An online repository of solvation thermodynamic and structural maps of SARS-CoV-2 targets

Brian Olson[3,4], Anthony Cruz[1,2], Lieyang Chen[1,3], Mossa Ghattas[1,2], Yeonji Ji[1,3], Kunhui Huang[1,3], Daniel J McKay[5], and Tom Kurtzman[1,2,3]

Correspondence to: Tom Kurtzman (E-mail: *thomas.kurtzman@lehman.cuny.edu*)

[1] Lehman College Department of Chemistry, 205 W Bedford Park Blvd Bronx, NY, United States of America, 10468
[2]Ph.D. Program in Chemistry, The Graduate Center of the City University of New York, 365 5th Avenue, New York New York, United States of America, 10016
[3]Ph.D. Program in Biochemistry, The Graduate Center of the City University of New York, 365 5th Avenue, New York New York, United States of America, 10016
[4]County College of Morris, Department of Biology and Chemistry, 214 Center Grove Rd, Randolph, NJ, United States of America, 07869
[5]Ventus Therapeutics, 7150 Frederick-Banting Montreal, Quebec H9S 2A1

ABSTRACT:

SARS-CoV-2 recently jumped species and rapidly spread via human-to-human transmission to cause a global outbreak of COVID-19. The lack of effective vaccine combined with the severity of the disease necessitates attempts to develop small molecule drugs to combat the virus. COVID19_GIST_HSA is a freely available online repository to provide solvation thermodynamic maps of COVID-19-related protein small molecule drug targets. Grid Inhomogeneous Solvation Theory maps were generated using AmberTools cpptraj-GIST and Hydration Site Analysis maps were created using SSTmap code. The resultant data can be applied to drug design efforts: scoring solvent displacement for docking, rational lead modification, prioritization of ligand- and protein- based pharmacophore elements, and creation of water-based pharmacophores. Herein, we demonstrate the use of the solvation thermodynamic mapping data. It is hoped that this freely provided data will aid in small molecule drug discovery efforts to defeat SARS-CoV-2.

1. Introduction

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) recently emerged and spread to cause a pandemic of coronavirus disease 2019 (COVID-19). Given the failure to contain the initial outbreak, the global failure to restrain the pandemic, and the absence of an effective vaccine, we may need to identify existing drugs or develop new drugs to interrupt COVID-19 at a critical juncture.

A number of targets may be of interest for the development of small molecule therapeutics for COVID-19: main protease (M^{pro}, 3CL^{pro}), helicase (Nsp13), endoribonuclease (Nsp15), and 2'-O-methyltransferase (Nsp10/16) are known viral protein drug targets for SARS-CoV-2. Small molecule drugs may target the substrate binding site of M^{pro}, the ADP binding site of Nsp13, the active site of Nsp15, or the S-adenosylmethionine (SAM) binding site of Nsp16.

Water is essential to the description of interactions between drugs and their biomolecular targets because solvation is a key contributor to molecular recognition and binding. Energies, entropies, and structural features of water molecules can be used to identify waters that may produce favorable or unfavorable contributions to the free energy of binding upon displacement and therefore aid in the identification of ligand interactions that may or may not be desirable. Water networks and tightly bound structural waters can affect ligand-receptor binding affinities. Information on water structure and thermodynamics may be useful to screen virtual compound databases, to identify new lead drug candidates, and inform rational lead modification to improve affinity and specificity for its target[1,2,3,4]. Ignoring water molecules in binding sites may reduce the chance that a drug design project will be successful.

Solvation thermodynamic mapping (STM) is widely used in academic studies of drug-protein interactions and has been widely integrated into the workflow of drug discovery and rational design efforts at major pharmaceutical companies. The utility of STM spans a number of areas in early-stage drug development efforts including virtual screening[1,2], formation or improvement of pharmacophores[2,5], docking[1,6], and rational lead modification[3,4]. While the utility of STM is apparent, there are significant obstacles to widespread use. Of particular concern is that many existing software packages for characterizing water properties are commercial and, hence, not available to all and/or they require computational expertise in molecular dynamics, computer modeling, and statistical mechanics in order to apply. This set of skills often does not exist in wet chemistry labs whose research is dedicated to discovering and optimizing new pharmaceutical compounds.

The goal of this publication is to remove these obstacles and make publicly available solvation thermodynamic and structural maps of SARS-CoV-2 targets as a resource to the academic and industrial drug design community to aid in their pursuit of identifying small molecule treatments for COVID-19. In order to aid in screening and modification of drugs, we offer a free public repository of solvation thermodynamic maps of significant small molecule COVID-19 drug targets. Here we present solvation maps of 7 targets that are likely viable for small molecule modulation. All maps and simulation data are publically available on the KurtzmanLab github (github): github.com/KurtzmanLab/COVID19_GIST_HSA.

2. Methods

2.1 Protein Preparation

Protein monomer structures were prepared using the Protein Preparation Wizard[7] in Maestro[8] with default settings. ACE and NMA groups were used to cap the protein termini. Active sites were visually inspected and compared to ligand-bound structures to ensure that protonation states and conformations were consistent with known ligand-protein interactions. All proteins were left as-is except for 6YB7, for which side chain rotamers were adjusted so as not to interfere with the binding of an aligned N3 ligand from 6LU7[9]. His164 was changed from being protonated in the delta position (HID) to the epsilon position (HIE) to reproduce the known protein ligand interaction. The protein preparation wizard also suggested two conformations for Met166 in 6YB7 and both were used. No changes were made for other proteins. Energy minimization for hydrogen atoms was then performed in Maestro.

A second set of structure models for SARS-CoV-2 M^{pro} (PDB IDs: 6YB7 and 6W63) were manually prepared by one of the authors (McKay). All histidine side chains were assigned as either HIE or HID given the local environment. All asparagine and glutamine side chains were examined and found to be in reasonable rotameric states. For these systems, the PARM@FROSST small molecule extension to ff14SB[10] and AM1-BCC[11] charges were used for the ligands.

2.2 Molecular dynamics simulations

Molecular dynamics simulations were performed in GPU accelerated AMBER 16[12] using the ff14SB[10] force field and the optimal point charge (OPC) model[13] of water. Ligand force field parameters were assigned with the general AMBER force field (GAFF)[14] using the Antechamber package[15] in AmberTools. Antechamber assigns charges, missing bonds, angles, dihedral angles and Lennard-Jones parameters for each atom. Ligand charges were assigned using AM1-BCC[11].

For systems with a co-crystalized ligand, the ligand was removed from the protein, and then the protein was solvated in a box of OPC water molecules with dimensions that ensured there were at least 10 Å between any atom of the protein and the box edge. Sodium or chlorine counterions were added accordingly to neutralize the system. Each system was then energetically minimized in a two-step process. The first minimization step was performed with 1500 steps of steepest descent with all protein atoms restrained harmonically using a force constant of 100 kcal/mol·Å^2. For the second minimization step, only main chain heavy atoms were restrained. Following minimization, the system was heated to 300 K in a 240 ps NVT simulation with the main chain heavy atoms restrained; the temperature was regulated by Langevin thermostat with collision frequency of 1 ps. This was followed by a 20 ns NPT simulation with the atom restraints declining from 100 Kcal/mol·Å^2 to 2.5 Kcal/mol·Å^2 in the first 10 ns. After that, a 50 ns NPT production simulation was conducted with the frames saved every 2 ps. In the production phase, the temperature was regulated via a Langevin thermostat set to 300 K with a collision frequency of 2 ps. The constant pressure (1 atm) was maintained by isotropic position scaling with a relaxation time of 0.5 ps.

2.3 GIST

GIST maps were created using the GPU port[16] of AmberTools cpptraj-GIST[17]. Analyses were performed on the complete 50 ns production trajectory for each system (25000 configurational snapshots). For each system, maps were created in a cubical region with 30 Å length sides centered on the geometric mean position of the co-crystalized ligand for the pdb (see Figure 1). The resolution of the grid was 0.5 Å (0.125 Å^3 per voxel). For structures with no co-crystalized ligand for the pdb entry, a homologous protein with a co-crystalized ligand was structurally aligned to the pdb structure and the geometric center of that ligand was used to define the GIST analysis region. In the case of 6JYT, the region was defined for HSA by a partial set of the residues found in the active site (K288, S289, D374, E375, R567). For the GIST analysis of 6JYT, the geometric center of ADP from a structurally aligned 2XZL was used as the center of the box. The ligands used for defining the GIST region for each structure can be found in the repository.

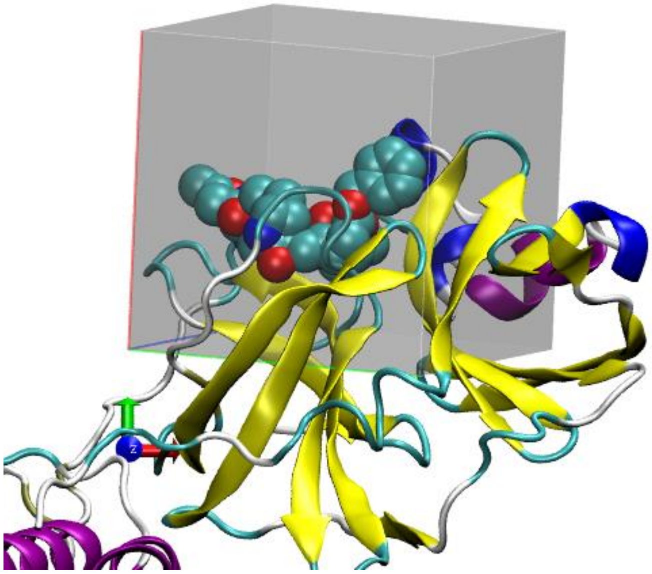

Figure 1: The co-crystalized structure of Mpro (cartoon) with ligand N3 from 6LU7. The GIST analysis was performed in the cubical region shaded in gray.

2.4 Hydration Site Analysis

Hydration Site Analysis (HSA)[18] was performed using the publicly available SSTmap code[19] with the default settings except for the region analysis which was set to within 10 Å of the ligand (*-d 10*). For each system, the analysis was per the first 20 ns (10,000 frames) of the MD production run for each protein.

Briefly, the method analyzes all the water positions from an MD trajectory and identifies high-density 1 Å radius spherical regions called hydration sites. In each hydration site, average quantities of the water molecules found in the hydration site are calculated and provide estimates for the local IST thermodynamic quantities. A number of measures that describe the local solvent structure and characterize the hydrogen-bonding environment of the water in each hydration site are also calculated. These measures can be used to characterize the enhancement or disruption of local water structure, describe the local enclosure, and describe the average hydrogen bonding interactions that water has in each hydration site with both its water neighbors and protein. Full details of the calculations are specified in a previous publication and the code is available on the github.

We also use newly developed code to determine the most probable orientations for water molecules in each hydration site. To do this, the orientations of all water molecules in each hydration site are clustered using a quaternion distance metric and the centroid orientation of each high-density cluster (generally at least 10% of the population) is recorded. The code and complete details of the method are in the github.

3 Repository contents

Structures 1-5 (SARS-CoV-2 structures)
Main Protease (M^{pro}, $3CL^{pro}$): 6LU7[9] (2.16 Å), 6YB7 (1.25 Å), 6M03 (2.00 Å), 6Y84 (1.39 Å), 6W63 (2.10 Å). Target the substrate binding site of M^{pro}.

Structure 6 (SARS-CoV-1 structure)
Helicase (Nsp13): 6JYT[20] (2.80 Å). Target (1) the ADP binding site but discourage (2) the nucleic acids binding site. No SARS-CoV-2 structure exists for this protein.

Structure 7 (SARS-CoV-2 structure)
Nsp16 (2'-O-methyltransferase, nsp 10/16): 6W4H (1.80 Å). Target the S-adenosylmethionine (SAM) binding site.

All files with prepared structures, topologies files, and molecular dynamics input and restart files are provided as well as solvation structural and thermodynamic maps described below.

3.1 Solvation thermodynamic maps

Inhomogeneous solvation theory (IST)[21,22,23] provides the statistical mechanical framework for the solvation thermodynamic quantities from explicit solvent molecular dynamics simulations. Here, we use two methods: Grid-based Inhomogeneous Solvation Theory (GIST)[24,25] and Hydration Site Analysis (HSA)[18] to localize the IST thermodynamic quantities onto a three-dimensional grid and onto high density 1 Å radius spherical "hydration sites", respectively. These localization approaches both process snapshots of the system configurations generated in molecular dynamics simulations to estimate local IST thermodynamic quantities including local energies, entropies, and number densities.

3.1.1 Grid based solvation maps

The repository contains grid-based solvation maps of calculated IST entropies, energies, and densities in Data Explorer (dx) format. The dx format enables visualization in standard graphics packages such as VMD and Pymol. For each target, energetic maps are provided for water's interactions with the protein, with other water molecules, and the total interactions of the water in each voxel with the system as a whole. GIST provides entropy maps for the total entropy as well maps that separately include the translational and orientational contributions to the total entropy. Maps are provided for all of the entropy and energy quantities for both normalized (per water quantities) and density (per voxel) quantities. A complete list of quantities can be found in table 1. Detailed descriptions of these quantities can be found in our prior work[17],[25].

Table 1 Key GIST quantities

Quantity	Description	Units
[a]TS_{trans}	Translational entropy density	kcal/mol/Å^3
[a]TS_{orient}	Orientational entropy density	kcal/mol/Å^3
[a]TS_{six}	Total entropy density	kcal/mol/Å^3
[a]TS_{trans}	Translational entropy density	kcal/mol/Å^3
[a]TS_{orient}	Orientational entropy density	kcal/mol/Å^3
[a]E_{ww}	Water-water energy density	kcal/mol/Å^3
[a]E_{sw}	Solute-water energy density	kcal/mol/Å^3
Neighbor count	Mean number of water neighbors[c]	molecules

[a] Corresponding normalized quantities also reported [b] Dipole moments are reported as time-averaged x,y, and z,components, along with the mean overall magnitude. [c] Neighbors are defined as two water molecules with an O-O distance of 3.5Å or less.

3.1.2 Hydration site solvation maps

For each target, the positions and calculated thermodynamic and structural quantities for the water in each hydration site are summarized in a space delimited spreadsheet file.

The same energetic quantities as calculated for GIST (above) are calculated for each hydration site and reported in per water (normalized) units. Additionally, the HSA data includes a breakdown of the total energy into contributions from Lennard-Jones, electrostatic, and first solvation shell water-water interactions.

SSTMap also calculates a number of quantities that are aimed at characterizing the local environment surrounding each hydration site. These are aimed at better describing local water structure and the interactions of the water in the hydration site with the protein surface.

Table 2 HSA structural quantities

Quantity	Description	Units
N_{nbr}	Average # first shell neighbors	None
N_{ww}^{HB}	Average # water-water hydrogen bonds	None
N_{sw}^{HB}	# solute-water hydrogen bonds	kcal/mol
E_{nbr}^{ww}	Average water-water interaction energy by neighbor	Kcal/nbr
$N_{ww}^{HB,don}$	# water-water hydrogen bonds donated	None
$N_{ww}^{HB,acc}$	# water-water hydrogen bonds accepted	None
$N_{sw}^{HB,don}$	# solute-water hydrogen bonds donated	None
$N_{sw}^{HB,acc}$	# solute-water hydrogen bonds accepted	None
f_{ww}^{HB}	Fraction of hydrogen-bonded neighbors	None

Quantities that provide a measure of local water structure include the average number of first shell neighbors each water has in its first solvation shell, the fraction of these neighbors to which the hydration site water is hydrogen bonded, and the average energy of interaction with each neighboring water. When compared to bulk water values, these quantities provide measures of whether the local water structure is enhanced or frustrated[26].

Additional quantities that characterize the interaction of the water in each hydration site with the protein include: (1) an enclosure parameter that describes how much of the region around the hydration site is protein and how much is water, (2) the average number of hydrogen bond donor and acceptor interactions that water molecules found in the hydration site have with the protein surface, and (3) lists of the protein residues that donate and accept hydrogen bonds to the water in the hydration site.

A list of thermodynamic and structural quantities can be found in Table 3. A text delimited spreadsheet file summarizing all calculated water properties is found in the HSA directory for each protein.

In addition, to facilitate visualization, each HSA directory includes pdb files that feature (1) the hydration site centers, (2) water molecules located at the center of each hydration site that have the most probable orientation, and (3) water molecules located at the center of each hydration site that include all probable orientation clusters.

4 Potential applications

Solvation thermodynamic mapping has been used in a variety of applications aimed at aiding the discovery and design of new pharmaceutical compounds. In docking, scoring terms have been added to explicitly account for solvent displacement upon ligand binding and the modified docking scoring functions have been used to help improve AUC, pose prediction, and identify novel binding ligands[1,6,27]. Solvation maps have also been used to create pharmacophores[2] as well as provide criteria to prioritize the selection of pharmacophore sites[5]. Both water thermodynamics and water interactions with protein surfaces have been used to direct lead modification[4,28].

Here, we describe by example several potential applications for the solvation maps provided in this repository.

4.1 Rational lead modification

The properties of water in and around the binding site may be used to direct the design of chemical modifications to a lead compound or fragment. The physical principles of this are that the displacement of thermodynamically unfavorable surface water upon the binding of a ligand will lead to favorable contributions to the free energy as the water is displaced to the more thermodynamically favorable environment of bulk biological water.

Here, we illustrate how solvation structural and thermodynamic solvation mapping in this repository can be used to provide insight into which modifications may lead to boosts in binding affinity.

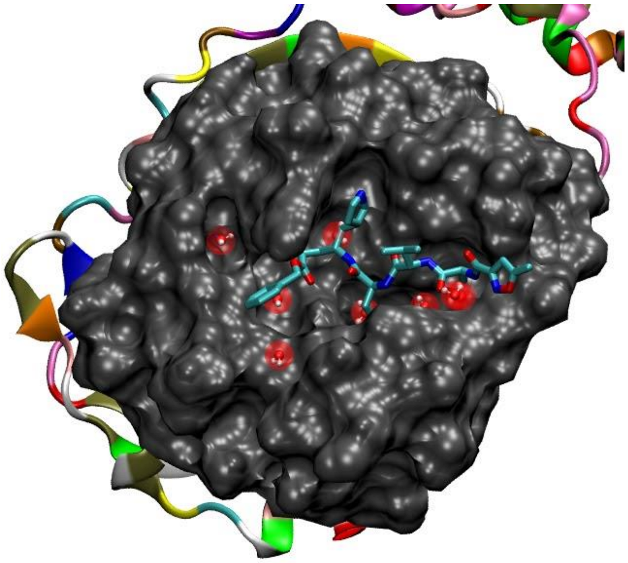

Figure 2: N3 bound to Mpro (PDB ID: 6LU7). Hydration sites that are located within 7.5 angstroms of N3 and have highly unfavorable energy ($\Delta E > 0.5$ kcal/mol with respect to neat water) are shown as transparent red spheres. The most probable water orientation for each hydration site is represented by a water molecule at the center of each sphere. The protein surface proximal (within 11 Å) to N3 is shown in gray.

The binding site of M^{pro} features a large number of energetically unfavorable hydration sites (see Figure 2). Prior work[29,30] suggests that the displacement of water from these hydration sites may be correlated with differences in binding affinities between congeneric pairs of ligands. Most of the hydration sites identified in figure 1 are displaced by N3. However, the two leftmost sites are not. We will focus on the upper left site, hydration site 7 (HS7), as it has an exceptionally unfavorable thermodynamic profile.

HS7 occupies a small cleft on the surface of the protein, which is formed by 7 different residues (28, 143, 119, 26, 118, and 145). The water in this cleft is resolved the crystal structures of 6LU7, 6W63, 6Y84, and 6YB7. However, this water is not reported in 6M03. The water is highly enclosed by the protein

(81.7%) having slightly less than one (0.96) water neighbor, on average, in its first solvation shell. Despite the hydration site being highly occupied (84.5% occupancy), the water is exceptionally unfavorable energetically (+2.6 kcal/mol) and entropically (-TS of 4.45 kcal/mol) by IST estimates. Its low entropy result is based on the water's highly restricted translational and orientational motion. The water's high enclosure in the protein cleft and its formation of two hydrogen bonds with the protein surface severely restrict the water's translational freedom leading to a translational entropic penalty of 2.11 kcals by IST estimates. The two hydrogen bonds it forms with the protein surface as well as forming a hydrogen bond 82 percent of the time with its water neighbor located above the cleft (HS56), further restrict its orientational freedom resulting in an entropic penalty of 2.33 kcals/mole.

Despite being on a hydrophilic surface (forming on average 2.00 hydrogen bonds with the protein), the water in HS7 cannot form a full complement of hydrogen bonds, instead forming only 2.85 geometric hydrogen bonds on average compared to a bulk OPC water which would form 3.62. This deficiency of more than three quarters of a hydrogen bond, on average, is a significant contribution to the unfavorable energetic profile (+2.6 kcals/mole overall) of HS7.

Both the unfavorable IST energy and entropy suggest that displacing this HS7 water could lead to gains in binding affinity. In order to displace this water, an optimal chemical group must replace interactions that the water makes with the protein without disrupting the hydrogen bond network that the water is making with its neighbors. As the water in HS7 is located in a cleft, any chemical group would also need to displace its water neighbor (h-bonded water in figure 3). The optimal chemical group would need to both donate a hydrogen bond to the backbone carbonyl of Gly143 and accept a hydrogen bond from the backbone amine of Asn119. A hydroxy group seems ideal for this.

All of the numerical data in the above analysis is located in the HSA summary spread sheet for 6LU7 (6LU7_apo_flex_hsa_ summary.csv). All the data for the visualizations is likewise located in the repository.

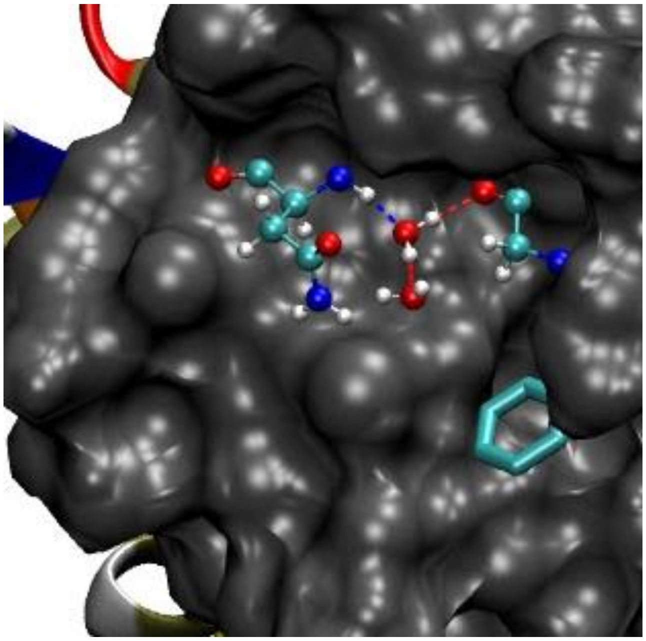

Figure 3: The most probable orientation of the water in HS7 donates a hydrogen bond (red dashed line) to the backbone carbonyl of Gly143, accepts a hydrogen bond (blue dashed line) from the backbone NH of Asn119, and donates a hydrogen bond to HS56 above the cleft wherein lies HS7.

4.2 Scoring Solvation Displacement in Docking

Four studies outline how solvation thermodynamic mapping can be used to aid in the discovery of new leads in docking. The first two of these studies are based on our prior work on Factor Xa[29,30] in which a displaced solvent functional used high energy and high density voxels as functional inputs to correlate with experimental measurements of differences in binding free energies between congeneric pairs of ligands[29,30]. The third docking study[6] by Uehara and Tanaka instead used a displaced solvent functional with free energetic maps created by GIST as input whereas the fourth study[1] by Balius et al. used the displacement of voxels with high energy densities as input. The third study showed improvements in pose prediction and enrichment and the fourth showed only nominal measurable improvements to docking enrichment and pose prediction, though the method was successfully used to prospectively identify new tightly binding compounds, including the tightest binding compound to cytochrome c peroxidase. A map showing related unfavorable and favorable energy density regions for M^{pro} is shown in figure 4.

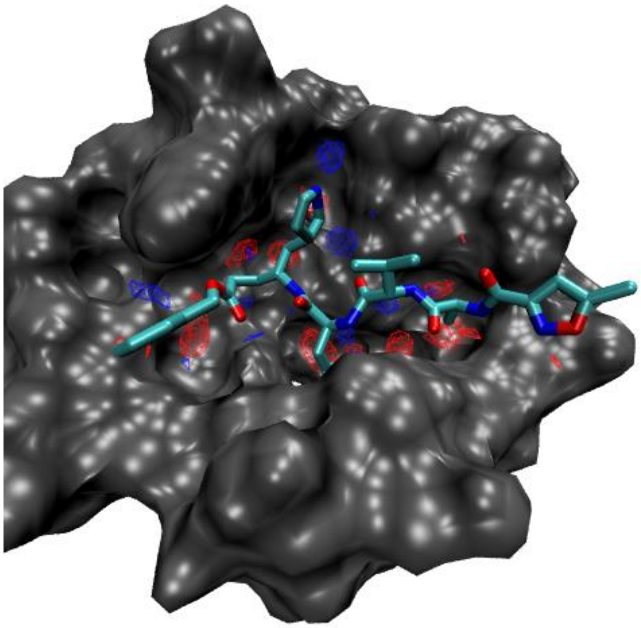

Figure 4: Unfavorable and Favorable Solvation Energy Density Map of M^{pro}. Regions of unfavorable energy density (E_{dens} > 0.1 kcal/mole/ Å^3) and favorable energy density ((E_{dens} > 0.1 kcal/mole/ Å^3) are shown in red or blue wireframe, respectively. The predicted score for a docked ligand would be penalized for displacing water from the favorable blue regions or given an affinity boost for displacing water from the red regions.

The GIST maps in this repository provide the data to create the maps used in all three of the GIST-based studies. Necessary modifications of the provided GIST dx maps (e.g. creating a free energy density map from the energy and entropy density maps) can be easily created using the GIST Post-Processing (GISTPP) code provided on the github.

4.3 Pharmacophore Creation

Solvation mapping can be used to generate water-based pharmacophore hypotheses[2] and to prioritize ligand- or protein-based pharmacophore sites[5]. Here we combine several interesting hydration sites with ligand-based pharmacophore elements.

Three pharmacophore sites were constructed using ligand-protein interactions based on analyses of co-crystalized ligands found inside the binding sites of SARS-CoV-2 M^{pro} structures (PDB ID: 6W63, 6LU7, 6Y2F, 6Y2G, and 6M2N). These ligand-based sites appear as dotted spheres in Figures 5 and 6.

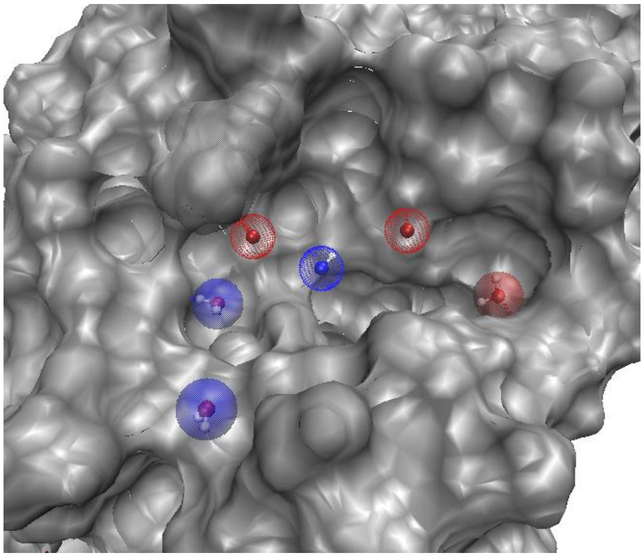

Figure 5: Hybrid ligand- and water-based pharmacophore within the binding site of M^{pro} (PDB ID: 6LU7). The ligand-based sites are shown as dotted spheres and the water-based sites are shaded spheres. Ligand-based sites have an NH group for donors or an oxygen for acceptors. The most probable water orientation is found at the center of each water-based pharmacophore site. Acceptor sites are red and donor sites are blue spheres.

The leftmost ligand-acceptor site (figures 5 & 6) lies inside the oxyanion hole. All five of the co-crystalized ligands accept a hydrogen bond from the backbone amino group of Gly143 while three of five (6Y2F, 6Y2G, and 6M2N) also accept a hydrogen bond from Cys146. The pharmacophore site shown in figure 6 shows both of these interactions. The middle ligand-based site donates a hydrogen bond to the backbone carbonyl of His164. Ligands from 6W63, 6Y2G and 6Y2F make this contact. The rightmost ligand site, inside the S1 subsite, accepts a hydrogen bond from the backbone amino group of Glu166. Four of the five (all except 6M2N) co-crystallized ligands accept a hydrogen bond from this group. Each ligand-based site is proximal to a hydration site and GIST high-density group of voxels but none have any significant thermodynamic signal for use in prioritization. These ligand-based sites were chosen by the fact that they were well conserved across the limited number of structures available with co-crystallized ligands.

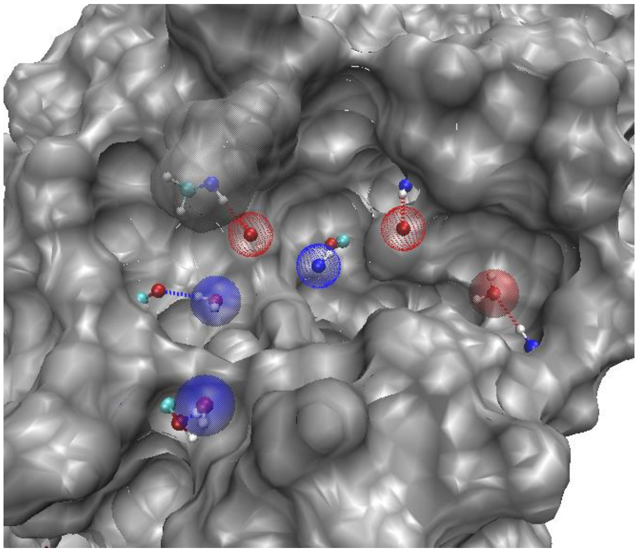

Figure 6: The same hybrid pharmacophore hypothesis as shown in Figure 5, except the interactions with chemical groups on the surface are shown explicitly. Blue dashed lines show the pharmacophore sites donation of hydrogen bonds and red dashed lines show acception.

We used hydration site analysis to add three additional ligand-based pharmacophores sites. These sites are shown in shaded spheres in figures 5 and 6. While water-based pharmacophore sites can be chosen using other criteria (as outlined in Jung et al.[2]), here we simply chose water-based sites that are energetically unfavorable and categorized them based on their donor/acceptor interactions with the protein surface.

The first site (on the far right of figures 5 and 6) is from HS9, which primarily accepts a hydrogen bond from the backbone amino group of residue Thr190 and has an unfavorable energy of -11.28 kcals/mole (almost 1 kcal above the bulk energy of -12.26 kcals/mole). The second site HS52 (middle left in Figures 5 & 6) has an energy of -10.46 kcals/mole (1.8 kcals above bulk energy) and donates a hydrogen bond to Thr26. The third site, HS56 has an energy of -11.69 kcals/mole (0.57 kcals less favorable than bulk) and donates a hydrogen bond to Thr25.

Together, the conserved ligand sites and the water-based sites create a pharmacophore hypothesis that can used to screen virtual compound databases.

While we arbitrarily chose three conserved sites from the ligand and three proximal hydration sites to construct the hypothesis outlined here. This approach allows a drug designer flexibility to choose ligand and water sites on virtually any solvent exposed surface of the protein, allowing different regions of the active site or potential allosteric sites to be targeted.

5. How to access data

All hydration site and GIST data is available on github with a readme.md that details directory structure and the descriptive file naming convention.

Briefly, each PDB structure has its own subdirectory named after its pdbid. Each pdbid subdirectory has further subdirectories for simulations with apo or complexed structures and different protein restraints. Additional subdirectories for each of these include the hydration site and GIST analyses, as well as the prepared protein input files and Amber MD restart files in case longer simulations are desired.

All of the above can be found on the github:

(github.com/KurtzmanLab/COVID19_GIST_HSA.)

5.1 Code availability

All water analysis code used to produce this data is open-source with extensive documentation and has been made available for download. These resources, combined with the provided prepared structures and input files, allow for all the data provided here to be reproduced.

Three sets of code were used for the water analysis in the repository: SSTMap, GIST-cpptraj, and GISTPP.

SSTMap was used for hydration site analyses, GIST-cpptraj was used for the GIST analyses, and GISTPP was used to make numerical manipulations to the GIST dx files. Usage tutorials and documentation can be found on the SSTMap project page (SSTMap.org) and on the AMBER website. GIST-cpptraj code is available on the Amber-MD github

https://github.com/Amber-MD

All other code is available on the github:

https://github.com/KurtzmanLab

Acknowledgments

We'd like to thank Daniel Roe for extensive help in integrating GIST into Amber Tools. José Duca, Camilo Velez-Vega, Callum Dickson, Andre Golosov, and the Novartis CADD team for provocative discussions on protein structure and preparation. This research was supported by the USA National Institutes of Health R01GM100946.

Keywords: SARS-CoV2, COVID-19, Solvation Thermodynamics, CADD, Drug Discovery, Drug Design, Docking, Virtual Screening, Rational Lead Modification, Lead Drug Design, Main Protease.

References and Notes

1. Balius, T. E. *et al.* Testing inhomogeneous solvation theory in structure-based ligand discovery. *Proc. Natl. Acad. Sci.* **114**, E6839–E6846 (2017).

2. Jung, S. W., Kim, M., Ramsey, S., Kurtzman, T. & Cho, A. E. Water Pharmacophore: Designing Ligands using Molecular Dynamics Simulations with Water. *Sci. Rep.* **8**, 10400 (2018).

3. Harriman, G. *et al.* Acetyl-CoA carboxylase inhibition by ND-630 reduces hepatic steatosis, improves insulin sensitivity, and modulates dyslipidemia in rats. *Proc. Natl. Acad. Sci.* **113**, E1796–E1805 (2016).

4. Collin, M.-P. *et al.* Discovery of Rogaratinib (BAY 1163877): a pan-FGFR. *ChemMedChem* **13**, 437–445 (2018).

5. Hu, B. & Lill, M. A. Protein Pharmacophore Selection Using Hydration-Site Analysis. *J. Chem. Inf. Model.* **52**, 1046–1060 (2012).

6. Uehara, S. & Tanaka, S. AutoDock-GIST: Incorporating Thermodynamics of Active-Site Water into Scoring Function for Accurate Protein-Ligand Docking. *Molecules* **21**, 1604 (2016).

7. Madhavi Sastry, G., Adzhigirey, M., Day, T., Annabhimoju, R. & Sherman, W. Protein and ligand preparation: parameters, protocols, and influence on virtual screening enrichments. *J. Comput. Aided Mol. Des.* **27**, 221–234 (2013).

8. *Schrödinger Release 2017-3*. (Schrödinger, LLC, 2017).

9. Jin, Z. *et al. Structure of Mpro from COVID-19 virus and discovery of its inhibitors*. http://biorxiv.org/lookup/doi/10.1101/2020.02.26.964882 (2020) doi:10.1101/2020.02.26.964882.

10. Maier, J. A. *et al.* ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *J. Chem. Theory Comput.* **11**, 3696–3713 (2015).

11. Jakalian, A., Jack, D. B. & Bayly, C. I. Fast, Efficient Generation of High Quality Atomic Charges. AM1 - BCC Model: II. Parameterization and Validation. *J Comput Chem* **23**, (2002).

12. D.A. Case, R.M. Betz, D.S. Cerutti, T.E. Cheatham, III, T.A. Darden, R.E. Duke, T.J. Giese, H. Gohlke, A.W. Goetz, N. Homeyer, S. Izadi, P. Janowski, J. Kaus, A. Kovalenko, T.S. Lee, S. LeGrand, P. Li, C., Lin, T. Luchko, R. Luo, B. Madej, D. Mermelstein, K.M. Merz, G. Monard, H. Nguyen, H.T. Nguyen, I., Omelyan, A. Onufriev, D.R. Roe, A. Roitberg, C. Sagui, C.L. Simmerling, W.M. Botello-Smith, J. Swails, & R.C. Walker, J. Wang, R.M. Wolf, X. Wu, L. Xiao and P.A. Kollman. *Amber 16*. (University of California, 2016).

13. Izadi, S., Anandakrishnan, R. & Onufriev, A. V. Building Water Models: A Different Approach. *J. Phys. Chem. Lett.* **5**, 3863–3871 (2014).

14. 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* **25**, (2004).

15. Wang, J., Wang, W., Kollman, P. A. & Case, D. A. Automatic atom type and bond type perception in molecular mechanical calculations. *J. Mol. Graph. Model.* **25**, 247–260 (2006).

16. Kraml, J., Kamenik, A. S., Waibl, F., Schauperl, M. & Liedl, K. R. Solvation Free Energy as a Measure of Hydrophobicity: Application to Serine Protease Binding Interfaces. *J. Chem. Theory Comput.* **15**, 5872–5882 (2019).

17. Ramsey, S. *et al.* Solvation thermodynamic mapping of molecular surfaces in AmberTools: GIST. *J. Comput. Chem.* **37**, 2029–2037 (2016).

18. Young, T., Abel, R., Kim, B., Berne, B. J. & Friesner, R. A. Motifs for molecular recognition exploiting hydrophobic enclosure in protein–ligand binding. *Proc. Natl. Acad. Sci.* **104**, 808–813 (2007).

19. Haider, K., Cruz, A., Ramsey, S., Gilson, M. K. & Kurtzman, T. Solvation Structure and Thermodynamic Mapping (SSTMap): An Open-Source, Flexible Package for the Analysis of Water in Molecular Dynamics Trajectories. *J. Chem. Theory Comput.* **14**, 418–425 (2018).

20. Jia, Z. *et al.* Delicate structural coordination of the Severe Acute Respiratory Syndrome coronavirus Nsp13 upon ATP hydrolysis. *Nucleic Acids Res.* **47**, 6538–6550 (2019).

21. Morita, T. & Hiroike, K. A New Approach to the Theory of Classical Fluids. III: General Treatment of Classical Systems. *Prog. Theor. Phys.* **25**, 537–578 (1961).

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

23. Lazaridis, T. Inhomogeneous Fluid Approach to Solvation Thermodynamics. 2. Applications to Simple Fluids. *J. Phys. Chem. B* **102**, 3542–3550 (1998).

24. Nguyen, C. N., Young, T. K. & Gilson, M. K. Grid inhomogeneous solvation theory: Hydration structure and thermodynamics of the miniature receptor cucurbit[7]uril. *J. Chem. Phys.* **137**, 044101–17 (2012).

25. Nguyen, C., Gilson, M. K. &. Young, T. Structure and Thermodynamics of Molecular Hydration via Grid Inhomogeneous Solvation Theory. *arXiv:1108.4876* (2011).

26. Haider, K., Wickstrom, L., Ramsey, S., Gilson, M. K. & Kurtzman, T. Enthalpic Breakdown of Water Structure on Protein

Active-Site Surfaces. *J. Phys. Chem. B* **120**, 8743–8756 (2016).

27. Murphy, R. B. *et al.* WScore: A Flexible and Accurate Treatment of Explicit Water Molecules in Ligand–Receptor Docking. *J. Med. Chem.* **59**, 4364–4384 (2016).

28. Abel, R. *et al.* Accelerating drug discovery through tight integration of expert molecular design and predictive scoring. *Curr. Opin. Struct. Biol.* **43**, 38–44 (2017).

29. 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.* **130**, 2817–2831 (2008).

30. Nguyen, C. N., Cruz, A., Gilson, M. K. & Kurtzman, T. Thermodynamics of Water in an Enzyme Active Site: Grid-Based Hydration Analysis of Coagulation Factor Xa. *J. Chem. Theory Comput.* **10**, 2769–2780 (2014).