Automated Orientation of Water Molecules in Neutron Structures of Proteins

Axl Eriksson[1], Octav Caldararu[1], Ulf Ryde[1] and Esko Oksanen*[2]

[1]Department of Theoretical Chemistry, Lund University, Chemical Centre, P. O. Box 124, SE-221 00 Lund, Sweden

[2] European Spallation Source ESS ERIC, P. O. Box 176, SE-221 00 Lund, Sweden

Abstract- The structure and function of proteins are strongly affected by the surrounding solvent water, *e.g.* by hydrogen bonds and the hydrophobic effect. These interactions depend not only on the position, but also on the orientation of the water molecules around the protein. Therefore, it is often vital to know the detailed orientations of the surrounding ordered water molecules. Such information can be obtained by neutron crystallography. However, it is tedious and time-consuming to determine the correct orientation of every water molecule in the structure (there are typically several hundreds of them), which today is done by manual evaluation. Here, we have developed a method that reliably automatizes the orienting of the water molecules in a simple and relatively fast way. First, we selected a quantitative quality measure, the real-space correlation coefficient, and a threshold that allows us to identify the water molecules that are clearly oriented. Second, we have optimised the refinement procedure by varying the refinement methods and parameters, thus finding settings that yielded the best results in terms of time and performance. It turned out to be favourable to employ only the neutron data and a fixed protein structure when reorienting the water molecules. Third, we have developed a method that identifies and reorients inadequately oriented water molecules systematically and automatically. The method has been tested on three proteins: galectin-3C, rubredoxin and inorganic pyrophosphatase and we show that it yields improved orientations of the water molecules for all three proteins in a shorter time than manual model building. It also led to an increased number of hydrogen bonds involving the water molecules for all proteins.

Keywords: neutron crystallography; automatic water orientation; hydrogen bonds; real-space correlation coefficient

1. Introduction

To understand the function of proteins, to design enzymes with new or improved catalytic functions or to develop potent inhibitors or drug molecules, structural information – most commonly from X-ray crystallography – is indispensable. The function is not determined by

the protein structure alone, but also by the surrounding solvent, through the hydrophobic effect and the formation of specific hydrogen bonds (Levy & Onuchic, 2006). The positions and orientation of the surrounding water molecules also affect the dynamics of the protein (Mattos, 2002) and are therefore, also of major interest (Nittinger *et al.*, 2015)

Unfortunately, since the hydrogen atoms contain only a single electron, it is problematic to precisely locate them in the electron density maps from X-ray crystallography. At ultra-high resolution (lower than 0.7 Å), the positions of some hydrogen atoms can be reliably determined (Neumann & Tittmann, 2014; Howard *et al.*, 2004; Blakeley *et al.*, 2015) but this is limited to particularly well-ordered atoms. This is a limitation in attempts to understand protein function, because the hydrogen positions determine the direction and polarity of hydrogen bonds. Because of these limitations in X-ray crystallography, the major source of structural information of hydrogen atoms comes from neutron crystallography. Neutrons are scattered by the nuclei (Sears, 1986) and deuterium scatters comparably to other atoms in proteins. However, neutron crystallography requires larger crystals and much longer exposure times than X-ray crystallography. Moreover, the ^{1}H hydrogen atoms need to be replaced by deuterium atoms (^{2}H or D) to limit the incoherent scattering background. Another complication is that the ^{1}H isotope of hydrogen has a negative scattering length, which can cancel the scattering contribution of the heavy atom they are bound to. D, on the other hand, has a positive scattering length, comparable in magnitude to that of C, N, O and S. 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 (O'Dell *et al.*, 2016). 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 (Afonine *et al.*, 2010).

Even if the D atoms are visible in neutron structures, it is a tedious procedure to go through every D atom in the structure to ensure that it is located in the best possible position. Crystallographic refinement can locally optimise the position around a given starting point, but it can often be trapped at local minima, so several starting positions need to be tested and in practice, the crystallographer ideally should visually judge every D atom in the structure. This is especially complicated for water molecules, where the D atoms can be placed anywhere on a sphere around the O atom (for protein residues, the position is either unambiguously determined by the structure or the search can be restricted to a circle). Therefore, the positions of water D atoms are often less well defined than other atoms in neutron structures. Phenix has a utility to solve this problem by performing a real-space refinement at sites with rotatable X-H/D bonds (Afonine *et al.*, 2010). However, this is not documented for water molecules, only

for Ser/Thr/Tyr OH groups, which are easier to solve. Furthermore, real-space refinement can often be stuck in a local minimum.

In this study, we try to address this problem by developing an automatic and systematic procedure to determine the orientation of water molecules (*i.e.* the positions of the D atoms) in neutron structures. This procedure is applied to three crystal structures, galectin-3C, rubredoxin and inorganic pyrophosphatase. We show that the approach increases the number of reasonably oriented water molecules and improves the fit of the model to the data.

2. Methods

2.1. Crystal structures

Three crystal structures were used as test cases in this study. The first was a 1.7 Å neutron structure of galectin-3C in complex with lactose (PDB ID: 6EYM) (Manzoni *et al.*, 2018) with 110 water molecules. The second was a 1.05 Å neutron structure of rubredoxin (PDB ID: 4AR3), an electron-transport protein (Cuypers *et al.*, 2013) with 149 water molecules. The third was a 2.3 Å neutron structure of inorganic pyrophosphatase (PDB ID: 5TY5) (Inoguchi *et al.*, 2017) with 385 water molecules. For all structures, the coordinates, atomic displacement factors (ADPs), occupancies, as well as the neutron and X-ray (only for galectin-3C) structure factors were obtained from the Protein Data Bank (Berman *et al.*, 2000).

All refinements were performed with the software phenix.refine (Adams *et al.*, 2010; Afonine *et al.*, 2012) with default settings, unless otherwise stated. The D atoms' coordinates, occupancies and ADPs were refined individually. H and exchanged H/D sites were treated as such and also refined individually. For galectin-3C, we tested to use either only the neutron data or the joint X-ray/neutron data.

The initial orientations of the water molecules in the crystal structures studies were obtained in three different ways. The first was to use coordinates in the deposited structures. If these were used, no initial refinement was performed. In the other two approaches, the deposited D atoms were removed and D atoms were added by either the ReadySet (Afonine *et al.*, 2012) or Maestro software (Schrödinger, 2019). The ReadySet is part of the Phenix software suite (Liebschner *et al.*, 2019), whereas Maestro is a separate software, designed for computational protein modelling. Both programs add D atoms to water molecules using their own energy-based algorithms, considering the local surroundings and possible hydrogen-bond networks. After adding D atoms with either software, a refinement was performed before doing any quantitative or qualitative evaluation.

In several cases, we evaluated qualitatively whether the water molecules had an adequate orientation – in some water molecules one or even both hydrogens can be disordered and hence

not visible in the maps although the oxygen is ordered: For every water molecule in the structure, we studied visually the water orientations in the nuclear and electron density maps in Coot. Both the $2mF_o - DF_c$ maps and $F_o - F_c$ maps were used for the qualitative evaluation. Furthermore, we used maps generated both from a model with and without water molecules, to assess the effect of model bias. For a fully ordered water in a correct and well-defined orientation, the D atoms are placed inside the neutron density around the water molecule and the density has a bent (banana-like) shape, enclosing the two D atoms. Figure 1a shows an example of a correct orientation of a water molecule, with both D atoms inside the density. Its O atom also participates in a hydrogen bond with the protein. Conversely, in improperly orientated water molecules, the D atoms are outside the nuclear density (Figure 1b). This evaluation was performed for the three deposited neutron structures, as well as for the same structures with the D atoms removed and then automatically added and refined with various strategies. These structures, in total ten, are listed in Table 1, showing the D-atom addition method and the refinement strategy.

We tested three metrics to quantitate the fit of the water D atoms to the data. The first was the real-space Z-difference (RSZD) score (Tickle, 2012), calculated by EDSTATS. It is considered the best measure to evaluate locally the goodness-of-fit for a group in a crystal structure and essentially evaluates the largest and smallest values in the $mF_o - DF_c$ map around the group of interest. The second was the real-space correlation coefficient (RSCC) between the $2mF_o - DF_c$ and the F_c maps. The $2mF_o - DF_c$ map was calculated both with water molecules included in the model and without (omit map), to assess the model bias. The RSCC can be calculated both by Phenix or EDSTATS (Afonine *et al.*, 2012; Tickle, 2012). The statistical significance of calculating RSCC on a small number of atoms might be questionable as the calculations use only ~30 grid points per map (depending on the resolution). To test if the RSCC is stable, we calculated water RSCC values in the deposited galectin-3C structure at five different grid spacings (from 0.8 Å to 0.25 Å). The average RSCC standard deviation is only 0.008, showing that RSCC is stable and most likely significant. The third metric we used was the difference of the ADPs of the two D atoms of the water molecules to the global mean ADP after refinement. Atomic groups with high ADP compared to the global mean ADP may suggest errors in the model. This has been studied for ligands (Deller & Rupp, 2015) but is logically equivalent for water molecules. However, the use of ADP values as quality metrics is somewhat reduced by the fact that they are normally refined using similarity restraints.

2.2. Script to systematically reorient water molecules

We constructed a script to systematically test different positions of the D atoms of the water molecules, called NWO (Neutron Water Orientation). NWO was written in Python and requires Biopython 1.70 or newer and Phenix 1.14 or newer. The script is freely available on Github (https://github.com/OCald/NeutronWaterOrientation).

NWO requires starting coordinates for both D atoms of all water molecules. They were obtained by ReadySet in this study. It takes as input arguments the number of cycles to run (n_{rot}) and an RSCC threshold (the minimum RSCC value for which a water molecule is considered to be in a good orientation). Optimal values of these parameters are discussed in the Results and Discussion section.

NWO first calculates the RSCC values for all water D atoms in the current structure. The RSCC calculation is done in Phenix and can be performed using two different $2mF_o - DF_c$ maps, either calculated starting from the full model or calculated starting from a model where water molecules are excluded to avoid model bias. For all water molecules with at least one D atom with a RSCC value below a threshold, NWO rotates the first D atom in the water molecule (D_1) an angle α_1 around the O–D_2 axis (where O and D_2 are the oxygen and the other D atom of the water molecule). Then, D_2 is rotated an angle α_2 around the O–D_1 axis. If the RSCC value of D_i (i = 1 or 2) is less than 0.2 below the RSCC threshold, α_i = 5º, otherwise α_i = 10º. Thus, the coordinates of the O atom, the D_1–O–D_2 angle and the O–D_i bond lengths are fixed to the values used in ReadySet (106.8º, and 0.96 Å). Next, the structure is optionally refined in reciprocal space with one macrocycle, using neutron data only and keeping the protein fixed. Finally, new RSCC values are calculated for the two D atom atoms. This is repeated until both RSCC values are above the threshold or a maximum number of rotations (n_{rot}) have been performed. NWO maximizes the RSCC of the entire water molecule and returns the structure for which the lowest of $RSCC_1$ and $RSCC_2$ (*i.e.* the RSCC of each D atom) is largest. If any of the final $RSCC_i$ values is still less than the threshold minus 0.2, that D atom is deleted from the model, in order to correctly model water molecules in which one or both D atoms are disordered while the oxygen atom is ordered.

2.3. Hydrogen bond analysis

The number of hydrogen bonds formed by water molecules was calculated using the CPPTRAJ module in AmberTools 19. The *hbond* command with default parameters was used, defining only water deuterium atoms as hydrogen-bond donors.

3. Results and Discussion

3.1. A quantitative quality measure

We have developed an automatic method to orient water molecules in neutron crystal structures of proteins based on a systematic test of a large number of positions for each water molecule in the structure. However, for such an approach, it is necessary to have a metric 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 metric.

To this end, we need a benchmark to compare with, *i.e.* a set of manually checked and confirmed water molecule orientations in a neutron crystal structure. We observed that not all the water molecules in the deposited structures are uncontroversially modelled (an example is shown in Figure 2). Therefore, we decided to manually evaluate each water molecule in the structures and mark for each water molecule whether it is ordered and correctly oriented or not. This was done for water molecules in the deposited structures, but also for structures with water D atoms deleted and then added again, either with the ReadySet or Maestro software (shown in Table 1). The qualitative evaluation of the deposited structure of galectin-3C, suggested that 39 of the 109 water molecules (36%) were in inadequate or improvable orientations. The corresponding numbers for the deposited structures of rubredoxin and pyrophosphatase were 98 of 149 (66%) and 150 of 385 (39%). For the galectin-3C structure with D atoms added by ReadySet or Maestro, 81 (74%) and 84 (77%) water molecules were poorly oriented, respectively, according to our qualitative evaluation. This shows that automatic H-addition algorithms perform rather poorly and they cannot replace experimental information from neutron diffraction. This is not unexpected, as neither uses any experimental data. It is also notable that ReadySet and Maestro give results of a similar quality.

Next, we calculated three common metrics for these structures, RSZD, RSCC and the ADP difference. Each metric gives a value for each D atom in all water molecules. Histograms for the distribution of the three metrics for the correctly and improperly oriented water molecules (according to the qualitative assessment) are shown in Figure 3. It can be seen that the RSZD score does not distinguish between the correct and improper orientations. In fact, there seems even to be some tendency that the improperly oriented D atoms give lower RSZD values than the good ones. For the ADP differences the results are more promising: the distribution for the water molecules with a correct orientation is more concentrated towards lower values, with a peak somewhat below the average (66), whereas the distribution is biased towards higher values for those with an improper orientation. However, there is a second peak at very high ADP differences for some correctly oriented water molecules.

The results for the RSCC score are even better: the histogram for the correctly oriented water molecules is strongly biased towards high values, whereas those with an improper orientation

have a wider distribution, peaking at somewhat lower values (although there is a quite large overlap). Therefore, we selected RSCC as our quality metric.

RSCC values can be calculated with both the Phenix and Edstats software. We chose to use Phenix because the difference between the two methods was minor and Phenix is already used for the refinement making the implementation simpler. Likewise, for placement of hydrogens, ReadySet from Phenix yielded results of a comparable quality as Maestro, so we chose to use ReadySet for hydrogen placement so that all calculations can be performed by the same software.

Next, we identified a threshold value for RSCC to determine whether a D atom is in a correct position or not. This was done by maximising the total number of correct predictions of RSCC, both for correctly and improperly oriented water molecules, using the manual evaluation as a benchmark. This gave a threshold of 0.89 for galectin-3C alone and 0.81 when all data from galectin-3C, rubredoxin and pyrophosphatase were combined. The threshold was based on the water molecules with two deuterium atoms in all manually evaluated structure listed in Table 1, in total 1480 water molecules.

240 water molecules had a RSCC ≥ 0.81 for both D atoms although the manual assessment suggested that they are improperly oriented or at least improvable, whereas 87 water molecules had RSCC < 0.81 for both D atoms, although they were evaluated as correctly oriented, resulting in 22% false positives and negatives. For 847 atoms, the RSCC results agree with the manual inspection that the orientation is correct and for 362 atoms the two evaluations agree that the orientation is improper.

A different RSCC threshold was found when calculating RSCC using omit $2mF_o - DF_c$ maps, using the same protocol described above. The threshold in this case was 0.69, which gave 30% false positives and negatives. Thus, using the omit $2mF_o - DF_c$ maps for calculating RSCC is slightly less reliable than calculating RSCC from the regular maps, with the water molecules included in the model.

3.2. Refinement settings

Our automatic method for water orientation requires a large number of refinements, so we tried different refinement settings in order to obtain a protocol that would yield a proper compromise between accuracy and speed. The joint X-ray/neutron refinement (for galectin-3C) was quite time-consuming and therefore we tested whether omission of the X-ray data would speed up the calculations Although the positions of the D atoms are almost entirely determined by the neutron data, the RSCC values decreased when the non-hydrogen atoms were not fixed. This was likely due to the lower data-to-parameter ratio. Neutron-only refinement with the

protein fixed increased the RSCC values significantly at the cost of slowing down the refinement slightly. However, the improvement was significant so we decided to continue with the fixed-protein strategy. With the protein fixed, the refinement was still appreciably faster than with the X-ray data, but the RSCC values were slightly worse. However, we judged the improvements outweighed the computational cost based on the data in Table 2. It can be seen that the neutron R_{free} value increases significantly when excluding the X-ray data, which is expected since less data are used to improve the model.

Table 3 presents all refinement strategies tested and their results for the neutron data. The best results were obtained with reciprocal-space refinement or with real-space refinement combined with reciprocal-space refinement, whereas real-space-only refinements yielded poor results.

Next, we investigated what number of refinement cycles was most effective in terms of result and time. The results in Table 4 show that 15 macrocycles were optimal, at least judged by R_{free}/R_{work} and the number of correctly oriented water molecules (n_{wat}).

In conclusion, our preferred approach was to use RSCC as the quality measure, to add D atoms with ReadySet and to run the refinement in reciprocal space with 15 macrocycles, using only neutron data and keeping the protein fixed. The values of the different parameters found herein are optimal for the proteins studied, but we can of course not guarantee that they are optimal also for other systems.

3.3. Automated water orientation

After having selected a metric (RSCC), a reasonable threshold (0.81 for regular maps and 0.69 for omit maps) and a proper refinement procedure, the next step was to automatically and systematically reorient the water molecules. We developed a script for this purpose, called NWO (Neutron Water Orientation), described in the Methods section. We tested three variants of NWO, without refinement after each orientation cycle (NWO-default), with refinement after each orientation cycle (NWO-refine) or using omit maps calculated from models without the water molecules to be oriented included (NWO-omit). First, we investigated if refinement in each reorientation cycle is necessary. The results in Table 5 show that the NWO-default and the NWO-refine variants gave similar results, probably because for the variant without refinement in each iteration, the final structures were refined after the best water orientation was found. Naturally, the NWO-default variant was much faster and therefore preferred. It can also be seen that the maximum number of orientation cycles should be set to 50 or 100. Timings of the NWO script and refinement procedures for all three proteins are shown in Table 6. NWO-default does not take longer than the refinement procedure itself in two of the three cases, thus showcasing

that it can be used alongside refinement without a large increase in computing time. Furthermore, the script timing is linear with respect to the number of water molecules present in the system.

Next, we compared the structures obtained using the reorientation script with $2mF_o - DF_c$ maps calculated without (NWO-omit) or with each water molecule included (NWO-default), performing 100 iterations of reorientation for each system. NWO-omit should avoid model bias, but on the other hand the RSCC threshold is less reliable. Table 7 shows the R values and n_{wat} resulting from our reorientation strategies. We also compare the NWO results with both the deposited structure and the results after refinement, but without running NWO, to separate the effect of NWO from the refinement procedure. It can be seen that NWO-default improved all structures in terms of n_{wat}. For galectin-3C, the deposited structure had 70 correctly oriented water molecules (out of 110 in total). After removing the D atoms and adding them again with ReadySet, but without any refinement, there were only 29 correctly oriented water molecules, which increased to 81 after refinement. With NWO-default, this increased further to 92 water molecules, illustrating the usefulness of the script. Figure 4 shows an example of the improvement of a water molecule in galectin-3C before and after using NWO. Likewise, for pyrophosphatase, the number of correctly oriented water molecules also increased from 235 in the deposited structure to 363 (out of 385) after running our script. For rubredoxin, the result number of correctly oriented water molecules also improved with our script, from 51 to 55 (49 for water molecules added by ReadySet), but this was only one third of the total number of water molecules, 149. The number did not change significantly if we doubled the number of orientation cycles to 200.

When using NWO-omit, the number of correctly oriented water molecules in galectin-3C was only 75, more than in the deposited structure but less than in the structure obtained with the default variant. Similarly, the rubredoxin structure from NWO-omit had 51 correctly oriented water molecules, the same number as in the deposited structure, whereas the pyrophosphatase structure contained 334 correctly oriented water molecules, an improvement compared to the deposited structure, but 29 fewer than the structure from NWO-default.

For galectin-3C, NWO-default hardly changed R_{work}, whereas the refinement procedure increased it slightly (from 0.168 to 0.171). A similar increase was also seen in R_{free} for both the refinement procedure and the reorientation script (from 0.211 to 0.224 and 0.228). NWO-omit resulted in slightly higher R_{work} and R_{free} values (0.179 and 0.230), suggesting the model obtained this way is of poorer quality. The differences in the values of the R factor may stem from the fact that the deposited structure was jointly refined with X-ray data, whereas we only perform neutron refinement.

For rubredoxin, with its appreciably higher resolution (1.05 Å), there was a similar increase in R_{work} caused by the refinement procedure (0.199 to 0.203), but no change with NWO-default. R_{free} changed neither by the refinement procedure, nor by NWO-default. Similar to galectin-3C, using omit maps increased the R_{work} and R_{free} values slightly (0.211 and 0.243). The increase in the R values for these two cases can be attributed to the less reliable RSCC threshold that could result in more false positives, i.e. a higher number of poorly oriented water molecules.

For inorganic pyrophosphatase, R_{work} decreased strongly by our refinement procedure (from 0.239 to 0.191). This was accompanied by similar increase in R_{free} (from 0.252 to 0.285), indicating a strong overfitting. It is not fully clear to us why we get such a large overfit for this structure. In fact, reciprocal-space refinement with default settings of Phenix for the deposited structure did not reproduce the deposited R_{work} and R_{free} values, but gave results closer to those obtained with our refinement strategy (0.177 and 0.281, respectively, i.e. with an even larger overfit). This may be due to differences in e.g. bulk-solvent or ADP models used by us and the original authors. The reorientation script continued this trend, but to a much smaller extent: R_{work} decreased by 0.003 to 0.188 and R_{free} increased by 0.008 to 0.293. This may be caused by the lower resolution (2.3 Å) and higher number of water molecules: Even if each water molecule increases R_{free} by a small amount, 385 water molecules may lead to a significant increase. Interestingly, the structure obtained with NWO-omit showed a decrease in the overfitting, with a similar R_{work} = 0.192 but a lower R_{free} = 0.278 (i.e. less than without the reorientation). This shows that for low-resolution structures, it may be advantageous to use omit maps for the reorientation script, whereas for our medium and high-resolution structures, NWO-default gives the best results.

D atoms with RSCC values more than 0.2 below the threshold were removed by the script. For NWO-default, no D atoms had so low values in galectin-3C and pyrophosphatase, but for rubredoxin, 23 water D atoms with low RSCC values were deleted. In contrast, in NWO-omit one D atom was deleted for galectin-3C, none for pyrophosphatase and 113 D atoms with low RSCC values were deleted for rubredoxin.

3.4. Hydrogen-bond analysis

Water molecules in correct orientations are expected to form hydrogen bonds with protein atoms or other water molecules. Thus, if the script work properly it should improve the water hydrogen-bonding network. Therefore, we counted the number of hydrogen bonds formed by water molecules with their deuterium atoms as donors, as only the deuterium atoms move during the reorientation script. This also avoids double-counting hydrogen bonds formed

between two water molecules. The number of hydrogen bonds in the various structures is shown in Table 8.

It can be seen that the number of hydrogen bonds improves after using NWO compared to the input ReadySet structures for all three systems, both for NWO-omit and NWO-default. Importantly, the structure obtained with NWO-default also shows more hydrogen bonds than the deposited structure. For pyrophosphatase, the hydrogen bonding network is drastically improved after reorientation compared to the deposited structure, 144 hydrogen bonds compared to 84. This shows that NWO works properly and actually leads to an improvement of the water structures. NWO-default gives more hydrogen bonds than NWO-omit for all three proteins. In fact, NWO-omit gives the same number or slightly fewer hydrogen bonds than the deposited structures for galectin-3 and rubredoxin.

Finally, it should be mentioned that we have checked that none of the reoriented water molecules gives rise to any steric clashes for any of the three proteins after employing NWO.

4. Conclusions

We have automated the hydrogen placement of water molecules in neutron structures using a script, NWO, that systematically reorients water molecules and decides which are correctly oriented using the RSCC as a validation metric. Our approach was tested on three proteins, galectin-3C, rubredoxin and pyrophosphatase, and we show that we increase the number of properly oriented water molecules and obtain improved or unchanged R_{work} for all proteins. Moreover, the number of hydrogen bonds involving the water molecules is increased for all three proteins. However, for two proteins, R_{free} increased slightly, indicating some degree of overfitting. This can be partially addressed by removing the water molecules to be reoriented from the $2mF_o - DF_c$ map, which improved the results for the structure with the lowest resolution (2.4 Å), but gave worse results for the other two proteins. In the future, we will try to develop an improved quantitative quality metric, combining several measures and developing new quantities. In addition, we will address the problem with alternate conformations and non-unity occupancy of water molecules.

Figure 1 Example of water molecules with a good orientation (left) and a poor orientation (right). The $2mF_o - DF_c$ nuclear-density omit maps are shown in blue at $\sigma = 1.0$.

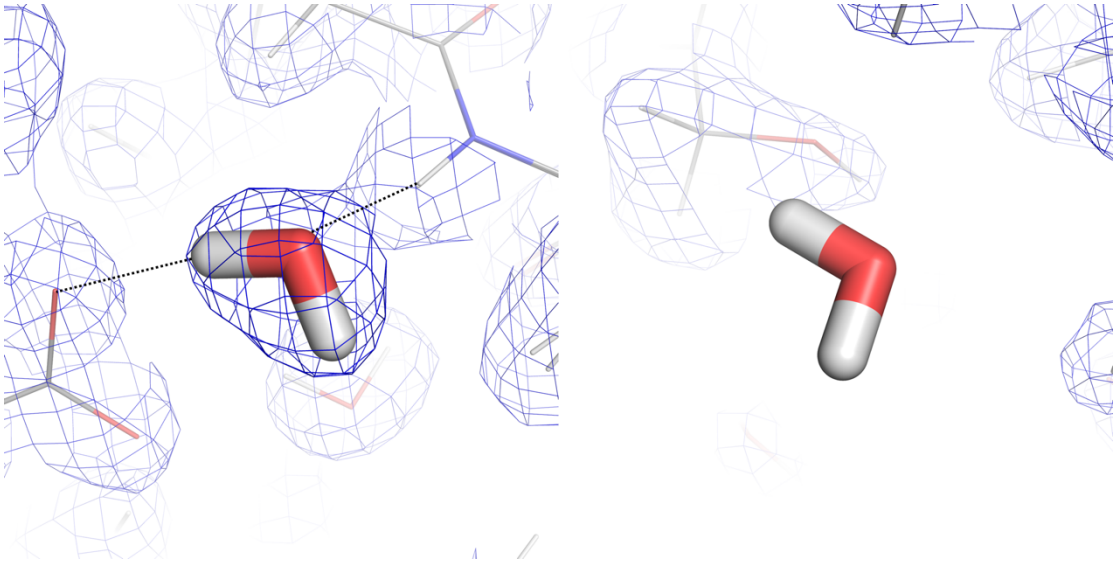

Figure 2 Example of a water molecule in the deposited structure of galectin-3C with a controversial orientation (blue) according to our qualitative evaluation. A better orientation is shown in brown (oriented manually) in two different views. The 2mF_o – DF_c nuclear-density omit map is shown in blue (σ = 1.0).

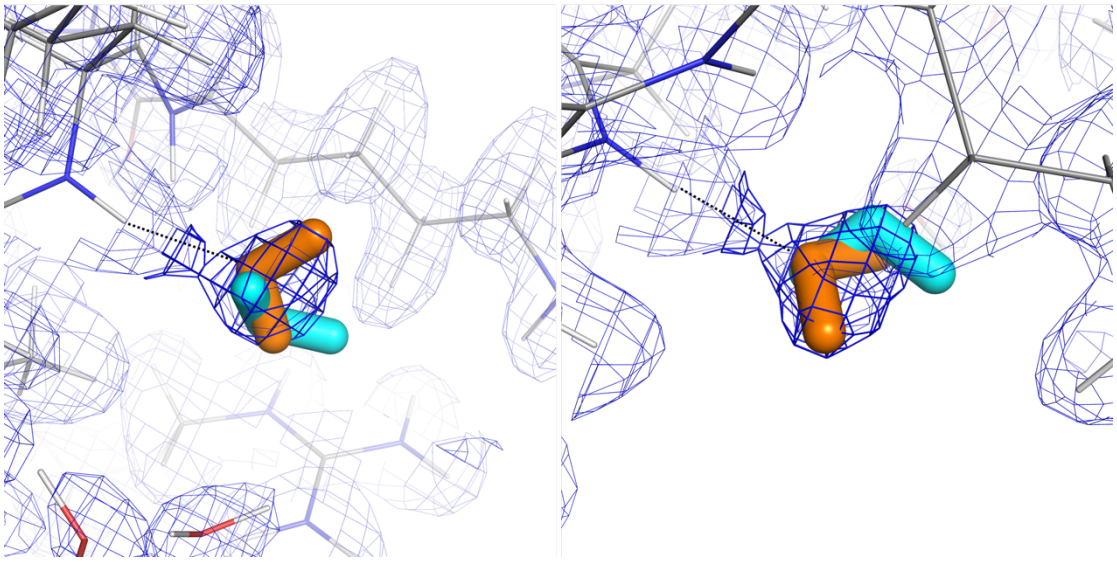

Figure 3 Histograms of RSZD, ADPs and RSCC values for the D atoms in water molecules. RSCC and RSZD was calculated with regular $2mF_o - DF_c$ maps. The histograms are calculated separately for water molecules with good or inadequate orientations, according to the qualitative assessment. The ideal RSZD value is zero, while a perfect RSCC is unity, and a good B-factor is around the average value (66). The RSZD and ADP histograms are from the qualitative evaluation of three galectin-3C structures: the deposited one, with hydrogens added in ReadySet and with hydrogens added in Maestro. The RSCC histogram contains data from all ten test cases, listed in Table 1.

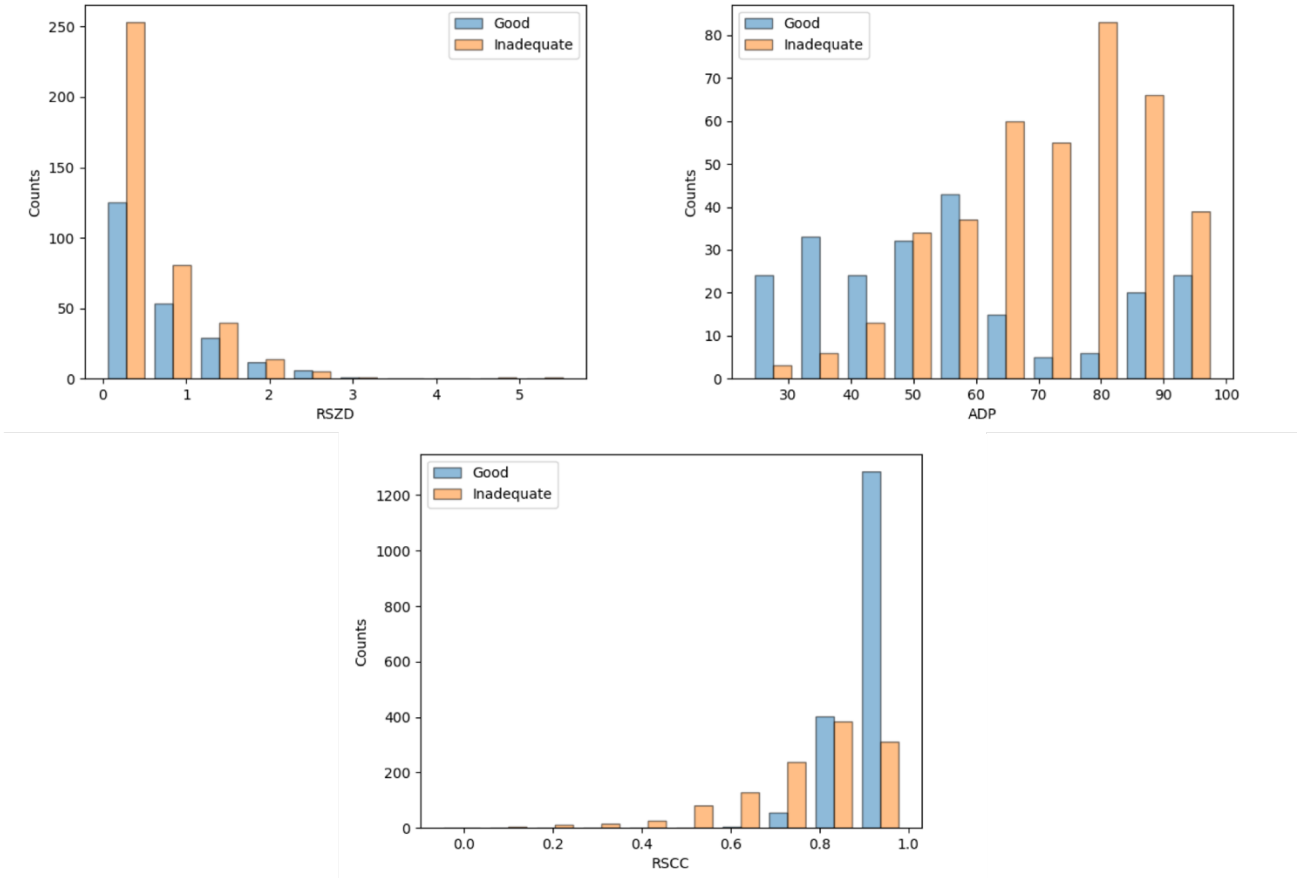

Figure 4 Example of a water molecule (number 617) in the structure of galectin-3C before (left) and after (right) running NWO-default. The $2mF_o - DF_c$ nuclear-density omit map is shown in blue (σ = 1.0).

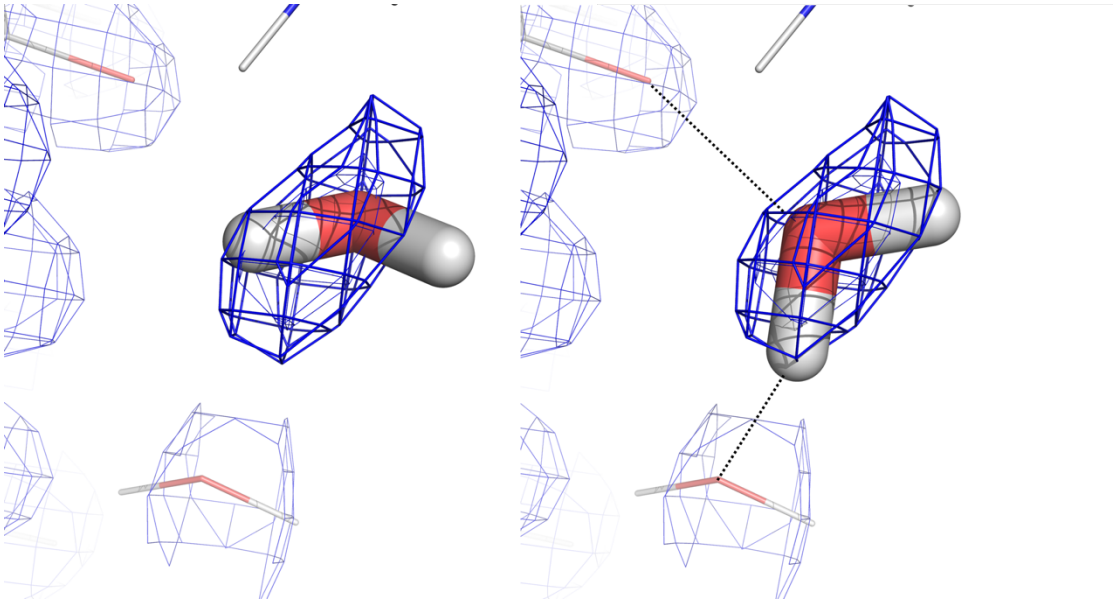

Table 1 Structures for which a qualitative assessment of the water molecules were performed. For each structure, the number of macrocycles in the refinement is given (0 for the deposited, which were not re-refined).

D atoms	refinement	number of macrocycles		
	strategy[a]	Galectin-3C	Rubredoxin	Pyrophosphatase
Deposited	–	0	0	0
ReadySet	1	3		
Maestro	1	3		
ReadySet	2	15	15	15
ReadySet	3	15		

Three types of refinement strategies were used: 1 – Joint X-ray/Neutron refinement with standard settings, 2 – Neutron refinement with a fixed protein, 3 – Neutron refinement with fixed protein, after running the water-orientation script.

Table 2 Results obtained for galectin-3C after a refinement with three macrocycles and standard settings, compared with the same refinement but using only neutron data and a fixed protein.

	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

Table 3 Performance of different refinement strategies for galectin-3C after re-generation of the water D atoms with ReadySet and three macrocycles. n_{wat} is the number of water molecules with correct orientation, i.e. with both D atoms having RSCC > 0.81.

refinement strategy						
data	space	protein	time (s)	n_{wat}	R_{work}	R_{free}
X-ray/neutron	reciprocal		845	54	0.2046	0.2184
X-ray/neutron	real		1820	35	0.2148	0.2252
neutron	reciprocal		203	69	0.1724	0.2288
neutron	real		201	52	0.1724	0.2288
neutron	real & reciprocal	fixed	214	71	0.1753	0.2280
neutron	reciprocal	fixed	213	71	0.1753	0.2280
neutron	real	fixed	135	50	0.1938	0.2305

Table 4 Dependence of the refinement results on the number of macrocycles (n_{mc}) for galectin-3C, refined in reciprocal space with a fixed protein and only neutron data. For each refinement, the refinement time (s), the number of correctly oriented water molecules (n_{wat}, i.e. having RSCC > 0.81for both D atoms), R_{work} and R_{free} are given.

n_{mc}	time	n_{wat}	R_{work}	R_{free}
3	283	71	0.1753	0.2280
5	329	75	0.1733	0.2258
7	440	77	0.1727	0.2262
10	607	79	0.1724	0.2249
12	716	78	0.1718	0.2251
14	799	80	0.1715	0.2244
15	953	81	0.1713	0.2245
18	1040	80	0.1710	0.2251
22	1255	81	0.1708	0.2258
25	1492	80	0.1711	0.2267
30	1702	81	0.1711	0.2280

Table 5 The performance of NWO-default and NWO-refine on galectin-3C regenerated by ReadySet, showing the timing (in seconds on a single computer), the number of water molecules with RSCC >0.81 (n_{wat}) and the two R factors as a function of the number of reorientation cycles.

		number of reorientation cycles			
	variant	10	50	100	150
time	default	420	720	1260	1740
	refine	4440	17880	23280	34200
n_{wat}	default	85	90	92	90
	refine	86	87	90	90
R_{work}	default	0.1721	0.1721	0.1720	0.1721
	refine	0.1723	0.1708	0.1711	0.1721
R_{free}	default	0.2276	0.2261	0.2283	0.2273
	refine	0.2312	0.2270	0.2308	0.2273

Table 6 The timing and performance times of NWO-default and phenix.refine on the three proteins. $n_{wat\text{-}tot}$ is the number of water molecules in each structure.

Protein	Galectin-3C	Rubredoxin	Pyrophosphatase
Time (s) NWO-refine	1260	1800	2820
Time (s) phenix.refine	1930	880	5082
$n_{wat\text{-}tot}$	110	149	385

Table 7 Comparison of the number of correctly oriented water molecules (n_{wat}) and the R factors for the deposited (dep) structures and the structures after refinement with or without NWO for the three proteins, using $2mF_o - DF_c$ maps with (default) or without (omit) water molecules included. n_{wat} is the number of correctly oriented water molecules with RSCC > 0.81. The number of water molecules correctly oriented according to the qualitative assessment is given in brackets. The refinement was run in reciprocal space for 15 macrocycles, using the neutron data only and keeping the protein fixed.

protein	galectin-3C				rubredoxin				pyrophosphatase			
	dep	without	default	omit	dep	without	default	omit	dep	without	default	omit
n_{wat}	81 (70)	81 (70)	92 (85)	75	44 (51)	49 (41)	55	51	313 (235)	351 (363)	363	334
R_{work}	0.168	0.171	0.172	0.179	0.199	0.203	0.203	0.211	0.239	0.191	0.188	0.192
R_{free}	0.211	0.224	0.228	0.230	0.237	0.237	0.238	0.243	0.252	0.285	0.293	0.278

Table 8 Number of hydrogen bonds formed with water deuterium atoms as donors in the deposited structure, the ReadySet structure and structures obtained with NWO-default and NWO-omit.

	deposited	ReadySet	NWO-default	NWO-omit
galectin-3C	44	31	47	42
rubredoxin	31	20	35	31
pyrophosphatase	84	104	144	140

Acknowledgements This investigation has been supported by grants from the Swedish research council (project 2018-05003), from Knut and Alice Wallenberg Foundation (KAW 2013.0022), from eSSENCE: the e-science collaboration and from the Royal Physiographic Society in Lund. The computations were performed on computer resources provided by the Swedish National Infrastructure for Computing (SNIC) at Lunarc at Lund University.

References

Adams, P. D., Afonine, P. V., Bunkóczi, G., Chen, V. B., Davis, I. W., Echols, N., Headd, J. J., Hung, L.-W., Kapral, G. J., Grosse-Kunstleve, R. W., McCoy, A. J., Moriarty, N. W., Oeffner, R., Read, R. J., Richardson, D. C., Richardson, J. S., Terwilliger, T. C., Zwart, P. H. & IUCr (2010). *Acta Crystallogr. Sect. D Biol. Crystallogr.* **66**, 213–221.

Afonine, P. V., Grosse-Kunstleve, R. W., Echols, N., Headd, J. J., Moriarty, N. W., Mustyakimov, M., Terwilliger, T. C., Urzhumtsev, A., Zwart, P. H., Adams, P. D., Ralf, W., Headd, J. J. & Thomas, C. (2012). *Acta Crystallogr. Sect. D Biol. Crystallogr.* **68**, 352–367.

Afonine, P. V, Mustyakimov, M., Grosse-Kunstleve, R. W., Moriarty, N. W., Langan, P. & Adams, P. D. (2010). *Acta Crystallogr. Sect. D*. **66**, 1153–1163.

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

Blakeley, M. P., Hasnain, S. S. & Antonyuk, S. V. (2015). *IUCrJ*. **2**, 464–474.

Cuypers, M. G., Mason, S. A., Blakeley, M. P., Mitchell, E. P., Haertlein, M. & Forsyth, V. T. (2013). *Angew. Chemie Int. Ed.* **52**, 1022–1025.

Deller, M. C. & Rupp, B. (2015). *J. Comput. Aided. Mol. Des.* **29**, 817–836.

Howard, E. I., Sanishvili, R., Cachau, R. E., Mitschler, A., Chevrier, B., Barth, P., Lamour, V., Van Zandt, M., Sibley, E., Bon, C., Moras, D., Schneider, T. R., Joachimiak, A. & Podjarny, A. (2004). *Proteins Struct. Funct. Bioinforma.* **55**, 792–804.

Inoguchi, N., Coates, L., Meilleur, F., Morris, M., Singhal, A., Barcena, J., Garcia-Ruiz, J., Pusey, M. & Ng, J. (2017). *Acta Crystallogr. Sect. A*. **73**, a132.

Levy, Y. & Onuchic, J. N. (2006). *Annu. Rev. Biophys. Biomol. Struct.* **35**, 389–415.

Liebschner, D., Afonine, P. V., Baker, M. L., Bunkoczi, G., Chen, V. B., Croll, T. I., Hintze, B., Hung, L. W., Jain, S., McCoy, A. J., Moriarty, N. W., Oeffner, R. D., Poon, B. K., Prisant, M. G., Read, R. J., Richardson, J. S., Richardson, D. C., Sammito, M. D., Sobolev, O. V., Stockwell, D. H., Terwilliger, T. C., Urzhumtsev, A. G., Videau, L. L., Williams, C. J. & Adams, P. D. (2019). *Acta Crystallogr. Sect. D Struct. Biol.* **75**, 861–877.

Manzoni, F., Wallerstein, J., Schrader, T. E., Ostermann, A., Coates, L., Akke, M., Blakeley, M. P., Oksanen, E. & Logan, D. T. (2018). *J. Med. Chem.* **61**, 4412–4420.

Mattos, C. (2002). *Trends Biochem. Sci.* **27**, 203–208.

Neumann, P. & Tittmann, K. (2014). *Curr. Opin. Struct. Biol.* **29**, 122–133.

Nittinger, E., Schneider, N., Lange, G. & Rarey, M. (2015). *J. Chem. Inf. Model.* **55**, 771–783.

O'Dell, W. B., Bodenheimer, A. M. & Meilleur, F. (2016). *Arch. Biochem. Biophys.* **602**, 48–60.

Schrödinger release 2019-2 (2019), Schrödinger LLC, New York,.

Sears, V. F. (1986). *Methods Exp. Phys.* **23**, 521–550.

Tickle, I. J. (2012). *Acta Crystallogr. Sect. D Biol. Crystallogr.* **68**, 454–467.