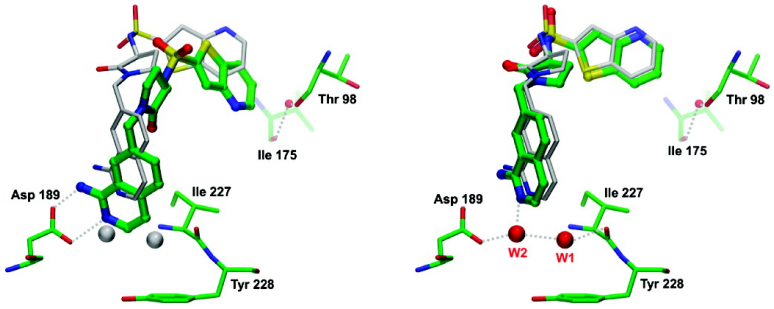
Journal of Medicinal Chemistry

Article

Modeling Water Molecules in Protein–Ligand Docking Using GOLD

Marcel L. Verdonk, Gianni Chessari, Jason C. Cole, Michael J. Hartshorn, Christopher W. Murray, J. Willem M. Nissink, Richard D. Taylor, and Robin Taylor

J. Med. Chem., **2005**, 48 (20), 6504-6515• DOI: 10.1021/jm050543p • Publication Date (Web): 14 September 2005



More About This Article

Additional resources and features associated with this article are available within the HTML version:

- Supporting Information
- Links to the 20 articles that cite this article, as of the time of this article download
- Access to high resolution figures
- Links to articles and content related to this article
- Copyright permission to reproduce figures and/or text from this article

View the Full Text HTML

Journal of Medicinal Chemistry is published by the American Chemical Society. 1155 Sixteenth Street N.W., Washington, DC 20036

Modeling Water Molecules in Protein−Ligand Docking Using GOLD

Marcel L. Verdonk,*,† Gianni Chessari,† Jason C. Cole,‡ Michael J. Hartshorn,† Christopher W. Murray,† J. Willem M. Nissink,‡ Richard D. Taylor,† and Robin Taylor‡

Astex Therapeutics Ltd, 436 Cambridge Science Park, Milton Road, Cambridge CB4 0QA, U.K., and Cambridge Crystallographic Data Centre, 12 Union Road, Cambridge CB2 1EZ, U.K.

Received June 9, 2005

We implemented a novel approach to score water mediation and displacement in the protein−ligand docking program GOLD. The method allows water molecules to switch on and off and to rotate around their three principal axes. A constant penalty, σ_p, representing the loss of rigid-body entropy, is added for water molecules that are switched on, hence rewarding water displacement. We tested the methodology in an extensive validation study. First, σ_p is optimized against a training set of 58 protein−ligand complexes. For this training set, our algorithm correctly predicts water mediation/displacement in ~92% of the cases. We observed small improvements in the quality of the predicted binding modes for water-mediated complexes. In the second part of this work, an entirely independent set of 225 complexes is used. For this test set, our algorithm correctly predicts water mediation/displacement in ~93% of the cases. Improvements in binding mode quality were observed for individual water-mediated complexes.

Introduction

There are three key remaining challenges facing the field of protein−ligand docking: accurate scoring and ranking of different compounds, dealing with protein flexibility and ligand-induced fit, and predicting the role of key water molecules in the protein−ligand interface. Here, we attempt to address the latter of these challenges. Water molecules can be involved in protein−ligand recognition either by forming mediating hydrogen bonds between the protein and the ligand or by being displaced by the ligand; both of these mechanisms have been shown to be of importance to drug discovery.[1] For example, the first-generation HIV-1 protease inhibitors were peptidic in nature and all formed hydrogen bonds to a conserved water molecule between the two central "flaps".[2] Subsequently, it was discovered that it is possible to displace this structural water molecule, which led to new inhibitor series.[3] Similarly, the benzamidine moieties in early factor Xa inhibitors interacted with a conserved water molecule situated above a tyrosine ring in the S1 pocket.[4] More recently, inhibitors binding with neutral moieties in the S1 pocket were shown to displace this water molecule.[5,6]

There could be several potential advantages to including water molecules in a protein−ligand docking program. First, if the compound interacts with the water molecule, including it could improve the predicted binding mode. Several studies have been reported in the literature where parallel dockings were done in the absence of water molecules and in the presence of some key water molecules. Some authors have reported significant improvements in docking performance when water molecules were included,[7,8] whereas others found that including water molecules had little effect on the quality of the dockings.[9,10] A second potential advantage of addressing water binding in a docking application is that it could distinguish between compounds that can displace a water molecule and compounds that cannot. Finally, correctly scoring water mediation and water displacement in scoring/energy functions could help in ranking compounds and, therefore, increase hit rates obtained from virtual screening.

Various applications have been reported in the literature for predicting potential water binding sites on proteins. For example, AQUARIUS[11] is a knowledge-based approach specifically aimed at identifying water sites in proteins; other applications including GRID,[12] MCSS,[13] SuperStar,[14] and CS-Map[15] can also be used for this purpose. However, such applications do not directly indicate which predicted water molecules are likely to be displaced by a ligand and which are likely to remain bound to the protein. Solving this issue is clearly of importance to structure-based design, as it would indicate whether compounds could be designed to displace the water or to interact favorably with a water molecule.

If a sufficient number of X-ray structures of protein−ligand complexes are available, displaceable and conserved waters can often be identified and a suitable design strategy can be adopted.[16] Consolv was developed by Raymer et al.[17] to automate the process of assigning conserved waters using the distribution of a number of structural parameters describing the water molecules in a training set of 13 diverse proteins. More recently, García-Sosa et al.[18] used a similar set of parameters in WaterScore to distinguish between conserved and displaceable water molecules.

When water molecules are known or assumed to play a role in protein−ligand recognition, the most common strategy is to perform separate docking runs in parallel, i.e., one in the absence of water molecules and a second

* To whom correspondence should be addressed. E-mail: m.verdonk@astex-therapeutics.com. Tel: +44 1223 226206. Fax: +44 1223 226201.
† Astex Therapeutics Ltd.
‡ Cambridge Crystallographic Data Centre.

10.1021/jm050543p CCC: $30.25
Published on Web 09/14/2005

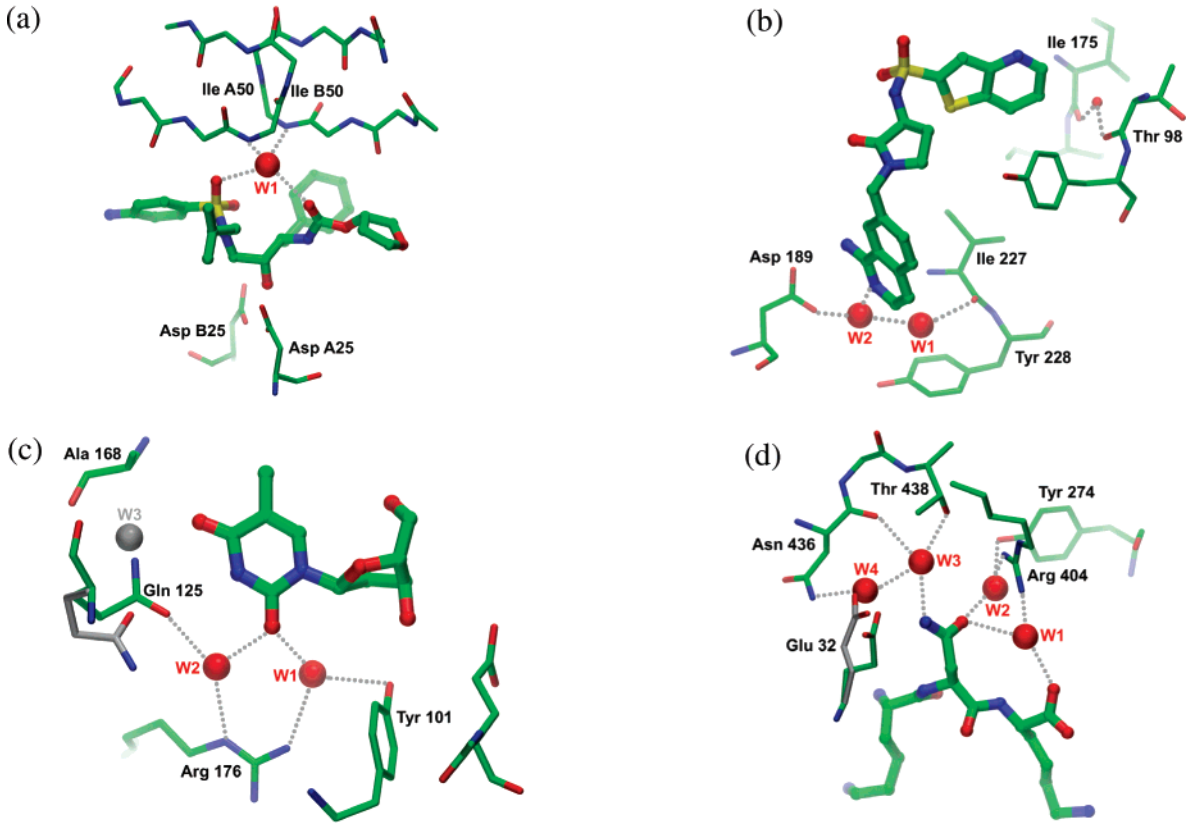


Figure 1. Four protein−ligand complex test cases and the water molecules used during docking. (a) HIV-1 (PDB entry 1hpv). (b) FXa (PDB entry 1f0r). (c) TK (PDB entry 1kim; PDB entry 2ki5 was used for the alternative conformation of Gln125 and for water W3 (both shown in gray)). (d) OppA (PDB entry 1b5i; PDB entry 1b3f was used for the alternative conformation of Glu32 (shown in gray)). All color figures were produced using AstexViewer 2.[43]

in the presence of one or more water molecules. However, these parallel runs need to be analyzed and some assessment of the cost of displacing a water molecule is required. Hence, it would be preferable if the docking program could assess both the bound and unbound states of water molecules. To address this, FlexX[19] can precalculate energetically favorable water sites;[20] "spherical" water molecules ("particles") can then be switched on at each of these positions during the docking protocol. In SLIDE,[21] Consolv is used to predict water molecules that are likely to be displaced, and these water molecules are removed from the binding site. The remaining water molecules can then be displaced during the docking at the cost of a penalty. AutoDock[22] can use multiple energy grids representing different states of the protein. Österberg et al.[23] created energy maps for different structures of HIV-1 protease, including one structure that contained the key water molecule interacting between the flaps, hence implicitly giving AutoDock the option to "choose" between the water-bound and the water-unbound state.

What we believe is missing in the above approaches is the concept that a water molecule that is displaced by a ligand gains rigid-body translational and rotational entropy and that this should therefore be *rewarded* in the scoring function used by the docking program. We also feel that predicting the positions of water molecules as well as their occupancies (i.e., whether they are bound or displaced) makes the problem unnecessarily challenging. In most structure-based drug discovery applications, the modeler will have access to knowledge about potential water sites and will be able to make an informed judgment on which water molecules to consider.

Here we present a novel method for dealing with key water molecules in protein−ligand docking and its implementation in the protein−ligand docking program GOLD.[24,25] The method represents each water molecule by an all-atom model and allows it to switch on and off (i.e., to be bound or displaced) and to rotate around its three principal axes; we have implemented the water placement model for two scoring functions: Goldscore and Chemscore. For a water molecule that is switched on, the interactions (both attractive and repulsive) it forms with the protein, ligand, and other water molecules (if present) are implemented using standard functional forms and parameters for the scoring function used. A constant penalty, σ_p, representing the loss of rigid-body entropy, is added for water molecules that are switched on, hence rewarding water displacement.

In the first part of this work we derive the optimum values for σ_p for the two scoring functions, using a training set of 58 protein−ligand complexes for four targets where water molecules play key roles in the recognition (see Figure 1); these targets are HIV-1 protease (HIV-1, 16 complexes), factor Xa (FXa, 14 complexes), thymidine kinase (TK, 15 complexes), and the oligopeptide-binding protein OppA (13 complexes); all complexes were taken from the Protein Data Bank[26] (PDB). Using the optimized penalties, we test the performance of GOLD at predicting which water molecules are displaced and which are not, and we inves-

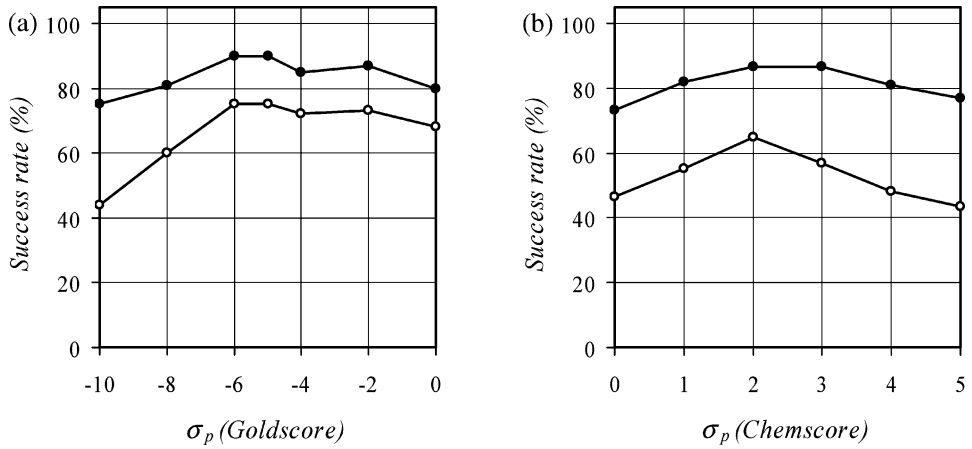


Figure 2. Success rates using (a) the Goldscore function and (b) the Chemscore function as a function of σ_p. The filled circles show the performance in terms of the prediction of water occupancies; this performance is defined as the percentage of water molecules for which the predicted occupancy is correct, averaged over the four targets and weighting each prediction such that occupied sites and unoccupied sites contribute equally to the average. The open circles show the performance in terms of the prediction of the overall binding modes; this performance is defined as the percentage of complexes for which the top-ranked GOLD solution is within 2.0 Å RMSD of the experimental solution *and* the water occupancies are predicted correctly, averaged over the four targets.

tigate the effect of including water molecules on the quality of the binding modes. This first part of the study was carried out at Astex Therapeutics Ltd.

The second part of this study was carried out at the Cambridge Crystallographic Data Centre (CCDC), where the σ_p values derived from the first phase were used in a validation on an entirely independent set of test structures. This set consists of three separate test sets of protein−ligand complexes, each a subset of the CCDC/Astex validation set:[25,27] a *primary mediating* waters set, a *displaced* waters set, and a *decoy* waters set. The primary mediating waters set contains 28 complexes where one or more water molecules form key mediating hydrogen bonds between the protein and the ligand. The displaced waters set consists of 55 complexes where the ligand has displaced at least one water molecule, observed in another X-ray structure of the same target. The decoy waters set contains 142 complexes for which one or more water molecules have been added in favorable positions identified by SuperStar.[14] For each of these test sets, we check the percentage of the water molecules that are switched on by our algorithm and analyze whether including the water molecules has an effect on the quality of the binding modes produced.

Training Set Results

Optimization of σ_p. For all four targets in the training set (HIV-1, FXa, TK, and OppA), dockings were performed using a range of σ_p values. We then calculated the fraction of correctly predicted water occupancies and the fraction of correctly predicted binding modes (i.e., for which the root-mean-square difference (RMSD) between the top-ranked GOLD solution and the experimental binding mode is below 2.0 Å). The results are shown in Figure 2. It is clear that the value of σ_p has a marked effect on the performance and that, both for the Goldscore and the Chemscore function, there is a clear optimum.

From Figure 2b, we selected $\sigma_p = +2.0$ as the optimal value for Chemscore. The Chemscore function is an energy-based scoring function, and the units of this scoring function are kJ/mol. Hence, the optimum value for σ_p translates to a free energy penalty associated with the loss of rigid-body entropy of 2.0 kJ/mol. This is roughly in line with the free energy cost of transferring a water molecule from solution to the protein derived by Dunitz,[28] 0−8.4 kJ/mol; the higher end of this range only applies for very tightly bound water molecules such as those coordinating to metal atoms.

From Figure 2a, we selected $\sigma_p = -5.0$ as the optimum value for Goldscore. The Goldscore function is not an energy-based scoring function, and favorable contributions to the score have positive signs. Hence, the fact that the Goldscore value for σ_p has the opposite sign to the Chemscore value is consistent. When Goldscore values are plotted against Chemscore values for a set of protein−ligand complexes, we have observed that Goldscore values are approximately twice as high as the corresponding Chemcore values (and have opposite signs). Hence, the value of $\sigma_p = -5.0$ for Goldscore is roughly consistent with the derived Chemscore value of $\sigma_p = +2.0$ and is also in line with the free energy cost of transferring a water molecule from the bulk solvent to the protein (see above).

Water Occupancies. Tables 1−4 show the performance in terms of the predicted water occupancies for each of the four targets. Individual cases will be discussed below, but it is clear from these tables that the predicted water occupancies are correct in the majority of the cases. A summary of the performance of our algorithm in terms of predicting water occupancies is given in Table 5. Using the Goldscore function, 94% of the water occupancies are predicted correctly, and the Chemscore function predicts 90% of the occupancies correctly. To quantify the significance of these results, we estimated the likelihood of obtaining such success rates by chance. As a null hypothesis, we assumed that, for each water site, the probability of it being occupied is equal to $N_{ON}/(N_{ON} + N_{OFF})$, where N_{ON} is the number of crystal structures in our test set for which this site is occupied and N_{OFF} is the number of structures for which the site is unoccupied. On the basis of this null hypothesis, we can estimate the probability

Table 1. HIV-1 Results for Water Occupancy and Binding Mode Predictions[a]

	Goldscore			Chemscore		
PDB code	W1	RMSD toggle	RMSD off	W1	RMSD toggle	RMSD off
1ajv	OFF	1.21	1.17	OFF	0.71	0.66
1ajx	OFF	0.63	0.46	OFF	0.84	0.69
1d4i	ON	0.71	0.74	ON	0.91	**4.41**
1ebz	ON	**3.59**	**10.51**	ON	0.77	**9.76**
1hpv	ON	0.96	**9.68**	ON	1.30	1.37
1hsg	ON	1.01	1.83	ON	1.09	**3.69**
1hvr	OFF	0.41	0.53	OFF	0.75	0.71
1hwr	OFF	0.53	0.52	OFF	0.62	0.58
1hxw	ON	**3.12**	**4.54**	ON	**3.26**	**3.85**
1npv	ON	0.67	0.57	ON	0.53	1.01
1ohr	ON	0.51	1.19	ON	1.41	**4.18**
1pro	OFF	0.33	0.50	OFF	0.80	0.54
1qbs	OFF	0.56	0.23	OFF	0.46	0.52
1sbg	ON	0.84	1.11	ON	0.78	0.53
2upj	OFF	**4.00**	**4.34**	OFF	1.80	1.67
7upj	OFF	0.93	1.06	OFF	1.10	1.03

[a] Erroneously predicted water occupancies and RMSD values > 2.0 Å are shown in bold.

Table 2. FXa Results for Water Occupancy and Binding Mode Predictions[a]

	Goldscore				Chemscore			
PDB code	W1	W2	RMSD toggle	RMSD off	W1	W2	RMSD toggle	RMSD off
1ezq	ON	OFF	0.68	0.50	ON	OFF	0.70	0.65
1f0r	ON	ON	0.59	**2.95**	ON	ON	0.79	1.27
1f0s	ON	ON	1.47	**2.26**	ON	ON	0.53	**2.53**
1fjs	ON	OFF	**2.56**	**2.62**	ON	OFF	1.78	**2.61**
1g2l	ON	OFF	0.67	1.91	ON	OFF	1.35	1.53
1ksn	ON	OFF	0.80	0.73	ON	OFF	0.46	0.46
1kye	ON	OFF	1.57	1.50	**OFF**	OFF	**3.34**	**3.53**
1mq5	OFF	OFF	0.78	0.72	OFF	OFF	1.50	0.56
1mq6	OFF	OFF	1.02	0.88	**ON**	**ON**	**7.35**	**7.60**
1nfu	**ON**	**ON**	**8.25**	**8.58**	**ON**	OFF	**8.26**	**8.55**
1nfw	OFF	OFF	0.81	0.83	OFF	OFF	**8.49**	**8.44**
1nfx	OFF	OFF	0.97	1.03	**ON**	**ON**	**8.90**	1.20
1nfy	OFF	OFF	0.50	0.66	**ON**	OFF	**8.57**	**8.46**
1xka	ON	OFF	1.26	**2.51**	**OFF**	OFF	1.25	1.05

[a] Erroneously predicted water occupancies and RMSD values > 2.0 Å are shown in bold.

p of obtaining N_{CORR} or more correctly predicted occupancies for any set of water sites, using a simple simulation. Table 5 shows the p values for the water sites of each target grouped together and also for all water sites combined. It is clear that in general the percentage of water molecules with correctly predicted occupancies is significantly higher than would be expected by chance. With the exception of the Chemscore predictions for the water occupancies in FXa, the probabilities of obtaining the achieved success rates (or better) by chance are less than 0.2%. The probability of obtaining the overall success rates of 94% and 90% (or better) by chance is less than 0.0001%.

Binding Modes. Tables 1−4 show the performance in terms of the quality of the predicted binding modes for each of the four targets; individual results will be discussed below. To investigate the effect of the inclusion of water molecules during the docking on the accuracy of the binding modes produced, we divided the HIV-1, FXa, TK, and OppA complexes in our training set into three categories: (i) *primary mediated complexes* are the 32 complexes where there is at least one hydrogen-bond donor or acceptor in the ligand that hydrogen bonds with a water molecule, but *not* directly with the protein; (ii) *secondary mediated complexes* are the 11 complexes where all ligand donors and acceptors that hydrogen bond with a water molecule are also involved in at least one direct hydrogen bond with the protein; and (iii) *nonmediated complexes* are the 15 complexes where the ligand displaces all water molecules in the binding site. All three categories of complexes contain representatives from each of the four targets in the training set. Our docking algorithm should have the best chance of improving the quality of the binding modes for the primary mediated complexes.

The success rates for predicting the binding modes are summarized in Table 6, both for runs in the absence of water molecules and for runs where our protocol was used to predict the occupancies and orientations of the water molecules. For both Goldscore and Chemscore, we see a clear improvement in the success rates for the primary mediated complexes when water molecules are included. For Goldscore, six complexes that are misdocked in the absence of water molecules are docked correctly when the water molecules are taken into account; for Chemscore, this number of "new successes" is even higher (nine complexes). We also see an improvement in the success rates for the secondary mediated complexes, but the number of complexes in this set is too small for this result to be statistically significant. For the nonmediated complexes, we observe no noticeable effect on the quality of the binding modes when water molecules are included.

There are three cases where including water molecules actually worsened the predicted binding modes: FXa, 1nfx, Chemscore; TK, 1e2n, Chemscore; and TK, 1ki2, Goldscore. It is interesting to note that, in each of these cases, one or more water molecules have been switched on by the algorithm, creating a new, top-ranking, incorrect binding mode.

HIV-1 Protease. Table 1 lists the results for the water occupancy and binding mode predictions for HIV-1. It is striking to notice that all water occupancies are predicted correctly by both scoring functions. Particularly for the Chemscore function, including the key water molecule between the flaps improves docking: four complexes (1di4, 1ebz, 1hsg, and 1ohr) that are mispredicted when the water molecule is left out are predicted correctly when the water molecule is included and allowed to spin around and toggle on/off. For the Goldscore function, only one mispredicted complex (1hpv) is predicted correctly upon inclusion of the water molecule. In all of these five complexes, the water molecule forms mediating hydrogen bonds between the protein and the ligand. In the 1hsg, Chemscore case, for example, the docking with waters turned off gave an RMSD of 3.69 Å. However, the docking with waters toggling gave an RMSD of 1.09 Å and the water was turned on. The difference in RMSD is attributed to the carbonyl amide on the ligand, which forms a hydrogen bond with the mediating water in the crystal structure. When the water is turned off, the carbonyl cannot form the mediating hydrogen bond and the amide adopts an alternative conformation.

Overall, when the water molecule is included, the performance is impressive, particularly for Chemscore, where the binding modes of 15 out of the 16 HIV-1 complexes are predicted within 2 Å of the X-ray binding

Table 3. TK Results for Water Occupancy and Binding Mode Predictions[a]

	Goldscore					Chemscore				
PDB code	W1	W2	W3	RMSD toggle	RMSD off	W1	W2	W3	RMSD toggle	RMSD off
1e2k	ON	ON		0.40	0.43	ON	ON		0.72	0.80
1e2m	ON	ON		0.81	0.81	ON	ON		0.86	**4.18**
1e2n	ON	ON		0.64	1.36	ON	ON		**2.24**	0.68
1e2p	ON	ON		0.92	1.02	ON	ON		1.45	1.50
1ki2	**ON**	**ON**	**ON**	**3.96**	1.89	**ON**	OFF	**ON**	1.91	**2.01**
1ki3	OFF	OFF	**ON**	0.87	0.72	OFF	OFF	**ON**	0.95	0.99
1ki4	ON	ON		0.72	1.25	ON	ON		0.44	0.52
1ki6	ON	ON		0.88	0.93	ON	ON		0.55	0.56
1ki7	ON	ON		0.64	0.57	ON	ON		0.52	0.49
1ki8	ON	ON		0.70	0.69	ON	ON		0.91	0.48
1kim	ON	ON		0.87	**3.21**	ON	ON		0.71	0.69
1qhi	OFF	OFF	ON	0.57	0.58	OFF	OFF	ON	0.47	0.51
1vtk	ON	ON		0.78	0.74	ON	ON		0.47	0.53
2ki5	OFF	OFF	ON	1.99	**2.03**	OFF	OFF	ON	1.77	1.77
3vtk	OFF	ON		1.02	1.04	OFF	ON		0.83	0.73

[a] Erroneously predicted water occupancies and RMSD values > 2.0 Å are shown in bold.

Table 4. OppA Results for Water Occupancy and Binding Mode Predictions[a]

	Goldscore						Chemscore					
PDB code	W1	W2	W3	W4	RMSD toggle	RMSD off	W1	W2	W3	W4	RMSD toggle	RMSD off
1b0h	ON	OFF	OFF	ON	0.83	0.93	ON	OFF	OFF	ON	0.84	**2.12**
1b1h	ON	ON	ON	ON	0.93	0.74	ON	ON	ON	ON	1.06	1.14
1b3f	ON	ON	**ON**		0.92	0.85	ON	ON	**ON**		1.24	1.13
1b3h	ON	ON	OFF	ON	1.14	1.03	ON	ON	OFF	ON	1.38	1.29
1b4h	ON	ON	ON	ON	0.94	0.98	ON	ON	ON	ON	1.18	1.18
1b4z	ON	ON	ON	ON	0.88	1.16	ON	ON	ON	ON	1.07	0.98
1b58	ON	ON	OFF	ON	0.46	0.92	ON	ON	OFF	ON	1.20	1.02
1b5i	ON	ON	ON	ON	1.13	1.27	ON	ON	ON	ON	0.70	1.25
1b5j	ON	ON	ON	ON	0.97	1.18	ON	ON	ON	ON	1.30	1.12
1jeu	ON	ON	ON	ON	0.93	1.43	ON	ON	ON	ON	0.96	0.95
1jev	ON	OFF	OFF		0.64	0.88	ON	OFF	OFF		1.05	0.81
1ola	ON	ON	**ON**	ON	0.99	0.73	ON	ON	**ON**	ON	0.89	0.74
1qka	ON	ON	OFF		0.56	0.74	ON	ON	OFF		0.89	0.88

[a] Erroneously predicted water occupancies and RMSD values > 2.0 Å are shown in bold.

Table 5. Success Rates and Estimated Significance Levels for the Water Occupancy Predictions, Both for the Training Set and the Test Set[a]

			Goldscore			Chemscore		
	N_{ON}	N_{OFF}	waters correct[b] (%)	entries correct[c] (%)	p	waters correct[b] (%)	entries correct[c] (%)	p
				Training Set				
HIV-1 (16)	8	8	100	100	0. 000016	100	100	0.000016
Fxa (14)	10	18	93	93	0.000085	71	46	0.21
Thym. K. (15)	23	11	88	87	0.000004	91	87	$<10^{-6}$
OppA (13)	40	9	96	85	0.00017	96	85	0.00017
overall	81	46	94	91	$<10^{-6}$	90	79	$<10^{-6}$
				Test Set				
primary mediating waters set (28)	40		95(1)	95(2)	$<10^{-6}$	91(1)	89(<1)	$<10^{-6}$
displaced waters set (55)		96	98(<1)	83(1)	$<10^{-6}$	83(2)	69(3)	$<10^{-6}$
decoy waters set (142)		214	95(1)	86(1)	$<10^{-6}$	94(1)	75(2)	$<10^{-6}$
overall	40	310	96(2)	86(2)	$<10^{-6}$	91(4)	75(2)	$<10^{-6}$

[a] Standard deviations for the test set results are given in parentheses. These are standard deviations in the success rates over five docking runs. These standard deviations only take into account the nondeterministic nature of the search algorithm; they do not include sampling errors, which are related to the size of the validation set (see ref 27). Assuming an overall success rate of 85%, this error is 6.8%, 4.8%, and 3.0% for the primary mediating waters set, the displaced waters set, and the decoy waters set, respectively. For the training sets for the individual targets (i.e., ~15 complexes), this error is approximately 9%. [b] Percentage of water molecules with correctly predicted occupancies. [c] Percentage of complexes for which the occupancies of all water molecules are predicted correctly.

mode. Österberg et al.[23] observed a similar performance when they used AutoDock to dock a set of 21 mostly peptidic HIV-1 inhibitors into their native X-ray structures. However, it needs to be pointed out that these authors kept the peptide main chain of the ligands rigid in the crystallographic conformation, which represents a significant reduction in the size of the search space.

Factor Xa. The results for the binding mode and water occupancy predictions for FXa are listed in Table 2. The Goldscore function produces good results for this

Table 6. Overview of the Success Rates Obtained for Binding Mode Predictions for Both the Training Set and the Test Set[a]

	Goldscore		Chemscore	
	without waters (%)	with waters (%)	without waters (%)	with waters (%)
Training Set				
primary mediated complexes (32)	78	91	75	94
secondary mediated complexes (11)	82	91	73	91
nonmediated complexes (15)	87	87	73	67
Test Set				
primary mediating waters set (28)	81(1)	89(2)	79(2)	81(1)
displaced waters set (55)	84(2)	74(3)	72(2)	67(2)
decoy waters set (142)	82(1)	75(1)	74(1)	68(<1)

[a] Standard deviations for the test set results are given in parentheses. These are standard deviations in the success rates over five docking runs. These standard deviations only take into account the nondeterministic nature of the search algorithm; they do not include sampling errors, which are related to the size of the validation set (see ref 27). Assuming an overall success rate of 75%, this error is 8.2%, 5.8%, and 3.6% for the primary mediating waters set, the displaced waters set, and the decoy waters set, respectively. For the primary mediated complexes, the secondary mediated complexes, and the nonmediated complexes in the training set, this error is 8.0%, 13.0%, and 11.1%, respectively.

target. Nearly all water occupancies are predicted correctly, and three complexes (1f0r, 1f0s, and 1xha, all forming water-mediated hydrogen bonds) are "corrected" by including the water molecules. When the water molecules are included, Goldscore correctly predicts the binding mode for 12 out of the 14 FXa complexes. The RMSD values were particularly improved for complexes 1f0r and 1f0s when the water molecules were included (see Figure 3). The aminoisoquinoline and the azaindole groups are both charged, and they interact with Asp189 through water mediation. Therefore, the presence of W2 in the docking experiments is fundamental in order to obtain the right binding mode. When W2 is not considered, the charged groups directly interact with Asp189, forming salt bridges.

The Chemscore function struggles to reproduce the binding modes and the water occupancies for this target. Still, the binding modes of two complexes (1f0s and 1fjs, both forming water-mediated hydrogen bonds) are corrected by including the water molecules. However, for the 1nfx complex, where both water molecules are displaced, the Chemscore prediction is correct when both water molecules are switched off but incorrect when the water molecules are included. Overall, the Chemscore function identifies the correct binding for only 8 of the 14 FXa complexes.

It has to be pointed out that this set of FXa complexes poses a real challenge for a scoring function. Particularly, the compounds that bind with a neutral moiety in the S1 pocket (1mq5, 1mq6, qnfu, 1nfw, 1nfx, and 1nfy) are tough test cases. These compounds also contain a basic group that binds in the S4 pocket, and most scoring functions will place that group in the S1 pocket instead. This makes the Goldscore results presented in Table 2 particularly impressive.

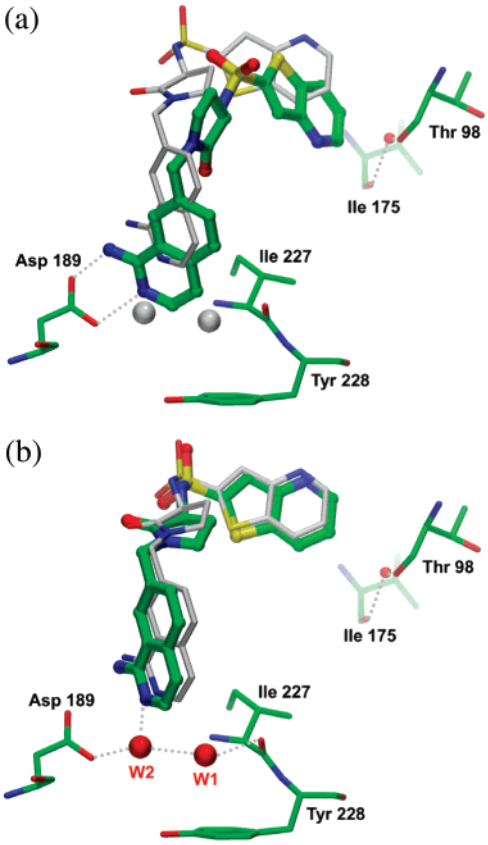


Figure 3. Training set example for FXa (PDB entry 1f0r) using Goldscore. (a) Docking performed in the absence of water molecules (X-ray water molecule positions shown in gray; RMSD = 2.95 Å). (b) Docking performed in the presence of water molecules (allowing them to toggle on/off and spin around their three principal axes; RMSD = 0.59 Å). The carbon atoms of the docking solutions are shown in green. The carbon atoms of the X-ray binding mode of the ligand are shown in gray.

Thymidine Kinase. The binding mode and water occupancy predictions for TK are listed in Table 3. Both Goldscore and Chemscore perform well against this target. The majority of the water occupancies are predicted correctly, and for both scoring functions, the binding modes of 14 out of the 15 complexes are predicted within 2 Å of their experimental binding mode. No significant improvements are observed in the quality of the binding modes when the water molecules are included. However, it is worth noting that both scoring functions perform very well in the absence of water molecules (13 out of 15 complexes are predicted correctly), so the scope for improvement was limited for this target. Various docking studies on TK have been reported in the literature. Most authors found that generally good binding modes can be produced without including the water molecules,[29−32] although Pospisil et al.[8] did observe small improvements in the RMSDs when the correct water molecules were included in the binding site.

OppA. It is clear from the results in Table 4 that both Goldscore and Chemscore perform well against OppA. Nearly all water occupancies are predicted correctly, and both scoring functions predict the binding modes of all compounds within 2 Å of their experimental binding mode. As in the case of TK, both scoring functions perform very well in the absence of water molecules, so

there was little scope for improvement of the binding modes by including water molecules.

Test Set Results

To test the performance of our methodology outside the training set, we put together three independent test sets, each of which is a subset of the CCDC/Astex test set of protein−ligand complexes.[27] We ensured that these three test sets do not contain any HIV-1, FXa, TK, or OppA complexes. All dockings were run using the optimized values for σ_p. The results for these three test sets are summarized in Tables 5 and 6. The p values for the water occupancy predictions for these three test sets were calculated assuming that, as a null hypothesis, for each water site, the probability of it being occupied is equal to the probability of it being unoccupied.

Primary Mediating Waters Set. The 28 complexes in this set contain at least one water molecule that is involved in a mediating hydrogen bond between the protein and the ligand; only water molecules for which the ligand donor or acceptor involved in the hydrogen bond to the water molecule does not form any direct hydrogen bonds with the protein were added. For both scoring functions, our docking protocol has correctly switched on the mediating water molecules in more than 90% of the cases. Upon including the water molecule(s), some improvements are observed in the quality of the predicted binding modes. For the Goldscore function, for example, the success rate for predicting the binding modes increases from 81% without water molecules to 89% when water molecules are included. It has to be noted that these improvements are not very significant from a statistical point of view, particularly if we take into account the sampling error associated with such a small test set (see Table 6). However, when all cases where we observe an improvement in the predicted binding mode are inspected, it is clear that the included water molecule(s) has/have caused the improvement. For example, when the Chemscore function is used to dock the ligand of PDB entry 1a4g from this test set against the empty (neuraminidase) binding site, the carboxylic acid part of the ligand is docked correctly, but the rest of the ligand is twisted with respect to the experimental binding mode (see Figure 4). When a key water molecule that is situated between Glu255 and Glu275 is included, the docking algorithm switches it on and allows it to form two quality hydrogen bonds with the ligand, hence almost exactly reproducing the experimental binding mode.

Displaced Waters Set. This set contains 55 complexes where the ligand has displaced at least one water molecule that was observed in another X-ray structure of the same target. Our docking protocol has correctly switched off ~90% of the water molecules in this test set, for both the Goldscore and the Chemscore function. A small drop-off (5−10%) in the success rates for the binding mode predictions is observed when these displaced water molecules are included. This is probably caused by false high-ranking docking solutions that involve interactions with one or more water molecules, as there is a high correlation between the erroneous switching on of water molecules and the mispredicted binding modes; of the eight complexes that are predicted correctly without waters but are mispredicted when waters are included, seven complexes have at least one water molecule erroneously switched on. On the other hand, there is one complex (2h4n, Goldscore) that is mispredicted in the absence of water but predicted correctly when waters are included; we can only ascribe this to chance because, if GOLD is allowed to toggle on/off water molecules that are not actually present in the experimental complex, this complicates the search space and generates potential decoy solutions.

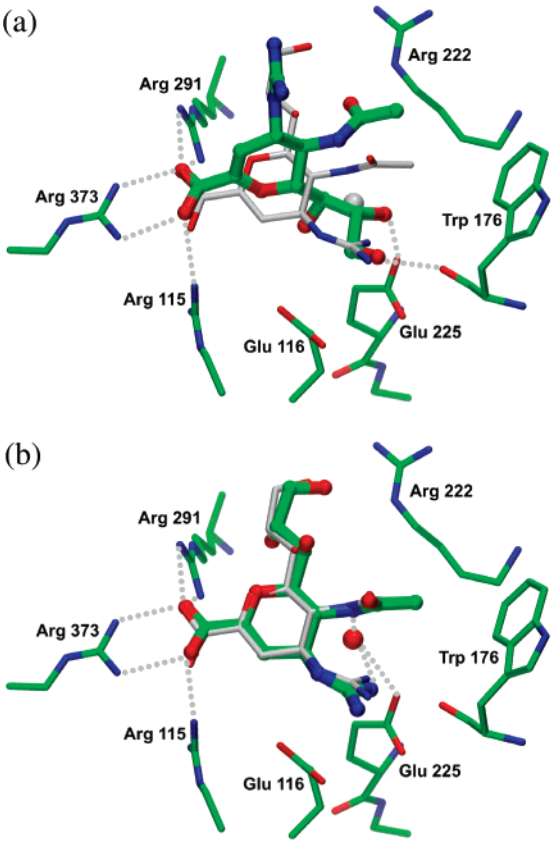


Figure 4. Test set example for neuraminidase (PDB entry 1a4g) using Chemscore. (a) Docking performed in the absence of water molecules (X-ray water molecule position shown in gray; RMSD = 4.67 Å). (b) Docking performed in the presence of a water molecule (allowing it to toggle on/off and spin around its three principal axes; RMSD = 0.38 Å). The carbon atoms of the docking solutions are shown in green. The carbon atoms of the X-ray binding mode of the ligand are shown in gray.

Decoy Waters Set. This set contains 142 complexes where one or more water molecules were added in positions that were identified by SuperStar[14] to be favorable water molecule sites and that overlap with the ligand in its experimental binding mode. Again, for both scoring functions, our algorithm has correctly switched off >90% of the water molecules in this test set. As for the displaced waters set, we see a small drop-off (~7%) in success rates for binding mode predictions, probably caused by false docking solutions that involve interactions between ligand and water molecule(s); of the 15 complexes that are predicted correctly without waters but are mispredicted when waters are included, 14 complexes have at least one water molecule erroneously switched on.

Discussion and Conclusions

We have implemented a novel method for dealing with key water molecules in protein−ligand docking.

This method takes into account the loss of rigid-body entropy when a water molecule binds to a protein, and allows explicit all-atom water molecules to appear and disappear and to spin around their three principal axes.

We have also presented an extensive and in-depth two-part validation study of the implemented methodologies. In the first part, a constant penalty, σ_p, representing the loss of rigid-body entropy, was optimized against a training set of 58 protein−ligand complexes; this was done for both the Goldscore and the Chemscore scoring functions. In reality, this term is not a constant but can be target dependent, or even water site dependent, as water molecules that bind very tightly to the protein lose more rigid-body entropy than weakly binding ones.[28,33] Although we have found that the algorithm performs well across a range of targets and water sites using a constant value for σ_p, it may be worth (or in some cases even necessary) reoptimizing the σ_p values in individual cases.

Using the optimized σ_p values to dock the compounds in the training set against their respective target structures and allowing all key water molecules to toggle on/off and to spin around, our algorithm correctly predicts whether a water molecule is present or displaced in approximately 92% of the cases (94% for Goldscore; 90% for Chemscore). We also observe a small but significant improvement in the quality of the predicted binding modes when water molecules are included for primary mediated complexes. No (statistically significant) effect on the quality of the predicted binding modes is observed for the secondary mediated and nonmediated complexes in the training set.

In the validation on an independent test set of a total of 225 complexes, water occupancies could be predicted with similarly high success rates as were obtained for the training set (approximately 90%). For the primary mediating waters set, clear improvements in the binding modes for individual complexes are observed upon including the water molecule(s). However, because of the small size of this test set, these improvements are not significant from a statistical point of view. For the displaced and decoy waters sets, a small drop-off (∼7%) in the success rates for predicting the binding modes is observed when the water molecules are included.

The fact that we only observe limited improvement in terms of the quality of the binding modes for the two pimary mediating sets (training and test sets) is not surprising in itself. The interactions a ligand forms with water molecules generally only represent a small fraction of the number of interactions the ligand forms in its bound state. Hence, it is not uncommon that a reasonable binding mode can be produced without including even key structural water molecules. Also, the fact that we observe a small drop-off in success rates for the displaced and decoy waters test sets is almost inevitable. All water molecules included in these sets are in good positions for a water molecule to interact with the protein and are, therefore, not easily displaced. Additionally, because the water molecules in these sets should be displaced by the ligand, including them can only distract the docking algorithm from identifying the correct binding mode. Each water molecule that is included increases both the search space and the likelihood of obtaining false positives. We, therefore, feel it is important to limit the number of water molecules included in a docking run and only include water molecules that are known to be crucial for ligand binding.

Compared to the training set, we obtain slightly worse results for the test set in terms of the quality of the predicted binding modes. We only see marginal improvement for the primary mediating waters set upon including water molecules, whereas this improvement is more pronounced for the primary mediated complexes in the training set. Also, we see a small drop-off in the success rates for the displaced and decoy waters sets when we include water molecules, whereas we do not see this drop-off for the nonmediated complexes in the training set. An obvious explanation for this result is that we have trained on the training set and, therefore, expect better results. Another possibility is that the training set only contains targets for which we know that the included water molecules are important for protein−ligand recognition, whereas the test sets were constructed automatically from the CCDC/Astex validation set and contain a wide range of targets. Also, most of the validation sets we have used here are quite small and, hence, the sampling errors are quite large (∼6%), which means that some of the differences observed may be due to random statistical variations.

What is probably more important than improvements in binding mode quality is that our docking algorithm can predict the water occupancies with a high degree of certainty (90%). Although this does not always have an impact on the quality of the binding modes produced, it will affect the scores of the docked compounds. Hence, using our protocol, including water molecules in the docking runs could improve the correlation of the scores with the affinities of the compounds and may improve the enrichments obtained in virtual screening experiments. We are keen to investigate this in a future study.

Materials and Methods

To predict whether a water molecule is present or absent in a protein−ligand complex, we need to estimate the free-energy change, ΔG_b, associated with transferring a water molecule from the bulk solvent to its binding site in a protein−ligand complex. Here, we define ΔG_b for a given water molecule w as follows:

$$\Delta G_b(w) = \Delta G_p(w) + \Delta G_i(w) \quad (1)$$

where $\Delta G_p(w)$ is the free energy associated with the loss of rigid-body entropy on binding to the target. $\Delta G_i(w)$ contains contributions resulting from the interactions that the water molecule forms with the protein and the ligand (relative to those it forms with bulk solvent) and also reflects any changes in the interactions between the protein and the ligand caused by the introduction of the water molecule. These interactions may be favorable (e.g., hydrogen bonds) or unfavorable (e.g., steric clashes). We will refer to $\Delta G_i(w)$ as the *intrinsic* binding affinity of a water molecule.

For a water molecule to bind to a protein−ligand complex, its intrinsic binding affinity needs to outweigh the loss of rigid-body entropy on binding. Therefore, to predict water mediation and water displacement, the balance of the two terms in eq 1 is critical. As far as we are aware, no protein−ligand docking scoring function takes the $\Delta G_p(w)$ term into account directly. However, in FlexX, Rarey et al.[20] did add a penalty for "vacant interactions" on water molecules, which could account for the

loss of rigid-body entropy, indirectly. In practice, $\Delta G_p(w)$ will vary for different water binding sites because tighter binding water molecules will lose more rigid-body entropy upon binding to the target than loosely binding water molecules.[28,33,34] However, to keep our model simple, we will assume that ΔG_p is a constant.

GOLD Implementation. We implemented code for the treatment of water mediation and displacement into the protein−ligand docking program GOLD.[24,25] We chose not to let the program predict water mediation sites but rather to use fixed positions for water molecules provided by the user. There are two reasons why we took this approach: (i) it reduces the search space drastically, particularly for water molecules with explicit hydrogen atoms such as we use here; (ii) we believe that, if a water molecule is important for binding, its position is generally known from experimental protein structures. Hence, for each water molecule in a binding site, two possible operations are added to the GOLD Genetic Algorithm (GA): the occupancy of each water molecule can switch between "ON" and "OFF", and each water molecule can be spun around three orthogonal axes.

GOLD uses fitting points on the protein and the ligand in order to place the ligand into the binding site.[24] When *active* water molecules (i.e., water molecules that may appear and disappear) are added to a binding site, the fitting points on the water molecules are simply added to the list of protein fitting points.

Scoring Functions. Two GOLD scoring functions, Goldscore and Chemscore, were extended to include the contributions from mediating or displaced water molecules. In both cases, the intrinsic binding affinity of a water molecule was modeled using the exact same functional forms as those used to model regular protein−ligand interactions. Hence, for Goldscore, the interaction of a water molecule with the protein and the ligand was described as the sum of a hydrogen-bond term and a van der Waals term, using the standard Goldscore parameters for both terms. For Chemscore, the interactions of a water molecule were described as the sum of a hydrogen-bond term, a metal term, and a clash term, using the same Chemscore parameters we used in a previous study.[35]

As explained above, we assumed that the free energy penalty associated with the loss of rigid-body entropy upon binding to the protein is constant for all water molecules. Therefore, for both scoring functions, the GOLD Fitness function can be written as

$$\text{Fitness} = \sigma_o + \sum_w o(w)(\sigma_p + \sigma_i(w)) \qquad (2)$$

where σ_o is the original score (i.e., Goldscore or Chemscore) for a given binding mode of the ligand; $o(w)$ is the occupancy of water molecule w and is equal to either 1 if the water is switched on or zero if it is switched off; σ_p represents the free energy penalty associated with the loss of rigid-body entropy (i.e., ΔG_p); $\sigma_i(w)$ reflects the intrinsic binding affinity of water molecule w (see above) and represents the summation over the interactions formed by this water molecule with all ligand atoms, protein atoms, and other water molecules w' for which $o(w') = 1$; and the summation is over all water molecules. The fact that we take into account interactions between water molecules means that the methodology should also be able to predict the orientation of water molecules that form part of a water network.

Docking Protocol. The current investigation is aimed at testing the ability of the two modified scoring functions to predict the presence/absence of water molecules and establishing whether including the water molecules improves the binding modes produced. Hence, we need to minimize the effect of other factors on the docking success rates. To eliminate, as much as possible, any dependencies of the performance on the search algorithm, we used long search settings for the GOLD GA: 100 dockings with 100 000 GA operations per docking; the algorithm was not allowed to terminate early when the same solution was produced repeatedly. Additional GA parameters were taken from the Default 1 GOLD GA settings. These GA settings were used for all docking runs presented in this study. Furthermore, each ligand was docked into its native protein structure to avoid cross-docking experiments that could complicate the analysis. All dockings were run in the absence of water molecules and in their presence, allowing them to spin around their three principal axes and to appear and disappear.

Training Set. The training set used in this work contains 58 protein−ligand complexes from four different targets: HIV-1, FXa, TK, and the oligopeptide-binding protein OppA. For each target, around 15 complexes were selected from the PDB, ensuring there were examples of water mediation and water displacement in each target test set. Each structure was visually checked to ensure it does not contain clashes between the protein and the ligand or contacts to crystallographically related protein chains. All complexes in the training set were determined at a resolution better than 2.5 Å, except for 1vtk (2.75 Å), 3vkt (3.0 Å), and 2upj (3.0 Å).

The preparation of the ligand and binding site was as follows: (i) for each target, the structures were superimposed based on the residues in the binding site; (ii) in the resulting frame of reference, the ligands were saved separately from the proteins; (iii) only the key structural water molecules were retained in the protein structures (if one of these key water molecules was displaced by the ligand, it was copied from the structure in the training set that is structurally most similar around the water site; however, if a water molecule was displaced by the *protein*, it was not included and it was not taken into account in the calculation of the success rates); (iv) hydrogen atoms were added to the protein and the ligand, taking considerable care to assign bond types and protonation/tautomeric states correctly; (v) all protein atoms within 6 Å of any non-hydrogen atom in the ligand were included in the binding site definitions.

For HIV-1, only structural water W1 (see Figure 1a) was included in the docking runs. In complexes where W1 is present, it forms hydrogen-bond bridges between the backbone amide protons of Ile A50 and Ile B50 and the inhibitor. From the PDB, eight HIV-1 complexes were chosen where W1 is mediating (1hpv, 1hsg, 1hxw, 1ohr, 1sbg, 1d4i, 1npv, and 1ebz) and eight complexes were chosen where W1 is displaced (1ajv, 1ajx, 1hvr, 1hwr, 1pro, 1qbs, 2upj, and 7upj). Following the suggestions of Brik and Wong,[36] one of the catalytic aspartic acids (Asp25:A) was protonated.

Figure 1b shows the two water molecules that we considered in each FXa binding site. W1 is located deep in the S1 pocket and forms a hydrogen bond with the backbone carbonyl of Ile227; W1 is also in van der Waals contact with the Tyr228 phenyl ring. W1 can mediate between the protein and the ligand (1ezq, 1f0s, 1fjs, 1g2l, 1g2m, 1ksn, 1kye, and 1xka), or it can be displaced (1mq5, 1mq6, 1nfu, 1nfw, 1nfx, and 1nfy). W2 is observed only in 1f0r and 1f0s, where it mediates between the ligand, Asp189, and W1. In all other structures, W2 is displaced by the ligand. A third water molecule is located in the S4 pocket, where it binds to the backbone carbonyls of Ile175 and Thr98. This water molecule is present in all structures, and its orientation is well conserved. Hence, we added this water molecule to all FXa binding sites, as part of the protein.

Crystal structures of TK show that Gln125 can adopt two different conformations (see Figure 1c). When Gln125 adopts conformation A (1e2k, 1e2m, 1e2n, 1e2p, 1ki4, 1ki6, 1ki7, 1ki8, 1kim, 1vtk, and 3vtk), water molecules W1 and W2 form mediating hydrogen bonds between the protein and the ligand. W1 is hydrogen bonded to the side chains of Tyr101 and Arg176; W2 hydrogen bonds to the side chains of Arg176 and Gln125. When Gln125 adopts conformation B (1ki2, 1ki3, 1qhi, and 2ki5), W1 and W2 are displaced by the ligand. In two of these complexes, a third water molecule, W3, forms mediating hydrogen bonds between the protein and the ligand; W3 forms hydrogen bonds to the backbone carbonyl of Ala168 and to the side chain of Gln125. Two water molecules (W1 and W2) were considered in the complexes where Gln123 adopts conformer

A, whereas three water molecules (W1, W2, and W3) were considered in the complexes where Gln125 adopts conformation B.

The OppA X-ray crystal structures used in this study are cocomplexes with tripeptides that have the sequence Lys-X-Lys. The various side chains at position X in the ligand bind in a hydrated pocket. Water molecules act as flexible adapters, matching the hydrogen-bonding requirements of the protein and the ligand. Figure 1d shows the four water molecules we considered in this study. W1 is present in all the complexes. It is tightly bound to the backbone of Gly415 and to Arg404, and it mediates with the ligands through the formation of a hydrogen bond with the carboxylic function at the C-terminus of the peptides. W2 mediates with W1, and it also binds to the side chain of Tyr274 and to the backbone of Gly415. W2 can mediate between the protein and the ligand (1b3f, 1b4h, 1b4z, 1b5i, 1b5j, 1jeu, 1ola, and 1qka) or can be displaced (1b0h and 1jev), and there are compounds with hydrophobic residues at position X (see above) that do not mediate or displace it (1b1h, 1b3h, and 1b58). W3 is hydrogen bonded with the carbonyl backbone of Asn436, with the side chain of Thr438, and with W4. W3 can mediate between the protein and the ligand (1b4h, 1b4z, 1b5i, 1b5j, and 1jeu), it can be displaced (1b0h, 1b3f, 1b3h, 1b58, 1jev, 1ola, and 1qka), and the hydrophobic residue at position X (see above) in the 1b1h ligand does not mediate or displace it. W4 mediates with W3, and it also binds to the side chains of Asn436 and Glu32. However, it was observed that Glu32 could displace W4 in order to directly interact with the ligand (1b3f, 1jer, and 1qka). In these complexes, W4 was not considered in our docking experiments. W4 can mediate between the protein and the ligand (1b4z, 1b5i, 1b5j, 1jeu, and 1ola), and there are compounds with hydrophobic residues at position X that do not mediate or displace it (1b0h, 1b1h, 1b3h, and 1b4h).

Test Sets. Three separate test sets of protein−ligand complexes were constructed for this work: a primary mediating waters set, a displaced waters set, and a decoy waters set. All three sets of complexes are subsets of the CCDC/Astex test set.[27] Only complexes from this test set that do not have protein−ligand clashes, crystallographic contacts, etc. were included here; these complexes are termed "clean". All entries of HIV-1, FXa, TK, and OppA complexes were excluded, leaving 186 clean complexes that are truly independent of the training set.

To construct the primary mediating waters set, each entry in the clean subset of the CCDC/Astex test set was checked for water molecules that form a hydrogen bond to both the protein and the ligand, ensuring that the ligand atom involved in the hydrogen bond to the water molecule does not form a direct hydrogen bond with the protein. These water molecules were protonated, merged into the protein structure files, and added to the definition of the binding site. This resulted in a primary mediating waters set of 28 complexes; 20 of these complexes contain exactly one mediating water molecule, and the remaining 8 complexes contain 2−4 water molecules. All complexes in this test set were determined at a resolution better than 2.5 Å, except for 1uvs (2.8 Å) and 1acj (2.8 Å).

The approach taken to construct the displaced waters set was as follows. For each entry in the clean subset of the CCDC/Astex test set, the amino acid sequence of the protein chain in contact with the ligand was used as a query in a FASTA[37] search against the PDB to identify entries that contain protein chains with a high degree of homology (>95%); only entries determined at a resolution better than 2.5 Å were accepted. Each matching chain was then superimposed (based on binding-site residues only) onto the protein chain of the reference test set entry; all water molecules in the PDB entry containing the matching chain were also transformed to this new frame of reference. This resulted in a distribution of water molecules in the binding site of the reference test set entry (see Figure 5). Water molecules that were >1.0 Å away from the ligand in the reference test set entry were removed from the distribution. If the resulting distribution contained less than four water molecules, the test set entry was rejected.

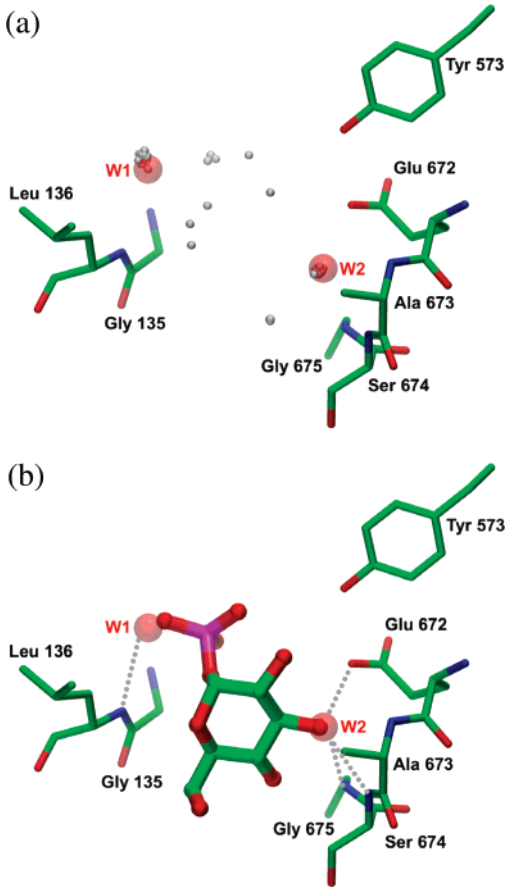


Figure 5. Example of a displaced waters set entry (PDB entry 3gpb, glycogen phosphorylase). (a) Distribution of water molecules extracted from structures with high sequence homology to PDB entry 3gpb; extracted water molecules (W1 and W2) are shown in red. (b) Extracted water molecules overlaid with the native ligand in the binding site of 3gpb.

Otherwise, the distribution of water molecules was converted to a three-dimensional map containing the density of water molecules at each grid point. Water molecule densities were set to zero at grid points where the density was <25% of the maximum density observed in the map. From each peak in the resulting density map (starting with the highest peak), a representative water molecule was selected (see Figure 5), ensuring that (i) the peak is made up of at least four water molecules; (ii) the selected water molecule does not clash with other selected water molecules or with protein atoms in the test set entry; (iii) the selected water molecule is within 3.3 Å of a protein atom in the test set entry; and (iv) the local protein environment of the water molecule in the reference is geometrically similar to the environment of the water in its native protein. Selected water molecules were protonated, merged into the protein structure file of the test set entry, and added to the definition of the binding site. This entire process was carried out using ReliScript (part of the ReliBase program[38]) and resulted in a displaced waters set of 55 complexes; 28 of these complexes contain one water molecule, 17 contain two water molecules, and the remaining 10 complexes contain 3−5 water molecules. All complexes in this test set were determined at a resolution better than 2.5 Å.

The decoy waters set of complexes was generated using the SuperStar modeling tool.[14,39−41] SuperStar is a knowledge-based tool that allows calculation of favorable positions of probes in a binding site. It uses crystallographic information on noncovalent contacts from the Cambridge Structural Database[42] to produce spatial maps that depict areas of possible interaction in the protein cavity. In this case, spatial propensity maps (see Figure 6a) were calculated for water oxygen interactions, and peak fitting of those maps was

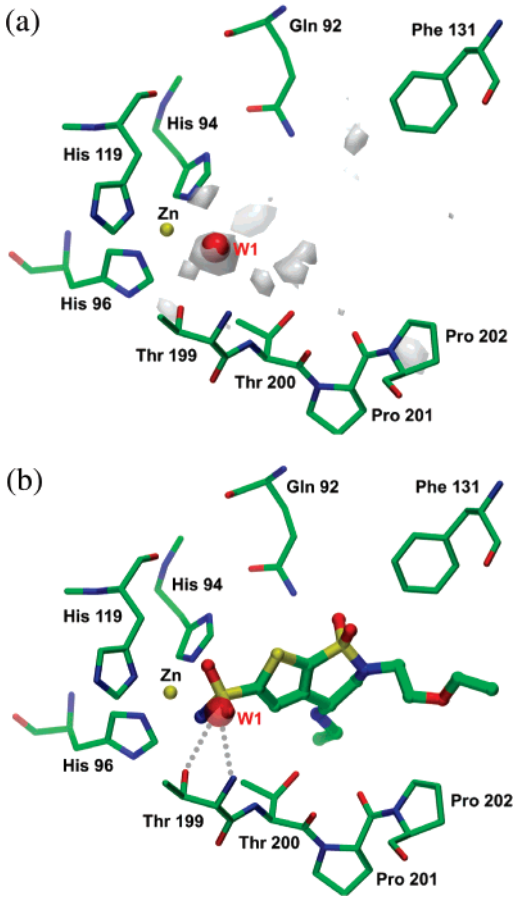


Figure 6. Example of a decoy waters set entry (PDB entry 1a42, carbonic anhydrase). (a) SuperStar propensity map for water oxygen atoms (contoured at propensity 2.8); the decoy water molecule (W1) generated from this map is shown in red. (b) Decoy water molecule overlaid with the native ligand in the binding site of 1a42.

performed to obtain favored positions (peaks in the maps) for hydrogen-bonding water oxygens. Favorable water positions were then identified by selecting peaks in order of the highest propensity, within 1.0 Å of any of the native ligand's heavy atoms and within hydrogen-bonding distance of the protein (see Figure 6b). A propensity threshold of 4.0 was applied to eliminate weakly binding waters. If more than one acceptable peak was found, checks were made for water−water clashes and only nonclashing peak positions were accepted. At the identified positions, water oxygen atoms were placed, which were then protonated, merged into the protein structure files, and added to the definition of the binding site. The above procedure resulted in a set of 142 complexes with decoy water molecules. Of the 142 complexes, 90 contain a single virtual water molecule; remaining complexes contain 2−5 waters. All complexes in this test set were determined at a resolution better than 2.5 Å, except for 1acm (2.8 Å).

Acknowledgment. The authors would like to thank Dr. Will Pitt for pointing out the thymidine kinase example and the roles played by water molecules in this system, Dr. Suzanne Brewerton for her help putting together the primary mediating waters set, and Louise Birch for running additional validation runs that are not presented in this study.

References

(1) 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.
(2) Chen, Z.; Li, Y.; Chen, E.; Hall, D. L.; Darke, P. L.; Culberson, C.; Shafer, J. A.; and Kuo, L. C. Crystal structure at 1.9-A resolution of human immunodeficiency virus (HIV) II protease complexed with L-735, 524, an orally bioavailable inhibitor of the HIV proteases. *J. Biol. Chem.* **1994**, *269*, 26344−26348.
(3) Lam, P. Y. S.; Jadhav, P. K.; Eyermann, C. J.; Hodge, C. N.; Ru, Y.; Bacheler, L. T.; Meek, O.; and Rayner, M. M. Rational design of potent, bioavailable, nonpeptide cyclic ureas as HIV protease inhibitors. *Science (Washington, D.C.)* **1994**, *263*, 380−384.
(4) Klein, S. I.; Czekaj, M.; Gardner, C. J.; Guertin, K. R.; Cheney, D. L.; Spada, A. P.; Bolton, S. A.; Brown, K.; Colussi, D.; Heran, C. L.; Morgan, S. R.; Leadley, R. J.; Dunwiddie, C. T.; Perrone, M. H.; Chu, V. Identification and initial structure−activity relationships of a novel class of nonpeptide inhibitors of blood coagulation factor Xa. *J. Med. Chem.* **1998**, *41*, 437−450.
(5) Herron, D. K.; Goodson, T., Jr.; Wiley, M. R.; Weir, L. C.; Kyle, J. A.; Yee, Y. K.; Tebbe, A. L.; Tinsley, J. M.; Mendel, D.; Masters, J. J.; Franciskovich, J. B.; Sawyer, J. S.; Beight, D. W.; Ratz, A. M.; Milot, G.; Hall, S. E.; Klimkowski, V. J.; Wikel, J. H.; Eastwood, B. J.; Towner, R. D.; Gifford-Moore, D. S.; Craft, T. J.; Smith, G. F. 1,2-Dibenzamidobenzene inhibitors of human factor Xa. *J. Med. Chem.* **2000**, *43*, 859−872.
(6) Yee, Y. K.; Tebbe, A. L.; Linebarger, J. H.; Beight, D. W.; Craft, T. J.; Gifford-Moore, D.; Goodson, T., Jr.; Herron, D. K.; Klimkowski, V. J.; Kyle, J. A.; Sawyer, J. S.; Smith, G. F.; Tinsley, J. M.; Towner, R. D.; Weir, L.; Wiley, M. R. N(2)−Aroylanthranilamide inhibitors of human factor Xa. *J. Med. Chem.* **2000**, *43*, 873−882.
(7) Yang, J. M.; Chen, C. C. GEMDOCK: a generic evolutionary method for molecular docking. *Proteins* **2004**, *55*, 288−304.
(8) Pospisil, P.; Kuoni, T.; Scapozza, L.; Folkers, G. Methodology and problems of protein−ligand docking: case study of dihydroorotate dehydrogenase, thymidine kinase, and phosphodiesterase 4. *J. Recept. Signal Transduction Res.* **2002**, *22*, 141−154.
(9) De Graaf, C.; Pospisil, P.; Pos, W.; Folkers, G.; Vermeulen, N. P. E. Binding Mode Prediction of Cytochrome P450 and Thymidine Kinase Protein−Ligand Complexes by Consideration of Water and Rescoring in Automated Docking. *J. Med. Chem.* **2005**, *48*, 2308−2318.
(10) Birch, L.; Murray, C. W.; Hartshorn, M. J.; Tickle, I. J.; Verdonk, M. L. Sensitivity of molecular docking to induced fit effects in influenza virus neuraminidase. *J. Comput.-Aided Mol. Des.* **2002**, *16*, 855−869.
(11) Pitt, W. R.; Goodfellow, J. M. Modelling of solvent positions around polar groups in proteins. *Protein Eng.* **1991**, *4*, 531−537.
(12) Goodford, P. J. A computational-procedure for determining energetically favorable binding-sites on biologically important macromolecules. *J. Med. Chem.* **1985**, *28*, 849−857.
(13) Miranker, A.; Karplus, M. Functionality maps of binding sites: a multiple copy simultaneous search method. *Proteins* **1991**, *11*, 29−34.
(14) 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.
(15) Kortvelyesi, T.; Dennis, S.; Silberstein, M.; Brown, L., III; Vajda, S. Algorithms for computational solvent mapping of proteins. *Proteins* **2003**, *51*, 340−351.
(16) Gunther, J.; Bergner, A.; Hendlich, M.; Klebe, G. Utilising structural knowledge in drug design strategies: applications using Relibase. *J. Mol. Biol.* **2003**, *326*, 621−636.
(17) Raymer, M. L.; Sanschagrin, P. C.; Punch, W. F.; Venkataraman, S.; Goodman, E. D.; Kuhn, L. A. Predicting conserved water-mediated and polar ligand interactions in proteins using a K-nearest-neighbors genetic algorithm. *J. Mol. Biol.* **1997**, *265*, 445−464.
(18) 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.
(19) 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.
(20) Rarey, M.; Kramer, B.; Lengauer, T. The particle concept: Placing discrete water molecules during protein−ligand docking predictions. *Proteins* **1999**, *34*, 17−28.
(21) Schnecke, V.; Kuhn, L. A. Virtual screening with solvation and ligand-induced complementarity. *Perspect. Drug Discovery Des.* **2000**, *20*, 171−190.
(22) Goodsell, D. S.; Olson, A. J. Automated docking of substrates to proteins by simulated annealing. *Proteins* **1990**, *8*, 195−202.
(23) Österberg, F.; Morris, G. M.; Sanner, M. F.; Olson, A. J.; Goodsell, D. S. Automated docking to multiple target structures: incorporation of protein mobility and structural water heterogeneity in AutoDock. *Proteins* **2002**, *46*, 34−40.

(24) Jones, G.; Willett, P.; Glen, R. C. Molecular recognition of receptor sites using a genetic algorithm with a description of desolvation. *J. Mol. Biol.* **1995**, *245*, 43−53.
(25) 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.
(26) 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.
(27) Nissink, J. W. M.; Murray, C. W.; Hartshorn, M. J.; Verdonk, M. L.; Cole, J. C.; Taylor, R. A new test set for validating predictions of protein−ligand interaction. *Proteins* **2002**, *49*, 457−471.
(28) Dunitz, J. D. The entropic cost of bound water in crystals and biomolecules. *Science (Washington, D.C.)* **1994**, *264*, 670.
(29) Bissantz, C.; Folkers, G.; Rognan, D. Protein-based virtual screening of chemical databases. 1. Evaluation of different docking/scoring combinations. *J. Med. Chem.* **2000**, *43*, 4759−4767.
(30) Friesner, R. A.; Banks, J. L.; Murphy, R. B.; Halgren, T. A.; Klicic, J. J.; Mainz, D. T.; Repasky, M. P.; Knoll, E. H.; Shelley, M.; Perry, J. K.; Shaw, D. E.; Francis, P.; Shenkin, P. S. Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *J. Med. Chem.* **2004**, *47*, 1739−1749.
(31) Jain, A. N. Surflex: fully automatic flexible molecular docking using a molecular similarity-based search engine. *J. Med. Chem.* **2003**, *46*, 499−511.
(32) Li, H.; Li, C.; Gui, C.; Luo, X.; Chen, K.; Shen, J.; Wang, X.; Jiang, H. GAsDock: a new approach for rapid flexible docking based on an improved multi-population genetic algorithm. *Bioorg. Med. Chem. Lett.* **2004**, *14*, 4671−4676.
(33) Dunitz, J. D. Win some, lose some: enthalpy−entropy compensation in weak intermolecular interactions. *Chem. Biol.* **1995**, *2*, 709−712.
(34) Renzoni, D.A.; Zvelebil, M. J. J. M.; Lundback, T.; Ladbury, J. E. *Exploring uncharted waters: water molecules in drug design strategies*; Springer-Verlag: Berlin, Germany, 1997; pp 161−180.
(35) Verdonk, M. L.; Cole, J. C.; Hartshorn, M.; Murray, C. W.; Taylor, R. D. Improved protein−ligand docking using GOLD. *Proteins* **2003**, *52*, 609−623.
(36) Brik, A.; Wong, C. H. HIV-1 protease: mechanism and drug discovery. *Org. Biomol. Chem.* **2003**, *1*, 5−14.
(37) Pearson, W. R. Lipman, D. J. Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci. U.S.A.* **1988**, *85*, 2444−2448.
(38) Hendlich, M.; Bergner, A.; Gunther, J.; Klebe, G. Relibase: design and development of a database for comprehensive analysis of protein−ligand interactions. *J. Mol. Biol.* **2003**, *326*, 607−620.
(39) Boer, D. R.; Kroon, J.; Cole, J. C.; Smith, B.; Verdonk, M. L. SuperStar: comparison of CSD and PDB-based interaction fields as a basis for the prediction of protein−ligand interactions. *J. Mol. Biol.* **2001**, *312*, 275−287.
(40) Nissink, J. W. M.; Taylor, R. Combined use of physicochemical data and small-molecule crystallographic contact propensities to predict interactions in protein binding sites. *Org. Biomol. Chem.* **2004**, *2*, 3238−3249.
(41) Verdonk, M. L.; Cole, J. C.; Watson, P.; Gillet, V.; Willett, P. SuperStar: improved knowledge-based interaction fields for protein binding sites. *J. Mol. Biol.* **2001**, *307*, 841−859.
(42) Allen, F. H. The Cambridge Structural Database: a quarter of a million crystal structures and rising. *Acta Crystallogr., Sect. B* **2002**, *58*, 380−388.
(43) Hartshorn, M. J. AstexViewer: a visualisation aid for structure-based drug design. *J. Comput.-Aided Mol. Des.* **2002**, *16*, 871−881.

JM050543P