HHS Public Access

Author manuscript

J Chem Theory Comput. Author manuscript; available in PMC 2019 January 09.

Published in final edited form as:

J Chem Theory Comput. 2018 January 09; 14(1): 418–425. doi:10.1021/acs.jctc.7b00592.

Solvation Structure and Thermodynamic Mapping (SSTMap): An open-source, flexible package for the analysis of water in molecular dynamics trajectories

Kamran Haider[†], **Anthony Cruz**[‡,¶], **Steven Ramsey**[‡,§], **Michael K Gilson**[‖], and **Tom Kurtzman**[*,‡,¶,§]

[†]Department of Physics, City College of New York, The City University of New York, 160 Convent Ave, New York, NY 10031

[‡]Department of Chemistry, Lehman College, The City University of New York, 250 Bedford Park Boulevard West, Bronx, New York, NY 10468

[¶]Ph.D. Program in Chemistry, The Graduate Center of The City University of New York, 365 Fifth Avenue, New York, New York, 10016, United States

[§]Ph.D. Program in Biochemistry, The Graduate Center of The City University of New York, 365 Fifth Avenue, New York, New York, 10016, United States

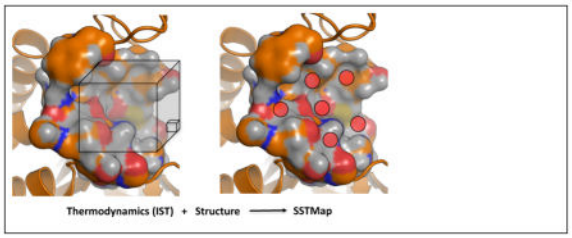
[‖]Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California San Diego, 9500 Gilman Drive, La Jolla, California, CA, 92093-0736

Abstract

We have developed SSTMap, a software package for mapping structural and thermodynamic water properties in molecular dynamics trajectories. The package introduces automated analysis and mapping of local measures of frustration and enhancement of water structure. The thermodynamic calculations are based on Inhomogeneous Fluid Solvation Theory (IST), which is implemented using both site-based and grid-based approaches. The package also extends the applicability of solvation analysis calculations to multiple molecular dynamics (MD) simulation programs by using existing cross-platform tools for parsing MD parameter and trajectory files. SSTMap is implemented in Python and contains both command-line tools and a Python module to facilitate flexibility in setting up calculations and for automated generation of large datasets involving analysis of multiple solutes. Output is generated in formats compatible with popular Python data science packages. This tool will be used by the molecular modeling community for computational analysis of water in problems of biophysical interest such as ligand binding and protein function.

Graphical Abstract

thomas.kurtzman@lehman.cuny.edu, Phone: 718-960-8832.



Introduction

The interactions between molecules, in biological settings, often occur in aqueous media where water plays a significant role.[1–6] Water molecules structure themselves on solute surfaces leading to differences in their enthalpy and entropy, relative to the bulk phase. Upon non-covalent association, water is displaced from the binding surfaces of the solutes into the bulk and accompanying changes in enthalpy and entropy of water molecules make significant contributions to the overall free energy of binding. In addition, water restructures itself around the bound complex and this reorganization has further thermodynamic consequences.[7–11]

As a result, characterizing active-site water structure and estimating the solvent contribution to the free energy of binding has become a key consideration in modern drug design.[3,12] Several computational tools have been developed in the last few years that focus on mapping out thermodynamic properties of water molecules in protein binding cavities (see reference[13] for a list and comparison of different tools). One class of tools that has gained favor in the molecular modeling community is based on Inhomogeneous fluid Solvation Theory (IST),[14,15] which provides a statistical mechanical framework to calculate solvation thermodynamics by analyzing molecular distribution functions from explicit solvent simulations of solute molecules.[14–17] Various implementations of IST includeWaterMap,[17] STOW,[18] GIST,[19,20] WatClust[21] and others.[22] Despite the utility of these tools in computer-aided drug design,[13,23] there are still limitations in their widespread application. For example, some of the tools are available only commercially and others are restricted to specific MD packages containing either grid-based or site-based implementation.

There has also been recent interest in how networks of water molecules in binding sites affect the binding of small molecules to protein targets.[24–27] The restructuring of water networks contributed to enthalpy-entropy compensation in ligand binding to carbonic anhydrase[25] and serine protease[27] and explained structure-activity relationships in a series of thermolysin ligands. [28] The strengthening of hydrogen-bonded water networks due to active-site mutations in carbonic anhydrase was shown to increase the enthalpic cost of binding.[26] In another study, ligands were designed based on a concept of stabilizing water networks around bound complexes, resulting in at least one example where the substituent with the most stabilized water network in the protein-ligand complex showed a significant increase in affinity.[29] We used measures of local water structure to investigate water molecules in active sites, and proposed that displacing water with disrupted local water structure could lead to enhancements in binding affinity, even when the local water thermodynamics is favorable

with respect to bulk.[30] Taken together, these results motivate further studies of the reorganization of water networks upon ligand binding and methods that facilitate analysis of water molecules in active sites.

Here we introduce SSTMap, an open-source computational tool that maps out measures of water structure and thermodynamics on solute surfaces, such as in protein binding pockets, using data from explicit solvent molecular dynamics trajectories. It broadens the application of solvation analysis calculations to a variety of popular MD trajectory and parameter file formats by interfacing with available cross-platform tools for trajectory and parameter file parsing, namely MDTraj[31] and ParmEd.[32] SSTMap calculations can be performed over a defined region on the surface that can be discretized into either a 3D grid of voxels (referred to as Grid-based Inhomogeneous Solvation Theory or GIST) or a number of discrete high density hydration sites (referred to as Hydration Site Analysis or HSA) similar to the WaterMap implementation of IFST (Figure 1). The full list of quantities calculated for each hydration site or voxel is reported in Table 1. The Python application programming interface (API) of SSTMap provides access to its individual functions, such as identification of hydration site clusters, calculation of voxel occupancies, energy, entropy, and hydrogen bonding calculations. Supporting documentation and tutorials for running SSTMap are available at sstmap.org.

Below, we describe the design and implementation features of SSTMap, followed by demonstration of its usage by sample applications to the protein Caspase 3. The first example demonstrates running SSTMap calculations through command-line tools or Python scripts. The second example shows different features for visualization and analysis of local water structure in active-sites. Finally, the third example demonstrates how data generated from SSTMap can be analyzed with Python's data analysis packages.

Implementation Features

Installation

SSTMap is distributed through conda, the package manager provided with the freely available Anaconda Python distribution, and can be installed on Linux and Mac OS operating systems, using the following `conda` command:

Box 1

Running SSTMap through command-line

```
conda install -c solvationtools -c omnia sstmap
```

The end-user can also download the development version of SSTMap from the Github repository and install it as follows:

Box 2

Running SSTMap through the command-line

```
git clone git@github.com:KurtzmanLab/SSTMap.git
cd SSTMap
python setup.py install
```

Coding Practices

SSTMap is implemented in Python 2.7, with linked C/C++ modules to speed up computationally intensive calculations. It interfaces with the cross-platform parameter and trajectory parsing tools, namely MDTraj[31] and ParmEd[32] so that the solvation structure and thermodynamic analyses can be applied to trajectories from a wide variety of MD packages including AMBER,[33] DESMOND,[34] Gromacs,[35] CHARMM,[36] OpenMM[37] and NAMD.[38]

SSTMap calculations utilize multidimensional NumPy arrays[39] for handling numeric data, such as coordinates and pairwise distances. The implementation of solvation analysis calculations follows an object-oriented model. The modular structure allows for easy maintenance and future enhancements. SSTMap is extensively documented through the use of Python docstrings, following the PEP257 conventions. Additionally, tutorials with step-by-step instructions on basic and advanced applications of the tool are available at sstmap.org. The source code of SSTMap is hosted on Github and is available under an open source license.

Methods

Simulation Details

SSTMap calculations require an explicit solvent molecular dynamics trajectory and the corresponding topology file. For demonstrative purposes, we simulated a Caspase 3 structure (PDB ID: 3H5I)[40] in the Amber14 molecular simulation package.[33] The starting structure contains a co-crystallized ligand, which was removed. The simulation was performed in a box of 17,102 TIP3P water molecules with Amber14SB[41] parameters for the protein structure. The solvated system was minimized with an initial 1500 steps of steepest descent with protein atoms restrained harmonically using a force constant of 100 kcal/mol$Å^2$, followed by another round of up to 2000 steps of conjugate gradient minimization, with the same restraints but applied only to protein heavy atoms. This was followed by heating the system to 300 K over 100 ps under the conditions of constant number of particles, volume and temperature (NVT). An equilibration simulation was then run in constant NPT conditions for 1 ns, with gradual removal of heavy atom restraints to 2.5 kcal/mol$Å^2$. The final production run was performed in constant NVT conditions at a temperature of 300 K for 10ns and with heavy atoms of protein restrained about their starting positions with a force constant of 2.5 kcal/mol$Å^2$, and with bonds involving hydrogen atoms constrained with the SHAKE algorithm.[42] Temperature was regulated using the Langevin thermostat[43] with a collision frequency of 1.0 ps^{-1} and pressure was regulated using the Berendsen

barostat[44] with isotropic scaling and a coupling constant of 1.0 ps. The snapshots of system coordinates were saved every 1ps, resulting in a trajectory file with 10,000 frames.

Capabilities

Setting up Solvation Analysis Calculations

SSTMap has two main command-line programs, named `run_hsa` and `run_gist`, that generate water structure and thermodynamic calculations in hydration sites and on a grid respectively. A generic example of these commands is shown below (Box 3).

Box 3

Running SSTMap through command-line

```
run_hsa -i topologyfilename -t trajectoryfilename -l ligandfilename.pdb -
f 10000
↪ -s 0 -o outputfilename
run_gist -i topologyfilename -t trajectoryfilename -l ligandfilename.pdb -
d 40 40
↪ 40 -f 10000 -s 0 -o outputfilename
# topology filename and trajectoryfileanme correspond to the
# topology and trajectory files from the MD package used for simulation.
```

The `-i` and `-t` flags are used to specify the topology and trajectory files, respectively, corresponding to the MD package used for simulation. The ligand PDB file, specified under `-l` flag, is used to define the binding site region for either clustering for HSA or creating 3D grids for GIST. The ligand can be either a co-crystallized molecule or a set of atoms in the binding site or even a set of dummy atoms to specify a region around which the water molecules from the entire simulations are used for analysis. There are optional flags for specifying starting frame and a prefix to name the output files (Table 2). Some MD packages have parameter files containing non-bonded interaction parameters. These can be specified under `-p` flag. Commands specific to each MD supported packages are provided in a tutorial on sstmap.org.

Both programs output a summary text file containing data for individual sites or voxels. This file contains averages of the thermodynamic and structural quantities, listed in Table 1. The `run_hsa` program also generates a PDB file containing the coordinates of all hydration sites identified during the calculation (Figure 2, left), and two PDB files for each hydration site, one consisting of the full set of water molecules found in the site during the entire length of simulation and another containing the most probable positions and orientations of water molecules found in the site.

The `run_gist` program calculates each quantity for each voxel inside a grid. The results are stored as Data Explorer (DX) format files for each quantity calculated over the grid; these can be visualized in standard molecular visualization packages such as PyMol[45] and

VMD.[46] The DX files can also be post-processed using previously developed GIST tools.[47] As an example, Figure 2 (right) shows the DX map for water density in the active site of Caspase 3, contoured at three times bulk density. The solvation analysis calculations in SSTMap can also be applied to protein-ligand complexes. As noted in our prior implementation,[19,47] a potentially useful application of GIST is to perform an end-states analysis between initial and final states, such as a complex with unsubstituted and substituted ligand. The thermodynamic impact of water reorganization upon ligand modification can then be calculated by taking the differences in the quantities in the initial and final states.

SSTMap can also be accessed through the API using Python code. The key benefit of using the API compared to the command-line tools is the flexibility in programmatically accessing SSTMap's capabilities and incorporating them into larger work flows. The code in Box 4 demonstrates an example Python code that is used to initialize and run customized GIST calculations, using the Caspase 3 example. In example 1, only energetic quantities are calculated for the grid (Box 4, line 8). In general, any combination of boolean flags can be used for energy, entropy, and hydrogen bonding calculations. In example 2, a GIST grid is initialized with a user-defined grid center, as opposed to the default where it is automatically derived from geometric center of the ligand molecule, Box 4, line 13.

Box 4

Using the `sstmap` API from Python code

```
 1 from sstmap.grid_water_analysis import GridWaterAnalysis
 2 # Example 1: Run GIST calculation with only energy calculations.
 3 gist = GridWaterAnalysis ("casp3.prmtop", "casp3.netcdf",
 4                                          start_frame=0,
num_frames=100000,
 5
ligand_file="casp3_ligand.pdb",
 6                                          grid_dimensions=[48, 48, 48],
 7                                          prefix="casp3")
 8 gist.calculate_grid_quantities(energy=True)
 9 # Example 2: Initialize GIST with a user-defined grid center position.
10 gist = GridWaterAnalysis("casp3.prmtop", "casp3.netcdf",
11                           start_frame=0, num_frames=100000,
12                           ligand_file="casp3_ligand.pdb",
13                           grid_center=[35.33, 52.23, 54.96],
14                           grid_dimensions=[48, 48, 48],
15                           prefix="casp3")
16 gist.calculate_grid_quantities()
```

Calculation and Visualization of Water Structure

SSTMap provides not only thermodynamic quantities based on the IST formalism, but also measures of the frustration or enhancement of water networks on solute surfaces. For example, SSTMap breaks down the water-water energy for a voxel or a hydration site into contributions from the different solvation shells. The contribution from the first solvation shell divided by the number of first shell neighbors provides a measure of enhancement or frustration in the local interactions for a hydration site or a voxel (E_{nbr}^{ww} in Table 1). A closely related measure of water structure is the fraction of hydrogen-bonded neighbors (f_{ww}^{HB}), the greater this fraction, the more structured the local network of water. A complete list of structural quantities can be found in the first half of Table 1. These quantities are calculated by default for hydration sites or GIST grids, using either the `run_hsa/run_gist` programs or the corresponding functions from the API (Box 4).

Additional functions are also available through the `sstmap` module for plotting water-water pair energy distributions, characterizing local hydrogen bonding environments, and calculating water-water energy contributions from successive solvation shells which enables investigation of long range water structure.

In the following, we demonstrate the functionality for visualizing local water structure around a hydration site, using the distributions of water-water pair energies in the first solvation shell. The distribution is obtained by storing individual pair interactions of every instance of a water molecule in a hydration site with its corresponding first shell neighbors. These data are stored during HSA calculations. The histograms of the pair interactions can be plotted as probability distributions of water-water pair-energies in the first shell, using utility functions in the `sstmap` module, as demonstrated by the code listed in Box 5. For example, a hydration site water that typically interacts favorably with one nearest-neighbor water, and unfavorably with another, will have a bimodal pair energy distribution. In this example, distributions for two neighboring hydration sites in Caspase 3 (top panel of Figure 3) is overlaid with the corresponding distribution from TIP3P water model. Unlike the bulk water distribution, these distributions have peaks indicating a significant population of both favorable and unfavorable water-water pair interaction energies. These distributions are for two neighboring water molecules which also form strong hydrogen bonds with residues on the protein surface, including two arginine, one histidine and one glutamine side chains. The water molecules in these sites predominantly form unfavorable interactions with each other while forming favorable interactions with their other water neighbors. This leads to the bimodal distribution seen in Figure 3. Our interpretation of this is that the water molecules are geometrically unable to simultaneously form hydrogen bonds with the protein surface and with each other and preferably form hydrogen bonds with the protein. This causes the interaction of the hydration site water molecules with each other to be positive on average which is reflected in in the rightmost peaks of Figure 3. We address the water in these sites in further detail in our prior publication.[30]

Box 5

Plotting functions to visualize distributions of water-water interactions

```
from sstmap.utils import plot_enbr, plot_rtheta
# Directory containing run_hsa output
data_1 = './casp3_hsa_data/'
# Use bulk water distribution for comparison
ref_enbr = 'bulkwat_enbr'
plot_enbr(data_1,
              site_indices=[0, 4],
              ref_data=ref_enbr)
# Directory containing angular structure data
data_2 = './casp3_angular_structure_data/'
# Use bulk water distribution for comparison
ref_rtheta = 'bulkwat_r_theta'
plot_rtheta(data_2,
            site_indices=[0, 4],
            ref_data=ref_rtheta)
```

In a similar manner, examples of plots for the joint probability distribution of distances and angles between water molecules in the hydration site and their neighbors, within a distance cutoff (6.0 A by default) is shown in Figure 3, bottom panel. These plots can be generated using Python statements shown in Box 5 and provide a visualization of the orientation of water molecules with respect to a hydration site. The orientation is described by the angle that corresponds to the minimum of four possible hydrogen bond angles between the site water and its neighboring water.[48] When seen in comparison with bulk water distribution (Figure 3, bottom left), site 0 (bottom middle) shows high angle peaks (> 30°) with its water neighbors found at 4.0 to 5.5 A sub-shell. Similarly, site 4 (bottom right) shows a high probability of finding poorly aligned water molecules in its first shell, consistent with its distribution in Figure 3 (top right). The added advantage in these plots is that water-water interactions in distant shells can be analyzed. The poor water-water interactions in the long-range for a given site can be further quantified by obtaining the measure $E_{ww}-(E_{nbr}^{ww}N_{nbr})$. Taken together, the plotting functions and the quantities describing breakdown in water-water interactions on the surface provide a useful approach to characterize water molecules in protein cavities, as noted previously.[30]

Analysis of Hydration Sites/GIST Grids

The `sstmap` module can be used with existing molecular simulation and data science packages for more advanced applications, such as automating solvation analysis and performing statistical analyses of hydration site datasets or GIST grids. The automation of solvation analysis for multiple trajectories or a large number of systems can be processed simply by looping over all the systems and using Python statements similar to the ones

shown in Box 4 for each iteration. For simulation programs that offer a Python API (e.g., OpenMM[49]), every step of the system preparation, simulation and analysis can be automated, thereby facilitating reproducibility of results.

Python's popular data analysis and visualization packages (e.g., `pandas, scipy.stats, scikit-learn, matplotlib, jupyter notebook`, see reference[50]) can be used to enhance the analysis of datasets of hydration sites or GIST grids obtained from SSTMap calculations. For example, determining characteristics of hydration sites that are involved in ligand binding can be valuable in understanding water displacement.

We demonstrate such usage of `sstmap` by an example code listed in Box 6. The goal in this analysis is to determine how interaction with charged groups on protein surfaces affects local water structure. Typically, high energy hydration sites are considered to contribute favorably to ligand binding upon displacement.[17,20] However, It is difficult to evaluate hydration sites where water interacts strongly with the protein and is, generally, enthalpically favorable relative to bulk. The Python script in Box 6 reads in a dataset of 218 hydration sites from six different proteins that was generated in a prior publication[30] and is publicly available. The script loads the dataset as a `pandas` data-frame, which facilitates extracting Individual columns for different subsets of hydration sites (Box 6, lines 8–14), which in this case is a set of hydration site categorized as charged sites (See[30] for definition).

Box 6

Analysis of hydration site datasets

```
 1 import pandas as pd
 2 from scipy import stats
 3 import matplotlib.pyplot as plt
 4 from matplotlib.backends.backend_pdf import PdfPages
 5 # Read in HSA dataset in Supplementary Information
 6 # of Haider et al, 2016.
 7 hsa_dataset = pd.read_excel(
 8               "https://ndownloader.figshare.com/files/5324491", None)
 9 charged_sites = hsa_dataset["All C"]
10 # Read solute-water energy
11 solute_water_energy = charged_sites["E_sw"]
12 # Read first shell water-water energy per neighbor
13 E_nbr = charged_sites["E^nbr_ww"]
14 # perform linear regression
15 slope, intercept, r_value, p_value, std_err = stats.linregress(
16            solute_water_energy.values, E_nbr.values)
17 line = slope * solute_water_energy.values + intercept
18 # plot formatting statements are omitted
19 plt.plot(solute_water_energy.values, line)
```

A scatter plot (Figure 4) and a linear regression model (Box 6, lines 15–17) is generated to analyze the relationship between solute-water interaction and perturbation in local water structure. As the interactions with the surface strengthen, local water structure is increasingly disrupted. In particular, sites above the dashed line have significantly less favorable E_{nbr} than bulk water, and form strong energetic interactions with the surface at the expense of poor interactions with first solvation shell water neighbors. It is expected such water molecules gain favorable first shell interactions in bulk, which may be a relevant contribution in displacement. (See reference[30] for more discussion on the topic).

We conjecture that similar analyses, combined with prospective studies of targeted water displacement, will generate new insights about the properties of water molecules in active-sites. SSTMap provides programmatic access to a range of thermodynamic and structural properties of binding site water, using both the hydration site and grid representations, and it processes trajectory data from multiple widely used simulation packages. It is therefore well-suited to facilitate such studies.

Selection of Simulation Parameters

Here we discuss choosing simulation parameters that are relevant for running HSA/GIST analysis.

Simulation length—The length of production simulation depends on the desired precision. Unless there are buried sites or cavities that are difficult for bulk water to access, a 2–10 ns simulation, with frames saved every picosecond (2–10K frames), is generally sufficient for qualitative maps that clearly and reliably show the spatial distribution of water density, energy, and entropy. For quantitative end-states calculations, convergence of system and/or regional quantities with multiple water molecules to within a desired precision requires longer simulations. Convergence of SSTMap quantities is identical to that reported in our previous work for GIST-cpptraj, [47] with entropy per water molecule converging to within 0.1 kcal/mol per water by 2 ns, 0.04 kcal/mol per water by 10 ns, and 0.004 kcal/mol per water by 100ns and energy being 0.1, 0.2 and 0.04 kcal/mol per water for the respective time scales.

Grid spacing—The convergence of thermodynamic quantities integrated over the entire grid does not depend on the grid spacing. However, the convergence of quantities in each voxel depends on the grid spacing as it is directly affected by the number of water molecules found in the voxel over the course of a simulation.

Constant pressure or constant volume simulation—Most applications of GIST/HSA focus on the solvation in a given sub-region (e.g. the active site of the protein which does not include the entire simulation box). In this case either NVT or NPT is appropriate. However, for thermodynamic end states analyses, which require a treatment of all the molecules in the system, simulations should be run in NVT, in order to avoid mismatches between the grid and the simulation box which would occur at constant pressure, due to fluctuations in the box dimensions. Therefore, after allowing the system to equilibrate at NPT to establish a volume appropriate to 300K and 1 atmosphere, the

production run is often done at constant volume. It is also better to run NVT simulations using the average box volume from the NPT simulation instead of choosing the final box volume. The final volume may correspond to a large volume fluctuation from the average in the NPT ensemble and can lead to a significantly different mean energy of the system.[51]

Solute atom restraints—IST equations are exact when all atoms of the protein are held fixed; see comments on limitations below. However, treating the protein as entirely rigid eliminates protein fluctuations that can lead to informative adjustments in water structure, such as ones that allow optimization of H-bonds between the protein and water molecules. In order to explore various possibilities, we may run simulations with all protein-heavy atoms restrained; only backbone atoms restrained; or with only restraints designed to prevent overall translation and rotation.

Limitations

The solvation mapping methods discussed here include some limitations and approximations, of which users should be aware. Perhaps most straightforward is the fact that the molecular simulations needed to generate the maps rely on potential functions, or force fields, that are not perfectly accurate. In addition, simulations need to be long enough to achieve adequate convergence. As detailed below, obtaining converging solvation thermodynamic quantities is usually straightforward for solvent-exposed sites, but standard simulation methods may not readily equilibrate the water population in a buried site or protein cavity. Several additional points are discussed in the following paragraphs.

First, IST provides an N-body expansion for the entropy, where N is the number of water molecules in the system, and applications of IST typically truncate this expansion after the first-order term. This truncation means that entropic contributions due to solute-induced changes in second- and higher-order water-water correlations are omitted. Computing these correlation terms requires longer simulations and more complex analysis methods, but studies of pure water suggest they are quantitatively important,[52,53] and progress has been made toward quantifying how they are perturbed by solutes. In one approach, the computational challenge is limited by computing orientational correlations only in the first hydration shell,[22] where they are likely to be most perturbed relative to bulk. In addition, we have recently reported progress in calculating entropic contributions from translational water-water correlations.[54] The contribution of water-water correlations to the entropy has also been estimated as proportional to the first order entropy,[55,56] but additional work is needed to assess the quality of this approximation.

Second, with a few exceptions,[10,25] applications of IST have used an initial states approximation. This involves analyzing the water in an initial state with an unoccupied binding site or a binding site containing an unmodified lead compound, without further analyzing water properties in a final state, following binding of a ligand or modification of the lead compound, respectively. Fully accounting for the thermodynamic consequences of water reorganization would require looking at both the initial and final states. Such calculations can be performed with SSTMap, though at the cost of an additional simulation

for each final state of interest. It is also worth noting that a number of studies indicate that the initial states approximation can have predictive value.[5,17,20,57]

Third, IST relies on Percus's source particle method[14,58,59] which treats the solute molecule, typically a protein, as a rigid object. This means that protein motions associated with water rearrangements are not accounted for. If one is using the IST results in the context of docking calculations, which often treat the protein as rigid, this may not be a major concern. However, if one wishes to account for protein flexibility, a natural approach is to simulate the protein without restraints, cluster the resulting conformations, and then run IST calculations for representative rigid snapshots of each highly populated cluster.

Concluding Remarks

The analysis of water has become an important consideration in modern drug design. This has led to an increased interest in developing theory and computational approaches for solvation analysis. The encouraging application of IST-based tools in several past studies prompted us to develop SSTMap, which offers capabilities beyond those of existing software packages for the analysis of surface water. Most notably, it combines thermodynamic analysis with structural analysis that can aid in understanding and evaluating the displacement of active-site water molecules. It provides both site-based and grid-based calculations in one package, with support for multiple MD packages and can be integrated into Python's scientific computing environment for advanced applications.

Acknowledgments

We kindly acknowledge funding by the National Institutes of Health through grant Nos. GM095417 and GM100946. M.K.G. has an equity interest in, and is a co-founder and scientific advisor of Vera Chem LLC. The contents of this publication are solely the responsibility of the authors and do not necessarily represent the official views of the NIH.

References

1. Ball P. Water as an active constituent in cell biology. Chem Rev. 2008; 108:74–108. [PubMed: 18095715]
2. Hummer G. Molecular binding: Under water's influence. Nat Chem. 2010; 2:906–907. [PubMed: 20966940]
3. Mancera RL. Molecular modeling of hydration in drug design. Curr Opin Drug Discov Devel Curr Opin Drug Discovery Dev. 2007; 10:275–280.
4. Krimmer SG, Betz M, Heine A, Klebe G. Methyl, Ethyl, Propyl, Butyl: Futile But Not for Water, as the Correlation of Structure and Thermodynamic Signature Shows in a Congeneric Series of Thermolysin Inhibitors. Chem Med Chem. 2014; 9:833–846. [PubMed: 24623396]
5. Daniel CW, Sherman TB. Calculating Water Thermodynamics in the Binding Site of Proteins - Applications of WaterMap to Drug Discovery. Current Topics in Med Chem. 2017; 17:2586–2598.
6. Geschwindner S, Ulander J, Johansson P. Ligand Binding Thermodynamics in Drug Discovery: Still a Hot Tip? J Med Chem. 2015; 58:6321–6335. [PubMed: 25915439]
7. Ladbury JE. Just add water! The effect of water on the specificity of protein-ligand binding sites and its potential application to drug design. Chemistry & Biology. 1996; 3:973–980. [PubMed: 9000013]
8. Levy Y, Onuchic JN. Water mediation in protein folding and molecular recognition. Annu Rev Biophys Biomol Struct. 2006; 35:389–415. [PubMed: 16689642]

9. Baron R, Setny P, Andrew McCammon J. Water in Cavity-Ligand Recognition. J Am Chem Soc. 2010; 132:12091–12097. [PubMed: 20695475]
10. Snyder PW, Mecinović J, Moustakas DT, Thomas SW, Harder M, Mack ET, Lockett MR, Heroux A, Sherman W, Whitesides GM. Mechanism of the hydrophobic effect in the biomolecular recognition of arylsulfonamides by carbonic anhydrase. Proc Natl Acad Sci US A. 2011; 108:17889–17894.
11. Snyder PW, Lockett MR, Moustakas DT, Whitesides GM. Is it the shape of the cavity, or the shape of the water in the cavity? Eur Phys J-spec Top. 2014; 223:853–891.
12. Wong SE, Lightstone FC. Accounting for water molecules in drug design. Expert Opin Drug Discov. 2011; 6:65–74. [PubMed: 22646827]
13. Bodnarchuk MS. Water, water, every where … It's time to stop and think. Drug Discov Today. 2016; 21:1139–1146. [PubMed: 27210724]
14. Lazaridis T. Inhomogeneous Fluid Approach to Solvation Thermodynamics. 1. Theory. J Phys Chem B. 1998; 102:3531–3541.
15. Lazaridis T. Inhomogeneous Fluid Approach to Solvation Thermodynamics. 2. Applications to Simple Fluids. J Phys Chem B. 1998; 102:3542–3550.
16. Young T, Abel R, Kim B, Berne BJ, Friesner RA. Motifs for molecular recognition exploiting hydrophobic enclosure in protein–ligand binding. Proc Natl Acad Sci US A. 2007; 104:808–813.
17. Abel R, Young T, Farid R, Berne BJ, Friesner RA. Role of the Active-Site Solvent in the Thermodynamics of Factor Xa Ligand Binding. J Am Chem Soc. 2008; 130:2817–2831. [PubMed: 18266362]
18. Li, Z., Lazaridis, T. Computational Drug Discovery and Design. In: Baron, R., editor. Methods in Molecular Biology 819. Springer; New York: 2012. p. 393-404.
19. Nguyen CN, Young TK, Gilson MK. Grid inhomogeneous solvation theory: Hydration structure and thermodynamics of the miniature receptor cucurbit[7]uril. J Chem Phys. 2012; 137:044101. [PubMed: 22852591]
20. Nguyen CN, Cruz A, Gilson MK, Kurtzman T. Thermodynamics of Water in an Enzyme Active Site: Grid-Based Hydration Analysis of Coagulation Factor Xa. J Chem Theory Comput. 2014; 10:2769–2780. [PubMed: 25018673]
21. Lopez ED, Arcon JP, Gauto DF, Petruk AA, Modenutti CP, Dumas VG, Marti MA, Turjanski AG. WATCLUST: a tool for improving the design of drugs based on protein-water interactions. Bioinformatics. 2015; 31:3697–3699. [PubMed: 26198103]
22. Huggins DJ, Payne MC. Assessing the Accuracy of Inhomogeneous Fluid Solvation Theory in Predicting Hydration Free Energies of Simple Solutes. J Phys Chem B. 2013; 117:8232–8244. [PubMed: 23763625]
23. Yang Y, Lightstone FC, Wong SE. Approaches to efficiently estimate solvation and explicit water energetics in ligand binding: the use of WaterMap. Expert Opin Drug Discov. 2013; 8:277–87. [PubMed: 23286874]
24. Biela A, Nasief NN, Betz M, Heine A, Hangauer D, Klebe G. Dissecting the Hydrophobic Effect on the Molecular Level: The Role of Water, Enthalpy, and Entropy in Ligand Binding to Thermolysin. Angew Chem Int Ed. 2013; 52:1822–1828.
25. Breiten B, Lockett MR, Sherman W, Fujita S, Al-Sayah M, Lange H, Bowers CM, Heroux A, Krilov G, Whitesides GM. Water Networks Contribute to Enthalpy/Entropy Compensation in Protein–Ligand Binding. J Am Chem Soc. 2013; 135:15579–15584. [PubMed: 24044696]
26. Fox JM, Kang K, Sastry M, Sherman W, Sankaran B, Zwart PH, Whitesides GM. Water-Restructuring Mutations Can Reverse the Thermodynamic Signature of Ligand Binding to Human Carbonic Anhydrase. Angew Chem Int Ed. 2017; 56:3833–3837.
27. Gopal SM, Klumpers F, Herrmann C, Schafer LV. Solvent effects on ligand binding to a serine protease. Phys Chem Chem Phys. 2017; 19:10753–10766. [PubMed: 28116375]
28. Betz M, Wulsdorf T, Krimmer SG, Klebe G. Impact of Surface Water Layers on Protein–Ligand Binding: How Well Are Experimental Data Reproduced by Molecular Dynamics Simulations in a Thermolysin Test Case? J Chem Inf Model. 2016; 56:223–233. [PubMed: 26691064]

29. Krimmer SG, Cramer J, Betz M, Fridh V, Karlsson R, Heine A, Klebe G. Rational Design of Thermodynamic and Kinetic Binding Profiles by Optimizing Surface Water Networks Coating Protein-Bound Ligands. J Med Chem. 2016; 59:10530–10548. [PubMed: 27933956]
30. Haider K, Wickstrom L, Ramsey S, Gilson MK, Kurtzman T. Enthalpic Breakdown of Water Structure on Protein Active-Site Surfaces. J Phys Chem B. 2016; 120:8743–8756. [PubMed: 27169482]
31. McGibbon RT, Beauchamp KA, Harrigan MP, Klein C, Swails JM, Hernandez CX, Schwantes CR, Wang LP, Lane TJ, Pande VS. MDTraj: A Modern Open Library for the Analysis of Molecular Dynamics Trajectories. Biophys J. 2015; 109:1528–1532. [PubMed: 26488642]
32. Swails, JM. ParmEd - Parameter/topology editor and molecular simulator. 2014. https://github.com/ParmEd/ParmEd
33. Case DA, Cheatham TE, Darden T, Gohlke H, Luo R, Merz KM, Onufriev A, Simmerling C, Wang B, Woods RJ. The Amber bioMol. Simul. programs. J Comput Chem. 2005; 26:1668–1688. [PubMed: 16200636]
34. Bowers, K., Chow, E., Xu, H., Dror, R., Eastwood, M., Gregersen, B., Klepeis, J., Kolossvary, I., Moraes, M., Sacerdoti, F., Salmon, J., Shan, Y., Shaw, D. Scalable Algorithms for Molecular Dynamics Simulations on Commodity Clusters. Proceedings of the ACM/IEEE SC 2006 Conference; 2006; p. 43-43.
35. Hess B, Kutzner C, van der Spoel D, Lindahl E. GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Mol. Simul. J Chem Theory Comput. 2008; 4:435–447. [PubMed: 26620784]
36. Brooks BR, et al. CHARMM: The biomolecular simul programs. J Comput Chem. 2009; 30:1545–1614. [PubMed: 19444816]
37. Eastman P, et al. OpenMM 4: A Reusable, Extensible, Hardware Independent Library for High Performance Mol. Simul. J Chem Theory Comput. 2013; 9:461–469. [PubMed: 23316124]
38. Phillips JC, Braun R, Wang W, Gumbart J, Tajkhorshid E, Villa E, Chipot C, Skeel RD, Kale L, Schulten K. Scalable molecular dynamics with NAMD. J Comput Chem. 2005; 26:1781–1802. [PubMed: 16222654]
39. Walt, Svd, Colbert, SC., Varoquaux, G. The NumPy Array: A Structure for Efficient Numerical Computation. Computing in Science & Engineering. 2011; 13:22–30.
40. Fang B, Boross PI, Tozser J, Weber IT. Structural and Kinetic Analysis of Caspase-3 Reveals Role for S5 Binding Site in Substrate Recognition. Journal of Mol Biol. 2006; 360:654–666. [PubMed: 16781734]
41. Maier JA, Martinez C, Kasavajhala K, Wickstrom L, Hauser KE, Simmerling C. ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. J Chem Theory Comput. 2015; 11:3696–3713. [PubMed: 26574453]
42. Ryckaert JP, Ciccotti G, Berendsen HJC. Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. J Comput Phys. 1977; 23:327–341.
43. Loncharich RJ, Brooks BR, Pastor RW. Langevin dynamics of peptides: The frictional dependence of isomerization rates of N-acetylalanyl-’-methylamide. Biopolymers. 1992; 32:523–535. [PubMed: 1515543]
44. Berendsen HJC, Postma JPM, van Gunsteren WF, DiNola A, Haak JR. Molecular dynamics with coupling to an external bath. J Chem Phys. 1984; 81:3684–3690.
45. The PyMOL Molecular Graphics System.
46. Humphrey W, Dalke A, Schulten K. VMD – Visual Molecular Dynamics. J Mol Graphics. 1996; 14:33–38.
47. Ramsey S, Nguyen C, Salomon-Ferrer R, Walker RC, Gilson MK, Kurtzman T. Solvation thermodynamic mapping of molecular surfaces in AmberTools: GIST. J Comput Chem. 2016; 37:2029–2037. [PubMed: 27317094]
48. Sharp KA, Vanderkooi JM. Water in the Half Shell: Structure of Water, Focusing on Angular Structure and Solvation. Acc Chem Res. 2010; 43:231–239. [PubMed: 19845327]
49. Eastman P, et al. OpenMM 4: A Reusable, Extensible, Hardware Independent Library for High Performance Mol. Simul J Chem Theory Comput. 2013; 9:461–469.

50. Oliphant TE. Python for Scientific Computing. Computing in Science Engineering. 2007; 9:10–20.
51. Fenley AT, Henriksen NM, Muddana HS, Gilson MK. Bridging Calorimetry and Simulation through Precise Calculations of Cucurbituril-Guest Binding Enthalpies. J Chem Theory Comput. 2014; 10:4069–4078. [PubMed: 25221445]
52. Lazaridis T, Karplus M. Orientational correlations and entropy in liquid water. J Chem Phys. 1996; 105:4294–4316.
53. Wang L, Abel R, Friesner RA, Berne BJ. Thermodynamic Properties of Liquid Water: An Application of a Nonparametric Approach to Computing the Entropy of a Neat Fluid. J Chem Theory Comput. 2009; 5:1462–1473. [PubMed: 19851475]
54. Nguyen CN, Kurtzman T, Gilson MK. Spatial Decomposition of Translational Water–Water Correlation Entropy in Binding Pockets. J Chem Theory Comput. 2016; 12:414–429. [PubMed: 26636620]
55. Huggins DJ. Estimating Translational and Orientational Entropies Using the k-Nearest Neighbors Algorithm. J Chem Theory Comput. 2014; 10:3617–3625. [PubMed: 26588506]
56. Raman EP, MacKerell AD. Spatial Analysis and Quantification of the Thermodynamic Driving Forces in Protein–Ligand Binding: Binding Site Variability. J Am Chem Soc. 2015; 137:2608–2621. [PubMed: 25625202]
57. Balius TE, Fischer M, Stein RM, Adler TB, Nguyen CN, Cruz A, Gilson MK, Kurtzman T, Shoichet BK. Testing inhomogeneous solvation theory in structure-based ligand discovery. Proc Natl Acad Sci US A. 2017; 114:E6839–E6846.
58. Frisch, HL., Lebowitz, JL. The equilibrium theory of classical fluids : a lecture note and reprint volume. New York: W.A. Benjamin; 1964.
59. Chandler D, McCoy JD, Singer SJ. Density functional theory of nonuniform polyatomic systems. I. General formulation. J Chem Phys. 1986; 85:5971–5976.

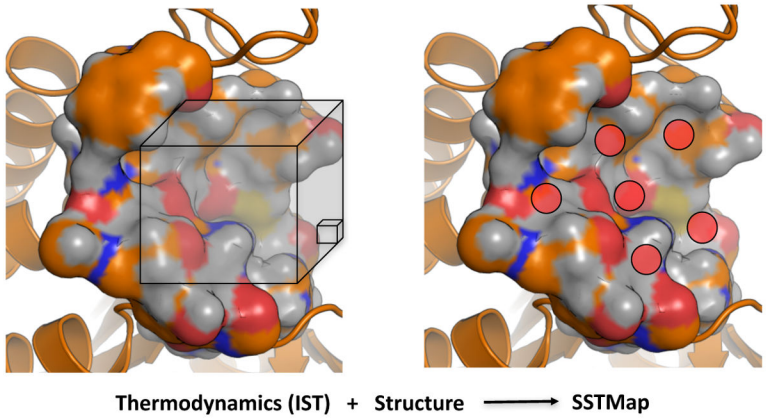


Figure 1. An illustration of SSTMap Functionality

SSTMap calculates structural and thermodynamic properties of water molecules on regions of solute surfaces that are represented as either a grid of voxels (left) or a set of high-density hydration sites (right).

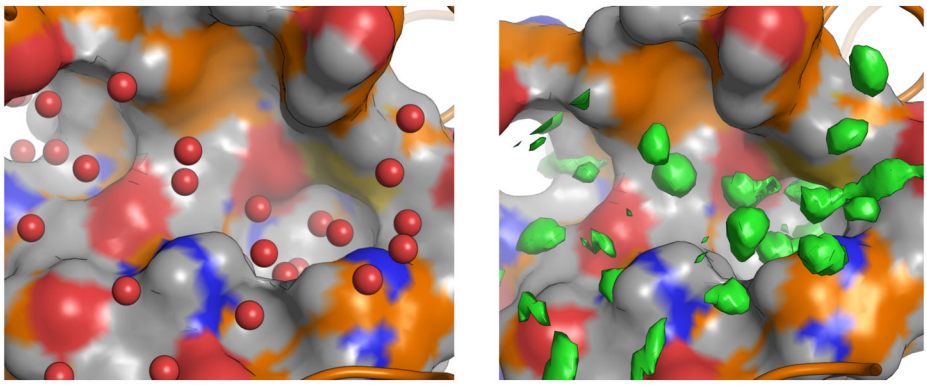


Figure 2. Visualization of SSTMap output for Caspase 3 active site

Left, Hydration sites identified by HSA calculations. Right, Three-dimensional maps of water density contoured at three times the bulk density, obtained by GIST calculations

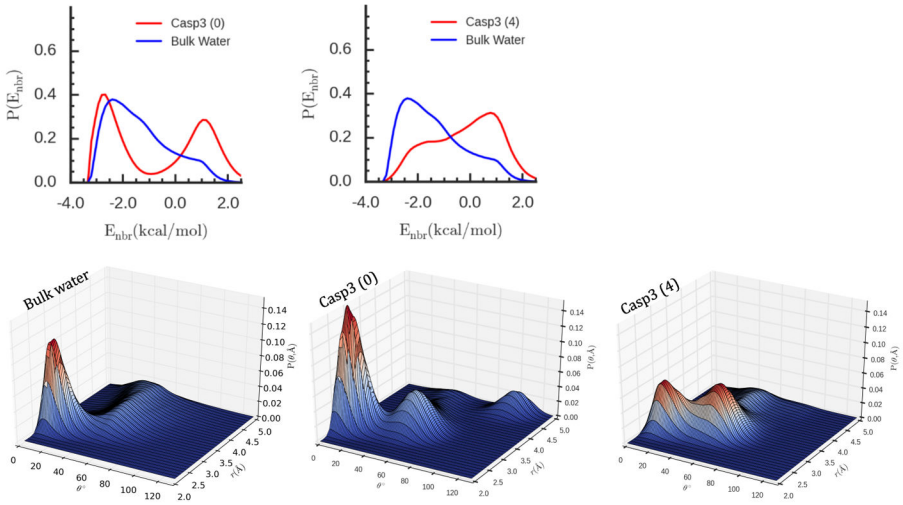


Figure 3. Analysis of local water structure in SSTMap

(Top) Probability distributions of pair energies of water molecules in hydration sites with their first shell neighbors (shown for Caspase 3 sites 0 and 4, overlaid with the same distribution for bulk water). Bottom row (from left to right): Two dimensional distributions of the angle and distance between hydration site water molecules and their neighbors for bulk water, Caspase 3 site 0, and 4.

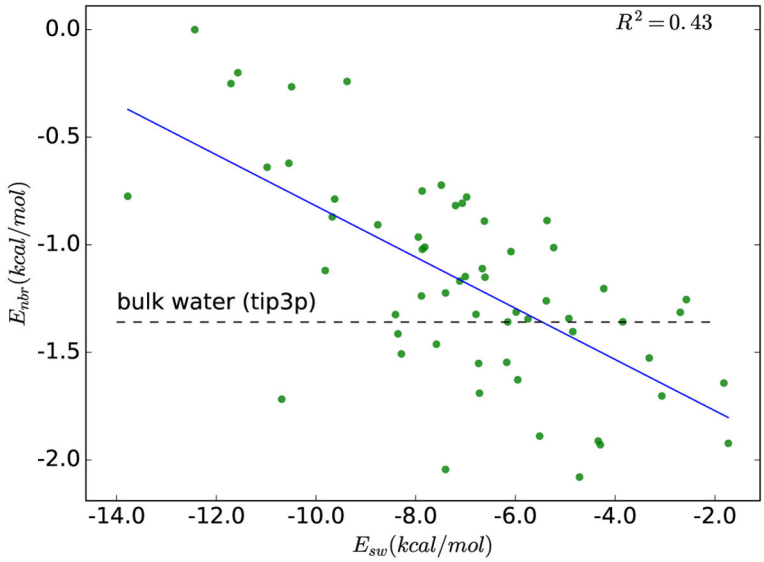


Figure 4. The effect of solute-water interaction on local water structure

The scatter plot is generated from a set of hydration sites where water molecules form hydrogen bonds with the charged side chains in active sites.

Table 1

Structural and thermodyamic quantities calculated by SSTMap.

Quantity	Description
Structural Quantities	
N_{nbr}	Average number of first shell neighbors
N_{ww}^{HB}	Average number of water-water hydrogen bonds
N_{sw}^{HB}	Average number of solute-water hydrogen bonds
E_{nbr}^{ww}	Average water-water interaction energy per neighbor
$N_{ww}^{HB,don}$	Number of water-water hydrogen bonds (donated)
$N_{ww}^{HB,acc}$	Number of water-water hydrogen bonds (accepted)
$N_{sw}^{HB,don}$	Number of solute-water hydrogen bonds (donated)
$N_{sw}^{HB,acc}$	Number of solute-water hydrogen bonds (accepted)
f_{ww}^{HB}	Fraction of hydrogen-bonded neighbors
Thermodynamic Quantities	
E_{ww}	Average water-water interaction energy
E_{sw}	Average solute-water interaction energy
E_{tot}	Average total energy
$TS_{sw,trans}$	Solute-water translational entropy
$TS_{sw,orient}$	Solute-water orientational entropy
$TS_{sw,tot}$	Total solute-water entropy

Table 2

Command-line arguments for running SSTMap programs `run_hsa` and `run_gist`.

Option	Description
Required	
-i	Input parameter file
-t	Input trajectory
-l	PDB file containing a ligand molecule
-g[‡]	Number of grid voxels in each direction
-f	Total number of frames to perform calculation on
Optional	
-p	Additional parameter files to extract non-bonded parameters (required for some MD packages)
-s	Starting frame for calculations (Default: 0)
-d	Reference bulk density (Default: 0.034, equivalent to 1g/cc)
-c[†]	PDB file containing coordinates of hydration site centers, if already determined.
-o	A prefix for output files.

[†]Option specific to `run_hsa`.

[‡]Option specific to `run_gist`.