JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

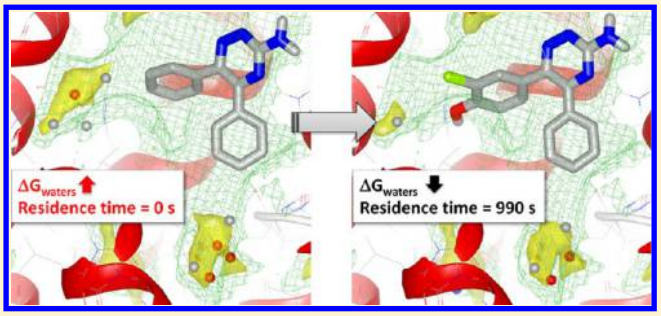
Water Network Perturbation in Ligand Binding: Adenosine A_{2A} Antagonists as a Case Study

Andrea Bortolato,*[,†] Ben G. Tehan,[†] Michael S. Bodnarchuk,[‡] Jonathan W. Essex,[‡] and Jonathan S. Mason[†]

[†]Heptares Therapeutics Ltd, BioPark, Broadwater Road, Welwyn Garden City, Herts, AL7 3AX, U.K.

[‡]School of Chemistry, University of Southampton, Highfield, Southampton, Hampshire, SO17 1BJ, U.K.

ABSTRACT: Recent efforts in the computational evaluation of the thermodynamic properties of water molecules have resulted in the development of promising new in silico methods to evaluate the role of water in ligand binding. These methods include WaterMap, SZMAP, GRID/CRY probe, and Grand Canonical Monte Carlo simulations. They allow the prediction of the position and relative free energy of the water molecule in the protein active site and the analysis of the perturbation of an explicit water network (WNP) as a consequence of ligand binding. We have for the first time extended these approaches toward the prediction of kinetics for small molecules and of relative free energy of binding with a focus on the perturbation of the water network and application to large diverse data sets. Our results support a qualitative correlation between the residence time of 12 related triazine adenosine A_{2A} receptor antagonists and the number and position of high energy trapped solvent molecules. From a quantitative viewpoint, we successfully applied these computational techniques as an implicit solvent alternative, in linear combination with a molecular mechanics force field, to predict the relative ligand free energy of binding (WNP-MMSA). The applicability of this linear method, based on the thermodynamics additivity principle, did not extend to 375 diverse A_{2A} receptor antagonists. However, a fast but effective method could be enabled by replacing the linear approach with a machine learning technique using probabilistic classification trees, which classified the binding affinity correctly for 90% of the ligands in the training set and 67% in the test set.



INTRODUCTION

Water represents a fundamental component for protein function with a crucial impact on protein plasticity, allosterism, mediation of ligand binding, and protein−protein interactions.[1] Displacement of unfavorable waters by the ligand, replacing them with groups complementary to the protein surface, represents a crucial driving force for protein−ligand binding.[2,3] Water molecules solvating protein active sites are often (A) entropically unfavorable due to the orientational and positional restraints imposed by the protein cavity and (B) energetically unfavorable due to the water molecule's inability to form a full complement of hydrogen (H−) bonds when solvating the protein surface. Knowledge of the desolvation penalty for a particular H-bonding group in the protein binding site enables the identification of H-bonding groups that could be key for ligand binding. Thus for both reasons it is preferable to include explicit water molecules in structure-based drug design approaches, and there are many examples of their importance, e.g. refs 4−7. In 1985, Goodford[8] developed the GRID software, using probes including water to energetically survey a protein binding site to identify areas of attraction (hotspots). Since then a number of studies have been published to try to model the role of water in ligand binding.[9−14] Recent efforts in the computational evaluation of the thermodynamic properties of water molecules resulted in the development of new promising in silico methods to evaluate the hydration properties of proteins, including in complex with small organic molecules.[15−27]

The importance of water for G protein-coupled receptors (GPCRs) has been supported by recent crystallographic data[20,28,29] and data from different studies[30−32] showing how ordered waters interact with residues that are important in disease states, receptor activation, and signaling. GPCRs represent one of the most important target classes in pharmaceutical research. Approximately 24% of all drugs reaching the market in the past decade target this protein family.[33] They constitute the largest family in the human genome, consisting of 390 GPCRs, excluding olfactory receptors.[34] Among GPCRs, adenosine receptors represent promising therapeutic targets for CNS diseases, cerebral and cardiac ischemic diseases, cancer, and immunological and inflammatory disorders.[35,36] Adenosine has a crucial role in regulating human brain functions such as the modulation of dopamine transmission, cognition, and memory, and therefore, adenosine receptor ligands represent potential therapeutic agents for the treatment of schizophrenia, panic disorder, anxiety, and Parkinson's disease.[37] For example, Preladenant, an

Received: March 8, 2013
Published: June 1, 2013

antagonist of the A_{2A} member of the adenosine receptor subfamily (A_{2A}), has recently progressed to phase III clinical trials for the treatment of Parkinson's disease.[38]

Several new GPCR structures in the active (agonist binding) and inactive (antagonist/inverse agonist binding) conformations have recently been published enabling detailed structural information to be used for hit identification and drug design. A total of 12 crystallographic structures of A_{2A} in complex with small organic molecules are now available in the Protein Data Bank (PDB).[39] They have been obtained by exploiting three diverse techniques: (A) the receptor has been fused to the T4 lysozyme[40,41] or apocytochrome b(562)RIL[29] between the intracellular ends of helices V and VI; (B) the receptor has been crystallized in complex with a mouse monoclonal antibody (Fab);[42] (C) a stabilized receptor for the desired state/conformation (active/inactive etc) has been obtained using the StaR approach,[43] by introducing a small combination of specially selected thermostabilizing mutations which allow crystallization in short-chain detergent and structure determination with both low and high affinity ligands,[44−46] but without affecting the binding (pharmacology) of ligands of that state.

In this study we analyzed the perturbation on the A_{2A} active site water network (WNP) as a consequence of ligand binding. We focused initially on 12 triazine A_{2A} receptor antagonists, for which reliable poses and protein conformations were available based on structural information for 2 ligand−protein complexes (crystallographic data) and on Biophysical Mapping (BPM)[46,47] for other ligands. In the BPM approach, additional single mutations are added to the StaR at positions that could be involved in small molecule interactions. The StaR and the panel of binding site mutants are captured onto Biacore chips to enable characterization of the binding of small molecule ligands using surface plasmon resonance (SPR) measurement. A matrix of binding data for a set of ligands versus each active site mutation is then generated, providing specific affinity and kinetic information (K_D, k_{on}, and k_{off}) of receptor−ligand interactions. This data set, in combination with molecular modeling and docking, is used to map the small molecule binding site for every compound.

We used WaterMap,[15,16,18] SZMAP,[48] Grand Canonical Monte Carlo simulations (GCMC),[9,49] and CRY, a novel GRID[8,50,51] probe, to evaluate the potential impact of the solvent in structure−activity and structure−kinetics relationship elucidations (Figure 1). Understanding of ligand binding kinetics at a molecular level and computational prediction of residence time (τ) represents a significant challenge for modern computational chemistry. Residence time, as quantified by the dissociative half-life (or off-rate) of the drug−target binary complex is an alternative perspective in drug optimization classically quantified by binding parameters such as IC_{50} or K_D. This aspect alone can in many cases define the duration of efficacy for a drug.[52−54] The residence time is a crucial metric of compound optimization and often is a key indicator of in vivo target efficacy and selectivity. Experimentally the in vitro cellular activity of compounds can be associated directly with the dissociative half-life of the receptor−ligand complex.[55,56]

The approaches used in this study are based on different and generally complementary theoretical frameworks. GRID energetically surveys protein binding sites by systematically calculating the interaction energy of chemical probes (ligand functional groups) with the protein. The water probe is able to make up to four hydrogen bonds with the protein. The CRY probe, a new combination of the DRY hydrophobic (including an entropic energy term) and C1= lipophilic probes, analyzes the hydrophobic and lipophilic/apolar character of the environment around the water molecules. The benefit of combining the two probes comes from the fact that the DRY probe is akin to using an inverse water probe and, thus, does not highlight regions where both water and hydrophobicity can both be favorable in adjacent positions. Combining the lipophilic C1= probe with the DRY probe ensures no areas of hydrophobicity are missed; for example in the factor Xa S1 binding site a lipophilic hotspot is identified with the C1= probe for the region where a critical water is displaced by a chloro group[15] whereas the DRY probe in a complementary way identifies the aromatic ring it is attached to. SZMAP and WaterMap try to estimate the water free energy including the breakdown into entropic and enthalpic contributions. SZMAP calculations are based on a semicontinuum approach: for every grid point, it places one explicit water molecule, treating the rest as a continuum, and samples the water orientations. In contrast, WaterMap exploits an all atom explicit solvent molecular dynamics simulation followed by a statistical thermodynamic analysis of water clusters. The Grand Canonical method computes the free energy of water binding to proteins by equilibrating concentrations between a reference state and a simulation cell that includes waters bound to a protein active site. The GCMC method locates ensembles of water positions consistent with a selected free-energy level. We tested the suitability of using these methods as an implicit solvent alternative, in linear combination with a molecular mechanic force field (WNP-MMSA), for the quantitative prediction of ligand free energy of binding (Figure 1). We extended the method's applicability to 375 A_{2A} receptor antagonists[46,57] substituting the linear combination method with a probabilistic classification tree[58] algorithm (ctWNP-MMSA), known to work for similarly related problems such as docking/scoring and protein−protein interaction predictions.[46,59−61]

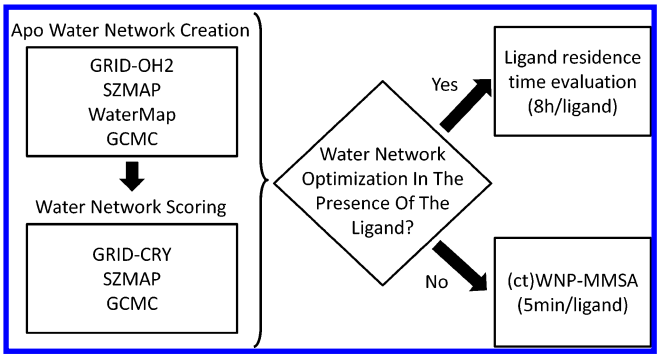


Figure 1. Schematic workflow diagram. Different computational methods can be used to create the starting water network and to evaluate the energy of every molecule of water with and without the different ligands. Optimizing the water network in the presence of every small molecule can be useful to understand the ligand residence time and requires generally an overnight calculation for every ligand assuming a standard 8 CPUs cluster. It is possible to estimate the free energy of binding quickly (5 min, 1 CPU) for every ligand using the (ct)WNP-MMSA methodology without optimizing the water positions after ligand binding.

RESULTS

Water Network Perturbation Prediction in Ligand Binding. The 3D coordinates of the A_{2A} receptor in complex

dx.doi.org/10.1021/ci4001458 | J. Chem. Inf. Model. 2013, 53, 1700−1713

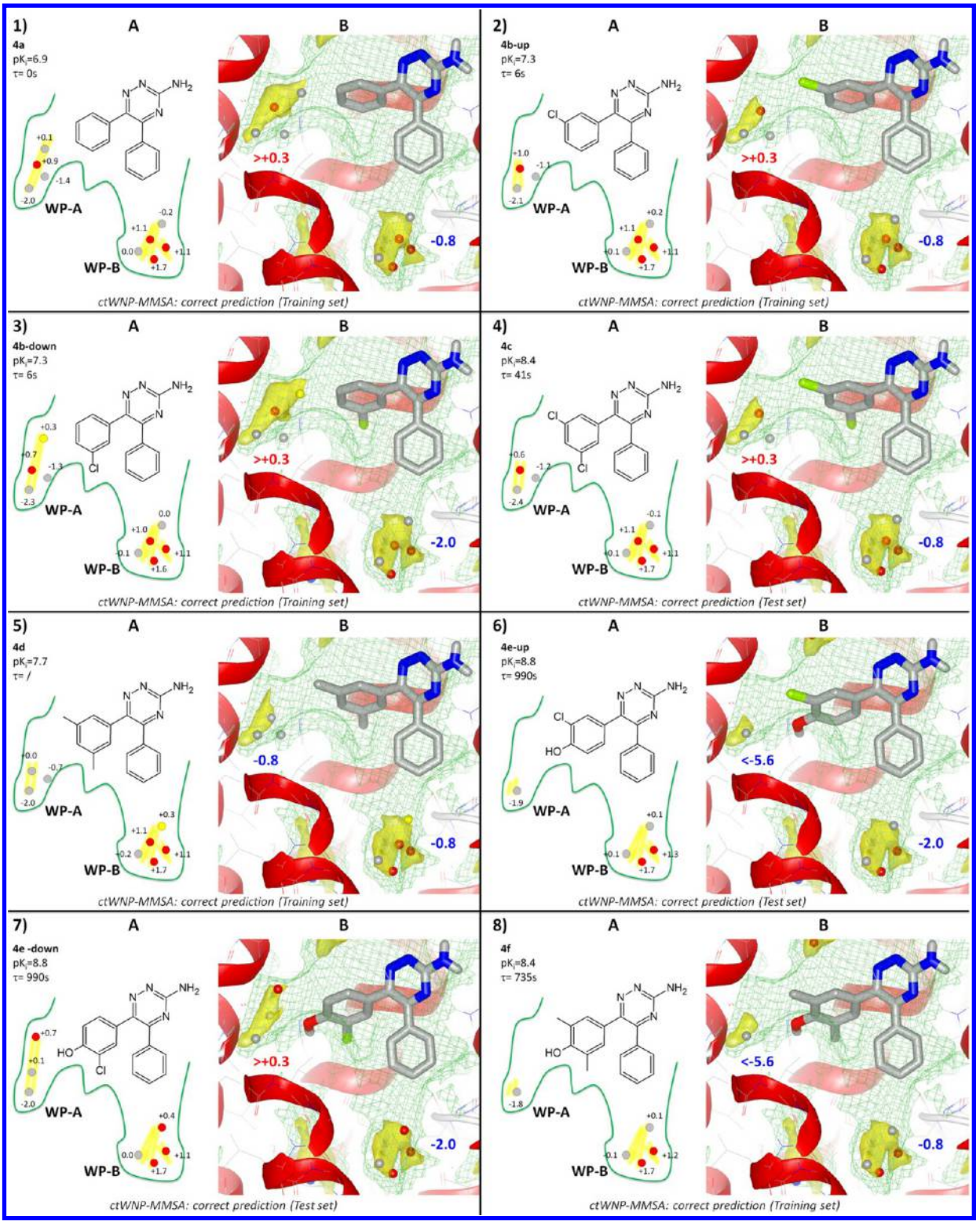


Figure 2. Predictions of the water network perturbation as consequence of a ligand (**4a**, **4b**, **4c**, **4d**, **4e**, **4f**) binding. (panels 1−6, section A) For every ligand the name, the pK_i, and the residence time (τ) in seconds are indicated on the top left. The residence time of **4d** has not been determined. The chemical structure is shown on the bottom left in a schematic representation of the A_{2A} pocket. The 2D pocket includes the predicted waters using the SZMAP/WaterMap protocol shown as circles and the corresponding free energy predictions in kilocalories per mole evaluated using the SZMAP water-neutral probe free energy difference. They have been color coded in red if predicted by SZMAP having a free energy >0.4 kcal/mol, in yellow if their predicted ΔG resulted between 0.2 and 0.4 kcal/mol, in gray if ΔG is 0.2 to −3.5 kcal/mol, and blue if $\Delta G <$ −3.5 kcal/mol. The approximate position of the CRY probe hot spots is also included in yellow. The two regions (WP-A and WP-B) of the binding site are labeled. (panel 1−6: section B) The crystallographic ligand pose or the predicted docked pose in the A_{2A} pocket are illustrated. The water accessible surface has been created using GRID OH2 probe, represented as green mesh and contoured at 1 kcal/mol. GRID CRY probe hot-spot regions are represented as a transparent yellow solid surface and contoured at −2.5 kcal/mol. Water molecules placed using the SZMAP/WaterMap protocol are represented as spheres and color coded as in the 2D representation (section A). The nearest water molecules to the ligand (first

dx.doi.org/10.1021/ci4001458 | J. Chem. Inf. Model. 2013, 53, 1700−1713

Figure 2. continued

hydration shell) in WP-A and WP-B have been rescored using the GCMC method. The resulting ΔG energy in kilocalories per mole is included near the corresponding pockets in red for positive values (corresponding to *unhappy* waters) and in blue for negative values (corresponding to *happy* waters). In the bottom of every panel, the resulting (correct or wrong) prediction of the classification trees method (ctWNP-MMSA) for the ligand is indicated.

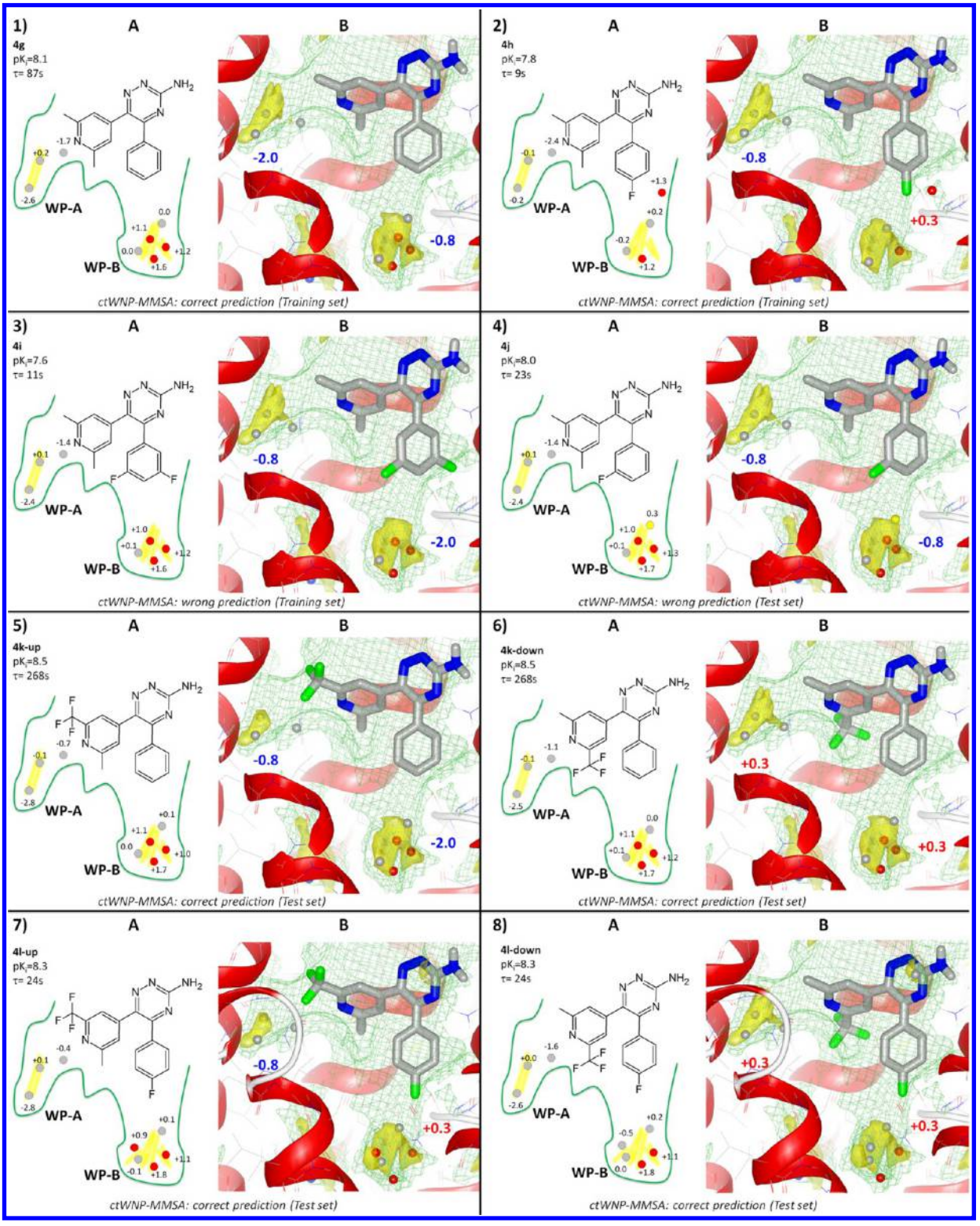


Figure 3. Predictions of the water network perturbation as a consequence of ligand (**4g**, **4h**, **4i**, **4j**, **4k**, **4l**) binding. The figure is organized as in Figure 2.

dx.doi.org/10.1021/ci4001458 | *J. Chem. Inf. Model.* 2013, 53, 1700−1713

with the triazine antagonists **4e**[45] (PDB[39] ID 3UZC, Figure 2-7) and **4g**[45] (PDB ID 3UZA, Figure 3-1) were used as reference ligand−protein structures. For the ligands **4e** and **4g**, the crystallographic binding mode (PDB IDs 3UZC and 3UZA)[45] has been used. For the other 10 triazines included in Figures 2 and 3, the high chemical similarity allows their binding mode to be modeled by molecular alignment to the most similar crystallographic ligand (**4e** or **4g**). A water network was created for these 12 triazines with the following optimized computational protocol: (A) For every selected ligand binding pose an initial water network in the binding site was created around the small molecule using an optimized protocol exploiting SZMAP.[48] (B) The system was completed (adding nonbinding site waters) and optimized using a customized WaterMap protocol (excluding the "solvate_pocket" step handled by step A). WaterMap calculations[15,16] consist of an all atom explicit solvent molecular dynamics simulation followed by a statistical thermodynamic analysis of water clusters (hydration sites). (C) The free energy of every water molecule near the ligand was additionally evaluated using SZMAP in the context of the protein−ligand complex. (D) The nearest water molecules to the ligand (first hydration shell) were also rescored using the GCMC method. The GCMC simulations for the water molecules were performed using the interacting particle method[62] at six different chemical potentials to observe the changes in water populations as a function of the binding free energy. The free energy of the waters has been estimated by comparing the position of the predicted waters (at the end of step B) to the GCMC water density grids.

Impact of Water Network Perturbation on Ligand Structure−Kinetics Relationships. The binding of **4e** (Figure 2) to the A_{2A} receptor results in the creation of two distinct regions enclosed between the ligand and the protein. These water pockets, labeled WP-A and WP-B in Figures 2 and 3, are not in contact with the solvent toward the extracellular side and are based on the crystallographic structures of **4e** and **4g** in complex with A_{2A} (PDB ID 3UZC and 3UZA).

WP-A protrudes between TM2 and TM3. In TM2 it is opposite to P61$^{2.59}$ corresponding to the kink in TM2 and involves A59$^{2.57}$, F62$^{2.60}$, A63$^{2.61}$, and I66$^{2.64}$ (superscripts refer to Ballesteros−Weinstein numbering[63]). In TM3 WP-A is facing the helix kink and includes I80$^{3.28}$, A81$^{3.29}$, F83$^{3.31}$, and V84$^{3.32}$. On TM7 the only relevant residue is H278$^{7.43}$, while the top of the pocket is delimited by F168 in the extracellular loop 2. GRID analysis using the CRY and OH2 probe shows that this pocket has a dual nature: highly hydrophobic as expected by the many apolar residues present, but at the same time also a hotspot for water molecules. The polar character is the consequence of H278$^{7.43}$ and of the kink in TM3 releasing the backbone amide atoms from the helix H-bond interactions. This dual hydrophilic−hydrophobic nature is also seen in WP-B due to the presence of polar and apolar residues. WP-B is delimited by TM3 (L85$^{3.33}$, A88$^{3.40}$, Q89$^{3.37}$, and I92$^{3.40}$) and TM4 (I135$^{4.56}$ and G136$^{4.57}$) facing the helix kink at P139$^{4.60}$, TM5 (Y176$^{5.37}$, M177$^{5.38}$, F180$^{5.41}$, N181$^{5.42}$, and C185$^{5.46}$), and TM6 (F242$^{6.44}$, W246$^{6.48}$, and H250$^{6.52}$).

We analyzed the effect of 12 related triazines A_{2A} antagonists (Figures 2 and 3) on the water molecules occupying these 2 pockets. They represent a near ideal test set for a computational chemistry structure-based approach to investigate receptor binding: (A) the binding affinity and kinetics properties of these small organic molecules has been previously measured;[46] (B) the binding mode prediction to the A_{2A} receptor can be based on the two crystal structures and Biophysical Mapping results.[46,47] When the experimental data were not sufficiently precise to select a unique binding mode, the two possible orientations have been analyzed. In particular two similar docking modes have been used for ligands with meta-substituents in the ring near WP-A (**4b**, **4e**, **4k**, and **4l**). We defined the **up** orientation as pointing toward the WP-A and the **down** binding mode toward TM7 as consequence of a rotation of 180° of the dihedral angle between the phenyl and the triazine rings.

An optimized computational protocol was then designed to estimate the number, positions, and free energies of the water molecules in the two pockets. For the initial water placement we created a 0.5 Å resolution grid in the region of interest and analyzed at every point the free energy of a water molecule using SZMAP. The most favorable nonclashing positions have been used to create a starting water network that has been optimized using WaterMap. The energy of the final water clusters have been further analyzed with SZMAP. For the high resolution A_{2A} crystal structure,[29] this protocol predicted the position of the 17 water molecules near the ligand in the pocket with an average error of 0.4 Å. Water molecules in bulk solvent were not predicted as accurately. This computational strategy resulted in robust, reproducible water networks which were easy to compare among the different ligands. This meant that if the part of the molecule interacting with a pocket was exactly the same and had the same location the resulting number, positions, and free energies of the water molecules in that pocket gave similar results. For example, the phenyl ring in the molecules **4a**, **4b**, **4c**, and **4d** interacting with the WP-B in a comparable fashion always resulted in five predicted water molecules in a very similar arrangement with similarly predicted ΔGs (Figure 2). The predicted water molecules in the protein active site are color coded, as defined in the Figure 2 legend, by the SZMAP water-neutral probe free energy difference and the previously[20] used terms of *happy* (blue spheres) and *unhappy* (red and yellow spheres) waters used for waters with significant negative and positive free energy differences. The results obtained were analyzed in comparison to the GCMC simulations and in particular for the nearest water molecules to the ligand (first hydration shell). The GCMC method locates ensembles of water positions consistent with a selected free-energy level. The free energy of the solvent molecules has been estimated by comparing the position of the SZMAP/WaterMap predicted waters to the GCMC water density grids with a 0.3 occupancy threshold.

A summary of the results for all 12 compounds is shown in Figures 2 and 3. **4a** was used as reference for the analysis of this set of molecules because (A) it has a very short residence time estimated at zero seconds; (B) it is the smallest compound; and (C) it has no substituents in both the phenyl moieties interacting with the water pockets. Using the SZMAP/WaterMap protocol a total of four and five solvent molecules were predicted in the WP-A and WP-B respectively for **4a**. One *unhappy* water is present in WP-A, while three are present in WP-B. The first shell water analysis based on the GCMC method suggests a possible direct interaction between the ligand and an *unhappy* water in WP-A that could be related to the very short residence time. Residence time is determined by the protein−ligand binding affinity and the free energy of activation of the ligand unbinding event. It is possible that in addition to having the least favorable protein−ligand binding free energy, breaking interactions to this water is easy, and so

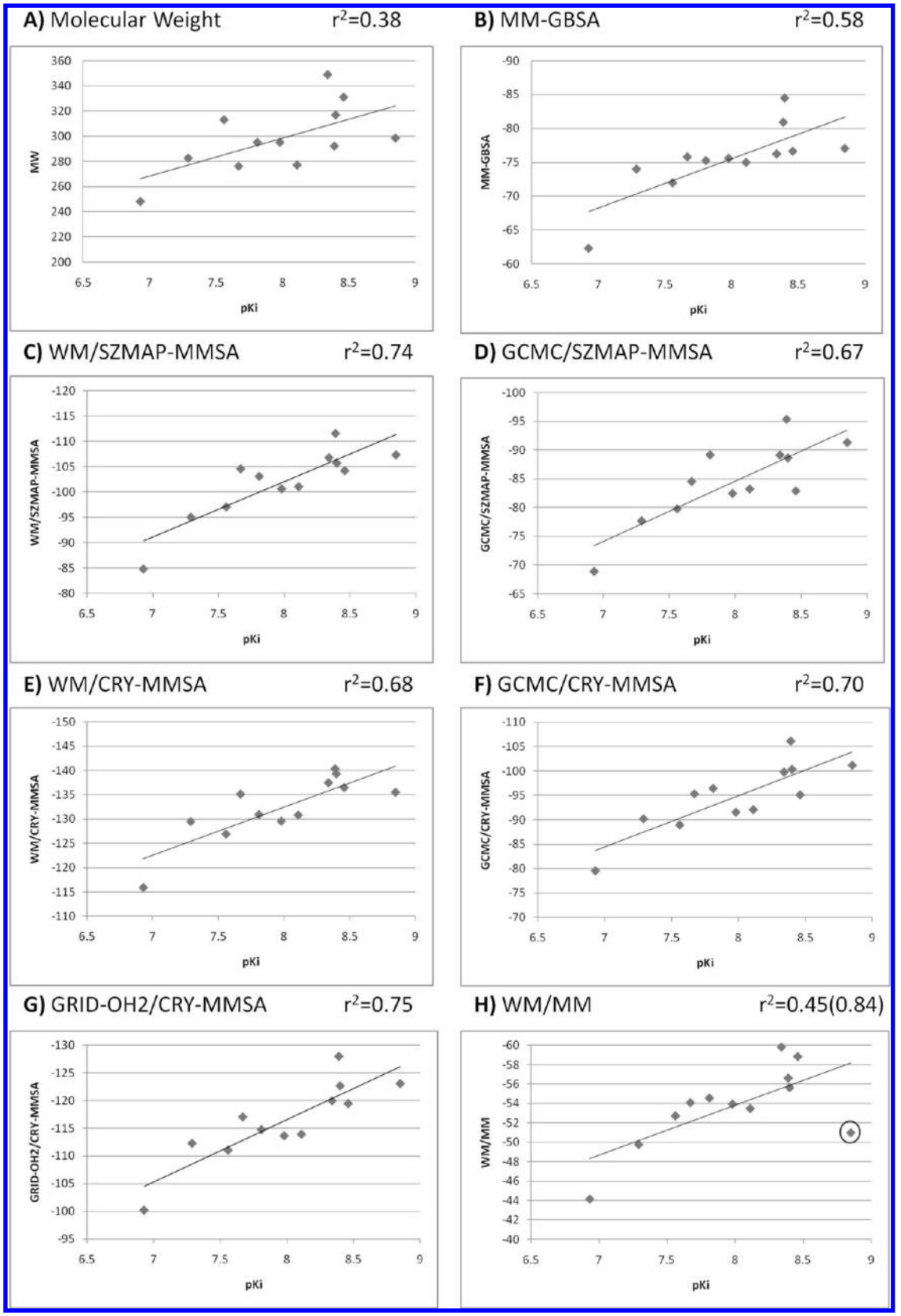


Figure 4. Correlation between ligand pK_i and molecular weight (A), MM-GBSA (B), explicit WNP-MMSA based methods (C−G) and WM/MM (H). Correlation factors are indicated as r^2. For WM/MM, r^2 corresponds to the correlation including an outlier indicated in the plot by a black circle. In brackets it is indicated the correlation factor without including it.

the free energy of activation is reduced. This model explains how weakly bound waters can affect residence times, and it is complementary to the study of Schmidtke et al.[64] supporting the importance of shielded hydrogen bonds as structural

dx.doi.org/10.1021/ci4001458 | J. Chem. Inf. Model. 2013, 53, 1700−1713

determinants of binding kinetics. **4b-up**, **4b-down**, **4c**, and **4d** include an unsubstituted benzene ring adjacent to WP-B. This is the same in **4a** and it results in comparable water properties for this pocket. They present instead differences in the substituents in the ring near WP-A in the meta position (one or two chlorine atoms or two methyl moieties). Chlorine atoms in ligands pointing toward WP-A (**4b-up** and **4c**) contribute to the hydrophobic environment occupied by one *unhappy* water. This could be linked to the short residence time of these ligands, in a comparable way to that described for **4a**. Similar conclusions can be drawn by the GCMC predicting in general no waters with negative ΔG_{bind} in WP-A in close contact with these ligands. Interestingly, the **4d** ligand with a smaller methyl moiety in the meta position near WP-A resulted in no *unhappy* waters in that region according to SZMAP/WaterMap, with the GCMC method also suggesting stabilization of the water in contact with the ligand in WP-A compared to the molecule **4a**.

The orientation **4b-down** seems comparable from a water energetics perspective to **4b-up** while the difference between the up and down conformation for **4e** is significantly different. **4e-up** and **4f** binding results in only one water in WP-A, that is stabilized by interaction with the ligand hydroxyl group. This observation can be linked to the much longer residence time of these small organic molecules: 990 and 735 s, respectively. **4e-down** seems less favorable than **4e-up** from a water prospective, and it is resulting in three waters in WP-A, one of which is *unhappy*. The first water shell analysis supports these results and in particular **4e-up** does not show any perturbation or destabilization of the adjacent waters upon binding, but on the contrary has only one very stable water in WP-A ($\Delta G < -5.6$ kcal/mol). These results are in agreement with a possible increase of the free energy of activation of the ligand unbinding event due to a tightly bound water linking the phenolic group of the ligand to the protein. For **4e-up/down**, WP-B contains water molecules with the same score using the GCMC, but slightly different predictions using SZMAP (especially 1 water is 0.6 kcal/mol less stable) even if the local environment is the same. In SZMAP this result could be due to a long-range electrostatic effect of the chlorine atom; however, this energy difference is probably close to the method sensitivity.

The introduction of a 2,6-dimethyl-pyridine results in a general stabilization of the water network in WP-A compared to **4a**. For **4g** the stabilization is estimated by the GCMC method to be less strong than **4e-up**, and this is in agreement with a shorter residence time (87 s vs 990 s). The trifluoromethyl derivative **4k** (τ = 268 s) in the **up** conformation is predicted by GCMC to stabilize more the water network than the **down** binding mode and stabilizes more WP-B than WP-A in contrast with **4g**. The ligands **4h**, **4i**, **4j**, **4l-up**, and **4l-down** with different fluorine substituents interacting with the solvent in WP-B have low residence time. The para-fluoro derivatives **4l** and **4h** gave results from the SZMAP/WaterMap (**4h**) and the GCMC (**4h** and **4l-up**) analysis of a destabilizing effect on the water network in WP-B. Interestingly the meta-fluoro moieties (**4i** and **4j**) according to the GCMC method seem to detrimentally affect the water stabilizing role of the pyridine, increasing the binding free energy of the water in WP-A by more than 1 kcal/mol compared to **4g**.

Impact of Water Network Perturbation on Ligand Free Energy of Binding Predictions. A popular approach to predict the ligand free energy of binding is based on a combination of the molecular mechanics force-field, the evaluation of the solvent accessible surface area and an implicit solvent model like Generalized Born (MM-GBSA) or Poisson–Boltzmann (MM-PBSA) method.[65,66] Alternative rapid free energy of binding predictions methods are the Linear Interaction Energy (LIE)[67−69] method and the Linear Response Approximation (LRA)[70,71] method. It should be noted these methods take into account only the bound and unbound states of the ligand–protein complex without considering intermediate states. Further to these methods we tested the possibility of replacing the MM-GBSA implicit solvent contribution with a water network perturbation energy term (WNP), which takes into account the effect of ligand binding on an explicit water network predicted within the protein active site. We call this new strategy WNP-MMSA. The initial solvent network and the water scoring can be generated with different methods resulting in different WNP-MMSA protocols. In this study the water network has been created with WaterMap, GCMC simulation, and the GRID water probe OH2. The free energy of every water was then evaluated with and without ligands using SZMAP and a novel optimized GRID probe, CRY. This apolar/hydrophobic molecular probe is based on the combination of the DRY and the sp^2 aromatic CH (C1=) probes. In WNP-MMSA the water positions are not modified or recalculated after ligand binding. Waters clashing with the ligand are considered displaced. The effect of the ligand on the remaining solvent molecules is analyzed with SZMAP or CRY. Avoiding the optimization of the water network as a consequence of ligand binding has the benefit of producing computational protocols of comparable speed to MM-GBSA; it also favors ligands that have a minimum impact on the arrangement on the water network, and thus more compatible with the "natural" solvent structure present in the apo protein active site.

A range of combinations of methods for the prediction of the ligand free energy of binding were examined. Only WNP-MMSA implementations resulting in better-than-random predictions are shown in Figure 4. They all include the MMSA term as calculated in MM-GBSA described above. The general nomenclature for the methods used in this work includes, the initial name of the method for the water network creation, followed by the method for water scoring, and finally the term MMSA: WM/SZMAP-MMSA, GCMC/SZMAP-MMSA, WM/CRY-MMSA, GCMC/CRY-MMSA, GRID-OH2/CRY-MMSA. For example WM/SZMAP-MMSA corresponds to a protocol using WaterMap to generate the water network, which is scored with SZMAP and combined with the molecular mechanics force field and the solvent accessible surface area. These methods have been compared with the WM/MM method.[16]

Predictions of the relative free energy of binding of the 12 triazines shown in Figures 1 and 2 were initially generated. These represent a good test set because their activity covers 100 fold from 1 to 100 nM and correlates poorly with the molecular weight (r^2 = 0.38, Figure 4-A). As a comparison, the MM-GBSA method[72] was evaluated resulting in an r^2 = 0.58 (Figure 4-B). The combination of the force field, the solvent accessible surface area, and a water network created using WaterMap and analyzed with SZMAP (WM/SZMAP-MMSA) resulted in correlation coefficients of 0.74 (Figure 4-C). The substitution of WaterMap to generate the water network with the GCMC method (GCMC/SZMAP-MMSA) gave an r^2 = 0.67 (Figure 4-D). The CRY probe energy function in combination with MMSA and WaterMap (WM/CRY-MMSA) or GCMC (GCMC/CRY-MMSA) for the solvent

network creation resulted in correlation coefficients of 0.68 and 0.70, respectively (Figure 4-E and -F). Substitution of the GCMC with a water network created using the GRID water probe (GRID-OH2/CRY-MMSA) further improved the predictions (r^2 = 0.75, Figure 4-G), resulting in a very fast methodology for both water network generation and rescoring. A similar published approach (WM/MM) using WaterMap in combination with energetic terms from a MM-GBSA calculation has been evaluated. It includes WaterMap and energy terms for protein−ligand van der Waals contacts, electrostatic interactions, ligand desolvation, and internal strain (ligand and protein). It resulted in a correlation factor of 0.45, 0.84 after the exclusion of one molecule (Figure 4-H). The most relevant difference between WM/MM and the WNP-MMSA methods is that the energy of every water not displaced by ligand binding is evaluated in the presence of the small organic molecule ligand for the WNP-MMSA methods.

This initial limited test showed promising results in the application of WNP-MMSA methods. The application to a larger diverse set of A_{2A} antagonists with known activity, including 335 triazines and 40 chromones, remained somewhat challenging. This data set includes compounds with K_i from low micromolar to low nanomolar, with approximately a Gaussian molecular weight distribution from 200 to 400 Da. A Gaussian distribution was also found for cLogP[73] predictions, spanning approximately a range from 0 to 5. The data set include diverse molecular structures: around 80% of the triazines have a FCFP_6 Tanimoto similarity lower than 0.5 relative to **4g**. FCFP_6 are based on the SciTegic extended-connectivity functional class fingerprints with a maximum bond distance of 6 from the central atom. For the chromones, about 50% of them had a Tanimoto similarity using FCFP_6 lower than 0.5 relative to the A_{2A} antagonist **12**.[57] All analyses have been done in Pipeline Pilot.[74] Triazines have been docked to the receptor using the most similar crystallographic ligand (**4e** or **4g**) as reference. For the docking of 39 chromone derivatives the BPM pose of the antagonist **12**[57] in complex with A_{2A} has been used as reference.

All the methods tested in this study (MM-GBSA, WNP-MMSA, and WM/MM) resulted in a poor prediction (worse than random) of the relative free energy of binding for this more realistic lead optimization medicinal chemistry test set (multiple series, limited structural data) that could be linked to wrong binding mode predictions. All the free energy of binding prediction methods described are parametrized based on an additivity approximation.[75−77] We evaluated an alternative to the linear combination method with a machine learning approach based on probabilistic classification trees[58,78] (ctWNP-MMSA). We used this novel approach treating the data as a two-class classification problem with 375 ligands divided into two groups based on their activity: weak pK_i = 6−7.5 and active pK_i = 7.5−9. Molecules were divided randomly into two groups of equal size: a training set and a test set. Five properties were used for the model creation: CRY total energy of the displaced waters, CRY total water perturbation energy (difference in CRY probe energies for water in apo vs halo structure), the change upon binding of solvent accessible surface area (ΔSA), Coulomb interaction energy (ΔCoulomb), and van der Waals interaction energy (ΔvdW). We applied the novel protocol to GRID-OH2/CRY-MMSA, the best performing method tested in the preceding study (r^2 = 0.75). The resulting model (Table 1) classified the binding affinity correctly for 90% of the molecules in the training set and 67% of the ligands in the test set (not included in the model creation). In the starting set of triazines, 10 out of 12 ligands have been classified correctly (Figures 2 and 3).

Table 1. Illustration of the Performance of the ctGRID/CRY-MMSA Method for the Test Set and Training Set

	ligands pK_i range	correct prediction	incorrect prediction
training set 188 ligands	6.0−7.5	107	10
	7.5−9.0	62	9
	total	**169**	**19**
test set 187 ligands	6.0−7.5	91	37
	7.5−9.0	35	24
	total	**126**	**61**

DISCUSSION

This study was performed with the intention to present and analyze a different view of the ligand binding event, moving the attention from just direct ligand−protein interactions and extending it to the water molecules also involved in the process. We believe this perspective to be complementary and important in understanding the molecular basis of ligand binding. The studies undertaken were focused on an understanding of how the solvent is affected by ligand binding and the potential impact this has on structure-based drug design. We believe this is a timely initial analysis of the broad range of complementary computational tools trying to tackle this difficult topic. In this study we included and combined WaterMap, SZMAP, GRID, and GCMC simulations (and in the second part probabilistic classification trees), delineating combinations that for the data set used were most effective (considering both time and accuracy). The water network perturbation in the A_{2A} receptor binding site upon ligand binding has been useful beyond structure−activity relationships, to understand the potential role of the solvent in the binding kinetics. Our results suggest an interesting qualitative correlation between *unhappy* trapped waters and ligand residence time. In particular the number and the position of *unhappy* waters seem relevant for the ligand k_{off}. High energy trapped solvent molecules in the first hydration shell of the ligand appeared to cause the biggest effect in decreasing the ligand residence time. These conclusions are still speculative and more validation is required, but they open a new research direction that can potentially drive both structure−residence time relationship analysis and *water*-based ligand design methods. While it is also thought that change in ligand residence time may also be due to protein movement, the high similarity of the ligands examined in detail here ensures that this factor should not affect our results to any great extent.

In the second part of this study, ways to incorporate these new computational tools into protocols to predict the free energy of binding of small molecules to the protein target (WNP-MMSA) were investigated. These methods provide an approach to translate the complexity of the protein active site into a representation using a precise water network composed of water molecules with unique characteristics inherited from the different regions of the protein. The outcome of ligand binding on the starting water network was then analyzed in a simple and fast way. In this initial approach modification of the apo-structure water positions has not been allowed; this protocol favors ligands that have a minimum impact on the

dx.doi.org/10.1021/ci4001458 | J. Chem. Inf. Model. 2013, 53, 1700−1713

arrangement on the water network and that are more compatible with the pre-existing solvent structure present in the apo protein. In the test set presented, the method gave computational protocols of comparable speed to MM-GBSA but with better ΔG_{bind} predictions. WNP-MMSA resulted in correlation factors between 0.67 and 0.75, compared to 0.58 of MM-GBSA. It has been possible to extend the applicability of WNP-MMSA method to a more realistic lead optimization medicinal chemistry test set (multiple series, limited structural data) by just replacing the thermodynamics additivity principle with a machine learning approach based on probabilistic classification trees (ctWNP-MMSA). Thermodynamic additivity is the principle that if two components, A and B, contribute independently to some process, then the total change in free energy (or enthalpy or entropy) is the sum of components, $\Delta G = \Delta G_A + \Delta G_B$. However, additivity only applies if components A and B are independent.[75−77] It is difficult to understand if there are "true" independent free energies for example for a hydrogen bond, a salt bridge, or a hydrophobic contact that we could add together to compute binding free energies. The concept of additivity is a fundamental premise that is widely taken for granted. It may be that it is inappropriate to sum free energies no matter what the relative weights and no matter what choices we make for the individual terms. The problem is perhaps additivity itself and the probabilistic classification trees method could represent a valid alternative. This strategy has been tested on a diverse set of 375 A_{2A} antagonists based on the triazine and chromone scaffolds with very promising results: the binding affinity predictions were correct for 90% of the molecules in the training set and 67% of the ligands in the test set. The general applicability of this strategy requires now more testing and validation.

CONCLUSION

In the binding event context, the solvent is as important as the ligand and the protein. Precise water modeling is not only an essential requirement for accurate free energy of binding prediction, but also potentially useful in understanding ligand binding kinetics. In this data set and study, WaterMap, GRID, SZMAP, and GCMC were found to be valuable and complementary tools, that helped us to better understand structure−activity relationships, both qualitatively (e.g., focusing on *unhappy* waters and ligand residence time) and quantitatively (prediction of relative free energies of binding). They can be used to map the unique characteristics of a protein binding site that can quite often be too complex to model correctly with an implicit solvent method. They represent a powerful support for the visual inspection/analysis of docking results. We also believe it is important to take into account in an explicit way the perturbation of the water network resulting from ligand binding. In particular *unhappy* waters trapped between the ligand and the protein seem related to the small molecule residence time and can be used to generate working hypotheses to prioritize the medicinal chemistry efforts in a hit-to-lead or lead optimization program. Ligand binding kinetics prediction still remains a challenge for modern computational chemistry. While these computational approaches are relatively new and more validation is still needed, the promising results described in this study give us a better understanding of the ligand binding event from a molecular viewpoint. In particular further validation in the ability to accurately predict the position of the water molecules can be based on the analysis of high resolution protein crystal structures. The energy evaluation of the water network and its effect on ligand free energy of binding or residence time can be further tested indirectly using SAR information and experimental structural information. The complexity of the ligand binding event may even require the removal of the assumption of the thermodynamics additivity principle, based on the promising results seen when using alternatives like probabilistic classification trees.

METHODS

Molecular Models and Ligand Docking. The 3D coordinates of StaR A_{2A} in complex with the triazine antagonists **4e**[45] (PDB[39] ID 3UZC) and **4g**[45] (PDB ID 3UZA) were used as reference ligand−protein structures. This engineered receptor binds a range of structurally diverse antagonists with a similar affinity to the wild-type receptor,[44] inferring that the use of molecular models based on the X-ray structures should not affect the relative comparison of the ligand free energy of binding predictions;[44] additionally the A_{2A} StaR system was used to determine the SPR measurements from which the kinetics were obtained.[46,47]

The receptor has been prepared with the Protein Preparation Wizard in Maestro:[24] hydrogen atoms have been added and the H-bond network has been optimized through an exhaustive sampling of hydroxyl and thiol moieties, tautomeric and ionic state of His and 180° rotations of the terminal dihedral angle of amide groups of Asp and Gln. The tautomer with the hydrogen in the δ nitrogen has been considered for His278$^{7.43}$ (superscripts refer to Ballesteros−Weinstein numbering[63]). Hydrogen atoms have been energy minimized using the OPLS2005 force field.

Ligands[46,57] were prepared exploiting a Pipeline Pilot[74] protocol including the software MoKa[79,80] and CORINA.[81] The tautomeric states with a minimum abundance threshold of 20% have been generated with TAUTHOR,[80] and the ionic states with a minimum abundance threshold of 10% at pH 7.4 were created with BLABBER.[79] All low energy 3D conformations were created with CORINA[81] including up to 10 ring conformations per molecule in an energy window of ∼2 kcal/mol. For the ligands **4e** and **4g**, the crystallographic binding mode (PDB IDs 3UZC and 3UZA)[45] has been used. For the other 10 triazines included in Figures 2 and 3, the high chemical similarity allows their binding mode to be modeled by molecular alignment to the most similar crystallographic ligand (**4e** or **4g**). For the remaining 321 triazine ligands considered in this study, all the generated tautomeric, ionic states, and ring conformers have been docked using Glide SP.[82,83] The core pattern comparison has been used during the docking. Exploiting this method the common ligand core has been restricted to the most similar reference ligand position with a tolerance of 1.0 Å. For the docking of 39 chromone derivatives the BPM pose of the antagonist **12**[57] in complex with A_{2A} has been used as reference. Three different poses for every ligand have been visually inspected in Vida using the GRID[8] maps generated by the methyl CH3 sp^3 probe (C3), the sp^2 aromatic CH probe (C1=), the water probe (OH2), the neutral planar NH probe (N1), and the sp^2 carbonyl oxygen (O).

Grand Canonical Monte Carlo Simulation. For the GCMC analysis, the protein (PDB ID 3UZC) was prepared by adding polar hydrogen atoms to the structure using WHAT-IF,[84] with nonpolar hydrogens added using LEaP. The protein conformation, including side chain rotamers, protonation states, and tautomers, was identical to those used for the other methods applied in this study. Ligands were parametrized for

the GAFF forcefield[85] using the antechamber module in AMBER, with the partial charges assigned using the AM1-BCC[86] model. To reduce the computational cost, only protein residues that have a heavy atom within 20 Å of the ligands were retained. The complex was solvated by a sphere of TIP4P[87] water molecules of 23 Å radius centered upon the ligand region. The resulting complex was then equilibrated for 10 million moves in the NVT ensemble to remove bad contacts. The amber99 forcefield[88] was used, with a temperature of 25 °C and a nonbonded cutoff of 10 Å. The bond angles and torsions for the side chains of residues within 15 Å of any heavy atom of the ligands and all the bond angles and torsions of the ligand were sampled during the simulation. GCMC simulations[49] were performed upon a 2100 Å^3 region of the binding pocket, incorporating the volume of the ligands and the nearby protein residues. Simulations were performed at six different chemical potentials to observe the changes in water populations as a function of the binding free energy. The dimensionless parameter B was used to control the chemical potential, as defined by Woo et al.[89] The B values used in the simulations of −6, −8, −10, −12, −14, and −16 correspond approximately to a maximum binding free energy of the water molecules observed at these potentials of 0.3, −0.8, −2.0, −3.2, −4.4, and −5.6 kcal/mol, respectively. The value of B can be related to the binding free energy using the expression below:

$$\Delta G_{\text{bind}} = \Delta G_{\text{hyd}} + k_{\text{B}}T(B - \ln \bar{n}) \quad (1)$$

In the above expression, the value of the hydration free energy of water, $\Delta G_{\text{hyd}'}$, is taken to be +6.4 kcal/mol. $\bar{n}$ is the expected number of particles in the system, given the volume of the simulated region and the number density of bulk water.

The GCMC simulations for the water molecules were performed using the interacting particle method as defined by Clark et al.,[62] using a modified version of ProtoMS2.[90] Each B value was simulated for 40 million MC moves, divided into 800 blocks of 50 000 steps each. At the end of each simulation the average water population across the entire simulation was recorded. The same protein structure has been used for all ligand complexes. Insertion and deletion attempts in the binding pocket were attempted with a probability of 3%, protein moves with a probability of 7%, solute moves with a probability of 1% with the remaining moves sampling the bulk solvent and the waters within the defined binding cavity. Density grids were created by analyzing the oxygen atom coordinates of all water molecules within the region of interest across 800 equally spaced simulation snapshots. This data was used to create a density map across a grid with 1 Å spacing, which was then normalized according to the most populated grid point. The binding free energy of a particular water was estimated from the B value at which it was present in fewer than 30% of the sampled configurations.

Water Network Perturbation Scoring and Prediction. A water network was created for the 12 triazines shown in Figures 2 and 3 with the following optimized computational protocol:

(A) For the binding site for every selected ligand binding pose a grid of 0.5 Å spacing was created around the small molecule.
(B) SZMAP[48] was used to analyze every point in the grid evaluating the effect of the protein and the antagonist.
(C) A Python script was used to post process the SZMAP output and create a water network selecting water positions starting from the most stable and going to the less favorable. All the water molecules to be accepted had to be at a distance of more than 2.4 Å from already selected water positions. This value corresponds to the minimum distance between the oxygen atoms in two water molecules found in the GPCR crystal structures solved until now.
(D) The hydrogens of this binding site water network were optimized in the presence of the protein−ligand complex in Maestro[91] using the Protein Preparation Wizard.
(E) The final system was optimized using a customized WaterMap protocol. WaterMap calculations[15,16] consist of an all atom explicit solvent molecular dynamics simulation followed by a statistical thermodynamic analysis of water clusters (hydration sites). For the WaterMap calculations presented here, the system from step D was not truncated and solvated in a TIP4P water box extending at least 5 Å in all directions. The "solvate_pocket" step for the binding site that uses an embedded GCMC method was excluded, being replaced by steps (A−D) using SZMAP. A restraint was applied to the protein heavy atoms and the system was relaxed with an initial minimization followed by a short molecular dynamics simulation heating the system from 10 to 300 K. A final preproduction simulation of 120 ps was run at 300 K. The production simulation was run for 2 ns at 300 K in the NTP ensemble. WaterMap statistical thermodynamic analysis of hydration sites were carried out in a region of 10 Å around the ligands.
(F) The free energy of every water molecule was additionally evaluated using SZMAP in the context of the protein−ligand complex.
(G) The nearest water molecules to the ligand (first hydration shell) were also rescored using the GCMC method. The GCMC simulation results were analyzed in Discovery Studio.[92] The free energy of the waters has been estimated by comparing the position of the predicted waters (at the end of step E) to the GCMC water density grids. The lowest B value resulting in water density (at 0.3 occupancy) for the considered water has been selected. The ΔG_{bind} of the water was estimated as the free energy corresponding to that B value. Water molecules resulting in a location without water density at $B = -6$ (0.3 occupancy) were considered to have a ΔG greater than 0.3 kcal/mol. Water molecules detected in all grids including the water density grid at $B = 16$ (0.3 occupancy) were considered to have a ΔG lower than −5.6 kcal/mol.

Ligand Free Energy of Binding Predictions. The molecular mechanics generalized Born surface area (MM-GBSA) method[65,66] considers only the bound and unbound states of the ligand−protein complex without taking into account the intermediate states. The free energy of binding is therefore estimated as

$$\Delta G_{\text{bind}} = G_{\text{complex}} - (G_{\text{ligand}} + G_{\text{protein}}) \quad (2)$$

The individual energy terms are decomposed into a gas phase component (G_{gas}) calculated using the force field and a solvation energy term ($G_{\text{solvation}}$). For ligands, an entropic contribution can be included ($\text{TS}_{\text{ligand}}$):

$$G_{\text{ligand}} = G_{\text{gas}} + G_{\text{solvation}} - \text{TS}_{\text{ligand}} \quad (3)$$

dx.doi.org/10.1021/ci4001458 | J. Chem. Inf. Model. 2013, 53, 1700−1713

The electrostatic contribution to the free energy of solvation is evaluated using an implicit solvent model: in this case the generalized Born equation (G_{GB}). The hydrophobic contribution to the free energy of solvation is taken into account by evaluating the solvent accessible surface area (G_{SA}) of the molecule.

$$G_{solvation} = G_{GB} + G_{SA} \quad (4)$$

In this study the vibrational entropy has not been taken into account for the ligands because of the high computational cost and the risk of producing large errors. We used a validated protocol based on the MM-GBSA application in Prime[93] using OPLS2005 and energy minimization of the ligands.[72]

We evaluated the possibility of replacing the generalized Born implicit solvent with a water network perturbation (WNP) term. This energy term is based on the estimation of the effect of ligand binding on an explicit water network in the protein active site. We therefore quantify the WNP as follows:

$$\Delta G_{WNP} = \Delta G_{perturbedwaters} - \sum G_{displacedwater} \quad (5)$$

$$\Delta G_{perturbedwaters} = \sum (G_{water\_complex} - G_{water\_apo}) \quad (6)$$

where ΔG_{WNP} is the water network perturbation energy term, $G_{displacedwater}$ are the individual free energies of binding of the water molecules displaced by the ligand binding, and $\Delta G_{perturbedwaters}$ is the difference in free energy of binding of the remaining water molecules with and without the ligand. The computational methods used to evaluate the free energy of binding of the waters are described below (Water Network Scoring).

Water Network Creation. In this study we tested the suitability of creating a water network in the apo active site with three different methods: WaterMap, GRID, and GCMC simulations. (A) In WaterMap, the default protocol used included 100 000 steps of grand canonical Monte Carlo simulation to populate the binding site with water molecules and a 2 ns all-atom explicit solvent molecular dynamics simulation followed by water clustering. (B) In GRID the water probe (OH2) was used to evaluate a grid of 0.5 Å spacing in the active site and a Python script was used to create a water network selecting water positions starting from the most stable and going to the least favorable, as mentioned previously. (C) GCMC simulation for the apo protein was applied as described above in the Grand Canonical Monte Carlo Simulation section. The results were analyzed in Discovery Studio with the water network being created from analysis of the water density grids at B values of −6, −8, −10, −12, −14, and −16, starting from the waters with higher occupancy and lower B values. Only nonclashing water molecules with at least 0.3 occupancy were included.

Water Network Scoring. The energy of every water molecule in the apo state and in the presence of the ligands was evaluated with SZMAP and a novel optimized GRID probe, CRY. This apolar/hydrophobic molecular probe is based on the combination of the DRY and the sp^2 aromatic CH (C1=) probes. To create a protocol comparable in speed to the described MM-GBSA, the calculation of the water network is performed only on the pseudoapo structure, removing the ligand. Water positions were not optimized for every ligand−protein complex. The energy of every water molecule was evaluated in the same position in the apo structure and in the presence of every ligand. Displaced waters (clashing with the ligand) are reported by SZMAP and result in unnaturally high energy in GRID.

A range of combinations of methods for the prediction of the ligand free energy of binding were examined. Only WNP-MMSA implementations resulting in better-than-random predictions are shown in Figure 4. They all include the MMSA term as calculated in MM-GBSA described above. The general nomenclature for the methods used in this work includes, the initial name of the method for the water network creation, followed by the method for water scoring, and finally the term MMSA: WM/SZMAP-MMSA, GCMC/SZMAP-MMSA, WM/CRY-MMSA, GCMC/CRY-MMSA, GRID-OH2/CRY-MMSA. For example WM/SZMAP-MMSA corresponds to a protocol using WaterMap to generate the water network, which is scored with SZMAP and combined with the molecular mechanics force field and the solvent accessible surface area. Similarly, GRID-OH2/CRY-MMSA corresponds to a protocol using the GRID probe OH2 to generate the water network which is scored with the CRY probe and finally combined with MMSA term.

These methods have been compared with the WM/MM method.[16] Briefly this approach incorporates energetic terms from an MM-GBSA calculation, such as protein−ligand van der Waals contacts, electrostatic interactions, ligand desolvation, and internal strain (ligand and protein) with the full WaterMap. WaterMap molecular dynamics simulation is followed by a statistical thermodynamic analysis of hydration sites to compute their enthalpy, entropy, and free energy relative to bulk water. A final preproduction simulation of 120 ps at 300 K is followed by a production simulation of 2 ns at 300 K in the NTP ensemble. In WaterMap, the excess entropy is computed by numerically integrating a local expansion of spatial and orientational correlation functions.[15] The enthalpy is computed by averaging the molecular mechanics energies of the water molecules in each hydration site over all frames of the molecule dynamics simulation. The default options have been applied using a simple minimization of the ligand only.

Probabilistic Classification Trees Based Free Energy of Binding Predictions. All the free energy of binding prediction methods described are parametrized based on the additivity approximation.[75] We evaluated the possibility of substituting the linear combination method with a machine learning approach based on probabilistic classification trees[58] (ctWNP-MMSA). We applied the novel protocol to GRID-OH2/CRY-MMSA, the best performing method tested in the preceding study (r^2 = 0.75). We analyzed the novel scoring function, denominated ctGRID-OH2/CRY-MMSA, using 375 A_{2A} antagonists divided into two groups based on their activity: weak (pK_i 6−7.5) and active (pK_i 7.5−9). The ligands were prepared and docked as described above. One pose per ligand has been selected after visual inspection. We used Canvas[78] to create a probabilistic classifier based on 100 decisional trees, with a minimum leaf size of five compounds. The splitting was based on the *information gain* algorithm. Molecules have been divided randomly into two groups of equal size: a training set and a test set. To solve a bug in the current version of Canvas, molecules have been sorted including first the training set and then the test set. Five properties have been used for the model creation: CRY total energy of the displaced waters, CRY total water perturbation energy, the change upon binding of solvent accessible surface (ΔSA), Coulomb interaction energy (ΔCoulomb), and van der Waals interaction energy (ΔvdW).

dx.doi.org/10.1021/ci4001458 | J. Chem. Inf. Model. 2013, 53, 1700−1713

AUTHOR INFORMATION

Corresponding Author

*Phone: +44(0)1707 358646. Fax: +44(0)1707 358640. E-mail: andrea.bortolato@heptares.com.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The authors want to thank Francesca Deflorian, Ulf Norinder, and Fiona Marshall for critical reading of the manuscript, Simon Cross and Massimo Baroni for support on the CRY probe development, and Thijs Beuming, Robert Abel, and Mike Word for useful discussions.

REFERENCES

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

(2) Clarke, C.; Woods, R. J.; Gluska, J.; Cooper, A.; Margaret, A.; Boons, G. J. Involvement of water in carbohydrate-protein binding. *J. Am. Chem. Soc.* **2001**, *123* (49), 12238−12247.

(3) Lam, P. Y.; 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. Rational design of potent, bioavailable, nonpeptide cyclic ureas as HIV protease inhibitors. *Science* **1994**, *263* (5145), 380−384.

(4) de Beer, S. B.; Vermeulen, N. P.; Oostenbrink, C. The role of water molecules in computational drug design. *Curr. Top. Med. Chem.* **2010**, *10* (1), 55−66.

(5) Mancera, R. L. Molecular modeling of hydration in drug design. *Curr. Opin. Drug Discov. Devel.* **2007**, *10* (3), 275−280.

(6) Wong, S. E.; Lightstone, F. C. Accounting for water molecules in drug design. *Expert. Opin. Drug Discov.* **2011**, *6* (1), 65−74.

(7) Battistutta, R.; Mazzorana, M.; Cendron, L.; Bortolato, A.; Sarno, S.; Kazimierczuk, Z.; Zanotti, G.; Moro, S.; Pinna, L. A. The ATP-binding site of protein kinase CK2 holds a positive electrostatic area and conserved water molecules. *Chembiochem.* **2007**, *8* (15), 1804−1809.

(8) Goodford, P. J. A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. *J. Med. Chem.* **1985**, *28* (7), 849−857.

(9) Guarnieri, F.; Mezei, M. Simulated Annealing of Chemical-Potential-A General Procedure for Locating Bound Waters-Application to the Study of the Differential Hydration Propensities of the Major and Minor Grooves of DNA. *J. Am. Chem. Soc.* **1996**, *118* (35), 8493−8494.

(10) Michel, J.; Tirado-Rives, J.; Jorgensen, W. L. Prediction of the water content in protein binding sites. *J. Phys. Chem. B* **2009**, *113* (40), 13337−13346.

(11) Barillari, C.; Taylor, J.; Viner, R.; Essex, J. W. Classification of water molecules in protein binding sites. *J. Am. Chem. Soc.* **2007**, *129* (9), 2577−2587.

(12) Imai, T.; Hiraoka, R.; Kovalenko, A.; Hirata, F. Locating missing water molecules in protein cavities by the three-dimensional reference interaction site model theory of molecular solvation. *Proteins* **2007**, *66* (4), 804−813.

(13) Rossato, G.; Ernst, B.; Vedani, A.; Smiesko, M. AcquaAlta: a directional approach to the solvation of ligand-protein complexes. *J. Chem. Inf. Model.* **2011**, *51* (8), 1867−1881.

(14) Ross, G. A.; Morris, G. M.; Biggin, P. C. Rapid and accurate prediction and scoring of water molecules in protein binding sites. *PLoS. One.* **2012**, *7* (3), e32036.

(15) Abel, R.; Young, T.; Farid, R.; Berne, B. J.; Friesner, R. A. Role of the active-site solvent in the thermodynamics of factor Xa ligand binding. *J. Am. Chem. Soc.* **2008**, *130* (9), 2817−2831.

(16) Abel, R.; Salam, N. K.; Shelley, J.; Farid, R.; Friesner, R. A.; Sherman, W. Contribution of explicit solvent effects to the binding affinity of small-molecule inhibitors in blood coagulation factor serine proteases. *ChemMedChem.* **2011**, *6* (6), 1049−1066.

(17) Guimaraes, C. R.; Mathiowetz, A. M. Addressing limitations with the MM-GB/SA scoring procedure using the WaterMap method and free energy perturbation calculations. *J. Chem. Inf. Model.* **2010**, *50* (4), 547−559.

(18) Higgs, C.; Beuming, T.; Sherman, W. Hydration Site Thermodynamics Explain SARs for Triazolylpurines Analogues Binding to the A2A Receptor. *ACS Med. Chem. Lett.* **2010**, *1* (4), 160−164.

(19) Laha, J. K.; Zhang, X.; Qiao, L.; Liu, M.; Chatterjee, S.; Robinson, S.; Kosik, K. S.; Cuny, G. D. Structure-activity relationship study of 2,4-diaminothiazoles as Cdk5/p25 kinase inhibitors. *Bioorg. Med. Chem. Lett.* **2011**, *21* (7), 2098−2101.

(20) Mason, J. S.; Bortolato, A.; Congreve, M.; Marshall, F. H. New insights from structural biology into the druggability of G protein-coupled receptors. *Trends Pharmacol. Sci.* **2012**, *33* (5), 249−260.

(21) Pearlstein, R. A.; Hu, Q. Y.; Zhou, J.; Yowe, D.; Levell, J.; Dale, B.; Kaushik, V. K.; Daniels, D.; Hanrahan, S.; Sherman, W.; Abel, R. New hypotheses about the structure-function of proprotein convertase subtilisin/kexin type 9: analysis of the epidermal growth factor-like repeat A docking site using WaterMap. *Proteins* **2010**, *78* (12), 2571−2586.

(22) Repasky, M. P.; Murphy, R. B.; Banks, J. L.; Greenwood, J. R.; Tubert-Brohman, I.; Bhat, S.; Friesner, R. A. Docking performance of the glide program as evaluated on the Astex and DUD datasets: a complete set of glide SP results and selected results for a new scoring function integrating WaterMap and glide. *J. Comput. Aided Mol. Des.* **2012**, *26* (6), 787−799.

(23) Wallnoefer, H. G.; Liedl, K. R.; Fox, T. A GRID-derived water network stabilizes molecular dynamics computer simulations of a protease. *J. Chem. Inf. Model.* **2011**, *51* (11), 2860−2867.

(24) Wang, L.; Berne, B. J.; Friesner, R. A. Ligand binding to protein-binding pockets with wet and dry regions. *Proc. Natl. Acad. Sci. U.S.A.* **2011**, *108* (4), 1326−1330.

(25) Sindhikara, D. J.; Yoshida, N.; Hirata, F. Placevent: an algorithm for prediction of explicit solvent atom distribution-application to HIV-1 protease and F-ATP synthase. *J. Comput. Chem.* **2012**, *33* (18), 1536−1543.

(26) Bren, U.; Janezic, D. Individual degrees of freedom and the solvation properties of water. *J. Chem. Phys.* **2012**, *137* (2), 024108.

(27) Pearlstein, R. A.; Sherman, W.; Abel, R. Contributions of water transfer energy to protein-ligand association and dissociation barriers: WaterMap analysis of a series of p38alpha MAP kinase inhibitors. *Proteins* **2013**, DOI: 10.1002/prot.24276.

(28) Haga, K.; Kruse, A. C.; Asada, H.; Yurugi-Kobayashi, T.; Shiroishi, M.; Zhang, C.; Weis, W. I.; Okada, T.; Kobilka, B. K.; Haga, T.; Kobayashi, T. Structure of the human M2 muscarinic acetylcholine receptor bound to an antagonist. *Nature* **2012**, *482* (7386), 547−551.

(29) Liu, W.; Chun, E.; Thompson, A. A.; Chubukov, P.; Xu, F.; Katritch, V.; Han, G. W.; Roth, C. B.; Heitman, L. H.; Ijzerman, A. P.; Cherezov, V.; Stevens, R. C. Structural basis for allosteric regulation of GPCRs by sodium ions. *Science* **2012**, *337* (6091), 232−236.

(30) Pardo, L.; Deupi, X.; Dolker, N.; Lopez-Rodriguez, M. L.; Campillo, M. The role of internal water molecules in the structure and function of the rhodopsin family of G protein-coupled receptors. *Chembiochem.* **2007**, *8* (1), 19−24.

(31) Angel, T. E.; Gupta, S.; Jastrzebska, B.; Palczewski, K.; Chance, M. R. Structural waters define a functional channel mediating activation of the GPCR, rhodopsin. *Proc. Natl. Acad. Sci. U.S.A.* **2009**, *106* (34), 14367−14372.

(32) Jastrzebska, B.; Palczewski, K.; Golczak, M. Role of bulk water in hydrolysis of the rhodopsin chromophore. *J. Biol. Chem.* **2011**, *286* (21), 18930−18937.

(33) Congreve, M.; Langmead, C. J.; Mason, J. S.; Marshall, F. H. Progress in structure based drug design for g protein-coupled receptors. *J. Med. Chem.* **2011**, *54* (13), 4283−4311.

(34) Lagerstrom, M. C.; Schioth, H. B. Structural diversity of G protein-coupled receptors and significance for drug discovery. *Nat. Rev. Drug Discov.* **2008**, *7* (4), 339−357.

(35) Hasko, G.; Linden, J.; Cronstein, B.; Pacher, P. Adenosine receptors: therapeutic aspects for inflammatory and immune diseases. *Nat. Rev. Drug Discov.* **2008**, *7* (9), 759−770.
(36) Jacobson, K. A.; Gao, Z. G. Adenosine receptors as therapeutic targets. *Nat. Rev. Drug Discov.* **2006**, *5* (3), 247−264.
(37) Lopes, L. V.; Sebastiao, A. M.; Ribeiro, J. A. Adenosine and related drugs in brain diseases: present and future in clinical trials. *Curr. Top. Med. Chem.* **2011**, *11* (8), 1087−1101.
(38) Hauser, R. A.; Cantillon, M.; Pourcher, E.; Micheli, F.; Mok, V.; Onofrj, M.; Huyck, S.; Wolski, K. Preladenant in patients with Parkinson's disease and motor fluctuations: a phase 2, double-blind, randomised trial. *Lancet Neurol.* **2011**, *10* (3), 221−229.
(39) 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* (1), 235−242.
(40) Jaakola, V. P.; Griffith, M. T.; Hanson, M. A.; Cherezov, V.; Chien, E. Y.; Lane, J. R.; Ijzerman, A. P.; Stevens, R. C. The 2.6 angstrom crystal structure of a human A2A adenosine receptor bound to an antagonist. *Science* **2008**, *322* (5905), 1211−1217.
(41) Xu, F.; Wu, H.; Katritch, V.; Han, G. W.; Jacobson, K. A.; Gao, Z. G.; Cherezov, V.; Stevens, R. C. Structure of an agonist-bound human A2A adenosine receptor. *Science* **2011**, *332* (6027), 322−327.
(42) Hino, T.; Arakawa, T.; Iwanari, H.; Yurugi-Kobayashi, T.; Ikeda-Suno, C.; Nakada-Nakura, Y.; Kusano-Arai, O.; Weyand, S.; Shimamura, T.; Nomura, N.; Cameron, A. D.; Kobayashi, T.; Hamakubo, T.; Iwata, S.; Murata, T. G-protein-coupled receptor inactivation by an allosteric inverse-agonist antibody. *Nature* **2012**, *482* (7384), 237−240.
(43) Lebon, G.; Bennett, K.; Jazayeri, A.; Tate, C. G. Thermostabilisation of an agonist-bound conformation of the human adenosine A(2A) receptor. *J. Mol. Biol.* **2011**, *409* (3), 298−310.
(44) Dore, A. S.; Robertson, N.; Errey, J. C.; Ng, I.; Hollenstein, K.; Tehan, B.; Hurrell, E.; Bennett, K.; Congreve, M.; Magnani, F.; Tate, C. G.; Weir, M.; Marshall, F. H. Structure of the adenosine A(2A) receptor in complex with ZM241385 and the xanthines XAC and caffeine. *Structure* **2011**, *19* (9), 1283−1293.
(45) Lebon, G.; Warne, T.; Edwards, P. C.; Bennett, K.; Langmead, C. J.; Leslie, A. G.; Tate, C. G. Agonist-bound adenosine A2A receptor structures reveal common features of GPCR activation. *Nature* **2011**, *474* (7352), 521−525.
(46) Congreve, M.; Andrews, S. P.; Dore, A. S.; Hollenstein, K.; Hurrell, E.; Langmead, C. J.; Mason, J. S.; Ng, I. W.; Tehan, B.; Zhukov, A.; Weir, M.; Marshall, F. H. Discovery of 1,2,4-triazine derivatives as adenosine A(2A) antagonists using structure based drug design. *J. Med. Chem.* **2012**, *55* (5), 1898−1903.
(47) Zhukov, A.; Andrews, S. P.; Errey, J. C.; Robertson, N.; Tehan, B.; Mason, J. S.; Marshall, F. H.; Weir, M.; Congreve, M. Biophysical mapping of the adenosine A2A receptor. *J. Med. Chem.* **2011**, *54* (13), 4312−4323.
(48) *SZMAP*, version 1. 0; Openeye Scientific Software: Santa Fe, NM, 2011.
(49) Adams, D. J. Chemical potential of hard-sphere fluids by Monte Carlo methods. *Mol. Phys.* **1974**, *28* (5), 1241−1252.
(50) Sciabola, S.; Stanton, R. V.; Mills, J. E.; Flocco, M. M.; Baroni, M.; Cruciani, G.; Perruccio, F.; Mason, J. S. High-throughput virtual screening of proteins using GRID molecular interaction fields. *J. Chem. Inf. Model.* **2010**, *50* (1), 155−169.
(51) Baroni, M.; Cruciani, G.; Sciabola, S.; Perruccio, F.; Mason, J. S. A common reference framework for analyzing/comparing proteins and ligands. Fingerprints for Ligands and Proteins (FLAP): theory and application. *J. Chem. Inf. Model.* **2007**, *47* (2), 279−294.
(52) Tummino, P. J.; Copeland, R. A. Residence time of receptor-ligand complexes and its effect on biological function. *Biochemistry* **2008**, *47* (20), 5481−5492.
(53) Copeland, R. A.; Pompliano, D. L.; Meek, T. D. Drug-target residence time and its implications for lead optimization. *Nat. Rev. Drug Discov.* **2006**, *5* (9), 730−739.
(54) Pan, A. C.; Borhani, D. W.; Dror, R. O.; Shaw, D. E. Molecular determinants of drug-receptor binding kinetics. *Drug Discov. Today* **2013**, DOI: 10.1016/j.drudis.2013.02.007.
(55) Copeland, R. A. Conformational adaptation in drug-target interactions and residence time. *Future. Med. Chem.* **2011**, *3* (12), 1491−1501.
(56) Copeland, R. A. The dynamics of drug-target interactions: drug-target residence time and its impact on efficacy and safety. *Expert. Opin. Drug Discov.* **2010**, *5* (4), 305−310.
(57) Langmead, C. J.; Andrews, S. P.; Congreve, M.; Errey, J. C.; Hurrell, E.; Marshall, F. H.; Mason, J. S.; Richardson, C. M.; Robertson, N.; Zhukov, A.; Weir, M. Identification of novel adenosine A(2A) receptor antagonists by virtual screening. *J. Med. Chem.* **2012**, *55* (5), 1904−1909.
(58) Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45* (1), 5−32.
(59) Chen, X. W.; Liu, M. Prediction of protein-protein interactions using random decision forest framework. *Bioinformatics* **2005**, *21* (24), 4394−4400.
(60) Sato, T.; Honma, T.; Yokoyama, S. Combining machine learning and pharmacophore-based interaction fingerprint for in silico screening. *J. Chem. Inf. Model.* **2010**, *50* (1), 170−185.
(61) Ballester, P. J.; Mitchell, J. B. A machine learning approach to predicting protein-ligand binding affinity with applications to molecular docking. *Bioinformatics* **2010**, *26* (9), 1169−1175.
(62) Clark, M.; Meshkat, S.; Talbot, G. T.; Carnevali, P.; Wiseman, J. S. Fragment-based computation of binding free energies by systematic sampling. *J. Chem. Inf. Model.* **2009**, *49* (8), 1901−1913.
(63) Ballesteros, J. A.; Weinstein, H. Integrated methods for the construction of three dimensional models and computational probing of structure function relations in G protein-coupled receptors. *Methods Neurosci.* **1995**, *25*, 366−428.
(64) Schmidtke, P.; Luque, F. J.; Murray, J. B.; Barril, X. Shielded hydrogen bonds as structural determinants of binding kinetics: application in drug design. *J. Am. Chem. Soc.* **2011**, *133* (46), 18903−18910.
(65) Kollman, P. A.; Massova, I.; Reyes, C.; Kuhn, B.; Huo, S.; Chong, L.; Lee, M.; Lee, T.; Duan, Y.; Wang, W.; Donini, O.; Cieplak, P.; Srinivasan, J.; Case, D. A.; Cheatham, T. E., III. Calculating structures and free energies of complex molecules: combining molecular mechanics and continuum models. *Acc. Chem. Res.* **2000**, *33* (12), 889−897.
(66) Gohlke, H.; Case, D. A. Converging free energy estimates: MM-PB(GB)SA studies on the protein-protein complex Ras-Raf. *J. Comput. Chem.* **2004**, *25* (2), 238−250.
(67) Aqvist, J.; Medina, C.; Samuelsson, J. E. A new method for predicting binding affinity in computer-aided drug design. *Protein Eng.* **1994**, *7* (3), 385−391.
(68) Bren, U.; Oostenbrink, C. Cytochrome P450 3A4 inhibition by ketoconazole: tackling the problem of ligand cooperativity using molecular dynamics simulations and free-energy calculations. *J. Chem. Inf. Model.* **2012**, *52* (6), 1573−1582.
(69) Bortolato, A.; Moro, S. In silico binding free energy predictability by using the linear interaction energy (LIE) method: bromobenzimidazole CK2 inhibitors as a case study. *J. Chem. Inf. Model.* **2007**, *47* (2), 572−582.
(70) Lee, F. S.; Chu, Z. T.; Bolger, M. B.; Warshel, A. Calculations of antibody-antigen interactions: microscopic and semi-microscopic evaluation of the free energies of binding of phosphorylcholine analogs to McPC603. *Protein Eng.* **1992**, *5* (3), 215−228.
(71) Bren, U.; Lah, J.; Bren, M.; Martinek, V.; Florian, J. DNA duplex stability: the role of preorganized electrostatics. *J. Phys. Chem. B* **2010**, *114* (8), 2876−2885.
(72) Lyne, P. D.; Lamb, M. L.; Saeh, J. C. Accurate prediction of the relative potencies of members of a series of kinase inhibitors using molecular docking and MM-GBSA scoring. *J. Med. Chem.* **2006**, *49* (16), 4805−4808.
(73) Hansch, C.; Leo, A. *Substituent constants for correlation analysis in chemistry and biology*; John Wiley: New York, 1979.
(74) *Pipeline Pilot*, version 8. 5; Accelrys: San Diego, CA, 2011.

(75) Dill, K. A. Additivity principles in biochemistry. *J. Biol. Chem.* **1997**, *272* (2), 701−704.
(76) Bren, U.; Martinek, V.; Florian, J. Decomposition of the solvation free energies of deoxyribonucleoside triphosphates using the free energy perturbation method. *J. Phys. Chem. B* **2006**, *110* (25), 12782−12788.
(77) Bren, M.; Florian, J.; Mavri, J.; Bren, U. Do all pieces make a whole? Thiele cumulants and the free energy decomposition. *Theor. Chem. Acc.* **2007**, *117* (4), 535−540.
(78) *Canvas*, version 1. 5; Schrödinger: New York, 2012.
(79) Milletti, F.; Storchi, L.; Sforna, G.; Cruciani, G. New and original pKa prediction method using grid molecular interaction fields. *J. Chem. Inf. Model.* **2007**, *47* (6), 2172−2181.
(80) Milletti, F.; Storchi, L.; Sforna, G.; Cross, S.; Cruciani, G. Tautomer enumeration and stability prediction for virtual screening on large chemical databases. *J. Chem. Inf. Model.* **2009**, *49* (1), 68−75.
(81) *Corina*, version 2. 1; Molecular Networks: Erlanger, Germany, 2011.
(82) 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* (7), 1739−1749.
(83) *Glide*, version 5. 7; Schrödinger: New York, 2011.
(84) Vriend, G. WHAT IF: a molecular modeling and drug design program. *J. Mol. Graph.* **1990**, *8* (1), 52−6−29.
(85) Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. Development and testing of a general amber force field. *J. Comput. Chem.* **2004**, *25* (9), 1157−1174.
(86) Jakalian, A.; Bush, B. L.; Jack, D. B.; Bayly, C. I. Fast, efficient generation of high quality atomic Charges. AM1−BCC model: I. Method. *J. Comput. Chem.* **2000**, *21* (2), 132−146.
(87) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **1983**, *79*, 926.
(88) Wang, J.; Cieplak, P.; Kollman, P. A. How well does a restrained electrostatic potential (RESP) model perform in calculating conformational energies of organic and biological molecules? *J. Comput. Chem.* **2000**, *21* (12), 1049−1074.
(89) Woo, H. J.; Dinner, A. R.; Roux, B. Grand canonical Monte Carlo simulations of water in protein environments. *J. Chem. Phys.* **2004**, *121*, 6392.
(90) Woods, C. J.; Michel, J. *ProtoMS2*, 2007.
(91) *Maestro*, version 9.2; Schrödinger: New York, 2011.
(92) *Discovery Studio*, version 3.1; Accelrys: San Diego, CA, 2011.
(93) *Prime*, version 3.0; Schrödinger: New York, 2011.