapter 4

Conclusion

The conducted experiments show that errors within models of macromolecular structures are far more common than one might expect. Considering the observations made throughout the experiments in this thesis, it becomes clear that the majority of PDB files profit from treatment prior to their use in further computational studies. Fortunately, a range of software packages and web services, capable of detecting and correcting these errors, are available. Based on the results obtained by analyzing a test set of PDB structures, a workflow was created to guide users through the process of choosing models of good quality and preparing them for intended experiments. This workflow deals with the most common errors found in PDB structures and includes six steps dealing with problems to which special attention should be paid:

1. Step - Validation of input structures

For reliable and unbiased results of experiments, selection of good models is crucial. Therefore it was found useful to check selected structures for gross errors and overall model quality. The WHAT IF webserver offers a great way of getting an overall impression on PDB files by calculating a variety of validation parameters and detecting outliers with deviations from expected values. Although detected errors do not necessarily indicate bad model quality, this step can prevent users from working with potentially poor models.

2. Step - Evaluation of binding sites and ligands

For the evaluation of the fit to the experimentally determined electron density the software VHELIBS is used. This program analyses binding sites and ligands by different adjustable parameters and divides them into the categories “Good”, “Dubious” and “Bad”. Evaluation of 50 PDB files showed that 12% of the investigated binding sites and

6% of the ligands were considered to have a “Bad” fit to their electron density. Values nearly doubled when applying three additional parameters to the default settings used for PDB files by the software. This underlies the importance of carefully analyzing structures used for simulations, especially when interactions between ligand and binding pocket are in the focus of interest. While a visual inspection of models containing “Dubious” elements is recommended, structures with “Bad” elements should be discarded immediately.

3. Step – Modelling missing coordinates

Analysis of the investigated PDB files show that 84% of all models contain at least one missing residue and in 36% of the structures sidechain atoms are missing. Missing residues, especially those located internally, need to be modelled in order to prevent unintended chain breaks during simulations or biased results. Missing atoms and terminal strings of missing residues, but should also be added to incomplete structures, despite being less severe problems. The program MODELLER is used to complete missing parts of inspected proteins by using ab initio prediction.

4. Step – Correcting ASN and GLN rotamers

Incorrectly assigned asparagine and glutamine rotamers are analyzed and corrected by the NQ-Flipper web service. This webserver calculates energy differences between the two corresponding sidechain rotamers. An evaluation of all examined structures showed, that on average 23% of the ASN and GLN residues within PDB files have unfavorable interaction energies. Rotamers with the flip indicator “F” are flipped.

5. Step – Calculating pK_a values of ionizable residues and ligands

The software PROPKA 3.1 is used to calculate pK_a values of each residue. Results of the calculations carried out for all PDB files within the test set, suggest that 31% of all ionizable residues have their pK_a values shifted so far from their intrinsic pK_a to result in protonation states other than usually expected for this amino acids at physiological pH values. These deviant protonation states need to be assigned before conducting a simulation using PDB structures.

6. Step – Applying formal charges to ligands

It was shown that 31% of the PDB files contain ligands with ionizable functional groups. Hydrogens automatically added when generating topology- and parameter files are allocated based on bond orders given in SDF files without taking electrostatics into account. As a consequence protonation states of ligands have to be assigned manually. Ligands with ionizable functional groups are detected by using the PROPKA pK_a reports. The MAESTRO software suite is used for assigning formal charges and adding the correct amount of hydrogen atoms to the ligand.

Bibliography

Acharya, K. Ravi, and Matthew D. Lloyd. 2005. “The Advantages and Limitations of Protein Crystal Structures.” *Trends in Pharmacological Sciences*. doi:10.1016/j.tips.2004.10.011.

Anandakrishnan, Ramu, Boris Aguilar, and Alexey V. Onufriev. 2012. “H++ 3.0: Automating pK Prediction and the Preparation of Biomolecular Structures for Atomistic Molecular Modeling and Simulations.” *Nucleic Acids Research* 40 (W1): 537–41. doi:10.1093/nar/gks375.

Barker, Nick, and Hans Clevers. 2000. “Quality Control in Databanks for Molecular Biology.” *BioEssays* 22 (11): 1024–34. doi:10.1002/1521-1878(200011)22:11<1024::AID-BIES9>3.0.CO;2-W.

Berman, Helen M. 2007. “The Protein Data Bank: A Historical Perspective.” *Acta Crystallographica Section A: Foundations of Crystallography* 64 (1). International Union of Crystallography: 88–95. doi:10.1107/S0108767307035623.

Berman, Helen M., John D. Westbrook, Zukang Feng, Gary Gilliland, T. N. Bhat, Helge Weissig, Ilya N. Shindyalov, and Philip E. Bourne. 2000. “The Protein Data Bank.” *Nucl. Acids Res.* 28 (1): 235–42. doi:10.1093/nar/28.1.235.

Brandt, Bernd W., Jaap Heringa, and Jack A M Leunissen. 2008. “SEQATOMS: A Web Tool for Identifying Missing Regions in PDB in Sequence Context.” *Nucleic Acids Research* 36 (Web Server issue): 255–59. doi:10.1093/nar/gkn237.

Brown, Eric N., and S. Ramaswamy. 2007. “Quality of Protein Crystal Structures.” *Acta Crystallographica Section D: Biological Crystallography* 63 (9). International Union of Crystallography: 941–50. doi:10.1107/S0907444907033847.

Brünger, A. T., P. D. Adams, G. M. Clore, W. L. DeLano, P. Gros, R. W. Grosse-Kunstleve, J. S. Jiang, et al. 1998. “Crystallography & NMR System: A New Software Suite for Macromolecular Structure Determination.” *Acta Crystallographica Section D Biological Crystallography* 54 (5): 905–21. doi:10.1107/S0907444998003254.

Brünger, Axel T. 1992. “Free R Value: A Novel Statistical Quantity for Assessing the Accuracy of Crystal Structures.” *Nature* 355 (6359): 472–75. doi:10.1038/355472a0.

Cereto-Massagué, Adrià, María José Ojeda, Robbie P. Joosten, Cristina Valls, Miquel Mulero, M. Josepa Salvado, Anna Arola-Arnal, Lluís Arola, Santiago Garcia-Vallvé, and Gerard Pujadas. 2013. “The Good, the Bad and the Dubious: VHELIBS, a Validation Helper for Ligands and Binding Sites.” *Journal of Cheminformatics* 5 (7): 1–9. doi:10.1186/1758-2946-5-36.

Cooper, DR. 2012. “NIH Public Access.” *Expert Opin Drug Discov* 6 (8): 771–82. doi:10.1517/17460441.2011.585154.

Dauter, Zbigniew, Alexander Wlodawer, Wladek Minor, Mariusz Jaskolski, and Bernhard Rupp. 2014. “Avoidable Errors in Deposited Macromolecular Structures: An Impediment to Efficient Data Mining.” *IUCrJ* 1. International Union of

Crystallography: 179–93. doi:10.1107/S2052252514005442.

Davis, Andrew M., Stephen A. St-Gallay, and Gerard J. Kleywegt. 2008. “Limitations and Lessons in the Use of X-Ray Structural Information in Drug Design.” *Drug Discovery Today* 13 (19–20): 831–41. doi:10.1016/j.drudis.2008.06.006.

Davis, Andrew M., Simon J. Teague, and Gerard J. Kleywegt. 2003. “Application and Limitations of X-Ray Crystallographic Data in Structure-Based Ligand and Drug Design.” *Angewandte Chemie - International Edition* 42 (24): 2718–36. doi:10.1002/anie.200200539.

Deller, Marc C., and Bernhard Rupp. 2015. “Models of Protein-Ligand Crystal Structures: Trust, but Verify.” *Journal of Computer-Aided Molecular Design* 29 (9). Springer International Publishing: 817–36. doi:10.1007/s10822-015-9833-8.

Djinovic-Carugo, Kristina, and Oliviero Carugo. 2015. “Missing Strings of Residues in Protein Crystal Structures.” *Intrinsically Disordered Proteins* 3 (1): 1–7. doi:10.1080/21690707.2015.1095697.

Dosztányi, Zsuzsanna, Bálint Mészáros, and István Simon. 2009. “Bioinformatical Approaches to Characterize Intrinsically Disordered/unstructured Proteins.” *Briefings in Bioinformatics* 11 (2): 225–43. doi:10.1093/bib/bbp061.

Felli, Isabella C., and Roberta Pierattelli. 2015. “Intrinsically Disordered Proteins Studied by NMR Spectroscopy.” *Advances in Experimental Medicine and Biology0065-2598* 870 (Chapter 159): 361–62. doi:10.1007/978-3-319-20164-1.

Fiser, András, and Andrej Sali. 2003. “ModLoop: Automated Modeling of Loops in Protein Structures.” *Bioinformatics* 19 (18): 2500–2501. doi:10.1093/bioinformatics/btg362.

Galaktionov, Stan, Gregory V Nikiforovich, and Garland R Marshall. 2001. “Ab Initio Modeling of Small, Medium, and Large Loops in Proteins.” *Biopolymers* 60 (2): 153–68. doi:10.1002/1097-0282(2001)60:2<153::AID-BIP1010>3.0.CO;2-6.

Gore, Swanand, Sameer Velankar, and Gerard J. Kleywegt. 2012. “Implementing an X-Ray Validation Pipeline for the Protein Data Bank.” *Acta Crystallographica Section D: Biological Crystallography* 68 (4). International Union of Crystallography: 478–83. doi:10.1107/S0907444911050359.

Grimsley, Gerald R., J. Martin Scholtz, and C. Nick Pace. 2009. “A Summary of the Measured pK Values of the Ionizable Groups in Folded Proteins.” *Protein Science* 18 (1): 247–51. doi:10.1002/pro.19.

Hekkelman, M. L., T. A H te Beek, S. R. Pettifer, D. Thorne, T. K. Attwood, and G. Vriend. 2010. “WIWS: A Protein Structure Bioinformatics Web Service Collection.” *Nucleic Acids Research* 38 (SUPPL. 2): 719–23. doi:10.1093/nar/gkq453.

Helliwell, John R. 1992. *Macromolecular Crystallography with Synchrotron Radiation*. Cambridge, UK: Cambridge University Press. doi:10.1017/CBO9780511524264.

Hooft, R W, C Sander, and G Vriend. 1997. “Objectively Judging the Quality of a Protein Structure from a Ramachandran Plot.” *Computer Applications in the Biosciences : CABIOS* 13 (4): 425–30. doi:10.1093/bioinformatics/13.4.425.

Hooft, R W W, C Sander, and G Vriend. 1996. “Positioning Hydrogen Atoms by Optimmizing Hydrogen-Bond Networks in Protein Strutures.” *Proteins: Struture,*

Function, and Genetics 26 (January): 363–76. doi:10.1002/(SICI)1097-0134(199612)26:4<363::AID-PROT1>3.0.CO;2-D.

Jamroz, Michal, and Andrzej Kolinski. 2010. “Modeling of Loops in Proteins: A Multi-Method Approach.” *BMC Structural Biology* 10: 5. doi:10.1186/1472-6807-10-5.

Karp, Daniel A., Apostolos G. Gittis, Mary R. Stahley, Carolyn A. Fitch, Wesley E. Stites, and Bertrand García-Moreno E. 2007. “High Apparent Dielectric Constant Inside a Protein Reflects Structural Reorganization Coupled to the Ionization of an Internal Asp.” *Biophysical Journal* 92 (6): 2041–53. doi:10.1529/biophysj.106.090266.

Kim, Meekyum Olivia, Sara E. Nichols, Yi Wang, and J. Andrew McCammon. 2013. “Effects of Histidine Protonation and Rotameric States on Virtual Screening of M. Tuberculosis RmlC.” *Journal of Computer-Aided Molecular Design* 27 (3): 235–46. doi:10.1007/s10822-013-9643-9.

Kleywegt, Gerard J. 2000. “Validation of Protein Crystal Structures.” *Acta Crystallographica Section D: Biological Crystallography* 56 (3): 249–65. doi:10.1107/S0907444999016364.

Kleywegt, Gerard J., and Axel T. Brünger. 1996. “Checking Your Imagination: Applications of the Free R Value.” *Structure* 4 (8): 897–904. doi:10.1016/S0969-2126(96)00097-4.

Kleywegt, Gerard J., Mark R. Harris, Jin Yu Zou, Thomas C. Taylor, Anders W??hlby, and T. Alwyn Jones. 2004. “The Uppsala Electron-Density Server.” *Acta Crystallographica Section D: Biological Crystallography* 60 (12 I): 2240–49. doi:10.1107/S0907444904013253.

Kleywegt, Gerard J., and T. Alwyn Jones. 1995. “Where Freedom Is Given, Liberties Are Taken.” *Structure* 3 (6): 535–40. doi:10.1016/S0969-2126(01)00187-3.

Kleywegt, Gerard J., and T. Alwyn Jones. 1997. “Model Building and Refinement Practice.” *Methods in Enzymology* 277: 208–30. doi:10.1016/S0076-6879(97)77013-7.

Kleywegt, Gerard J, and T Alwyn Jones. 1996. “Phi/Psi-Chology: Ramachandran Revisited.” *Structure* 4 (12): 1395–1400. doi:10.1016/S0969-2126(96)00147-5.

Lovell, Simon C, Ian W Davis, W B Adrendall, P I W de Bakker, J Michael Word, Michael G Prisant, J S Richardson, and David C Richardson. 2003. “Structure Validation by C Alpha Geometry: Phi,psi and C Beta Deviation.” *Proteins-Structure Function and Genetics* 50 (August 2002): 437–50. doi:10.1002/prot.10286.

McDonald, I K, and J M Thornton. 1995. “The Application of Hydrogen Bonding Analysis in X-Ray Crystallography to Help Orientate Asparagine, Glutamine and Histamine Side Chains.” *Protein Engineering* 8 (3): 217–24. doi: 10.1093/protein/8.3.217.

Morris, A. L., M. W. MacArthur, E. G. Hutchinson, and J. M. Thornton. 1992. “Stereochemical Quality of Protein Structure Coordinates.” *Proteins: Structure, Function and Genetics* 12 (4): 345–64. doi:10.1002/prot.340120407.

Morris, Richard J., and Gérard Bricogne. 2003. “Sheldrick’s 1.2 Å Rule and beyond.” *Acta Crystallographica - Section D Biological Crystallography* 59 (3): 615–17. doi:10.1107/S090744490300163X.

Mysinger, Michael M., Michael Carchia, John J. Irwin, and Brian K. Shoichet. 2012. “Directory of Useful Decoys, Enhanced (DUD-E): Better Ligands and Decoys for Better Benchmarking.” *Journal of Medicinal Chemistry* 55 (14): 6582–94. doi:10.1021/jm300687e.

Olsson, Mats H M, Chresten R. SØndergaard, Michal Rostkowski, and Jan H. Jensen. 2011. “PROPKA3: Consistent Treatment of Internal and Surface Residues in Empirical P K a Predictions.” *Journal of Chemical Theory and Computation* 7: 525–37. doi:10.1021/ct100578z.

Pace, C. Nick, Gerald R. Grimsley, and J. Martin Scholtz. 2009. “Protein Ionizable Groups: pK Values and Their Contribution to Protein Stability and Solubility.” *Journal of Biological Chemistry* 284 (20): 13285–89. doi:10.1074/jbc.R800080200.

Park, Hahnbeom, Gyu Rie Lee, Lim Heo, and Chaok Seok. 2014. “Protein Loop Modeling Using a New Hybrid Energy Function and Its Application to Modeling in Inaccurate Structural Environments.” *PLoS ONE* 9 (11): 1–18. doi:10.1371/journal.pone.0113811.

Radivojac, Predrag, Zoran Obradovic, David K Smith, Guang Zhu, Slobodan Vucetic, Celeste J Brown, J David Lawson, and a Keith Dunker. 2004. “Protein Flexibility and Intrinsic Disorder.” *Protein Sci* 13 (1): 71–80. doi:10.1110/ps.03128904.

Read, Randy J., Paul D. Adams, W. Bryan Arendall, Axel T. Brunger, Paul Emsley, Robbie P. Joosten, Gerard J. Kleywegt, et al. 2011. “A New Generation of Crystallographic Validation Tools for the Protein Data Bank.” *Structure* 19 (10): 1395–1412. doi:10.1016/j.str.2011.08.006.

Rhodes, Gale. 2006. *Crystallography Made Crystal Clear: A Guide for Users of Macromolecular Models*. *Complementary Science Series*. Vol. 35. doi:10.1002/bmb.89.

Rose, Peter W., Andreas Prlić, Chunxiao Bi, Wolfgang F. Bluhm, Cole H. Christie, Shuchismita Dutta, Rachel Kramer Green, et al. 2015. “The RCSB Protein Data Bank: Views of Structural Biology for Basic and Applied Research and Education.” *Nucleic Acids Research* 43 (D1): D345–56. doi:10.1093/nar/gku1214.

Rostkowski, Michal, Mats H M Olsson, Chresten R Soendergaard, and Jan H Jensen. 2011. “Graphical Analysis of pH-Dependent Properties of Proteins Predicted Using PROPKA.” *BMC Struct. Biol.* 11 (Cc): 6. doi:10.1186/1472-6807-11-6.

Rupp, Bernhard. 2009. *Biomolecular Crystallography: Principles, Practice, and Application to Structural Biology*. https://books.google.co.il/books?id=gTAWBAAAQBAJ&dq=torsion+angles+chirality&source=gbs_navlinks_s.

Sali, A, and T L Blundell. 1993. “Comparative Protein Modelling by Satisfaction of Spatial Restraints.” *Journal of Molecular Biology* 234 (3): 779–815. doi:10.1006/jmbi.1993.1626.

Smyth, M S, and J H J Martin. 2000. “Review X Ray Crystallography.” *J Clin Pathol: Mol Pathol* 53 (1): 8–14. doi:10.1136/mp.53.1.8.

Søndergaard, Chresten R., Mats H M Olsson, Michał Rostkowski, and Jan H. Jensen. 2011. “Improved Treatment of Ligands and Coupling Effects in Empirical

Calculation and Rationalization of P K a Values." *Journal of Chemical Theory and Computation* 7 (7): 2284–95. doi:10.1021/ct200133y.

Thurlkill, R L, R L Thurlkill, G R Grimsley, G R Grimsley, J M Scholtz, J M Scholtz, C N Pace, and C N Pace. 2006. "pK Values of the Ionizable Groups of Proteins." *Protein Sci* 15 (5): 1214–18. doi:10.1110/ps.051840806.

Warren, Gregory L., Thanh D. Do, Brian P. Kelley, Anthony Nicholls, and Stephen D. Warren. 2012. "Essential Considerations for Using Protein-Ligand Structures in Drug Discovery." *Drug Discovery Today* 17 (23–24). Elsevier Ltd: 1270–81. doi:10.1016/j.drudis.2012.06.011.

Weichenberger, Christian X., Piotr Byzia, and Manfred J. Sippl. 2008. "Visualization of Unfavorable Interactions in Protein Folds." *Bioinformatics* 24 (9): 1206–7. doi:10.1093/bioinformatics/btn108.

Weichenberger, Christian X., and Manfred J. Sippl. 2006. "Self-Consistent Assignment of Asparagine and Glutamine Amide Rotamers in Protein Crystal Structures." *Structure* 14 (6): 967–72. doi:10.1016/j.str.2006.04.002.

Weichenberger, Christian X., and Manfred J. Sippl. 2007. "NQ-Flipper: Recognition and Correction of Erroneous Asparagine and Glutamine Side-Chain Rotamers in Protein Structures." *Nucleic Acids Research* 35 (SUPPL.2): 403–6. doi:10.1093/nar/gkm263.

Wlodawer, Alexander, Wladek Minor, Zbigniew Dauter, and Mariusz Jaskolski. 2008. "Protein Crystallography for Non-Crystallographers, or How to Get the Best (but Not More) from Published Macromolecular Structures." *FEBS Journal* 275 (1): 1–21. doi:10.1111/j.1742-4658.2007.06178.x.

Word, J M, S C Lovell, J S Richardson, and D C Richardson. 1999. "Asparagine and Glutamine: Using Hydrogen Atom Contacts in the Choice of Side-Chain Amide Orientation." *Journal of Molecular Biology* 285 (4): 1735–47. doi:10.1006/jmbi.1998.2401.

Yuan, Zheng, Timothy L. Bailey, and Rohan D. Teasdale. 2005. "Prediction of Protein B-Factor Profiles." *Proteins: Structure, Function and Genetics* 58 (4): 905–12. doi:10.1002/prot.20375.

Zhou, Alice Qinhua, Corey S. O'Hern, and Lynne Regan. 2011. "Revisiting the Ramachandran Plot from a New Angle." *Protein Science* 20 (7): 1166–71. doi:10.1002/pro.644.