J|A|C|S
ARTICLES
Published on Web 02/09/2007

Classification of Water Molecules in Protein Binding Sites

Caterina Barillari,[†] Justine Taylor,[†] Russell Viner,[‡] and Jonathan W. Essex*,[†]

Contribution from the School of Chemistry, University of Southampton, Highfield, Southampton, SO17 1BJ, U.K., and Syngenta, Jealott's Hill International Research Centre, Bracknell, RG42 6EY, U.K.

Received September 28, 2006; E-mail: j.w.essex@soton.ac.uk

Abstract: Water molecules play a crucial role in mediating the interaction between a ligand and a macromolecular receptor. An understanding of the nature and role of each water molecule in the active site of a protein could greatly increase the efficiency of rational drug design approaches: if the propensity of a water molecule for displacement can be determined, then synthetic effort may be most profitably applied to the design of specific ligands with the displacement of this water molecule in mind. In this paper, a thermodynamic analysis of water molecules in the binding sites of six proteins, each complexed with a number of inhibitors, is presented. Two classes of water molecules were identified: those conserved and not displaced by any of the ligands, and those that are displaced by some ligands. The absolute binding free energies of 54 water molecules were calculated using the double decoupling method, with replica exchange thermodynamic integration in Monte Carlo simulations. It was found that conserved water molecules are on average more tightly bound than displaced water molecules. In addition, Bayesian statistics is used to calculate the probability that a particular water molecule may be displaced by an appropriately designed ligand, given the calculated binding free energy of the water molecule. This approach therefore allows the numerical assessment of whether or not a given water molecule should be targeted for displacement as part of a rational drug design strategy.

1. Introduction

The importance of water molecules in protein−ligand binding has been increasingly recognized over the years. Water molecules can stabilize the complex between a protein and a ligand by hydrogen bonding with the two components, as is observed, for example, in human dihydrofolate reductase complexed with methotrexate[1] and in hen egg-white lysozyme complexed with the Fv fragment of the monoclonal antibody.[2] In some cases, water molecules can be displaced upon ligand binding, and this can lead to an increase in binding affinity owing to a favorable gain in entropy associated with the release of a well-ordered water molecule into bulk solvent.[3] The classical example of this situation is that of HIV-1 protease, where cyclic urea derivatives were designed to displace a conserved water molecule observed in all crystal structures of the protease complexed with linear peptido-mimetic inhibitors.[4] These cyclic urea derivatives were found to be potent inhibitors of the protease. There are also cases where the displacement of water molecules has been associated with a decrease in binding affinity of the ligands displacing the water molecules. This was observed, for example, in OppA binding to different tripeptides of the form Lys-X-Lys, where the water pattern in the binding site changes according to the nature of X, and when X displaces water molecules there is a decrease in binding affinity.[5]

Until a few years ago it was common practice to ignore water molecules in protein binding sites, but recently a few papers describing docking[6−9] and drug design[10−12] with inclusion of water molecules have been published, and they showed that results are much more accurate when water molecules are taken into account. The main problem for the consideration of water molecules in drug design is to know which molecules are important in mediating the interaction between a protein and a ligand and which, instead, can be targeted for displacement.

Poornima and Dean[3,13,14] published three papers on the problem of hydration in drug design, where they identified common characteristics for water molecules in the binding sites of proteins complexed with inhibitors. These water molecules

[†] University of Southampton.
[‡] Syngenta, Jealott's Hill International Research Centre.

(1) Meiering, E. M.; Wagner, G. *J. Mol. Biol.* **1995**, *247*, 294−308.
(2) Bhat, T. N.; Bentley, G. A.; Boulot, G.; Greene, M. I.; Tello, D.; Dall'Acqua, W.; Souchon, H.; Schwarz, F. P.; Mariuzza, R. A.; Poljak, R. J. *Proc. Natl. Acad. Sci. U.S.A.* **1994**, *91*, 1089−1093.
(3) Poornima, C. S.; Dean, P. M. *J. Comput.-Aided Mol. Des.* **1995**, *9*, 500−512.
(4) Lam, P. Y. S.; Jadhav, P. K.; Eyermann, C. J.; Hodge, C. N.; Ru, Y.; Bacheler, L. T.; Meek, J. L.; Otto, M. J.; Rayner, M. M.; Wong, Y. N.; Chang, C.; Weber, P. C.; Jackson, D. A.; Sharpe, T. R.; Erickson-Viitanen, S. *Science* **1994**, *263*, 380−384.
(5) Ladbury, J. E. *Chem. Biol.* **1996**, *3*, 973−980.
(6) Minke, W. E.; Diller, D. J.; Hol, W. G. J.; Verlinde, C. L. M. J. *J. Med. Chem.* **1999**, *42*, 1778−1788.
(7) Rarey, M.; Kramer, B.; Lengauer, T. *Proteins* **1999**, *34*, 17−28.
(8) de Graaf, C.; Pospisil, P.; Pos, W.; Folkers, G.; Vermeulen, N. P. E. *J. Med. Chem.* **2005**, *48*, 2308−2318.
(9) Verdonk, M. L.; Chessari, G.; Cole, J. C.; Hartshorn, M. J.; Murray, C. W.; Nissink, J. W. M.; Taylor, R. D.; Taylor, R. *J. Med. Chem.* **2005**, *48*, 6504−6515.
(10) Mancera, R. L. *J. Comput.-Aided Mol. Des.* **2002**, *16*, 479−499.
(11) Garcia-Sosa, A. T.; Firth-Clark, S.; Mancera, R. L. *J. Chem. Inf. Model.* **2005**, *45*, 624−633.
(12) Garcia-Sosa, A. T.; Mancera, R. L. *J. Mol. Model.* **2006**, *12*, 422−431.
(13) Poornima, C. S.; Dean, P. M. *J. Comput.-Aided Mol. Des.* **1995**, *9*, 513−520.
(14) Poornima, C. S.; Dean, P. M. *J. Comput.-Aided Mol. Des.* **1995**, *9*, 521−531.

10.1021/ja066980q CCC: $37.00

can make at least three hydrogen bonds, they have low B factors, and they are normally localized in grooves on the protein surface. In this study all water molecules in protein binding sites were considered, and no distinction was made between water molecules that are conserved in several structures and water molecules that can instead be displaced by some other ligands.

The program Consolv was developed to predict if water molecules in the binding sites of apo-proteins are likely to be conserved or displaced upon ligand binding. A combined *k*-nearest neighbor classifier and genetic algorithm was used to develop the program using four descriptors that characterize the water environment: atomic density, atomic hydrophobicity, hydrogen bonds, and crystallographic B factor.[15] Consolv, though, fails to predict as displaceable those water molecules that can be displaced by a polar group in a ligand.

WaterScore is another program that was developed to predict the conserved and displaceable nature of water molecules in protein apo-structures.[16] Using multivariate logistic regression analysis, it was found that the B factor of a water molecule, the solvent contact surface area, the total hydrogen bond energy, and the number of protein-water contacts, can distinguish between *bound* and *displaceable* water molecules. The main limitation of this program, though, is in the way displaceable water molecules are defined: these are water molecules that are present in the apo-structure and not in the holo-structure of the protein, but they are not sterically displaced by the ligand. Water molecules sterically displaced by ligands were ignored in the development of the knowledge-base of the software, while the identification of these water molecules would be a very useful tool in drug design, as demonstrated by the HIV-1 protease example mentioned above.[4]

More recently, Amadasi et al. reported the use of the HINT force field[17] and of the Rank algorithm,[18] which assesses potential hydrogen bonding, to identify conserved and displaceable water molecules in protein apo-structures.[19] A set of proteins for which pairs of apo- and holo-structures are available was selected, and water molecules in the binding sites of apo-proteins were classified into seven categories, and their average HINT and Rank scores were calculated. It was found that water molecules with moderate Rank and high HINT score are important for interactions with a ligand, while water molecules with low Rank and HINT scores are likely to be sterically displaced by a ligand.

Previous work on water molecules in protein binding sites has therefore been based on predicting whether a water molecule is retained on moving from an apo-structure to a ligand-bound holo-structure. Often, though, the hydration pattern in the binding site of the protein can be very different in the empty pocket and in the pocket with a ligand bound, as indeed can the protein structure itself. OppA is such an example which will be discussed in this work.

Thus, to date, no studies have addressed the issue of identifying water molecules in protein−ligand (holo) structures that may be displaced through rational modification of the ligand (as was the case for HIV-1 protease), and those that may not. Through such a process, synthetic effort may be most profitably devoted to making those molecules targeting the most displaceable waters, and maximizing interactions with the least displaceable. This is the aim of the work reported here.

A dataset of six proteins was selected and for each protein, complexes with five to seven different ligands were considered. Water molecules conserved in all the structures, and water molecules displaced by some of the ligands, were identified by superimposing the structures of the same protein complexed with different ligands using Relibase+.[20] The absolute binding free energy of 54 water molecules was calculated using the double decoupling method,[21] with replica exchange thermodynamic integration (RETI),[22,23] in Monte Carlo simulations. Correlations between the calculated water binding free energies and whether or not the water molecule is observed to be displaced have then been sought. In addition, Bayesian statistics have been applied, such that the probability that a water molecule may be displaced is calculated, given its free energy of binding. The knowledge of the nature of each water molecule can then be used to design new ligands ad hoc, to maximize the interactions with conserved water molecules, and to target those that can be displaced. Finally, correlations between the calculated water binding free energy, and the experimentally observed change in ligand binding affinity that is seen when the ligand is modified and the water molecule displaced, have been sought. This is intended to address the fundamental question of whether it is really worthwhile to target water molecules for displacement through rational design.

2. Methods

2.1. Data-Set Selection. Six proteins, for which specific water molecules are known to be important in mediating the interaction with the ligands, were selected: HIV-1 protease, neuraminidase, trypsin, factor Xa, scytalone dehydratase, OppA. For each protein, all structures with resolution better than 2.5 Å were retrieved using Relibase+;[20] water molecules conserved in all available complexes and water molecules that can be sterically displaced by a ligand were identified. We should emphasize that only proteins for which the ligand binding mode and the protein structure are largely unaffected by water displacement were selected for this study. While there are examples where small changes in the ligand cause large changes in binding mode, these are not considered here.

HIV-1 Protease. It was mentioned in the introduction that potent cyclic urea inhibitors of this protease were designed so that they could displace a water molecule present in complexes with linear peptido mimetic inhibitors.[4] To study this water molecule, labeled as **Wat A**, the reference structure used for the search in Relibase+ was 1hpx. A total of 173 entries were retrieved: four of them belong to apo-structures of the protein while the rest belong to complexes with inhibitors. Twenty-four entries do not have water molecules in their structure. The water molecule in question, that is, Wat 301 in pdb 1hpx, is conserved in all the apo-structures and in 111 of the holo-structures; the water molecule is not present in complexes where the ligand is a cyclic urea derivative, which displaces it. Five structures of the protease

(15) Raymer, M. L.; Shanschagrin, P. C.; Punch, W. F.; Venkataraman, S.; Goodman, E. D.; Kuhn, L. A. *J. Mol. Biol.* **1997**, *265*, 445−464.
(16) Garcia-Sosa, A. T.; Mancera, R. L.; Dean, P. M. *J. Mol. Model.* **2003**, *9*, 172−182.
(17) Kellogg, G. E.; Semus, S. F.; Abraham, D. J. *J. Comput.-Aided Mol. Des.* **1991**, *5*, 545−552.
(18) Chen, D. L.; Kellogg, G. E. *J. Comput.-Aided Mol. Des.* **2005**, *19*, 69−82.
(19) Amadasi, A.; Spyrakis, F.; Cozzini, P.; Abraham, D. J.; Kellogg, G. E.; Mozzarelli, A. *J. Mol. Biol.* **2006**, *358*, 289−309.
(20) Gunther, J.; Bergner, A.; Hendlich, M.; Klebe, G. *J. Mol. Biol.* **2003**, *326*, 621−636.
(21) Gilson, M. K.; Given, J. A.; Bush, B. L.; McCammon, J. A. *Biophys. J.* **1997**, *72*, 1047−1069.
(22) Woods, C. J.; Essex, J. W.; King, M. A. *J. Phys. Chem. B* **2003**, *107*, 13703−13710.
(23) Woods, C. J.; Essex, J. W.; King, M. A. *J. Phys. Chem. B* **2003**, *107*, 13711−13718.

in complex with linear peptido-mimetic inhibitors were selected: 1hpx, 1ec0, 1ec1, 1ebw, 1eby.

Neuraminidase. The displacement of one water molecule by some ligands in neuraminidase has been associated with their higher binding affinity.[24] The complex between subtype N9 neuraminidase and sialic acid (pdb 1mwe) was chosen as reference structure for searching in Relibase+. In this complex two neighboring water molecules bridging between the protein and the ligand can be identified: Wat 327 (**Wat C**) and Wat 328 (**Wat B**). A total of 35 entries were found: 14 belong to apo-structures and the remaining 21 belong to complexes with different inhibitors. Wat 328 is conserved in 5 of the 14 apo-structures, while Wat 327 is conserved in 3 of them. When the holo-structures are considered, Wat 327 is conserved in all of them, while Wat 328 is conserved in 13 and it is displaced by the ligand in the remaining ones. Seven complexes were selected: 1f8b, 1f8c, 1mwe, 2qwj, 2qwk, 1nnc, 1l7f.

Trypsin. The displacement of one water molecule in the binding site of trypsin has been used as a strategy to design inhibitors selective for this serine-protease.[25] To study this water molecule, labeled as **Wat D**, the structure contained in pdb 1az8 was selected as reference for searches in Relibase+. A total of 160 entries were retrieved, of which 41 are apo-structures and 119 are complexes. The water molecule mentioned above is conserved in 38 apo-structures and in 107 holo-structures, while it is displaced by a halogen atom present in the ligand in 6 holo-structures. Seven structures of trypsin in complex with inhibitors were selected: 1az8, 1bty, 1c1q, 1c5t, 1gi1, 1o2j, 1fou.

Factor Xa. In the X-ray structures of this protein complexed with different classes of potent inhibitors, two water mediated interactions can be detected.[26−28] The reference structure chosen for the search in Relibase+ was pdb 1ezq, where both water molecules are present, that is, Wat 100 (**Wat E**) and Wat 115 (**Wat F**). A total of 21 entries were retrieved, of which 2 are apo-structures and 19 are holo-structures. Wat 100 is conserved in 1 apo-structure and in 13 holo-structures while it is displaced by a halogen atom in the ligand in the remaining 6. Wat 115 is not present in the apo-structures and it is conserved in only one other holo-structure while it is displaced by the ligand in nine of the others. Complexes of factor Xa with five inhibitors were selected: 1ezq, 1ksn, 1lpg, 1lpz, 1f0s.

Scytalone Dehydratase. Two water molecules are present in the binding site of this protein: inhibitors were designed to specifically displace one of them, while the other is conserved in all the structures.[29] The reference structure chosen for the search in Relibase+ was pdb 4std, in which both water molecules are present, that is, Wat 54 (**Wat H**) and Wat 64 (**Wat G**). A total of 19 entries were retrieved, 3 of which belong to an apo-structure while the others belong to holo-structures. No water molecules are present in the apo-structure; Wat 64 is conserved in 13 of the holo-structures and it is displaced by the ligand in the other 3; Wat 54, instead, is conserved in all the holo-structures. Complexes of scytalone dehydratase (SD) with five inhibitors were considered: 3std, 4std, 5std, 6std, 7std.

OppA. Tripeptides of the form Lys-X-Lys bind to the protein and, according to the nature of X, the water pattern changes in the binding site. It was found that when X displaces water molecules, the binding affinity of the peptides decreases.[5,30] The reference structure for searches

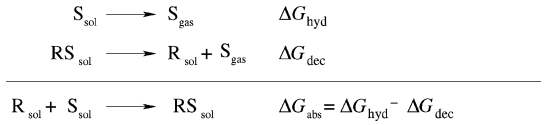

Figure 1. Thermodynamic cycle describing the double decoupling method.

in Relibase+ was the complex between OppA and KDK (pdb 1b4z). Three water molecules mediating the interaction between the protein and the ligand can be identified in this complex: Wat 10 (**Wat J**), Wat 11 (**Wat I**), and Wat 455 (**Wat K**). A total of 33 entries were retrieved, of which only 1 is an apo-structure. The hydration pattern of the binding site in the apo- and holo-structures is very different, and the three water molecules found in 1b4z mentioned above are not present in the apo-structure. When the holo-structures are considered, Wat 11 is conserved in all of them, Wat 10 is conserved in 30 and displaced by the ligand in 2, and Wat 455 is conserved in 18 structures and displaced by the ligand in 9. Six complexes of OppA with different tripeptides were selected: 1b3l, 1b4z, 1jev, 1jeu, 1jet, 2olb.

2.2. Double Decoupling Method. According to the double decoupling method, the absolute binding free energy of a substrate, *S*, to a receptor, *R*, can be calculated by performing two simulations, one in which the substrate is decoupled from the solvent and one in which it is decoupled from the receptor.[21,31] The thermodynamic cycle for this process can be represented as shown in Figure 1.

If the substrate is water, as in the present case, the first simulation is the calculation of the free energy cost of decoupling a water molecule from bulk solvent (ΔG_{hyd}) and the second is the calculation of the free energy for decoupling the water molecule of interest from the protein−ligand complex (ΔG_{dec}). This methodology has often been used to calculate binding free energies of water molecules buried in cavities of proteins,[32−34] and a full description of the theory has been extensively reported elsewhere.[21,34,35]

To guarantee reversibility of the process during the decoupling of the water molecule from the protein−ligand complex, the water molecule needs to be restrained in its position. A flat-bottomed harmonic well potential was used by Helms and Wade in a study of water molecules mediating protein−ligand interactions in cytochrome P450cam;[36] in many other cases, a harmonic restraint was preferred.[32−34,37] In the present study, we decided to constrain the water molecule with a hard-wall potential with the following form:

$$U(r) = \begin{cases} \infty & \text{for } d < r_{HW} \\ \infty & \text{for } d_{wat} > r_{HW} \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

where r_{HW} is the radius of the hard-wall, d is the distance between solvent, protein, and ligand atoms and the center of the hard wall, d_{wat} is the distance between the water molecule being annihilated and the center of the hard-wall. The advantage of using a hard-wall potential instead of a harmonic restraint is that it prevents other solvent molecules and protein and ligand atoms from occupying the position that is left vacant by the water molecule. The arrival of new water molecules in the cavity left by the deletion of a particular water molecule was observed in some cases when a harmonic restraint was used (data not shown). The hard-wall potential was centered on the center of mass of

(24) von Itzstein, M.; et al. *Nature* **1993**, *363*, 418−423.
(25) Mackman, R. L.; Katz, B. A.; Breitenbucher, J. G.; Hui, H. C.; Verner, E.; Luong, C.; Liu, L.; Sprengeler, P. A. *J. Med. Chem.* **2001**, *44*, 3856−3871.
(26) Guertin, K. R.; et al. *Bioorg. Med. Chem. Lett.* **2002**, *12*, 1671−1674.
(27) Matter, H.; Defossa, E.; Heinelt, U.; Blohm, P.; Schneider, D.; Muller, A.; Herok, S.; Schreuder, H.; Liesum, A.; Brachvogel, V.; Lonze, P.; Walser, A.; Al-Obeidi, F.; Wildgoose, P. *J. Med. Chem.* **2002**, *45*, 2749−2769.
(28) Maignan, S.; Guilloteau, J. P.; Choi-Sledeski, Y. M.; Becker, M. R.; Ewing, W. R.; Pauls, H. W.; Spada, A. P.; Mikol, V. *J. Med. Chem.* **2003**, *46*, 685−690.
(29) Chen, J. M.; Xu, S. L.; Wawrzak, Z.; Basarab, G. S.; Jordan, D. B. *Biochemistry* **1998**, *37*, 17735−17744.
(30) Sleigh, S. H.; Seavers, P. R.; Wilkinson, A. J.; Ladbury, J. E.; Tame, J. R. H. *J. Mol. Biol.* **1999**, *291*, 393−415.
(31) Jorgensen, W. L.; Buckner, J. K.; Boudon, S.; Tirado-Rives, J. *J. Chem. Phys.* **1988**, *89*, 3742−3746.
(32) Zhang, L.; Hermans, J. *Proteins* **1996**, *24*, 433−438.
(33) Olano, L. R.; Rick, S. W. *J. Am. Chem. Soc.* **2004**, *126*, 7991−8000.
(34) Hamelberg, D.; McCammon, J. A. *J. Am. Chem. Soc.* **2004**, *126*, 7683−7689.
(35) Boresch, S.; Tettinger, F.; Leitgeb, M.; Karplus, M. *J. Phys. Chem. B* **2003**, *107*, 9535−9551.
(36) Helms, V.; Wade, R. C. *Biophys. J.* **1995**, *69*, 810−824.
(37) Roux, B.; Nina, M.; Pomes, R.; Smith, J. C. *Biophys. J.* **1996**, *71*, 670−681.

the water molecule of interest and a radius of 1.4 Å was chosen, on the basis of an analysis of the O−H radial distribution function of TIP4P water.[38]

It was first demonstrated by Gilson and co-workers[21] that the process for decoupling a substrate from a receptor depends on the standard concentration and that the computed free energy for decoupling needs to be corrected to account for this. The free energy for decoupling a substrate from a receptor can be obtained from the following equation:[21]

$$\Delta G_{\mathrm{dec}} = \Delta G_{\mathrm{comp}} + \Delta G_{\mathrm{rest}} - RT \ln\frac{\sigma_{\mathrm{RS}}}{\sigma_{\mathrm{R}}\sigma_{\mathrm{S}}} + P^0(V_{\mathrm{R}} - V_{\mathrm{RS}}) \quad (2)$$

where ΔG_{comp} is the computed free energy; ΔG_{rest} is the correction term due to constraining the substrate during the simulation; R is the gas constant; T is the temperature; σ_{RS} is the symmetry number of the complex; σ_{R} is the symmetry number of the receptor; σ_{S} is the symmetry number of the substrate; P^0 is the standard pressure; $V_{\mathrm{R}} - V_{\mathrm{RS}}$ represents the volume change of the system when the substrate is decoupled from the receptor in a constant pressure simulation.

The correction term appropriate for a hard-wall potential, with no rotational restraints, is given by[21,40]

$$\Delta G_{\mathrm{rest}} = RT \ln\frac{V_{\mathrm{eff}}}{V^0} \quad (3)$$

where V^0 is the volume at the standard concentration and V_{eff} is the effective volume which, in this case, is the volume on which the hard-wall potential acts. For a standard concentration of 1 mol L^{-1}, $V^0 =$ 1660 $Å^3$;[35,39] the hard-wall potential is a sphere of radius 1.4 Å and, as such, its volume is 11.5 $Å^3$. The correction term for the constraint used in the present study is then −2.9 kcal mol^{-1}. The third term of eq 2 accounts for the symmetry of the substrate; water has a symmetry number of 2, and thus this term in the equation has a value of −0.4.[34,41] The last term in eq 2 can be considered negligible at normal pressures.[21,34,41]

2.3. System Preparation. The same procedure was used for the preparation of all protein−ligand complexes used in this study. The HB2 routine in the HBONDS option of the program WHAT IF[42] was used to add hydrogen atoms onto oxygens of water molecules and onto heavy atoms of proteins. All residues were considered in their protonation state at pH 7. For HIV-1 protease, two aspartate residues in the binding site, Asp 25 and Asp 25′, can have different protonation states according to the ligand bound to the protein.[43] NMR studies and calculations both suggest that symmetric and neutral cyclic urea derivatives that have a central diol moiety bind to the protein where both the aspartate residues are protonated.[43,44] Four of the five selected complexes (1ec1, 1ec0, 1ebw, 1eby) contain C2-symmetric ligands, with a central diol moiety, which is present in the cyclic ureas as well. Even though no data is available for the protonation state of the protein in these complexes, analysis of the crystal structures suggests that the mechanism of binding of the diol group is similar to that of cyclic ureas and, as such, it is quite likely that both aspartates will be protonated. In the remaining complex (1hpx), the inhibitor is KNI272, which is neutral and asymmetric, and it is known that, in its complex with the protease, Asp 25 is protonated and Asp 25′ is deprotonated.[43]

For histidine residues the protonation state was assigned based on literature data, or, if no data were available, it was based on the WHAT IF results. Hydrogen atoms on inhibitors were added using BABEL.[45]

The proteins were modeled using the AMBER 99 force field,[46] while inhibitors were modeled using the generalized AMBER force field (GAFF).[47] Partial charges on the ligands were calculated with the AM1-BCC method, using the program Antechamber in the AMBER 7 package.[48,49] AM1-BCC was preferred over HF/6-31G* RESP for reasons of computational expediency. Indeed, AM1-BCC charges were parametrized to reproduce HF/6-31G* RESP charges.[49] Water molecules were modeled using the TIP4P potential.[38] Electrostatic neutrality of the systems was ensured by adding the correct number of ions using the LEaP program in the AMBER 7 package. Each system, including crystallographic water molecules, was solvated in an orthorhombic box of water which was constructed in such a way that the minimum distance between the protein and the edge of the box was 10 Å. Details for each system have been provided in the Supporting Information.

2.4. Simulation Procedure for Water Decoupling. Calculations were performed by gradually switching off first electrostatic and then Lennard-Jones interactions between the water molecule of interest and the rest of the system, in the isothermal−isobaric ensemble at $T =$ 298 K and $P = 1$ atm, with our group's MC program, ProtoMC.[50] The free energy method used was replica exchange thermodynamic integration (RETI).[22,23] Periodic boundary conditions were applied, with a nonbonded cutoff of 10 Å. MC simulations were preferred over molecular dynamics (MD) simulations for a number of reasons, including the ease of applying internal coordinate constraints, our observation that MC yields well-behaved free energy gradients, and that the hard-wall potential to constrain the annihilating water molecules is trivial to implement in MC.

Free Energy of Water Decoupling from Bulk Solvent. A water molecule in the center of a box containing 597 water molecules, with dimensions 26.99 × 26.74 × 26.55 $Å^3$, was decoupled from the system in two steps: in a first simulation the electrostatic interactions between the water molecule and its surroundings were switched off and in a second simulation the Lennard-Jones interactions were switched off.

The system was initially equilibrated for 10 million (M) MC steps. Each of the two successive simulations was then broken up into 21 evenly separated λ windows, with a value of $\Delta\lambda$ of 0.001, and calculations for all λ values were run in parallel. At the beginning of the simulations, an additional equilibration with 500 thousand (K) MC moves was performed for each replica at each λ value.

The annihilation of both the electrostatic and Lennard-Jones interactions was performed in 10 M MC steps divided into 200 blocks of 50 K steps. A λ swap move between neighboring λ values was attempted at the end of each block. Data were collected over the last 7 M steps for both the electrostatic and Lennard-Jones decoupling.

Free Energy of Water Decoupling from Protein−Ligand Complexes. Each system was initially equilibrated with the following procedure: 10 M MC steps of solvent moves only, 1 M MC steps of protein moves only, 6 M MC steps of general equilibration.

To increase flexibility, the backbone of the proteins was allowed to move by rotation/translation around C_α, as well as the side chains. For the ligand and the protein, bonds were constrained, whereas full sampling of the angles and dihedrals was allowed, with the exception of ring systems which were constrained. Cysteine residues involved in disulfide bonds were kept fixed.

(38) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *J. Chem. Phys.* **1983**, *79*, 926−935.
(39) Hermans, J.; Wang, L. *J. Am. Chem. Soc.* **1997**, *119*, 2707−2714.
(40) Mann, G.; Hermans, J. *J. Mol. Biol.* **2000**, *302*, 979−989.
(41) Lu, Y.; Yang, C.-Y.; Wang, S. *J. Am. Chem. Soc.* **2006**, *128*, 11830−11839.
(42) Vriend, G. *J. Mol. Graphics.* **1990**, *8*, 52−56.
(43) Trylska, J.; Antosiewicz, J.; Geller, M.; Hodge, C. N.; Klabe, R. M.; Head, M. S.; Gilson, M. K. *Protein Sci.* **1999**, *8*, 180−195.
(44) Yamazaki, T.; Nicholson, L. K.; Torchia, D. A.; Wingfield, P.; Stahl, S. J.; Kaufman, J. D.; Eyerman, C. J.; Hodge, C. N.; Lam, P. Y. S.; Ru, Y.; Jadhav, P. K.; Chang, C.; Weber, P. C. *J. Am. Chem. Soc.* **1994**, *116*, 10791−10792.
(45) Walters, M.; Stahl, M. BABEL, version 1.6. http://smog.com/chem/babel.
(46) Wang, J.; Cieplak, P.; Kollman, P. A. *J. Comput. Chem.* **2000**, *21*, 1049−1074.
(47) Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. *J. Comput. Chem.* **2004**, *25*, 1157−1174.
(48) Jakalian, A.; Bush, B. L.; Jack, D. B.; Bayly, C. I. *J. Comput. Chem.* **2000**, *21*, 132−146.
(49) Jakalian, A.; David, B. J.; Bayly, C. I. *J. Comput. Chem.* **2002**, *23*, 1623−1641.
(50) Woods, C. J. *ProtoMC 1.0;* University of Southampton: Southampton, U.K., 2003.

Table 1. Binding Free Energy for Water Molecules Buried in SC-EC

water[a]	ΔG_{abs}[b]	ΔG_{abs}^{lit}[c]
412	-2.2 ± 0.5	-0.4 ± 2
417	-6.3 ± 0.5	-6.3 ± 2
418	-1.0 ± 0.5	-2.6 ± 2
804	$+0.4 \pm 0.5$	$+0.1 \pm 2$

[a] Water id in pdb 1cse. [b] Calculated binding free energy of water molecule. [c] This value was obtained by correcting the computed decoupling free energy of the water molecule reported in literature[32] with the correction term discussed in the text. All values are in kcal mol^{-1}.

Each equilibrated structure was taken as starting structure for the calculations. Simulations were performed using RETI.[22,23] The free energy cost of decoupling the water molecule of interest from the protein was calculated in two steps: first the electrostatic interactions between the water molecule and the environment were switched off and then the Lennard-Jones interactions were switched off. The water molecule was constrained in its position with a hard-wall potential of radius 1.4 Å centered on the center-of-mass of the water molecule in the equilibrated structure. For each simulation, 21 evenly spaced λ windows were used and, at the beginning of the simulations, an additional equilibration with 500 K MC moves was performed for each replica at each λ value.

The annihilation of the electrostatic interactions was usually performed in 15 M MC steps divided into 300 blocks of 50 K steps each; the annihilation of the Lennard-Jones interactions was performed in 10 M MC steps divided into 200 blocks of 50 K steps each. A λ swap move between neighboring pairs was attempted at the end of each block. Data were collected and averaged over the last 10 M steps for the electrostatic decoupling and over the last 7 M steps for the Lennard-Jones decoupling. For the neuraminidase complexes, it was found that longer runs were needed to converge the calculated free energies, and so the electrostatic decoupling was performed in 20 M MC steps and data were collected and averaged over the last 15 M MC steps.

2.5. Method Validation. Owing to the double nature of the water molecule as ligand and solvent, the binding free energy of a specific water molecule in a protein cannot be measured experimentally. For this reason, to validate our methodology, two test systems for which literature simulation data is available, were selected: the Subtilisin Carlsberg-Eglin C (SC-EC) complex and BPTI.

The binding free energy of Wat 122 in BPTI (pdb 5pti) was calculated to be -4.1 ± 0.5 kcal mol^{-1}, which is in good agreement with the value reported by Olano et al.[33] at 298 K, that is, -4.7 ± 1 kcal mol^{-1}.

The binding free energies of four water molecules buried in SC-EC (pdb 1cse) were calculated and their values are compared to those reported by Zhang and Hermans[32] in Table 1. It should be noted that in their paper, Zhang and Hermans report a value of -2.3 kcal mol^{-1} for the correction term to apply to the computed free energy of water decoupling from the protein, ΔG_{comp}, but they do not give any detail on its calculation. On the basis of successive literature data,[21,34,41] we believe that this value is wrong. If we consider eq 2, the second term, ΔG_{rest}, has a value of -4.7 kcal mol^{-1} for the use of a harmonic restraint with force constant K_{harm} equal to 5 kcal mol^{-1} Å^{-2} (which was used in their paper) and the third term has a value of -0.4.[34,41] The final correction to the computed binding free energies should be thus -4.3 kcal mol^{-1}. There is good agreement between our calculated results and published results derived using related methods.

2.6. Error Analysis in Simulations. To estimate the error on the free energy calculations, standard errors over block averages of 50 K steps were calculated for each value of λ, with the errors being propagated across λ by calculating the maximum and minimum possible free energies.

3. Results

3.1. Free Energy of Water Decoupling from Bulk Solvent. The calculated free energy cost for the elimination of the electrostatic interactions is $+8.5 \pm 0.3$ kcal mol^{-1}, while the free energy for the elimination of the Lennard-Jones interactions is -2.0 ± 0.3 kcal mol^{-1}, yielding a total free energy of decoupling a water molecule from bulk solvent of $+6.5 \pm 0.4$ kcal mol^{-1}. This value is in good agreement with the experimentally determined value of $+6.3$ kcal mol^{-1},[51] and with the excess chemical potential of water of $+6.1$ kcal mol^{-1} calculated for TIP4P water.[38]

3.2. Binding Free Energy of Water Molecules in Protein−Ligand Complexes. The absolute binding free energies of 54 water molecules were calculated. Results are reported in Table 2 and they will now be discussed: only free energies calculated using the forward finite difference gradients are given because the difference with the reverse gradients is negligible. Figures showing the free energy gradients as a function of the coupling parameter for one example of the annihilation of a displaceable water molecule, and one example for a conserved water molecule, are provided in the Supporting Information.

HIV-1 Protease. Wat A is very tightly bound in all five complexes studied and the binding is mainly due to favorable electrostatic interactions. This water molecule is located in a tight cavity, the environment is highly polar, and the water molecule is tetrahedrally coordinated, as it can donate hydrogen bonds to two carbonyl groups in the ligand and it can accept two hydrogen bonds from Ile 50 and Ile 50′ in the protein, as shown in Figure 2.

The binding free energy of Wat 301 in HIV-1 protease complexed with KNI272 was calculated to be -3.3 kcal mol^{-1} by Hamelberg and McCammon[34] in the protein with a doubly deprotonated aspartic acid dyad, which is not in agreement with our result of -10.0 ± 0.5 kcal mol^{-1}. More recently, Lu et al. have repeated this calculation using two different protonation states for the aspartic acid dyad, and obtained essentially the same answer as Hamelberg and McCammon.[41] We do not have a definitive explanation for this discrepancy, although there are a number of possible reasons. First, the electrostatic parameters for the ligands were calculated using different levels of theory, AM1-BCC in this study versus RESP and HF/6-31G* in the others. Second, we have adopted MC simulations to sample the configurations of the system, whereas the other studies use molecular dynamics. The use of MC will restrict the ability of the protein to sample cooperative large-scale motions. However, this is not an issue here since we explicitly do not want the protein to reorganize on annihilation of the water molecule, that is, we want the water cavity to be retained. Third, we have adopted a hard-wall constraint when annihilating the water molecules, since in doing so we prevent system collapse, or the arrival of another water molecule, into the cavity we create. This decision means that the free energies we calculate reflect the binding free energy of a water molecule in a pre-existing cavity, whereas the results obtained using the harmonic restraint may also reflect the free energy cost associated with cavity formation, that is, the two free energies are not directly comparable. In cases where the cavity is solvent exposed, or the water binding pocket is amenable to collapse, the free

(51) Ben-Naim, A.; Marcus, Y. *J. Chem. Phys.* **1984**, *81*, 2016−2027.

Table 2. Free Energy of Binding of Water Molecules in Protein−Ligand Complexes

protein[a]	water[b]	ΔG_{el}[c]	ΔG_{lj}[d]	ΔG_{dec}[e]	ΔG_{abs}[f]	protein	water	ΔG_{el}	ΔG_{lj}	ΔG_{dec}	ΔG_{abs}
HIV−1 protease	**Wat A**					**FXa**	**Wat E**				
1hpx	301	+19.1 ± 0.1	−0.1 ± 0.1	+16.5 ± 0.2	−10.0 ± 0.5	1f0s	68	+8.7 ± 0.1	−1.4 ± 0.2	+4.8 ± 0.2	+1.7 ± 0.5
1ec1	614	+19.4 ± 0.2	−0.7 ± 0.2	+16.2 ± 0.3	−9.7 ± 0.5	**FXa**	**Wat F**				
1ec0	627	+18.5 ± 0.2	−1.3 ± 0.2	+14.7 ± 0.3	−8.2 ± 0.5	1ezq	115	+9.6 ± 0.2	+1.8 ± 0.1	+8.9 ± 0.2	−2.4 ± 0.5
1ebw	319	+17.9 ± 0.2	−1.2 ± 0.2	+14.2 ± 0.3	−7.7 ± 0.5	1ksn	133	+9.3 ± 0.2	+0.7 ± 0.2	+7.5 ± 0.2	−1.0 ± 0.5
1eby	316	+18.4 ± 0.2	−2.3 ± 0.2	+13.6 ± 0.3	−7.1 ± 0.5	**SD**	**Wat G**				
neuraminidase	**Wat B**					4std	64	+15.6 ± 0.2	−3.5 ± 0.3	+9.6 ± 0.3	−3.1 ± 0.5
1f8b	1401	+7.3 ± 0.1	+0.8 ± 0.1	+5.6 ± 0.2	+0.9 ± 0.5	5std	537	+14.2 ± 0.2	+0.5 ± 0.2	+12.2 ± 0.2	−5.7 ± 0.5
1f8c	1001	+8.1 ± 0.1	−0.8 ± 0.2	+4.8 ± 0.2	+1.7 ± 0.5	6std	57	+14.1 ± 0.2	+0.4 ± 0.2	+12.0 ± 0.2	−5.5 ± 0.5
1mwe	328	+7.9 ± 0.1	+0.4 ± 0.2	+5.8 ± 0.2	+0.7 ± 0.5	7std	91	+15.4 ± 0.2	−1.4 ± 0.2	+11.5 ± 0.3	−5.0 ± 0.5
2qwj	8R	+9.7 ± 0.1	−1.5 ± 0.2	+5.7 ± 0.2	+0.8 ± 0.5	**SD**	**Wat H**				
2qwk	2S	+9.8 ± 0.2	−3.1 ± 0.2	+4.2 ± 0.3	+2.3 ± 0.5	3std	36	+13.5 ± 0.2	−2.0 ± 0.2	+9.0 ± 0.2	−2.5 ± 0.5
neuraminidase	**Wat C**					4std	54	+15.0 ± 0.2	−0.6 ± 0.2	+11.9 ± 0.2	−5.4 ± 0.5
1f8b	2221	+19.1 ± 0.2	−0.6 ± 0.2	+16.0 ± 0.3	−9.5 ± 0.5	5std	538	+16.2 ± 0.2	−1.5 ± 0.2	+12.2 ± 0.3	−5.7 ± 0.5
1f8c	1021	+19.3 ± 0.2	−1.4 ± 0.2	+15.4 ± 0.3	−8.9 ± 0.5	6std	3	+13.0 ± 0.2	−1.9 ± 0.2	+8.6 ± 0.3	−2.1 ± 0.5
1mwe	327	+17.5 ± 0.3	+0.2 ± 0.2	+15.2 ± 0.3	−8.7 ± 0.5	7std	20	+15.0 ± 0.2	−2.0 ± 0.2	+10.5 ± 0.3	−4.0 ± 0.5
2qwj	1R	+19.1 ± 0.2	+0.3 ± 0.2	+16.9 ± 0.3	−10.4 ± 0.5	**OppA**	**Wat I**				
2qwk	1S	+21.5 ± 0.2	−4.9 ± 0.3	+14.1 ± 0.3	−7.6 ± 0.5	1b4z	11	+11.0 ± 0.2	+2.0 ± 0.1	+10.5 ± 0.2	−4.0 ± 0.5
1nnc	121	+20.5 ± 0.2	−0.3 ± 0.2	+17.7 ± 0.3	−11.2 ± 0.5	1jet	60	+10.5 ± 0.2	+2.4 ± 0.1	+10.4 ± 0.2	−3.9 ± 0.5
1l7f	35	+29.7 ± 0.2	−11.6 ± 0.3	+15.6 ± 0.4	−9.1 ± 0.6	1jeu	1	+12.2 ± 0.2	+1.5 ± 0.1	+11.2 ± 0.2	−4.7 ± 0.5
trypsin	**Wat D**					1b3l	115	+14.7 ± 0.2	+0.4 ± 0.2	+12.6 ± 0.3	−6.1 ± 0.5
1az8	638	+10.9 ± 0.1	+0.6 ± 0.2	+9.0 ± 0.2	−2.5 ± 0.5	2olb	25	+18.1 ± 0.2	+1.3 ± 0.2	+16.9 ± 0.3	−10.4 ± 0.5
1bty	268	+9.0 ± 0.1	+1.5 ± 0.1	+8.0 ± 0.2	−1.5 ± 0.5	1jev	14	+6.3 ± 0.2	+0.1 ± 0.2	+3.9 ± 0.3	+2.6 ± 0.5
1c5t	325	+9.8 ± 0.1	+0.3 ± 0.1	+7.6 ± 0.2	−1.1 ± 0.5	**OppA**	**Wat J**				
1c1q	325	+11.3 ± 0.1	−0.4 ± 0.1	+8.4 ± 0.2	−1.9 ± 0.5	1b4z	10	+15.9 ± 0.2	−1.5 ± 0.2	+11.9 ± 0.3	−5.4 ± 0.5
1gi1	268	+11.3 ± 0.1	+1.3 ± 0.1	+10.1 ± 0.2	−3.6 ± 0.5	1jet	85	+15.0 ± 0.2	+0.3 ± 0.2	+12.8 ± 0.3	−6.3 ± 0.5
1f0u	6	+9.4 ± 0.1	+1.4 ± 0.1	+8.3 ± 0.2	−1.8 ± 0.5	1jeu	55	+15.6 ± 0.2	−0.2 ± 0.2	+12.9 ± 0.3	−6.4 ± 0.5
1o2j	705	+9.8 ± 0.1	+0.8 ± 0.2	+8.1 ± 0.2	−1.6 ± 0.5	1b3l	72	+19.0 ± 0.2	−1.7 ± 0.2	+14.8 ± 0.3	−8.3 ± 0.5
FXa	**Wat E**					2olb	16	+16.4 ± 0.2	−2.1 ± 0.2	+11.8 ± 0.3	−5.3 ± 0.5
1ezq	100	+11.7 ± 0.1	+1.7 ± 0.1	+10.9 ± 0.2	−4.4 ± 0.5	**OppA**	**Wat K**				
1ksn	16	+11.4 ± 0.1	+1.3 ± 0.1	+10.2 ± 0.2	−3.7 ± 0.5	1b4z	455	+12.9 ± 0.2	+0.3 ± 0.2	+10.7 ± 0.2	−4.2 ± 0.5
1lpg	215	+9.9 ± 0.1	+1.4 ± 0.1	+8.8 ± 0.2	−2.3 ± 0.5	1jet	557	+15.7 ± 0.2	−1.4 ± 0.2	+11.8 ± 0.3	−5.3 ± 0.5
1lpz	200	+11.7 ± 0.1	+0.6 ± 0.1	+9.8 ± 0.2	−3.3 ± 0.5	1b3l	45	+19.2 ± 0.2	−2.3 ± 0.3	+14.4 ± 0.4	−7.9 ± 0.6

[a] For each protein, the pdb code of the complexes used in the study is given. [b] ID number of the water molecule in the pdb file. [c] Free energy for decoupling of electrostatic interactions. [d] Free energy for decoupling of Lennard-Jones interactions [e] Total free energy for water decoupling in protein−ligand complexes. [f] Absolute binding free energy of the water molecule. All ΔG values are in kcal mol^{-1}.

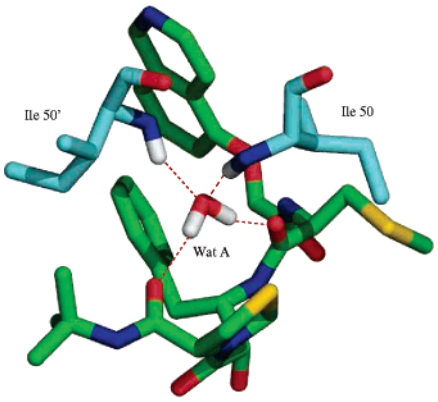

Figure 2. **Wat A** at the interface between HIV-1 protease and ligand KNI272 (Wat 301 in pdb 1hpx). The ligand is colored in green, while protein residues are colored in blue; the dashed red lines represent hydrogen bonds. All figures have been generated with PyMOL.[52]

energies calculated using the two approaches may differ. For an isolated pocket whose integrity is maintained on water annihilation, the two approaches should yield identical results to within error. Our decision to adopt the hard-wall constraint was driven partly by pragmatic concerns, namely, the observation of other water molecules refilling cavities during the annihilation process with a harmonic restraint and the view that the free energy calculated using the hard-wall constraint is the free energy in which we are more interested. Our objective is to calculate the binding free energies of water molecules that may or may not be displaced by rational modification of the ligand. If ligand modification displaces a water molecule, then ligand atoms occupy the volume left by the water molecule, that is, the ligand fills the vacant cavity. Separating the free energy of cavity formation from the calculated binding free energy of each water molecule therefore gives a clearer assessment of the water molecule's intrinsic binding free energy and its relationship with ligand modification.

Neuraminidase. Wat B has an unfavorable binding free energy in all the systems under study and both electrostatic and Lennard-Jones interactions between the water molecule and the environment are quite weak. Analysis of the X-ray structures of the complexes highlights that the site hosting this water molecule is actually partially apolar as the methyl group of the acetamide in the ligand, the alkyl chain of Glu 119 and the side chain of Leu 134 are within 4 Å of the water molecule, as can be seen in Figure 3. This water molecule can only accept one hydrogen bond from the ligand, and it can donate one to the carbonyl of Trp 178.

The problem of the contrast between the experimental evidence of water molecules located in apolar cavities and calculated positive binding free energies has already been reported in the literature. Zhang and Hermans found positive binding free energies for crystallographic water molecules located in apolar cavities of the Subtilisin Carlsberg-Eglin C complex and stated that those water molecules are just an artifact of the refinement process and should not be there.[32] Olano and co-workers[33] recently calculated the binding free energy of a crystallographic water molecule located in a partially apolar cavity in a mutant of barnase and found it to be positive as well; in this case the authors suggested that the positive value

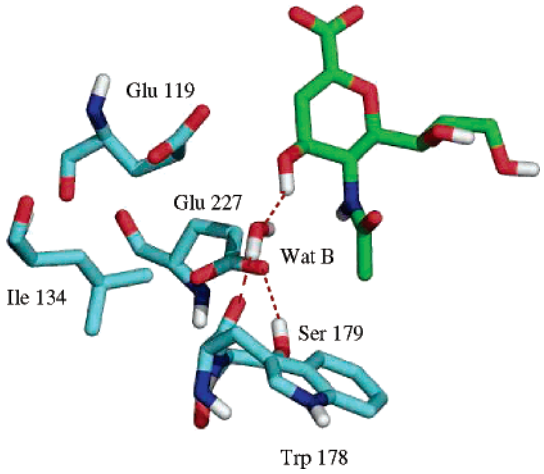

Figure 3. **Wat B** mediating the interaction between neuraminidase and inhibitor DANA (Wat 1401 in pdb 1f8b). The ligand is colored in green, while protein residues are colored in blue; the dashed red lines represent hydrogen bonds.

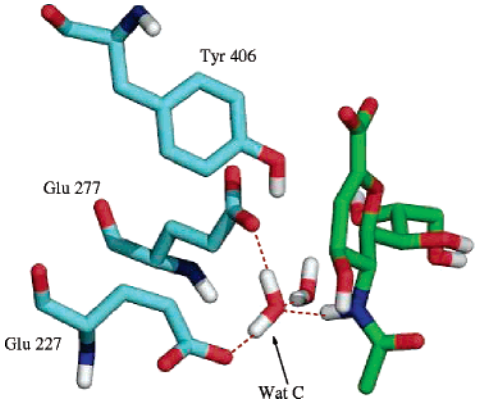

Figure 4. **Wat C** mediating the interaction between neuraminidase and inhibitor DANA (Wat 2221 in pdb 1f8b). The ligand is colored in green, while protein residues are colored in blue; the dashed red lines represent hydrogen bonds.

could be due to a deficiency of nonpolarizable force-fields normally used in the calculations, as the interactions between water molecules and nonpolar groups are normally underestimated by these force fields.

In the case of neuraminidase, **Wat B** is observed in many crystal structures and this makes it difficult to accept Zhang and Hermans's explanation. The contrast between experimental evidence of the water molecule and the calculated positive binding free energy is highly likely to be due to a deficiency of the force field, as postulated by Olano and co-workers.[33]

Wat C is conserved in all the structures and it is tightly bound in all of them. The electrostatic interactions between this water molecule and the environment are very strong. Analysis of the X-ray structures of the complexes shows that this water molecule is tetrahedrally coordinated, as it can accept one hydrogen bond from the amide NH of the ligands and from a neighboring water molecule, while it can donate two hydrogen bonds to two glutamate residues in the active site, Glu 227 and Glu 277, as shown in Figure 4. Moreover, this water molecule acts as a bridge between the two negatively charged glutamate residues and, as such, it has a very important role in stabilizing the structure.

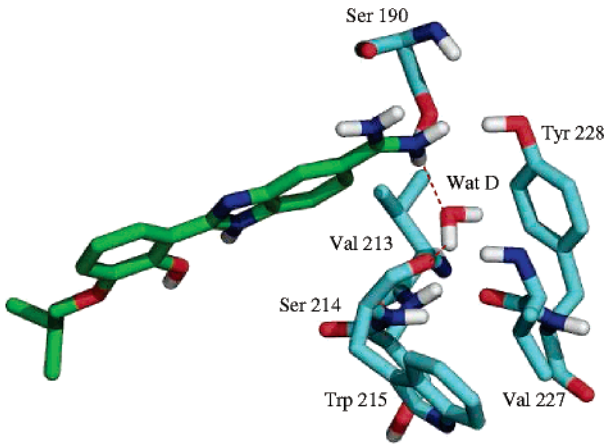

Figure 5. **Wat D** mediating the interaction between trypsin and one inhibitor (Wat 705 in pdb 1o2j). The ligand is colored in green, while protein residues are colored in blue; the dashed red lines represent hydrogen bonds.

Trypsin. The binding free energy of **Wat D** is favorable in all seven trypsin-inhibitor complexes considered. The water molecule has similar values of binding free energies in all the complexes, which is due to the similarity of the water environment in all the complexes. In fact, all ligands share the same aromatic amidino moiety which interacts with the water molecule.

Even though the binding free energy of the water molecule is favorable, the water molecule is not particularly tightly bound in the complexes and both electrostatic and Lennard-Jones interactions are not very strong. Analysis of the X-ray structures shows that the environment is partially apolar with the side-chain of Val 213 and the aromatic ring of Tyr 228 close to the water molecule, as shown in Figure 5.

Hamelberg and McCammon reported a value of -1.9 ± 0.5 kcal mol^{-1} for the binding free energy of this water molecule in trypsin complexed with benzamidine (pdb 1ane).[34] We calculated the binding free energy of this water molecule in the same complex using another structure (pdb 1bty): the binding site is conserved in the two structures and our result of -1.5 ± 0.5 kcal mol^{-1} is consistent with that previously published.

Factor Xa. The binding of **Wat E** is favorable in all the FXa complexes with the exception of 1f0s. The inhibitor present in this complex is different from the other inhibitors. In the structures where the binding free energy is favorable, **Wat E** directly interacts with the amidino group in the ligands and the carbonyl group of Ile 227 and the hydroxyl group of Tyr 228, as shown in Figure 6. In all complexes electrostatic and Lennard-Jones interactions are not very strong, because of the partial apolarity of the environment, which is due to the vicinity of Ala 190, Val 213, and the aromatic ring of Tyr 228, but overall the binding is favorable. In 1f0s, **Wat E** is not directly bridging between the ligand and the protein: in this case there are two water molecules between the ligand and the protein. **Wat E** can make two hydrogen bonds with the protein and one with the other water molecule, not with the ligand as in the other complexes. In this case electrostatic interactions are weaker, and the Lennard-Jones interactions are unfavorable. This water molecule is located in a more open cavity than the same water molecule in the other structures, and it also has a higher B factor, which means that it is less localized.

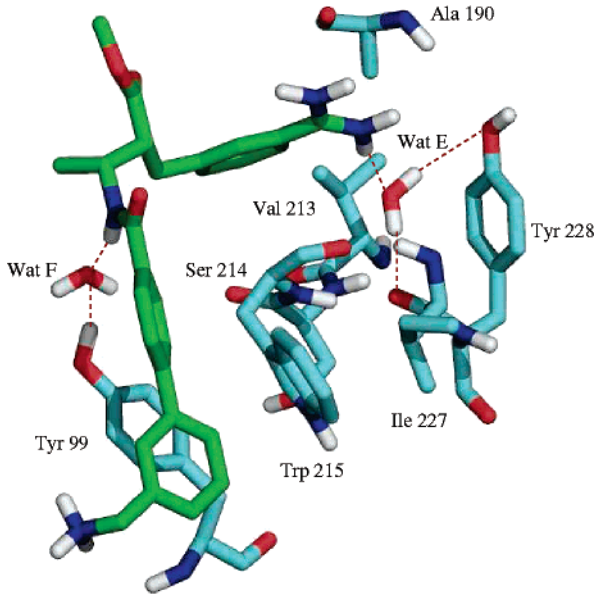

Figure 6. **Wat E** and **Wat F** mediating the interaction between factor Xa and an inhibitor (Wat 100 and Wat 115 in pdb 1ezq). The ligand is colored in green, while protein residues are colored in blue; the dashed red lines represent hydrogen bonds.

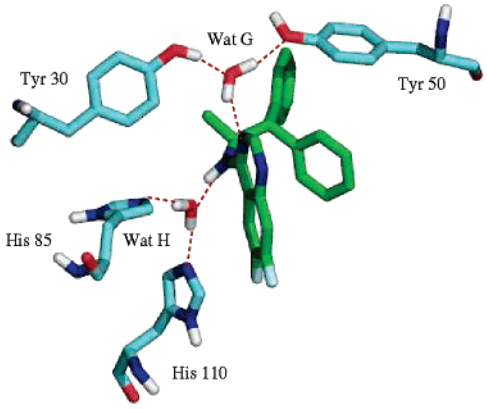

Figure 7. **Wat G** and **Wat H** mediating the interaction between scytalone dehydratase and one inhibitor (Wat 537 and Wat 538 in pdb 5std). The ligand is colored in green, while protein residues are colored in blue; the dashed red lines represent hydrogen bonds.

Wat F is favorably bound in both complexes, but both electrostatic and Lennard-Jones interactions are not very strong. This water molecule is located in an open, solvent exposed cavity and it can only form two hydrogen bonds with the NH in the ligand and with Tyr 99 in the protein, as shown in Figure 6.

Scytalone Dehydratase. Wat G has a favorable binding free energy in all complexes. This water molecule can make one hydrogen bond with the ligand and two hydrogen bonds with Tyr 30 and Tyr 50 in the protein, and that is why the electrostatic interactions are quite strong. **Wat H**, which is conserved in all the structures, also has favorable binding in all complexes. This water molecule accepts one hydrogen bond from the ligand and donates two hydrogen bonds to His 85 and His 110 in the protein. The two water-mediated interactions between scytalone and one inhibitor are shown in Figure 7.

OppA. Wat I is favorably bound in all the complexes with the exception of 1jev. In this complex the central residue in the ligand is a tryptophan: the introduction of this bulky, hydrophobic group disrupts the water network in the binding site, and **Wat I** is isolated from other solvent molecules in this structure. The environment for this water molecule is not favorable, as demonstrated by the very weak electrostatic interactions. **Wat I** is favorably bound in the remaining five complexes but the binding free energy ranges from −3.9 to −10.4 kcal mol^{-1}. Analysis of the structures of all ligands in complex with OppA shows that this is likely to be due to the interaction with one particular residue in the protein, that is, Arg 404. In 2olb Arg 404 can strongly interact with the water molecule, as there are no other possible partners in its vicinity. In 1b4z the central residue in the ligand is replaced by a negatively charged aspartate; Arg 404 in this case strongly interacts with this residue, and this causes a decrease in electrostatic interactions for the water molecule. The same is true for 1jeu, where the central residue in the ligand is a glutamate. Finally, in 1b3l and 1jet the central residue of the ligand is, respectively, a glycine and an alanine, which means that the cavity previously filled by the central residue is now replaced by water molecules. The arginine residue in this case is much more free to move compared to the other structures, and in fact during the simulation the side chain of the arginine moves toward Glu 32 and the distance between the two is reduced from more than 6.0 to 3.9 Å. In 1jet the electrostatic interactions are weaker than in 1b3l due to the presence of the methyl group of the alanine as central residue.

Wat J is favorably bound in all the complexes. Also for this water molecule the electrostatic interactions are favorable, and this is due to the polarity of the environment and to the fact that it can form several hydrogen bonds with the inhibitors, the protein and other water molecules filling the cavity.

Wat K is favorably bound in the three complexes in which it is conserved, but the free energy is more negative in 1b3l than in the other two and this is due to stronger electrostatic interactions. In this case the water molecule can interact with Thr 438 and Asn 436, and it is part of a water network. In 1b4z, the water network is disrupted by the presence of the central aspartate residue: this residue mainly interacts with Arg 404 and with another water molecule that bridges between this residue and Glu 32. In 1jet, the central group is instead alanine; the presence of the apolar methyl group disrupts the water network present in 1b3l and makes the electrostatic interactions for this water molecule weaker.

Figure 8 shows **Wat I**, **Wat J**, and **Wat K** at the interface between OppA and ligand KDK.

4. Discussion

From the results discussed above, it appears that the binding free energy of each water molecule is dependent on the nature of the environment in which it is located. Water molecules which are strongly bound are generally found in polar cavities, and they can form at least three hydrogen bonds with both the protein and the ligand; water molecules which are loosely bound, instead, can generally form less than three hydrogen bonds, and they are usually located in partially apolar environments.

The main aim of this work was to investigate the possibility of distinguishing water molecules that are always conserved in protein−ligand complexes from water molecules that can instead be displaced by some ligands. A total of 54 water molecules was studied, of which 18 can be classified as conserved (**Wat**

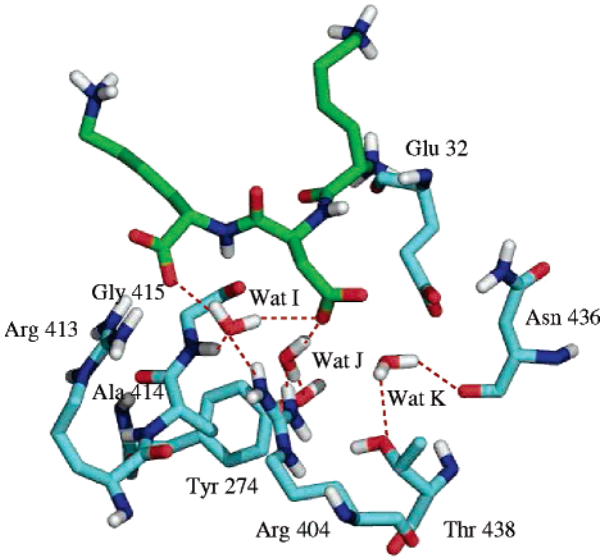

Figure 8. Water mediated interactions between OppA and KDK (Wat 10, Wat 11, and Wat 455 in pdb 1b4z). The ligand is colored in green, while protein residues are colored in blue; the dashed red lines represent hydrogen bonds.

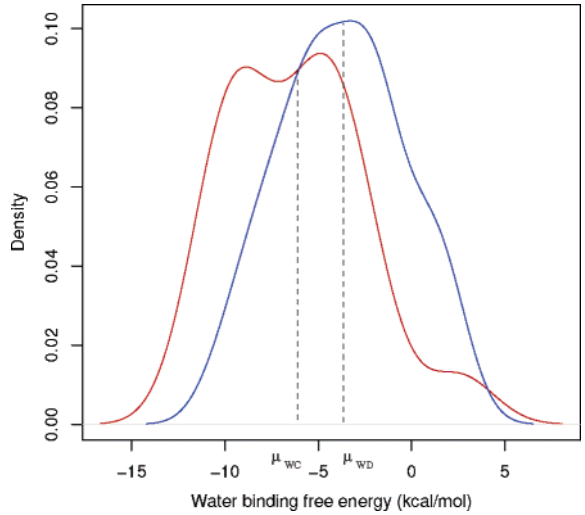

Figure 9. Density distribution of conserved water molecules (red line) and water molecules displaced by ligands (blue line). μ_{WC} is the mean free energy of conserved water molecules; μ_{WD} is the mean free energy of displaceable water molecules. The figure was obtained using the *density* function in the R program.

C, **Wat H**, and **Wat I**) and 36 can be classified as displaceable (**Wat A**, **Wat B**, **Wat D**, **Wat E**, **Wat F**, **Wat G**, **Wat J**, **Wat K**). Figures 2−8 show the locations of these water molecules and their interactions with proteins and ligands in the different protein−ligand complexes. The average binding free energy of conserved water molecules is −6.2 kcal mol^{-1} and that of water molecules displaced by ligands is −3.7 kcal mol^{-1}. A *t* test assuming equal variances was performed in *R*[53] to see at what level of confidence the observed difference of mean free energies among the two classes of water molecules is statistically significant. As can be seen from Table 3, the difference is statistically significant at the 95% confidence interval (α = 0.05).

Table 3. *t*-Test Results

	conserved vs displaced water molecules
t_{calcd}	2.511
$T_{critical}$	2.007
α	0.050
$P(T \leq t)$	0.015

The density distributions of conserved water molecules and water molecules that can be displaced by a ligand are shown in Figure 9. For conserved water molecules the mean free energy, μ, is −6.2 kcal mol^{-1} and the standard deviation, σ, is 3.5 kcal mol^{-1}, while for displaceable water molecules $\mu = -3.7$ kcal mol^{-1} and $\sigma = 3.3$ kcal mol^{-1}.

If we have a new water molecule, not belonging to the dataset, with binding free energy *x*, then the probability that the water molecule belongs to the class of conserved water molecules can be calculated using Bayes' formula as follows:[54]

$$P(C|x) = \frac{p(x|C)\,P(C)}{p(x)} \tag{4}$$

where $P(C|x)$ is the probability of the water molecule being conserved; $P(C)$ is the a priori probability, given by the ratio between the number of conserved water molecules in the training set (18) and the total number of observations in the training set (54); $p(x)$ is the marginal density of the binding free energy; $p(x|C)$ is the conditional density function of the binding free energy for conserved water molecules, which is taken to be the Gaussian density as follows:

$$p(x|C) = \frac{1}{\sigma\sqrt{2\pi}} \exp(-((x-\mu)^2/2\sigma^2)) \tag{5}$$

where μ is the mean free energy of conserved water molecules, σ is the standard deviation, and *x* is the free energy of the new water molecule.

Similarly, the probability of the new water molecule belonging to the class of water molecules that can be displaced by a ligand can be calculated as

$$P(D|x) = \frac{p(x|D)\,P(D)}{p(x)} \tag{6}$$

where $P(D|x)$ is the probability of the water molecule of being displaced; $P(D)$ is the a priori probability of displaced water molecules, given by the ratio between the number of displaced water molecules in the training set (36) and the total number of observations in the training set (54); $p(x|D)$ is the conditional density function of the binding free energy for displaceable water molecules, which is taken to be the Gaussian density as discussed above.

The marginal density of the binding free energy, $p(x)$, is calculated as follows:

$$p(x) = p(x|C)\,P(C) + p(x|D)\,P(D) \tag{7}$$

The new water molecule can be classified as conserved if $P(C|x) > P(D|x)$, while it can be classified as displaceable by a ligand if $P(D|x) > P(C|x)$.

(52) DeLano, W. L. *The PyMOL Molecular Graphics System*; DeLano Scientific: San Carlos, CA, 2002; http://www.pymol.org.
(53) Ihaka, R.; Gentleman, R. *J. Comput. Graphic. Stat.* **1996**, *5*, 299−314.
(54) Crawshaw, J.; Chambers, J. *A Concise Course in Advanced Level Statistics*; Nelson Thornes: Cheltenham, U.K., 2001.

Table 4. Displacement/Conservation Probabilities for Test Water Molecules

pdb code	water[a]	ΔG_{abs}[b]	$P(C\|x)$	$P(D\|x)$
1h01	107	−2.2	0.21	0.79
1h08	71	−2.4	0.22	0.78
1v1k	108	3.1	0.10	0.90

[a] ID number of the water molecule in the pdb file. [b] Absolute binding free energy of the water molecule (kcal mol^{-1}).

As an example of its use, this approach was tested on three water molecules in protein CDK2 complexed with three different ligands. One water molecule mediates the interaction between the ligand and the protein in those complexes and it can be sterically displaced by apolar groups present in some inhibitors. The structures considered are pdb 1h01, 1h08, and 1v1k and the water molecule of interest in each structure is respectively Wat 107, Wat 71, and Wat 108. The binding free energies of these three water molecules were calculated using the same program and the same conditions previously described in the methods section. In structure 1h01 there are two enantiomers for the ligand, indexed as FAL and FBL in the pdb: the free energy reported here refers to calculations done on enantiomer FAL. Also in structure 1h08 there are two enantiomers for the ligand, BYP and BWP, and in this case the free energy reported refers to calculations done on BWP. The calculated free energies and probabilities for displacement/conservation are reported in Table 4.

As can be seen from Table 4, for Wat 107 there is 79% probability of belonging to the class of displaceable water molecules and 21% probability of belonging to the class of conserved water molecules; for Wat 71 the probability of belonging to the class of displaceable water molecules is 78% and that of belonging to the class of conserved water molecules is 22%; for Wat 108 the probability of belonging to the class of displaceable water molecules is 90% and that of belonging to the class of conserved water molecules is 10%. For all water molecules $P(D|x)$ is higher than $P(C|x)$, so we can say that the water molecules can be displaced by a ligand, and this is in agreement with experimental data as discussed before.

It must be noted that this approach cannot say definitively whether a particular water molecule is or is not displaceable by a ligand, but merely gives the likelihood for this event based on previous experience. Our dataset will inevitably contain bias, arising in part from its limited size, and also from the extent to which synthetic chemists have invested effort in designing molecules to target particular water molecules. In the case of **Wat A** in HIV-1 protease, for example, this statistical approach gives a low displacement probability for this water molecule, whereas we know from experience that this water molecule can be displaced by a specifically designed cyclic urea derivative. The statistics do, however, give an indication of the difficulty associated with displacing a particular water molecule, and perhaps some indication of the synthetic effort required in designing an appropriate ligand to do so.

4.1. Water Displacement and Change in Ligand Binding Affinity. The existence of a relationship between the binding free energies of water molecules that are displaced by ligands and the change in ligand binding affinity after water displacement was investigated. There is a common notion that displacing water molecules in protein−ligand complexes is likely to be advantageous in terms of ligand binding free energy, because of the associated increase in entropy.

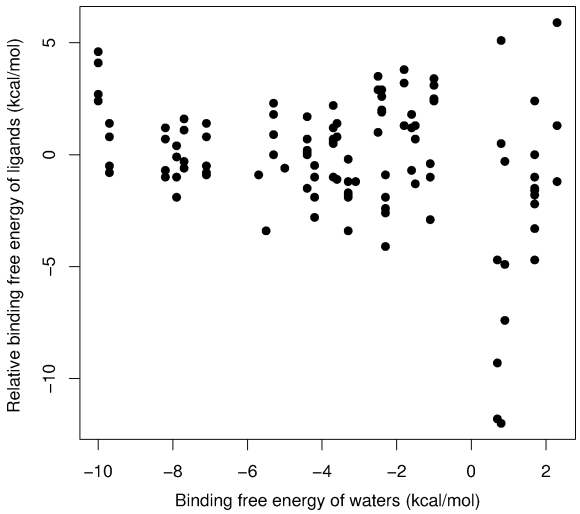

Figure 10. Experimental relative binding free energies of ligands vs calculated binding free energies of water molecules.

Protein−ligand complexes containing water molecules that can be displaced by a ligand and for which the binding free energy was calculated were considered. The binding free energies of the ligands in those complexes were calculated from experimental K_i values using the relation

$$\Delta G = -RT \ln K_i \tag{8}$$

For each protein, a series of ligands displacing the water molecule of interest, for which binding affinities were available, were selected using Relibase+, and they have been provided in the Supporting Information. The difference between the binding free energy of the ligands displacing the water molecule and the ligands interacting with the water molecule ($\Delta\Delta G_{lig}$) was correlated to the calculated binding free energy of the water molecule itself (ΔG_{wat}). The results are presented in Figure 10.

It is evident that no linear correlation exists between the binding free energies of water molecules and the change in binding affinity of ligands displacing the water molecules. There are cases where a ligand displaces a very tightly bound water molecule, but there is a decrease in binding affinity (top left corner of graph) and cases where a ligand displaces a loosely bound water molecule but there is a large increase in binding affinity (bottom right corner of graph). This is not really surprising: we could probably expect to see a correlation if the comparison involved pairs of ligands which have the same structure, but differ only in one functional group, that in one case can interact with the water molecule and in the other displaces it. The structures compared here are all different, one from the other, and in some cases ligands displace more than one water molecule. This analysis is of course complicated by the difficulty in obtaining consistent experimental binding free energy data between different assays, and any real trend will be partly obscured by this problem.

Analysis of the data used to generate Figure 10 shows that there are also cases where some ligands displacing a particular water molecule are more potent than the ligand interacting with that water molecule, but some other ligands displacing the same water molecule are less potent. This is clear evidence of the

fact that, if we needed reminding, binding is a complex process, of which water displacement is only one element, and as such designing a ligand to displace a particular water molecule need not necessarily lead to an increase in potency. Indeed, Figure 10 suggests that water displacement is just as likely to lead to reduced potency, as increase it. Of course, there are other reasons for displacing water molecules, including the desire for greater ligand specificity. We should also note that the protein−ligand complexes specifically chosen in this study do not show significant changes in binding mode or reorganization of the bridging waters on ligand modification. These additional factors will impose further limitations on the methodology described here.

5. Conclusions

In this work, the calculation of the binding free energies of 54 water molecules in different protein−ligand complexes of a selected dataset has been discussed. Monte Carlo computer simulations using replica exchange thermodynamic integration and the double decoupling approach have been used to calculate the free energies with a precision (i.e., statistical error) of the order of 0.5 kcal mol^{-1}.

It was demonstrated that the binding free energies of water molecules are dependent on the nature of the environment: tightly bound water molecules are generally located in highly polar cavities and they can make three or four hydrogen bonds with the protein and the ligand; loosely bound water molecules are generally located in partially apolar cavities and they can make less than three hydrogen bonds with the protein and the ligand.

The difference in average binding free energy between molecules that are known to be conserved, and those that may be sterically displaced by ligands, was shown to be statistically significant at the 95% level of confidence. Bayesian statistics was then applied to demonstrate that, given the binding free energy of a water molecule, it is possible to calculate the probability of the water molecule being conserved or displaced by a ligand. This knowledge may then be used to focus the synthesis of ligands ad hoc, to maximize the interactions with conserved water molecules and target those that may be displaced.

In many studies reported in the literature, the increase or decrease in binding affinity of ligands has been associated with the displacement of a particular water molecule.[4,5,24] With this mind, the existence of a relationship between the calculated binding free energies of water molecules and the change in binding affinity of ligands displacing those water molecules was investigated. No direct relationship was found, partly because water displacement is only one part of the binding process and other factors are obviously important, and partly because of the difficulties in obtaining consistent experimental ligand binding affinities.

As discussed in the introduction, in the context of rational drug design it is important to identify conserved water molecules in the binding site of a protein, as inclusion of these water molecules in docking[6,8,9] and de novo drug design[10,11] can greatly improve the results. The problem is *how can these molecules be identified?* On the basis of this work, two situations may be considered:

(1) If the structure of a target protein in the apo-form is available, but there are no holo-structures, then the available information is quite limited. The hydration pattern in the binding site of the protein can be very different in the empty pocket and in the pocket with a ligand bound, as indeed can the protein structure itself. OppA is just such an example discussed in the present work. Software like Consolv[15] and WaterScore,[16] can be used to identify conserved water molecules in the binding site of the free protein. Programs like GRID[55] or SuperStar[56] can also be used to identify favorable hydration positions.

(2) If at least one crystallographic structure of a target protein complexed with an inhibitor is available and water molecules are present in the binding site, then the binding free energy of those water molecules can be calculated. The probability of each water molecule being conserved or displaced by a ligand can then be evaluated using eqs 4 and 6 as discussed above. This information can then be used to generate new ligands that maximize the interactions with conserved water molecules and target water molecules that can be displaced.

The difficulty of this last approach, however, is that expensive and technically difficult free energy calculations are required. As intimated in this paper, there is a correlation between the calculated binding free energies of the water molecules and the nature of their binding pocket. Using the free energy data reported here, we have therefore used a simple and fast statistical approach, based on molecular descriptors of the binding pocket, to successfully predict the water-binding free energies. These results will be reported elsewhere.

Acknowledgment. We would like to thank Syngenta for funding this project, and Dr. Christopher J. Woods and Professor Alan Welsh for helpful discussions.

Supporting Information Available: Complete list of authors for references 24 and 26; details for the set up of all systems; figures showing the free energy gradients as a function of the coupling parameter for one example of the annihilation of a displaceable water molecule and one example for a conserved water molecule; the list of ligands displacing water molecules for which the binding free energy is known and that were used to generate the data in Figure 10. This material is available free of charge via the Internet at http://pubs.acs.org.

JA066980Q

(55) Goodford, P. J. *J. Med. Chem.* **1985**, *28*, 849−857.
(56) Verdonk, M. L.; Cole, J. C.; Taylor, R. *J. Mol. Biol.* **1999**, *289*, 1093−1108.