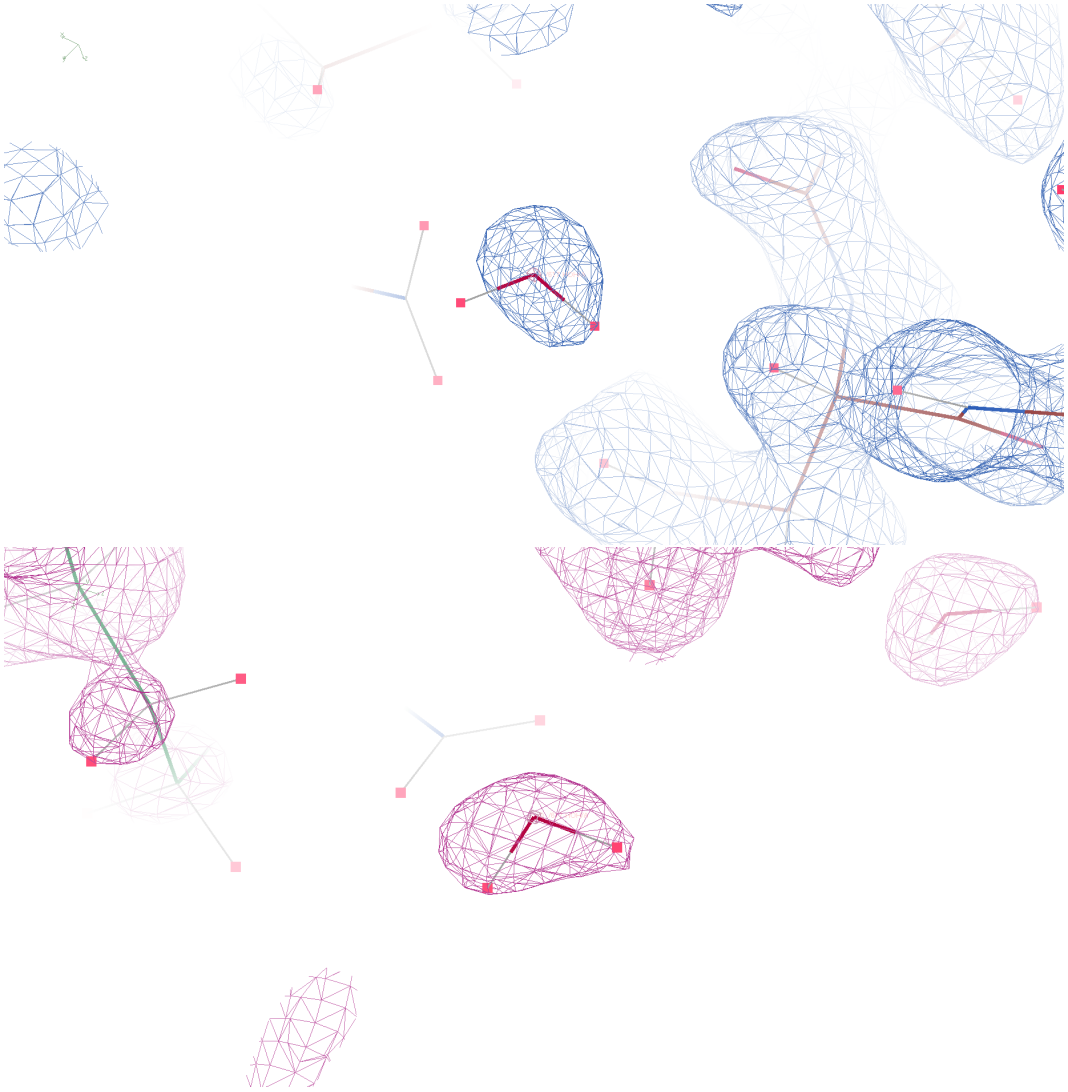
LUND UNIVERSITY

Automated Orienting of Water Molecules in Neutron Crystal Structures

Written by: Axl Eriksson
Supervisor: Ulf Ryde
Examiner: Jan Forsman
Summer 2019 (15 hp)
2019-06-24



Lund University
Department of Theoretical Chemistry
Faculty of Science

Abstract

A protein's function is directly related to its structure, which is in a water medium where it is affected by hydrogen bonds and the hydrophobic effect. These interactions are in turn dependent on the water molecules' orientations around the protein. Therefore, it is vital to have correct orientations for the water molecules. Such information can be obtained by neutron crystallography. However, even in such structures, the correct orientation of water requires a manual evaluation and possible re-orientation of each water molecule. This is a tedious and time-consuming procedure since proteins typically contain hundreds of water molecules in their lattice and can have several sub-units. Therefore, we have here tried to develop a method that reliably automatizes the orienting of the water molecules in a simple and relatively fast way. As test cases, we used the proteins galectin-3C, rubredoxin and pyrophosphatase. We evaluated the water molecules' orientations both quantitatively, with RSCC values, and qualitatively, by studying the orientations in density maps. We have optimized the refinement by varying the optimization methods and refinement parameters, thus finding the settings that yielded the best results in terms of time and performance. In particular, we have constructed two scripts that identify and re-orient inadequately oriented water molecules. Ultimately, we performed the refinement and re-orienting using only neutron data. We show that our approach yields improved orientations of the water molecules for all three proteins, in a shorter time than a manual orientation.

Populärvetenskaplig sammanfattning på svenska

Vid läkemedelsutveckling används proteiner vars funktion är beroende på hur det är veckat (dess konformation). Denna veckning påverkas i sin tur av vätebindningar från vatten, som bildar vätskemiljön i kroppen där medicinen ska verka och styrkan av vätebindningar påverkas i sin tur av vattenmolekylernas orientering kring proteinet. För att kunna förutse proteinets funktion i kroppen är det därför viktigt att känna till vattenmolekylernas orientering i förhållande till proteinet. Efter att ha bildat en kristall av proteinet och utfört t.ex neutronkristallografi, kan man ta reda på detta genom att utföra en kristallografisk förfining. I denna förfining orienteras vattenmolekylerna automatiskt baserat på en modell och kärntäthetskartorna från neutrondatan. Men detta steg är inte perfekt och kräver därför en manuell omorientering av felaktigt orienterade vattenmolekyler. Detta brukar ta lång tid och kräver mycket arbete eftersom proteinstrukturer kan innehålla flera hundra vattenmolekyler och kan bestå av flera subenheter. Eftersom omorienteringen görs manuellt, kan resultaten vara lite partiska och därmed finns inte heller någon tydlig gräns på vad som gör en vattenmolekyls orientering bra eller dålig. Därför vill vi utveckla en objektiv metod som på ett pålitligt sätt automatiserar orienteringen av vattenmolekyler på ett snabbt och enkelt sätt. För att göra metoden så objektiv som möjligt införde vi real-space correlation coefficient (RSCC) som en numerisk validering på vattenmolekylerna. Som testproteiner använde vi galectin-3C, rubredoxin och pyrofosfatas. Vi har optimerat förfiningen genom att variera optimeringsmetoden och parametrarna. Sedan jämförde vi RSCC med vår manuella studie av de modellerade vattenmolekylerna. När vi hittat den optimerade förfiningsmetoden gjorde vi ett program som automatisk roterade dåligt orienterade vattenmolekyler baserat på deras RSCC. Vår metod gav förbättrad orientering bland vattenmolekylerna hos alla tre proteiner och på kortare tid än en manuell orientering.

Contents

1 Theory

1.1 Introduction

Proteins are vital bio-molecules with a vast variety of functions, such as transporting chemicals in the cell or forming/breaking down metabolites in anabolic or catabolic processes. A protein's function varies with its conformation and its interactions with water affects the conformation due to hydrogen bonds and the hydrophobic effect [1]. Since water molecules form the natural surroundings of the proteins, they can affect the dynamics of the protein and also stabilize protein folding [2, 3]. The strengths of these protein-water interactions depend on the orientations of the water molecules in relation to the protein complex. Hence, to fully understand and determine a protein's function, which is vital in for example drug development, it is important to determine the orientations of the water molecules in the protein complex [4].

1.2 X-ray crystallography

When studying protein structures, X-ray crystallography is the most common method to obtain the three-dimensional structure, with a high resolution as techniques and equipment improve [5]. With X-ray crystallography, details in the bonding around atoms can be revealed and studied with the electron density map. An example is the Fe^{2+} ion in hemoglobin in Figure 1, for which the electron density map indicates where the Fe^{2+} ion and the other heavy atoms are located, based on the electron density. The figure shows the electron density of the heme ring and the contours are closer together (denser) where the electron density is higher. On top of the density map, the heme group is illustrated. In the center of the ring, the Fe^{2+} ion is located, where the electron density is highest [6].

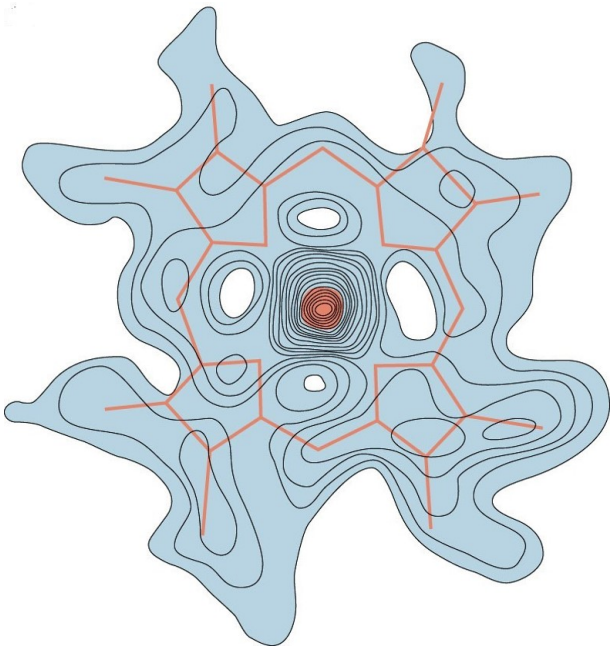


Figure 1: An electron density map of the heme group in hemoglobin, generated by X-ray crystallography. The Fe^{2+} ion in the center has the highest electron density. Illustration from the book *Fundamentals of Biochemistry* by Voet.

With X-ray diffraction, the reflection of the protein is detected with an electronic detector, as a diffraction pattern. This reflection is called the *reciprocal space* and is described by the coordinates *h*, *k* and *l*, which describe the position of an individual reflection in reciprocal space of the diffraction pattern [7]. The center of the detector is the origin with the coordinates *hkl = 000* and the other reflections are simply integers in *hkl*. In X-ray diffraction, each diffracted X-ray can be described as a sum of all diffraction contributions from all atoms in the unit cell. This sum is called the *structure factor equation* and the sum from the reflection *hkl* is called a *structure factor* (F_{hkl}). The structure factor can be calculated by Equation 1:

$$F_{hkl} = f_A + f_B + ... \tag{1}$$

Equation 1 shows that every reflection on the detector results from diffractions of all atoms in the lattice, where f_A is the diffraction by atom A [7]. When mapping a protein, experimental structure factors ($F_{o,\,hkl}$) obtained from the X-ray data and calculated structure factors ($F_{c,\,hkl}$) obtained from the current model are compared. The model is obtained by first starting from a crude model of the structure built into the electron density, calculated by Equation 2 with observed intensities from the X-ray data, which are proportional to ${F_{o,\,hkl}}^2$, and estimated phases (α_{calc}). The crude model can originate from an already solved protein with a similar structure as the target protein [8].

$$\rho(x,y,z) = \frac{1}{V}\Sigma_h\Sigma_k\Sigma_l|F_{o,\,hkl}|e^{-2\pi i(hx+ky+lz-\alpha_{calc})} \tag{2}$$

where V is the cell volume. Next, the electron density map is visualized and molecular groups or atoms are identified. From this model, structure factors can be calculated by Equation 3 with the

information from the second step and the crude model [7].

$$F_{c,\,hkl} = \int_h \int_k \int_l \rho(x,y,z) e^{2\pi i(hx+ky+lz)} dxdydz \tag{3}$$

A flowchart of the model improvement just described, is illustrated in Figure 2 where the crude model is the initial model used to develop develop the electron density map. The improved model then becomes the new initial model and the cycle repeats.

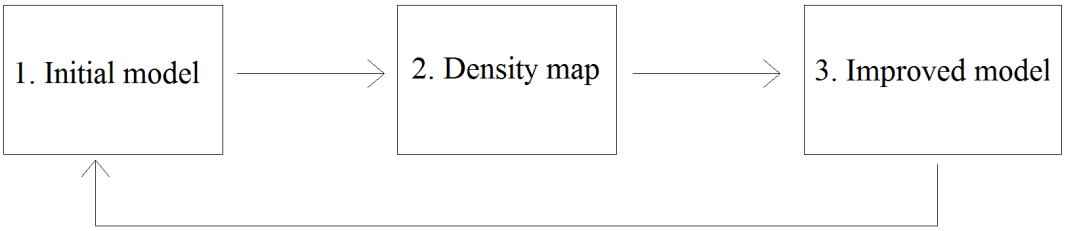


Figure 2: Flowchart of the refinement of the model.

The model is normally improved by performing crystallographic refinement. This usually involves a minimization and consists of several cycles, in which the *refinement parameters* and *refinement target* are optimized by the *optimization method*.

The refinement parameters are the variables that are optimized in the minimization and they describe the crystal with its properties. These refinement parameters combine to form the total calculated structure factors of the model (as described by Equation 3), and should resemble the total structure factors of the observed data as closely as possible. The refinement parameters typically include the coordinates, the atomic displacement parameters (ADP, further discussed below) and the occupation numbers of each atom. However, they can also focus on non-atomic parameters like twinning or anisotropy [9].

Hydrogen atoms pose a special problem to crystallography, because they involve only a single electron and therefore are hard to discern in X-ray structures. Therefore, they are normally not included in the models. At very high resolution, they start to be visible and might be included in the model. However, it is more common to include them as riding hydrogens. This means that the hydrogens are placed in ideal positions around the heavy atom. This allows the hydrogens to be included in the model even if they cannot be detected due to insufficient resolution. However, all hydrogens are not in their ideal position and when there are neutron data available, or if the resolution is good enough, the hydrogens should be refined individually [10].

The refinement target is the mathematical function (the target function) that is minimized during the refinement. This function is constructed to decrease the more the model agrees with the observed data. Hence, crystallographic refinement is about modifying the model parameters to

reduce the target function and thus optimizing the model [9]. The various model parameters are typically refined in separate steps, rather than simultaneously. A typical refinement function is shown in Equation 4.

$$T_{xyz} = WXC_{scale} * WXC * T_{exp} + WC * T_{xyz,restraints} \quad (4)$$

where T_{xyz} is the target function with focus on positional parameters, keeping the other parameters fixed. T_{exp} is the term that describes how close the current model is to the experimental data, while $T_{xyz,restraints}$ is a restraint that compensates for lack of experimental data by introducing empirical knowledge (the ideal bond lengths and angles). WXC_{scale}, WXC and WC are weights used to balance the relative contributions from experimental and restraint terms. These weights can be optimized for the data and geometry restraints on the model. [9].

The optimization method is simply the method used for the optimization and there are plenty of different methods with a varying speed and applicability on the model parameters [9]. One example is simulated annealing which runs a molecular dynamics simulation at a high temperature. This allows the atoms and water molecules to overcome energy barriers that prevent them from accessing more favorable positions [9].

1.3 Neutron crystallography

As already mentioned, X-ray crystallography can yield structural information of macromolecules with a high resolution. However, it is insufficient to precisely locate the electron density around hydrogen atoms even at high resolution [5]. Even at ultra-high resolution (better than 1 Å), it is hard to reliably determine the electron density around the hydrogens, as they only contain a single electron. But knowing the hydrogen positions in a protein is often important, as they can determine the direction and polarity of hydrogen bonds, and play key roles in enzymatic reactions such as hydrolysis [11, 12]. Therefore, the protein can be deuterated when forming the crystal and then analyzed with neutron crystallography which reveals the exact position of each deuterium, through nuclear scattering [13]. The sample can either be perdeuterated, in which all hydrogens in the structure have been exchanged for deuterium, or it can be partially deuterated, in which only the solvent is deuterated, meaning that only solvent-exposed exchangeable (i.e. polar) hydrogens are replaced by deuterium (which is typically about half of the hydrogens) [14]. If not all the hydrogen atoms are exchanged at a specific site, the signals may be cancelled with the deuteriums because hydrogens have negative scattering lengths whereas deuteriums (and atoms like C, N, O and S) have positive scattering lengths. Furthermore, the magnitude of a hydrogen's scattering length is about half of that of deuterium leading to a complete cancellation if the H/D ratio is 2:1 [15]. Another advantage with neutron crystallography is that it does not cause radiation damage to the protein, so multiple measurements can be performed at room temperature [10]. Since X-ray and neutron crystallography give different but complementary information about the protein, they are often used together on the same crystal, followed by a *joint X-ray/neutron refinement*. This allows the study of the protein's X-ray and neutron density maps simultaneously [16].

1.4 Density maps

The proteins' neutron and X-ray maps can be studied in software like Coot or PyMOL. By studying the maps for $2F_o - F_c$ and $F_o - F_c$, where F_o is the observed structure factors and F_c is the calculated

from Equation 3, the quality of the current model can be judged [17, 18]. In other words, this is a comparison between the experimental data and the model. $F_o - F_c$ compares the observed structure factors with those of the model and highlights where there are too much or too little electron density. $2F_o - F_c$ mimics an F_o map but adds an extra $(F_o - F_c)$-term to compensate for model bias in the phases used to convert from reciprocal to real space. Figure 3 illustrates $2F_o - F_c$ and $F_o - F_c$ nuclear-density maps from neutron data, but also the electron density generated from the X-ray data.

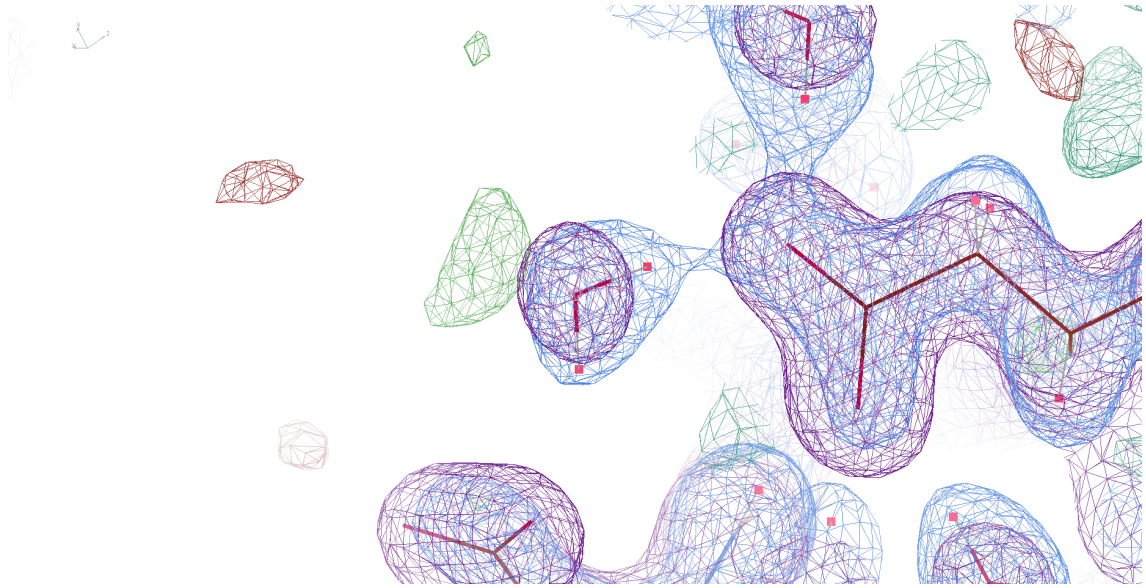


Figure 3: Illustration of a water molecule (in the center) in the protein galectin-3C (PDB ID 6EYM), with the electron density given by the $2F_o - F_c$ X-ray map (σ = 1.0) as a purple sphere around the oxygen atom. The $2F_o - F_c$ neutron map (σ = 1.0) is indicated in light blue (note that there is significant nuclear density also around the deuterium atoms) and the $F_o - F_c$ neutron difference map is indicated in green (σ = 3.0) where F_o is greater than F_c, and in red (σ = -3.0) where F_c is greater than F_o.

1.5 Validation metrics

The most common and effective parameter to use in numerical evaluations of the local quality of the current model is the real-space Z-difference (RSZD) score because it locally measures the accuracy of the model and the errors in the data, by using the F_o - F_c difference maps. However, the comparison between F_o and F_c might not be sensitive enough to judge data from water molecules, which is typically too weak. Edstats is the only widely used software to calculate RSZD [19].

Another quality measure is the real-space correlation coefficient (RSCC), which describes the relation between the calculated electron density from the model and that of experimental data. The RSCC can be used on arbitrary sets of atoms and can be calculated directly in the Phenix

software from

$$RSCC = \frac{\sum |\rho_{obs} - < \rho_{obs} > | \sum |\rho_{calc} - < \rho_{calc} > |}{(\sum |\rho_{obs} - < \rho_{obs} > |^2 \sum |\rho_{calc} - < \rho_{calc} > |^2)^{1/2}} \quad (5)$$

where ρ_{obs} and ρ_{calc} are the observed and calculated electron densities, respectively [20].

The ADP or *B-factor* (sometimes called temperature factor), is a model parameter that describes how much an atom vibrates around its position in the model. Atoms in the sidechains typically exhibit larger vibrations than those in the mainchain. Since diffraction is affected by vibrating atoms, every atom j should be assigned a unique temperature factor (B_j) [7]. However, B_j does not only reflect the thermal motion, but also disorder i.e. atoms that do not occupy the same position in every unit cell. This makes it difficult to distinguish these two sources to the B-factor [7]. However, an unusually high B_j tends to be dominated by errors which makes it difficult to accurately determine large vibrations. In addition, the atom is likely modeled inadequately if the B-factor of an atom is much greater than the average B-factor of surrounding atoms.

The measured structure factor amplitudes will be equal to the calculated structure factor amplitudes if the model is good. The residual index (the *R*-factor) is the most common way to measure this convergence with the general formula:

$$R = \frac{\sum ||F_{obs}| - |F_{calc}||}{\sum |F_{obs}|} \quad (6)$$

Thus, the *R*-factor measures the accuracy of the model. Two different R-factors are normally calculated: R_{work} and R_{free}. R_{work} is calculated for all structure factors and measures the agreement between the current model and experimental data. On the other hand, for the calculation of R_{free}, only about 5% of all structure factors are used and these are not employed in the refinement. Consequently, R_{free} estimates the overfitting in the model [21]. Therefore, R_{work} is normally smaller than R_{free}. If R_{free} is much greater, the model is over-fitted [7, 22]. Since the *R*-factors measure the model globally, they are not good validation metrics for individual atoms or groups.

2 Method

The crystal structure we used to develop this method was of galectin-3C (PDB ID: 6EYM), with the resolution 1.7 Å and 110 water molecules. Galectin-3C is a lactose-binding protein. The coordinates, as well as the neutron and X-ray maps were obtained from the Protein Data Bank (PDB) [23]. We used two other proteins to test the model, one being rubredoxin (PDB ID 4AR3), an electron-transport protein, with the resolution 1.05 Å and 149 water molecules. The other was pyrophosphatase (PDB ID 5TY5), a hydrolase, with the resolution 2.2 Å and 385 water molecules.

In several cases, we evaluated qualitatively whether the water molecules had an adequate orientation: For every water molecule in the structure, we studied visually the water orientations in the neutron and electron density maps using Coot. If the orientation is good, the deuteriums have been placed inside the neutron density around the water molecule. For the inadequate orientations the deuteriums are placed where no neutron density is present or where it is lower than the noise. Figure 4 shows an example of a good orientation of a water molecule where both deuteriums are where the density is the highest. One of its deuteriums also participates in a hydrogen bond with the protein.

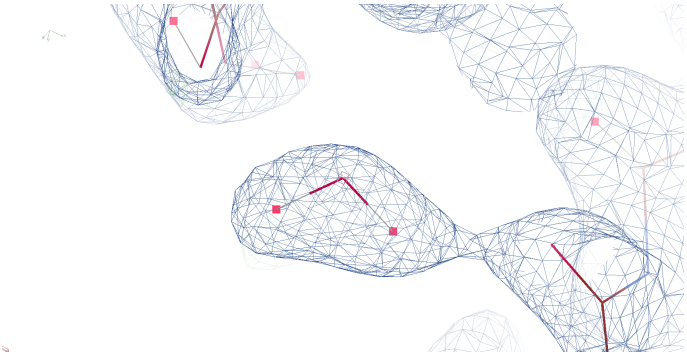


Figure 4: Example of a $2F_o - F_c$ neutron map, indicated in blue, (σ = 1.0) with a good orientation of the water molecule.

Figure 5 shows an example of an inadequate orientation of a water molecule where the deuteriums are outside of the neutron density.

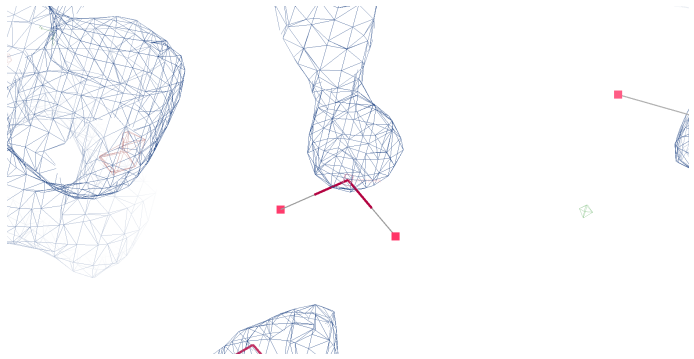


Figure 5: Example of a $2F_o - F_c$ neutron map, indicated in blue, (σ = 0.80) with a poorly oriented water molecule with both deuteriums outside of the neutron density.

For galectin-3C, the water orientations of the deposited structure were evaluated in Coot. Then the hydrogens were removed from the water molecules and re-added with Ready Set and Maestro, respectively [9, 24]. After a joint X-ray/neutron refinement of 3 macro-cycles, the structure was evaluated again. Both Ready Set and Maestro add hydrogens (or deuteriums) to water using their own energy-based algorithms. Maestro is widely used in computational chemistry where correct hydrogen positions are important, so it was expected to perform best. The structures were then refined with phenix.refine using different settings. We did refinements in reciprocal space, real space and in combined reciprocal and real space. In some calculations, the X-ray data was excluded and the refinement used only neutron data, with the protein fixed and only the water molecules (including the oxygens) allowed to move. We varied the number of macro-cycles with the goal to find the best result while minimizing the computational cost, both with and without simulated annealing. The deuteriums were refined individually because we used the neutron data.

The qualitative evaluation was performed visually in Coot after the refinements in phenix.refine (except structures that were deposited to PDB). Owing to the vast amount of refinements, we did not qualitatively evaluate all of the generated structures. Hence, the structures that miss qualitative data have only been evaluated based on the RSCC values. The structures that were analyzed qualitatively are listed in Table 1.

Table 1: The structures with qualitative data.

Protein	Waters	Macro-cycles	Refinement settings
Galectin-3C	Deposited	/	/
Galectin-3C	Ready Set	3	Joint X-ray/Neutron with standard settings
Galectin-3C	Maestro	3	Joint X-ray/Neutron with standard settings
Galectin-3C	Ready Set	15	Neutron refinement on fixed protein
Galectin-3C	Ready Set	15	Neutron refinement on fixed protein, after orientation script
Rubredoxin	Deposited	/	/
Rubredoxin	Ready Set	3	Neutron refinement on fixed protein
Rubredoxin	Ready Set	15	Neutron refinement on fixed protein
Pyrophosphatase	Deposited	/	/
Pyrophosphatase	Ready Set	15	Neutron refinement on fixed protein

When evaluating the deposited structure of galectin-3C, we found that 40 out of 110 (36%) of the water molecules were in inadequate orientations according to our qualitative assessment. For rubredoxin 98 out of 149 (66%) were in inadequate orientations and for pyrophosphatase it was 150 out of 385 (39%), according to our evaluation. An example is given in Figure 6. From the galectin-3C structure generated by Ready Set 81 out of 110 water molecules were in poor orientations. This shows that our qualitative evaluation is reasonable and gives results like that obtained by other groups.

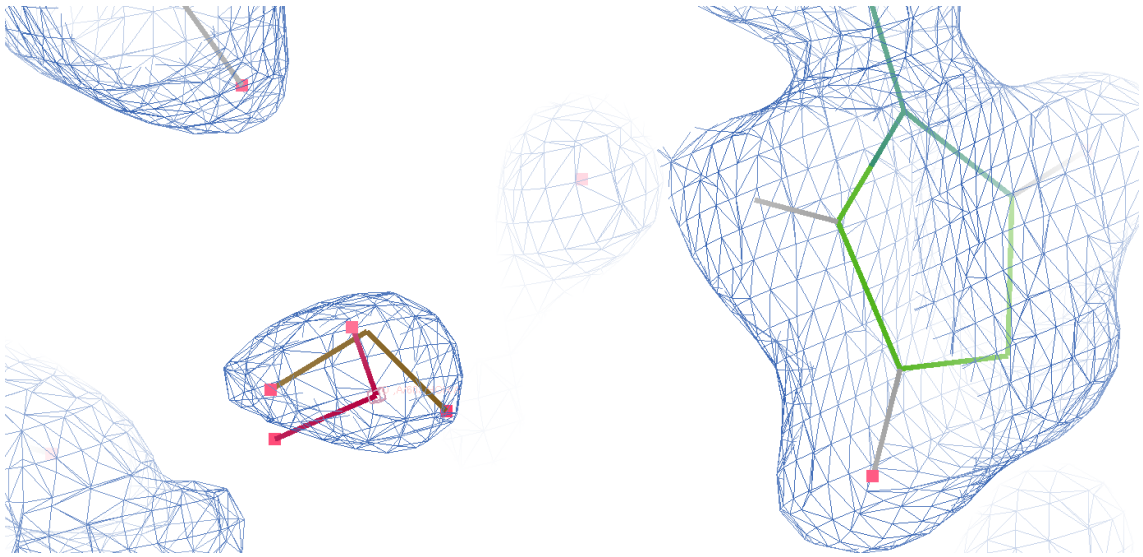


Figure 6: Example of a water molecule in the deposited structure of galectin-3C with an incorrect orientation (red) according to our qualitative evaluation. A better orientation, according to our assessment, is shown in brown (oriented manually). The $2F_o - F_c$ neutron density map is shown in blue ($\sigma = 0.89$).

To obtain quantitative data, we evaluated the water molecules based on their RSZD, B-factors and RSCC. As will be discussed in the Results section, RSCC gave the most reliable results. As a threshold for RSCC, we selected the value that gave the largest number of correct predictions of both correctly and incorrectly oriented water molecules, compared to the qualitative judgement.

Next, we made a script to systematically improve the positions of the deuterium atoms of the water molecules. The script starts by calculating the RSCC values for all water-deuteriums in an initial structure (obtained from PDB, Ready Set or Maestro). For all water molecules with at least one deuterium atom with a RSCC value below the threshold, the script first rotates one of the deuterium atoms (D1) an angle α around the O–D2 axis (where O and D2 are the oxygen and the other deuterium atom of the water molecule). Then, D2 is rotated an angle α' around the O–D1 axis. The angles α and α' may or may not be the same depending on the RSCC values of D1 and D2. If the RSCC value is 0.2 below the RSCC threshold, the deuterium is rotated 5°, otherwise it is rotated 10°. Thus, the oxygen coordinate, the D-O-D bond angle and the O-D bond lengths are fixed, using data from Ready Set (106.8°, and 0.96 Å). There are two versions of the script with a single difference: One refines the water molecules after this rotation (Script Ref), whereas the other does not refine (Script No-ref). The refinement was in reciprocal space, with one macro-cycle and used only neutron data. The scripts then calculate new RSCC values for the two deuterium atoms and if the lowest of the two new RSCC values is the highest measured, the coordinates are stored. If the new RSCC values are above the threshold for both deuteriums, the corresponding water molecule is accepted, and no more conformations are tried. Otherwise, the re-orientation cycle is repeated. This continues until both deuteriums of all water molecules have RSCC values above the threshold, or if a maximum number of cycles have been performed. If, at the end, no water orientation gave RSCC values above the threshold, the best orientation is used and deuteriums with RSCC values below the threshold – 0.2 are removed from the model. In other words, the script models poorly described water molecules without any deuterium atoms. Thus, the script systematically goes through orientations for the water molecules, without consideration of the energy, thereby bypassing any energy barriers until the RSCC is optimized.

For both scripts, four maximum numbers of orientation cycles were tested: 10, 50, 100 and 150. When the scripts were run on the galectin-3C, the structure was refined with phenix.refine, using the best optimization methods. Then, we calculated the RSCC values of the structure. After that, the RSCC values and procedure times of the refined structure from Script Ref were compared with those of Script No-ref. For the best performing script, we decided what number of orientation cycles was the most effective based on the RSCC data and performance time. Then, we qualitatively evaluated the protein structure with the best script and optimization strategy. The best performing script and optimization strategy were used on rubredoxin and pyrophosphatase using only neutron data, and their RSCC values were calculated.

3 Results and Discussion

3.1 Optimization methods

In this study, we have tried to develop an automatic method to orient water molecules in neutron crystal structures. This can be done by systematically trying a large number of positions for each water molecule in the structure. However, for such an approach, it is necessary to have an automatic quality measure that can be used to determine which orientation is best and whether it is acceptable or not. Therefore, our first step was to find such a measure by comparing three reasonable quantitative quality measures, RSZD, RSCC and the B-factor of each water-deuterium. The three measures are compared to the results of the qualitative (manual) assessment in Figure 7.

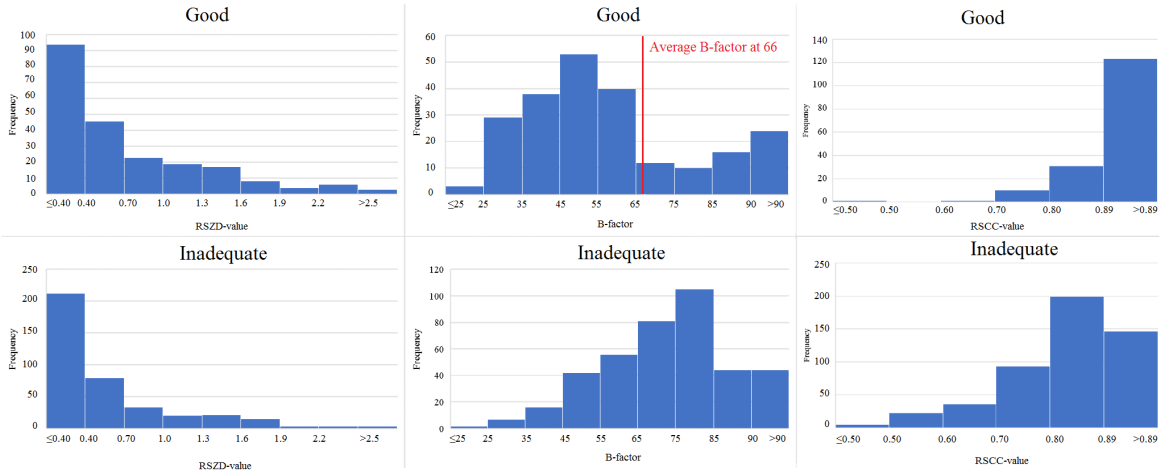


Figure 7: Histograms of RSCC, B-factors and RSZD, all calculated by Phenix on galectin-3C after 3 macro-cycles of refinement with standard settings. The histograms are calculated separately for water molecules with good (top row) or inadequate orientations (bottom row), according to the qualitative assessment. The ideal RSZD value is zero, while a perfect RSCC is one, and a good B-factor is around the average value (marked with a red line). The histograms contain 252 deuterium atoms with good orientation and 408 deuterium atoms with inadequate orientation, in total 660.

From Figure 7, it can be seen that the RSZD score does not give consistent results as the histograms for the good and inadequate orientations look almost the same. Furthermore, there seems to be some tendency that the poor orientations give lower RSZD values than the good ones. For the B-factors, the results are more promising: The distribution for the water molecules with proper orientation is more concentrated toward lower values, with a peak somewhat below the average (66), whereas the distribution is biased towards higher values for those with a poor orientation. However, there is a second peak at very high B-factors for some properly oriented water molecules. The results for the RSCC score is even better: The result for the properly oriented water molecules are strongly biased towards high values, whereas those with a poor orientation have a wider distribution, peaking on somewhat lower values (although there is a quite large overlap). Thus, we

selected RSCC as our quantitative measure, especially as it also gives a simple implementation in the scripts.

The RSCC values can be calculated with both Phenix and Edstats. We chose to use Phenix because the difference between the two methods was minor. Furthermore, Phenix is already used for the refinement so this choice would make the method simpler. Likewise, for placement of hydrogens, Ready Set from Phenix yielded results of similar quality as Maestro despite the expectation that Maestro would perform better. Hence, we chose to use Ready Set for hydrogen placement so that all calculations can be performed by the same software.

Next, we have to select a threshold value for RSCC to determine whether a deuterium atom is in a proper position or not, as described in the Methods section. This was done by maximizing the total number of correct predictions of RSCC, both for correctly and incorrectly oriented water molecules, compared to the qualitative (manual) evaluation. This gave a threshold of 0.89 for galectin-3C and 0.81 when all data from galectin-3C, rubredoxin and pyrophosphatase were combined. There are a total of 660 deuteriums and 146 of these are poorly oriented deuteriums that receive good RSCC values (above the threshold 0.89), and 123 deuteriums receive good scores and orientations. This distinction is not great but 146 out of 660 (about 22%) is quite low, and while 123 out of 660 is lower, it can be improved by varying the refinement settings.

We first intended to use both X-ray and neutron data in a joint X-ray/neutron refinement but using the X-ray data slowed the refinement significantly. However, removing the X-ray data decreased the RSCC values probably because the ratio between data and parameters is reduced. To account for this we ran the neutron-only refinement with the protein fixed, which increased the RSCC values significantly at the cost of slowing down the refinement slightly. However, the improvement was significant enough to continue with the fixed-protein strategy. With the protein fixed, the refinement was still faster than with the X-ray data (without a fixed protein), but the RSCC values were also slightly worse. However, we judged the improvements outweighed the computational cost based on the data in Table 2.

Table 2: Galectin-3C with a refinement of 3 macro-cycles and standard settings compared with the same refinement but with only neutron data and a fixed protein.

Refinement	Joint X-ray/Neutron	Neutron with fixed protein
Time (s)	845	213
R_{work} (X-ray)	0.1327	/
R_{free} (X-ray)	0.1437	/
R_{work} (Neutron)	0.2046	0.1753
R_{free} (Neutron)	0.2184	0.2280

In Table 2, the R_{free} value increases significantly when excluding the X-ray data, which was expected since less data was used to improve the model. Performing simulated annealing (SA) did not improve the results and made the refinement slower. Table 3 presents all the strategies tested with their respective results for the neutron data. It can be seen that the refinements performed best in reciprocal space or with real space combined with reciprocal space.

Table 3: Performance of different optimization strategies for galectin-3C generated from Ready Set and 3 macro-cycles in the refinement. A water molecule is considered to be in the correct orientation if both deuterium atoms have RSCC values above the threshold 0.89.

Strategy	Time (s)	Number of water molecules in good orientations	R_{work}	R_{free}
Joint X-ray/neutron (reciprocal space)	844.8	34	0.2046	0.2184
Joint X-ray/neutron (real space)	1820	16	0.2148	0.2252
Neutron (real and reciprocal space)	203.2	31	0.1724	0.2288
Neutron (reciprocal space)	201.1	43	0.1724	0.2288
Neutron (real and reciprocal space, fixed protein)	214.3	52	0.1753	0.2280
Neutron (reciprocal space, fixed protein)	213.3	52	0.1753	0.2280
Neutron (real space, fixed protein)	134.7	24	0.1938	0.2305
Neutron (reciprocal space, fixed protein, SA)	584.6	51	0.1718	0.2379

Next we investigated what number of cycles that was most effective in terms of result and time. The results are presented in Table 4, which shows the time, *R*-factors and the number of water molecules in correct orientations, of each refinement (a water molecules is considered to be in the correct orientation if both deuterium atoms have RSCC values above the threshold 0.89). It can be seen that 15 macro-cycles is best because R_{free} stops to improve after 15 cycles. R_{work} and the number of properly-oriented water molecules increase slightly with more macro-cycles, but the improvement is uneven and small, compared to the increase in computer time.

Table 4: Dependence of the results on the number of macro-cycles for galectin-3C, refined in reciprocal space with a fixed protein and only neutron data. For each refinement, the refinement time, the number of correctly oriented water molecules (i.e. with RSCC > 0.89 for both deuterium atoms), R_{work} and R_{free} are given.

Macro-cycles	Time (s)	Number of water molecules in correct orientations	R_{work}	R_{free}
3	283	52	0.2028	0.2410
5	329	56	0.1753	0.2280
7	440	57	0.1727	0.2262
10	607	58	0.1724	0.2249
12	716	59	0.1718	0.2251
14	799	60	0.1715	0.2244
15	953	61	0.1713	0.2245
18	1040	63	0.1710	0.2251
22	1255	62	0.1708	0.2258
25	1492	64	0.1711	0.2267
30	1702	64	0.1711	0.2280
35	1927	65	0.1706	0.2278

If the time does not matter, one can increase the number of macro-cycles until it no longer makes any difference, which was 35 macro-cycles for galectin-3C. The X-ray data can also be included to make a joint X-ray/neutron refinement, if the protein is not fixed. But this limits the number of structures from the PDB that can be used and this approach has not yet been tested. In conclusion, the best performing optimization methods and parameters for the refinement were to use the RSCC and Ready Set on a fixed protein in reciprocal space with 15 macro-cycles and neutron data only.

3.2 Automated Water Orientation Script

After having selected a proper quantitative quality measure (RSCC), a proper threshold (0.81) and a proper refinement procedure, the next step was to develop an approach to automatically and systematically re-orient the water molecules. We developed two different scripts to this aim: Script Ref, which refines after each orientation cycle, and Script No-ref, which does not refine. The results in Table 5 show that the two scripts gave similar results, probably because the generated structures were refined with phenix.refine after the best water orientations were found. Naturally, Script No-ref was significantly faster tha Script Ref. As for the maximum number of orientation cycles for the script, 100 proved to be the best in terms of time and results, although 50 cycles gave nearly as good results.

Table 5: The performance of the two re-orientation scripts on galectin-3C regenerated by Ready Set, showing the timing (on a single computer) and the number of water molecules with RSCC values above the threshold and the two *R*-factors.

Orientation cycles	10	50	100	150
Time (s, Script No-ref)	420	720	1260	1740
Time (s, Script Ref)	4440	17880	23280	34200
RSCC above 0.81 (Script No-ref)	85	90	92	90
RSCC above 0.81 (Script Ref)	86	87	90	90
R_{work} (Script No-ref)	0.1721	0.1721	0.1720	0.1721
R_{work} (Script Ref)	0.1723	0.1708	0.1711	0.1721
R_{free} (Script No-ref)	0.2276	0.2261	0.2283	0.2273
R_{free} (Script Ref)	0.2312	0.2270	0.2308	0.2273

Table 6 shows the performance times of Script No-ref on the three proteins using 100 cycles and the RSCC threshold of 0.81. For comparison, the timing of phenix.refine with 15 macro-cycles and fixed proteins in reciprocal space with neutron data only is also presented.

Table 6: The performance times of Script No-ref and phenix.refine on the three proteins.

Protein	Galectin-3C	Rubredoxin	Pyrophosphatase
Time (s, script)	1260	1800	2820
Time (s, phenix.refine)	1930	880	5082
Number of water molecules	110	149	385

Table 7 presents the *R*-factors and number of water molecules with RSCC values above the threshold resulting from our optimized strategy with phenix.refine (neutron data only, with the protein fixed, in reciprocal space for 15 macro-cycles). We compare the results obtained with or without running our water-orientation script. The corresponding data for the deposited structure are also included.

Table 7: Comparison of the deposited structures and the structures after refinement with or without the water-orientation script for the three proteins. For each protein, the *R*-factors and the number of properly oriented water molecules are given (i.e. with RSCC > 0.81 for the two structures after refinement, but properly oriented according to the qualitative assessment for the deposited structure, and total number of water in brackets).

Protein	Galectin-3C			Rubredoxin			Pyrophosphatase		
Deposited/Without/With Script	Deposited	Without	With	Deposited	Without	With	Deposited	Without	With
RSCC above 0.81	70 (110)	86	92	51 (149)	49	55	235 (385)	351	363
R_{work}	0.1680	0.1713	0.1720	0.1990	0.2030	0.1998	0.2390	0.1912	0.1877
R_{free}	0.2110	0.2245	0.2283	0.2370	0.2375	0.2495	0.2520	0.2847	0.2932

Deuterium atoms with RSCC values below 0.61 (0.2 below the threshold of 0.81) were removed by the script, but all hydrogens in galectin-3C and pyrophosphatase had RSCC values above 0.61. However, rubredoxin had 23 hydrogens with RSCC values below 0.61 which were deleted. The script improved all structures in terms of RSCC values of the water molecules and of the R_{work} values (essentially unchanged for Galectin-3C), but R_{free} was deteriorated for all structures. 55 correctly oriented water molecules out of 149 for rubredoxin is rather disappointing although with manual orientation, only 51 are identified. Therefore, we doubled the number of orientation cycles to 200 for Script No-ref and then proceeded with the same refinement settings used in Table 7. However, the improvement was insignificant. Furthermore, R_{free} increased significantly for pyrophosphatase after using the best optimization strategy and script. The reason could be that the optimization strategy and script were made primarily for galectin-3C whose structure has been determined with X-ray crystallography, whereas for the other two proteins, there is no X-ray structure. Also, keeping in mind that the X-ray data was removed from galectin-3C, the increase in R_{work} is understandable. The increase partly arises from the script but also from phenix.refine as indicated in Table 7.

The script improved the number of properly oriented water molecules for all proteins but the protein structure for pyrophosphatase (whose structure was determined using only neutron crystallography) degraded as shown in the R_{free} in Table 7. By not fixing the protein for pyrophospahatase, the *R*-factors could be improved because the protein structure would then also be considered with the refinement of the water molecules. There may also be other parameters that affect the results since the structure for rubredoxin did not degrade like pyrophosphatase, despite the lack of an X-ray structure. One could be that pyrophosphatase has 385 water molecules, and if each water molecule increases the R_{free} by a slight amount, then 385 water molecules will have a more significant increase. This explains the poor R_{free} results in Table 7.

The script may also be improved if it could consider alternate water conformations, which would be a good improvement to phenix.refine as well. Furthermore, since the evaluation of the water orientations was qualitative, the results may be biased. This bias would arise from the script that optimizes the thresholds as it has both the quantitative and qualitative data as inputs. Since Script No-ref and Script Ref both use the generated threshold, the resulting structures are most likely biased by an unknown amount. However, the systematic error in the evaluation has been kept consistent as all qualitative evaluations have been performed by the same person.

For galectin-3C, the deposited structure had 70 qualitatively good water molecules. After removing the deuteriums and re-adding them with Ready Set without any refinement, there were only 29 qualitatively good water molecules. Then after applying Script No-ref and the best optimization strategy, 85 water molecules had qualitatively good orientations, illustrating the strength of the script. Figure 8 shows an example of the improvement of water molecules in galectin-3C before and after using the script.

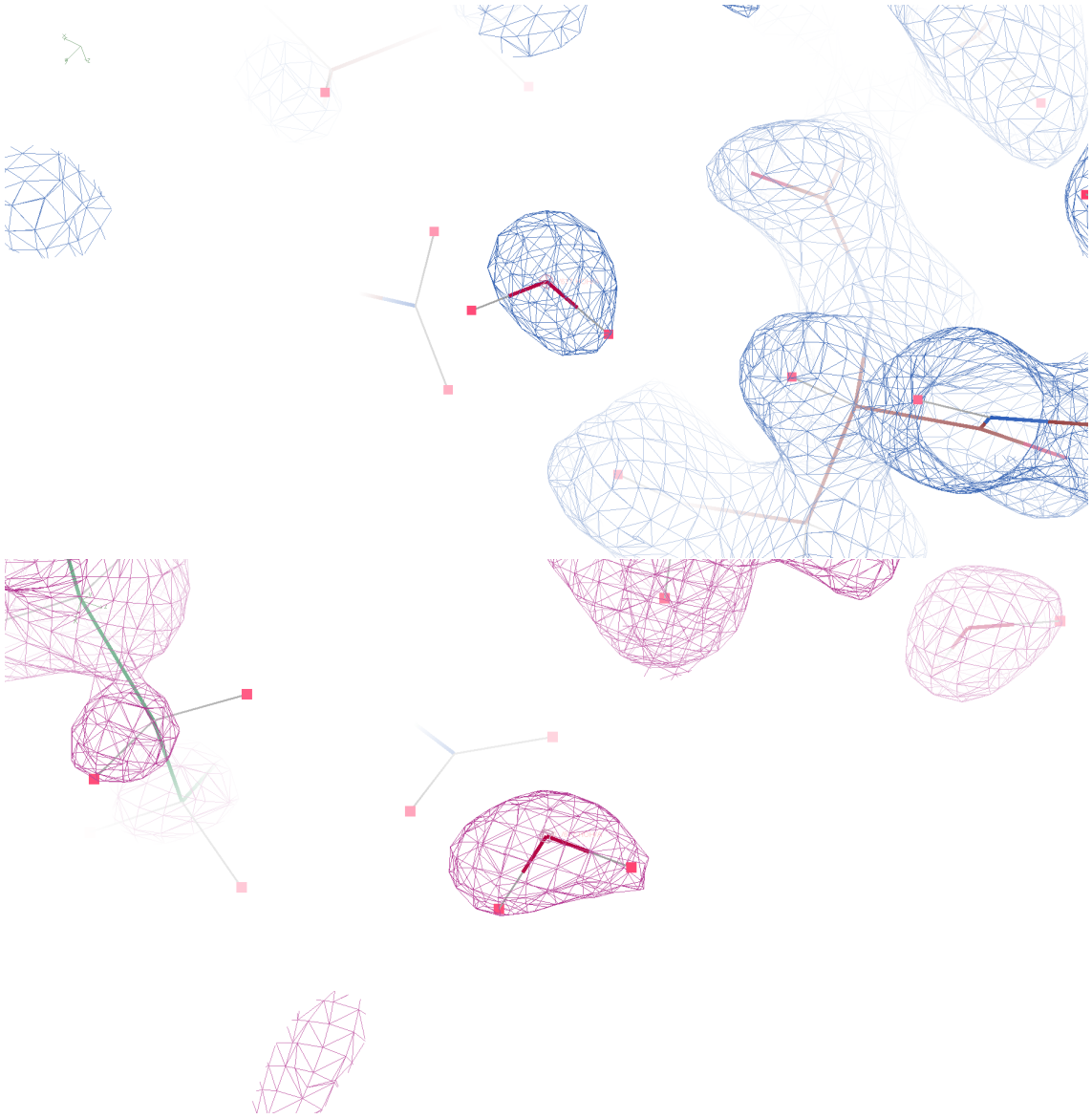


Figure 8: Example of the improvement of the orientation of a water molecule (number 617 in the crystal structure) with our script. Top: starting structure after using the optimized strategy from phenix.refine, with the neutron density given by $2F_o - F_c$ neutron map ($\sigma = 1.0$), in blue. Bottom: final structure obtained with the script and the optimized strategy, with the neutron density given by $2F_o - F_c$ neutron map ($\sigma = 1.0$), in purple.

4 Conclusion

We have automated the hydrogen placement of water molecules in neutron structures using a script that re-orients water molecules with the help of phenix.refine. The script worked best with 100 orientation cycles with no refinement after each cycle and an RSCC threshold of 0.81. We found that the best performing settings with phenix.refine was with neutron data only, with the protein fixed in reciprocal space and 15 macro-cycles. When using Ready Set, the script and phenix.refine with those settings on the three proteins galectin-3C, rubredoxin and pyrophosphatase, the quantitative data agreed best with the qualitative results. However, the strategy worked best on galectin-3C since the strategy was developed for that protein.

Acknowledgements

A big thank you goes to the Department of Theoretical Chemistry for the warm welcome, and special thanks to Octav Caldararu for his crystal clear instructions and assistance. I also want to thank Ulf Ryde for showing his care and his much valued inputs on this project.

References

[1] Jeremy Mark Berg, John L Tymoczko, and Lubert Stryer. Biochemistry, 2012.

[2] Yaakov Levy and José N Onuchic. Water mediation in protein folding and molecular recognition. *Annu. Rev. Biophys. Biomol. Struct.*, 35:389–415, 2006.

[3] Carla Mattos. Protein–water interactions in a dynamic world. *Trends in biochemical sciences*, 27(4):203–208, 2002.

[4] Eva Nittinger, Nadine Schneider, Gudrun Lange, and Matthias Rarey. Evidence of Water Molecules - A Statistical Evaluation of Water Molecules Based on Electron Density. *Journal of chemical information and modeling*, 55(4):771–783, 2015.

[5] Piotr Neumann and Kai Tittmann. Marvels of enzyme catalysis at true atomic resolution: distortions, bond elongations, hidden flips, protonation states and atom identities. *Current opinion in structural biology*, 29:122–133, 2014.

[6] Jain JL, Jain Sunjay, and Jain Nitin. *Fundamentals of biochemistry*. S. Chand Publishing, 2004.

[7] Gale Rhodes. *Crystallography made crystal clear*. Academic Press, 2 edition, 2006.

[8] Yang Zhang. Progress and challenges in protein structure prediction. *Current opinion in structural biology*, 18(3):342–348, 2008.

[9] Pavel V Afonine, Ralf W Grosse-Kunstleve, Nathaniel Echols, Jeffrey J Headd, Nigel W Moriarty, Marat Mustyakimov, Thomas C Terwilliger, Alexandre Urzhumtsev, Peter H Zwart, and Paul D Adams. Towards automated crystallographic structure refinement with phenix. refine. *Acta Crystallographica Section D: Biological Crystallography*.

[10] Pavel V Afonine, Marat Mustyakimov, Ralf W Grosse-Kunstleve, Nigel W Moriarty, Paul Langan, and Paul D Adams. Joint x-ray and neutron refinement with phenix. refine. *Acta Crystallographica Section D: Biological Crystallography*, 66(11):1153–1163, 2010.

[11] EI Howard, R Sanishvili, RE Cachau, A Mitschler, B Chevrier, P Barth, V Lamour, M Van Zandt, E Sibley, C Bon, et al. Ultrahigh resolution drug design i: details of interactions in human aldose reductase–inhibitor complex at 0.66 å. *Proteins: Structure, Function, and Bioinformatics*, 55(4):792–804, 2004.

[12] Kosei Kawasaki, Hidemasa Kondo, Mamoru Suzuki, S Ohgiyai, and Sakae Tsuda. Alternate conformations observed in catalytic serine of bacillus subtilis lipase determined at 1.3 å resolution. *Acta Crystallographica Section D: Biological Crystallography*, 58(7):1168–1174, 2002.

[13] Varley F Sears. Neutron scattering lengths and cross sections. *Neutron news*, 3(3):26–37, 1992.

[14] William B O'Dell, Annette M Bodenheimer, and Flora Meilleur. Neutron protein crystallography: A complementary tool for locating hydrogens in proteins. *Archives of Biochemistry and Biophysics*, 602:48–60, 2016.

[15] Jarjis Habash, James Raftery, Rachel Nuttall, Helen J Price, Clive Wilkinson, JR Helliwell, et al. Direct determination of the positions of the deuterium atoms of the bound water in concanavalin A by neutron Laue crystallography. *Acta Crystallographica Section D: Biological Crystallography*, 56(5):541–550, 2000.

[16] Z Richard Korszun. [15] neutron macromolecular crystallography. In *Methods in enzymology*, volume 276, pages 218–232. Elsevier, 1997.

[17] Paul Emsley, Bernhard Lohkamp, William G. Scott, and Kevin Cowtan. Features and development of coot. *Acta Crystallographica Section D - Biological Crystallography*, 66:486–501, 2010.

[18] Schrödinger, LLC. The PyMOL molecular graphics system, version 1.8. November 2015.

[19] EDSTATS. `http://www.ccp4.ac.uk/html/edstats.html`. Accessed 2019-07-05.

[20] Online Dictionary of Crystallography: Real-space correlation coefficient. `http://reference.iucr.org/dictionary/Real-space_correlation_coefficient`. Accessed 2019-07-05.

[21] Anne Louise Morris, Malcolm W MacArthur, E Gail Hutchinson, and Janet M Thornton. Stereochemical quality of protein structure coordinates. *Proteins: Structure, Function, and Bioinformatics*, 12(4):345–364, 1992.

[22] R factor. `http://reference.iucr.org/dictionary/R_factor`. Accessed 2019-07-14.

[23] Helen M. Berman, John Westbrook, Zukang Feng, Gary Gilliland, T. N. Bhat, Helge Weissig, Ilya N. Shindyalov, and Philip E. Bourne. The Protein Data Bank. *Nucleic Acids Research*, 28(1):235–242, 01 2000.

[24] Schrödinger Release 2019-2: Maestro, Schrödinger, LLC, New York, NY, 2019.