JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

Application Note

GridSolvate: A Web Server for the Prediction of Biomolecular Hydration Properties

Piotr Setny*

Cite This: *J. Chem. Inf. Model.* 2020, 60, 5907−5911

Read Online

ACCESS | Metrics & More | Article Recommendations | SI Supporting Information

ABSTRACT: We present a novel web server, named gridSolvate, dedicated to the prediction of biomolecular hydration properties. Given a solute in atomic representation, such as a protein or protein−ligand complex, the server determines positions and excess chemical potential of buried and first hydration shell water molecules. Calculations are based on our semiexplicit hydration model that provides computational efficiency close to implicit solvent approaches, yet captures a number of physical effects unique to explicit solvent representation. The model was introduced and validated before in the context of bulk hydration of drug-like solutes and determination of protein hydration sites. Current methodological developments merge those two avenues into a single, easily accessible tool. Here, we focus on the server's ability to predict water distribution and affinity within protein−ligand interfaces. We demonstrate that with possibly minimal user intervention the server correctly predicts the locations of 77% of interface water molecules in an external set of test structures. The server is freely available at https://gsolvate.biomod.cent.uw.edu.pl.
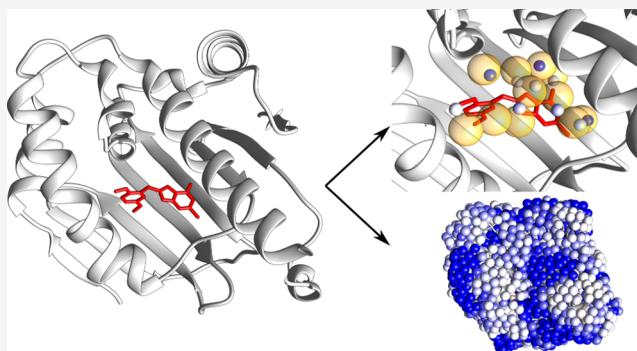
INTRODUCTION

The interplay between the aqueous environment and living matter has been actively studied for decades.[1] The role of hydration effects in shaping biomolecular structures, dynamics, and interactions is now well recognized, as is the importance of their quantitative treatment in practical applications, for instance, in drug design.[2] Computational methods offer increasingly reliable description of macromolecular hydration.[3] A wide array of available approaches starts with explicit solvent simulations typically combined with various postprocessing techniques aimed at general description of solvent thermodynamic properties,[4,5] estimation of water binding free energy to specific sites,[6,7] bound water detection,[8] or scoring.[9−11] Computationally less demanding, yet also less accurate, are implicit solvent models.[12] Finally, there are numerous specialized algorithms designed for description of surface hydration,[13−15] water annotation in crystal structures,[16−18] or its placement and scoring,[19−22] also in the context of protein−ligand docking.[23−26]

Recently, we developed a semiexplicit solvent model, applicable to biological macromolecules (proteins, nucleic acids), drug-like compounds, and their complexes,[27−29] that provides information concerning locations of buried and first hydration shell water molecules together with estimates of their excess chemical potentials. The calculations are much faster than explicit solvent simulations, yet still capture a number of physical effects that are neglected in typical simplified approaches: directionality of water hydrogen bonds, entropic penalty due to

limited rotational freedom of bound water, or the asymmetry of charge distribution within water molecule. In particular, in comparison to most other grid-based approaches for typing protein hydration sites, the model captures mutual interactions within clusters of buried water molecules. For a detailed discussion, we direct the reader to our previous reports concerning theoretical model assumptions,[27] estimation of hydration free energies of drug-like compounds,[28] and prediction of protein internal hydration sites.[29] Here, we present a gridSolvate web server that allows automatized calculations for protein−ligand complexes and makes the method available for external users.

METHODS

Hydration Model. The server relies on our semiexplicit hydration model.[27−29] Briefly (see Supporting Information, SI, for details), it combines an atomistic description of solute(s)−water interactions with mean field treatment of intrasolvent interactions. As an input, it requires a structure of the system to be hydrated with nonbonded force field parameters assigned to

each atom. The aqueous environment is represented by a discrete lattice whose nodes serve to map spatial water distribution and excess chemical potential. The latter is defined at each point, **r**, by an effective Hamiltonian of a rotatable, atomistic water probe

$$H_{eff}(\mathbf{r}, \{n\}) = -k_{\mathrm{B}}T \, \ln\!\left(\sum_{\theta} e^{-\beta H_{UV}(\mathbf{r},\boldsymbol{\theta}) - \beta H_{VV}(\mathbf{r},\boldsymbol{\theta},\{n\})}\right) \quad (1)$$

where $\{n\}$ denotes an instantaneous distribution of occupied and empty lattice points, $H_{UV}(\mathbf{r},\theta)$ is a solute−solvent interaction energy combining electrostatic and Lennard-Jones (LJ) contributions, $H_{VV}(\mathbf{r},\theta,\{n\})$ is a mean-field solvent−solvent term, $k_{\mathrm{B}}T = \beta^{-1}$ are the Boltzmann constant times temperature (300 K), and the summation extends over 12 probe orientations, θ. It is assumed than any lattice point whose H_{eff} is greater than a certain threshold value becomes vacated. A stationary solvent distribution in the presence of a solute is reached iteratively, using a fast, cellular automata-based algorithm.[27] The resulting spatial map of excess water chemical potential can be used to assess local hydration propensity of a solute surface or can be partitioned into the most probable locations of individual water molecules.

GridSolvate Server. The server processes input structures, handles hydration calculations, and returns results on a dedicated web page or through an email (Figure 1).
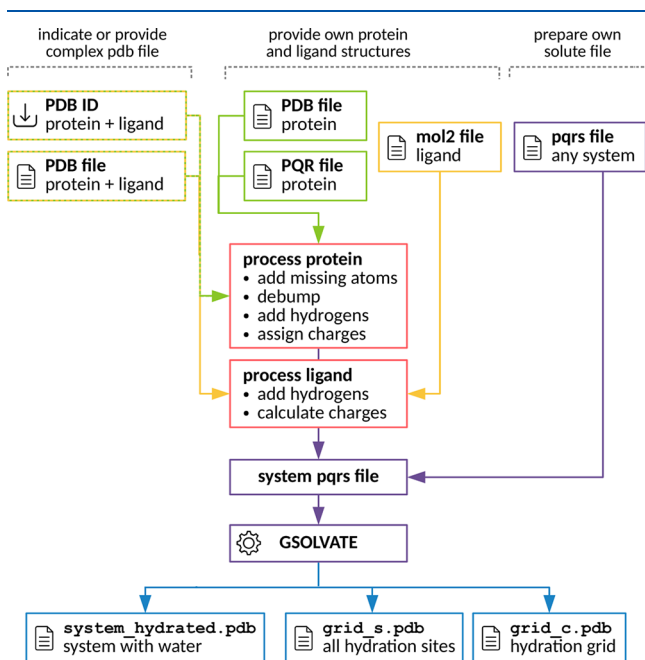

Figure 1. Workflow of the GridSolvate server.

The Input. The following input options are available:

- PDB[30] id for download. The structure can include a protein or protein−ligand complex. The chains that should be considered for calculations can be specified by the user, or by default the first chain found will be processed. Ligand structure can be selected for automatic processing (see below) by providing a ligand residue name. Crystallographic water oxygen atoms can be left within the system in order to be included in the calculation.

- PDB file directly uploaded to the server. Such a file will be processed with analogous options as an automatically downloaded PDB file.

- PQR file containing a structure prepared with the use of the PDB2PQR program.[31] Such an option gives the possibility to prepare custom protonation states and desired positions of hydrogen atoms.

- MOL2 file, containing a separate ligand structure. It can already include hydrogen atoms and partial atomic charges, or they can be calculated by the server.

- PQRS file with an arbitrary system of interest. The file adopts an expanded PQR format, with one additional column containing the atomic LJ potential well depth (ϵ parameter) in kcal/mol (detailed description in the SI). A PQRS file does not require any processing before hydration calculations and can describe just a part of a macromolecule (e.g., binding site region) or any artificial atomic assembly. It can be manually created by the user, or one can modify and reuse a file previously generated by the server.

An additional parameter is the number of hydration calculations, N, that are performed in order to obtain the final statistics. Typically, the convergence is achieved for $100 < N < 1000$, as discussed before.[29]

Job Processing. Protein structures are analyzed with the use of the PDB2PQR program[31] in order to add hydrogen atoms and to assign nonbonded force field parameters. Optionally, they can be optimized, which includes the addition of missing heavy atoms and resolution of steric clashes. Protein structures that are submitted already in PQR format are left untouched, and only ϵ atomic LJ parameters are added based on default atom types.

Ligand structures that are extracted automatically from PDB files are supplemented with hydrogen atoms by the openbabel[32] program, converted to mol2 format, and submitted to the antechamber program[33] for partial charges calculation. If automatic ligand processing fails, a separate MOL2 file with an already protonated ligand structure can be submitted. Alternatively, both the protonation state and partial charges can be defined within the MOL2 file.

The Output. The system_hydrated.pdb file contains the submitted system along with predicted locations of bound water molecules and their estimated binding free energies, ΔG, in kcal/mol (stored in the occupancy column). *All* detected hydration sites, including those only partially hydrated (with $\Delta G > 0$), are listed in a separate file named grid_s.pdb. A spatial distribution of excess chemical potential mapped on a 0.5 Å spaced Cartesian grid is available in the grid_c.pdb file, with the μ_{ex} value present in the occupancy column. An input structure processed by the server and assembled into a calculation-ready form is included in the system.pqrs file.

Test Set and Test Calculations. To assess the server's ability to predict hydration patterns at protein−ligand binding interfaces, we considered a benchmark set of ~1500 high resolution (≤1.5 Å) crystal structures with validated electron densities for water molecules.[34,35] Out of this set, we selected complexes that (a) were formed by protein chains shorter than 600 residues, (b) did not have any nonprotein or nonwater atoms or atoms assigned to a different chain within 8 Å from ligand heavy atoms, and (c) did not have any water molecule closer than 2 Å to ligand heavy atoms. The resulting 463 protein−ligand complexes and considered interface water

molecules, defined as those at most a 4 Å distance from both protein and ligand heavy atoms, are listed in the SI.

In each case, we first attempted to perform calculations in fully automatic mode by providing the PDB code for download, desired chain id, ligand residue name, and ligand charge of 0 e, using default N = 500 repetitions. If this procedure was unsuccessful (mostly because of problems with ligand processing), we extracted ligand structures, assigned protonation states using Discovery Studio Visualizer (Dassault Systèmes, BIOVIA; custom script), and submitted to the server protein chains and ligand structures as PDB and MOL2 files, respectively. If subsequent calculations were still unsuccessful, the complex was discarded.

We analyzed the number of truly predicted (TP) water molecules and the number of false predictions (FP). The former were identified as those having a crystallographic binding site water oxygen atom within a certain cutoff distance, R_c. The latter were defined as those not having *any* crystallographic water oxygen within R_c. TP and FP numbers were transformed to TP and FP fractions, TPf and FPf, respectively, defined such that TPf = TP/Xw and FPf = FP/Nw, where Xw denotes the number of crystallographic interface water molecules, and Nw denotes the number of water molecules placed by the server within the interface. We further stratified TPf and FPf depending on the fraction of solvent accessible surface area (SASAf = SASA/$4\pi r^2$, where r = 1.4 Å) of the respective water oxygen atoms.

■ RESULTS

The output offers insight into local surface hydration (Figure 2a), as well as into the distribution and affinity of buried water molecules (Figure 2b). In this latter respect, a sample output shows a number of reasonably well placed, strongly bound water molecules within buried regions. Solvent exposed areas of the interface are typically populated by less strongly bound water,
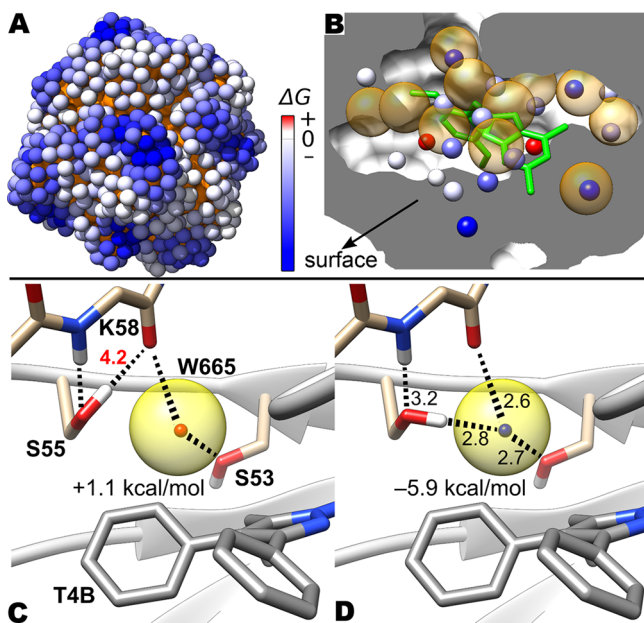
Figure 2. Sample gridSolvate results: (A) protein surface hydration (PDB: 2nnq) and (B) protein−ligand interface hydration (PDB: 1uyg). Prediction (C) before and (D) after manual adjustment of S55 hydroxyl group orientation in fatty acid binding protein (PDB: 2nnq). Yellow sphere of 1.4 Å radius indicates the position of W665; small spheres indicate predicted water locations. Hydrogen bond distances are in Å.

often misplaced with respect to crystallographic water oxygens. One may expect, however, that water molecules located there are less important for specific protein−ligand binding and also that their network captured in a crystal lattice may not fully hold in physiologic conditions. The presence of two predicted water molecules with $\Delta G > 0$ (red spheres) at positions not occupied by crystal water highlights the fact that binding interfaces are not tightly packed but contain void regions. Their correct identification is equally important as the accurate placement of the true solvent.

Out of the 463 protein−ligand complexes, the server managed to automatically process 253. A further 150 were processed following the alternative, yet still automatic, protonation scheme. Out of the remaining 60 cases, 48 failed in partial charge calculations and 12 due to incomplete protein structures, preventing PQR file generation.

The assessment of model predictions as a function of R_c is presented in Figure 3a. As can be expected, TPf rises as less
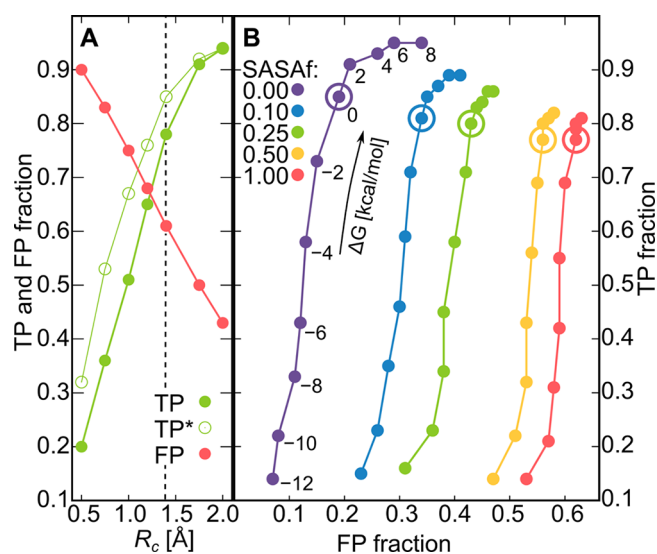
Figure 3. (A) TPf and FPf for predictions with $\Delta G < 0$. TP* refers to water molecules forming at least two hydrogen bonds with a protein−ligand environment.[35] (B) TPf and FPf as a function of free energy cutoff assessed based on water molecules with SASAf up to a given value.

accurate water placements are accepted. We assume that the upper limit of reasonable accuracy is R_c = 1.4 Å, which corresponds to a customary radius of a water molecule. Here, an average TPf reaches 0.77, with values for individual complexes varying from 0.0 to 1.0.

In general, the model places more water molecules within the interface region than are present in the crystal structures (Nw = 8094, Xw = 3060, respectively, Table SI). This translates to seemingly high FPf = 0.62 for R_c = 1.4 Å. It should not necessarily indicate truly erroneous results, since individual water molecules in solvent-exposed areas are often not resolved in crystal structures and can be also missing in buried regions. Indeed, if SASAf distributions among X-ray- and model-predicted binding site water molecules are compared (SI, Figure 1), it is evident that the model more likely covers solvent-exposed areas of binding interfaces, whose hydration is underrepresented in crystal structures. For reasonable water placement in such regions, it is important to maintain correct spacing between individual solvent molecules. In this respect,

the model was parametrized to reproduce distance distribution within surface water obtained with explicit solvent MD simulations.[29] Consequently, we assume that FPf is increasingly overestimated as more exposed hydration sites are taken into account. This is reflected by a significant drop of FPf for increasingly buried hydration sites (Figure 3b). For the most buried water molecules, 0.85 of all crystallographic locations are recovered, with a 0.19 rate of potentially false placements (Figure 3b, violet circle).

The likelihood of correct water placement and its importance for protein−ligand binding can be gauged based on the estimated ΔG. Water molecules predicted to bind with high affinity are less likely to be false predictions, which is particularly evident for most buried hydration sites (Figure 3b). At the same time, the threshold of $\Delta G = 0$ seems to be reasonable for discriminating between occupied and vacated hydration sites, as relatively little TPf gain is achieved for higher ΔG values at a price of increased FPf. This is in agreement with our previous results concerning protein cavities, showing that sites with positive ΔG tend to constitute true negative predictions.[29] Still, their consideration in drug design may be important, since they indicate sterically accessible areas whose targeting with hydrophobic ligand groups may increase its binding strength.

The overall quality of results is similar as delivered by related approaches[8,21,25,35] that typically report TPf in the 0.7−0.8 range. A straightforward comparison is hampered by various criteria used to confirm successful water placement (e.g., a threshold used to distinguish occupied and empty locations), the level of user involvement in calculations (automatic vs supervised), and inclusion or not of FPf into the assessment. Of importance is also the nature of the test set; as can be seen from Figure 3a, if the set is limited to water molecules that form at least two hydrogen bonds with a protein−ligand environment as in the original work by Nittinger et al.,[35] TPf is increased by more than 0.1 with respect to predictions for our standard interface definition. Still, the results are a bit worse compared to that original work, in which a TPf of 0.8 is reported for $R_c = 1.0$ Å. This may result to some extent from a relatively coarse 0.5 Å grid used in our method, the fact that the prediction of water placement in our method is based on a general force field model rather than trained directly on crystal structures, and also from a possibly inferior generic scheme of system protonation.

An important limitation of the proposed approach is fixed solute geometry considered for calculations. Most problematic is the orientation of rotatable hydroxyl groups since it is determined in the absence of water molecules and may be not optimal. Such a situation is evident in the W665 water binding site in fatty acid binding protein from the test set (Figure 2c,d). An incorrect orientation of the S55 hydroxyl group results in the prediction of water $\Delta G \simeq +1$ kcal/mol and, hence, the treatment of the site as empty. Manual rotation of the S55 hydroxyl group results in $\Delta G = -6$ kcal/mol and correct assignment of a buried water molecule, indicating room for significant improvement if the user's knowledge is involved in system preparation.

SUMMARY

We presented a novel web server, GridSolvate, for the assessment of interaction between water and biomolecular solutes such as proteins or protein−ligand complexes. It is based on our previously introduced, semiexplicit hydration model. The server accepts and automatically handles atomistic input structures provided in standard, easily available formats but leaves also the possibility to manually specify and tune their

details. The output includes the distribution of surface and internal solvent molecules in the context of a solute, together with estimates of their excess chemical potential. It can be used to analyze individual water molecules at protein−ligand binding sites for the sake of drug design studies, as well as to derive descriptors of surface hydration propensity for the prediction of protein aggregation or macromolecular interface regions.

■ ASSOCIATED CONTENT

ⓢ Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jcim.0c00779.

Information as mentioned in the text (TXT)

Information as mentioned in the text (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Piotr Setny − *Centre of New Technologies, University of Warsaw, 02-097 Warsaw, Poland;* ⓞ orcid.org/0000-0002-0769-0943; Email: p.setny@cent.uw.edu.pl

Complete contact information is available at:
https://pubs.acs.org/10.1021/acs.jcim.0c00779

Notes

The author declares no competing financial interest.

■ ACKNOWLEDGMENTS

■ REFERENCES

(1) Hummer, G.; Tokmakoff, A. Preface: Special Topic on Biological Water. *J. Chem. Phys.* **2014**, *141*, 22D101.

(2) Spyrakis, F.; Ahmed, M. H.; Bayden, A. S.; Cozzini, P.; Mozzarelli, A.; Kellogg, G. E. The Roles of Water in the Protein Matrix: A Largely Untapped Resource for Drug Discovery. *J. Med. Chem.* **2017**, *60*, 6781−6828.

(3) Onufriev, A. V.; Izadi, S. Water models for biomolecular simulations. *WIREs Comput. Mol. Sci.* **2018**, *8*, No. e1347.

(4) Lazaridis, T. Inhomogeneous Fluid Approach to Solvation Thermodynamics. 1. Theory. *J. Phys. Chem. B* **1998**, *102*, 3531−3541.

(5) Michel, J.; Henchman, R. H.; Gerogiokas, G.; Southey, M. W. Y.; Mazanetz, M. P.; Law, R. J. Evaluation of Host−Guest Binding Thermodynamics of Model Cavities with Grid Cell Theory. *J. Chem. Theory Comput.* **2014**, *10*, 4055−4068.

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

(7) Ross, G. A.; Bodnarchuk, M. S.; Essex, J. W. Water Sites, Networks, and Free Energies with Grand Canonical Monte Carlo. *J. Am. Chem. Soc.* **2015**, *137*, 14930−14943.

(8) Rudling, A.; Orro, A.; Carlsson, J. Prediction of Ordered Water Molecules in Protein Binding Sites from Molecular Dynamics Simulations: The Impact of Ligand Binding on Hydration Networks. *J. Chem. Inf. Model.* **2018**, *58*, 350−361.

(9) Cui, G.; Swails, J. M.; Manas, E. S. SPAM: A Simple Approach for Profiling Bound Water Molecules. *J. Chem. Theory Comput.* **2013**, *9*, 5539−5549.

(10) Hu, B.; Lill, M. A. WATsite: Hydration site prediction program with PyMOL interface. *J. Comput. Chem.* **2014**, *35*, 1255−1260.

(11) Cuzzolin, A.; Deganutti, G.; Salmaso, V.; Sturlese, M.; Moro, S. AquaMMapS: An Alternative Tool to Monitor the Role of Water

Molecules During Protein−Ligand Association. *ChemMedChem* **2018**, *13*, 522−531.

(12) Roux, B.; Simonson, T. Implicit solvent models. *Biophys. Chem.* **1999**, *78*, 1−20.

(13) Pitt, W. R.; Murray-Rust, J.; Goodfellow, J. M. AQUARIUS2: Knowledge-based modeling of solvent sites around proteins. *J. Comput. Chem.* **1993**, *14*, 1007−1018.

(14) Jiang, L.; Kuhlman, B.; Kortemme, T.; Baker, D. A "solvated rotamer" approach to modeling water-mediated hydrogen bonds at protein-protein interfaces. *Proteins: Struct., Funct., Genet.* **2005**, *58*, 893−904.

(15) Fiorucci, S.; Zacharias, M. Prediction of protein-protein interaction sites using electrostatic desolvation profiles. *Biophys. J.* **2010**, *98*, 1921−30.

(16) Raymer, M. L.; Sanschagrin, P. C.; Punch, W. F.; Venkataraman, S.; Goodman, E. D.; Kuhn, L. a. Predicting conserved water-mediated and polar ligand interactions in proteins using a K-nearest-neighbors genetic algorithm. *J. Mol. Biol.* **1997**, *265*, 445−64.

(17) García-Sosa, A. T.; Mancera, R. L.; Dean, P. M. WaterScore: a novel method for distinguishing between bound and displaceable water molecules in the crystal structure of the binding site of protein-ligand complexes. *J. Mol. Model.* **2003**, *9*, 172−82.

(18) Amadasi, A.; Surface, J. A.; Spyrakis, F.; Cozzini, P.; Mozzarelli, A.; Kellogg, G. E. Robust classification of "relevant" water molecules in putative protein binding sites. *J. Med. Chem.* **2008**, *51*, 1063−7.

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

(20) Schymkowitz, J. W. H.; Rousseau, F.; Martins, I. C.; Ferkinghoff-Borg, J.; Stricher, F.; Serrano, L. Prediction of water and metal binding sites and their affinities by using the Fold-X force field. *Proc. Natl. Acad. Sci. U. S. A.* **2005**, *102*, 10147−52.

(21) 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*, 1867−81.

(22) *SZMAP 1.5.0.2*; OpenEye Scientific Software: Santa Fe, NM, 2013. http://www.eyesopen.com (accessed October 2020).

(23) Rarey, M.; Kramer, B.; Lengauer, T. The particle concept: placing discrete water molecules during protein-ligand docking predictions. *Proteins: Struct., Funct., Genet.* **1999**, *34*, 17−28.

(24) Verdonk, M. L.; Chessari, G.; Cole, J. C.; Hartshorn, M. J.; Murray, C. W.; Nissink, J. W. M.; Taylor, R. D.; Taylor, R. Modeling water molecules in protein-ligand docking using GOLD. *J. Med. Chem.* **2005**, *48*, 6504−15.

(25) 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*, No. e32036.

(26) Forli, S.; Olson, A. J. A force field with discrete displaceable waters and desolvation entropy for hydrated ligand docking. *J. Med. Chem.* **2012**, *55*, 623−38.

(27) Setny, P.; Zacharias, M. Hydration in discrete water. A mean field, cellular automata based approach to calculating hydration free energies. *J. Phys. Chem. B* **2010**, *114*, 8667−75.

(28) Setny, P. Hydration in Discrete Water (II): From Neutral to Charged Solutes. *J. Phys. Chem. B* **2015**, *119*, 5970−5978.

(29) Setny, P. Prediction of water binding to protein hydration sites with discrete, semi-explicit solvent model. *J. Chem. Theory Comput.* **2015**, *11*, 5961−5972.

(30) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The protein data bank. *Nucleic Acids Res.* **2000**, *28*, 235−242.

(31) Dolinsky, T. J.; Nielsen, J. E.; McCammon, J. A.; Baker, N. A. PDB2PQR: an automated pipeline for the setup of Poisson-Boltzmann electrostatics calculations. *Nucleic Acids Res.* **2004**, *32*, W665−W667.

(32) O'Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R. Open Babel: An open chemical toolbox. *J. Cheminf.* **2011**, *3*, 33.

(33) Wang, J.; Wang, W.; Kollman, P. A.; Case, D. A. Automatic atom type and bond type perception in molecular mechanical calculations. *J. Mol. Graphics Modell.* **2006**, *25*, 247−260.

(34) Nittinger, E.; Schneider, N.; Lange, G.; Rarey, M. Evidence of water molecules - A statistical evaluation of water molecules based on electron density. *J. Chem. Inf. Model.* **2015**, *55*, 771−783.

(35) Nittinger, E.; Flachsenberg, F.; Bietz, S.; Lange, G.; Klein, R.; Rarey, M. Placement of Water Molecules in Protein Structures: From Large-Scale Evaluations to Single-Case Examples. *J. Chem. Inf. Model.* **2018**, *58*, 1625−1637.