JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

Evidence of Water Molecules—A Statistical Evaluation of Water Molecules Based on Electron Density

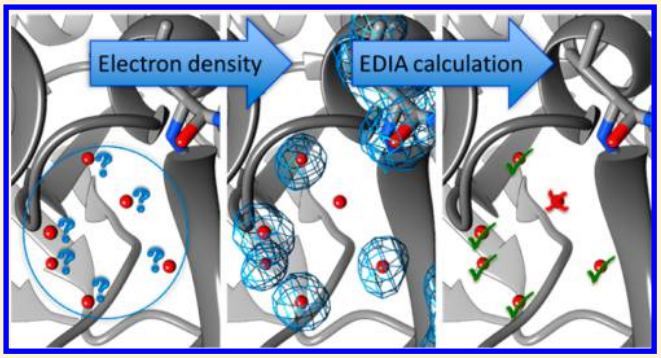
Eva Nittinger,[†] Nadine Schneider,[†] Gudrun Lange,[‡] and Matthias Rarey*,[†]

[†]Center for Bioinformatics, University of Hamburg, Bundesstraße 43, 20146 Hamburg, Germany

[‡]Bayer CropScience AG, Industriepark Hoechst, G836, 65926 Frankfurt am Main, Germany

Supporting Information

ABSTRACT: Water molecules play important roles in many biological processes, especially when mediating protein−ligand interactions. Dehydration and the hydrophobic effect are of central importance for estimating binding affinities. Due to the specific geometric characteristics of hydrogen bond functions of water molecules, meaning two acceptor and two donor functions in a tetrahedral arrangement, they have to be modeled accurately. Despite many attempts in the past years, accurate prediction of water molecules—structurally as well as energetically—remains a grand challenge. One reason is certainly the lack of experimental data, since energetic contributions of water molecules can only be measured indirectly. However, on the structural side, the electron density clearly shows the positions of stable water molecules. This information has the potential to improve models on water structure and energy in proteins and protein interfaces. On the basis of a high-resolution subset of the Protein Data Bank, we have conducted an extensive statistical analysis of 2.3 million water molecules, discriminating those water molecules that are well resolved and those without much evidence of electron density. In order to perform this classification, we introduce a new measurement of electron density around an individual atom enabling the automatic quantification of experimental support. On the basis of this measurement, we present an analysis of water molecules with a detailed profile of geometric and structural features. This data, which is freely available, can be applied to not only modeling and validation of new water models in structural biology but also in molecular design.



INTRODUCTION

In order to fully understand complex biomolecular structures, the role of water molecules needs to be comprehended in greater detail. Water molecules form part of the environment of biological macromolecules, in which they can on the one hand stabilize protein folding by mediating interactions and on the other sustain the dynamics of the protein.[1−5] Enzymatic reactions often directly involve one water molecule, i.e., as reactant for hydrolysis reactions,[6,7] or as steric hindrance to guide stereoselectivity[8]. Further, water molecules stabilize biological complexes by mediating protein−protein or protein−ligand interactions, in which mediated hydrogen bonds are often as abundant as direct interactions.[9−12]

In addition to the aforementioned biological processes, water molecules also play an essential role in energetic effects upon binding of, for example protein and ligand or protein and protein, due to their contribution to hydration and dehydration as well as the hydrophobic effect.[13−19] On the one hand, energy is needed in order to dehydrate hydrophilic atoms of the protein and the ligand. On the other hand, energy is also gained by releasing water molecules into the bulk solvent and the hydrophobic effect. Therefore, in order to correctly estimate protein−ligand binding affinities, water molecules have to be taken into account when developing drugs. However, it is not yet understood how water molecules exactly contribute to the binding affinity.[20,21] It is hardly possible to experimentally measure the energy contribution of an individual water molecule. Even if a water molecule in the binding site can be displaced, for instance by an extension of the ligand, the resulting binding affinity is a combination of those two effects, replacement and substitution.[22−24]

Different computational methods have been developed lately that try estimate the energetic cost or gain of water molecule displacement to guide rational drug design. Herein, two different approaches exist: first, classification of X-ray crystallographic water molecules[25−28] and, second, computer based prediction of water molecule positions[29−36] (Table 1). As early as 1985, Goodford developed a method, which calculates energies between diverse probe groups and a protein in a grid-based manner.[30] One of these probes resembled water, enabling the prediction of water molecule positions in three examples in good agreement to X-ray crystallographic water molecules. However, its overall applicability has not been proven.

Received: October 31, 2014
Published: March 5, 2015

ACS Publications © 2015 American Chemical Society

DOI: 10.1021/ci500662d
J. Chem. Inf. Model. 2015, 55, 771−783

Table 1. Methods for Positioning Water Molecules within Biological Complexes and Prediction of Water Molecule Stability—Structurally and Energetically

method type	method name	prediction task
simulation-based	Grand Canonical Monte Carlo simulation (GCMC)[29]	binding free energy estimation
	GIST[35]	displaceability, thermodynamic analysis
	JAWS[34]	binding affinity estimation
	SZMAP[31]	orientation and displaceability
	WaterMap[32,33]	energy (enthalpic and entropic) contribution
docking-based	WaterDock[36]	conserved vs displaceable
grid-based	GRID[30]	energy (enthalpic) contribution
	WaterFlap	water score ("happiness")

Recently, more elaborate and time-expensive methods have been developed, from which some use molecular dynamics simulations in order to place water molecules and estimate their binding affinity contributions (e.g., WaterMap[32,33]). These programs can be separated into three main types—simulation-based, docking-based, and grid-based. Furthermore, they can be classified according to their overall prediction aim (See Table 1). Most of the time it is not known which water molecules will be displaced upon binding and which ones remain in the protein binding site, thus mediating between protein and ligand. However, it is assumed that those water molecules remaining in the protein complex are more stable than the ones being displaced upon complex formation.

Diverse characterizations of water molecules have been conducted, ranging from structural analysis to thermodynamic description.[37−44] Water molecules in protein−ligand interactions have been of great interest and different analyses have shown that those bridging protein and ligand often have three or more interactions compared to only one interaction on average in protein−protein interfaces.[37,38] Furthermore, water molecules prefer interactions to the backbone rather than to the side chain, indicated by the number of interactions as well as thermodynamically.[38−40] Dunitz has approximated the entropic gain of transferring a water molecule from protein into bulk water to be up to 7 cal/mol at room temperature.[41] Using inhomogeneous fluid solvation theory the impact of α-helix and β-sheet on the thermodynamic properties of water molecules has been analyzed. Herein, the thermodynamics of water molecules are affected up to a distance of 4 Å from β-sheets and 4.3 Å from α-helices.[42,43] In an application of the method WaterMap, it was estimated that charged amino acids display the most favorable hydration sites, whereas backbone, aromatic, and aliphatic amino acids are less favorable.[44] In this publication we want to keep the focus on structural characteristics of crystallographic water molecules. More detailed information about thermodynamics of water molecules can be found in ref 45.

In order to retrieve experimental data for water molecules in biological complexes, crystal structures are the major source. However, X-ray crystallographic experiments result in diffraction patterns, which have to be interpreted further in order to acquire the molecular structure. The temperature factor (B-factor) is an indicator of thermal motion of each atom and is often taken as a criterion to identify flexible regions in protein structures. However, the B-factor depends on the refinement procedure: its interpretation can be artifactural if crystal contacts are neglected and it varies between different structures. The B-factor does not inform whether an atom is resolved by electron density, but it does indicate its structural flexibility and disorder. Electron density, which is available for many structures nowadays, is the fundamental experimental data available for water molecules (See Table 2). Two measurements exist that describe electron density for specific parts of a structure: the real-space R-factor (RSR) and the real-space correlation coefficient (RSCC).[46] The RSR was developed as an objective interpretation of electron density maps and for the localization of errors during density map interpretation. It is calculated by using the observed electron density from the crystallographic experiment and the calculated electron density derived from the built structure. The lower the RSR is (in a range from 0 to 1) the better the fit of the structure in the electron density. The RSCC, in contrast to the RSR, is the correlation coefficient between the two maps resulting in a value between −1 (complete anticorrelation) and 1 (complete correlation). One drawback of the RSCC arises when it is calculated for atoms with weak densities, but correct intensity distributions. This is especially problematic for water molecules since even a good score might be achieved with low resolution. For protein−ligand complexes, it was shown lately by Hawkins et al.[47] that neither RSR nor RSCC can adequately capture the difference between observed and calculated data.

Table 2. Values Used as Structural Quality Criteria for Identification of Modeling Errors and Structural Uncertainties

value	advantage	disadvantage
B-factor	• included in PDB file • indicator of thermal motion	• can be misinterpreted, e.g. due to crystal packing
RSR	• normalized value [0 (good) to 1 (bad)]	• resolution and density threshold dependent
RSCC	• no density threshold used • normalized value [−1 (anticorrelation) to 1 (correlation)]	• weak density with correct intensity distribution might lead to a good score

Water molecules resolved by X-ray crystallography exist at local energy minima of the position.[48] However, not all atoms present in molecular structures are supported by electron density. Most water molecules are too flexible to be resolved. Furthermore, water molecules found in the crystallographic structure vary strongly in their experimental support. For a detailed study of water molecules in proteins and at protein interfaces, a quantitative measure of electron density support is therefore mandatory to exclude noise arising from unresolved water molecules.

In order to exploit electron density for this task, we developed a new value, called EDIA (= **E**lectron **D**ensity for **I**ndividual **A**toms), which describes the experimental electron density around a single atom, for instance around a single water molecule. Using EDIA, we conducted an extensive evaluation of water molecules from a high-resolution subset of the PDB[49], containing 5485 PDB structures and over 2.3 million water molecules. In the following sections, we give a detailed description of the selected data set and its diversity. The newly developed EDIA measure will be explained in detail. According to the EDIA value, the water molecules of the data set are examined with regard to several structural and geometric

DOI: 10.1021/ci500662d
J. Chem. Inf. Model. 2015, 55, 771−783

characteristics, such as hydrogen bonding preferences and their structural environment. Finally, we show detailed examples of frequently discussed issues like "hydrophobic bubbles"[40] or modeling errors, which can still be found in high-resolution data.

MATERIALS AND METHODS

Data Set. A high-resolution Protein Data Bank[49] subset was compiled using the following advanced search criteria: resolution ≤ 1.5 Å, experimental method = X-ray, molecule type = protein. All structures between 2000-01-01 and 2014-02-01 were selected that had an external link to the EDS server[50], ensuring the availability of electron density data for all chosen structures. Using all criteria, a data set of 5526 PDB structures was compiled (Figure 1, date of download: February 1, 2014).

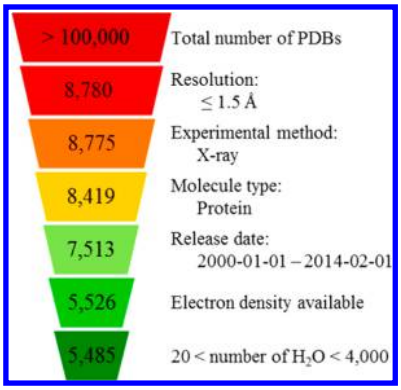


Figure 1. Data set compilation and the effect of each search criterion on the number of PDB structures.

In the next step, two extremes were discarded: those PDB structures with less than 20 water molecules and those with more than 4000, resulting in a final data set of 5485 PDB structures with 2 330 581 water molecules (See Table S1).

Electron Density-Based Value. The electron density is provided as a 3D grid for the asymmetric unit, the smallest unit of volume that by application of symmetry operations is able to reconstruct the unit cell. The unit cell on the other hand is the smallest volume that only by translational application can recreate its pattern in space. We developed an automated estimation of electron density around a single atom not covalently bound to other heavy atoms, called Electron Density for Individual Atoms (EDIA). EDIA is the weighted sum of experimental electron density values around a single atom a in its van der Waals radius:

$$\mathrm{EDIA}(a) = \frac{1}{\sum_{p\in S(a)} \omega(p)\cdot\sigma} \sum_{p\in S(a)} \omega(p)\cdot(f(p) - \mu) \quad (1)$$

where $\omega(p)$ describes a weight function for grid point p, σ, the electron density threshold, and $S(a)$ the subset of grid points around an atom a. The function $f(p)$ reflects the density value at grid point p, and μ, the mean density of the electron density map. In order to allow comparisons between different structures, the EDIA is normalized by the standard deviation σ of the respective electron density map of the asymmetric unit.

Each grid point of the electron density map is associated with a measured density value. In eq 1 $S(a)$ is the subset of all grid points G that are within the van der Waals radius of atom a:

$$S(a) = \{p \in G | |\overrightarrow{px_a}| \le r_{\mathrm{vdW}(a)}\} \quad (2)$$

The distance in angstroms of a grid point p to the atom center x_a is $|\overrightarrow{px_a}|$. The distribution of electron density around an atom, caused by the vibration of the atom itself as well as the distribution of electrons around an atom, is resembled using a Gaussian weight:

$$\omega(p) = \mathrm{e}^{-1/2(|\overrightarrow{px_a}|/\delta)^2} \quad (3)$$

The width of the Gaussian bell δ was defined as the covalent radius of the atom. This results in a weight of 0.5 when the distance of a grid point to the center of the atom is equal to its covalent radius. The Gaussian distribution was combined with a linear function $g(p)$ in order to get no further density contributions of grid points with a distance greater than the van der Waals radius of the atom (Figure 2a):

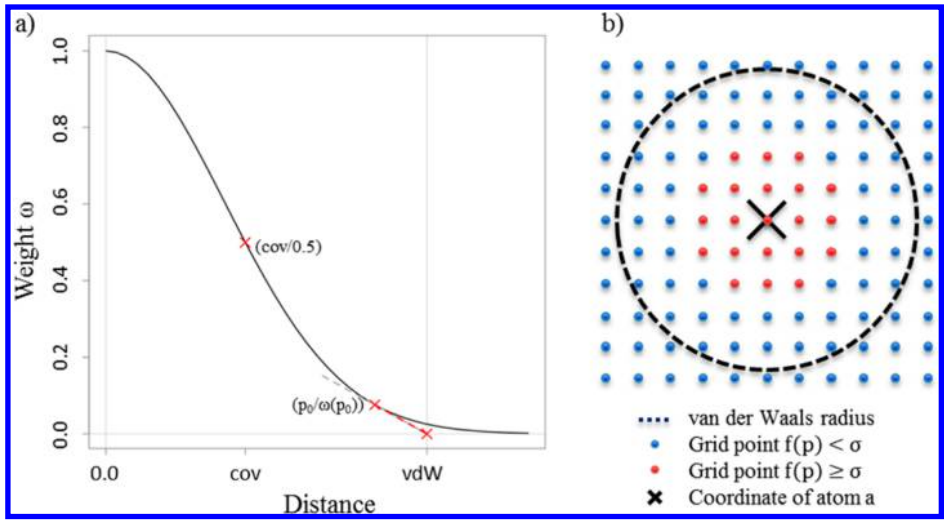


Figure 2. (a) Gaussian weight combined with linear function for EDIA calculation. (b) 2D scheme of an atom and its surrounding electron density grid, with grid points contributing (red dots) and not contributing (blue dots) to EDIA: cov = covalent radius of atom a, vdW = van der Waals radius of atom a, $(p_0/\omega(p_0))$ = starting point of linear function (red dashed line).

DOI: 10.1021/ci500662d
J. Chem. Inf. Model. 2015, 55, 771−783

$$g(p) = \frac{\omega(p_0)}{p_0 - r_{vdW}}(p - r_{vdW}) \quad (4)$$

The slope of the linear function was selected such that the overall function remains continuously differentiable by passing through the points $(p_0, \omega(p_0))$ and $(r_{vdW}, 0.0)$, with $p_0 = (r_{vdW}/2) + [(r_{vdW}^2/4) - \delta^2]^{1/2}$.

Only density values ρ at grid points p that are above the density threshold σ (= map mean density + standard deviation, threshold typically applied for density visualization) were added up (Figure 2b):

$$f(p) = \begin{cases} \rho(p), & \rho(p) \geq \sigma \\ 0, & \rho(p) < \sigma \end{cases} \quad (5)$$

By excluding very low density values, the noise in the EDIA can be substantially reduced.

Preprocessing of PDB Structures. In order to analyze the hydrogen bond network, hydrogen positions were calculated using Protoss.[51] Protoss adds hydrogens and places them into the structure optimizing the hydrogen bond network accounting for tautomers and protonation states. Herein, it not only considers the amino acids of the protein but also water molecules, metals and ligands. Using an empirical scoring function, Protoss generates an optimal hydrogen bond network for a biological complex. This preprocessing step was applied once for each PDB complex. Afterward, each water molecule of the complex was analyzed individually in its surrounding (4.5 Å radius around the center of the water oxygen).

Descriptor Calculation. For the characterization of water molecules within their surrounding environment, several descriptors were calculated (See Table 3).

Table 3. Summary of Descriptors Used to Characterize Water Molecules

descriptor type	description	details
hydrogen bond descriptors	number of hydrogen bonds	• to all modeled atoms • to protein and ligand atoms as well as metals
	mean length of hydrogen bonds	• to all modeled atoms • to protein and ligand atoms as well as metals
	hydrogen bond partners	• atom type • functional group • acceptor or donor • side chain or backbone
proximity-based descriptors	hydrophobicity	• proportion of hydrophobic atoms within 4.5 Å radius
	hydrophobic surface	• proportion of hydrophobic surface pointing toward the water molecule
	water clusters	• water molecules within 3.5 Å radius

Hydrogen Bond Descriptors. Hydrogen bonds of water molecules to donors or acceptors were identified if the opening angle between ideal donor and acceptor direction was less than 50° and the distance between hydrogen and acceptor was within a range of 1.9 ± 0.5 Å. For each hydrogen bonding function only the geometrically best hydrogen bond was accepted. In this way, bifurcate hydrogen bonds were excluded. The total number of hydrogen bonds and the mean length of the hydrogen bonds were calculated for each water molecule. Additionally, the hydrogen bonding partners were recorded, i.e., atom type, functional group, acceptor or donor, backbone or side chain.

Hydrophobicity-Based Descriptors. In order to classify the surrounding of a water molecule, two types of hydrophobicity-related values were applied. First, the fraction of hydrophobic atoms in a 4.5 Å surrounding sphere was calculated (Figure 3a):

$$\text{hydrophobicity} = \frac{\text{number of hydrophobic atoms surrounding a water molecule (4.5 Å)}}{\text{total number of atoms surrounding a water molecule (4.5 Å)}} \quad (6)$$

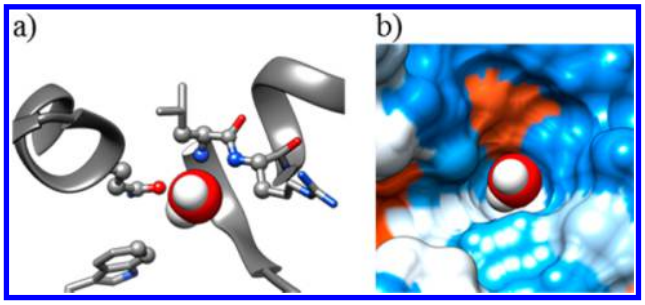


Figure 3. Hydrophobicity-based descriptors. (a) Ball-and-stick represented atoms are within 4.5 Å distance and contribute to hydrophobicity (hydrophobicity = 0.623). (b) Molecular surface of a water molecule surrounded by protein (hydrophobic surface = 0.372): orange = hydrophilic, blue = hydrophobic (molecular graphics were created using UCSF Chimera[52]).

Second, the size of the hydrophobic surface patches of surrounding atoms (≤4.5 Å) was calculated. Here, only those surface patches being closer than 2 Å to the water molecule were considered. For normalization, the fraction was used. (Figure 3b):

$$\text{hydrophobic surface} = \frac{\text{surface area of hydrophobic atoms pointing towards a water molecule (2 Å)}}{\text{surface area of all atoms pointing towards a water molecule (2 Å)}} \quad (7)$$

Water Cluster. The water content of the surroundings of a water molecule was analyzed using water clusters. Herein, within a distance of 3.5 Å, the surrounding was checked for other water molecules. If another water molecule is present, another test, which checked for the presence of further water molecules, was performed. This procedure was prolonged until no further water molecule was identified. Finally, the total number of water molecules within one cluster was counted.

Allocation of Water Molecules. Since water molecules occupy different positions within a biological complex, a further classification into surface (S), protein−ligand interface (PLI), protein−protein interface (PPI), and captured (C, also often called "buried") was performed. This categorization was carried out using the molecular surface area (MSA, Figure 4). Water molecules were classified as PLI, if protein and ligand atoms were found within a 4.5 Å radius around the oxygen atom. Analogous classification was performed for PPI water molecules, which have atoms of two different protein chains

DOI: 10.1021/ci500662d
J. Chem. Inf. Model. 2015, 55, 771−783

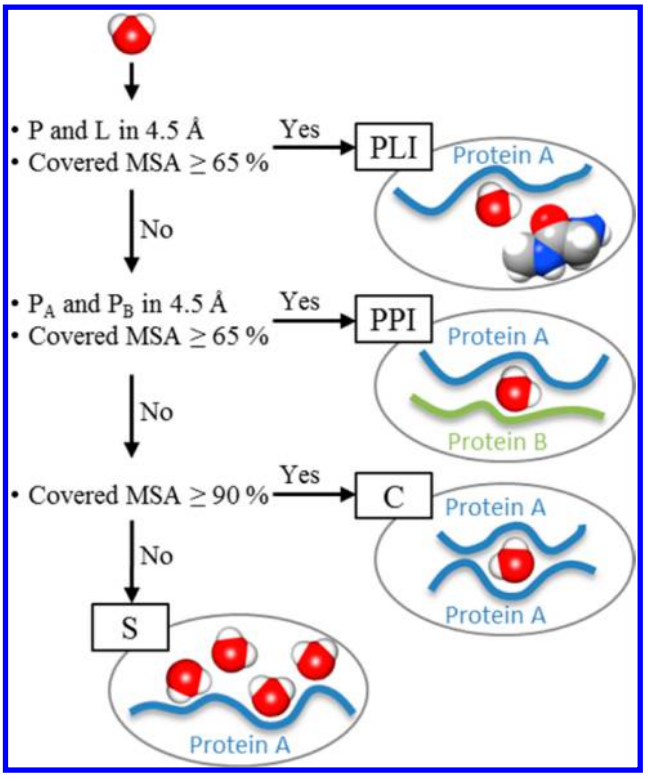


Figure 4. Classification of water molecules: P = protein, L = ligand, P_A = protein A, P_B = protein B.

within a radius of 4.5 Å. In both cases, we additionally checked that at least 65% of their MSA was covered, since they would otherwise lie in the outer rim of either PLI or PPI. Water molecules were classified as C if more than 90% of their MSA was covered by a single protein chain, which means it has more than three covered hydrogen bonding functions. All remaining water molecules were classified as S.

Data Set Compositions. The high-resolution data set was divided into different subsets to allow a detailed structural analysis of different classes of water molecules. Apart from the classification into S, C, PLI, and PPI, the data set was further classified according to diverse structural criteria (See Table 4).

RESULTS AND DISCUSSION

Data Set Composition. The data set is highly diverse with respect to the water content of the PDB structures (Figure 5a).

Table 4. Subsets of the High Resolution Data Set, Their Compositions, and Abbreviations Used for the Analysis of Water Molecules

data set abbreviation	data set composition
$Hbond_{all}$	water molecules interacting with protein, ligand, and water molecules
$Hbond_{PL}$	water molecules interacting with protein and ligand
$Hbond_{H2O}$	water molecules interacting with water molecules
HB	water molecules that cannot form any hydrogen bond and lie within a highly hydrophobic surrounding (= hydrophobic bubbles)
$WIND_{all}$	well-integrated water molecules (≥three hydrogen bonds) without electron density interacting with protein, ligand, and water molecules
$WIND_{PL}$	well-integrated water molecules (≥three hydrogen bonds) without electron density interacting with protein or ligand

Some complexes have very low water content with less than one water molecule per two amino acids (13%). Others in contrast are more highly hydrated with more than three water molecules per two side chains (19%). One of the "dry" proteins is a heat-labile enterotoxin of *E. coli* with one water molecule per 100 amino acids (PDB ID: 4fo2). Its biological unit is a pentamer; mostly water molecules of the inner protein parts were modeled while an outer solvation layer is nearly absent.

A structure of a "wet" protein is an antifreeze protein of *L. dearborni*, which consists of only one very small subunit (63 amino acids) and up to four water molecules per amino acid (PDB ID: 1ucs). This protein does not have any inner water molecules (C, PLI, or PPI), but an extensive solvation layer with many close water molecules. The structure is very well-resolved (0.62 Å) and even hydrogen atoms could be modeled. However, the distances between water molecules are often very small, if not even clashing. The intention of the authors was to model different interaction networks, which are indicated by reduced occupancies of water molecules.[53] Multiple water molecules are modeled into the electron density of oxygen and hydrogen atoms (Figure 6a). This example (PDB ID: 1ucs) shows that even if the structures are of high-resolution, it should not be taken as a guarantee for easily interpretable data. Few other modeling errors also captured our interest, such as the fusion of a water molecule with an amino acid side chain (Figure 6b). This error would not be detected with EDIA, because electron density from the amino acid is available. However, it can be detected using either the electron density difference map ($f_o - f_c$ map), since too many electrons are available; or the difference of Gaussian operator, since the position of the water molecule has clearly no circular density distribution. Overlaps of water molecules were also identified in 11 PDB structures. Those contain multiple water molecules with identical coordinates of the oxygen atom (PDB ids: 2ghc, 2hc1, 2yqb, 3t6f, 3ziy, 3zjp, 4a8n, 4af9, 4ayp, 4ayr, 4b8x). Furthermore, two structures included overlaid ligands with different occupancies (overlap of biotin and biotin-D-sulfoxide (PDB id: 3t6f), overlay of β-D-glucose and α-D-glucose (PDB id: 4af9), in both cases the ligand with the higher occupancy was kept for further analysis) and 36 structures contained unknown ligands that were represented by oxygen atoms only (See Table S1).

In literature it is often mentioned that the number of observed water molecules depends on the resolution of the structure.[10,48,54,55] Figure 5b shows that within structures of our high-resolution data set, the ratio of water molecules hardly varies. In all cases, the median number of water molecules per amino acid is close to one, with a minimum of one water molecule per 100 amino acids (PDB id: 4fo2) and a maximum of nearly four water molecules per one amino acid (PDB id: 1ucs).

The data set compromises a wide range of protein complexes including complexes without any ligands (25%) and those containing ligands, cofactors, and crystallization buffer molecules (75%). Protein complexes range from small monomers to multimers, with more than 40% of the data set containing more than one protein chain. The size of the proteins ranges from 12 (peptides) to 5480 amino acids. Using the classification of water molecules described in the Materials and Methods section, the data set contains 127 988 PPI and 80 717 PLI water molecules, 264 584 captured (C), and 1 857 292 surface (S) water molecules (Figure 7).

DOI: 10.1021/ci500662d
J. Chem. Inf. Model. 2015, 55, 771−783

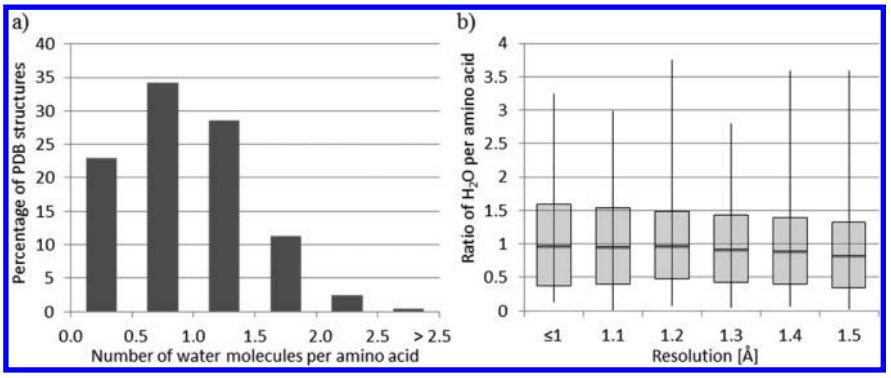


Figure 5. (a) Histogram of the number of water molecules per amino acid for the data set of 5485 PDB structures: median = 0.89, standard deviation = 0.51. (b) Box plot of resolution dependent number of water molecules per amino acid: number of water molecules per category (≤1 Å) 177, (1.1 Å) 296, (1.2 Å) 487, (1.3 Å) 826, (1.4 Å) 1349, (1 Å) 2340. Box limits are median ± one standard deviation.

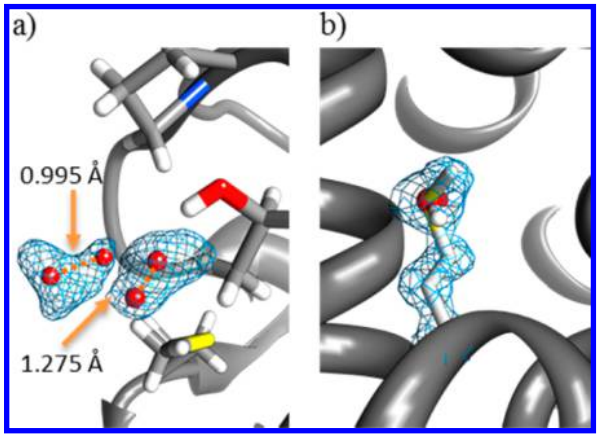


Figure 6. (a) Two alternative water molecules are modeled into the density of one water molecule, H_2O-A-256—H_2O-A-182: 1.275 Å, H_2O-A-284—H_2O-A-136: 0.995 Å (PDB id: 1ucs). (b) Oxygen atom of the water molecule B-2179 fuses with sulfur atom of methionine B-208 (PDB id: 2jae). Electron density is only shown for methionine and water molecule. Difference electron density (not displayed) indicates too many electrons at the water position: blue = electron density map ($2f_o - f_c$) at 1σ (molecular graphics were created using UCSF Chimera[52]).

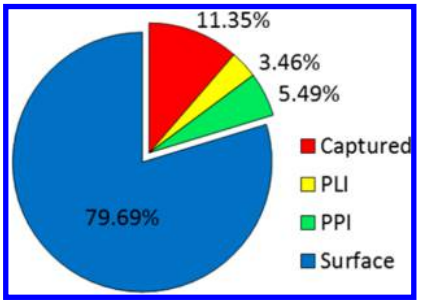


Figure 7. Classification of water molecules according to their position in the complex and their contribution to the data set of 2.3 million water molecules is displayed.

EDIA Values. The histogram of all EDIA values for the high-resolution data set approximately displays an extreme value distribution (Figure 8a). The individual EDIA values resemble well the graphically observable electron density (Figure 8b–f), with zero meaning no electron density and values above one describing clear electron density. Visual inspection of water molecules, their corresponding electron density, and the EDIA give confidence that the EDIA captures the measured electron density.

EDIA was further used to differentiate between well-resolved (with clear electron density) and insufficiently resolved (insufficient electron density) water molecules. As a cutoff value we used $EDIA_{Thrs}$ = 0.24, which is the median EDIA value of all water molecules from the data set minus one standard deviation. The median was chosen, because it is less affected by outliers, especially arising from a few very high EDIA values. Note that a cutoff value of zero was not used because surrounding and very close atoms might cause a slight increase of a water's EDIA value. While this has very little effect on resolved water molecules, it leads to a very small EDIA for unresolved water molecules. Visual inspection confirmed the chosen threshold of 0.24 (see Figure 8c and d), above which water molecules are considered as sufficiently resolved. This leads to 8.9% (208 052) of all water molecules of the data set being classified as insufficiently resolved by electron density, from which the majority (93.4%) belongs to the group of surface water molecules (See Table 5). This meets previous expectations and confirms the applicability of the EDIA.

Hydrogen Bonding Characteristics of Water Molecules. Water molecules were analyzed for their hydrogen bonding characteristics, wherein a maximum of four hydrogen bonds, two donor and two acceptor functions, was assumed. Bifurcated hydrogen bonds were excluded by considering only the geometrically best hydrogen bond for each hydrogen bonding function (see the Materials and Methods section).

Three separate statistics were created: (1) all hydrogen bonds to explicitly modeled atoms were counted for all water molecules, (2) only water molecules sufficiently resolved by electron density ($EDIA \geq EDIA_{Thrs}$) were considered, (3) only water molecules unresolved by electron density ($EDIA < EDIA_{Thrs}$) were taken into account (see Table 6).

The first statistic, which includes all water molecules, results in a mean number of hydrogen bonds of 2.15 ($Hbond_{total}$) from which 1.22 ($Hbond_{H2O}$) are formed to other water molecules. As expected, the second statistic, describing only water molecules with clear electron density, does not vary much from the first one. However, looking at water molecules without clear electron density in the third statistic, the mean number of hydrogen bonds decreases significantly ($Hbond_{total}$ = 1.65). In both sets, meaning water molecules resolved by electron density and unresolved ones, the high proportion of surface water molecules causes a bias in the mean number of hydrogen bonds (See Table 5 and Figure 9).

DOI: 10.1021/ci500662d
J. Chem. Inf. Model. 2015, 55, 771–783

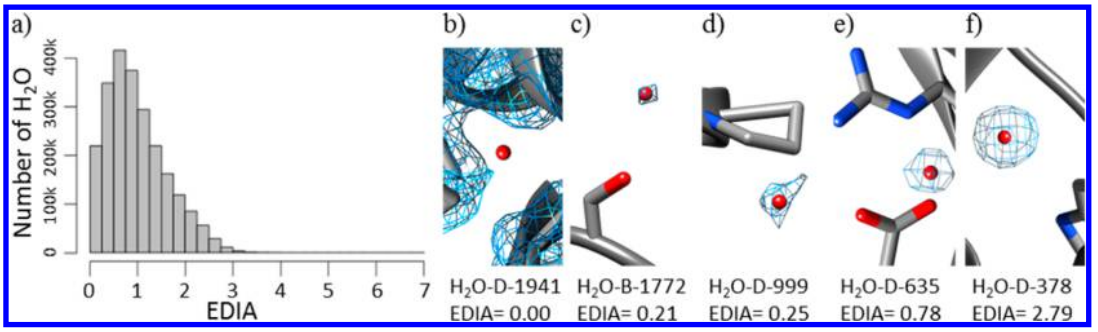


Figure 8. (a) EDIA histogram for all water molecules in the data set, mean = 0.981, median = 0.868, standard deviation = 0.625. (b–f) Examples of EDIA values and the corresponding visualization (3fpc, 1.4 Å). (b) Electron density map ($2f_o - f_c$ map, blue mesh) at 1σ for the whole region is shown. (c–f) $2f_o - f_c$ at 1σ is only shown for the water molecule itself (molecular graphics were created using UCSF Chimera[52]).

Table 5. Numbers of Water Molecules in Each Class[a] after Separation According to $EDIA_{Thrs}$

	well-resolved H_2O ($EDIA(H_2O) \geq EDIA_{Thrs}$)	insufficiently resolved H_2O ($EDIA(H_2O) < EDIA_{Thrs}$)
C	258106 (97.6%)	6478 (2.4%)
PLI	77673 (96.2%)	3044 (3.8%)
PPI	123,731 (96.7%)	4257 (3.3%)
S	1663019 (89.5%)	194273 (10.5%)

[a]Captured C, protein–ligand interface PLI, protein–protein interface PPI, surface S.

Table 6. Hydrogen Bonding Characteristics of Water Molecules[a]

	H_2O position	$Hbond_{total}$ ± stdev	$Hbond_{H2O}$ ± stdev
	all	2.15 ± 0.97	1.22 ± 0.93
H_2O EDIA ≥ $EDIA_{Thrs}$	all	2.12 ± 0.96	1.15 ± 0.92
	S	1.98 ± 0.94	1.19 ± 0.93
	C	2.70 ± 0.83	0.87 ± 0.79
	PLI	2.47 ± 0.92	1.03 ± 0.89
	PPI	2.60 ± 0.90	1.31 ± 0.94
H_2O EDIA < $EDIA_{Thrs}$	all	1.65 ± 0.89	1.14 ± 0.89
	S	1.62 ± 0.89	1.15 ± 0.89
	C	1.88 ± 0.96	0.76 ± 0.79
	PLI	2.03 ± 0.92	1.04 ± 0.88
	PPI	2.09 ± 0.95	1.24 ± 0.95

[a]Captured C, protein–ligand interface PLI, protein–protein interface PPI, surface S. $Hbond_{total}$ = all hydrogen bonds to protein. Ligand or other water molecules were considered. $Hbond_{H2O}$ = hydrogen bonds only to other water molecules were considered. stdev = one standard deviation.

Therefore, water molecules were further classified according to their position in the biological complex (S, C, PLI, PPI, see the Materials and Methods section), allowing a more detailed view on their hydrogen bonding characteristics (see Figure 9). As expected, in all four categories the mean number of hydrogen bonds ($Hbond_{total}$) decreases from resolved to unresolved water molecules (see Table 5). The most drastic decrease can be seen for unresolved captured water molecules, which form about one hydrogen bond less in comparison to resolved captured water molecules. These results show that modeled water molecules in positions without experimental proof are less integrated into the hydrogen bonding network.

The likelihood of "missing partners" is a bias of surface water molecules. Either further shells of water molecules are not resolved and remain unmodeled or they might interact with neighboring protein chains that are not in the asymmetric unit considered here. Analyzing the number of bulk water accessible hydrogen bonding functions, surface water molecules have a mean of 1.69 and 2.06 accessible hydrogen bonding functions for resolved and insufficiently resolved water molecules. These numbers might be slightly overestimated, given that we tested whether 75% of the volume of a water molecule would fit in the ideal direction of a hydrogen bonding function. In total this leads to 3.67 hydrogen bonds for both resolved as well as unresolved surface water molecules, close to the number of about 3.5 hydrogen bonds on average in bulk at 298 K.[56−58]

Hydrophobic Bubbles. Surprisingly, we found captured water molecules that are resolved by electron density ($mean_{EDIA}$ = 0.88 ± 0.56) but do not form any hydrogen bonds with protein, ligand atoms, or water molecules (Figures 9a and 10). These water molecules resemble so-called hydrophobic bubbles.[40] In total, only 1438 (0.54%) of all captured water molecules are hydrophobic bubbles sufficiently resolved by electron density (0.06% of the whole data set). They are highly constrained in their position inside the protein, which is probably the reason why they are resolved by electron density; but display a higher thermal motion than PPI, PLI, or other captured water molecules and about the same B-factor as surface water molecules (see Table 7). Since these hydrophobic bubbles must be highly energetically unfavorable, as they are spatially constrained and cannot compensate the enthalpic loss by hydrogen bond formation, they might present predetermined breaking points of protein structures. An alternative explanation for some of the hydrophobic bubbles would be that the electron density comes from a noble gas, which was used to solve the phase problem.[59−61] The electron density of a noble gas would be hardly differentiable from a water molecule especially if the position is only partially occupied.

Well-Integrated Unresolved Water Molecules. Another interesting result is given by water molecules that are not resolved ($EDIA < EDIA_{Thrs}$) but are very well integrated into a hydrogen bonding network forming three or four hydrogen bonds to protein or ligand (see Figure 11). In total 769 of those are found in our data set (referred to as $WIND_{PL}$). The number increases substantially to 34 217 if surrounding water molecules are considered as a hydrogen bonding partner (referred to as $WIND_{all}$). Simply accounting for the high number of formed hydrogen bonds, the water molecule would be expected to be resolved by electron density. However, there is little or no experimental proof for the water molecule to be in this place. One of the reasons for the lack of electron density might be their hydrogen bonding partner. Compared to resolved water molecules, $WIND_{all}$ water molecules build 13% to 40% more

DOI: 10.1021/ci500662d
J. Chem. Inf. Model. 2015, 55, 771−783

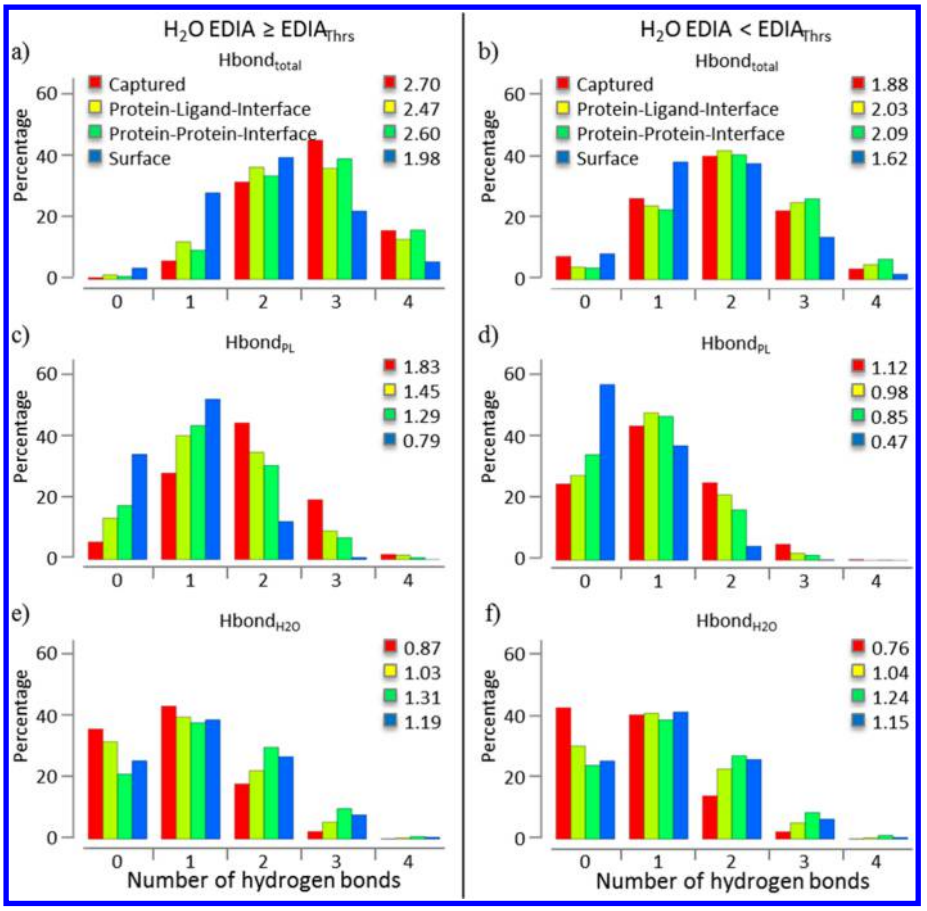


Figure 9. Histograms of hydrogen bonds formed by all water molecules from the data set. Data is separated for each class of water molecules, each class represents 100% of its water molecules. Hydrogen bonds of water molecules (a, c, e) with electron density (EDIA ≥ $EDIA_{Thrs}$) and (b, d, f) without electron density (EDIA < $EDIA_{Thrs}$). (a and b) All hydrogen bonds (to protein, ligand, and other water molecules) are counted. (c and d) Only hydrogen bonds to explicit partners, protein or ligand, are counted. (e and f) Only hydrogen bonds to other water molecules are counted. Numbers in the legend are the mean number of hydrogen bonds for each class.

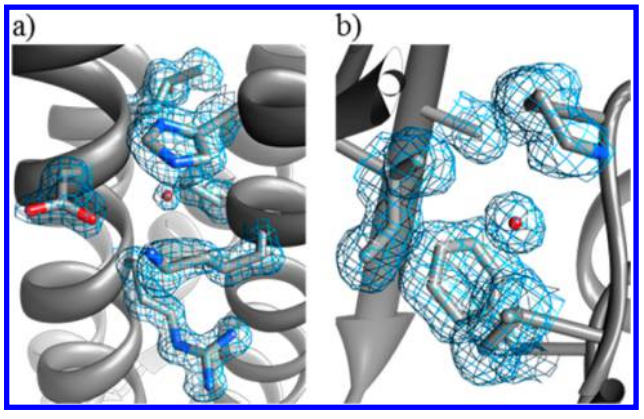


Figure 10. Examples for hydrophobic bubbles. (a) H_2O-I-1613, EDIA = 0.45 (PDB ID: 3ak8) and (b) H_2O-A-535, EDIA = 1.58 (PDB ID: 4h5i): blue = $2f_o - f_c$ map at 1σ (molecular graphics were created using UCSF Chimera[52]).

hydrogen bonds to other water molecules (see Table 8). These water molecules may not have a favored position, since they interact less often with amino acids, and their molecular motion might therefore be too high to be detected during X-ray crystallography (see Table 7). A closer look at the 769 $WIND_{PL}$ as well as $WIND_{all}$ water molecules and their surroundings often reveals highly flexible regions in which the water molecules are incorporated (see Table 7).

Hydrogen Bonding Partners and Proximity Preferences of Water Molecules. Classified water molecules were further analyzed concerning hydrogen bonding partner preferences and their favored surroundings.

First, we analyzed whether water molecules primarily interact with backbone or side chain functional groups. Therefore, we calculated the ratio of hydrogen bonding functions from backbone to side chain in our high-resolution data set. The distribution of backbone hydrogen bonding functions as opposed to side chain ones is shifted toward the backbone by

Table 7. Average B-Factor for Water Molecules with EDIA ≥ $EDIA_{Thrs}$ (total, S, C, PLI, PPI, HB), Well-Integrated Water Molecules with EDIA < $EDIA_{Thrs}$[a] and for All High-Resolution Proteins

	total	S	C	PLI	PPI	$WIND_{all}$	$WIND_{PL}$	HB	Protein
B-factor	27.42	29.35	18.35	22.71	23.40	44.03	45.60	28.97	15.66

[a]$WIND_{all}$ ≥ three hydrogen bonds to protein, ligand, or other water molecules; $WIND_{PL}$ ≥ three hydrogen bonds to protein or ligand. HB = hydrophobic bubbles.

DOI: 10.1021/ci500662d
J. Chem. Inf. Model. 2015, 55, 771−783

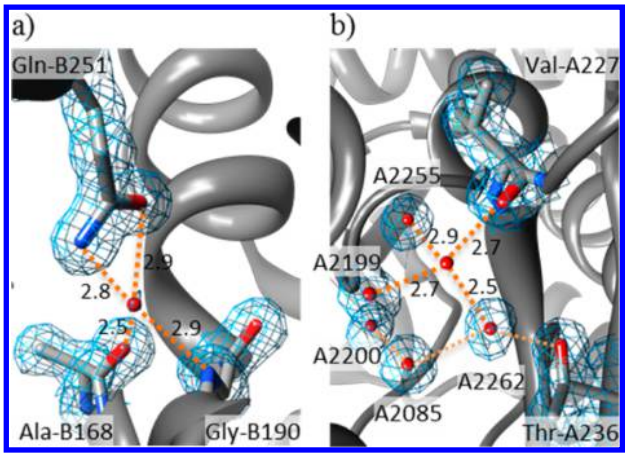


Figure 11. Examples for well-integrated, unresolved (EDIA < $EDIA_{Thrs}$) water molecules (a) H_2O-B-2135, EDIA = 0.16 (PDB ID: 4bgb) and (b) H_2O-A-2254, EDIA = 0.00 (PDB ID: 1of8); blue = $2f_o - f_c$ map at 1σ, orange dashed lines = hydrogen bonds with distances in angstroms (molecular graphics were created using UCSF Chimera[52]).

Table 8. Hydrogen Bonding Partners of Water Molecules with Sufficient Electron Density (C, PLI, PPI, S: EDIA ≥ $EDIA_{Thrs}$) Compared to Well-Integrated Unresolved Water Molecules[a]

	Hbond to H_2O	Hbond to PL
C	32.27%	67.73%
PLI	41.55%	58.45%
PPI	50.37%	49.63%
S	59.98%	40.02%
$WIND_{all}$	73.06%	26.94%
$WIND_{PL}$	6.93%	93.07%

[a]$WIND_{all}$ ≥ three hydrogen bonds to protein, ligand, or other water molecules; $WIND_{PL}$ ≥ three hydrogen bonds to protein or ligand; EDIA < $EDIA_{Thrs}$; Hbond = hydrogen bond; PL = hydrogen bond formed with either protein or ligand.

65%. Most of the hydrogen bonding functions of the protein backbone are satisfied due to their secondary structure patterns (α-helices and β-sheets). Therefore, it is not surprising, that PLI and PPI water molecules are more likely to interact with amino acid side chains (see Table 9). Especially PLI water molecules satisfy hydrogen bonding functions of side chains, whereas captured water molecules preferably interact with the protein backbone. The latter was also found in previous studies.[39,62] We found that PPI water molecules slightly favor side chain interaction (52%), however to a smaller extent than estimated by Ahmed et al. (78.5%)[40] but more than found by Rodier (45%)[10].

Second, atom-based preferences were dissected. It is noticeable that all water molecules are more likely to interact with oxygen atoms than the proportion of nitrogen to oxygen atoms of the high-resolution proteins would suggest. Similar to the backbone to side chain preferences, captured and PLI water molecules display the most extremes, with PLI water molecules interacting by 73% with oxygen atoms (See Table 9).

Third, since water molecules have two acceptor and donor functions each, we examined whether they do have any bias toward acceptor or donor interaction partners. Herein, the proportion is for all water molecules highly similar to the acceptor/donor distribution of the high-resolution proteins. Only PLI water molecules and well-integrated water molecules with insufficient electron density (referred to as $WIND_{all}$) show a greater bias toward acceptor interaction partners than water molecules from the other categories (see Table 9).

The last two results are directly connected to each other. Most of the time oxygen atoms are acceptors, wherein nitrogen atoms are more often present as donors. The proportion of oxygen acceptors in the high-resolution data set is nearly double the amount of nitrogen donors, 18 times the number of oxygen donors, and more than 100 times the amount of nitrogen acceptors. Therefore, it is most likely that water molecules interact with oxygen acceptors. Additionally, these results are in accordance with previous findings using quantum mechanical calculations.[63] Interactions from water molecules to either nitrogen or oxygen do not lead to significant differences in the estimated binding energy. However, water molecules interacting with acidic groups leads to an increase in binding affinity. This correlates well with the finding, that glutamate and aspartate are frequent interaction partners (see Table 10).

Given the different classes of water molecules we further investigated their functional group and amino acid preferences. Six different groups were formed, wherein each corresponds to one or more amino acid types (see Table 10). The preferences were then compared to the mean occurrence of amino acids in the high-resolution protein structures. Most noticeable is the high proportion of hydrogen bonds of water molecules to aspartate and glutamate residues (for C, PLI, and PPI). This probability is compared to the normal occurrence of aspartate and glutamate residues in proteins, which are the most abundant ones. Within protein−protein interfaces water molecules have a high probability to interact with arginine residues. Herein, the frequency is nearly as high as the normal occurrence of arginine residues within the protein. These findings are consistent with previous studies.[10,40] The results

Table 9. Hydrogen Bonding Partner Preferences of Water Molecules with Sufficient Density (total, C, PLI, PPI, S: EDIA ≥ $EDIA_{Thrs}$) Compared to Well-Integrated Unresolved Water Molecules[a] and the Respective Occurrence in the High-Resolution Proteins

	total	S	C	PLI	PPI	$WIND_{all}$	$WIND_{PL}$	protein
BB/SC	0.529	0.511	0.605	0.473	0.483	0.398	0.457	0.653
N/O	0.312	0.311	0.325	0.275	0.316	0.251	0.385	0.471
Don/Acc	0.325	0.322	0.341	0.293	0.329	0.254	0.396	0.363

[a]$WIND_{all}$ ≥ three hydrogen bonds to protein, ligand, or other water molecules; $WIND_{PL}$ ≥ three hydrogen bonds to protein or ligand; EDIA < $EDIA_{Thrs}$; BB/SC = ratio of backbone to side chain interactions; N/O = ratio of interactions to nitrogen or oxygen atoms of the protein, Don/Acc = ratio of hydrogen bonds to donor or acceptor functions of the protein. A ratio of 0.5 means equal distribution of the interaction partners. For the different categories of water molecules hydrogen bond partners are considered, wherein for the protein column available functions/atoms of the high-resolution data set are counted.

DOI: 10.1021/ci500662d
J. Chem. Inf. Model. 2015, 55, 771−783

Table 10. Functional Group and Corresponding Amino Acid Preferences of Water Molecules with Sufficient Density (total, C, PLI, PPI, S: EDIA ≥ $EDIA_{Thrs}$) Compared to Well-Integrated Unresolved Water Molecules[a] and the Respective Occurrence in the High-Resolution Proteins

interaction partner		total	S	C	PLI	PPI	$WIND_{all}$	$WIND_{PL}$	protein
functional group (amino acid)	amide (N, Q)	4	3	5	4	4	3	8	8
	imidazole/indole (H/W)	1	1	2	1	1	0	1	4
	amine (K)	2	2	1	2	2	2	8	6
	carboxyl (D, E)	7	7	8	8	8	6	13	12
	guanidine (R)	2	2	3	2	4	2	9	5
	hydroxyl (S, T, Y)	5	4	8	6	6	3	12	15
hydrophobic amino acids		–	–	–	–	–	–	–	50
ligand		1	0	0	16	0	0	0	–
H_2O		54	60	32	42	50	73	7	–
backbone		24	20	41	20	24	10	41	–

[a]$WIND_{all}$ ≥ three hydrogen bonds to protein, ligand, or other water molecules; $WIND_{PL}$ ≥ three hydrogen bonds to protein or ligand; EDIA < $EDIA_{Thrs}$. For the water molecules hydrogen bond partners are considered (in percent), wherein for the protein the mean proportion of the respecting functional group is given (in percent). (–) No contribution.

Table 11. Proximity Preferences of Water Molecules with Sufficient Density (total, C, PLI, PPI, S: EDIA ≥ $EDIA_{Thrs}$) Compared to Well-Integrated Unresolved Water Molecules[a]

	total	S	C	PLI	PPI	$WIND_{all}$	$WIND_{PL}$
hydrophobicity	0.617	0.613	0.639	0.607	0.625	0.594	0.609
hydrophobic surface	0.567	0.579	0.513	0.505	0.544	0.502	0.298
water cluster size	18	17	13	17	31	92	13

[a]$WIND_{all}$ ≥ three hydrogen bonds to protein, ligand, or other water molecules; $WIND_{PL}$ ≥ three hydrogen bonds to protein or ligand; EDIA < $EDIA_{Thrs}$.

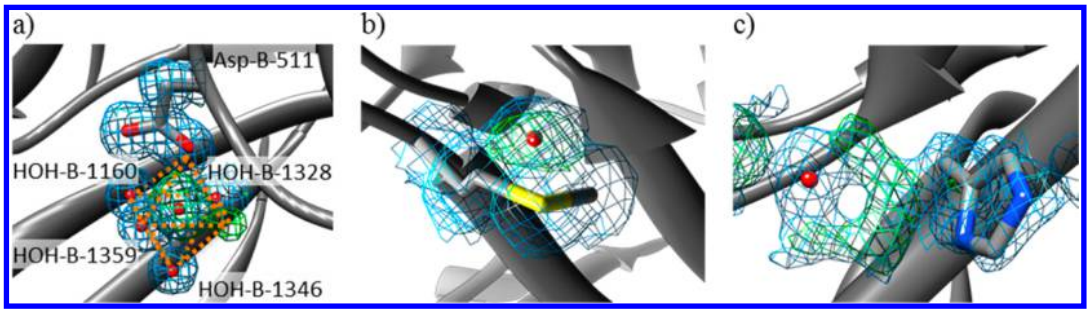


Figure 12. Modeling errors. (a) H_2O-B-1166 modeled into electron density that would better be suited for an ion (EDIA = 5.83, electron density difference map ($f_o - f_c$) indicates too little electrons modeled). Orange lines indicate possible ion coordination geometry (PDB ID: 1hyo). (b) H_2O-A-352 modeled into the electron density of methionine A-239 (distance 1.725 Å). The electron density difference map indicates missing electrons at the water position and too many electrons at the sulfur of methionine, EDIA = 4.38 (PDB ID: 2rdq). (c) H_2O-B-1328 modeled into alternative site chain conformation of Histidine B-985, EDIA = 1.73 (PDB ID: 1hyo). blue = $2f_o - f_c$ map at 1σ, green = $f_o - f_c$ map at 3σ (molecular graphics were created using UCSF Chimera[52]).

are in accordance with the above analysis, in which oxygen atoms as well as acceptors dominated the interaction partners of PLI water molecules. Only well-integrated water molecules with insufficient electron density (referred to as $WIND_{PL}$) show a higher number of hydrogen bonds to carboxyl groups. Interestingly, the latter water molecules also have more interactions with lysine and arginine than their mean occurrence in the high-resolution protein structures (see Table 10 $WIND_{PL}$). Another noticeable result is the small number of hydrogen bonds of surface water molecules with histidine or tryptophan, in accordance with previous studies that have shown that those amino acids are found less often on protein surfaces.[64,65]

Finally, water molecules were analyzed for proximity preferences as described in the Materials and Methods section. The hydrophobicity of their direct surrounding, their hydrophobic surface area, and the size of water clusters was examined (see Table 11). Noticeable is the relatively high proportion of hydrophobic atoms in the surrounding of captured water molecules, while the hydrophobic surface around them is comparably low. The latter characteristic allows captured water molecules to be well integrated into the protein complex. The difference between hydrophobicity and hydrophobic surface is even more remarkable, when looking at $WIND_{PL}$ water molecules with 0.609 and 0.298 respectively. This shows, in accordance to the very high B-factor, that those areas are highly flexible (see Figure 11b).

Water clusters have on average 18 water molecules for all water molecules with an EDIA value greater than $EDIA_{Thrs}$ (see Table 11). This number might appear quite large in comparison to previous studies.[66,67] However, previous analyses have focused on water clusters in cavities and not the whole protein. Additionally, the water clusters analyzed here have been detected based on a distance criterion only (see the Materials

DOI: 10.1021/ci500662d
J. Chem. Inf. Model. 2015, 55, 771–783

and Methods section). Interestingly, PPI water molecules show the biggest average number of water molecules within one water cluster among the four classes. Unsurprisingly, well-integrated, unresolved water molecules have a huge number of water molecules in one cluster if hydrogen bonds to water molecules are taken into account, relating again to their higher mobility.

Identification of Modeling Errors. Astonishingly, more than 1200 water molecules show very high EDIA values above median plus four standard deviations (EDIA > 3.3). Visual inspection of a random sample of 10% of those locations suggests modeling errors in over 75% (Figure 12a), with the electron density difference map showing too few electrons modeled in the position where the water molecule was placed. Most of those water molecules would better be substituted by ions, for which in at least 20% very good coordination geometries, such as octahedral or tetrahedral, can be found.

A further misinterpretation of the electron density was detected as water molecules may be built into the electron density of alternative amino acid configurations (Figure 12b and c). Identification of those modeling errors is more complex, as electron density supposed for other atoms is available. Automated identification of modeling errors or misinterpretation will be approached in future EDIA development.

CONCLUSION

Water molecules play an important role in many biological aspects, not only in mediating protein−ligand interactions, but also contributing fundamentally to binding affinity by dehydration and the hydrophobic effect. As those water molecules resolved by X-ray crystallography exist at local energy minima it would be advantageous to reliably predict those positions upon modeling molecular complexes.

In order to analyze the characteristics of water molecules a high-resolution subset consisting of 5485 structures from the PDB was compiled. Our evaluation has shown that high resolution itself is no guarantee for electron density support of each individual water molecule. Analyzing the electron density is unavoidable to differentiate between well resolved and unresolved water molecules. Therefore, a new measure based on electron density was developed, called EDIA. Advantages compared to already existing measurements, like B-factor, RSR, and RSCC are a direct comparison between modeled structure and electron density, as well as an intuitive interpretation of the value itself. Normalization by the standard deviation of the electron density map allows direct comparison of water molecules from different structures.

In order to detect misinterpretations of the electron density map and modeling errors the EDIA could be enhanced by taking the electron density difference map into account. In this way, further areas of too little or too many electrons in the modeled structure could be detected in an automated manner. Further water molecules misleadingly placed in electron density supposed for alternative amino acid or ligand conformations could be detected using a Difference of Gaussian filter.[68] Herein, two different sigma levels would be applied to the electron density map. Subtracting one image from the other preserves positions with drastic shifts, but discards all points that are at continuous areas, thus eliminating noise. As water molecules have a fairly circular, secluded distribution of electron density, it would become apparent if the underlying electron density, in which the water molecule is placed, is more stretched out and extensive as it is the case for amino acid side chains or ligands. Both aspects will be evaluated in our future development of EDIA.

The new measure EDIA can support the differentiation of water molecules that should be excluded from an analysis due to insufficient electron density support (EDIA < 0.24), from those that actually have implications for further modeling. Water molecules with unrealistically high EDIA values (EDIA > 3.3) need more attention due to a high probability of a wrong interpretation of the electron density. Even though very rarely, hydrophobic bubbles with good EDIA values were observed in this data set (1438 ≙ 0.52%) and showed that they may be of biological relevance. Otherwise, such highly unfavorable locations for water molecules would not be expected to exist. The observation from this evaluation provide further support for validating water molecules in crystal structures as well as implications for further characterization and modeling. In a subsequent analysis it would be highly interesting to investigate the underlying thermodynamics in order to understand why experimentally observed water molecules seem to be stable in their surroundings. This refers in particular to water molecules in a hydrophobic environment.

Many computational methods that aim to predict water molecule locations have been lately developed. However, little has been undertaken for the validation of water molecules in protein structures. As proven by the number of water molecules without electron density (208 052 ≙ 8.9% of the data set), a simple comparison with structurally modeled water molecules might not be sufficient. Herein, this well characterized high-resolution data set allows an extensive evaluation of water prediction methods, including the possibility to differentiate between water molecules well-resolved by electron density and those not supported by electron density.

In summary, the EDIA serves two purposes. First, properties and functions of meaningful modeled water molecules in crystal structures can be characterized and comprehended. Second, it can support the validation of computational methods for placing water molecules.

ASSOCIATED CONTENT

Supporting Information

Table S1: Data set of all PDBids. Table S2: Table of captured, PLI, and PPI water molecules including EDIA and diverse descriptors. This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author

*E-mail: rarey@zbh.uni-hamburg.de.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The project is part of the Biokatalyse2021 cluster and is funded by the Federal Ministry of Education and Research (BMBF) under grant 031A1838. We would like to thank our cooperation partner BioSolveIT GmbH, especially Holger Claussen and Christian Lemmen for fruitful discussions. Special thanks to Robert Klein for helpful information about the electron density map format. We also thank Karen Schomburg for critical reading and revising the manuscript.

DOI: 10.1021/ci500662d
J. Chem. Inf. Model. 2015, 55, 771−783

ABBREVIATIONS

EDIA, electron density for individual atoms; $2f_o - f_c$, composite electron density map; $f_o - f_c$, electron density difference map; MSA, molecular surface area; PLI, protein−ligand interface; PPI, protein−protein interface; RSCC, real-space correlation coefficient; RSR, real-space R-factor

REFERENCES

(1) Timasheff, S. N. The Control of Protein Stability and Association by Weak Interactions with Water: How Do Solvents Affect These Processes? *Annu. Rev. Biophys. Biomol. Struct.* **1993**, *22*, 67−97.

(2) Zhang, L.; Yang, Y.; Kao, Y.-T.; Wang, L.; Zhong, D. Protein Hydration Dynamics and Molecular Mechanism of Coupled Water-Protein Fluctuations. *J. Am. Chem. Soc.* **2009**, *131*, 10677−10691.

(3) Levy, Y.; Onuchic, J. N. Water Mediation in Protein Folding and Molecular Recognition. *Annu. Rev. Biophys. Biomol. Struct.* **2006**, *35*, 389−415.

(4) Mattos, C. Protein−water Interactions in a Dynamic World. *Trends Biochem. Sci.* **2002**, *27*, 203−208.

(5) Ahmad, S.; Kamal, M. Z.; Sankaranarayanan, R.; Rao, N. M. Thermostable Bacillus Subtilis Lipases: In Vitro Evolution and Structural Insight. *J. Mol. Biol.* **2008**, *381*, 324−340.

(6) Kawasaki, K.; Kondo, H.; Suzuki, M.; Ohgiya, S.; Tsuda, S. Alternate Conformations Observed in Catalytic Serine of Bacillus Subtilis Lipase Determined at 1.3 Å Resolution. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **2002**, *58*, 1168−1174.

(7) Rühlmann, A.; Kukla, D.; Schwager, P.; Bartels, K.; Huber, R. Structure of the Complex Formed by Bovine Trypsin and Bovine Pancreatic Trypsin Inhibitor. *J. Mol. Biol.* **1973**, *77*, 417−436.

(8) Phillips, R. S. How Does Active Site Water Affect Enzymatic Stereorecognition? *J. Mol. Catal. B Enzym.* **2002**, *19*−*20*, 103−107.

(9) Langhorst, U.; Backmann, J.; Loris, R.; Steyaert, J. Analysis of a Water Mediated Protein−Protein Interactions within RNase T1. *Biochemistry* **2000**, *39*, 6586−6593.

(10) Rodier, F.; Bahadur, R. P.; Chakrabarti, P.; Janin, J. Hydration of Protein-Protein Interfaces. *Proteins* **2005**, *60*, 36−45.

(11) Janin, J. Wet and Dry Interfaces: The Role of Solvent in Protein−protein and protein−DNA Recognition. *Structure* **1999**, *7*, R277−R279.

(12) Ahmad, M.; Gu, W.; Geyer, T.; Helms, V. Adhesive Water Networks Facilitate Binding of Protein Interfaces. *Nat. Commun.* **2011**, *2*, 261.

(13) Chothia, C.; Janin, J. Principles of Protein−protein Recognition. *Nature* **1975**, *256*, 705−708.

(14) Young, L.; Jernigan, R. L.; Covell, D. G. A Role for Surface Hydrophobicity in Protein-Protein Recognition. *Protein Sci.* **1994**, *3*, 717−729.

(15) Tsai, C. J.; Lin, S. L.; Wolfson, H. J.; Nussinov, R. Studies of Protein-Protein Interfaces: A Statistical Analysis of the Hydrophobic Effect. *Protein Sci.* **1997**, *6*, 53−64.

(16) Chandler, D. Interfaces and the Driving Force of Hydrophobic Assembly. *Nature* **2005**, *437*, 640−647.

(17) Shimokhina, N.; Bronowska, A.; Homans, S. W. Contribution of Ligand Desolvation to Binding Thermodynamics in a Ligand-Protein Interaction. *Angew. Chem., Int. Ed. Engl.* **2006**, *45*, 6374−6376.

(18) Snyder, P. W.; Mecinovic, J.; Moustakas, D. T.; Thomas, S. W.; Harder, M.; Mack, E. T.; Lockett, M. R.; Héroux, A.; Sherman, W.; Whitesides, G. M. Mechanism of the Hydrophobic Effect in the Biomolecular Recognition of Arylsulfonamides by Carbonic Anhydrase. *Proc. Natl. Acad. Sci. U. S. A.* **2011**, *108*, 17889−17894.

(19) Biela, A.; Nasief, N. N.; Betz, M.; Heine, A.; Hangauer, D.; Klebe, G. Dissecting the Hydrophobic Effect on the Molecular Level: The Role of Water, Enthalpy, and Entropy in Ligand Binding to Thermolysin. *Angew. Chem., Int. Ed. Engl.* **2013**, *52*, 1822−1828.

(20) Ladbury, J. E. Just Add Water! The Effect of Water on the Specificity of Protein-Ligand Binding Sites and Its Potential Application to Drug Design. *Chem. Biol.* **1996**, *3*, 973−980.

(21) Biela, A.; Khayat, M.; Tan, H.; Kong, J.; Heine, A.; Hangauer, D.; Klebe, G. Impact of Ligand and Protein Desolvation on Ligand Binding to the S1 Pocket of Thrombin. *J. Mol. Biol.* **2012**, *418*, 350−366.

(22) Chen, J. M.; Xu, S. L.; Wawrzak, Z.; Basarab, G. S.; Jordan, D. B. Structure-Based Design of Potent Inhibitors of Scytalone Dehydratase: Displacement of a Water Molecule from the Active Site. *Biochemistry* **1998**, *37*, 17735−17744.

(23) Wissner, A.; Berger, D. M.; Boschelli, D. H.; Floyd, M. B.; Greenberger, L. M.; Gruber, B. C.; Johnson, B. D.; Mamuya, N.; Nilakantan, R.; Reich, M. F.; Shen, R.; Tsou, H.-R.; Upeslacis, E.; Wang, Y. F.; Wu, B.; Ye, F.; Zhang, N. 4-Anilino-6,7-Dialkoxyquinoline-3-Carbonitrile Inhibitors of Epidermal Growth Factor Receptor Kinase and Their Bioisosteric Relationship to the 4-Anilino-6,7-Dialkoxyquinazoline Inhibitors. *J. Med. Chem.* **2000**, *43*, 3244−3256.

(24) Seo, J.; Igarashi, J.; Li, H.; Martasek, P.; Roman, L. J.; Poulos, T. L.; Silverman, R. B. Structure-Based Design and Synthesis of N(omega)-Nitro-L-Arginine-Containing Peptidomimetics as Selective Inhibitors of Neuronal Nitric Oxide Synthase. Displacement of the Heme Structural Water. *J. Med. Chem.* **2007**, *50*, 2089−2099.

(25) Zhang, L.; Hermans, J. Hydrophilicity of Cavities in Proteins. *Proteins* **1996**, *24*, 433−438.

(26) García-Sosa, A. T.; Mancera, R. L.; Dean, P. M. WaterScore: A Novel Method for Distinguishing between Bound and Displaceable Water Molecules in the Crystal Structure of the Binding Site of Protein-Ligand Complexes. *J. Mol. Model.* **2003**, *9*, 172−182.

(27) Barillari, C.; Taylor, J.; Viner, R.; Essex, J. W. Classification of Water Molecules in Protein Binding Sites. *J. Am. Chem. Soc.* **2007**, *129*, 2577−2587.

(28) Amadasi, A.; Surface, J. A.; Spyrakis, F.; Cozzini, P.; Mozzarelli, A.; Kellogg, G. E. Robust Classification of "Relevant" Water Molecules in Putative Protein Binding Sites. *J. Med. Chem.* **2008**, *51*, 1063−1067.

(29) Adams, D. J. Grand Canonical Ensemble Monte Carlo for a Lennard-Jones Fluid. *Mol. Phys.* **1975**, *29*, 307−311.

(30) Goodford, P. J. A Computational Procedure for Determining Energetically Favorable Binding Sites on Biologically Important Macromolecules. *J. Med. Chem.* **1985**, *28*, 849−857.

(31) Grant, J. A.; Pickup, B. T.; Nicholls, A. A Smooth Permittivity Function for Poisson-Boltzmann Solvation Methods. *J. Comput. Chem.* **2001**, *22*, 608−640.

(32) Abel, R.; Young, T.; Farid, R.; Berne, B. J.; Friesner, R. A. Role of the Active-Site Solvent in the Thermodynamics of Factor Xa Ligand Binding. *J. Am. Chem. Soc.* **2008**, *130*, 2817−2831.

(33) Abel, R.; Salam, N. K.; Shelley, J.; Farid, R.; Friesner, R. A.; Sherman, W. Contribution of Explicit Solvent Effects to the Binding Affinity of Small-Molecule Inhibitors in Blood Coagulation Factor Serine Proteases. *ChemMedChem.* **2011**, *6*, 1049−1066.

(34) Michel, J.; Tirado-Rives, J.; Jorgensen, W. L. Prediction of the Water Content in Protein Binding Sites. *J. Phys. Chem. B* **2009**, *113*, 13337−13346.

(35) Nguyen, C. N.; Young, T. K.; Gilson, M. K. Grid Inhomogeneous Solvation Theory: Hydration Structure and Thermodynamics of the Miniature Receptor cucurbit[7]uril. *J. Chem. Phys.* **2012**, *137*, 044101.

(36) Ross, G. A.; Morris, G. M.; Biggin, P. C. Rapid and Accurate Prediction and Scoring of Water Molecules in Protein Binding Sites. *PLoS One* **2012**, *7*, e32036.

(37) Poornima, C. S.; Dean, P. M. Hydration in Drug Design. 1. Multiple Hydrogen-Bonding Features of Water Molecules in Mediating Protein-Ligand Interactions. *J. Comput. Aided. Mol. Des.* **1995**, *9*, 500−512.

(38) Lu, Y.; Wang, R.; Yang, C.-Y.; Wang, S. Analysis of Ligand-Bound Water Molecules in High-Resolution Crystal Structures of Protein-Ligand Complexes. *J. Chem. Inf. Model.* **2007**, *47*, 668−675.

(39) Park, S.; Saven, J. G. Statistical and Molecular Dynamics Studies of Buried Waters in Globular Proteins. *Proteins* **2005**, *60*, 450−463.

(40) Ahmed, M. H.; Spyrakis, F.; Cozzini, P.; Tripathi, P. K.; Mozzarelli, A.; Scarsdale, J. N.; Safo, M. A.; Kellogg, G. E. Bound

DOI: 10.1021/ci500662d
J. Chem. Inf. Model. 2015, 55, 771−783

Water at Protein-Protein Interfaces: Partners, Roles and Hydrophobic Bubbles as a Conserved Motif. *PLoS One* **2011**, *6*, e24712.
(41) Dunitz, J. D. The Entropic Cost of Bound Water in Crystals and Biomolecules. *Science* **1994**, *264*, 670.
(42) Huggins, D. J. Benchmarking the Thermodynamic Analysis of Water Molecules around a Model Beta Sheet. *J. Comput. Chem.* **2012**, *33*, 1383−1392.
(43) Huggins, D. J. Application of Inhomogeneous Fluid Solvation Theory to Model the Distribution and Thermodynamics of Water Molecules around Biomolecules. *Phys. Chem. Chem. Phys.* **2012**, *14*, 15106−15117.
(44) Beuming, T.; Che, Y.; Abel, R.; Kim, B.; Shanmugasundaram, V.; Sherman, W. Thermodynamic Analysis of Water Molecules at the Surface of Proteins and Applications to Binding Site Prediction and Characterization. *Proteins* **2012**, *80*, 871−883.
(45) Kinoshita, M. Importance of Translational Entropy of Water in Biological Self-Assembly Processes like Protein Folding. *Int. J. Mol. Sci.* **2009**, *10*, 1064−1080.
(46) Jones, T. A.; Zou, J. Y.; Cowan, S. W.; Kjeldgaard, M. Improved Methods for Building Protein Models in Electron Density Maps and the Location of Errors in These Models. *Acta Crystallogr. Sect. A Found. Crystallogr.* **1991**, *47*, 110−119.
(47) Hawkins, P. C. D.; Kelley, B. P.; Warren, G. L. The Application of Statistical Methods to Cognate Docking: A Path Forward? *J. Chem. Inf. Model.* **2014**, *54*, 1339−1355.
(48) Levitt, M.; Park, B. H. Water: Now You See It, Now You Don't. *Structure* **1993**, *1*, 223−226.
(49) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235−242.
(50) Kleywegt, G. J.; Harris, M. R.; Zou, J. Y.; Taylor, T. C.; Wählby, A.; Jones, T. A. The Uppsala Electron-Density Server. *Acta Crystallogr. D. Biol. Crystallogr.* **2004**, *60*, 2240−2249.
(51) Bietz, S.; Urbaczek, S.; Schulz, B.; Rarey, M. Protoss: A Holistic Approach to Predict Tautomers and Protonation States in Protein-Ligand Complexes. *J. Cheminform.* **2014**, *6*, 12.
(52) Pettersen, E. F.; Goddard, T. D.; Huang, C. C.; Couch, G. S.; Greenblatt, D. M.; Meng, E. C.; Ferrin, T. E. UCSF Chimera–a Visualization System for Exploratory Research and Analysis. *J. Comput. Chem.* **2004**, *25*, 1605−1612.
(53) Ko, T.-P.; Robinson, H.; Gao, Y.-G.; Cheng, C.-H. C.; DeVries, A. L.; Wang, A. H.-J. The Refined Crystal Structure of an Eel Pout Type III Antifreeze Protein RD1 at 0.62-A Resolution Reveals Structural Microheterogeneity of Protein and Solvation. *Biophys. J.* **2003**, *84*, 1228−1237.
(54) Karplus, P. A.; Faerman, C. Ordered Water in Macromolecular Structure. *Curr. Opin. Struct. Biol.* **1994**, *4*, 770−776.
(55) Carugo, O.; Bordo, D. How Many Water Molecules Can Be Detected by Protein Crystallography? *Acta Crystallogr. Sect. D Biol. Crystallogr.* **1999**, *55*, 479−483.
(56) Hoffmann, M. M.; Conradi, M. S. Are There Hydrogen Bonds in Supercritical Water? *J. Am. Chem. Soc.* **1997**, *119*, 3811−3817.
(57) Soper, A. K.; Bruni, F.; Ricci, M. A. Site−site Pair Correlation Functions of Water from 25 to 400 °C: Revised Analysis of New and Old Diffraction Data. *J. Chem. Phys.* **1997**, *106*, 247.
(58) Wernet, P.; Nordlund, D.; Bergmann, U.; Cavalleri, M.; Odelius, M.; Ogasawara, H.; Näslund, L. A.; Hirsch, T. K.; Ojamäe, L.; Glatzel, P.; Pettersson, L. G. M.; Nilsson, A. The Structure of the First Coordination Shell in Liquid Water. *Science* **2004**, *304*, 995−999.
(59) Tilton, R. F.; Kuntz, I. D.; Petsko, G. A. Cavities in Proteins: Structure of a Metmyoglobin Xenon Complex Solved to 1.9.ANG. *Biochemistry* **1984**, *23*, 2849−2857.
(60) Prangé, T.; Schiltz, M.; Pernot, L.; Colloc'h, N.; Longhi, S.; Bourguet, W.; Fourme, R. Exploring Hydrophobic Sites in Proteins with Xenon or Krypton. *Proteins* **1998**, *30*, 61−73.
(61) Schiltz, M.; Fourme, R.; Prangé, T. Use of Noble Gases Xenon and Krypton as Heavy Atoms in Protein Structure Determination. *Methods Enzymol.* **2003**, *374*, 83−119.
(62) Williams, M. A.; Goodfellow, J. M.; Thornton, J. M. Buried Waters and Internal Cavities in Monomeric Proteins. *Protein Sci.* **1994**, *3*, 1224−1235.
(63) Rezáč, J.; Riley, K. E.; Hobza, P. S66: A Well-Balanced Database of Benchmark Interaction Energies Relevant to Biomolecular Structures. *J. Chem. Theory Comput.* **2011**, *7*, 2427−2438.
(64) Miller, S.; Janin, J.; Lesk, A. M.; Chothia, C. Interior and Surface of Monomeric Proteins. *J. Mol. Biol.* **1987**, *196*, 641−656.
(65) Fukuchi, S.; Nishikawa, K. Protein Surface Amino Acid Compositions Distinctively Differ between Thermophilic and Mesophilic Bacteria. *J. Mol. Biol.* **2001**, *309*, 835−843.
(66) Yin, H.; Hummer, G.; Rasaiah, J. C. Metastable Water Clusters in the Nonpolar Cavities of the Thermostable Protein Tetrabrachion. *J. Am. Chem. Soc.* **2007**, *129*, 7369−7377.
(67) Vaitheeswaran, S.; Yin, H.; Rasaiah, J. C.; Hummer, G. Water Clusters in Nonpolar Cavities. *Proc. Natl. Acad. Sci. U. S. A.* **2004**, *101*, 17002−17005.
(68) Marr, D.; Hildreth, E. Theory of Edge Detection. *Proc. R. Soc. London B. Biol. Sci.* **1980**, *207*, 187−217.

DOI: 10.1021/ci500662d
J. Chem. Inf. Model. 2015, 55, 771−783