JOURNAL OF CHEMICAL INFORMATION AND MODELING

ARTICLE

pubs.acs.org/jcim

Molecular Docking with Ligand Attached Water Molecules

Mette A. Lie,[*,†,§] René Thomsen,[‡] Christian N. S. Pedersen,[†,∥] Birgit Schiøtt,[*,§] and Mikael H. Christensen[‡]

[†]Bioinformatics Research Centre (BiRC), Faculty of Science and Technology, Aarhus University, Denmark

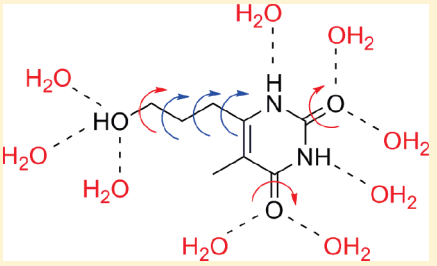
[‡]Molegro ApS, Aarhus, Denmark

[§]Centre for Insoluble Protein Structures (*in*SPIN) and Interdisciplinary Nanoscience Centre (*i*NANO), Department of Chemistry, Aarhus University, Denmark

[∥]Centre for Membrane Pumps in Cells and Disease (PUMPKIN), Department of Computer Science, Aarhus University, Denmark

Supporting Information

ABSTRACT: A novel approach to incorporate water molecules in protein−ligand docking is proposed. In this method, the water molecules display the same flexibility during the docking simulation as the ligand. The method solvates the ligand with the maximum number of water molecules, and these are then retained or displaced depending on energy contributions during the docking simulation. Instead of being a static part of the receptor, each water molecule is a flexible on/off part of the ligand and is treated with the same flexibility as the ligand itself. To favor exclusion of the water molecules, a constant entropy penalty is added for each included water molecule. The method was evaluated using 12 structurally diverse protein−ligand complexes from the PDB, where several water molecules bridge the ligand and the protein. A considerable improvement in successful docking simulations was found when including flexible water molecules solvating hydrogen bonding groups of the ligand. The method has been implemented in the docking program Molegro Virtual Docker (MVD).



INTRODUCTION

A protein molecule and a ligand in solution are covered by water molecules. When a ligand and a protein receptor form a complex, desolvation must take place, and the ligand and the protein interact through direct interactions. Also, in some cases, contacts are mediated through discrete water molecules. These water molecules stabilize the protein−ligand complex by forming a hydrogen-bonded network, mediating the interactions between the ligand and the receptor.[1−9] Furthermore, a hydrogen-bonded network of water molecules may stabilize the complex formed with one ligand but not another, thereby contributing to the specificity of ligand recognition.

Several protein−ligand docking studies have been performed to elucidate that the presence of water molecules in a ligand binding site plays a key role in protein−ligand recognition. In 2008, Murray and co-workers[10] published a docking study where the inclusion of all crystallographic water molecules within 6.0 Å of any ligand atom resulted in a large increase in docking accuracy. By including all water molecules in the binding site, the search space is however biased toward the correct binding mode, which is also discussed in the paper.[10]

In 2008, Mancera and co-workers[11] published a comprehensive redocking study to investigate the importance of water molecules for the accuracy of protein−ligand docking predictions. In that study, any crystal water molecule that is capable of forming mediating hydrogen bonds between the ligand and the receptor, i.e., any crystallographic water within 2.5 Å from any ligand atom and 3.0 Å from any protein atom, was included as a static part of the receptor structure. The study found that the efficacy of the docking simulations and the accuracy of the docking predictions were significantly improved with the inclusion of the crystallographic water molecules in the binding site. The redocking study[11] was in 2010 followed by a cross-docking study.[12] Six different protein targets with between three and 13 available protein−ligand PDB structures were considered. For each of the six targets, a common set of the crystallographic water molecules was found. Also in this study, a significant improvement in the accuracy of the predicted binding modes was observed with the inclusion of the conserved water molecules.

In the study by Murray and co-workers,[10] the presence of all crystallographic water molecules in the binding site biased the search space by physically restricting the number of possible binding modes. In both studies by Mancera and co-workers, the redocking study[11] as well as the cross-docking study,[12] all water molecules that were capable of interacting with both the ligand and the receptor protein were included by indiscriminately including all water molecules that fulfilled the distance criteria to both ligand and protein, and on that basis, it was concluded that water molecules play a key role in protein−ligand recognition because the efficacy and accuracy of the docking simulations were improved. Thus, the results obtained from all three of the mentioned studies are artificially constrained by including all nearby water molecules,[10] all possibly mediating water molecules,[11] or all conserved possibly mediating water molecules.[12] However, despite these limitations, it is obvious

Received: December 30, 2010
Published: March 31, 2011

that water molecules in some cases are important for the binding of a ligand in a protein receptor. Water molecules are thus important in structure-based drug design, where the binding affinity of ligand molecules can be improved by mimicking, displacing, and targeting bound water molecules.[13]

When a water molecule is known or assumed to play a role in protein−ligand recognition, a simple way to incorporate it in a docking problem is to perform one docking simulation, where the water molecule is included as a static part of the receptor structure and another where it is absent. This strategy is feasible if only a few water molecules are potentially important, but when n water molecules are assumed to play a role in protein−ligand recognition, this approach will ultimately sum up to 2^n separate docking simulations in parallel. An alternative is a displaceable water model where the individual crystal water molecules from the PDB structure can be toggled on/off automatically during the docking simulation, so that a ligand can keep favorable water molecules and displace nonfavorable water molecules. A displaceable water approach was implemented in the docking program GOLD[14] in 2005 as well as in the docking program Molegro Virtual Docker (MVD).[15]

The docking program FlexX[16] introduced a method for placing water molecules during the protein−ligand docking simulation, named The Particle Concept.[17] The Particle Concept places water molecules between the ligand and the protein during the docking simulation and allows the water molecules to form hydrogen bonds and steric interactions to both the ligand and the protein. In a preprocessing phase, energetically favorable positions for water molecules inside the active site of the protein are calculated. During the docking calculation, the water molecules are then switched on and off at these positions, and thus, the positions of the final water molecules are decided solely on the basis of the protein. The overall improvement of the docking results using The Particle Concept is small (27.5% of the test cases improve, 23% of them deteriorate).

In the aforementioned approaches, the included water molecules, which are either part of the X-ray crystal structure or placed on potential water binding sites that can be predicted by a number of programs,[18−25] are static during the docking simulation, although they are sometimes displaceable.[14,15] What we believe is missing so far is an approach where the water molecules do not hold a static position defined by the protein receptor structure but instead display the same dynamics as the ligand molecule. We thus propose a model where water molecules are attached to the ligand molecule in such a way that they can be displaced by the protein, and furthermore, displacement of a water molecule by the protein should be rewarded by the scoring function because a displaced water molecule gains rigid-body translational and rotational entropy.

In this paper, we present such a novel method for dealing with water molecules in protein−ligand docking and its implementation in the protein−ligand docking program MVD.[26] The Attached Water Model (AWM) approach fully solvates the ligand molecule with the maximum number of water molecules around hydrogen bonding functional groups, and these attached water (AW) molecules are then allowed to be included or displaced during the docking simulation. A large number of AW molecules are in this way attached to a ligand; each carbonyl group increases the number of AW molecules by two, where the carbonyl oxygen accepts hydrogen bonds from two AWs. Similarly, each hydroxyl group increases the number of AW molecules by three because it accepts hydrogen bonds from two AWs and donates a hydrogen bond to one AW. We must however emphasize that the nature of the AW molecule is different from the nature of the ligand atoms. Whereas the presence and the connectivity of the atoms in the ligand is unchanged during the docking calculation, the AW molecules are not necessarily part of the final pose. The AW molecules are displaceable during the docking run. For an AW molecule to become part of the final pose, the total energy from the interactions (both attractive and repulsive) it forms with the protein, cofactor, ligand, and other AW molecules must be favorable (negative). Notice that the hydrogen bond attaching the water molecule to the ligand is not included in this energy sum. A positive constant penalty, S_p, representing the loss of rigid-body entropy, is added for each water molecule that is switched on, hence rewarding water displacement.

In this paper, the AWM approach is benchmarked on a set of 12 protein−ligand complexes with mediating water molecules. For each of the successful docking simulations, we find the fraction of the included AW molecules that overlay the position of the crystallographic water molecules. Similarly, we investigate the ability of the AWM to find the positions of the crystallographic water molecules by calculating the fraction of the crystallographic water molecules between the ligand and the receptor molecule that is predicted by the AWM. To select these water molecules, a 3.0 Å cutoff from any ligand atom and any protein atom were jointly applied because only water molecules that fulfill these hydrogen bonding distance criteria are able to mediate interactions between ligand and receptor. Furthermore, the AWM is tested on 12 protein−ligand complexes that do not require water molecules for correct docking in MVD. This is done to test whether or not the AWM scoring function deteriorates the docking accuracy on protein−ligand complexes where water molecules are not needed for correct docking. The methodology is described in detail in the Theoretical Methodology section whereas the technical details and the scoring function are discussed in the Methods section.

THEORETICAL METHODOLOGY

Entropy Considerations. Binding of a water molecule is associated with a loss of rigid-body translational and rotational freedom. This must be modeled in the docking program, otherwise the predicted solutions tend to have the ligands surrounded by a solvation shell of water molecules, which mediates interactions to the protein.

A water molecule will bind to a protein−ligand complex only if its loss of rigid-body entropy on binding is outweighed by its gain in binding enthalpy. In other words, a water molecule that is displaced by a ligand gains rigid-body translational and rotational entropy, and this should therefore be rewarded in the scoring function by the docking program. One way to model this is by adding a constant and positive penalty for each included AW molecule to model the loss of rigid-body entropy. GOLD uses this approach in their displaceable water model.[14] However, as also stated by the authors, the entropy penalty is not a constant in reality. Because water molecules that bind very tightly to the protein lose more rigid-body entropy than weakly binding ones, the entropy penalty is water site dependent.[27−29]

In The Particle Concept[17] approach, implemented in the FlexX[16] docking program, no entropy penalty value is added for included water molecules. Instead a penalty is added per vacant interaction site at an included water molecule. Furthermore, a hydrogen bond between an included water molecule and

the ligand contributes less to the interaction energy than a hydrogen bond between the ligand and the protein. This vacant interaction penalty indirectly accounts for the loss of rigid-body entropy.

In our AWM approach, we use the simplified entropy model of introducing a fixed constant per included water molecule. The following values have been tested for the entropy penalty: −2.5, 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, and 10. In MVD the units are arbitrary, but an ideal hydrogen bond contributes to the overall energy by −2.5. Thus, the S_p value ranges from zero to four times the energy of a hydrogen bond (with opposite signs). Furthermore an S_p value of −2.5 is included because this is energetically equivalent to a setup, where the interaction between the ligand and the AW molecule contributes to the score by −2.5 corresponding to an ideal hydrogen bond in combination with an entropy penalty of zero. In our current model, an AW molecule is included if the sum of its interactions with the protein, cofactor, other AW molecules, and the ligand is negative. However, the hydrogen bond attaching the water molecule to the ligand is chosen not to be taken into account in this evaluation, as it is a constant for all water−ligand hydrogen bonds.

Selection of Protein−Ligand Benchmark Complexes. To test our novel AWM, we needed a set of complexes where one or more water molecules are needed to mediate the important interactions between the ligand and the receptor. On the basis of the list of complexes used for evaluation of previous water models,[14,17] we initially collected a number of complexes from the protein data bank (PDB),[30,31] where it by visual inspection seems most likely that crystallographic water molecules form hydrogen bond bridges between the ligand and the protein receptor. These complexes were then subjected to redocking experiments in MVD, where all crystallographic water molecules were deleted. No success in this "no-water" redocking experiment implies that water molecules are needed for successful docking either because (a) they mediate important interactions that are needed for the ligand to recognize the receptor or (b) the water molecules simply bias the search space toward the correct binding mode architecture. Irrespective of the role of the water molecules, we must also ensure that the docking program is able to predict the correct binding mode in the presence of all water molecules, and thus, we also performed redocking experiments with the inclusion of all the crystallographic water molecules. Complexes that fail the "no-water" test and have success in the "all-water" experiment are candidates for our AWM because water molecules are shown to be important either because of (a) or (b) mentioned above or a combination of the two. The resulting benchmark set consists of 12 complexes and is shown in Chart 1. The 12 complexes display a huge diversity of ligands, ranging from the two-substrate mimicking inhibitor P^1,P^5-bis-(adenosine-5′-)pentaphosphate, comprising 57 heavy atoms in PDB entry 1ake and tripeptides in PDB entries 1b5i and 1jeu, to small molecules in PDB entries 2xis and 1did, with a xylitol and an imino-glucitol derivative, comprising 10 and 11 heavy atoms, respectively.

Protein−Ligand Docking. Protein−ligand complexes were taken directly from the PDB. Charges and protonation states were assigned according to the standard MVD procedure.[26] During docking, the scoring function does not take explicit hydrogen positions into account, and all bonds that only rotate hydrogen atoms were therefore kept rigid. All other acyclic single bounds were set flexible. Thus, a C−OH single bond is kept rigid in the conventional MVD setup, but is rotatable in the AWM setup. Solvation by the AW molecules thereby increases the number of flexible torsions (Figure 1). The docking runs were performed with default settings using MVD version 4.0.

Evaluation Procedure. For each benchmark complex, 10 different solutions (poses) were obtained, resulting from 10 independent runs with the guided differential evolution algorithm. After re-ranking with more detailed calculation of hydrogen bonding and van der Waals interactions,[26] of the 10 solutions, the standard Cartesian root-mean-square deviation (rmsd) measure between similar atoms in the highest-ranked solution and the known experimental structure was calculated. A pose was considered successful if the rmsd between the top-ranked pose and the experimentally known ligand was less than 2.0 Å. This success criterion has been used in a number of docking studies.[16,32−36]

METHODS

Representation of the Ligand and the Attached Water Molecules. In our model, a ligand configuration (a pose) is represented by position, orientation, and a number of torsional angles. The torsional angles represent the internal degrees of freedom of the ligand. They can be found by decomposing the ligand into rigid fragments, which are connected to each other through only a single covalent bond. These fragments are naturally arranged in a hierarchical structure: a torsion tree. The atom coordinates for a given configuration may then be calculated by starting from the root fragment and recursively traversing the torsion tree while applying the corresponding rotation (Figure 2).

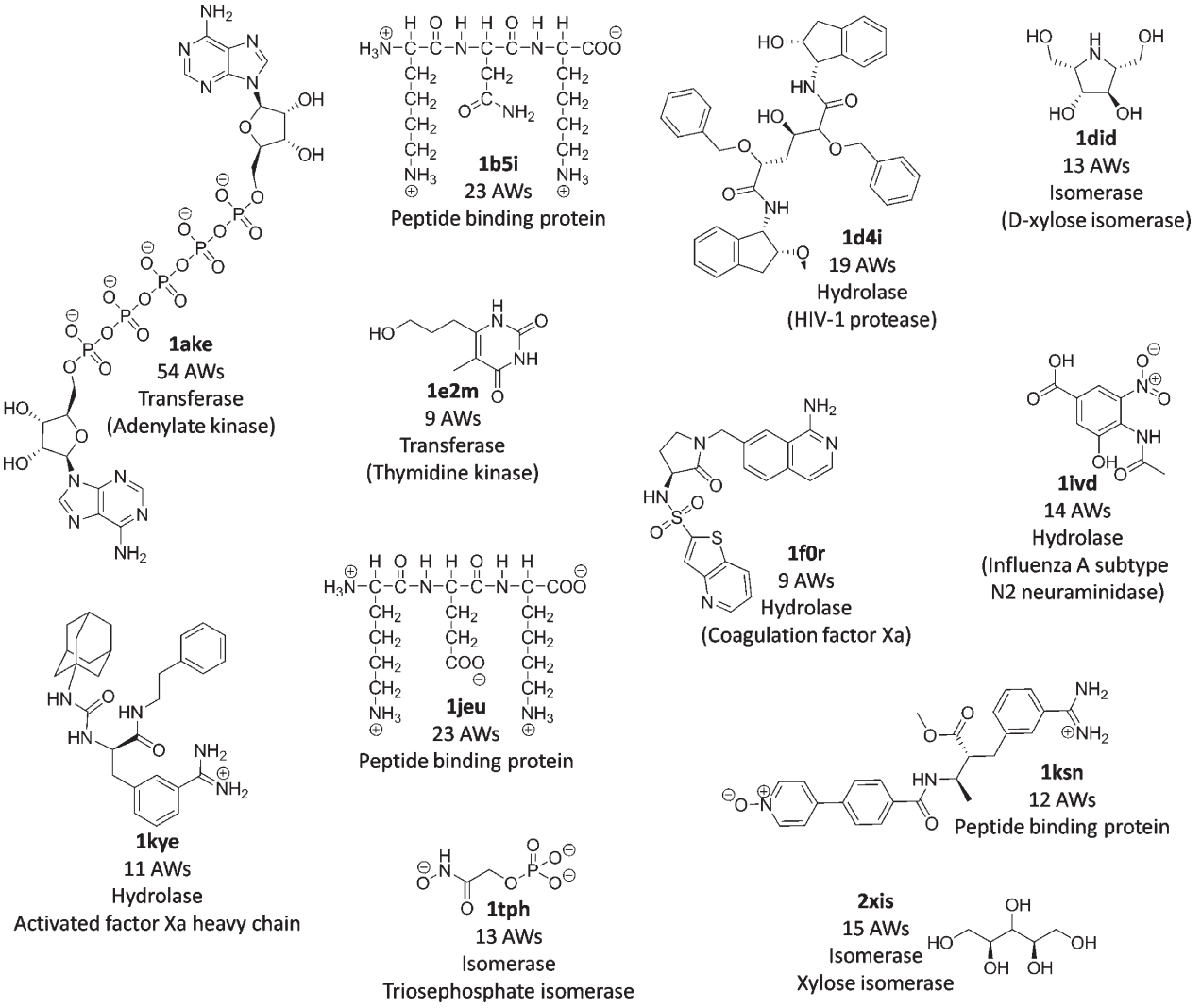
The AWM uses full ligand flexibility, whereas the protein is kept rigid during the docking simulation. This is a common simplification also used in previous studies of explicit water molecules in protein−ligand docking.[14,17]

In MVD, all bonds that only rotate hydrogen atoms are kept rigid, and thus a C−OH single bond is kept rigid in a conventional MVD setup. Solvation of the ligand results in a large number of AW molecules surrounding the ligand, and solvation by three AW molecules of a hydroxyl group turns the C−OH single bond into a rotatable torsion in the AWM setup (Figure 2). The same is the case for C−NH_2 single bonds. Notice that the attaching hydrogen bond itself (between ligand and AW molecule) is not a torsional degree of freedom; the hydrogens of the water molecules are not allowed to rotate. In our model, the directionality of the water hydrogens and lone pairs are not taken into account. Geometric constraints are imposed only for the hydrogen bond donor and acceptor atoms in the ligand and the protein (please see the original paper describing MolDock[26] for a description of how hydrogen bond directionality is handled).

The AWM extends the torsion tree model by storing potentially interacting water molecules as part of the representation (Figure 2). The water molecules become nodes in the torsion tree, connected by a new type of edge representing a hydrogen bond. The AW molecules do not necessarily participate in the interaction. If the energy of an AW molecule is positive, e.g., if the water molecule clashes with the protein, the water molecule is considered displaced and does not enter the energy calculation.

The ligand is solvated by AW molecules in the following way. For all heavy atoms in the ligand capable of acting as a hydrogen bond donor, we add an AW molecule for each of their hydrogens. The AW molecule is placed on the line defined by the heavy

Chart 1. Benchmark Set: Ligand Structure, PDB Entry, Number of AW Molecules Used To Solvate the Ligand, and Protein Classification



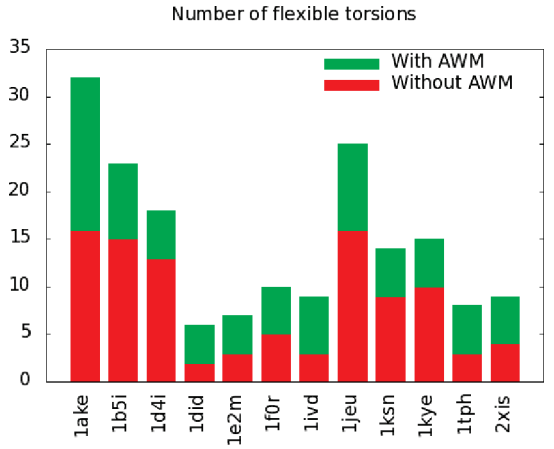


Figure 1. AWM increases the number of flexible torsions. Red columns show the number of torsions for the 12 complexes in a conventional setup in MVD. Green parts of the columns illustrate the additional torsions when the AW molecules are included.

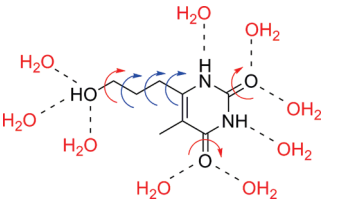


Figure 2. Example of a ligand together with its AW molecules (in red). Torsional degrees of freedom are indicated by the circular arrows. Blue arrows indicate the rotatable torsions in a conventional MVD setup, and red arrows illustrate the additional torsional degrees of freedom imposed by the solvating AW molecules.

donor atom and its attached hydrogen, with a distance between the heavy donor atom and the oxygen in the water of 2.8 Å. Hydrogen bond acceptors in the ligand are handled in a similar manner. First the lone pair positions are identified. For sp^2 hybridized atoms, these are assumed to be located at 120° angles to the covalent bond and in the same plane as the neighbors of the heavy acceptor atom. For sp^3 hybridized atoms, we assume an

dx.doi.org/10.1021/ci100510m |J. Chem. Inf. Model. 2011, 51, 909–917

angle of 109.45° between covalent bonds and lone pairs. The AW molecules are then placed on the line defined by the heavy donor atom and its lone pair, again with a distance between the heavy acceptor atom and the oxygen in the water of 2.8 Å. It is possible for an atom to act both as an acceptor and donor, e.g., hydroxyl. In this case, water molecules are attached to both the hydrogen and the lone pairs. Notice that in some cases the position of the water molecules relative to the ligand is fixed, for instance in the case of aromatic nitrogens, e.g., pyrrole type nitrogens. In other cases, e.g., hydroxyl, the direction of the hydrogen atom (and the AW molecules) may be rotated around the covalent bond; thus, the hydroxyl C−O bond is chosen rotatable in the AWM. Furthermore, in the special case of carboxyl, the direction of the lone pairs (and accordingly the AW molecules) would be restricted to the plane of the neighboring atoms, but this was found to be contradicted by experimental evidence.[37,38] Therefore, we chose to treat the carboxyl double-bond as a rotatable bond (see ligand example in Figure 2).

Evaluation of the Ligand and Attached Water Molecules. For evaluating the protein−ligand interactions, we use the MolDock Score.[26] The AW evaluation consists of the following terms

$$E_{AW} = E_{Protein} + E_{Ligand} + E_{Other\text{-}AW} + S_p$$

Except for the entropy term, the terms in the expression are all pairwise PLP potentials of the form described in the original paper describing MolDock.[26] They account for hydrogen bonds and steric forces between heavy atoms. $E_{Protein}$ is the energy between the AW molecule and the atoms in the protein (this also includes all cofactors in the complex). E_{Ligand} is the interaction energy between the atoms in the ligand and the AW molecule. Notice that the hydrogen bond attaching the water molecule to the ligand is not included in this energy. $E_{Other\text{-}AW}$ is the energy between a given AW molecule and other AW molecules. This term ensures that AW molecules do not occupy the same space (if two AW molecules are clashing, we arbitrarily choose to keep the first one to be evaluated). Finally, an entropy penalty term, S_p, is included because a bound molecule will experience a loss of rotational and translational freedom, and thus, according to the definition of the Gibbs free energy, will be less favorable. As a simplified entropy model, we model this entropy penalty using a positive fixed constant for each included water molecule. The magnitude of this constant is determined using computational experiments (see below).

RESULTS AND DISCUSSION

It must again be emphasized that for the 12 complexes in our benchmark set, it is not possible to predict the binding modes correctly (rmsd < 2 Å) in the absence of the crystallographic water molecules (Figure 3), but it is possible in the presence of all crystallographic water molecules.

Docking simulations were performed for the 12 complexes in Chart 1 using a range of values for the entropy penalty, S_p. Table 1 summarizes the results of the docking simulations for each of the complexes using the AWM. Data for the highest ranked pose for each complex are included in the table, together with data for the first successful pose if the highest ranked pose is unsuccessful (rmsd > 2 Å). The rmsd value of the pose, along with its rank, is reported in the table. A successful pose is reported in bold.

Using our AWM, it is possible to correctly predict the binding of 50% (6/12) of the complexes as the first-ranked pose, including the poses up to rank 5 increases the success rate to 67% (8/12). We furthermore calculated the fraction of the included AW molecules that correctly predicts crystallographic water positions in the X-ray structure. In addition, we found the fraction of the crystallographic water molecules between the ligand and the receptor, which is predicted by an AW molecule. Here, a crystallographic water molecule is considered if its shortest distance to both a receptor heavy atom as well as a ligand heavy atom is less than 3 Å and thus potentially mediating a contact between the receptor and the ligand. The results for these water prediction measures are shown in Table 2. Only poses where the highest ranked pose represents the crystallographic binding mode (rmsd < 2 Å) are analyzed in detail. The numbers in bold display the number of the "included" AW molecules that are within 2 Å of any crystallographic water molecule. The nonbold numbers describe the fraction of the crystallographic water molecules between both ligand and receptor (3 Å to each) that is within 2 Å of an included AW molecule.

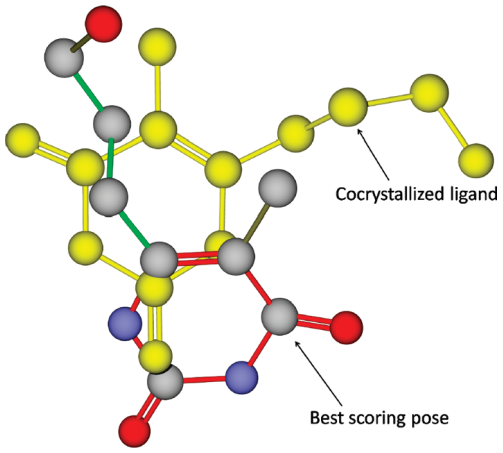


Figure 3. Unsuccessful docking using conventional docking protocol in MVD. Best pose (rmsd = 4.25 Å) for docking of 6-hydroxypropylthymine into its receptor (thymidine kinase, PDB entry 1e2m) and cocrystallized ligand in yellow. Flexible torsions are displayed as green bonds.

For PDB entry 1e2m, the correct binding mode is found as the first-ranked pose for 10 out of 12 values of S_p. The AWM is also successful in finding correct binding modes as the first-ranked pose for the PDB complexes 1jeu, 1b5i, and 1f0r for a large range of the tested entropy penalty values. For PDB entry 1jeu, a correct binding mode is found as the highest ranked pose for entropy penalty values, S_p, ranging from 3 to 9, using the AWM. Similarly, correct binding modes are predicted as the highest ranked pose for S_p ranging from 1 to 10 for PDB entry 1b5i. This is also the case for the PDB structure 1f0r, with the entropy penalty values 2−6 and 9.

Figure 4 illustrates the highest-ranked pose obtained for redocking of the ligand from PDB entry 1e2m using the AWM approach with an entropy penalty of 3. The crystallographic binding mode is reproduced with an rmsd of 0.80 Å. Of the four included AW molecules (green), two overlay with crystallographic water molecules. Furthermore, of the three crystallographic water molecules within 3 Å of both ligand and the protein receptor, two are within 2 Å of an included AW molecule. The third crystal water is in hydrogen bonding distance (3.1 Å) to the ligand carbonyl oxygen; however, it is positioned 3.96 Å from an AW molecule. Thus, in the crystal structure, the two crystal water molecules coordinating the carbonyl oxygen and the carbonyl

Table 1. Results for Docking Simulations using the Attached Water Model[a]

	rmsd (Å)											
PDB	$S_p = -2.5$	$S_p = 0$	$S_p = 1$	$S_p = 2$	$S_p = 3$	$S_p = 4$	$S_p = 5$	$S_p = 6$	$S_p = 7$	$S_p = 8$	$S_p = 9$	$S_p = 10$
1ake	(01) 23.20	(01) 4.75	(01) 22.66	**(01) 1.93**	*(01) 10.11*	(01) 11.61	(01) 3.44	(01) 6.50	(01) 9.85	(01) 3.95	(01) 2.64	(01) 4.63
1b5i	(01) 3.51	(01) 4.25	**(01) 1.32**	**(01) 1.93**	***(01) 1.64***	**(01) 1.83**	**(01) 1.16**	**(01) 1.62**	**(01) 1.62**	**(01) 1.95**	**(01) 1.24**	**(01) 1.21**
	(05) 1.54											
1d4i	(01) 9.50	(01) 2.36	(01) 12.03	(01) 11.49	*(01) 12.25*	(01) 9.31	(01) 6.93	(01) 11.06	(01) 5.11	(01) 4.50	(01) 12.64	(01) 10.07
1did	(01) 5.43	(01) 4.86	(01) 4.77	(01) 4.47	*(01) 3.50*	(01) 4.32	(01) 3.51	(01) 3.55	(01) 3.55	**(01) 1.57**	(01) 3.43	(01) 3.72
					(05) 1.06							**(02) 0.65**
1e2m	**(01) 0.85**	(01) 4.66	**(01) 0.80**	**(01) 0.80**	***(01) 0.80***	**(01) 0.80**	**(01) 0.84**	**(01) 0.85**	**(01) 0.85**	**(01) 0.82**	**(03) 0.85**	**(01) 0.77**
		(04) 1.01										
1f0r	(01) 3.46	(01) 3.20	(01) 7.13	**(01) 1.58**	***(01) 1.64***	**(01) 1.79**	**(01) 1.64**	**(01) 1.91**	(01) 4.35	(01) 2.43	**(01) 1.64**	(01) 8.85
									(03) 1.04			
1ivd	(01) 3.28	(01) 2.10	(01) 2.26	(01) 3.71	*(01) 5.00*	**(01) 1.76**	(01) 5.20	**(01) 1.80**	(01) 5.01	(01) 5.06	(01) 5.04	(01) 4.98
		(02) 1.92	**(04) 1.96**	**(02) 1.06**	***(02) 1.50***		**(02) 1.82**		**(02) 1.83**	**(02) 1.79**	**(02) 1.81**	**(02) 1.80**
1jeu	(01) 2.71	(01) 2.69	(01) 3.87	(01) 8.92	***(01) 1.53***	**(01) 1.42**	**(01) 1.79**	**(01) 1.38**	**(01) 1.71**	**(01) 1.73**	**(01) 1.78**	(01) 5.57
		(02) 1.74	**(03) 1.80**									
1ksn	(01) 8.90	(01) 3.90	(01) 6.73	(01) 10.50	*(01) 5.95*	(01) 4.30	(01) 3.04	(01) 11.41	(01) 8.61	(01) 9.69	(01) 8.88	(01) 8.87
1kye	(01) 6.02	(01) 6.57	(01) 7.43	(01) 8.11	*(01) 3.00*	(01) 3.76	(01) 7.53	(01) 3.65	(01) 7.77	(01) 8.38	(01) 3.09	(01) 4.04
										(10) 1.86		
1tph	(01) 14.82	(01) 5.35	(01) 4.95	(01) 5.83	***(01) 1.33***	(01) 4.28	(01) 3.64	**(01) 0.47**	(01) 5.14	**(01) 0.90**	**(01) 0.49**	(01) 5.88
						(04) 1.82	**(06) 1.79**					
2xis	(01) 8.96	(01) 9.59	(01) 12.19	(01) 2.71	***(01) 1.80***	(01) 2.70	(01) 2.67	(01) 2.67	(01) 2.86	(01) 2.61	(01) 2.10	**(01) 1.73**
						(03) 1.92	**(03) 1.76**			**(09) 1.45**	**(06) 1.96**	

[a] Successful docking simulations, i.e., with rmsd values below 2 Å, are shown in bold. The number in parenthesis is the rank of the reported pose. The column with the numbers in italic corresponds to the optimum value of the entropy penalty.

Table 2. Water Matching Accuracy[a]

PDB	$S_p = -2.5$	$S_p = 0$	$S_p = 1$	$S_p = 2$	$S_p = 3$	$S_p = 4$	$S_p = 5$	$S_p = 6$	$S_p = 7$	$S_p = 8$	$S_p = 9$	$S_p = 10$
1ake				**3 of 20**								
				3 of 9								
1b5i			**5 of 8**	**3 of 7**	***3 of 6***	**3 of 5**	**3 of 7**	**3 of 4**	**2 of 2**	**2 of 2**	**3 of 3**	**3 of 3**
			4 of 10	3 of 10	*3 of 10*	4 of 10	4 of 10	4 of 10	2 of 10	1 of 10	3 of 10	2 of 10
1d4i												
1did										**1 of 1**		
										0 of 2		
1e2m	**3 of 6**		**2 of 4**	**2 of 4**	***2 of 4***	**2 of 4**	**2 of 3**	**2 of 3**	**2 of 3**	**2 of 3**		**2 of 2**
	2 of 3		2 of 3	2 of 3	*2 of 3*	2 of 3	2 of 3	2 of 3	2 of 3	2 of 3		2 of 3
1f0r				**0 of 1**	***0 of 4***	**0 of 2**	**0 of 2**	**0 of 2**			**0 of 0**	
				0 of 1	*0 of 1*	0 of 1	0 of 1	0 of 1			0 of 1	
1ivd						**1 of 5**		**1 of 4**				
						2 of 2		2 of 2				
1jeu					***6 of 9***	**6 of 7**	**4 of 5**	**3 of 6**	**3 of 3**	**1 of 1**	**2 of 3**	
					6 of 12	5 of 12	2 of 12	3 of 12	4 of 12	0 of 12	1 of 12	
1ksn												
1kye												
1tph					***2 of 4***			**3 of 3**		**1 of 1**	**2 of 2**	
					3 of 3			3 of 3		1 of 3	2 of 3	
2xis					***1 of 9***							**1 of 2**
					1 of 2							0 of 2

[a] The numbers in bold correspond to the fraction of included AW molecules that are within approximately 2 Å of any crystallographic water molecule. The nonbold numbers correspond to the fraction of crystallographic water molecules that is predicted by the included AW molecules. Only successful docking simulations have been examined, i.e., rmsd below 2 Å for the highest ranked pose (Table 1). The column with the numbers in italic corresponds to the optimum value of the entropy penalty.

dx.doi.org/10.1021/ci100510m |J. Chem. Inf. Model. 2011, 51, 909–917

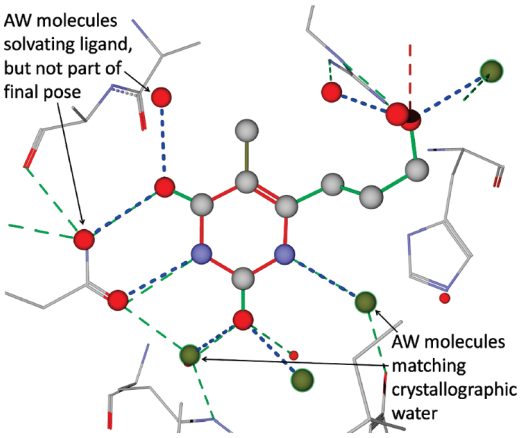


Figure 4. Best pose (rmsd = 0.80 Å) for docking of 6-hydroxypropylthymine into its receptor (thymidine kinase, PDB entry 1e2m), using the AWM docking protocol with the entropy penalty $S_p = 3$. The ligand was initially solvated by nine AW molecules, and four of these were included in the final pose (green). The two included AW molecules in the bottom of the picture match crystallographic water molecules. Flexible torsions are displayed as green bonds.

carbon are not fulfilling a planar trigonal geometry around a carbonyl oxygen, which our AWM is deemed to fulfill. These observations show that our AWM is too simple for reproducing the crystallographic binding mode for this particular case. For AWM to predict the binding mode of PDB entry 1e2m, deviation from a trigonal planar geometry around sp^2 hybridized oxygen atoms should be allowed; however, this would further increase the degrees of freedom.

The number of included AW molecules decrease with increasing value of the entropy penalty S_p. As expected, the degree of successful predictions of included AW molecules increases as shown by the numbers in bold in Table 2, corresponding to the fraction of the included AW molecules, which matches crystallographic water molecules. Furthermore, as shown by the non-bold numbers in Table 2, the fraction of the potentially mediating crystallographic water molecules, i.e., within 3 Å of both ligand and receptor, that are matched by an included AW molecule in general decreases with increasing entropy penalty, and accordingly, a decreasing number of included AW molecules are found. This is also as expected.

The complex with PDB entry 1f0r represents an interesting example because the AWM is able to correctly predict the crystallographic binding mode as the first-ranked pose for the entropy penalty range 2−6 and 9 (Table 1), but on the other hand, there is no agreement between the included AW molecules and the crystal water molecules. In the crystal structure, only one water molecule fulfills the 3 Å distance to both ligand and receptor, and this water is in perfect hydrogen bonding geometry with the nitrogen atom in the isoquinoline moiety of the ligand. In the correctly predicted first-ranked pose (rmsd = 1.64 Å, S_p = 3) using AWM, the isoquinoline nitrogen has moved 1.55 Å and is no longer in hydrogen bonding distance to this crystal water. Instead the AWM finds it more favorable to include four AW molecules elsewhere on the ligand that interact favorably with the protein. Thus, the AWM is able to find correct binding modes for a range of S_p values, even though it is not able to predict the positions of the crystallographic water molecules.

For most complexes, the AWM is able to predict the correct crystallographic binding mode, however, not always as the highest ranked pose. A correct binding mode is predicted for PDB entry 1ivd for $S_p = 2-10$, however, as the top ranked pose only for $S_p = 4$ and 6. For the complex with PDB entry 1tph, correct binding modes are found for $S_p = 3-6$ and 8−9; however, only as the first-ranked pose for S_p = 3, 6, 8, and 9. Similarly, for PDB entry 2xis, correct binding modes are predicted for $S_p = 3-5$ and 8−10, but only as a first-ranked pose for S_p = 3 and 10.

For a few complexes, the AWM was not able to predict the crystallographic binding mode. For PDB entries 1d4i and 1ksn, the correct binding mode was not predicted in any of the docking runs, irrespective of the value of the entropy penalty S_p. For PDB entry 1kye, a correct binding mode was predicted as rank 10 for $S_p = 8$, but for all other values of the entropy penalty no correct binding mode was predicted. Similarly, for PDB entry 1did, a correct binding mode was predicted only for the entropy values 3, 8, and 10 with rank 5, 1, and 2, respectively. Finally, no correct binding mode was found for PDB entry 1ake, except for the entropy value $S_p = 3$ with rank 1.

There is no correlation between the docking success using the AWM and the number of flexible torsions in the ligands. Of the four complexes where the AWM is able to predict the binding mode for a range of entropy values (PDB entries 1e2m, 1jeu, 1b5i, and 1f0r), only structures with PDB entries 1e2m and 1f0r are cocrystallized with small ligands having 6 and 9 flexible torsions after solvation with AW molecules, respectively, whereas ligands from PDB entries 1b5i and 1jeu have 23 and 25 flexible torsions after solvation. On the other hand, exposure of the binding site to the surface plays a role. It was impossible or very difficult to predict the correct binding mode of the PDB complexes 1d4i, 1ksn, 1kye, and 1did, of which two PDB entries, 1ksn and 1kye, have the binding sites positioned in an open groove on the protein surface.

Some of the complexes have several crystallographic water molecules fulfilling the 3 Å criteria to both receptor and cocrystallized ligand, for example, 10 and 12 water molecules for the PDB entries 1b5i and 1jeu, respectively. Using our AWM, we find the correct binding mode as the first-ranked pose for most of the entropy penalty range, and importantly, with a lower number of included AW molecules. These observations point in the direction that not all the X-ray crystal waters play a key role in the recognition between ligand and receptor but merely fill the empty space.

Figure 5 displays the success rate of the performed docking simulations as a function of the value of the entropy penalty, S_p. The brown curve corresponds to the results for the highest ranked poses (top 1), and the additional curves correspond to the results when lower-ranked poses are considered. A steep increase in the success rate is observed when the entropy penalty value increases until a value of 3. Further increasing the entropy penalty, S_p, results in an overall small drop in docking performance. An entropy penalty optimum of 3 is thus observed. Using this optimum entropy value, the AWM is able to correctly predict the binding for 50% (6/12) of the complexes as the first-ranked pose, including the poses up to rank 5 increases the success rate to 67% (8/12).

Because the AWM introduces additional degrees of freedom and modifies the scoring function, it must be checked if these modifications deteriorate the docking accuracy on protein−ligand complexes where explicit water molecules are not relevant to the binding mode. Thus, as a further validation of our approach, we have tested our AWM using the optimum entropy value of 3 on a test set consisting of 12 complexes where correct

dx.doi.org/10.1021/ci100510m |J. Chem. Inf. Model. 2011, 51, 909–917

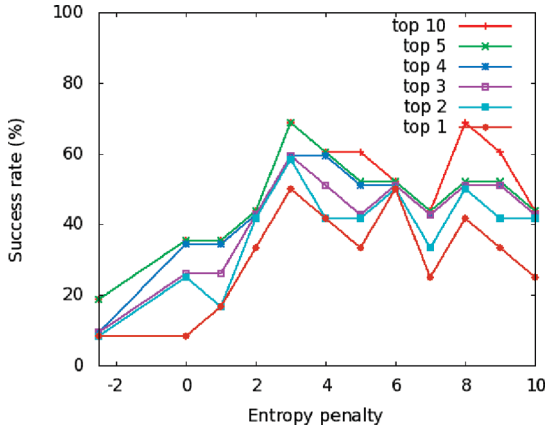


Figure 5. Success rate as a function of the entropy penalty S_p. The success rate is defined as the number of complexes for which at least one of the top n solutions is within 2.0 Å rmsd of the experimental solution.

binding mode prediction is possible using the conventional docking approach without inclusion of water molecules. The complexes constitute the PDB entries 1gpk, 1hnn, 1k3u, 1m2z, 1of1, 1s19, 1t40, 1uou, 1w2g, 1yv3, 1ywr, and 2bsm, and the ligands are solvated by 6−13 AW molecules. Using the optimum entropy value of 3, the AWM is able to correctly predict the binding mode for all of these complexes as the highest-ranked pose. The rmsd values for both conventional docking simulations in MVD as well as for docking using the AWM docking protocol are tabulated in the Supporting Information, together with ligand structure and the number of flexible torsions.

CONCLUSION

We have implemented a novel method for incorporating key water molecules in protein−ligand docking. First, the method fully solvates the ligand with attached water (AW) molecules, and these are then included during the docking calculation, if the interaction energy between the AW molecule and the surroundings is favorable (negative). The loss of rigid-body entropy when a water molecule binds to a protein is taken into account by adding a constant (positive) entropy penalty value per included AW molecule. From the training set consisting of 12 diverse complexes, an optimum is found for the entropy penalty value of $S_p = 3$. This optimum is based on a relatively small, however, diverse training set and can be used as a starting point. In reality, the entropy penalty term depends on a number of variables, as it is water site dependent.[27−29] Thus, a more complex model for the entropy penalty may be needed in some cases. The entropy penalty value of 3 is energetically equivalent to a penalty of 5.5 in a setup where the interaction between the AW molecule and the ligand atom it is attached to contributes to the score. Thus, the entropy penalty is approximately twice the strength of an ideal hydrogen bond with opposite sign in MVD, which is in agreement with the size of entropy penalty for including water molecules in GOLD's displaceable water model.[14]

The method has a clear advantage: the AW molecules are attached to the ligand and display the same flexibility as the ligand itself during the docking simulation. The AW molecules are thus not restricted to a fixed position defined by the receptor, but they have a dynamic nature. To the best of our knowledge no other docking program has implemented a method to incorporate key water molecules which positions are not defined by the structure of the protein receptor. Instead, water molecules are included as a static part of the receptor structure, and their positions are taken directly from the crystal structure or placed on favorable calculated water binding sites.

Using our approach, we have demonstrated that it is possible to correctly predict the binding mode in complexes where key water molecules are needed for the ligand to recognize the receptor. We were able to predict the binding mode (rmsd < 2 Å) for half of the complexes as the first-ranked pose, and increasing the number of poses to 5 increases the success rate to 67%. It was not possibly to predict the correct binding mode for four of the complexes, irrespective of the value of the entropy penalty. It should however be mentioned that half of these have surface-exposed binding sites, which makes correct prediction even more difficult.

Solvation of a ligand by the AW molecules results in a large amount of AW molecules surrounding the ligand. Each carbonyl group increases the number of AW molecules by two, and each hydroxyl group increases the number of AW molecules by three, etc. In the AWM, the largest possible number of water molecules is attached to the ligand, which in the present study ranges from 9 to 54 and, thus, increases the degrees of freedom. The consequences are that the search space and thus the number of potential binding modes increases, which might induce more false positives. The question that still remains is whether our AWM is also able to correctly predict the binding mode in complexes where no prior knowledge of water molecules exists. Thus, as a further validation of our AWM, we have included 12 protein−ligand complexes where correct prediction of the binding mode is possible in the absence of water molecules, using conventional docking in MVD. Using our AWM with the optimum entropy penalty value of 3, we were able to predict the correct binding mode for all of these complexes. This suggests that our AWM model can be used without any knowledge of whether water mediated interactions will be present in a protein−ligand complex or not. Thus, no loss in accuracy was observed when using the AWM scoring function on protein−ligand complexes where water molecules are not needed for correct docking. We therefore see applications of this method in computational screening of large libraries of chemicals to identify compounds that complement a known target binding site (virtual screening experiments).[39,40] In the near future, we will investigate if the AWM can improve enrichment in virtual screening experiments.

The AWM docking protocol using the optimum entropy penalty value of 3 is thus superior to conventional docking in MVD within the tested complexes because the AWM scoring function did not deteriorate the docking accuracy on protein−ligand complexes where water molecules are not needed for correct docking, and by using the AWM it was furthermore possible to correctly predict the binding mode for 50% of the complexes where it was not possible with conventional docking in MVD.

ASSOCIATED CONTENT

Supporting Information. Results of docking simulations of the 12 complexes where correct binding mode prediction is possible using the conventional docking approach in MVD and without inclusion of crystallographic water molecules. The rmsd values for both conventional docking in MVD and AWM docking are tabulated, together with the number of flexible torsions. This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author

*E-mail: lie@birc.au.dk, fax +45 8942 3109, phone +45 8942 3109 (M.A.L) or birgit@chem.au.dk, fax+45 8619 6199, phone +45 8942 3953 (B.S.).

ACKNOWLEDGMENT

Financial support from the Danish Natural Science Research Council and The Danish Council for Strategic Research's Programme Commision on Strategic Growth Technologies is acknowledged. Grants from the Carlsberg, Lundbeck, and Novo Nordisk Foundations are furthermore thanked.

REFERENCES

(1) Ball, P. Water as an active constituent in cell biology. *Chem. Rev.* **2008**, *108*, 74–108.

(2) Okada, T.; Fujiyoshi, Y.; Silow, M.; Navarro, J.; Landau, E. M.; Shichida, Y. Functional role of internal water molecules in rhodopsin revealed by X-ray crystallography. *Proc. Natl. Acad. Sci. U. S. A.* **2002**, *99*, 5982–5987.

(3) 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.

(4) 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.

(5) Palomer, A.; Perez, J. J.; Navea, S.; Llorens, O.; Pascual, J.; Garcia, L.; Mauleon, D. Modeling cyclooxygenase inhibition. Implication of active site hydration on the selectivity of ketoprofen analogues. *J. Med. Chem.* **2000**, *43*, 2280–2284.

(6) Levy, Y.; Onuchic, J. N. Water mediation in protein folding and molecular recognition. *Annu. Rev. Biophys. Biomol. Struct.* **2006**, *35*, 389–415.

(7) Vogt, J.; Perozzo, R.; Pautsch, A.; Prota, A.; Schelling, P.; Pilger, B.; Folkers, G.; Scapozza, L.; Schulz, G. E. Nucleoside binding site of herpes simplex type 1 thymidine kinase analyzed by X-Ray crystallography. *Proteins* **2000**, *41*, 545–553.

(8) Ni, H.; Sotriffer, C. A.; McCammon, J. A. Ordered water and ligand mobility in the HIV-1 integrase−5CITEP complex: A molecular dynamics study. *J. Med. Chem.* **2001**, *44*, 3043–3047.

(9) Chung, E.; Henriques, D.; Renzoni, D.; Zvelebil, M.; Bradshaw, J. M.; Waksman, G.; Robinson, C. V.; Ladbury, J. E. Mass spectrometric and thermodynamic studies reveal the role of water molecules in complexes formed between SH2 domains and tyrosyl phosphopeptides. *Structure* **1998**, *6*, 1141–1151.

(10) Hatshorn, M. J.; Verdonk, M. L.; Chessari, G.; Brewerton, S. C.; Mooij, W. T. M.; Mortenson, P. N.; Murray, C. W. Diverse, high-quality test set for the validation of protein−ligand docking performance. *J. Med. Chem.* **2007**, *50*, 726–741.

(11) Roberts, B. C.; Mancera, R. L. Ligand−protein docking with water molecules. *J. Chem. Inf. Model.* **2008**, *48*, 397–408.

(12) Thilagavathi, R.; Mancera, R. L. Ligand−Protein cross-docking with water molecules. *J. Chem. Inf. Model.* **2010**, *50*, 415–421.

(13) Marrone, T. J.; Briggs, J. M.; McCammon, J. A. Structure-based drug design: Computational advances. *Annu. Rev. Pharmacol. Toxicol.* **1997**, *37*, 71–90.

(14) Verdonk, M. L.; Chessari, G.; Cole, J. C.; Hartshorn, M. J.; Murray, C. W.; Nissink, J. W. M.; Taylor, R. D.; Taylor, R. Modeling water molecules in protein−ligand docking using GOLD. *J. Med. Chem.* **2005**, *48*, 6504–6515.

(15) Thomsen, R.; Christensen, M. Molegro Virtual Docker 4.0 User Manual; Molegro ApS: Aarhus, Denmark, 2009; 125−133.

(16) Rarey, M.; Kramer, B.; Lengauer, T.; Klebe, G. A fast flexible docking method using an incremental construction algorithm. *J. Mol. Biol.* **1996**, *261*, 470–489.

(17) Rarey, M.; Kramer, B.; Lengauer, T. The particle concept: Placing discrete water molecules during protein−ligand docking predictions. *Proteins* **1999**, *34*, 17–28.

(18) Abel, R.; Young, T.; Farid, R.; Berne, J. B.; 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.

(19) Beuming, T.; Farid, R.; Sherman, W. High-energy water sites determine peptide binding affinity and specificity of PDZ domains. *Protein Sci.* **2009**, *18*, 1609–1619.

(20) Young, T.; Abel, R.; Kim, B.; Berne, B. J.; Friesner, R. A. Motifs for molecular recognition exploiting hydrophobic enclosure in protein−ligand binding. *Proc. Natl. Acad. Sci. U. S. A.* **2007**, *104*, 808–813.

(21) Pitt, W. R.; Goodfellow, J. M. Modelling of solvent positions around polar groups in proteins. *Protein Eng.* **1991**, *4*, 531–537.

(22) Goodford, P. J. A computational procedure for determining energetically favorable binding-sites on biologically important macromolecules. *J. Med. Chem.* **1985**, *28*, 849–857.

(23) Miranker, A.; Karplus, M. Functionality maps of binding sites: A multiple copy simultaneous search method. *Proteins* **1991**, *11*, 29–34.

(24) Verdonk, M. L.; Cole, J. C.; Taylor, R. SuperStar: A knowledge-based approach for identifying interaction sites in proteins. *J. Mol. Biol.* **1999**, *289*, 1093–1108.

(25) Kortvelyesi, T.; Dennis, S.; Silberstein, M.; Brown, L., III; Vajda, S. Algorithms for computational solvent mapping of proteins. *Proteins* **2003**, *51*, 340–351.

(26) Thomsen, R.; Christensen, M. H. MolDock: A new technique for high-accuracy molecular docking. *J. Med. Chem.* **2006**, *49*, 3315–3321.

(27) Gohlke, H.; Klebe, G. Approaches to the description and prediction of the binding affinity of small-molecule ligands to macromolecular receptors. *Angew. Chem., Int. Ed.* **2002**, *41*, 2644–2676.

(28) Dunitz, J. D. Win some, lose some: Enthalpy−entropy compensation in weak intermolecular interactions. *Chem. Biol.* **1995**, *2*, 709–712.

(29) Dunitz, J. D. The entropic cost of bound water in crystals and biomolecules. *Science* **1994**, *264*, 670.

(30) 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.

(31) Berman, H. M.; Henrick, K.; Nakamura, H. Announcing the worldwide Protein Data Bank. *Nat. Struct. Biol.* **2003**, *10*, 980.

(32) Friesner, R. A.; Banks, J. L.; Murphy, R. B.; Halgren, T. A. Glide: A new approach for rapid accurate docking and scoring. 1. Method and assessment of docking accuracy. *J. Med. Chem.* **2004**, *47*, 1739–1749.

(33) Jain, A. N. Surflex: Fully automatic flexible molecular docking using a molecular similarity-based search engine. *J. Med. Chem.* **2003**, *46*, 499–511.

(34) Jones, G.; Willett, P.; Glen, R. C.; Leach, A. R.; Taylor, R. Development and validation of a genetic algorithm for flexible docking. *J. Mol. Biol.* **1997**, *267*, 727–748.

(35) Nissink, J. W. M.; Murray, C.; Hartshorn, M.; Verdonk, M. L.; Cole, J. C.; Taylor, R. A New test set for validating predictions of protein−ligand interaction. *Proteins* **2002**, *49*, 457–471.

(36) Kramer, B.; Rarey, M.; Lengauer, T. Evaluation of the FlexX incremental construction algorithm for protein−ligand docking. *Proteins* **1999**, *37*, 228–241.

(37) Apaya, R. P.; Bondí, M.; Price, S. L. The orientation of N−H···O=C and N−H···N hydrogen bonds in biological systems: How good is a point charge as a model for a hydrogen bonding atom? *J. Comp.-Aided Mol. Des.* **1997**, *11*, 479–490.

(38) Taylor, R.; Kennard, O.; Versichel, W. Geometry of the imino-carbonyl (N−H···O=C) hydrogen bond. 1. Lone-pair directionality. *J. Am. Chem. Soc.* **1983**, *105*, 5761–5766.

(39) Shoichet, B. K. Virtual screening of chemical libraries. *Nature* **2004**, *432*, 862–865.

(40) Vasudevan, S. R.; Churchill, G. C. Mining free compound databases to identify candidates selected by virtual screening. *Expert. Opin. Drug Discovery* **2009**, *4*, 901–906.