5-2018

Computational Modeling of (De)-Solvation Effects and Protein Flexibility in Protein-Ligand Binding using Molecular Dynamics Simulations

Ying Yang
Purdue University

COMPUTATIONAL MODELING OF (DE)-SOLVATION EFFECTS AND PROTEIN FLEXIBILITY IN PROTEIN-LIGAND BINDING USING MOLECULAR DYNAMICS SIMULATIONS

A Dissertation

Submitted to the Faculty

of

Purdue University

by

Ying Yang

In Partial Fulfillment of the

Requirements for the Degree

of

Doctor of Philosophy

May 2018

Purdue University

West Lafayette, Indiana

THE PURDUE UNIVERSITY GRADUATE SCHOOL
STATEMENT OF DISSERTATION APPROVAL

Dr. Markus Lill, Chair

Department of Medicinal Chemistry and Molecular Pharmacology

Dr. Chiwook Park

Department of Medicinal Chemistry and Molecular Pharmacology

Dr. Jean-Christophe Rochet

Department of Medicinal Chemistry and Molecular Pharmacology

Dr. Daisuke Kihara

Department of Biological Sciences and Department of Computer Science

Approved by:

Dr. Zhongyin Zhang

Head of the Department Graduate Program

Dedicated to my late grandmother Zequan Zhou,
Who loved me and the family with all her life.
To my beloved family members,
Especially my mother Min Li and my father Yuan Yang,
Who have always supported me to purse my passion.
献给我已辞世的外婆周泽全,
感谢她用尽一生来爱我和我们的大家族,
至我敬爱的家人，尤其是我的父母杨源、李敏,
感谢他们始终鼓励我追求我的兴趣所在。

ACKNOWLEDGMENTS

It has been almost six years since I first come to Purdue University pursuing the doctoral degree. Through the journey, I have learned a lot academically and grown up personally. Thanks to family and friends for their constant encouragement and unconditional support along the way.

I owe my deepest gratitude to Dr. Markus Lill whose guidance and support made this work possible. He showed me how to think and work as a researcher. Also, I sincerely thank my graduate advisory committee Dr. Chiwook Park, Dr. Jean-Christophe Rochet, and Dr. Daisuke Kihara for their advices.

Many thanks go to all the Lill lab fellows. I am especially grateful to my senior lab member Dr. Bingjie Hu from who I gained and learned a lot. Also thanks to Amr, Laura, Neha for being good friends and creating collaborative and encouraging environment in the lab. I greatly appreciate you reading my manuscripts and giving valuable comments. I am also grateful to the undergraduate students Matt Masters, Drew McNutt, and Mark Choi making contribution to this work.

And finally, I would like to thank Xufeng Wang who loves the imperfect of me, and supports me during my most difficult time.

TABLE OF CONTENTS

Page

LIST OF TABLES . . . ix

LIST OF FIGURES . . . x

ABBREVIATIONS . . . xvi

GLOSSARY . . . xviii

ABSTRACT . . . xix

PUBLICATIONS . . . xxi

1 INTRODUCTION . . . 1

1.1 The Drug Discovery Process . . . 1

1.1.1 Computer-aided Drug Design . . . 1

1.1.2 A Water World: Accounting for Water Molecules in SBDD . . . 2

1.1.3 Everything is in Motion: Accounting for Protein Flexibility in SBDD . . . 2

1.1.4 Molecular Dynamics (MD) Simulations . . . 4

1.2 Individual Water Molecules in Drug Discovery . . . 5

1.3 Hydration Site Prediction with WATsite . . . 9

1.3.1 Molecular Dynamics Simulation . . . 9

1.3.2 Hydration Site Identification . . . 10

1.3.3 De-solvation Free Energy Estimation for Hydration Sites . . . 13

1.4 Examples of Hydration Site Prediction using WATsite . . . 16

1.4.1 Hydration Site Prediction with Ligand: Water at Binding Interface . . . 18

1.4.2 Hydration Site Prediction without Ligand: Water Displacement upon Ligand Binding . . . 18

1.5 Research Summary . . . 24

Page

2 FACTORS INFLUENCING HYDRATION SITE PREDICTION BASED ON MD SIMULATIONS 25

2.1 Introduction 25

2.2 Materials and Methods 26

2.2.1 Protein Systems and Preparation 26

2.2.2 MD Simulation and Theory of Hydration Site Analysis 26

2.2.3 Comparison between Hydration Sites 26

2.2.4 Dependence of Hydration Site Analysis on Simulation Length . 27

2.2.5 Generation of Different Starting Conformations 28

2.2.6 Estimation of De-solvation Free Energy of the Protein upon Ligand Binding 28

2.3 Results and Discussion 29

2.3.1 Dependence of Hydration Site Analysis on Simulation Length . 29

2.3.2 Sensitivity of Hydration Site Prediction on Initial Protein Structure32

2.3.3 Sensitivity of Protein De-solvation Free Energy Estimation on Initial Protein Structure and Distance Cutoff between Ligand Atoms and Hydration Sites 40

2.4 Conclusion 40

3 INCORPORATING PROTEIN FLEXIBILITY INTO EXPLICIT HYDRATION SITE PREDICTION 44

3.1 Introduction 44

3.2 Materials and Methods 46

3.2.1 Protein Systems and Preparation 47

3.2.2 MTMD Simulations of Ligand-Induced Protein Conformational Change 47

3.2.3 Extracting Trajectories of Overlapping MTMD Windows 49

3.2.4 Theory of Hydration Site Identification and de-solvation Free Energy Prediction 50

3.2.5 Development of Two Methods for the Hydration Site Analysis Involving Protein Flexibility 50

Page

3.2.6 Estimation of Buried Apolar Surface Area of Protein-Ligand Complexes . . . 57

3.2.7 Estimation of de-solvation Free Energy of the Protein upon Ligand Binding . . . 57

3.3 Results and Discussion . . . 59

3.3.1 Protein Conformational Change from the Apo to the Holo Form Simulated by MTMD . . . 60

3.3.2 Method I: Identify Hydration Site Transition Pathways through MTMD windows . . . 61

3.3.3 Method II: Associate Hydration Sites in Apo and Holo Forms using HSSCS . . . 69

3.3.4 Comparison Between Method I and Method II . . . 70

3.3.5 Method IIb: multiple-to-multiple hydration-site pairings . . . 72

3.3.6 Prediction of Protein De-solvation Free Energy including Protein Flexibility . . . 77

3.4 Conclusion . . . 80

4 GPU-ACCELERATION OF HYDRATION SITE PREDICTION PROGRAM WATSITE . . . 83

4.1 Introduction . . . 83

4.2 Materials and Methods . . . 84

4.2.1 Implementation of Water Interaction-Energy Computation in OpenMM . . . 84

4.2.2 Preparation of Protein Systems . . . 84

4.2.3 System Truncation . . . 85

4.2.4 MD Simulations . . . 85

4.2.5 Hydration Site Analysis . . . 86

4.2.6 Grid-based Water Analysis . . . 86

4.2.7 Convergence . . . 87

4.3 Results . . . 88

4.3.1 Validation of Hydration Site Prediction . . . 88

4.3.2 Grid-based Analysis of Water Energy . . . 88

Page

4.3.3 Energy Convergence of Hydration Site and Grid-based Analysis 91
4.3.4 System Truncation 93
4.3.5 Speed Increase with GPU-Accelerated Implementation of WATsite93
5 MODELING OF HALOGEN-PROTEIN INTERACTIONS IN CO-SOLVENT MOLECULAR DYNAMICS SIMULATIONS 97
5.1 Introduction 97
5.2 Materials and Methods 100
5.2.1 Preparation of Protein Systems and Co-solvent Probes 100
5.2.2 MD simulations 102
5.2.3 Fragment Map and Free Energy Map Generation 107
5.2.4 Ligand Scoring and Local MC Sampling 108
5.3 Results and Discussion 109
5.3.1 Convergence of Fragment FEMap 109
5.3.2 Inhibitor Halogen Atom Location Revealed by Fragment Map 111
5.3.3 Correspondence between LGFE and Experimental Affinity . . 111
5.4 Conclusion 113
6 FUTURE DIRECTIONS 118
6.1 Research Summary 118
6.2 Methodological Improvements for Hydration Site Prediction 119
6.2.1 Occluded Binding Sites 119
6.2.2 Protonation states of binding site residues 119
6.2.3 Influence of Polarization 120
6.3 Routine Consideration of Explicit Water Molecules in Drug Discovery Projects 120
6.3.1 MM-GB/SA 121
6.3.2 Docking and Scoring 121
REFERENCES 123
A WATsite User Guide with PyMOL Interface 134
VITA 152

LIST OF TABLES

Table Page

1.1 E_{bulk} estimated for five water models. Unit: kcal/mol. 15

3.1 Three types of hydration sites identified in the GUA system using method I.65

3.2 Three types of hydration sites identified in the SH2 system using method I. 66

3.3 Comparison of matched hydration sites between method I and II for the GUA system. 71

3.4 Heat map showing all hydration-site pairs identified using method IIb allowing for multiple pairing with the same hydration site. 76

3.5 Comparison of matched hydration sites between method I and II for the SH2 system. 78

3.6 Heat map showing all hydration-site pairs identified using method IIb allowing for multiple pairing with the same hydration site (SH2 system). . 79

4.1 Overlap coefficients (OC) for occupancy and energy grids of HIV-1 protease with a total of 100 ns simulation time. 92

4.2 Effects of truncation on the accuracy of grid energies and the speed of simulation. 94

5.1 Fragment map type and correspondence to co-solvent probe atoms 101

5.2 RESP fitted partial charges of PhX probes used in the co-solvent simulation. 103

5.3 Co-solvent System Setup. 104

5.4 Overlap coefficient (OC) calculated for all fragment FEMaps. 109

5.5 Three collections of FEMaps used in the MC sampling. Exclusion map are always included. 112

LIST OF FIGURES

Figure Page

1.1 Illustration of the roles binding site water molecules play in protein-ligand binding. Protein binding sites are mostly filled with water. Displacement of energetically unfavorable water from the hydrophobic region upon ligand binding can lead to large potency gains which is also the driving force of protein-ligand binding. The inter-facial water molecules can mediate protein-ligand interactions via hydrogen bond. Release of such water molecule may lead to energetic penalty depends on whether the loss of water mediated energy can be compensated. 3

1.2 Example of structure-based drug design by targeting water molecules in the binding site of SD and EGFR kinase. (A) N-(3,3-diphenylpropyl)-4-benz-1,2,3-triazine-amine (lig 1) and 3-cyano-N-(3,3-diphenylpropyl)-4-cinnolin-amine (lig 2) bound to scytalone dehydratase (SD) in the presence or absence of an interfacial water molecule. Displacement of the interfacial water molecule by the additional cyano group of lig 2 results in a 2.0 kcal/mol gain in binding free energy. (B) 4-anilino-6,7-dialkoxyquinazoline inhibitor (lig 3) and 4-anilino-6,7-dialkoxyquinoline-3-carbonitrile inhibitor (lig 4) bound to EGFR kinase. Similar displacement of the interfacial water molecule by the cyano group results in unfavorable binding free energy. This figure is adapted from [26]. 6

1.3 Overall procedure of hydration site prediction using WATsite. (A) The protein is solvated in an rectangular water box. (B) Water's oxygen atom is projected onto the 3D grid inside protein active site. (C) 3D grids are filtered with less than twice the bulk water occupancy. (D) A clustering algorithm is used to identify the locations of hydration sites. (E) (F) De-solvation free energy is estimated for each hydration site. 12

1.4 Example of bound water in HIV-1 protease.Residue Ile-50 and Ile-50' from two sub-units is shown in grey sticks, and inhibitor KNI in green sticks. Water 301 is shown as red sphere, and yellow dashed lines represent hydrogen bonds (PDB ID: 1hpx). 17

1.5 Hydration site result predicted with the presence of ligand. (A) Choose the "WATsite.out" file and select all options to load the results. (B) The result of example case of HIV-1 protease with bound ligand. The interfacial water molecule is selected. 19

Figure Page

1.6 Representation of predicted hydration sites and bound ligands in SD. (A) Difference between lig 1 and lig 2 experimentally. (B) Overlay of lig 1 and hydration sites in the active site. (C) Overlay of lig 2 and hydration sites in the active site. The free energy of the additional hydration site displaced by lig 2 is labeled. 21

1.7 Representation of predicted hydration sites and bound ligand in EGFR kinase. (A) Difference between lig 3 and lig 4 experimentally. (B) Overlay of lig 3 and hydration sites. (C) Overlay of lig 4 and hydration sites. The free energy of the additional hydration site displaced by lig 4 is labeled. The free energy of the additional hydration site displaced by lig 4 is labeled. 22

1.8 The protein de-solvation free energy estimated for ligands in SD and EGFR using the PyMOL plugin. (A) The de-solvation free energy difference (2.39 kcal/mol) between lig 1 and lig 2 results from one additional hydration site displaced by lig 2. (B) The de-solvation free energy difference (-0.65 kcal/mol) between lig 3 and lig 4 also results from one additional hydration site displaced by lig 4. 23

2.1 Correlation of energy values of the paired hydration sites obtained from the 20 ns MD simulations and from shorter simulation lengths. A: 1 ns, B: 1.5 ns, C: 2 ns, D: 2.5 ns, E: 3 ns, F: 4 ns, G: 5 ns, H: 10 ns. The correlation coefficients (R^2) is calculated to the regression line with the slope = 1 and zero point = 0, i.e. y=x. (Left) De-solvation free energy ΔG (kcal/mol), (middle) enthalpy ΔH (kcal/mol), and (right) entropy $-T\Delta S$ (kcal/mol). The distances between paired hydration sites are color coded according to the color bar. 31

2.2 Superimposition of hydration sites in the binding site of pyridoxine 5'-phosphate oxidase (PDB: 2AQ6). A: 0 Å <RMSD< 0.5 Å; B: 0.5 Å <RMSD< 1 Å; C: 1 Å <RMSD< 1.5 Å; D: 1.5 Å <RMSD< 2 Å. For clarity, only hydration sites within 1 Åto any atoms of the ligand are shown. The hydration sites are colored differently for different initial protein conformation. 33

2.3 Superimposition of hydration sites in the binding site of pyridoxine 5'-phosphate oxidase (PDB: 1XXO) 34

2.4 Superimposition of hydration sites in the binding site of goose egg-white lysozyme (PDB: 153L). 35

2.5 Superimposition of hydration sites in the binding site of goose egg-white lysozyme (PDB: 154L). 36

2.6 The percentage of paired hydration sites out of all predicted hydration sites found in different RMSD groups of each protein system. 37

Figure Page

2.7 Pairwise comparison between two trials of simulations within the 0 Å $<$ RMSD $<$ 0.5 Å group of goose lysozyme (PDB: 154L). 38

2.8 The RMSE distribution of thermodynamic properties (ΔG, ΔH, and $-T\Delta S$) for four RMSD groups representing different similarity levels. A: GEWL system (apo, PDB: 153L); B: GEWL system (holo, PDB: 154L); C: PLP system (apo, PDB: 1XXO); D: PLP system (holo, PDB: 2AQ6). . 39

2.9 The variation of de-solvation free energy involved in replacing water molecules upon ligand binding for four RMSD groups using different distance cutoff values. A: Goose egg-white lysozyme system (GEWL) (PDB: 153L, 154L). B: pyridoxine 5'-phosphate oxidase system (PLP) (PDB: 1XXO, 2AQ6). . 41

3.1 Scheme for different scenarios of the influence of conformation change of the protein on explicit water molecules involved in protein-ligand binding. Free energy of each hydration site scales from blue (more favorable) to red (less favorable binding). (A) Positional change of hydration sites upon conformational change upon ligand binding: α leaves, β and ϵ enter the protein active site. Upon ligand binding, water β, although replaced by the ligand in the holo state, does not contribute to the de-solvation energy as it is in the bulk solvent in the apo state. In contrary, water α, although not directly replaced in the holo state by the ligand, contributes to the protein's de-solvation free energy as it is replaced by the conformational change induced by ligand binding. Contrary, water ϵ contributes to the de-solvation free energy. (B) Change of thermodynamic profiles of hydration sites during conformational change upon ligand binding: γ becomes less stable, and δ becomes more stable. The de-solvation energy of those waters in the apo state (not holo state) needs to be taken into consideration for the free energy of binding calculation. 45

3.2 Illustration of the 28 ns MTMD simulation of the protein conformational change from apo to holo structure. RMSD restraints is applied to both the binding site residues and the whole protein structure of GUA system. Using 0.5 ns intervals, 49 overlapping windows with a length of 4 ns were generated. 49

3.3 Overview of the two methods for the analysis of hydration site locations and thermodynamic profiles during protein conformational change upon ligand binding. 51

Figure Page

3.4 Illustration of the Hydration Site-Specific Coordinate System (HSSCS) method. (A) All combinations of three out of eight closest residues are used to define a HSSCS. (B) The hydration sites associated with THR-34 and THR-35 are 4.7 Å apart between apo (cyan) and holo (magenta) form when measured in the Cartesian coordinate system of aligned apo and holo protein structure. (C)(D) Three red spheres (a, b and c) defining a plane R are used to create a HSSCS coordinate system. The x-axis is defined by spheres a and b, the y-axis is on the same plane and perpendicular to the x-axis, and the z-axis is perpendicular to the plane R. The two hydration sites are essentially co-localized in terms of the HSSCS. 54

3.5 Representative protein conformations of each 4 ns MTMD window. (A) GUA system; (B) SH2 system; representing the conformational transition pathway simulated by MTMD simulation (purple: apo structure, red: holo structure). 61

3.6 Overlay of hydration sites identified in the GUA system. (A) Hydration sites predicted using the apo (Purple) and the holo (Red) protein conformation. (B) Hydration sites of all MTMD windows. (Purple: apo protein conformation, Red: holo protein conformation) (C) Surface representation of the apo protein conformation with the predicted hydration sites and the extracted ligand from the holo form. (D) Surface representation of the holo protein form with the predicted hydration sites and the bound ligand. The ligand binding pocket is observed in the holo form. 62

3.7 Overlay of hydration sites identified in the SH2 domain. (A) Hydration sites predicted using the apo (Purple) and the holo (Red) protein conformation. (B) Hydration sites of all MTMD windows. (Purple: apo protein conformation, Red: holo protein conformation) (C) Surface representation of the apo protein conformation with the predicted hydration sites and the extracted ligand from the holo form. (D) Surface representation of the holo protein form with the predicted hydration sites and the bound ligand. The ligand binding site is solvent exposed. 63

3.8 Examples of hydration site transition pathways for the GUA system. (A) type (i) hydration sites that remain in the binding site throughout the conformational transition; (B) type (ii) hydration sites that disappear throughout transition from apo to holo form; (C) type (iii) hydration sites that appear throughout the conformational change. The graphs on the right side show the thermodynamic properties (ΔG, $-T\Delta S$, and ΔH) (in kcal/mol) and occupancy of the hydration site throughout the transition (window 0 represents the apo protein structure, and window 48 the holo protein structure). 68

Figure Page

3.9 Example of additional hydration site pair identified using method II which disappeared at intermediate windows using method I. (A) Hydration site #28 in the apo form (purple), and #9 in the holo form (red) are interacting with SER-34 via hydrogen bonding. (B) Transition trajectories of hydration site #28 in the apo form and #9 in the holo form using method I highlights the missing hydration site throughout the transition. (C) Hydration sites for window 11 and 16 are shown together with the water density at window 14 at which no hydration site was identified. (D) Overlay of thermodynamic properties and occupancy of hydration site #28 and #9. 73

3.10 Example of different hydration-site pairs identified from method I and II. More hydration sites interact with GLU-69 and SP-100 in solvent-exposed apo form (A) than in the holo form(B). (C)-(F) Water density of each hydration site around GLU-69 or ASP-100. Less pronounced water density is observed in the protein apo form(C,E) than in the holo form(D,F). . . . 75

3.11 The correlation between buried apolar surface area and predicted protein de-solvation free energies (ΔG_{desolv}) versus the experimentally measured pIC$_{50}$ values. (A) The buried apolar surface area is estimated using PyMol as described in the method section. (B) ΔG_{desolv} using the rigid protein method is estimated with the hydration sites predicted using the protein holo form. (C) ΔG_{desolv} from the flexible protein method is estimated using all identified hydration-site pairs between apoHS and holoHS using the equations 3.3 and 3.4. 81

4.1 Comparison of hydration sites prediction from WATsite2.0 based on GROMACS and WATsite3.0 based on OpenMM. 89

4.2 Overlay of hydration sites and energy grids at two absolute energy cutoffs. 90

4.3 Convergence of hydration site energies for HIV-1 protease at 5 ns and 10 ns comparing to 100 ns. 91

4.4 Comparison of hydration sites predicted with 12 Å system truncation versus full system. 94

4.5 Simulation performance of WATsite with and without GPU-acceleration. . 95

4.6 Total time cost of MD simulation and post hydration site analysis. 96

Figure Page

5.1 Structures of the human cathepsin L (hCatL) inhibitors and their interaction scheme. Inhibitors of hCatL undergo reversible covalent binding to the thiol group of CYS-25 in the S1 pocket under formation of thioimidates. In addition, hydrogen bonds are also formed to the backbone NH group of GLY-68, and C=O group of ASP-162 (green dashed lines). The XB interaction to GLY-61 is highlighted in red. (adapted from Hardegger et al. [98]) 98

5.2 Crystal structures showing the interaction between the backbone carbonyl of GLY-61 in the S3 pocket of hCatL with F, Cl, Br, and I. PDB ID: F 2XU4, Cl 2YJC, Br 2YJ2, I 2YJ8. Color: F turquoise, Cl green, Br brown, I purple. Figure was generated using PyMOL [99]. 99

5.3 Electrostatic-potential surfaces of five PhX probes used in the co-solvent simulation. The electrostatic potential is mapped onto the isosurface at an electron density value of 0.0004 e au^{-3} 102

5.4 Artificial repulsive force between all hydrophobic probes. One example shown for propane and iodo-benzene. 106

5.5 Qualitative comparison of fragment maps. Fragment FEMaps from trajectories 1-5 shown in green, and 6-10 shown in red. FEMaps are shown at free energy values of -1.5 kcal/mol for ACC, DON, HPHOB,ALI, HPHOB,ARO, and -2.5 kcal/mol for the remaining atom types. 110

5.6 Individual halogen FEMaps at two free energy isolevels. Left: -2.6 kcal/mol; Right: -3.2 kcal/mol. Color: F turquoise, Cl green, Br brown, I purple. 115

5.7 Bivariate fit of experiment ΔG by the predicted LGFE for all ligands. LGFE predicted with (Left) explicit halogen probes with EP, (Middle) explicit halogen probes without EP, (Right) only general FEMaps. 116

5.8 Bivariate fit of experiment ΔG by the predicted LGFE for Methyl, Cl, Br, I ligands. LGFE predicted with (Left) explicit halogen probes with EP, (Middle) explicit halogen probes without EP, (Right) only general FEMaps. 116

5.9 Bivariate fit of experiment ΔG by the predicted LGFE for only Cl, Br, I ligands. LGFE predicted with (Left) explicit halogen probes with EP, (Middle) explicit halogen probes without EP, (Right) only general FEMaps. 117

ABBREVIATIONS

Å	Ångström
AMBER	Assisted Model Building with Energy Refinement
apoHS	Hydration sites predicted in the protein apo form
CADD	Computer-aided drug design
DBSCAN	Density-based spatial clustering of applications with noise
ESP	Electrostatic potential
fs	Femtosecond (10^{-15} seconds)
GPU	Graphical Processing Unit
GUA	Yeast guanylate kinase
holoHS	Hydration sites predicted in the protein holo form
HS	Hydration Site
HSSCS	Hydration-site specific coordinate systems
K	Kelvin temperature
kcal/mol	Kilocalories per mole
LJ	Lennard-Jones
MD	Molecular Dynamics
MM/GBSA	Molecular-mechanics generalized Born surface area
MM/PBSA	Molecular-mechanics Poisson Boltzmann surface area
MTMD	Multiply-targeted molecular dynamics
NMR	Nuclear magnetic resonance
ns	Nanoseconds (10^{-9} seconds)
PDB	Protein data bank
PDF	Probability distribution function
QT	Quality threshold clustering

RESP	Restrained electrostatic potential
RMSD	Root-mean-square deviation
RMSD_{BS}	Root mean square deviation in the binding site
SH2	SH2 domain of (Pp60) Src kinase system
SAR	Structure-activity relationship
SASA	Solvent accessible surface area
$\Delta G_{binding}$	Free-energy of binding
ΔG_{desolv}	Protein de-solvation free energy

GLOSSARY

Apo protein structure: The protein structure in its ligand-free form.

Conformation: The spatial arrangement of the atoms affording distinction between stereoisomers which can be interconverted by torsional rotations about formally single bonds.

Holo protein structure: The protein structure in its ligand-bound form.

Hydration Site: The highest probable position of a water molecule in the coordinate system of the protein.

ABSTRACT

Yang, Ying. Ph.D., Purdue University, May 2018. Computational Modeling of (De)-Solvation Effects and Protein Flexibility in Protein-Ligand Binding using Molecular Dynamics Simulations. Major Professor: Markus A. Lill.

Water is a crucial participant in virtually all cellular functions. Evidently, water molecules in the binding site contribute significantly to the strength of intermolecular interactions in the aqueous phase by mediating protein-ligand interactions, solvating and de-solvating both ligand and protein upon protein-ligand dissociation and association. Recently many published studies use water distributions in the binding site to retrospectively explain and rationalize unexpected trends in structure-activity relationships (SAR). However, traditional approaches cannot quantitatively predict the thermodynamic properties of water molecules in the binding sites and its associated contribution to the binding free energy of a ligand.

We have developed and validated a computational method named WATsite to exploit high-resolution solvation maps and thermodynamic profiles to elucidate the water molecules' potential contribution to protein-ligand and protein-protein binding. We have also demonstrated the utility of the computational method WATsite to help direct medicinal chemistry efforts by using explicit water de-solvation.

In addition, protein conformational change is typically involved in the ligand-binding process which may completely change the position and thermodynamic properties of the water molecules in the binding site before or upon ligand binding. We have shown the interplay between protein flexibility and solvent reorganization, and we provide a quantitative estimation of the influence of protein flexibility on de-solvation free energy and, therefore, protein-ligand binding.

Different ligands binding to the same target protein can induce different conformational adaptations. In order to apply WATsite to an ensemble of different protein conformations, a more efficient implementation of WATsite based on GPU-acceleration and system truncation has been developed. Lastly, by extending the simulation protocol from pure water to mixed water-organic probes simulations, accurate modeling of halogen atom-protein interactions has been achieved.

PUBLICATIONS

The text of Chapter 2 includes content from:

Yang, Y.; Hu, B.; Lill, M. A., WATsite2.0 with PyMOL Plugin: Hydration Site Prediction and Visualization. Methods Mol. Biol. 2017, 1611, 123-134.

Yang, Y.; Abdallah, A. H. A.; Lill, M. A., Calculation of Thermodynamic Properties of Bound Water Molecules. Methods Mol. Biol. 2018, 1762, 389-402.

The text of Chapter 3 is a reprint of the material from:

Yang, Y.; Hu, B.; Lill, M. A., Analysis of factors influencing hydration site prediction based on molecular dynamics simulations. J. Chem. Inf. Model. 2014, 54 (10), 2987-95.

The text of Chapter 4 is a reprint of the material from:

Yang, Y.; Lill, M. A., Dissecting the Influence of Protein Flexibility on the Location and Thermodynamic Profile of Explicit Water Molecules in Protein-Ligand Binding. J. Chem. Theory Comput. 2016, 12 (9), 4578-92.

Portions of Chapter 5 are a preprint of the material prepared for publication as:

Yang, Y.; Masters, M.; Abdallah, A. H.; Lill, M. A., GPU-accelerated Hydration Site Analysis Tool WATsite and Application to Protein-Ligand Binding Affinity Prediction. 2018.

The text of Chapter 6 is a preprint of the material prepared for publication as:

Yang, Y.; Abdallah, A. H.; Lill, M. A., Modeling of halogen-protein interactions in co-solvent molecular dynamics simulations. 2018.

1. INTRODUCTION

1.1 The Drug Discovery Process

Drug discovery and development is in general a time-consuming and expensive process, which takes 10-15 years and approximately $1.7 billion dollars to bring a drug to market [1]. The pre-clinical stages include 1) target identification and validation; 2) high throughput screening, hit identification; 3) lead optimization and selection of a candidate molecule for clinical development [2].

1.1.1 Computer-aided Drug Design

In an attempt to reduce the time and cost associated with drug discovery, computer-aided drug design (CADD) techniques have become an integral part of this process. CADD is routinely used in the initial discovery phase focusing on reducing the number of ligands to be experimentally tested (hit to lead generation phase) and toward the end of the discovery phase emphasizing on optimizing the affinity and specificity of a selected number of compounds [3]. CADD methods can be classified into two major categories: structure-based drug design (SBDD) and ligand-based drug design (LBDD) [4,5].

In SBDD, the 3D structure information of the macromolecule, usually from X-ray crystallography or NMR, is used to rationally optimize the ligand's interaction with the target protein. In contrast, LBDD typically derives statistical models learning from a collection of molecules with known potency to rationally select and optimize chemical features for the desired biological outcome. Despite advances in method development, the consistent and accurate prediction of free energies of binding for protein-ligand complexes remains one of the major issues in CADD.

Two important challenges for the accurate prediction of binding free energies are the correct modeling of (de)-solvation effects and protein flexibility.

1.1.2 A Water World: Accounting for Water Molecules in SBDD

Water, acting as solvent, reactant, catalyst, lubricant, is a crucial participant,in virtually all ligand-binding processes in biology. For example, water can act as well as the origin of the hydrophobic effect [6]. Evidently, water molecules in a biomolecular system usually have a significant, if not dominant, contribution to the strength of intermolecular interactions in the aqueous phase by mediating protein–ligand interactions, solvating and de-solvating both ligand and protein upon protein-ligand dissociation and association [7–12]. For example, the de-solvation of hydrophobic ligand and/or protein moieties during protein-ligand association is typically associated with a gain in water entropy and/or enthalpy, frequently the major driving force for protein-ligand binding [13].

As shown in Figure 1.1, water molecules can be displaced upon ligand binding and can also mediate the interaction between protein and ligand via hydrogen bonds. Nevertheless, water molecules are often under-appreciated and even ignored in ligand docking studies. One reason for this neglect is that our understanding of the effect of water thermodynamics on ligand-protein binding free energies is still limited.

1.1.3 Everything is in Motion: Accounting for Protein Flexibility in SBDD

Over the years, several theoretical frameworks were developed to describe protein-ligand binding, including Fischer's "lock and key" model in 1894 [14]. Since then, it has been recognized that proteins are dynamic molecules, and their dynamic behavior plays a vital role for their functions including the ligand recognition process [15]. The historical lock-and-key model is too limited to reflect the fact that protein flexibility associated with ligand binding can be observed for the majority of proteins. Koshland

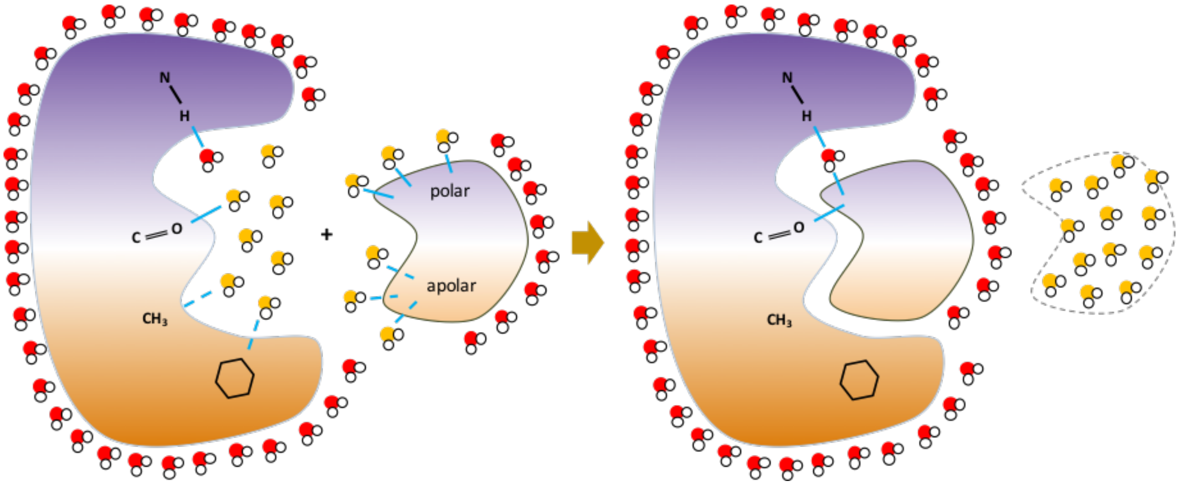


Fig. 1.1. Illustration of the roles binding site water molecules play in protein-ligand binding. Protein binding sites are mostly filled with water. Displacement of energetically unfavorable water from the hydrophobic region upon ligand binding can lead to large potency gains which is also the driving force of protein-ligand binding. The inter-facial water molecules can mediate protein-ligand interactions via hydrogen bond. Release of such water molecule may lead to energetic penalty depends on whether the loss of water mediated energy can be compensated.

proposed the “induced-fit” model in 1958 taking, for the first time, receptor flexibility into account [16]. Later in the 1990s the ”conformational selection” (or “population shift”) model was developed supported by numerous experiments [17]. Recent studies indicate the co-existence of induced-fit and conformational-selection models in most protein-ligand systems [18].

As conformational changes in proteins can involve many protein residues or whole protein domains, computational modeling of protein flexibility associated with ligand binding is still a challenging task in SBDD. Molecular dynamics (MD) simulations and variants thereof are the most frequently used approaches in SBDD nowadays to model protein flexibility.

1.1.4 Molecular Dynamics (MD) Simulations

MD simulations numerically solve the classical Newton’s equations of motion for the atoms of the protein-ligand-solvent system [5]. The system’s potential energy and forces at a given set of atom coordinates can be obtained by pre-parametrized classical force fields. This approach assumes a molecular mechanics description of the system, including force field terms for bonded and non-bonded interactions. The atomic forces are converted into changes in atom velocities, which are used to change the atom coordinates throughout the next discrete time step. Microscopic dynamics of a system under thermal equilibration can so be computed using MD simulations. From the simulations, the time-average over a trajectory equates an ensemble average according to the ergodic hypothesis [19]. A more detailed description of MD and the algorithms associated can be found elsewhere in the literature [20–23].

Since the first MD simulation on protein BPTI [24], MD techniques have become popular in SBDD due to the ability of tracking molecular movements with atomic precision. Protein and ligand flexibility as well as solvent effects are simultaneously considered in MD simulations. As a direct consequence of graphical processing unit (GPU) acceleration and new algorithmic developments, unprecedented speed of MD simulations are nowadays achieved [25]. Given the improvement in the accuracy of molecular force fields, modeling of protein flexibility, even slow motions at long timescales, can be achieved.

1.2 Individual Water Molecules in Drug Discovery

Protein-ligand complexes with high resolution structure from X-ray crystallography have shed light on the impact of individual water molecule to their binding free energy. For example, a single water was displaced by a cyano group for two systems: scytalone dehydratase (SD) and EGFR kinase (EGFR) (Figure 1.2) [26–28]. A water molecule is mediating the interaction between lig 1 and the Tyr-30 and Tyr-50 residues of SD via hydrogen bonding, while the conversion of the benzotriazine substructure in lig 1 to the 3-cyanocinnoline in lig 2 leads to the displacement of the inter-facial water molecule and a 30-fold improvement in K_i value [27]. Similarly, an inter-facial water molecule between lig 3 and residue Thr-766 of EGFR kinase is replaced by modifying the quinazoline in lig 3 to 3-cyanoquinoline in lig 4. Despite the same strategy, a 3-fold decrease in activity was observed for the EGFR kinase system [28].

For a rational decision on whether or not to replace a water molecule by a ligand's functional group, it is important to know the position as well as the associated thermodynamic property.

Experiment

The localized positions of water molecules in the binding site, i.e. hydration sites, can be partially identified in X-ray crystal structures. A standard experimental approach to study the impact of displacing a water molecule is via ligand modification like the example shown in Figure 1.2. However, it is impossible to directly measure the energy contribution of individual water molecule via experiment, and such approach is not practical to apply in real drug discovery projects. Therefore, numer-

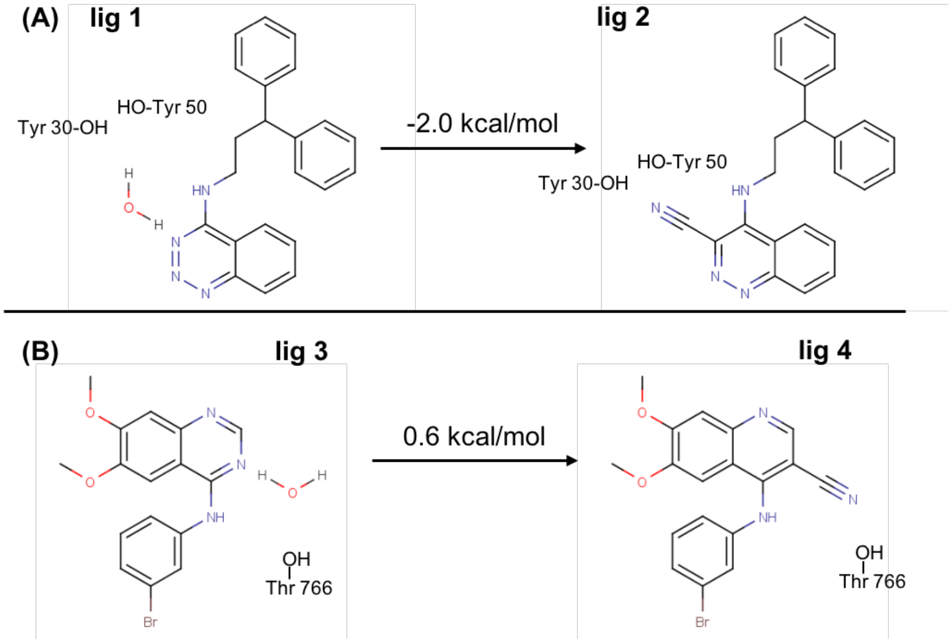


Fig. 1.2. Example of structure-based drug design by targeting water molecules in the binding site of SD and EGFR kinase. (A) N-(3,3-diphenylpropyl)-4-benz-1,2,3-triazine-amine (lig 1) and 3-cyano-N-(3,3-diphenylpropyl)-4-cinnolin-amine (lig 2) bound to scytalone dehydratase (SD) in the presence or absence of an interfacial water molecule. Displacement of the interfacial water molecule by the additional cyano group of lig 2 results in a 2.0 kcal/mol gain in binding free energy. (B) 4-anilino-6,7-dialkoxyquinazoline inhibitor (lig 3) and 4-anilino-6,7-dialkoxyquinoline-3-carbonitrile inhibitor (lig 4) bound to EGFR kinase. Similar displacement of the interfacial water molecule by the cyano group results in unfavorable binding free energy. This figure is adapted from [26].

ous computational approaches have been developed to predict the location and/or thermodynamics of water molecules [6].

Knowledge-based

Knowledge-based approaches, such as AQUARIUS [29] and SuperStar [30], predict likely hydration sites around polar or charged groups in proteins using experimentally derived algorithms on preferred geometries of water molecules around different amino acids from crystal structural data. AcquaAlta is another algorithm that specifies rules for favorable water geometries using an extensive search of the Cambridge Structural Database (CSD) and also uses ab initio calculations for the hydration propensities of functional groups. Those rules are then used to identify the location of water molecules bridging polar groups between the protein and the ligand [31]. Water potential mean forces (wPMF) [32] is another method based on 3946 non-redundant high resolution crystal structures where water pattern and residue hydrophilicities were extracted.

Energy-based

Energy-based methods, such as GRID [33] and CARTE [34], calculate the interaction energy between a water molecule and the protein to estimate the energetic favorability of water molecules in the binding site of a protein. WaterFLAP [35] takes the energetic minima from GRID OH2 molecular interaction field and predicts the water locations.

A more recent WaterDock [36] approach can be used to predict the locations of hydration sites and the likelihood each hydration site being displaced or conserved via repeated, independent docking of a water molecule into a protein cavity, followed by a filtering and clustering procedure.

SZMAP (Solvent Zap MAP) [37] developed by OpenEye is a hybrid explicit/implicit solvent approach. SZMAP takes one explicit water molecule, and the other water molecules are treated implicitly Poisson-Boltzmann solvent. Due to the design of only one explicit water, SZMAP has difficulty with predicting water networks and hydrogen bonding patterns in water clusters.

Statistical mechanics-based

3D reference interaction site model (3D-RISM) is an integral theory approach which produces an approximate average solvent distribution around a rigid solute using liquid state integral equations where the high dielectric polarization, the detailed interactions with a solute, and the multibody correlations of the solvent structure are taken into consideration [38].

Monte Carlo

Examples of Monte Carlo based methods include GCMC (Grand Canonical Monte Carlo) from Essex group [39] and JAWS (Just Add Water moleculeS) from Jorgensen group [40]. In GCMC, movement of "insertion" and "deletion" are possible, where the probability to accept or remove such move is controlled by the chemical potential. JAWS applies the double decoupling method to compare the removal of a water molecule from the bulk and from the binding site, so that to determine the ΔG_{bind} of that water molecule.

Molecular Dynamics

Molecular dynamic (MD)-based methods become popular for analyzing hydration sites. The protein is simulated with explicit water molecules and subsequent physics-based analysis is used to predict the location of water molecules in the binding site and the corresponding thermodynamic profile. Developing and using the inhomogeneous fluid solvation theory (IFST), Li and Lazaridis used MD simulations to calculate the thermodynamic properties of water molecules in the protein binding site including enthalpic and entropic contributions [41–43]. On the basis of IFST, WaterMap [44] was developed to identify hydration sites in binding pockets, and to evaluate the favorability of their displacement using an empirical formula based on the computed enthalpic and entropic contributions. Without a discrete hydration site definition, GIST(Grid Inhomogeneous Solvation Theory) [45] was developed to compute water density and thermodynamics on a 3D grid.

1.3 Hydration Site Prediction with WATsite

Along with other computational methods, MD-based hydration site analysis program WATsite was initially developed by a previous student in the group Bingjie Hu [46, 47]. Since then I have added additional features to the program such as another clustering algorithm to define hydration site, more options of force fields and water models, output energy grids instead of hydration sites, and GPU-acceleration.

In general, WATsite identifies hydration sites using a MD trajectory. The thermodynamic profile of each hydration site is then estimated by computing the enthalpy and entropy of the water molecule occupying a hydration site throughout the simulation. This section will detail the theory and method of this program.

1.3.1 Molecular Dynamics Simulation

System preparation

The ligand in each protein-ligand complex can be removed or kept in the binding site of protein, depending on the purpose of hydration site prediction. The crystallographic water molecules are usually kept. The protein will then be solvated in an rectangular water box with a minimum of 10 Å between any protein atom to the edge of the box (Figure 1.3 A). Chlorine and sodium ions were then added to neutralize the systems.

MD simulation protocol

MD simulations are performed using GROMACS [44] (WATsite2.0) or OpenMM [48] (WATsite3.0) with the AMBER force field (Amber99sb-ildn or Amber14SB) [49]. The SHAKE algorithm [50] was applied to constrain bonds including hydrogen atoms to their equilibrium lengths and maintain rigid water geometries. Long-range electrostatic interactions were treated with the Particle Mesh Ewald method [51] with a cutoff of 10 Å for the direct interactions. The Lennard-Jones interactions were truncated at a distance of 10 Å, and a long-range isotropic correction was applied to the pressure representing Lennard-Jones interactions beyond the cutoff.

Each system was first energy minimized for 5000 steps using the steepest descent algorithm. With all heavy atoms harmonically restrained (spring constants of 10 kJ mol^{-1} Å^{-2}), the system was then equilibrated for 1.25 ns with periodic boundary condition in all three dimensions. Finally, for hydration site identification and analysis, a production simulations will be performed with the same settings as the equilibration run.

In WATsite2.0 with GROMACS, temperature coupling was performed using the Nose-Hoover thermostat at 300 K, and the Parrinello-Rahman method was used for pressure coupling at 1 bar. Whereas for WATsite3.0 with OpenMM, a Langevin integrator with a time step of 2 fs was used together with a stochastic thermostat collision frequency of 1 ps^{-1}. The pressure control was implemented via isotropic box edge adjusting by MC moves every 25 time steps simulating the effect of constant pressure.

1.3.2 Hydration Site Identification

Hydration sites will be identified using all snapshots generated throughout the production run of each MD simulation.

First, the protein binding site was defined as a box surrounding its original ligand plus 3 Å in each dimension. A 3D grid was placed over the binding site using a grid spacing of 0.25 Å (Figure 1.3 B). In each snapshot, the positions of all the waters' oxygen atoms in the binding site were determined. A Gaussian distribution function centered on the oxygen atom centroid was distributed onto the 3D grid. To keep consistent with the definition of the 1 Å radius hydration sites described below, we used 0.33 Å as the standard deviation of this Gaussian function such that the Gaussian distribution covers 99.7% of the water occupancy within a 1 Å (three times the standard deviation) radius sphere. The distribution function was averaged over the MD trajectory and pronounced peaks (red grid in Figure 1.3 C) in this averaged function represent tightly binding water molecules which maintain their position throughout the MD simulation.

Next, a clustering algorithm is used to identify the locations of hydration sites (Figure 1.3 D).

When using the quality threshold **(QT) clustering algorithm**, for each grid point all other grid points that are within a 1 Å radius sphere are identified. The sphere that has the maximum occupancy (summation of the probabilities over all grid points in that sphere) was selected as first hydration site and all grid points contained in this sphere were removed from subsequent QT clustering steps. This clustering process was repeated until the occupancy in an identified hydration site becomes less than twice the expected occupancy of a 1 Å radius sphere in bulk solvent. The latter was determined by analyzing the pseudo-hydration sites in a MD simulation of bulk solvent. A pseudo-hydration site was defined as a randomly selected 1 Å radius sphere in the bulk solvent. The same Gaussian distribution functions were used to compute the occupancy probability of each grid point. The occupancy of a pseudo-hydration site was thus a simple summation of the probabilities on the grid points inside the defined sphere. Water molecules from the MD trajectory were assigned to each hydration site if its oxygen position is within the hydration site sphere. The 1 Å radius sphere, which has been used in previous hydration site studies ensures there is at most one water molecule in each hydration site per MD snapshot.

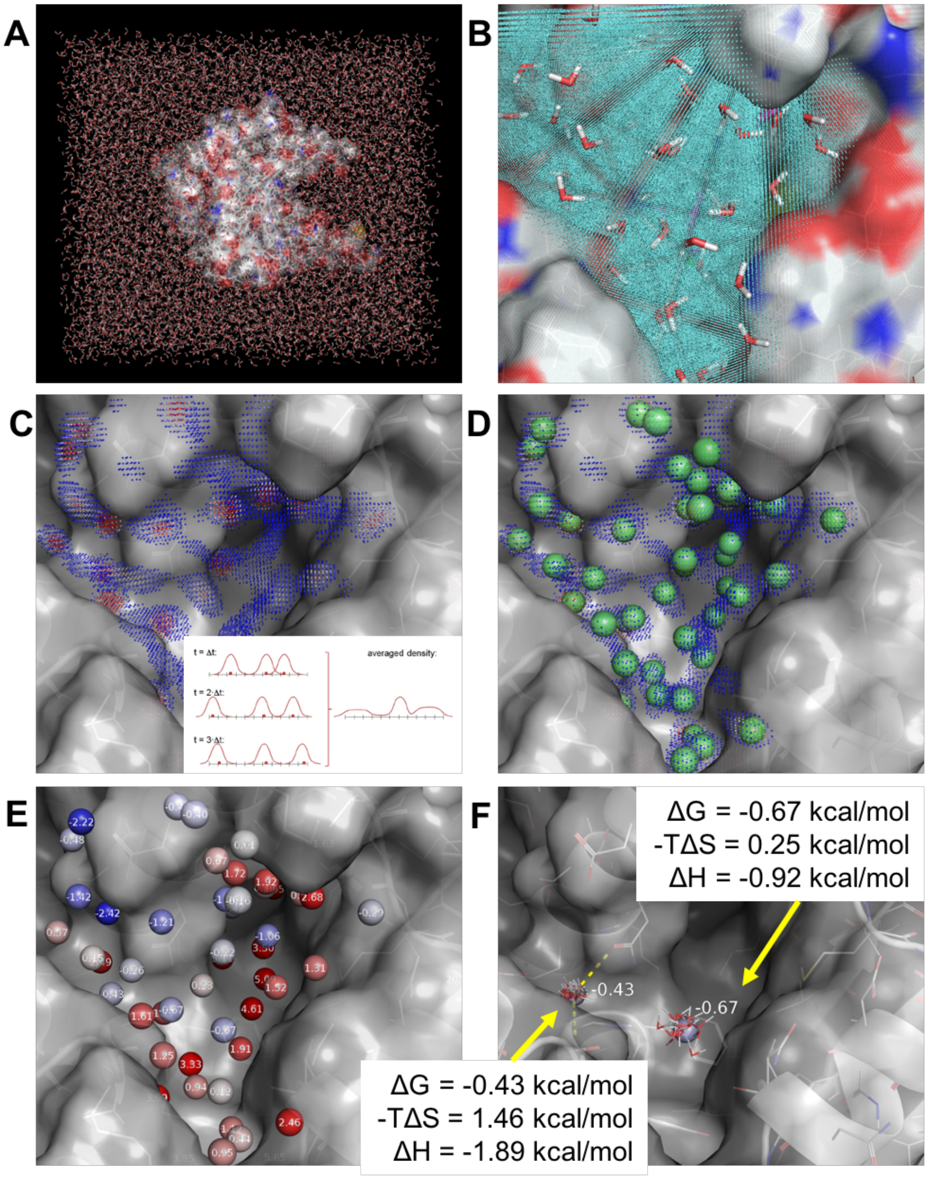


Fig. 1.3. Overall procedure of hydration site prediction using WATsite. (A) The protein is solvated in an rectangular water box. (B) Water's oxygen atom is projected onto the 3D grid inside protein active site. (C) 3D grids are filtered with less than twice the bulk water occupancy. (D) A clustering algorithm is used to identify the locations of hydration sites. (E) (F) De-solvation free energy is estimated for each hydration site.

When using the density-based spatial clustering of applications with noise **(DBSCAN) clustering algorithm**, grid points were filtered out if their occupancy is lower than twice the corresponding value in bulk solvent. A hydration site is defined if the cluster contains a minimum of 80 grid points that were not filtered out in the previous step. This cutoff value was defined based on the analysis of several X-ray structures (PDBID: 3T8G, 3T74, 3T87, 3T8H, 3T8C, 3T8D, 4H57, 4D9W) for which we were able to reproduce 90% of the water locations in the crystal structures using this criterion. Increasing the minimum number of grid points in a cluster as criterion, would result in too many crystallographic water molecules not being reproduced by the predicted hydration sites. On the other hand, if the minimum number of grid points in a cluster is reduced, we may generate hydration sites with relatively low occupancy increasing the noise in predicting favorable water locations.

1.3.3 De-solvation Free Energy Estimation for Hydration Sites

The de-solvation free energy of each hydration site (Figure 1.3 E & F) was determined by separately analyzing the enthalpy and entropy contributions of the water molecules inside a hydration site using:

$$\Delta G_{hs} = \Delta H_{hs} - T\Delta S_{hs} \tag{1.1}$$

ΔH_{hs} and ΔS_{hs} are the enthalpic and entropic change of transferring a water molecule from the bulk solvent into the hydration site of the protein binding site.

The change of the pressure-volume work associated with a volume change can be neglected. Thus the enthalpic change can be estimated by the change of the interaction energies:

$$\Delta H_{hs} \approx \Delta E_{hs} = E_{hs} - E_{bulk} \tag{1.2}$$

E_{hs} is the interaction energy of a water molecule in the hydration site with the surrounding protein and water atoms. It was determined based on the average sum of van der Waals and electrostatic interactions between each water molecule inside a given hydration site with the protein and all the other water molecules. E_{bulk} is the interaction energy of a water molecule with its surrounding environment in the bulk solvent. The average interaction energies of a water molecule with their surrounding environment estimated for five water models are listed in Table 1.1.

Assuming no change in the momenta part of the partition function upon transferring a water molecule from the bulk solvent into the protein cavity, ΔS_{hs} can be estimated by:

$$\Delta S_{hs} = S_{hs} - S_{bulk} \tag{1.3}$$

$$S = -R \iint\!\!\Bigg(\rho_{ext}(q) \ln \rho_{ext}(q) dq \tag{1.4}$$

R is the gas constant, and $\rho_{ext}(q)$ is the external mode probability density function (PDF) of the water molecules' translational and rotational motions during the molecular dynamics simulation. It should be noted that higher-order correlations between water molecules in the binding site are neglected in this approach.

To estimate $\rho_{ext}(q)$ for each hydration site, we analyzed the translational and rotational motions of the water molecules in that hydration site using a method adapted from McCammon and co-workers. For each hydration site, the translational degrees of freedom of water molecules in this site were defined by the fluctuation of the position of its center oxygen in the protein coordinate system. The Euler angles representing the spatial orientation of the water molecules in reference to the Cartesian coordinate system were used to calculate the rotational degrees of freedom. In detail, the rotated system (X, Y, Z) for quantifying the rotation of a water molecule was defined based on its $H_1 - O - H_2$ plane: the unit vector in the direction of $O - H_1$ defines X, the unit vector orthogonal to X in the $H_1 - O - H_2$ plane defines Y, and the unit vector orthogonal to the $H_1 - O - H_2$ plane defines Z. The Euler angles were then computed based on this rotated system. Two 3×3 zero-mean covariance matrices were constructed for the translational and rotational motions respectively assuming decoupled translational and rotational motions. One 6×6 zero-mean covariance matrix was also constructed assuming the translational and rotational motions are coupled. Principal components analysis was performed by diagonalizing the zero-mean covariance

Water Model	E_{bulk}
SPC/E	-21.15
TIP3P	-17.95
OPC	-23.14
TIP4P	-18.70
TIP4PEW	-20.95

Table 1.1.
E_{bulk} estimated for five water models. Unit: kcal/mol.

matrices and the original coordinates from snapshots were projected onto each of the principle component dimensions. A histogram was constructed for each principle component dimension with 70 bins to allow $\rho_{ext}(q)$ to be calculated by normalizing the histogram. The configurational entropy of each dimension was then numerically integrated using the composite Simpson's rule. The overall configurational entropy is then summed over all the principal component dimensions. No significant difference of the estimated configurational entropy was observed between using the two 3×3 matrices and the one 6×6 matrix by performing the paired t-tests at the significance level of 0.01 for five tested protein structures. Therefore, all results presented in this thesis utilize the two 3×3 matrices.

$T\Delta S_{bulk}$ is the entropy of a pseudo-hydration site in the bulk solvent. To estimate $T\Delta S_{bulk}$, a 100 ns simulation of a water box with 13734 expicit SPC water molecules was performed following the same procedures and parameters as the solvated protein simulation described above. Five pseudo-hydration sites are randomly chosen, and the average $-T\Delta S_{bulk}$ has been -3.8 kcal/mol (standard error: 0.035 kcal/mol).

1.4 Examples of Hydration Site Prediction using WATsite

There are two water displacement problems in SBDD: (1) displacing water molecules from the hydrophobic site of protein by inserting the ligand; and (2) displacing interfacial water molecules by growing substituent group on the ligand usually encountered in lead optimization.

In the context of first problem, binding site water displacement is a significant contribution, if not the driving force, of protein-ligand binding. A major application of WATsite is to use the predicted hydration sites to estimate the de-solvation free energies involved in replacing binding site water molecules by ligand binding.

For the second problem, a ligand is often modified to displace ordered water molecules in the binding site. Due to the inherent entropic contributions, releasing an ordered water molecule from the binding site into the bulk solvent is thought to be favorable for protein-ligand binding. However, in some cases the enthalpic gain from extra water-mediated hydrogen bonds exceeds the entropic loss for immobilizing the water involved. Thus, the thermodynamics of water molecules in protein active sites is important for understanding protein-ligand interactions for drug design.

Development of the cyclic inhibitors of human immunodeficiency virus (HIV-1) protease is a well-know example (Figure 1.4). A conserved water molecule (water 301) is located on the HIV-1 protease symmetry axis. This water molecule forms two hydrogen bonds to residue Ile-50 and Ile-50' on two sub-units and another two hydrogen bonds to the inhibitor. In another example, a similar strategy was used to displace a single water molecule by a cyano group for two different protein systems: scytalone dehydratase (SD) and EGFR kinase (Figure 1.2).

Using these examples, we will show in the following sections the ability of WATsite to predict the locations of water molecules observed in crystal structures and to estimate the protein de-solvation free energy of the bound ligands.

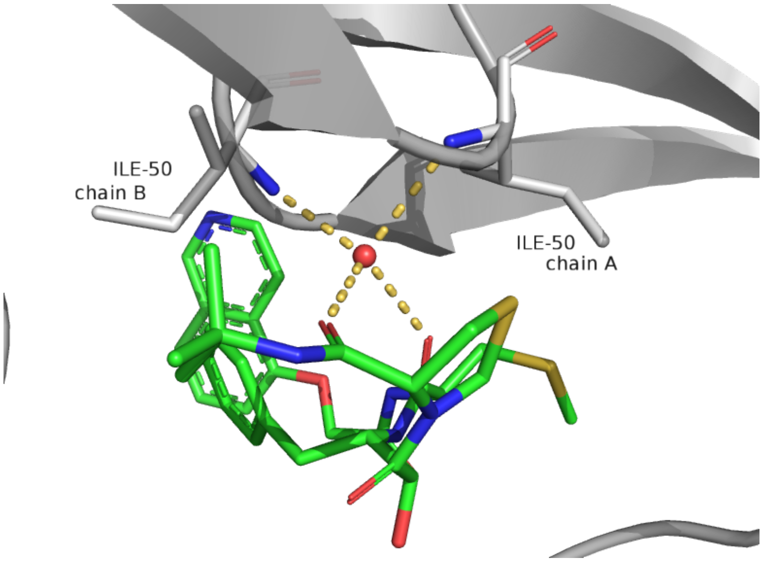


Fig. 1.4. Example of bound water in HIV-1 protease.Residue Ile-50 and Ile-50' from two sub-units is shown in grey sticks, and inhibitor KNI in green sticks. Water 301 is shown as red sphere, and yellow dashed lines represent hydrogen bonds (PDB ID: 1hpx).

1.4.1 Hydration Site Prediction with Ligand: Water at Binding Interface

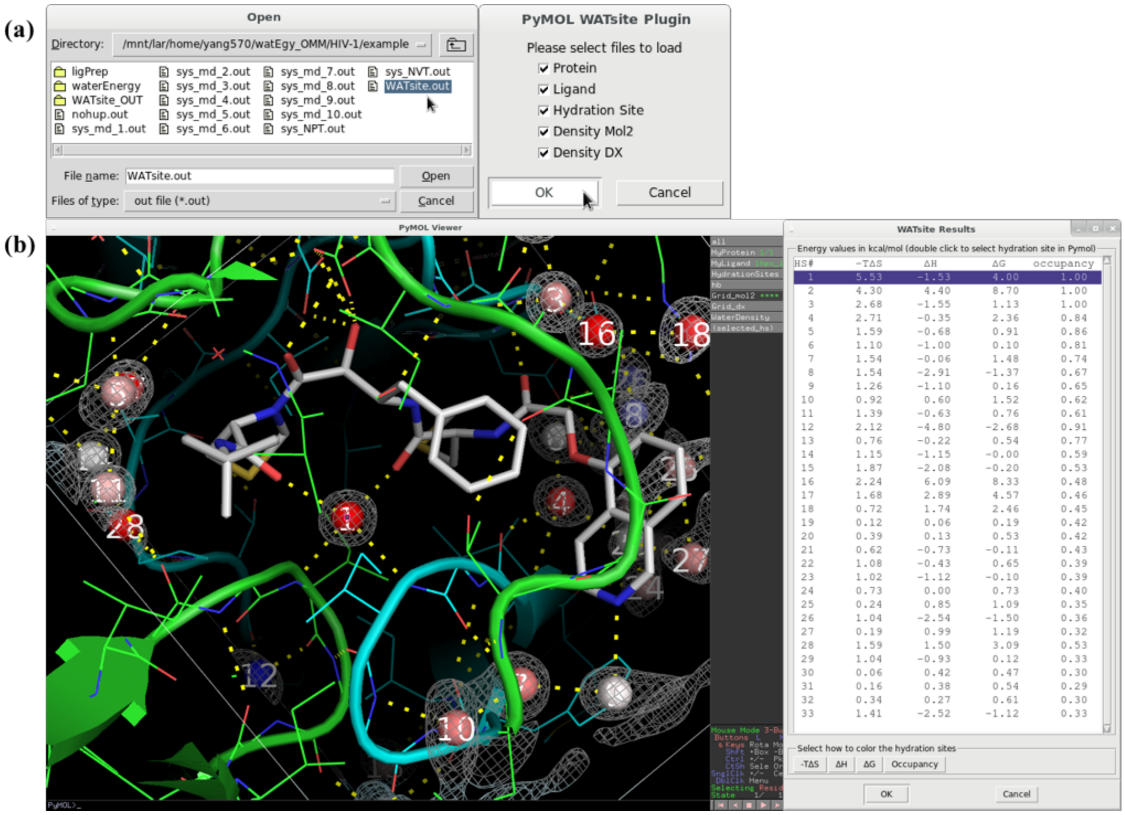
HIV-1 protease

When a lead compound is already known for a specific target, WATsite can be useful in suggesting ligand modification in order to improve affinity due to the displacement of hydration sites with unfavorable free energies. We performed a hydration site prediction with the presence of a bound inhibitor (KNI) for the HIV-1 protease (PDB: 1HPX). Here, we want to investigate water molecules at the binding interface between protein and ligand, so we select 'Protein', 'Ligand', and 'Hydration Site' to load into PyMOL. The result of the example case of HIV-1 protease are shown in Figure 1.5. The crystal waters are all predicted, and the inter-facial water mediating the protein-ligand interaction via hydrogen bonding is shown as the center red sphere in Figure 1.5B.

The hydration sites are shown as small spheres and colored in this example based on their ΔG values in a blue-white-red spectrum where blue indicates relatively low ΔG values and red indicates relatively high ΔG values. A hydration site with a more positive ΔG value (darker red) indicates an unfavorable environment of the water molecule in the binding site. Therefore, a gain in free energy of binding can be expected if the water in that hydration site is replaced by a ligand. The "occupancy" values indicate the probability a water molecule is observed in the given hydration site during the MD simulation.

1.4.2 Hydration Site Prediction without Ligand: Water Displacement upon Ligand Binding

We can also perform hydration site prediction with the ligand removed from the protein binding site. This can be useful to compare and evaluate the different protein de-solvation free energies from a congeneric series of ligands.



Directory: /mnt/lar/home/yang570/watEgy_OMM/HIV-1/example

File name: WATsite.out

Files of type: out file (*.out)

Please select files to load

☑ Protein
☑ Ligand
☑ Hydration Site
☑ Density Mol2
☑ Density DX

Energy values in kcal/mol (double click to select hydration site in Pymol)

HS#	-TΔS	ΔH	ΔG	occupancy
1	5.53	-1.53	4.00	1.00
2	4.30	4.40	8.70	1.00
3	2.68	-1.55	1.13	1.00
4	2.71	-0.35	2.36	0.84
5	1.59	-0.68	0.91	0.86
6	1.10	-1.00	0.10	0.81
7	1.54	-0.06	1.48	0.74
8	1.54	-2.91	-1.37	0.67
9	1.26	-1.10	0.16	0.65
10	0.92	0.60	1.52	0.62
11	1.39	-0.63	0.76	0.61
12	2.12	-4.80	-2.68	0.91
13	0.76	-0.22	0.54	0.77
14	1.15	-1.15	-0.00	0.59
15	1.87	-2.08	-0.20	0.53
16	2.24	6.09	8.33	0.48
17	1.68	2.89	4.57	0.46
18	0.72	1.74	2.46	0.45
19	0.12	0.06	0.19	0.42
20	0.39	0.13	0.53	0.42
21	0.62	-0.73	-0.11	0.43
22	1.08	-0.43	0.65	0.39
23	1.02	-1.12	-0.10	0.39
24	0.73	0.00	0.73	0.40
25	0.24	0.85	1.09	0.35
26	1.04	-2.54	-1.50	0.36
27	0.19	0.99	1.19	0.32
28	1.59	1.50	3.09	0.53
29	1.04	-0.93	0.12	0.33
30	0.06	0.42	0.47	0.30
31	0.16	0.38	0.54	0.29
32	0.34	0.27	0.61	0.30
33	1.41	-2.52	-1.12	0.33

Select how to color the hydration sites

Fig. 1.5. Hydration site result predicted with the presence of ligand. (A) Choose the ”WATsite.out” file and select all options to load the results. (B) The result of example case of HIV-1 protease with bound ligand. The inter-facial water molecule is selected.

Scytalone dehydratase and EGFR kinase

In order to estimate the difference of water displacement and de-solvation free energies between lig 1 & 2 and lig 3 & 4 as shown in the previous schematic Figure1.2. Hydration sites were predicted without the presence of bound ligand. The input pseudo-apo protein structure of SD was generated by removing the bound cyanocinnoline inhibitor from the complex (PDBID: 3STD). The crystal structure of EGFR kinase (PDBID: 1M17) was used to generate the input protein structure.

The hydration sites were predicted for the example systems SD (Figure 1.6 A) and EGFR (Figure 1.7 A). A more positive value means a more favorable contribution to the protein-ligand binding free energy.

The protein de-solvation free energies were estimated using the PyMOL plugin by adding up the free energies of hydration site overlapping with each ligand (Figure 1.8). The predicted protein de-solvation free energy of lig 2 is larger than lig 1, thus pointing to a favorable contribution to the binding free energy to SD for lig 2 compared to lig 1. Similarly, the protein de-solvation free energy of lig 4 is smaller than lig 3, and an unfavorable contribution to the binding free energy is expected. These results agree well with the relative binding free energies documented in the literature [26–28].

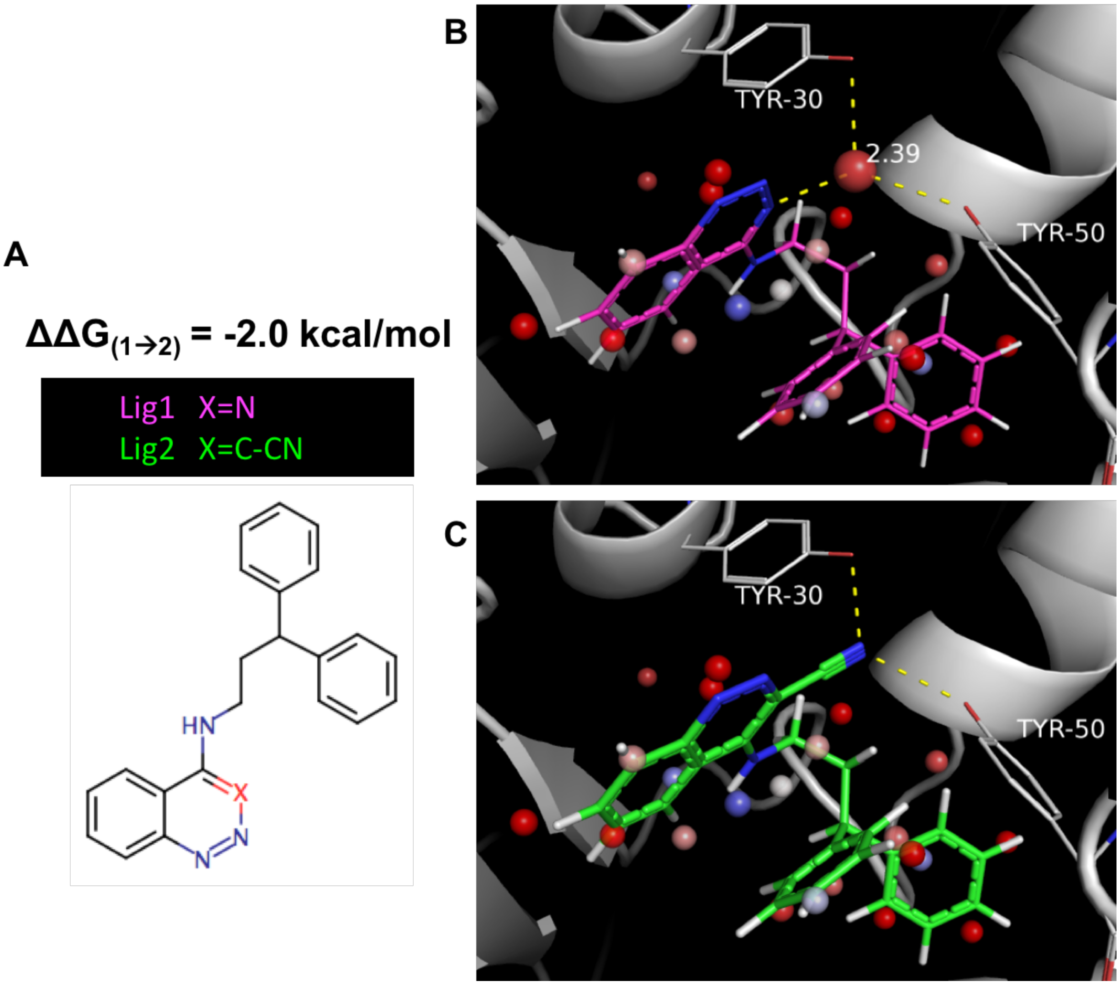


Fig. 1.6. Representation of predicted hydration sites and bound ligands in SD. (A) Difference between lig 1 and lig 2 experimentally. (B) Overlay of lig 1 and hydration sites in the active site. (C) Overlay of lig 2 and hydration sites in the active site. The free energy of the additional hydration site displaced by lig 2 is labeled.

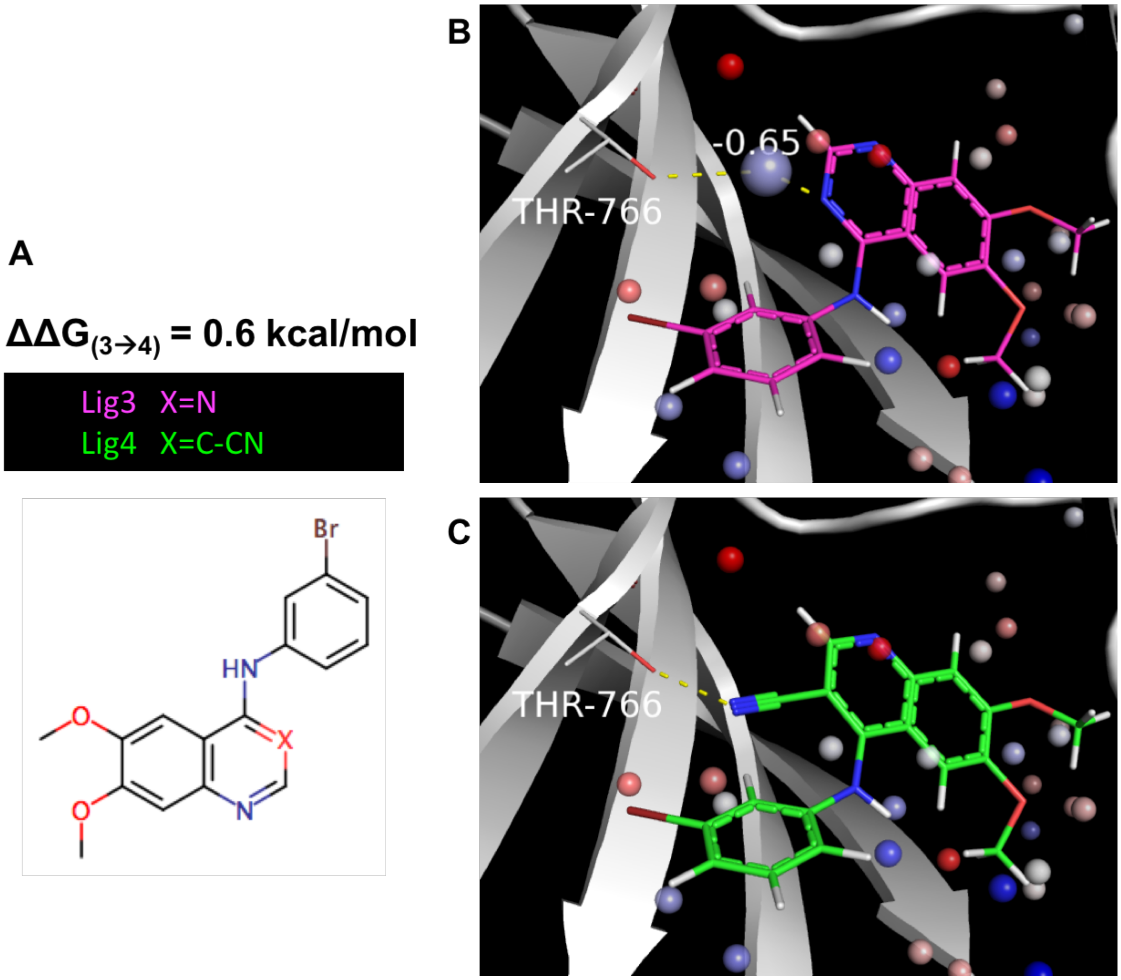
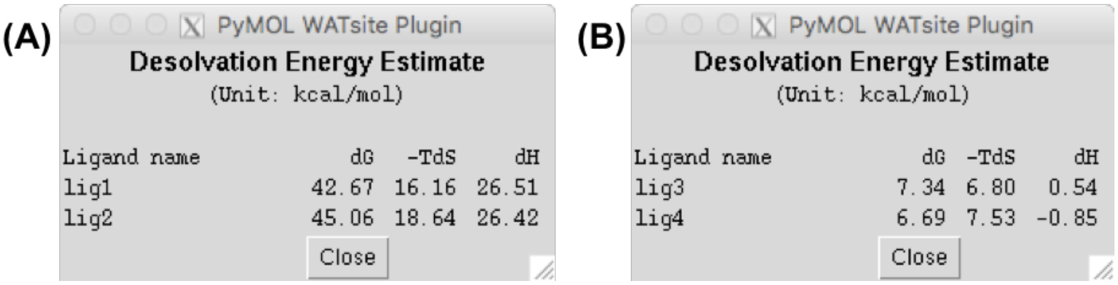


Fig. 1.7. Representation of predicted hydration sites and bound ligand in EGFR kinase. (A) Difference between lig 3 and lig 4 experimentally. (B) Overlay of lig 3 and hydration sites. (C) Overlay of lig 4 and hydration sites. The free energy of the additional hydration site displaced by lig 4 is labeled. The free energy of the additional hydration site displaced by lig 4 is labeled.



(A)

Desolvation Energy Estimate

```
(Unit: kcal/mol)

Ligand name               dG   -TdS     dH
lig1                   42.67  16.16  26.51
lig2                   45.06  18.64  26.42
```

(B)

Desolvation Energy Estimate

```
(Unit: kcal/mol)

Ligand name           dG  -TdS     dH
lig3                7.34  6.80   0.54
lig4                6.69  7.53  -0.85
```

Fig. 1.8. The protein de-solvation free energy estimated for ligands in SD and EGFR using the PyMOL plugin. (A) The de-solvation free energy difference (2.39 kcal/mol) between lig 1 and lig 2 results from one additional hydration site displaced by lig 2. (B) The de-solvation free energy difference (-0.65 kcal/mol) between lig 3 and lig 4 also results from one additional hydration site displaced by lig 4.

1.5 Research Summary

The overall goal of this thesis is to address how free energies of individual water molecules under consideration of protein flexibility can be incorporated into the prediction of thermodynamic profiles of protein-ligand binding. In this chapter, we summarized the most common computational methods for locating water molecules in the binding site of proteins. The methodological details of our in-house, MD-based hydration site prediction program WATsite were presented. We also discussed two different types of scenarios in SBDD in which hydration site prediction can be useful.

Multiple factors such as simulation length and initial protein conformations can affect hydration site prediction. Chapter 2 will present a detailed analysis of those factors. As previously mentioned, proteins are dynamic molecules, and their flexibility plays a vital role for their functions including the ligand recognition process. The interplay between protein flexibility and solvent reorganization will be discussed in Chapter 3. Chapter 4 will report two efforts to speed up hydration site analysis: GPU-acceleration and system truncation. Chapter 5 describes a slightly different direction by extending the simulation protocol from pure water to mixed water-organic probes simulations where accurate modeling of halogen atom-protein interactions has been achieved.

2. FACTORS INFLUENCING HYDRATION SITE PREDICTION BASED ON MD SIMULATIONS

2.1 Introduction

The hydration site analysis programs using MD simulations have become popular in the last few years [44, 46, 52], but many questions concerning the simulation protocol and its effect on hydration site identification and thermodynamic profiling remain unanswered. For example, the binding site may not be ideally hydrated at the beginning of the MD simulation and water molecules need to diffuse into or out of the binding site. This diffusion of water molecules into and out of binding cavities may be slow, especially with buried active sites. In addition, most water molecules typically are not well ordered in the binding site. Furthermore, it is well known that the convergence of entropy is often notoriously slow in MD simulations [53, 54]. Considering these issues the question arises for how long MD simulation should be performed to accurately predict hydration sites and their thermodynamic profile? Also, hydration sites may be predicted based on different X-ray structures or homology models representing different starting protein conformations. Thus, it is important to investigate how similar the predicted hydration sites and associated free energies are for different initial protein conformations.

In this chapter, we will approach these issues by 1) studying the influence of simulation lengths on hydration site analysis, and 2) determining the sensitivity of hydration site profiling and de-solvation free energy prediction on differences in starting protein conformations.

2.2 Materials and Methods

2.2.1 Protein Systems and Preparation

Four conformations from two protein systems have been chosen: Goose egg-white lysozyme (GEWL) (PDB code: 153L, 154L) [55], and mycobacterium tuberculosis pyridoxine 5'-phosphate oxidase (PLP) (PDB code: 1XXO, 2AQ6) [56]. For each system, the ligands from the holo structures were removed and the crystallographic water molecules were kept. The program Reduce [57] was used to adjust the side-chain conformations of ASN, GLN, and HIS, and tautomers and protonation states of HIS residues. The protein was then solvated in an octahedron of water molecules using the SPC water model [58] with a minimum distance of 10 Å between any protein atom and the faces of the octahedron. Chlorine and sodium ions were then added to neutralize the systems.

2.2.2 MD Simulation and Theory of Hydration Site Analysis

The detail has been described in Chapter 2. Here, 20 ns production simulation were performed to test convergence of hydration site locations, enthalpy and entropy calculations.

2.2.3 Comparison between Hydration Sites

To compare the relative locations of hydration sites between different simulations of the same protein, the last frame of each MD trajectory was aligned to the corresponding binding site in the X-ray structure using PyMOL. The last frame was arbitrarily chosen for the alignment process. As the protein is restrained during the MD simulation, the alignment process is fairly independent of the selection of a specific snapshot from the same MD trajectory. The predicted locations of the hydration sites were then shifted using the same transformation. The similarity of hydration site locations from two different simulation runs was determined by calculating all pairwise distances between hydration sites of two different simulations. The pair of hydration sites with smallest distance was identified and subsequently removed from further analysis. This process was continued until no additional hydration site pairs with a distance smaller than 1 Å could be identified. The 1 Å threshold for defining similar hydration site locations was chosen as the hydration sites are defined as spheres with radius of 1 Å [44, 46]. Each identified pair of hydration sites was considered to represent the same hydration site of a protein.

To compare the thermodynamic profiles of hydration sites between different simulations of the same protein, the free energy values of all pairs of the same hydration site were plotted against each other. The correlation coefficients (R^2) to the regression line with slope = 1 and zero point = 0, i.e. y = x were then calculated. Also, the root-mean squared error (RMSE) of energy values of all paired hydration sites was calculated.

2.2.4 Dependence of Hydration Site Analysis on Simulation Length

To study the influence of simulation length on calculated enthalpy and entropy values for each hydration site, different time points throughout the MD simulations were selected and the enthalpy and entropy values of each hydration site up to this time point were calculated. Analysis was performed for the first 1 ns, 1.5 ns, 2 ns, 2.5 ns, 3 ns, 4 ns, 5 ns, and 10 ns from the 20 ns simulation. For each simulation length, the enthalpy, entropy and free energy values were compared for each hydration site to the corresponding values of the 20 ns simulation, assuming that the energy values reached convergence after 20 ns simulation. The correlation between the energy values of two different simulation was quantified using Pearson correlation coefficients R2 for the linear regression line with slope = 1 and zero point = 0, e.g. $\Delta G_i^{20ns} = \Delta G_i^{1ns}$ for all hydration sites i. The specific regression line was chosen because we are studying the convergence properties of the absolute values of the hydration energies over simulation length.

2.2.5 Generation of Different Starting Conformations

In order to study the sensitivity of hydration site prediction on initial protein structure, 1 ns MD simulations without harmonic restrain were performed to sample different protein conformations.

The Root-Mean Square Deviation (RMSD) between binding site residues of each frame to every other frames from the entire trajectory was calculated. Conformation pairs were distributed into four different bins with RMSD values of 0-0.5 Å, 0.5-1 Å, 1-1.5 Å, and 1.5-2 Å respectively. From each bin, five conformations were selected to define four RMSD groups representing different levels of similarity. A group with higher RMSD values contains conformations with larger structural variations. Then, with heavy atoms harmonically restrained another 4 ns MD simulation was performed for all selected conformations, and those trajectories were used to predict the hydration sites for further analysis.

2.2.6 Estimation of De-solvation Free Energy of the Protein upon Ligand Binding

Using the predicted hydration sites and the PyMOL plugin of WATsite [47], the de-solvation free energy of the protein due to replacing water molecules in the protein binding site upon ligand binding was estimated. Different distance cutoffs (1 Å, 1.5 Å, 2 Å, and 2.5 Å) are specified to identify hydration sites within the specified distance to any of the ligands' heavy atoms. Larger distance cutoffs usually identify more hydration sites that are displaced by the ligand. The de-solvation free energy is then estimated by summing up the free energies of those identified hydration sites that are displaced upon ligand binding.

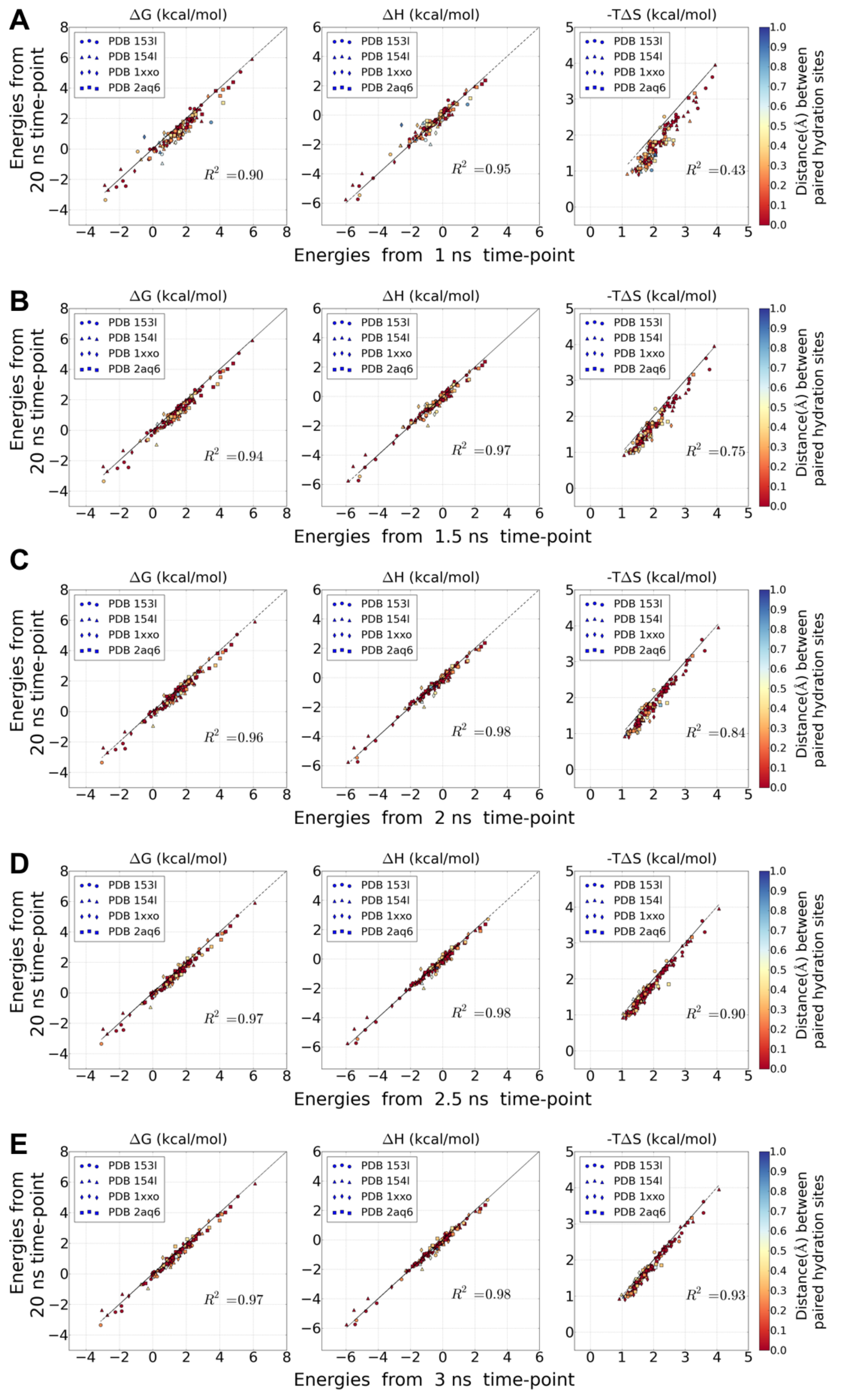
2.3 Results and Discussion

2.3.1 Dependence of Hydration Site Analysis on Simulation Length

To study the influence of simulation length on calculated enthalpy and entropy for each hydration site, different time points throughout the MD simulations were selected and the enthalpy and entropy values of each hydration site up to this time point were calculated.

The correlation between the energy values of different time points of simulations (1 ns, 1.5 ns, 2 ns, 2.5 ns, 3 ns, 4 ns, 5 ns, and 10 ns) and the energy values of the entire 20 ns simulation were calculated. As described in the Materials and Methods section, the paired hydration sites between two simulations were first determined, and estimated energy values of the same hydration site were pairwise compared. In order to study the convergence of the energy values, the Pearson correlation coefficients R^2 to the regression line with slope = 1 and zero point = 0, i.e. y=x were then calculated as shown in Figure 2.1. The geometric distances between paired hydration sites were color coded, ranging from red (identical position) to blue (1 Å distance). For protein PLP (PDB: 1XXO and 2AQ6), using 24 processors the required computation time for the three experiments (1 ns, 2.5 ns, and 4 ns) was about 12 h, 30 h, and 52 h respectively.

While high correlations for the enthalpy and free energy values of the 20 ns simulations was achieved already with using the 1 ns trajectories, a comparable correlation for the entropy values between these two time-points was rather low (Figure 2.1 A). With only one exception, the entropy values obtained throughout the 1 ns simulations are generally larger than those of the 20 ns simulations. This is most likely due to insufficient sampling at shorter simulation lengths overestimating the entropy loss upon binding into the binding site. With increasing simulation length , the correlation for the entropy values quickly improves, reaching a R^2 value of 0.9 at 2.5 ns (the R^2 values of enthalpy and free energy are 0.9 or larger for all comparisons) (Figure 2.1 D). The greater than 0.95 R^2 values of the 4 ns versus 20 ns comparison


A
ΔG (kcal/mol)
ΔH (kcal/mol)
-TΔS (kcal/mol)
PDB 153l
PDB 154l
PDB 1xxo
PDB 2aq6
R^2 =0.90
R^2 =0.95
R^2 =0.43
Energies from 20 ns time-point
Energies from 1 ns time-point
Distance(Å) between paired hydration sites
B
R^2 =0.94
R^2 =0.97
R^2 =0.75
Energies from 1.5 ns time-point
C
R^2 =0.96
R^2 =0.98
R^2 =0.84
Energies from 2 ns time-point
D
R^2 =0.97
R^2 =0.98
R^2 =0.90
Energies from 2.5 ns time-point
E
R^2 =0.97
R^2 =0.98
R^2 =0.93
Energies from 3 ns time-point
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0

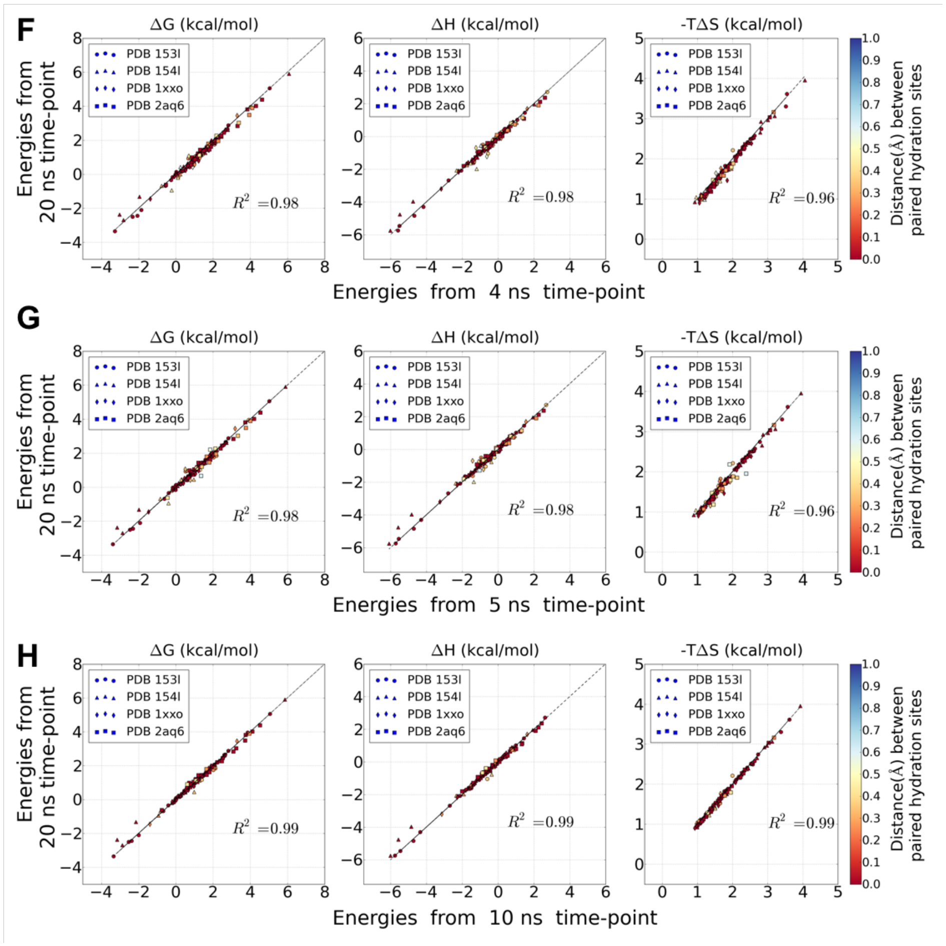


Fig. 2.1. Correlation of energy values of the paired hydration sites obtained from the 20 ns MD simulations and from shorter simulation lengths. A: 1 ns, B: 1.5 ns, C: 2 ns, D: 2.5 ns, E: 3 ns, F: 4 ns, G: 5 ns, H: 10 ns. The correlation coefficients (R^2) is calculated to the regression line with the slope = 1 and zero point = 0, i.e. y=x. (Left) De-solvation free energy ΔG (kcal/mol), (middle) enthalpy ΔH (kcal/mol), and (right) entropy $-T\Delta S$ (kcal/mol). The distances between paired hydration sites are color coded according to the color bar.

(Figure 2.1 F, 0.96 for entropy, 0.98 for enthalpy and 0.98 for free energy) indicate that 4 ns seems to be sufficient to generate converged thermodynamic profiles for all hydration sites compared to the 20 ns reference simulation. Therefore, we decided to use a simulation length of 4 ns for the rest of this study.

2.3.2 Sensitivity of Hydration Site Prediction on Initial Protein Structure

We also investigated if the starting conformations of a protein system for MD simulations have significant influence on the prediction of hydration sites. We hypothesized that the conformations of the binding site residues influence the prediction of the position and thermodynamic profile of hydration sites. Thus for each protein system, we constructed four RMSD groups of conformations representing different levels of binding site similarity as described in the Methods section. Then, within each RMSD group, the five sets of predicted hydration sites were aligned. A superimposition of those hydration sites for PLP (PDB: 2AQ6) is displayed in Figure 2.2. The hydration sites are colored for different initial protein conformation. The predicted locations of hydration sites using the least variant initial structures (RMSD 0-0.5 Å) are quite similar (Fig 2.2 A), while the positions of hydration sites overlap less with increasing RMSD (Fig 2.2 B,C,D. Corresponding results for the other three protein systems used in our study are displayed in Figure 2.3, 2.4, 2.5.

To quantitatively analyze how similar the hydration sites are predicted in each RMSD group, pairwise hydration site comparisons were carried out within each RMSD group, resulting in 10 pairwise comparisons per RMSD group. For each comparison, we identified paired hydration sites and calculated the percentage of paired hydration sites from all predicted hydration sites. This distribution of paired hydration sites for all four protein systems is displayed in the form of a box-plot graph for each RMSD group in Figure 2.6. As expected, more paired hydration sites were found in the group with smaller conformational variation than those with larger initial RMSD. On average more than 80% of all hydration sites have similar locations when

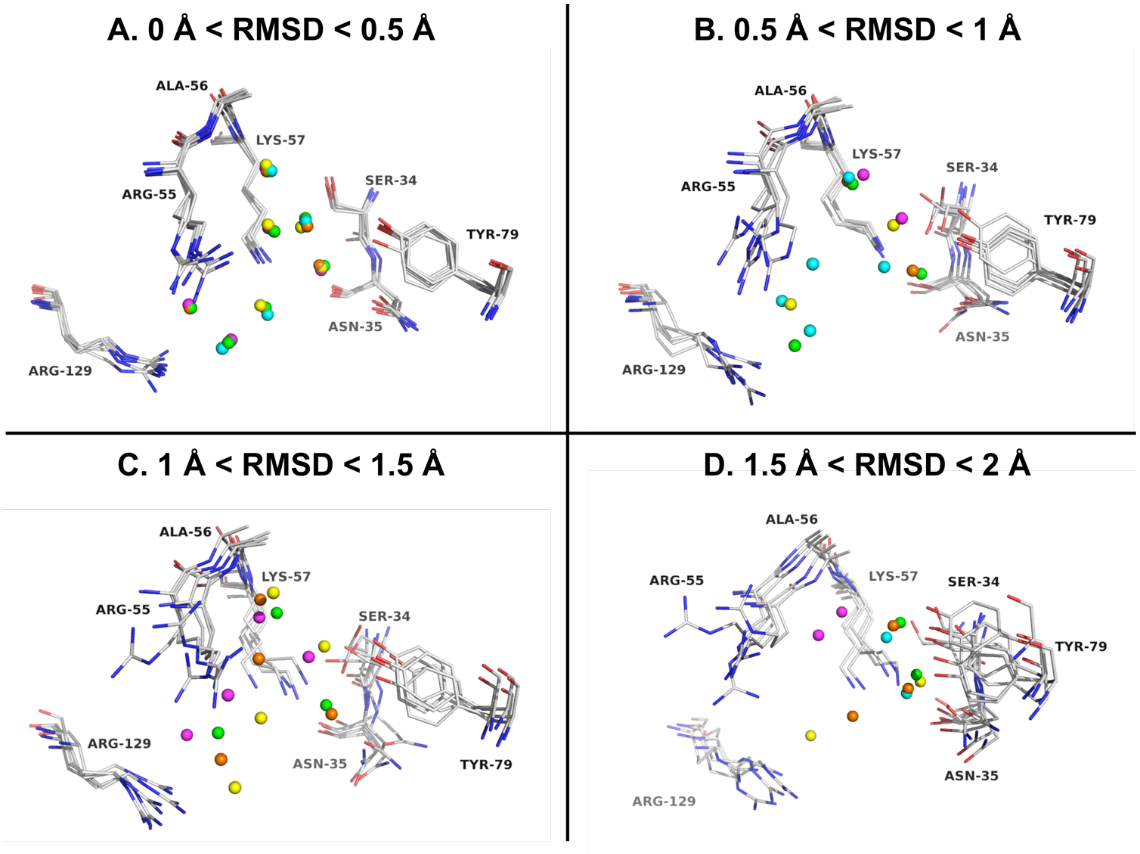


Fig. 2.2. Superimposition of hydration sites in the binding site of pyridoxine 5'-phosphate oxidase (PDB: 2AQ6). A: 0 Å <RMSD< 0.5 Å; B: 0.5 Å <RMSD< 1 Å; C: 1 Å <RMSD< 1.5 Å; D: 1.5 Å <RMSD< 2 Å. For clarity, only hydration sites within 1 Åto any atoms of the ligand are shown. The hydration sites are colored differently for different initial protein conformation.

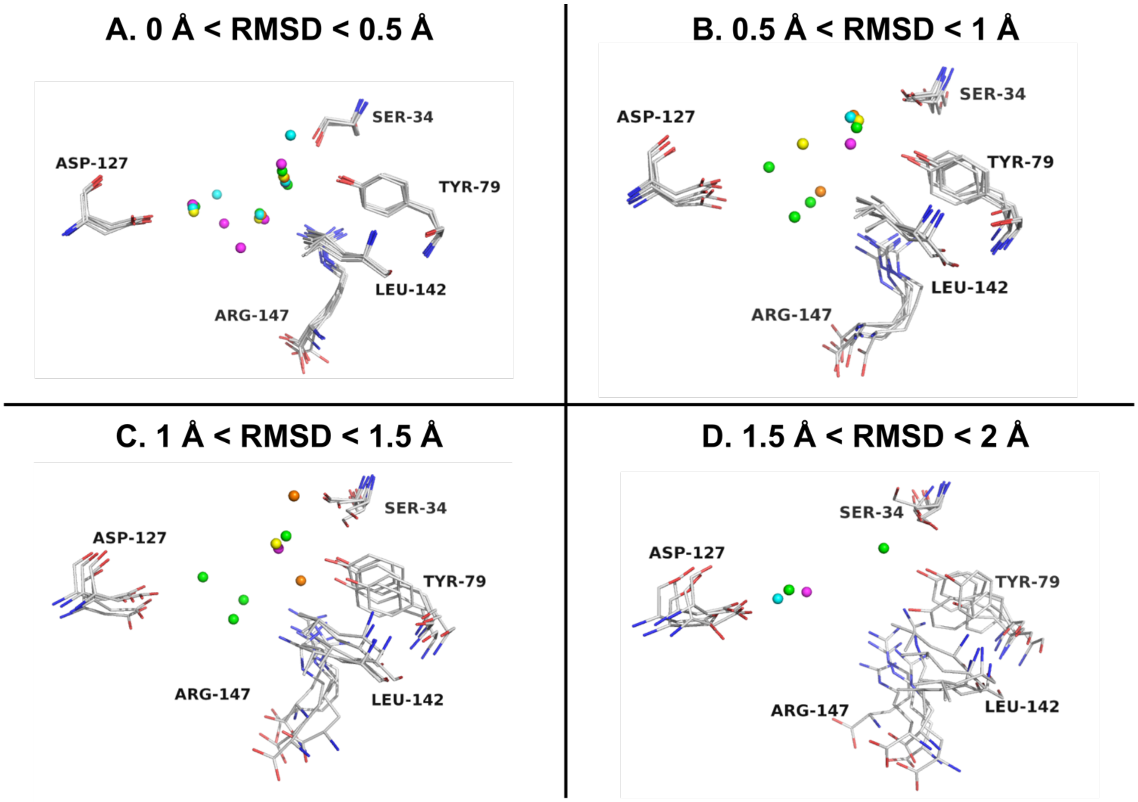


Fig. 2.3. Superimposition of hydration sites in the binding site of pyridoxine 5'-phosphate oxidase (PDB: 1XXO)

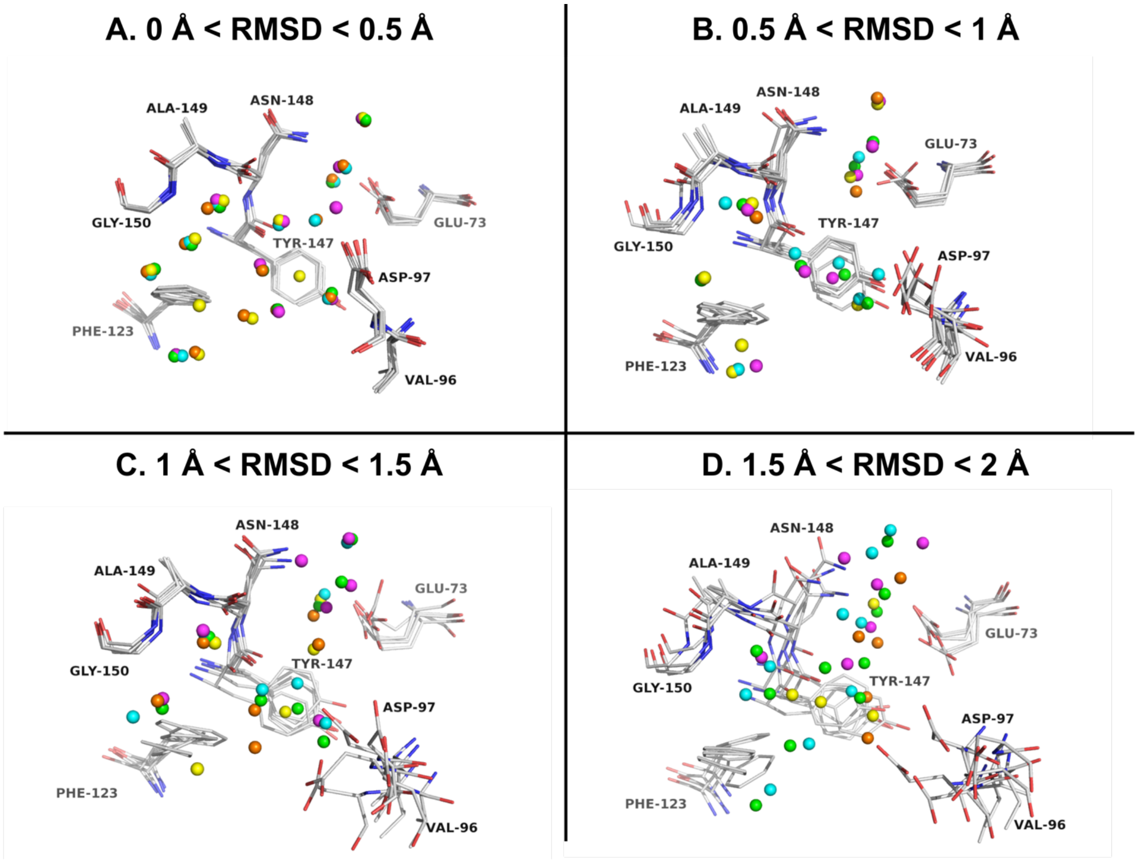


Fig. 2.4. Superimposition of hydration sites in the binding site of goose egg-white lysozyme (PDB: 153L).

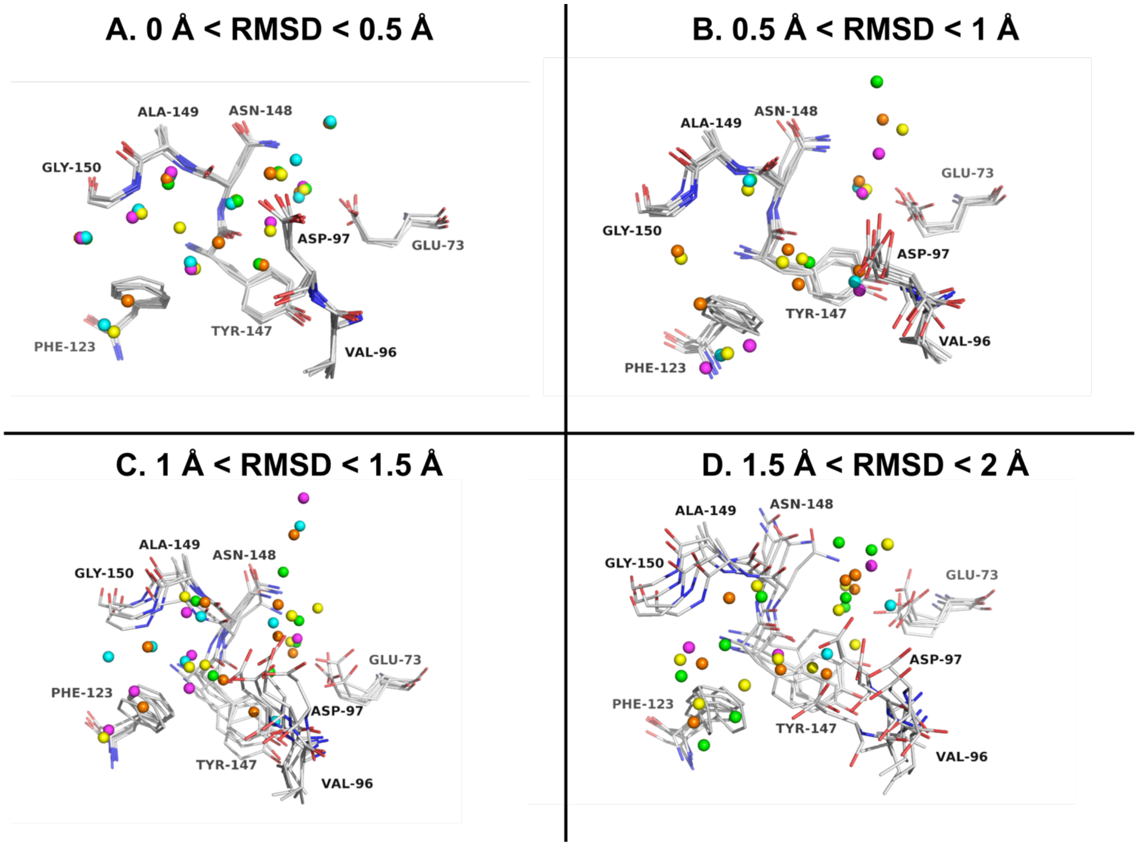


Fig. 2.5. Superimposition of hydration sites in the binding site of goose egg-white lysozyme (PDB: 154L).

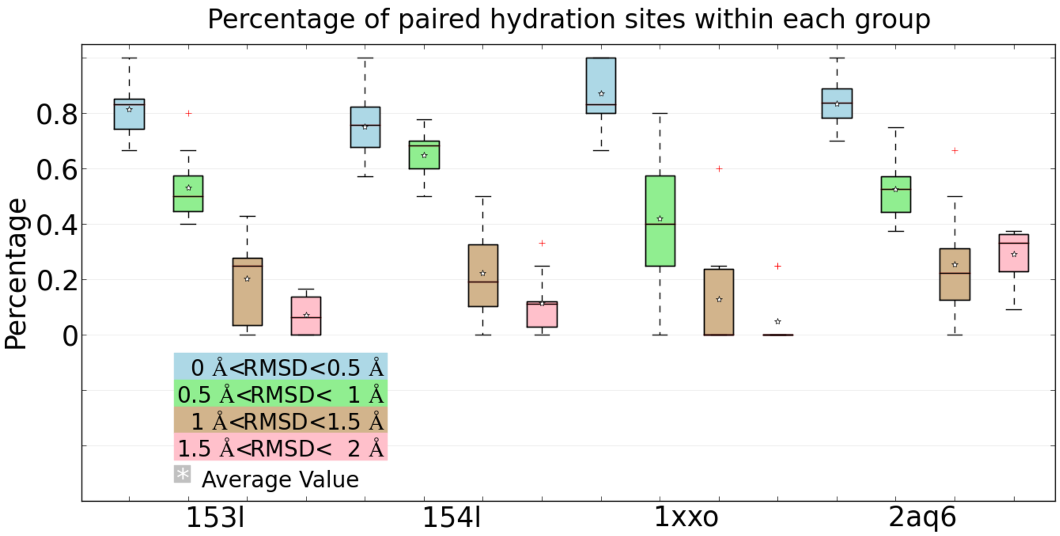


Fig. 2.6. The percentage of paired hydration sites out of all predicted hydration sites found in different RMSD groups of each protein system.

the starting protein structures are very similar (RMSD 0-0.5 Å), while only about a third of the hydration sites have similar locations if the starting structures deviate by 1-1.5 Å RMSD. This demonstrates the high sensitivity of WATsite and likely other MD-based hydration site programs on the starting protein structure.

We also analyzed how similar the estimated free energy values were for the paired hydration sites. After the pairs of hydration sites were identified, the distances between paired hydration sites were distributed into bins with a size of 0.1 Å. One example is shown in Figure 2.7. Most pairs of hydration sites have well conserved locations with a distance smaller than 0.5 Å (Figure 2.7A). Only a few hydration sites demonstrate a larger deviation. Also the correlation of the energy values of two different simulations was plotted in Figure 2.7 B for one randomly selected pair of comparisons for GEWL (PDB: 154L) from the group 0 Å $<$ RMSD $<$ 0.5 Å. As the high R^2 indicates, the de-solvation free energy of paired hydration sites estimated from similar initial protein conformations correlate well with each other.

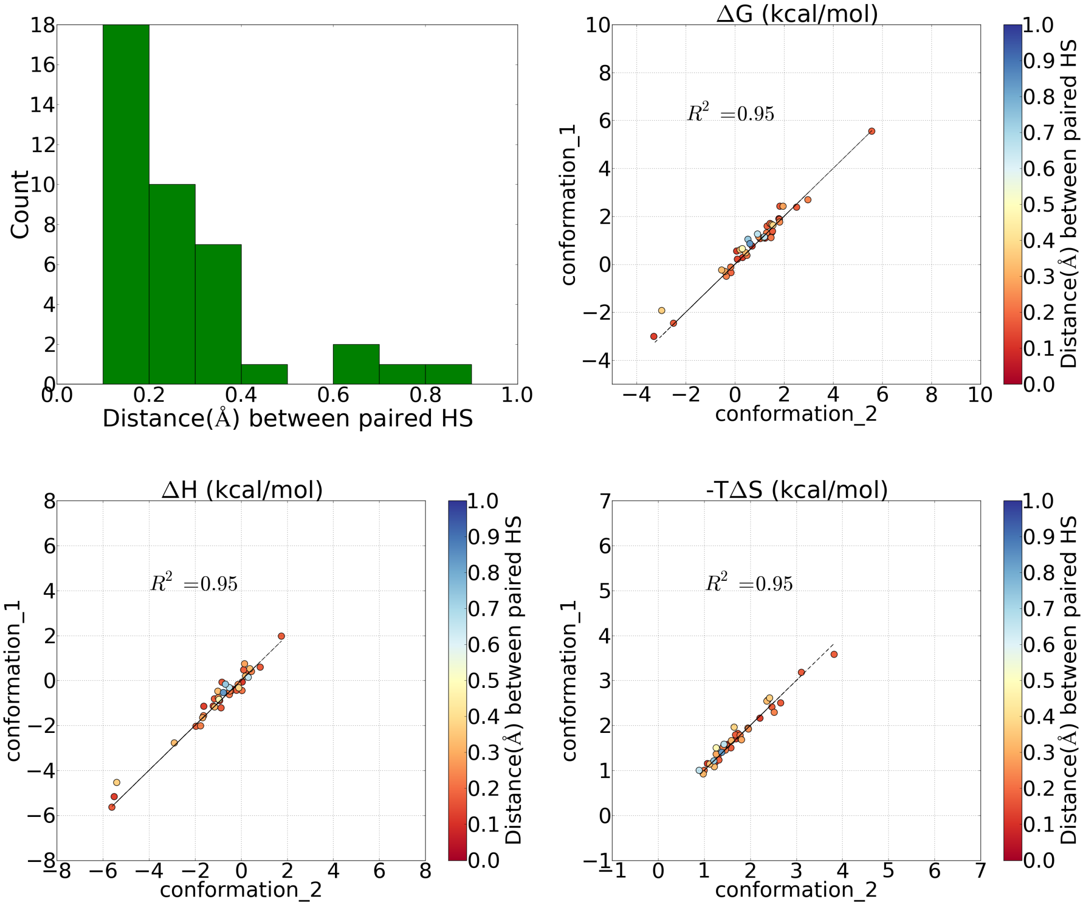


Fig. 2.7. Pairwise comparison between two trials of simulations within the 0 Å < RMSD < 0.5 Å group of goose lysozyme (PDB: 154L).

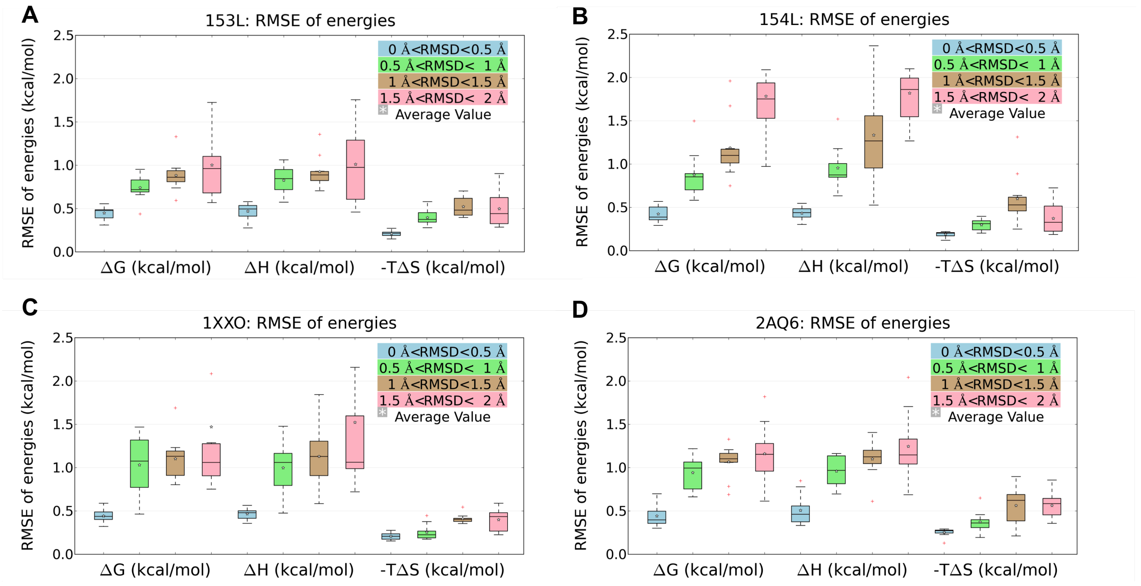


Fig. 2.8. The RMSE distribution of thermodynamic properties (ΔG, ΔH, and $-T\Delta S$) for four RMSD groups representing different similarity levels. A: GEWL system (apo, PDB: 153L); B: GEWL system (holo, PDB: 154L); C: PLP system (apo, PDB: 1XXO); D: PLP system (holo, PDB: 2AQ6).

To quantitatively analyze the similarity of all five sets of hydration sites in each RMSD group, the root mean square error (RMSE) for each thermodynamic property of interest (ΔG, $-T\Delta S$, and ΔH) was calculated for any two comparisons. The distribution of RMSE values of all pairwise comparisons for each protein system was obtained and is displayed in form of box-plots in Figure 2.8. Within the group with most similar starting conformations (0 Å $<$ RMSD $<$ 0.5 Å), all individual MD simulations generate consistent estimates of enthalpy, entropy and free energy independent of the starting structure. The RMSE for entropy is relatively small compared to the other two properties due to the small range of entropy values. In general, as the RMSD increases, the values of RMSE significantly increase due to the conformational variations of binding site residues, but the strength of dependency is system-dependent.

2.3.3 Sensitivity of Protein De-solvation Free Energy Estimation on Initial Protein Structure and Distance Cutoff between Ligand Atoms and Hydration Sites

In the last section, we studied the effect of different initial protein structures on the estimation of de-solvation free energy of the protein upon ligand binding. This quantity is computed as described in the Materials and Methods section. Different distance cutoffs between hydration site and the crystal ligands' heavy atoms were chosen to identify those hydration sites that are replaced upon ligand binding. Distance cutoffs of 1.0 Å, 1.5 Å, 2.0 Å, and 2.5 Å were chosen. The sum of the free energies of these hydration sites provides an estimate for the de-solvation free energy of the protein for each ligand. Thus, for each RMSD group we computed the de-solvation free energy for all five sets of predicted hydration sites. The maximum, minimum, and average values of the five de-solvation energies for each RMSD group are plotted in Figure 2.9. MD simulations of the group with the smallest conformational variation (RMSD < 0.5 Å) estimate the de-solvation energies consistently. Furthermore with increasing distance cutoff, more hydration sites are considered to be replaced upon ligand binding and therefore result in larger de-solvation energies. Finally and not surprisingly, larger variation in the predicted de-solvation free energies can be observed for the groups with more diverse initial protein structures compared to more similar initial protein conformations. Whereas the standard error for the group with RMSD < 0.5 Å is on average 0.84 kcal/mol, it is on average 2.10 kcal/mol for the group with RMSD between 1.5 and 2.0 Å.

2.4 Conclusion

In this chapter, we validated that the locations and thermodynamic properties of hydration sites can be reliably predicted using an MD simulation with a length of only 4 ns, which provides similar hydration site data compared to those of longer 20 ns simulations.

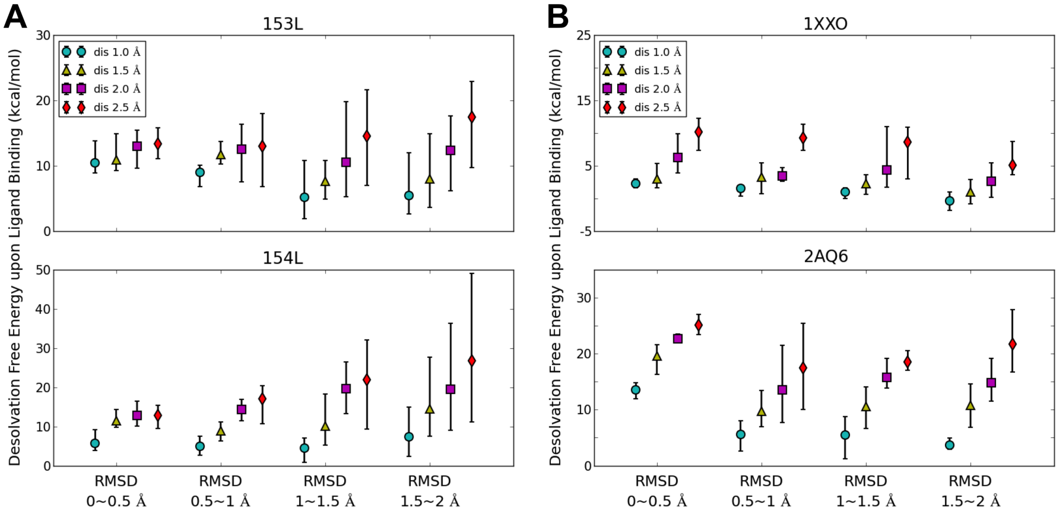


Fig. 2.9. The variation of de-solvation free energy involved in replacing water molecules upon ligand binding for four RMSD groups using different distance cutoff values. A: Goose egg-white lysozyme system (GEWL) (PDB: 153L, 154L). B: pyridoxine 5'-phosphate oxidase system (PLP) (PDB: 1XXO, 2AQ6).

Our study also demonstrates that the conformations of binding site residues significantly influence the prediction of hydration site locations and thermodynamic profiles and thus the de-solvation free energies associated with replacing water molecules upon ligand binding. The predicted locations of hydration sites and the computed free energies for all paired hydration sites are only consistent if the binding site residues have similar conformations (RMSD < 0.5 Å). More than 80% of the hydration sites have similar locations if the structures of the binding site are similar (RMSD 0 - 0.5 Å) but this percentage declined significantly with increasing deviations in the starting protein conformations. Thus, our study provides guidance on how similar protein structures need to be in order to obtain consistent hydration site predictions.

This sensitivity has important implications in drug discovery although it is typically not sufficiently considered by practitioners in the field. Often a limited set of X-ray structures with different types of ligands for a target protein is available. Furthermore, sometimes protein structures are significantly different dependent on the bound ligand. Thus, the question arises if a holo crystal structure for one lead compound can be used to predict hydration sites and use those for analysis of another lead series. Or can an apo structure be useful for hydration site prediction for a ligand-bound form of the same protein? The results of our study provide a first guidance to users of MD-based hydration-site programs with respect to those questions. An alternative grid-based approach, the grid inhomogeneous solvation theory (GIST) has been recently designed [45] potentially overcoming some of the observed sensitivity of the hydration site approaches. GIST computes de-solvation energies on individual grid points covering the binding site of the protein. For different protein conformations, different water density contours and different de-solvation energies are likely to be observed in GIST, too. However, GIST does not require a definition of hydration site. For localized high water density spots, hydration sites can be reliably predicted using clustering techniques. In those cases, conformational changes in the protein are equally resembled in positional changes in the high density spots and the representing hydration sites. For areas in the binding site with less pronounced water density peaks, e.g. more mobile water molecules, the definition of the hydration sites is sensitive to the clustering algorithm. As a consequence, small conformational changes of the protein can result in quite different hydration site positions. This may be a case where grid-based approaches could have advantages as the sensitivity of the clustering algorithm on small changes in non-localized water density is removed from the analysis. It would be interesting to perform studies similar to ours using those grid-based approaches to validate or falsify the hypothesis that grid-based approaches may be less susceptible to conformational differences in protein structure.

Whereas the influence of protein conformation on hydration site location and profiling is not surprising, our study provides a first quantification of this effect. To incorporate protein flexibility into hydration site prediction, two simple approaches could be thought of. First, unrestrained MD simulations with explicit water molecules could be performed, and the trajectory can be clustered. The clustering procedure will generate clusters of similar protein structures (e.g. with RMSD < 0.5 Å between structures of each cluster). Since the protein structures within a cluster are similar any frame could be used as reference for alignment, and subsequently hydration sites would be predicted for each cluster or "sub-trajectory" separately. Second, alternative protein conformation could be generated first using MD simulations and clustering, and subsequent simulations with position restraint on protein atoms could be performed for each protein conformation to obtain hydration site information. The latter has been adopted in this study. In both scenarios, clustering of MD snapshots has to be performed to separate alternative conformations for separate hydration site analysis. Our study provides a first guideline on the cluster size that should be chosen to obtain consistent hydration site predictions. Our data suggests that very narrow clusters seem to be required to obtain consistent estimates for hydration site locations, thermodynamic profiles and therefore protein de-solvation energies. Even protein conformations that deviate about 1 Å in RMSD can result in an average of 6.1 kcal/mol variations in de-solvation estimates.

3. INCORPORATING PROTEIN FLEXIBILITY INTO EXPLICIT HYDRATION SITE PREDICTION

3.1 Introduction

The explicit consideration of water molecules has gained increasing attention in drug-design projects over the last decade. Several computational tools based on MD simulations have been developed, including WaterMap from Schrodinger Inc., GIST from Gilson et. al. and WATsite from Lill et. al., to predict the localized position and thermodynamic profile of water molecules (i.e. hydration site) in the active site using either the ligand-bound (holo) or ligand-free (apo) protein conformation. Typically a single starting structure, either apo or holo form, of the target protein with restraint on the non-hydrogen atoms is used for subsequent MD simulations and hydration site prediction. The previously studies implicitly assumed that structural differences between apo and holo structure are small and do not influence hydration site prediction.

Our recent study[ref], however, indicated hydration site prediction is largely affected by small conformational variations in the initial protein structure used for the computational analysis. For most protein systems the rigid receptor assumption is invalid and we have to assume that large variations in hydration site predictions exist for protein systems with significant conformational change between apo and holo structure. Previous hydration-site related studies neglect the conformational transition of the protein upon ligand binding. Thus, the use of hydration site information from a single protein conformation is most likely incorrect for ligand binding (Figure 3.1).

There may be hydration sites predicted in the apo protein conformation, which are not present in the holo form (water α in Figure 3.1 A). These hydration sites are

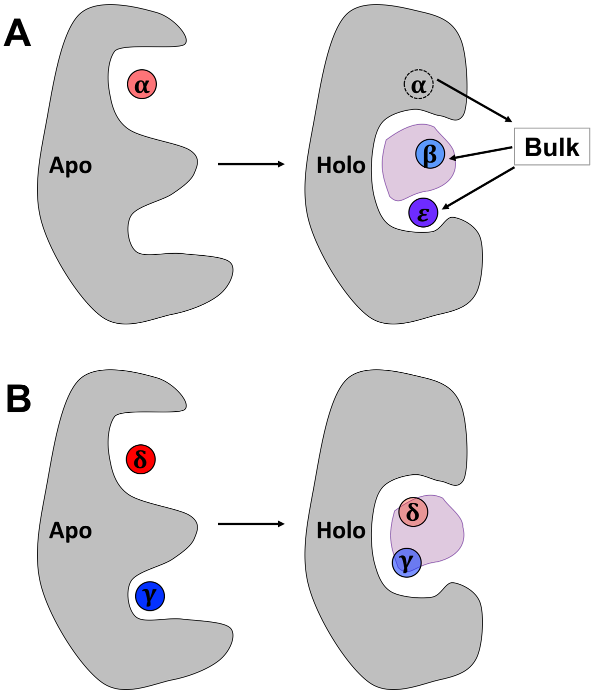


Fig. 3.1. Scheme for different scenarios of the influence of conformation change of the protein on explicit water molecules involved in protein-ligand binding. Free energy of each hydration site scales from blue (more favorable) to red (less favorable binding). (A) Positional change of hydration sites upon conformational change upon ligand binding: α leaves, β and ϵ enter the protein active site. Upon ligand binding, water β, although replaced by the ligand in the holo state, does not contribute to the de-solvation energy as it is in the bulk solvent in the apo state. In contrary, water α, although not directly replaced in the holo state by the ligand, contributes to the protein's de-solvation free energy as it is replaced by the conformational change induced by ligand binding. Contrary, water ϵ contributes to the de-solvation free energy. (B) Change of thermodynamic profiles of hydration sites during conformational change upon ligand binding: γ becomes less stable, and δ becomes more stable. The de-solvation energy of those waters in the apo state (not holo state) needs to be taken into consideration for the free energy of binding calculation.

not directly displaced by the bound ligand; however, they contribute to the protein de-solvation free energy since the ligand-induced conformational change causes their disappearance. On the other hand, hydration sites may be predicted in the holo form of the protein and displaced upon ligand binding (water β in Figure 3.1 A) although it is actually not present in the ligand-free apo state of the protein. Thus, adding water 's de-solvation energy to the free energy of binding upon replacement by the bound ligand is actually incorrect. Also, the protein's de-solvation free energy is the free energy difference between the water being released to the bulk solvent due to ligand binding versus the water bound in the ligand-free apo state. When the thermodynamic profile of a hydration site in apo and holo state differs, using the energies from the holo state is again incorrect for the estimation of the free energy of binding (water γ, ϵ in Figure 3.1 B).

3.2 Materials and Methods

With the aim of incorporating protein flexibility into hydration site prediction, we have developed two methods to dissect the changes in location and free energy of hydration sites upon protein conformational change. Method I involves the simulation of the conformational change by Multiply-Targeted Molecular Dynamics (MTMD), and provides a detailed transition of each hydration site throughout the trajectory. In method II, the locations of hydration sites are specified by local coordinate systems defined by nearby protein residues. Using these hydration-site specific coordinate systems (HSSCS), hydration sites in the apo protein structure are directly associated with those in the holo form without the necessity to follow the conformational transition path. In addition, we compare the results from both methods. Whereas method I provides a detailed explanation of appearance and disappearance of hydration sites during protein conformational change, method II is computationally more efficient, identifies a larger number of hydration-site pairs compared to method I and can be applied to a large number of ligands that bind to a diverse ensemble of protein conformational states.

3.2.1 Protein Systems and Preparation

The ligand-free and ligand-bound form of yeast guanylate kinase (GUA, PDBID: 1EX6, 1EX7) [59,60] and the SH2 domain of (Pp60) Src kinase system (SH2, PDBID: 1BKL, 1O42) [61] were used to test our methods. The GUA system undergoes large conformational change upon ligand binding, and the heavy-atom Root Mean Square Deviation in the binding site ($\mathrm{RMSD_{BS}}$) between the apo and holo form is 4.53 Å. On the other hand, with an RMSDBS of 1.62 Å, the SH2 system represents a system with smaller conformational change upon ligand binding. The binding site was defined to contain all residues which have at least one atom within 6 Å of any ligand atom in the X-ray structure of the holo form.

The apo protein conformations were prepared as input for Multiply-Targeted Molecular Dynamics (MTMD) simulations, keeping the crystallographic waters. The program Reduce [57] was used to adjust the side-chain conformations of ASN, GLN, and HIS, and tautomers and protonation states of the HIS residues. The protein was then solvated in an octahedron of water molecules using the SPC water model [58] with a minimum distance of 15 Å between any protein atom and the faces of the octahedron. Chlorine and sodium ions were added to neutralize the simulation systems. The holo protein structures with the ligand removed from the active site were only used as the end-point reference structure for MTMD.

3.2.2 MTMD Simulations of Ligand-Induced Protein Conformational Change

All simulations were performed with Amber14 [62] using an NPT ensemble with the Amber ff99SB-ILDN force field [63], periodic boundary conditions, and a timestep for integrating the equations of motion of 2 fs. The electrostatic interactions were calculated using the Particle Mesh Ewald method [64,65]. The Lennard-Jones interactions were truncated at a distance of 10 Å. The covalent bonds including a hydrogen atom were constrained using the SHAKE algorithm [50]. Temperature coupling was performed using a Langevin thermostat [66] with collision frequency of 2 ps^{-1} at 300 K, and isotropic position scaling with a pressure relaxation time of 1 ps for pressure coupling at 1 bar.

The prepared apo protein structure was first energy minimized for 5000 steps using the steepest descent algorithm. Throughout the equilibration process the temperature of the system was progressively increased from 0 K to 300 K within 50 ps, and then held constant for 500 ps. During the time-course of the MTMD simulation, an additional energy term based on the mass-weighted RMSD was added as a restraint term. The original apo protein conformation was used as the reference structure at the start of the MTMD simulation, and the holo conformation as the end-point reference structure.

A 4 ns MD simulation was performed with restraints on the apo conformation alone, followed by a 20 ns MTMD simulation where the RMSD restraints to the apo conformation were gradually decreased while increasing the corresponding RMSD restraints to the holo conformation. Finally, the MTMD simulation concluded with another 4 ns MD simulation with restraints only to the holo conformation. Since many different protein conformations may have the same RMSD and we are especially interested in the hydration sites in the protein binding site, both the binding site RMSD restraints and all heavy atoms RMSD restraints were applied in the MTMD simulation (Figure 3.2). Thus, four energy terms (both binding site and all heavy atoms RMSD for apo and holo protein) were added to the restraint term in the energy function. As shown in Figure 3.2, the MTMD simulation is initially restrained to the apo conformation for 4 ns, i.e. low RMSD to the apo conformation and high RMSD with respect to the holo form (heavy atom RMSD to binding site residues and all residues are 4.5 Å and 3.8 Å for the GUA system, respectively). During the next 20 ns the RMSD restraints to the holo protein conformation are gradually increased (RMSD to apo decreased) forcing the protein system slowly to the holo state, i.e. reduced

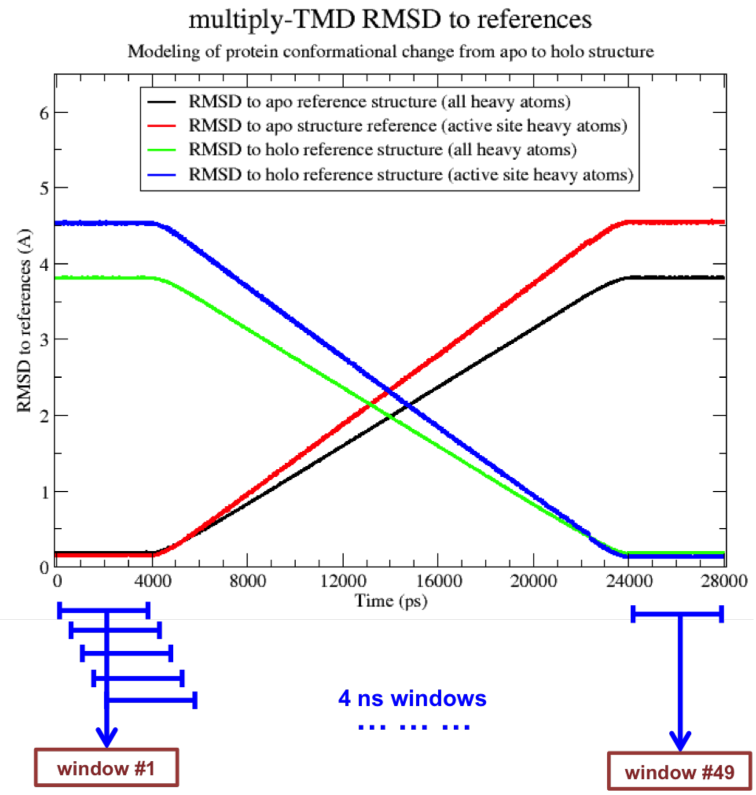


Fig. 3.2. Illustration of the 28 ns MTMD simulation of the protein conformational change from apo to holo structure. RMSD restraints is applied to both the binding site residues and the whole protein structure of GUA system. Using 0.5 ns intervals, 49 overlapping windows with a length of 4 ns were generated.

RMSD values with respect to holo structure (and increasing RMSD to apo form). In the last 4 ns period of the MTMD simulation the system remains restrained to the protein holo conformation. Coordinates were saved every picosecond, generating a protein conformational change trajectory with 28,000 frames.

3.2.3 Extracting Trajectories of Overlapping MTMD Windows

The MTMD trajectory was split into multiple overlapping 4 ns windows to predict the hydration sites. A 3.5 ns overlap to the previous window was chosen to ensure a gradual transition of hydration site locations. As shown in Figure 3.2, the first set of hydration sites was predicted using the first 4 ns of the MTMD trajectory (window #1), then the second set of hydration sites was predicted using the window between 0.5 ns and 4.5 ns (window #2), and the last set of hydration sites was predicted using the last 4 ns of the MTMD trajectory (window #49).

3.2.4 Theory of Hydration Site Identification and de-solvation Free Energy Prediction

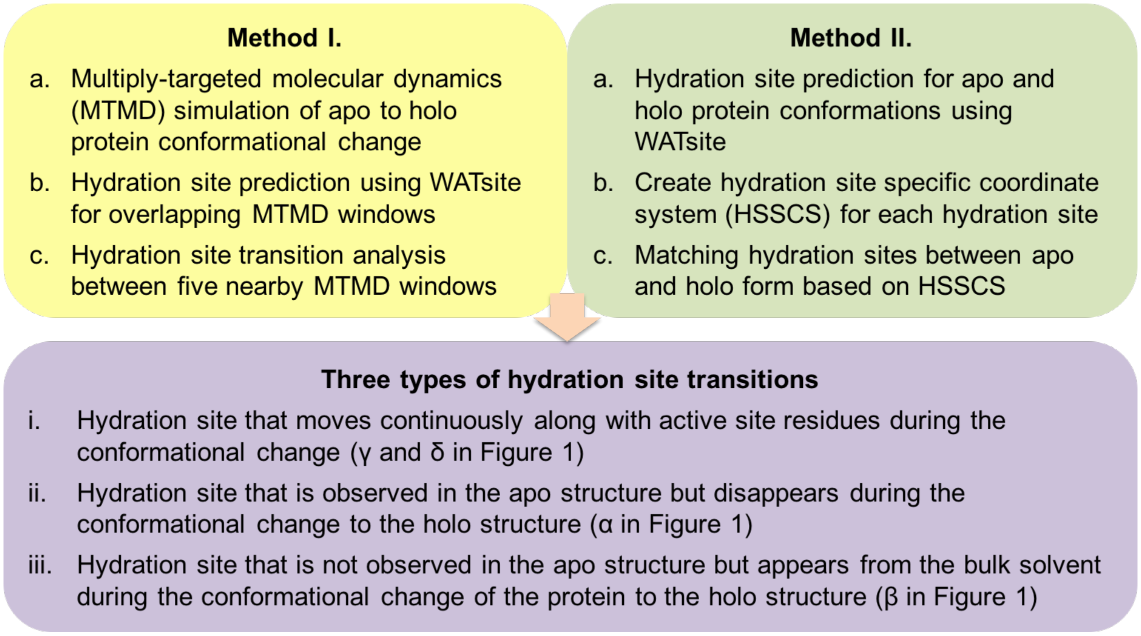
The theory of hydration site identification and de-solvation free energy prediction has been described in detail in Chapter 2.2.2 and Chapter 2.2.3.

3.2.5 Development of Two Methods for the Hydration Site Analysis Involving Protein Flexibility

Figure 3.3 shows the overall procedure of the two methods we have developed for the analysis of hydration site locations and thermodynamic profiles incorporating protein flexibility. Hydration sites predicted using the ligand unbound (apo) protein structure are called apoHS, and using the ligand bound (holo) protein conformation, holoHS.

Method I

Method I consists of three steps: a. MTMD simulation with explicit water molecules was used to simulate the conformational change of a protein from its apo structure to the holo form. b. The MTMD trajectory was then split into overlapping windows of 4 ns length with neighboring windows separated by 0.5 ns (Figure 3.2). Our hydration site prediction program WATsite [47] was used to predict the hydration sites of overlapping MTMD windows. c. The changes in hydration sites locations along the MTMD frames and their associated thermodynamic profiles were then an-



Method I.

a. Multiply-targeted molecular dynamics (MTMD) simulation of apo to holo protein conformational change

b. Hydration site prediction using WATsite for overlapping MTMD windows

c. Hydration site transition analysis between five nearby MTMD windows

Method II.

a. Hydration site prediction for apo and holo protein conformations using WATsite

b. Create hydration site specific coordinate system (HSSCS) for each hydration site

c. Matching hydration sites between apo and holo form based on HSSCS

Three types of hydration site transitions

i. Hydration site that moves continuously along with active site residues during the conformational change (γ and δ in Figure 1)

ii. Hydration site that is observed in the apo structure but disappears during the conformational change to the holo structure (α in Figure 1)

iii. Hydration site that is not observed in the apo structure but appears from the bulk solvent during the conformational change of the protein to the holo structure (β in Figure 1)

Fig. 3.3. Overview of the two methods for the analysis of hydration site locations and thermodynamic profiles during protein conformational change upon ligand binding.

alyzed. The potential disappearance, appearance and changes in thermodynamic profiles of hydration site waters during the conformational change were determined.

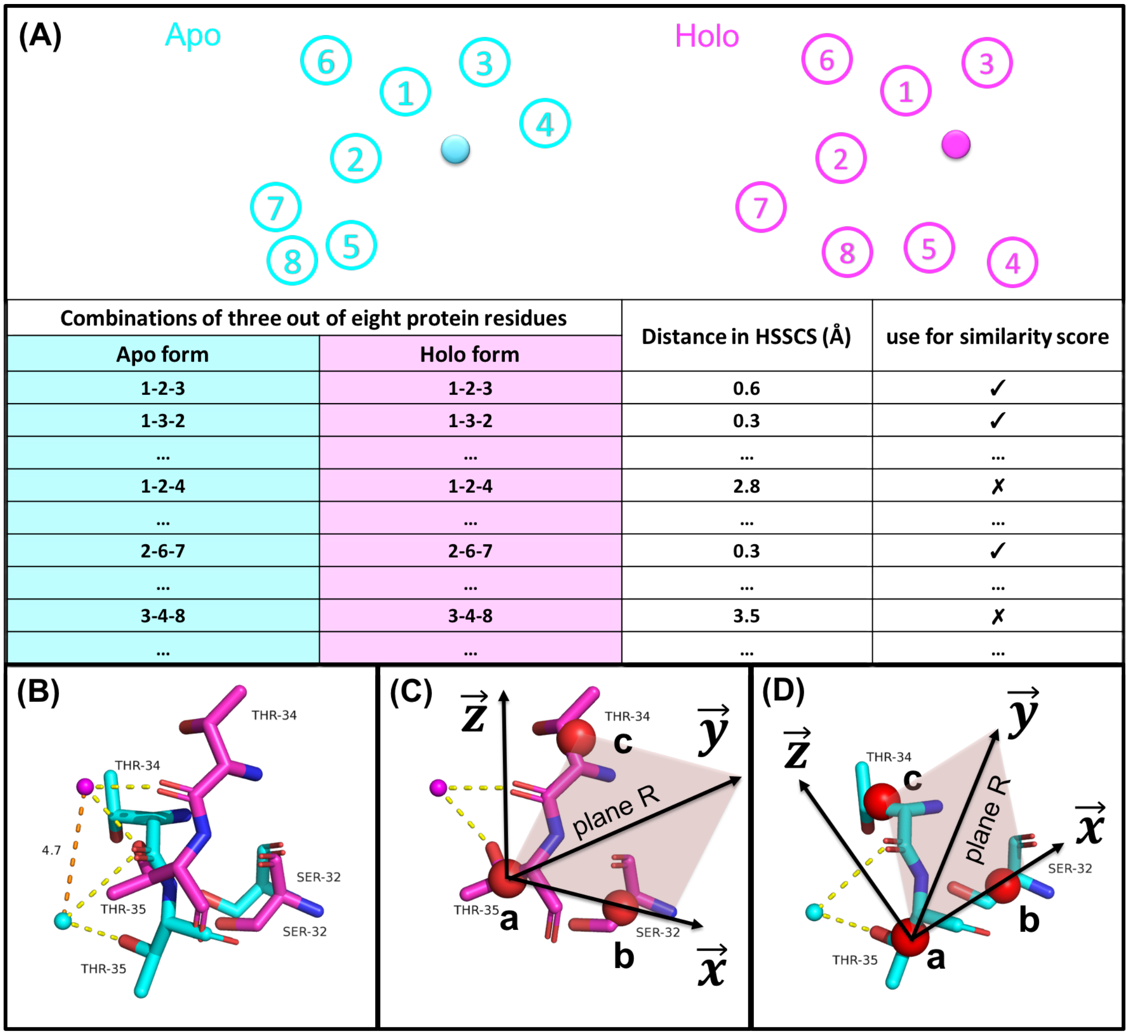
In detail, using the 49 sets of hydration sites predicted from the overlapping MTMD windows, we aimed to connect the hydration sites in the holo protein conformation to the corresponding hydration sites in the apo form defining continuous paths between the starting and ending locations. The transition of hydration sites was dissected by identifying the hydration-site pairs between neighboring MTMD windows. The closest two hydration sites between the current and the previous MTMD windows were considered as a hydration-site pair if the distance between them is smaller than 1 Å. The 1 Å threshold for defining similar hydration site locations was chosen as the hydration sites are defined as spheres with radius of 1 Å [46]. If no hydration-site pair was identified between the neighboring windows n and n-1, the hydration site comparison was performed between the current window n and window n-2. This protocol was continued (comparison n with n-3, then n with n-4, finally n with n-5) until hydration sites were paired or no pairing was identified up to fifth previous window n-5. As the RMSD of the protein conformational change from apo to holo for GUA system is about 4.5 Å, using 49 MTMD windows means that the protein changes about 0.5 Å in RMSD throughout five MTMD windows. As our previous study indicated [67], hydration sites are incomparable for protein binding site residues with RMSD larger than 0.5 Å. Thus, we limit the comparison of the positions of hydration sites to the previous five MTMD windows. Once all paired hydration sites are identified, the transition of hydration sites is tracked from the protein apo form to its holo form, and in opposite direction, incorporating the conformational change of the protein upon ligand binding. If one hydration site is not paired to any hydration site in the previous five windows, it is considered as a new hydration site water that appears from bulk solvent throughout the protein's conformational change (β in Figure 3.1). Hydration sites that are not paired with any hydration site in the subsequent five windows are considered as sites that disappear during conformational change to the bulk solvent, i.e. the ligand-induced conformational change of the protein displaced the hydration site water into bulk solvent (α in Figure 3.1).

Method II

Method II is based on the observation that a water molecule in the binding site is often stabilized by direct contacts with a single or small number of residues, especially due to specific hydrogen bond interactions. Thus, in this method hydration sites in apo and holo form of the protein are associated based on their similarity in interactions with nearby residues using a local coordinate system, named hydration-site specific coordinate systems (HSSCS). This allows to associate the hydration sites in the holo protein conformation to the corresponding hydration sites in the apo form without explicit simulation of the protein conformational change process.

In detail, the eight closest binding site residues are identified for each hydration site (Figure 3.4 A). Since three points are required to create a coordinate system, any three out of the eight residues are used for the definition of HSSCS, giving a total of 56 combinations. The centroids of the residues (denoted as a, b, c in Figure 3.4 B and C) define a plane R. The x-axis is defined by vector $\vec{ab}$ and the y-axis is on the plane R and perpendicular to vector $\vec{ab}$. The z-axis is the perpendicular vector to both the x- and y-axis in a right handed coordinate system. The Cartesian coordinates of each hydration site are projected onto the HSSCS, and the residues defining the HSSCS along with the local coordinates of the hydration site are stored. It should be noted that the sequence of three centroids does influence the local coordinate system. For example, the HSSCS with x-axis defined by vector $\vec{ab}$ or $\vec{bc}$ are different. Thus, there are six permutations for each combination of three residues; thus, a hydration site can be represented by 336 (6x56) HSSCS.

Centroids of residues instead of potential hydrogen bond donor/acceptor atoms are chosen to define the coordinate system for the following reasons: First, potential hydrogen bond donor/acceptor atoms may only be good reference points for hydration



Combinations of three out of eight protein residues		Distance in HSSCS (Å)	use for similarity score
Apo form	Holo form		
1-2-3	1-2-3	0.6	✓
1-3-2	1-3-2	0.3	✓
…	…	…	…
1-2-4	1-2-4	2.8	✗
…	…	…	…
2-6-7	2-6-7	0.3	✓
…	…	…	…
3-4-8	3-4-8	3.5	✗
…	…	…	…

Fig. 3.4. Illustration of the Hydration Site-Specific Coordinate System (HSSCS) method. (A) All combinations of three out of eight closest residues are used to define a HSSCS. (B) The hydration sites associated with THR-34 and THR-35 are 4.7 Å apart between apo (cyan) and holo (magenta) form when measured in the Cartesian coordinate system of aligned apo and holo protein structure. (C)(D) Three red spheres (a, b and c) defining a plane R are used to create a HSSCS coordinate system. The x-axis is defined by spheres a and b, the y-axis is on the same plane and perpendicular to the x-axis, and the z-axis is perpendicular to the plane R. The two hydration sites are essentially co-localized in terms of the HSSCS.

site that are stabilized predominantly via hydrogen bonding. However, for hydration sites in a hydrophobic cavity, it is unclear which atoms to pick as reference points for defining the coordinate system, as the cavity is formed by a multitude of atoms. Second, there are cases where a hydration site is close to one polar and two nonpolar residues in the apo form. However, in the holo form, this polar residue is no longer part of the protein binding site, while the other two nonpolar residues are still stabilizing the hydration site. Third, for cases where a hydration site is stabilized by hydrogen bonding with a single polar residue, using three atoms (e.g., Cα, hydrogen bond donor/acceptor atoms, or main chain N/O) on the same polar residue as reference points may cause numerical instabilities in defining the hydration site position in this particular coordinate system. The reason for this observation is that these atoms are sometimes closer to each other compared to the potentially hydrogen-bonded hydration site. Consequently, the coordinate axes may change significantly with small variance in the atom positions and the same hydration site may artificially have very different coordinates in the two Cartesian coordinate systems. Fourth, the positions of the atoms are thermally fluctuating much larger than the residue centroids. Thus, differences in the hydration site locations using atom-based coordinate systems may be representing the thermal fluctuation rather than conformational transitions between two stable sub-states associated with apo and holo form of the protein.

As Cartesian coordinate system can be very sensitive to the chosen points when they approach co-linearity, we checked the scalar product of the angle theta formed by vector $\vec{ab}$ and $\vec{ac}$, where a, b and c are the coordinates of the three centroids. For the holo protein conformation of the GUA system, none of the coordinate systems out of 13440 has a scalar product larger than 0.9 (cos25°=0.906). For the apo protein conformation of the GUA system, only 3.6% of all coordinate systems (354 out of 9744) have a scalar product larger than 0.9. We need to mention that a voting scheme is used here over all coordinate systems. Such a small percentage of sensitive coordinate system is not believed to significantly influence the overall outcome. To identify the associated hydration sites in the apo and holo protein conformation, all hydration sites from the apo form are compared to those in the holo form. For any hydration site comparison, distances in HSSCS coordinates are calculated for which the three residues and the permutation are the same. Any two hydration sites with a distance smaller than 2 Å were identified and stored as possible pair. We noticed that two close hydration sites between nearby windows in Cartesian coordinates with distance 0.2 Å have larger distance variations (from 0.1 Å to 0.9 Å) in HSSCS coordinates. Thus, we used 2 Å as the distance criteria for pairing hydration sites using HSSCS coordinates. Using these local coordinate systems, pairs are identified even if the RMSD of the three residues defining the HSSCS between apo and holo form is larger than 10 Å in the Cartesian coordinate system. Different pairs may be identified depending on the three residues defining the HSSCS (Figure 3.4 A). Furthermore, hydration-site pairs that are associated with HSSCS defined by close residues are typically of more importance than using HSSCS of more distant residues because the interaction strength is likely to be larger in the former case, especially when the hydration site forms hydrogen bonds with those residues. Thus, to identify the most likely hydration-site pair between apo and holo form, a hydration-site similarity score was defined that depends on the number of HSSCS in which the pair i-j is matched (Figure 3.4 A) and the distance ($d_{(i,j)-0}$) between the hydration site and the origin of the HSSCS averaged over the apo, d_{i0}, and holo conformation, d_{j0}, (i.e., centroid a in Figure 3.4 C and D). The pair with highest similarity score (Equation 3.1) were selected to be the associated hydration sites between the apo and the holo form. A distance cutoff (d_{cutoff}) of 4 Å was chosen. The highest similarity score might be different from apo-to-holo and holo-to-apo direction. For one hydration-site pair, 672 (336x2) is the highest possible similarity score if they are paired in all HSSCS and all distances to the origin of HSSCS, $d_{(i,j)-0}$, are smaller than the distance cutoff (d_{cutoff}). The similarity score was normalized to a range of 0% to 100% by dividing the raw value by the highest possible similarity score (672), yielding the percent similarity score $s_{i,j}$. Finally, all hydration-site pairs with a percent similarity score larger than 1% were used and discussed in the result section. We also extended method II, allowing one hydration site to be paired with multiple sites, and titled this modification method IIb.

$$S_{i,j} = \sum_{k}^{matched\ permutations} \begin{cases} 1, & d_{(i,j)-0} < d_{cutoff} \\ exp(-0.1(d_{(i,j)-0} - d_{cutoff})), & d_{(i,j)-0} \geq d_{cutoff} \end{cases} \tag{3.1}$$

3.2.6 Estimation of Buried Apolar Surface Area of Protein-Ligand Complexes

We estimated the buried apolar Solvent Accessible Surface Area (SASA) using the get_area command in PyMol [68]. The buried apolar SASA was computed by summing up the apolar SASA of the protein and the ligand, and then subtracting the apolar SASA of the complex.

3.2.7 Estimation of de-solvation Free Energy of the Protein upon Ligand Binding

In previous studies [44, 46], hydration sites are predicted using the holo protein structure alone, and thus the protein de-solvation free energy is estimated using the hydration sites in the holo form. Any hydration site to any of the ligands' heavy atoms with distance (d_{hs-lig}) smaller than a distance cutoff (e.g., $\mathrm{d}_{lig_{c}utoff}$ = 2.24 Å) are considered as been displaced upon ligand binding. The distance cutoff is adapted from Abel et. al. since it is very unlikely that the contact distance between a water-oxygen atom and a ligand carbon atom is less than 0.8 (1.4 Å + 1.4 Å) = 2.24 Å (assuming the radii of carbon and oxygen atoms are approximately 1.4 Å) [44]. The de-solvation free energy of the protein binding site (ΔG_{desolv_rigid}) (Equation 3.2) can then be estimated by summing up the free energies of the hydration sites in the holo form, with respect to bulk solvent, that are displaced upon ligand binding. The contribution of each displaced hydration site is based on the closeness to the ligand:

$$\Delta G_{desolv\_rigid} = \sum\nolimits_{holoHS,lig} \Delta G_{holoHS}(1 - \frac{d_{hs-lig}}{d_{lig\_cutoff}}) \tag{3.2}$$

However, when a protein conformational change is associated with ligand binding, the hydration site free energy profiles in the apo protein state need to be considered as reference for the ligand-free state, i.e. the thermodynamic properties of the apo hydration sites, that correspond to the holo hydration sites that are replaced by the ligand, have to be used to calculate the protein de-solvation free energy. To fully take the influence of protein conformational change on protein de-solvation into account, three types of hydration site transitions are considered for the estimation of protein de-solvation free energy:

Type i. Hydration sites that are displaced upon ligand binding in the protein holo form and match to a corresponding hydration site in the apo form (γ and δ in Figure 3.1 B). The free energy of the paired hydration sites in the apo form ($\Delta G_{\gamma/\delta\_apo}$) is used.

Type ii. Hydration sites that exist in the protein apo form but disappear in the protein holo form (α in Figure 3.1 A). These hydration sites are contributing to the free energy of protein de-solvation although they are not directly displaced by the ligand. The reason for this treatment is that ligand binding induces conformational change of the protein which is responsible for the de-solvation effect. Ligand binding therefore indirectly replaces those water molecules. This water replacement into bulk solvent, however, may cost or gain free energy (ΔG_{α}) to ligand binding and therefore has to be considered for computing the de-solvation free energy.

Type iii. If the hydration site does not have a corresponding site in the apo form, i.e. it only appears from bulk solvent during protein conformational change, its replacement does not contribute to the free energy of ligand binding (β in Figure 3.1 A), and ΔG_{β} is therefore not considered in the estimation of the free energy of protein de-solvation. If a type (iii) hydration site is not displaced by the ligand (ϵ in Figure 3.1 A), it unfavorably (-ΔG_{ϵ}) contributes to the free energy of protein de-solvation.

Thus, the overall de-solvation free energy is computed by Equation 3.3:

$$\begin{aligned}\Delta G_{desolv\_flexible} = &\sum\nolimits_{\Big(type\_i} \left[\Delta G_{\gamma/\delta\_apo}\ (1 - \frac{d_{hs-lig}}{d_{lig\_cutoff}})\right] \\ &+ \sum\nolimits_{type\_ii} \Delta G_{\alpha} - \sum\nolimits_{\Bigg(type\_iii} \Delta G_{\epsilon}\end{aligned} \tag{3.3}$$

In the modified method IIb, multiple pairings are allowed for one hydration site, the percent similarity score (s) of each pair is used to scale the contribution from different corresponding hydration-site pairs in the apoHS. For type i hydration sites ($\Delta G_{\gamma/\delta}$), for example, when multiple hydration sites in apoHS (e.g. l, m, n) are paired with a hydration site (e.g. k) in the holoHS, the contribution from displaced hydration site k to the protein de-solvation free energy is estimated by Equation 3.4:

$$\begin{aligned}\Delta G_k =& \Delta G_l \left(\frac{s_{k,l}}{s_{k,l} + s_{k,m} + s_{k,n}}\right)\Bigg(\\ &+ \Delta G_m \left(\frac{s_{k,m}}{s_{k,l} + s_{k,m} + s_{k,n}}\right)\Bigg(\\ &+ \Delta G_n \left(\frac{s_{k,n}}{s_{k,l} + s_{k,m} + s_{k,n}}\right)\Bigg(\end{aligned} \tag{3.4}$$

The de-solvation free energy from type i hydration sites are calculated using equation 3.4, and the contributions from type ii and iii hydration sites are calculated as in equation 3.3.

3.3 Results and Discussion

Current methods predict hydration sites and their thermodynamic profile based on an initial protein structure, typically a holo form of the protein. Our previous study [67] indicated that even for proteins with relatively small conformational change (0.5 Å ¡ RMSDBS ¡ 1.0 Å) between apo and holo conformations large variations in location and thermodynamic profile of the hydration sites is typically observed. Here, we propose two methods to identify hydration sites in the protein holo form (holoHS) that originated from corresponding hydration sites in the apo form (apoHS) but changed their location due to the conformational transition between both protein states. In method I, we simulate a protein conformational transition pathway, and predict hydration site changes observed throughout the transition. The hydration sites in the apo form are connected to those in the holo form via locally adjacent, intermediate hydration-site pairs. In method II, we assign hydration sites to the surrounding protein residues, and match the hydration site in the apo form to corresponding sites in the holo form based on their conserved interactions with those nearby residues.

3.3.1 Protein Conformational Change from the Apo to the Holo Form Simulated by MTMD

Protein systems that undergo significant conformational change upon ligand binding are commonly seen in drug discovery, and many of them are assumed to occur via the induced-fit mechanism. When ligands induce large conformational change, it is important to understand what influence this conformational change has on binding site water molecules. Here, we simulated the protein conformational change by applying RMSD restraints on the heavy atoms of the protein slowly shifting the system from apo to holo conformation.

Figure 3.5 shows an overlay of the midpoint MD frame from each 4 ns MTMD window representing the conformational transition pathway for the GUA and SH2 systems. The GUA system undergoes a large conformational change of the loop region that interacts with the phosphate group of the ligand (Figure 3.5 A). The $\mathrm{RMSD_{BS}}$ between the apo and holo form of the GUA protein system is 4.53 Å, thus the hydration sites predicted using these two protein conformations are not overlapping using the Cartesian coordinates of the superimposed protein structures (Figure 3.6 A). With an $\mathrm{RMSD_{BS}}$ of 1.62 Å, the conformational change of the SH2 system is relatively small compared to the GUA system, but the hydration sites predicted using the apo and the holo forms are still spatially distinct (Figure 3.7 A). As the SH2 system undergoes smaller conformational change compared to GUA, the MTMD simulation length was reduced to 20 ns including 12 ns with continuously

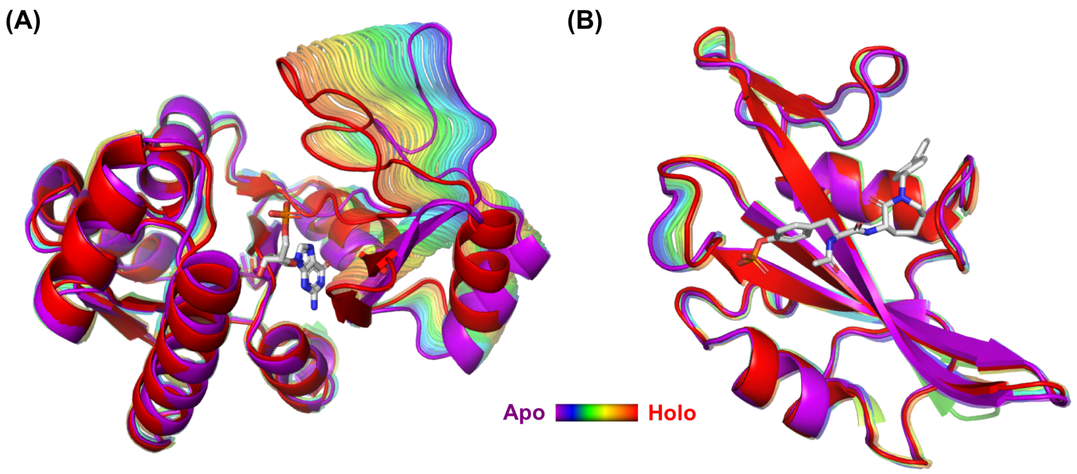


Fig. 3.5. Representative protein conformations of each 4 ns MTMD window. (A) GUA system; (B) SH2 system; representing the conformational transition pathway simulated by MTMD simulation (purple: apo structure, red: holo structure).

changing RMSD restraints to apo and the holo form (separate RMSD restraints for binding site and whole protein were used similarly as in the GUA simulations). In the first and the last 4 ns period of the MTMD simulation the system remains restrained to the protein apo and holo conformations, respectively.

3.3.2 Method I: Identify Hydration Site Transition Pathways through MTMD windows

For each 4 ns MTMD window, one set of hydration sites is predicted. However, a simple overlay of all predicted hydration sites from all windows is unable to directly identify the transition pathways for each site during the protein conformational change (Figure 3.6 B & 3.7 B). To identify the pathways of all hydration sites throughout the apo to holo transition pathway, we determined all hydration-site pairs between nearby MTMD windows. As described in the Materials and Methods section, a hydration-

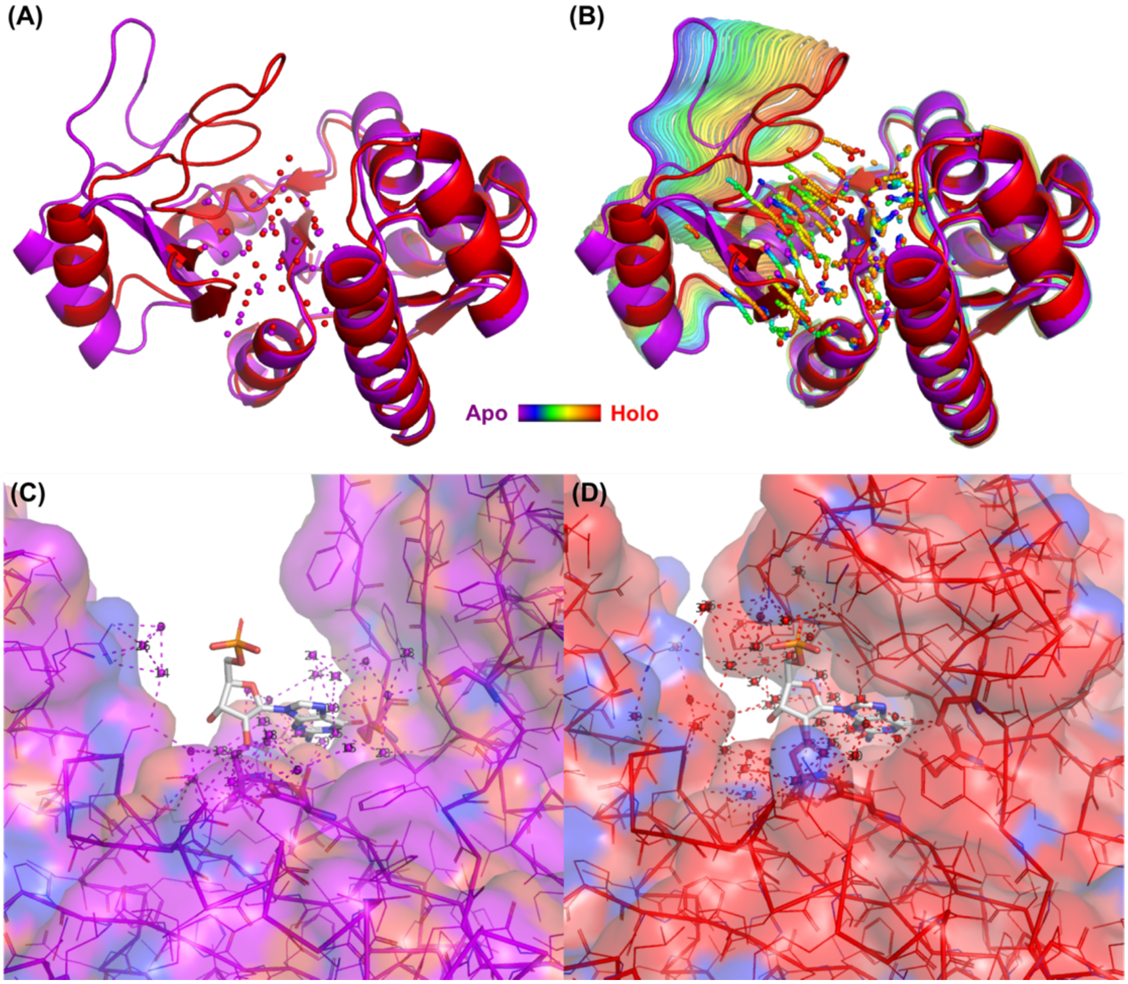


Fig. 3.6. Overlay of hydration sites identified in the GUA system. (A) Hydration sites predicted using the apo (Purple) and the holo (Red) protein conformation. (B) Hydration sites of all MTMD windows. (Purple: apo protein conformation, Red: holo protein conformation) (C) Surface representation of the apo protein conformation with the predicted hydration sites and the extracted ligand from the holo form. (D) Surface representation of the holo protein form with the predicted hydration sites and the bound ligand. The ligand binding pocket is observed in the holo form.

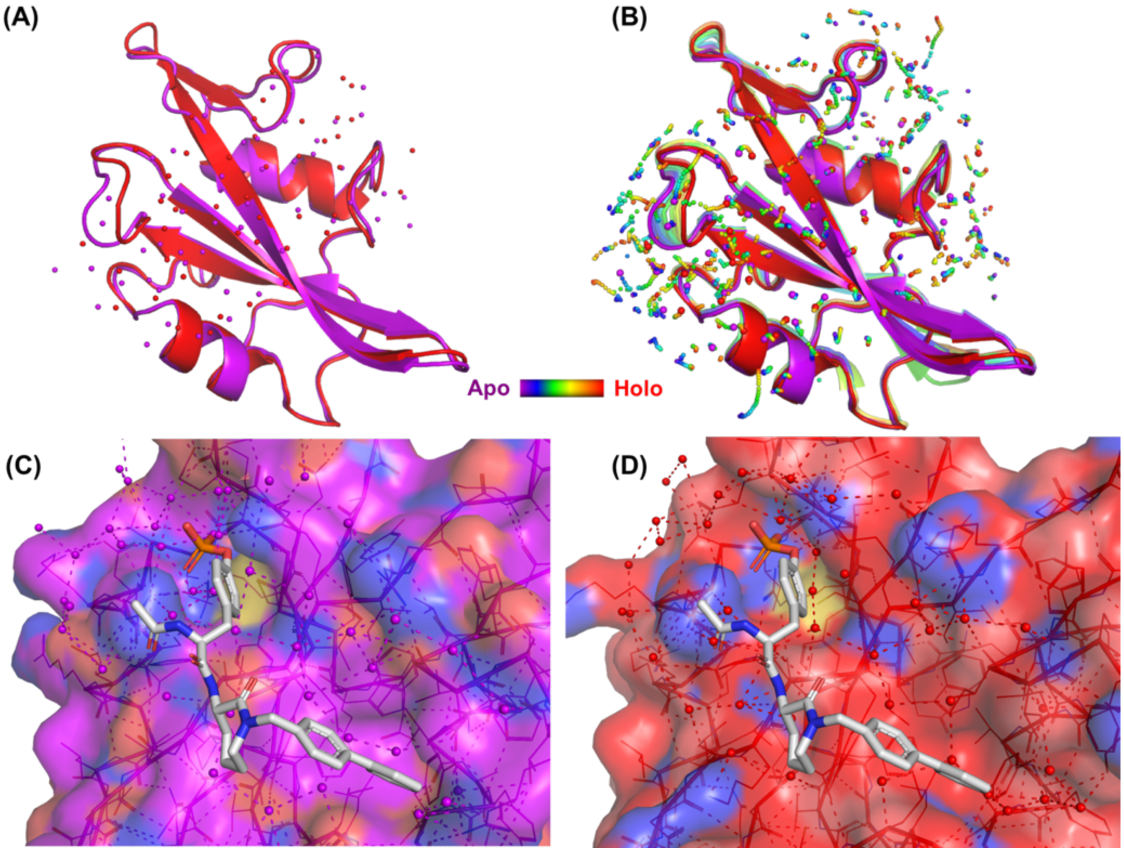


Fig. 3.7. Overlay of hydration sites identified in the SH2 domain. (A) Hydration sites predicted using the apo (Purple) and the holo (Red) protein conformation. (B) Hydration sites of all MTMD windows. (Purple: apo protein conformation, Red: holo protein conformation) (C) Surface representation of the apo protein conformation with the predicted hydration sites and the extracted ligand from the holo form. (D) Surface representation of the holo protein form with the predicted hydration sites and the bound ligand. The ligand binding site is solvent exposed.

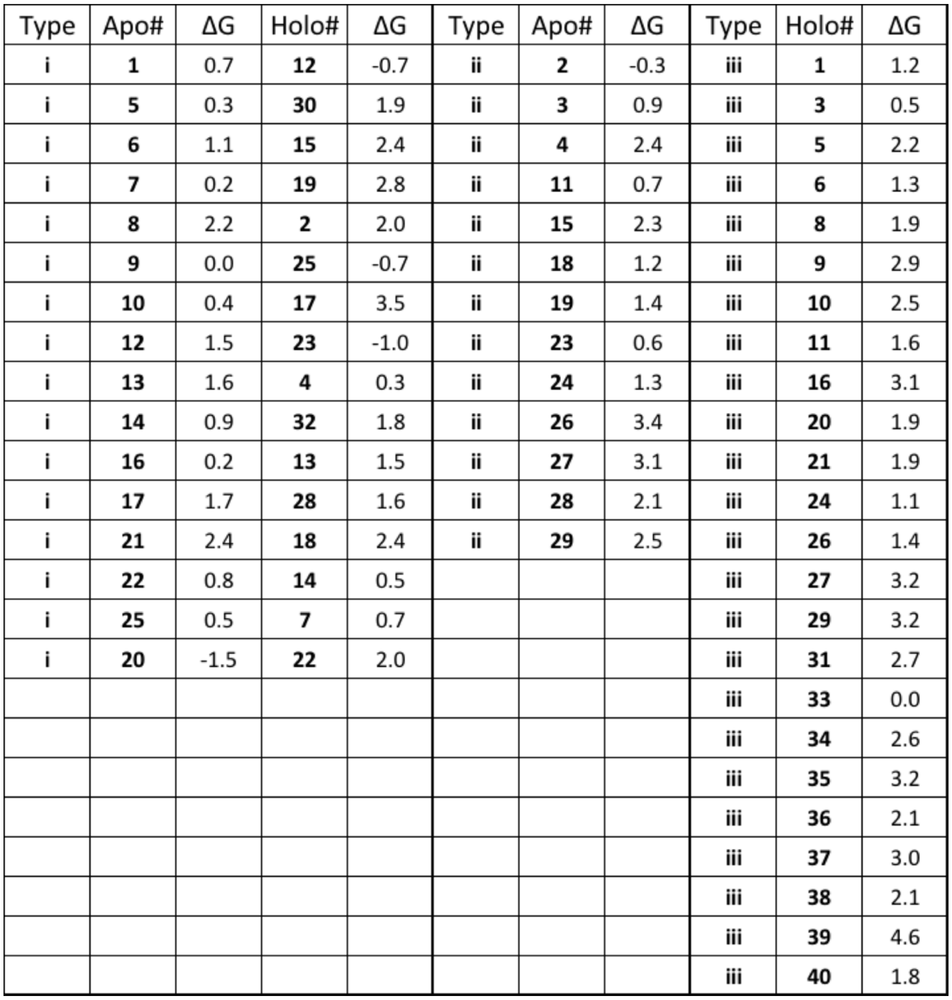
site pair is defined if the closest two hydration sites from two adjacent windows have a distance smaller than 1 Å.

Based on the spatial relocation, appearance and disappearance of hydration sites during the protein conformational change, three types of hydration site transitions are identified: (i) Hydration sites that move continuously along with active site residues during conformational change or remain largely unchanged in their location, (ii) hydration sites observed in the apo structure but that disappear during the conformational change to the holo form (i.e. transition into bulk solvent), and (iii) hydration sites that are not observed in the apo structure but appear during the conformational change to the holo form (i.e. transition from the bulk solvent to the binding site). For the calculation of de-solvation free energies, only water molecules that are located to hydration sites in the apo or holo form are relevant. Hydration sites that only appear in intermediate windows do not contribute to the computation of protein de-solvation free energy, and will therefore not be discussed here.

The detailed results for all hydrations site transitions are tabulated in Table 3.1 (GUA) and Table 3.2 (SH2). For the GUA system, our method predicted 29 hydration sites in the apo structure and 40 in the holo structure (Figure 3.6). 16 hydration sites belong to type (i), and 13 to type (ii), and 24 to type (iii) hydration sites. For the SH2 system (Figure 3.7), 51 hydration sites were predicted in the apo conformation, and 46 in the holo form. 25 hydration sites were classified as type (i), 26 as type (ii), and 21 as type (iii) hydration sites. Thus, only approximately half of the hydration sites are conserved throughout the conformational transition in the binding site, highlighting the significant dynamics in hydration site appearance and disappearance during the apo to holo transition.

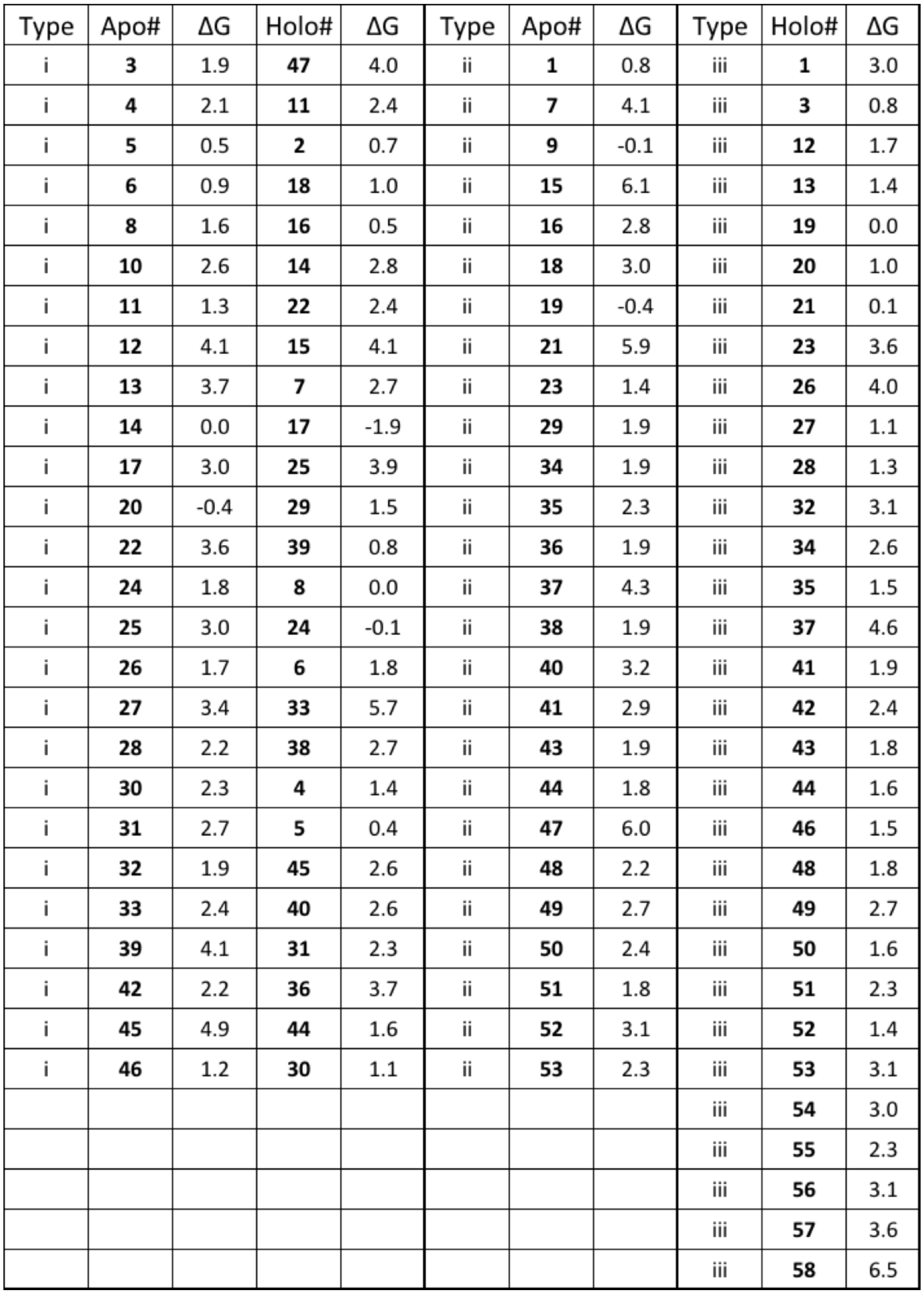
Figure 3.8 shows examples of hydration site transitions in the GUA system. The thermodynamic properties (ΔG, $-T\Delta S$, and ΔH) (in kcal/mol) and occupancy of each hydration site is shown on the right. Even for type (i) hydration sites (e.g. Figure 3.8) that are conserved in the binding site (no appearance or disappearance), the location of the hydration site with respect to the whole protein may alter significantly.

Table 3.1.
Three types of hydration sites identified in the GUA system using method I.



Type	Apo#	ΔG	Holo#	ΔG	Type	Apo#	ΔG	Type	Holo#	ΔG
i	1	0.7	12	-0.7	ii	2	-0.3	iii	1	1.2
i	5	0.3	30	1.9	ii	3	0.9	iii	3	0.5
i	6	1.1	15	2.4	ii	4	2.4	iii	5	2.2
i	7	0.2	19	2.8	ii	11	0.7	iii	6	1.3
i	8	2.2	2	2.0	ii	15	2.3	iii	8	1.9
i	9	0.0	25	-0.7	ii	18	1.2	iii	9	2.9
i	10	0.4	17	3.5	ii	19	1.4	iii	10	2.5
i	12	1.5	23	-1.0	ii	23	0.6	iii	11	1.6
i	13	1.6	4	0.3	ii	24	1.3	iii	16	3.1
i	14	0.9	32	1.8	ii	26	3.4	iii	20	1.9
i	16	0.2	13	1.5	ii	27	3.1	iii	21	1.9
i	17	1.7	28	1.6	ii	28	2.1	iii	24	1.1
i	21	2.4	18	2.4	ii	29	2.5	iii	26	1.4
i	22	0.8	14	0.5				iii	27	3.2
i	25	0.5	7	0.7				iii	29	3.2
i	20	-1.5	22	2.0				iii	31	2.7
								iii	33	0.0
								iii	34	2.6
								iii	35	3.2
								iii	36	2.1
								iii	37	3.0
								iii	38	2.1
								iii	39	4.6
								iii	40	1.8

Table 3.2.
Three types of hydration sites identified in the SH2 system using method I.



Type	Apo#	ΔG	Holo#	ΔG	Type	Apo#	ΔG	Type	Holo#	ΔG
i	**3**	1.9	**47**	4.0	ii	**1**	0.8	iii	**1**	3.0
i	**4**	2.1	**11**	2.4	ii	**7**	4.1	iii	**3**	0.8
i	**5**	0.5	**2**	0.7	ii	**9**	-0.1	iii	**12**	1.7
i	**6**	0.9	**18**	1.0	ii	**15**	6.1	iii	**13**	1.4
i	**8**	1.6	**16**	0.5	ii	**16**	2.8	iii	**19**	0.0
i	**10**	2.6	**14**	2.8	ii	**18**	3.0	iii	**20**	1.0
i	**11**	1.3	**22**	2.4	ii	**19**	-0.4	iii	**21**	0.1
i	**12**	4.1	**15**	4.1	ii	**21**	5.9	iii	**23**	3.6
i	**13**	3.7	**7**	2.7	ii	**23**	1.4	iii	**26**	4.0
i	**14**	0.0	**17**	-1.9	ii	**29**	1.9	iii	**27**	1.1
i	**17**	3.0	**25**	3.9	ii	**34**	1.9	iii	**28**	1.3
i	**20**	-0.4	**29**	1.5	ii	**35**	2.3	iii	**32**	3.1
i	**22**	3.6	**39**	0.8	ii	**36**	1.9	iii	**34**	2.6
i	**24**	1.8	**8**	0.0	ii	**37**	4.3	iii	**35**	1.5
i	**25**	3.0	**24**	-0.1	ii	**38**	1.9	iii	**37**	4.6
i	**26**	1.7	**6**	1.8	ii	**40**	3.2	iii	**41**	1.9
i	**27**	3.4	**33**	5.7	ii	**41**	2.9	iii	**42**	2.4
i	**28**	2.2	**38**	2.7	ii	**43**	1.9	iii	**43**	1.8
i	**30**	2.3	**4**	1.4	ii	**44**	1.8	iii	**44**	1.6
i	**31**	2.7	**5**	0.4	ii	**47**	6.0	iii	**46**	1.5
i	**32**	1.9	**45**	2.6	ii	**48**	2.2	iii	**48**	1.8
i	**33**	2.4	**40**	2.6	ii	**49**	2.7	iii	**49**	2.7
i	**39**	4.1	**31**	2.3	ii	**50**	2.4	iii	**50**	1.6
i	**42**	2.2	**36**	3.7	ii	**51**	1.8	iii	**51**	2.3
i	**45**	4.9	**44**	1.6	ii	**52**	3.1	iii	**52**	1.4
i	**46**	1.2	**30**	1.1	ii	**53**	2.3	iii	**53**	3.1
								iii	**54**	3.0
								iii	**55**	2.3
								iii	**56**	3.1
								iii	**57**	3.6
								iii	**58**	6.5

The water molecule in Figure 3.8 A, for example, interacts with ASP-100 in the apo protein structure via a hydrogen bond. Throughout the transition to the holo form, GLU-69 approaches the hydration site, and a second hydrogen bond is formed to GLU-69 in addition to the hydrogen bond to ASP-100. Due to the second hydrogen bond with a formally charged residue the enthalpy of solvation becomes more negative, associated with a loss in conformational freedom, i.e. increase in $-T\Delta S$ (Figure 3.8 A, right). Together with the increasing stability of the hydration site, the occupancy of this hydration site throughout a 4 ns MTMD window increases from about 60% in the apo conformation to 100% in the holo conformation. Although this hydration site is predicted to be conserved between apo and holo form, its predicted free energy of de-solvation in the apo form is 2.5 kcal/mol less favorable than in the holo form. Thus, when this hydration site is displaced by a binding ligand, its free energy predicted in the holo form should not be used, as the de-solvation free energy is defined with respect to the apo form of the protein.

An example of a type (ii) hydration site that disappears during the conformational change of the protein is shown in Figure 3.8 B. The hydration site in this example forms hydrogen-bond interactions with the carboxyl group of GLN-102 in the apo protein conformation. The carboxyl group flips during the conformational change to the holo form, and the hydration site gradually loses interaction with the group and disappears after the 17 ns MTMD window into bulk solvent. As this hydration site is not predicted in the holo form of the protein, it will not be taken into account in the estimation of the free energy of binding of a ligand, if only the holo protein conformation would be considered. The hydration site, however, actually transits into the bulk solvent during the conformational change induced by the binding ligand. Therefore, its de-solvation free energy needs to be considered as if being replaced by the binding ligand, i.e. direct replacement of a water molecule in the holo structure or due to induced conformational change equally contributes to the de-solvation free energy of the protein upon ligand binding.

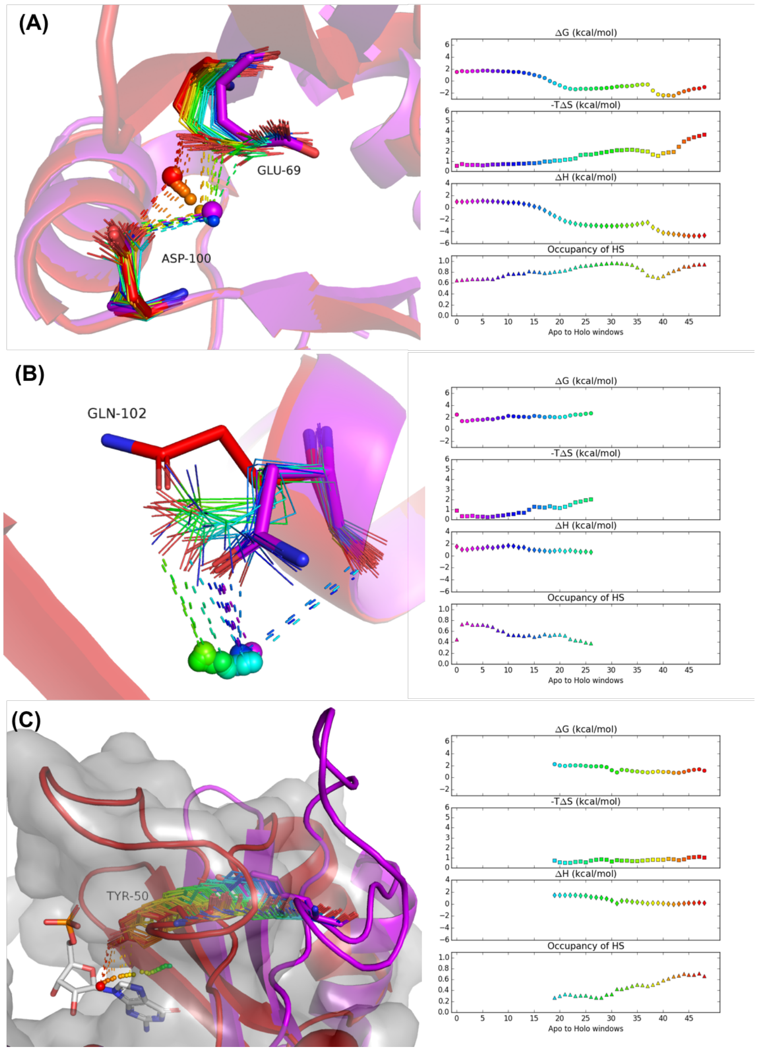


Fig. 3.8. Examples of hydration site transition pathways for the GUA system. (A) type (i) hydration sites that remain in the binding site throughout the conformational transition; (B) type (ii) hydration sites that disappear throughout transition from apo to holo form; (C) type (iii) hydration sites that appear throughout the conformational change. The graphs on the right side show the thermodynamic properties (ΔG, $-T\Delta S$, and ΔH) (in kcal/mol) and occupancy of the hydration site throughout the transition (window 0 represents the apo protein structure, and window 48 the holo protein structure).

Type (iii) hydration sites appear during the conformation change to the holo structure (Figure 3.8 C). The hydration site is not observed in the protein apo structure as the binding site residue TYR-50 on the flexible loop region is pointing to the bulk solvent. For clarity, we only show the surface of the holo conformation with the bound ligand in the active site. As the protein changes its conformation to the holo form, TYR-50 moves towards the active site, and a hydration site appears forming hydrogen bonds with TYR-50. If this hydration site is displaced by a ligand and released into the bulk solvent, it will not contribute to the free energy of ligand binding since it originates from bulk solvent in the apo protein conformation and ends in the bulk solvent again upon ligand binding. If it is, however, not replaced by a bound ligand, it will contribute to the free energy of binding, as the interaction energy and entropy of the water molecule in the binding site (in the holo state) may differ from the corresponding free energy contributions in bulk solvent (in the apo state).

3.3.3 Method II: Associate Hydration Sites in Apo and Holo Forms using HSSCS

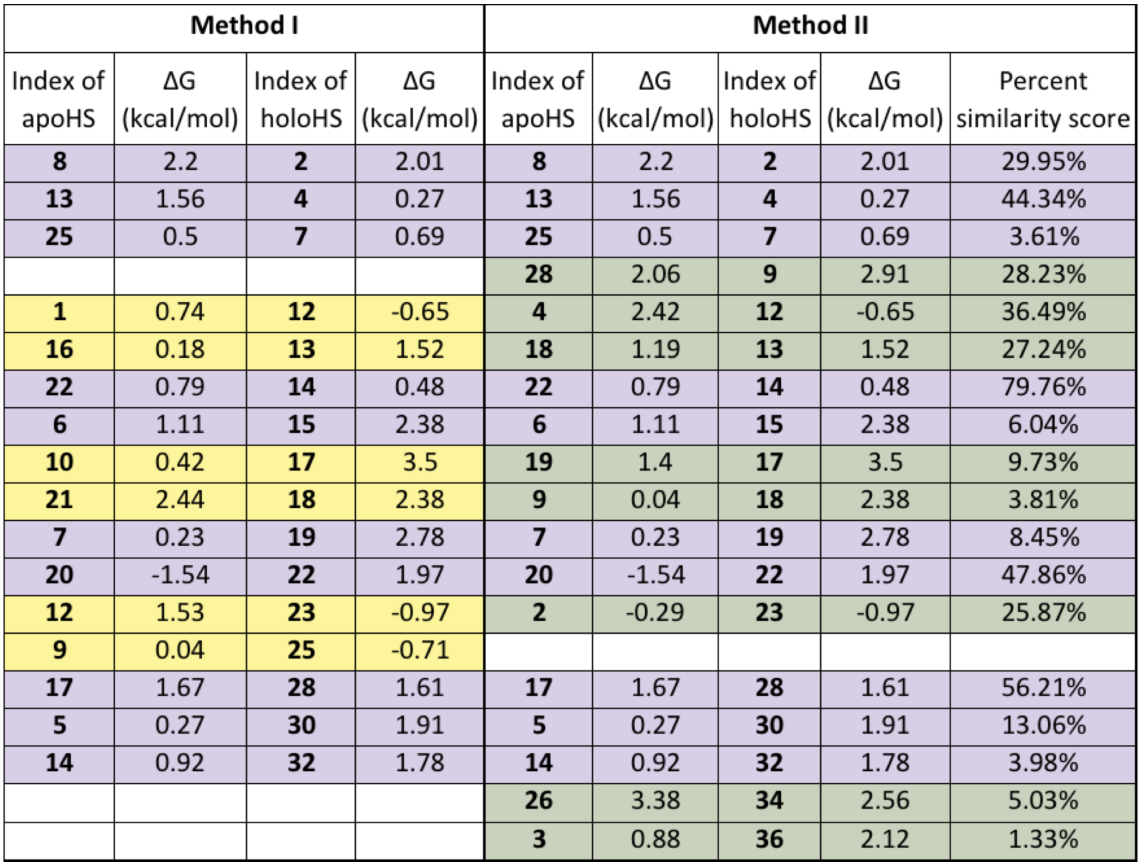
The protein de-solvation free energy obtained from the explicit hydration site information has been successfully used to replace the implicit de-solvation term in MMGBSA or in the scoring of docking poses [47, 69–71]. As demonstrated in the previous section, For protein systems with significant conformational change upon ligand binding, changes in hydration site location and thermodynamic profile need to be considered for a more accurate calculation of protein de-solvation free energies. Using MTMD simulations in method I, provides a continuous path between apo and holo hydration sites allowing to define clear associations between hydration sites in the two protein states. The method, however, is too computationally expensive to apply to a large ligand library, if different protein conformations are induced or stabilized by those bound ligands. To accelerate the process of predicting the de-solvation free energies of the protein upon ligand binding, a second method is proposed that requires only the hydration site identification and profiling for the end points (apo and holo form). To associate the hydration sites in the holo protein conformation to the corresponding hydration sites in the apo form, the definition of hydration site specific coordinate systems (HSSCS) for each hydration site is introduced. Using the HSSCS, locations and thermodynamic profiles of hydration sites between the apo and the holo form are compared and associated based on their similarity in interactions with nearby residues. The highest similarity score for each hydration site is used to identify a hydration-site pair.

3.3.4 Comparison Between Method I and Method II

A comparison of both methods for identifying associated hydration sites is shown in Table 3.3 for the GUA system. There are 16 type (i) hydration-site pairs identified using method I, and 18 pairs identified using method II. The majority of hydration sites identified by method I (10 from 16) is reproduced by method II (highlighted in light purple). Hydration-site pairs that are predicted differently by the two methods are highlighted in yellow (method I) and green (method II). The free energy differences of predicted hydration site de-solvation between the apo and holo protein conformations ranges from 0.1 to 3.5 kcal/mol. The percent similarity score is normalized to a range 0% to 100% and shown in Table 3.3.

Method II identifies hydration-site pairs for three additional holoHS (#9, #34, and #36) between apo and holo structure that are missed by method I. Analysis of those additional paired hydration sites highlights an issue encountered in method I. For example, a hydration-site pair (#28(apoHS)–#9(holoHS)) predicted only by method II is shown in Figure 3.9. Hydration site #28 in the apo form and #9 in the holo form are stabilized by a hydrogen bond to SER-34. The hydration sites are paired by method II using, for example, the local HSSCS defined by the three nearby residues SER-34, TYR-50, and TYR-78 (Figure 3.9 A). However, in method I, the hydration site disappears at window 12, and does not reappear before window

Table 3.3.
Comparison of matched hydration sites between method I and II for the GUA system.



Method I				Method II				
Index of apoHS	ΔG (kcal/mol)	Index of holoHS	ΔG (kcal/mol)	Index of apoHS	ΔG (kcal/mol)	Index of holoHS	ΔG (kcal/mol)	Percent similarity score
8	2.2	**2**	2.01	**8**	2.2	**2**	2.01	29.95%
13	1.56	**4**	0.27	**13**	1.56	**4**	0.27	44.34%
25	0.5	**7**	0.69	**25**	0.5	**7**	0.69	3.61%
				28	2.06	**9**	2.91	28.23%
1	0.74	**12**	-0.65	**4**	2.42	**12**	-0.65	36.49%
16	0.18	**13**	1.52	**18**	1.19	**13**	1.52	27.24%
22	0.79	**14**	0.48	**22**	0.79	**14**	0.48	79.76%
6	1.11	**15**	2.38	**6**	1.11	**15**	2.38	6.04%
10	0.42	**17**	3.5	**19**	1.4	**17**	3.5	9.73%
21	2.44	**18**	2.38	**9**	0.04	**18**	2.38	3.81%
7	0.23	**19**	2.78	**7**	0.23	**19**	2.78	8.45%
20	-1.54	**22**	1.97	**20**	-1.54	**22**	1.97	47.86%
12	1.53	**23**	-0.97	**2**	-0.29	**23**	-0.97	25.87%
9	0.04	**25**	-0.71					
17	1.67	**28**	1.61	**17**	1.67	**28**	1.61	56.21%
5	0.27	**30**	1.91	**5**	0.27	**30**	1.91	13.06%
14	0.92	**32**	1.78	**14**	0.92	**32**	1.78	3.98%
				26	3.38	**34**	2.56	5.03%
				3	0.88	**36**	2.12	1.33%

17 (Figure 3.9 B). The reason for this temporary disappearance is the decreasing water density that does not meet the criteria of defining a hydration site (Figure 3.9 C & D). Just before and after the hydration site's disappearance and re-appearance the occupancy of this hydration site is only about 20%. It should be noted that WATsite hydration site identification is based on the water density on grid points and the clustering method. For areas in the binding site with less pronounced water density peaks, e.g. more mobile water molecules, the definition of the hydration sites is sensitive to the clustering algorithm. The sensitivity to the water density, clustering algorithm, and the number of nearby MTMD windows of hydration-site pair identification is the main reason why hydration site paths become discontinuous throughout the MTMD trajectory.

3.3.5 Method IIb: multiple-to-multiple hydration-site pairings

We also observed that all the hydration-site pairs that differ in the two methods are located in proximity to charged residues such as GLU or ASP. Figure 3.10 A & B show an example of hydration sites directly interacting with ASP-100 and GLU-69. There are nine hydration sites that are in close proximity with these two residues in the apo form while only five hydration sites are observed in the holo form. As shown in Figure 3.10 C & E, less pronounced water density peaks are observed in the apoHS around GLU-69 and ASP-100 compared to the density of the comparable holoHS (Figure 3.10 D & F). The reason for this discrepancy is the flexibility of the solvent-exposed carboxylate group of a GLU or ASP residue; a slight rotation can result in different water density especially at solvent-exposed locations. Analysis of the 4 ns MD trajectory of apo protein conformation demonstrates the significant mobility of the water molecules in this region at the interface to bulk water, which results in the rather broad distribution of water density with weakly pronounced peaks. Due to the sensitivity of the clustering algorithm in those situations, five hydration sites are identified, although only 4-4 water molecules interact with residue GLU-69 at

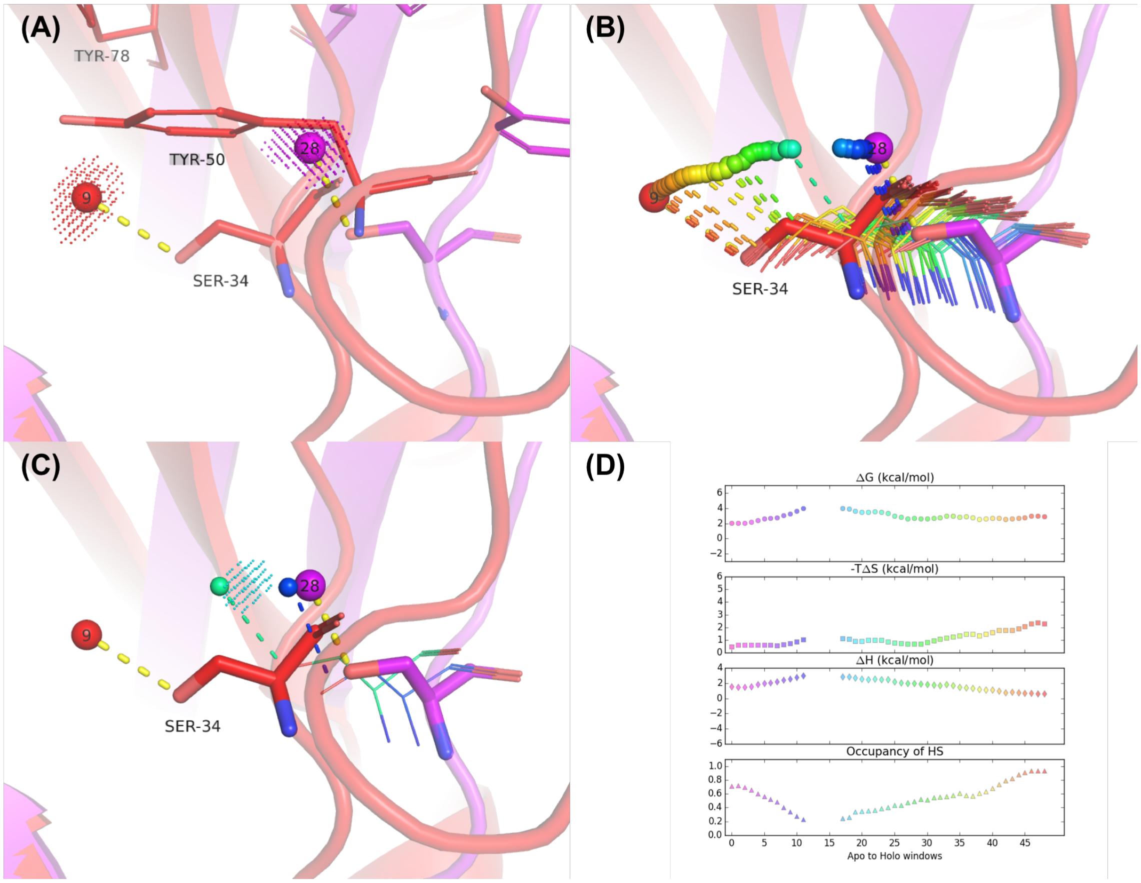


Fig. 3.9. Example of additional hydration site pair identified using method II which disappeared at intermediate windows using method I. (A) Hydration site #28 in the apo form (purple), and #9 in the holo form (red) are interacting with SER-34 via hydrogen bonding. (B) Transition trajectories of hydration site #28 in the apo form and #9 in the holo form using method I highlights the missing hydration site throughout the transition. (C) Hydration sites for window 11 and 16 are shown together with the water density at window 14 at which no hydration site was identified. (D) Overlay of thermodynamic properties and occupancy of hydration site #28 and #9.

the same time. This also explains why we observe different hydration-site pairs. For example, both #2 and #12 interact with the residue ASP-100 in the apoHS (Figure 3.10 A), whereas, #23 is observed to interact with the same oxygen atom in the holo form (Figure 3.10 B). Hydration-site pair #12–#23 is identified in method I, while method II identified #2–#23 as a pair. Similarly, different hydration-site pairs are identified around GLU-69 by both methods (#1–#12 in method I; #4–#12 in method II). Both associations are meaningful because both #1 and #4 in the apoHS interact with GLU-69, and #12 interacts with the same residue in the holoHS.

The above observation, that multiple hydration site in the apoHS can be paired with one site in the holoHS, led to the conclusion that one-to-one hydration-site pairing is not ideal. Therefore, we adjusted method II by allowing one hydration site to be paired with multiple sites instead of only using the pairing with the highest similarity score. Table 3.4 shows all hydration-site pairs with a similarity score larger than 1%. All hydration-site pairs identified by method I are now predicted by method II as well. HoloHS that are paired differently by method I and the original method II are circled in red (method I) and green (original method II; highest similarity score). It should be noted that the percentage similarity scores are significantly smaller that 100%. This is not surprising, as it is not possible that a hydration-site is within 4 Å of all eight residues used to define the HSSCS. Furthermore, it is unlikely that two hydration sites are paired in all 672 HSSCS, in particular if the protein changes conformation (cf. Figure 3.4 A).

The results of comparing both methods for the SH2 system are shown in Table 3.5 & 3.6. All hydration-site pairs that are predicted by method I but not by original method II (Table 3.5) are predicted when allowing multiple pairings for one hydration site in method IIb. These multiple-to-multiple hydration-site pairings identified here make more sense from a physical perspective since several identified hydration sites may actually be the result of the same water molecule interacting with rotated side chain of the protein residue. Thus, the sensitivity of hydration site identification on water density and clustering algorithm may be reduced by method IIb.

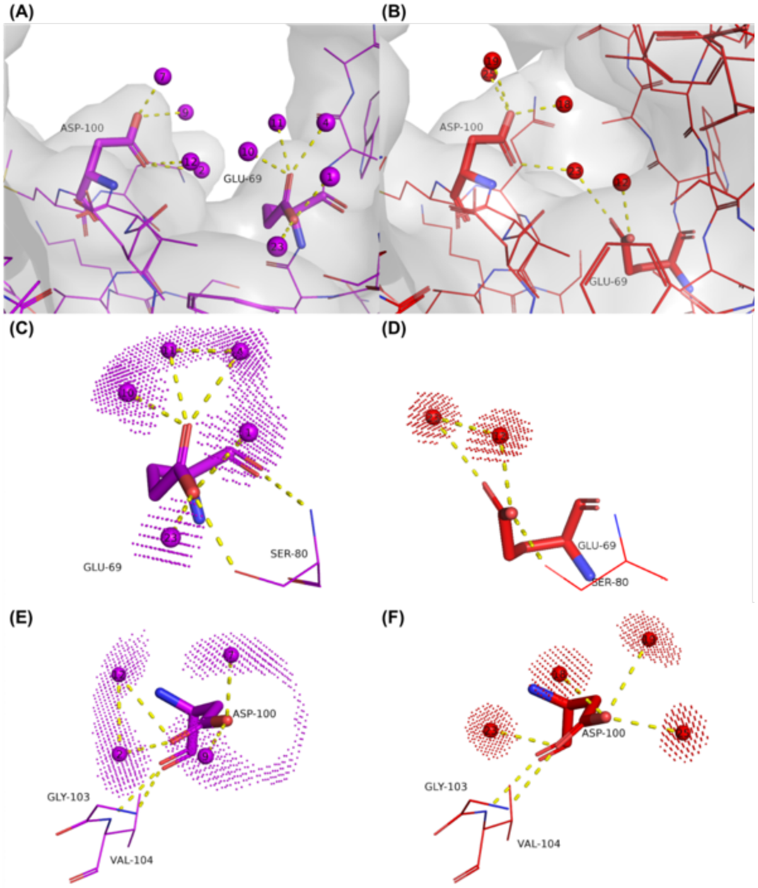
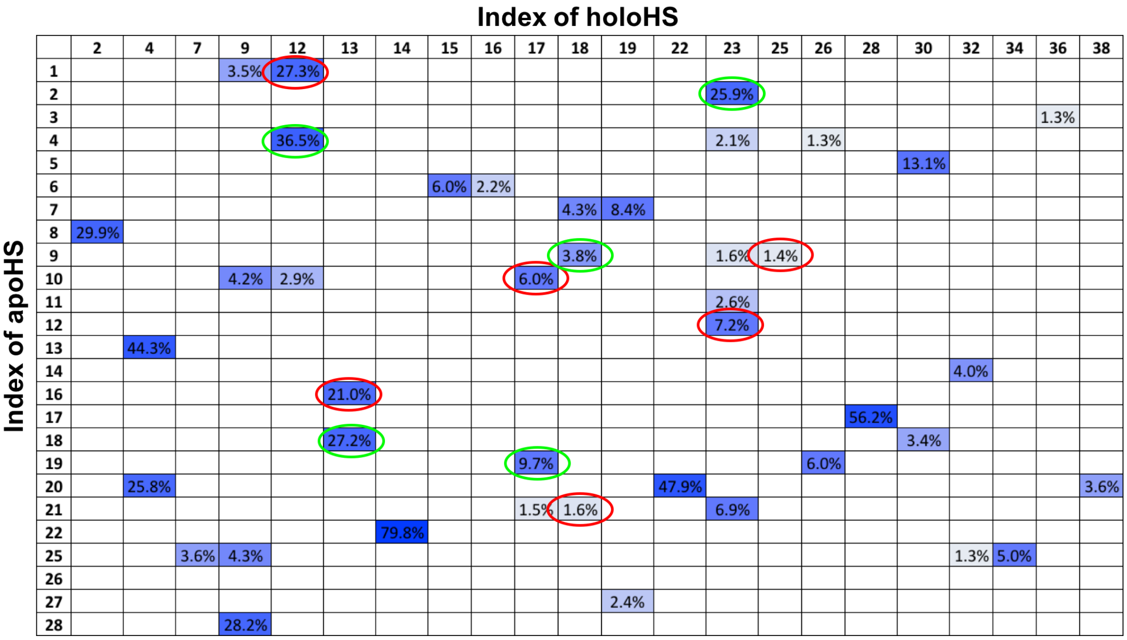


Fig. 3.10. Example of different hydration-site pairs identified from method I and II. More hydration sites interact with GLU-69 and SP-100 in solvent-exposed apo form (A) than in the holo form(B). (C)-(F) Water density of each hydration site around GLU-69 or ASP-100. Less pronounced water density is observed in the protein apo form(C,E) than in the holo form(D,F).

Table 3.4.
Heat map showing all hydration-site pairs identified using method IIb allowing for multiple pairing with the same hydration site.



Index of holoHS

Index of apoHS

	2	4	7	9	12	13	14	15	16	17	18	19	22	23	25	26	28	30	32	34	36	38
1				3.5%	27.3%																	
2														25.9%								
3																					1.3%	
4					36.5%									2.1%		1.3%						
5																		13.1%				
6								6.0%	2.2%													
7											4.3%	8.4%										
8	29.9%																					
9											3.8%			1.6%	1.4%							
10				4.2%	2.9%					6.0%												
11														2.6%								
12														7.2%								
13		44.3%																				
14																			4.0%			
16						21.0%																
17																	56.2%					
18						27.2%												3.4%				
19										9.7%						6.0%						
20		25.8%											47.9%									3.6%
21										1.5%	1.6%			6.9%								
22							79.8%															
25			3.6%	4.3%															1.3%	5.0%		
26																						
27												2.4%										
28				28.2%																		

*A percent similarity score of each hydration-site pair is shown in the table and color coded from white (0%) to blue (100%). Hydration-site pairs that identified differently by method I and II are circled in red (method I) and green (original method II).

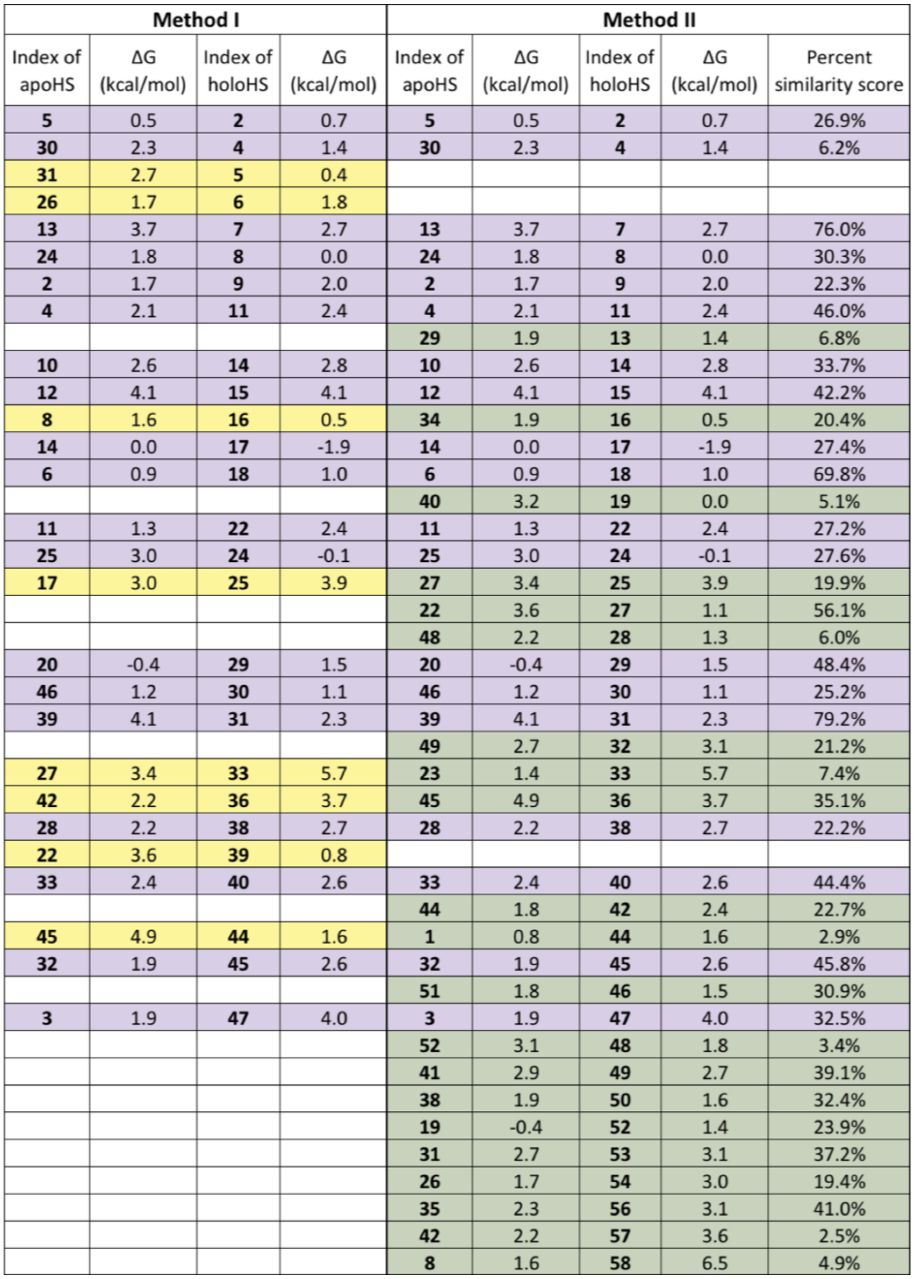
The analysis also highlighted that method I fails to identify many hydration-site pairs that are identified by method IIb. In method I, we pair hydration sites from the apo to the holo structure with the assumption of a one-to-one correspondence. Furthermore, method I is neglecting the temporary disappearance of hydration site due to changes in water density during the conformational transition. In this context, it should also be noted that simulating the conformational change using MTMD may result in intermediate protein conformations that are unrealistic and thus unfavorable for stable water positions (Figure 3.9).

3.3.6 Prediction of Protein De-solvation Free Energy including Protein Flexibility

As no experimental data is available on water energies, a direct validation of the above prediction is not possible. Thus, we tried to correlate the predicted protein de-solvation free energy with binding affinity even if the other components, such as the protein conformational energy and protein ligand interaction energy, are missing.

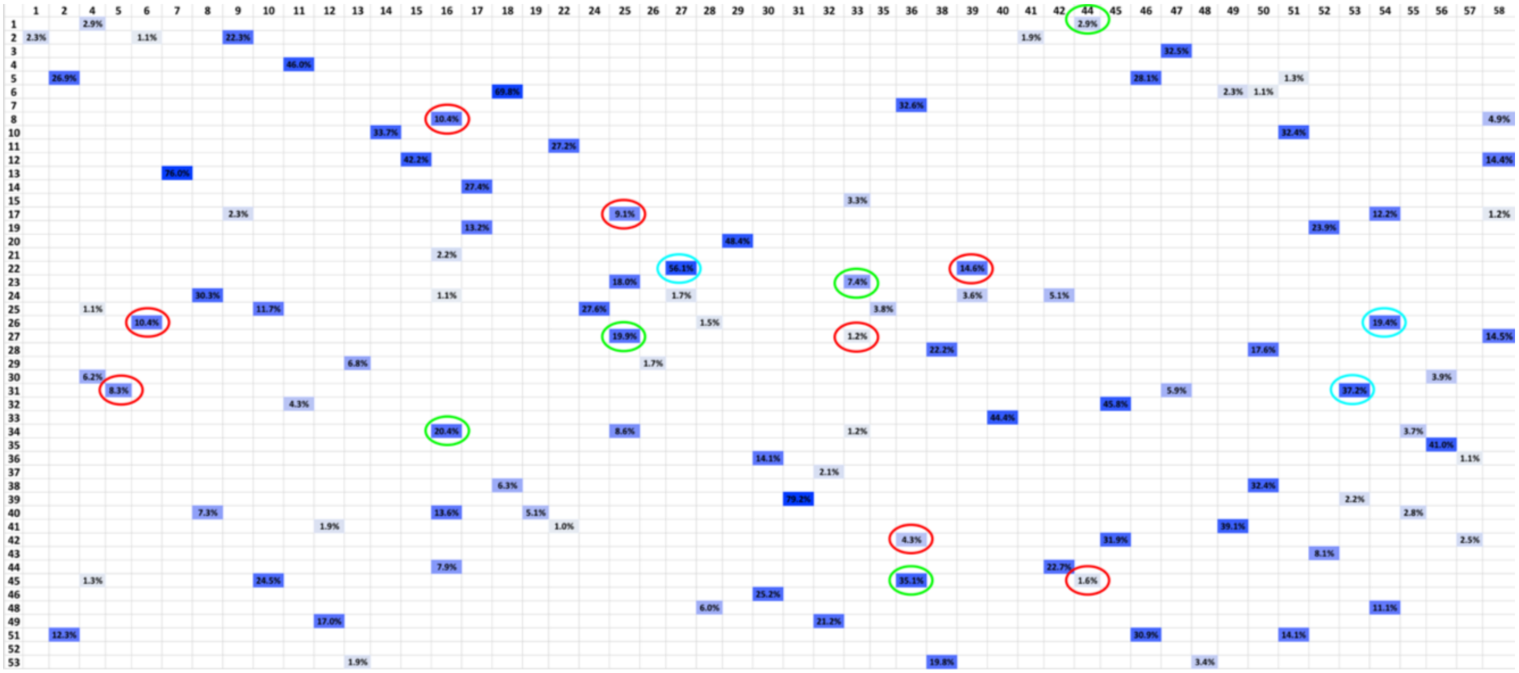
The de-solvation free energies of the protein active site upon binding of several ligands to the SH2 domain of (Pp60) Src protein system. The apoHS and holoHS, and their pairings, were used to estimate the protein de-solvation free energy. Ligands from different crystal structures (PDBID: 1O4A, 1O4B, 1O42, 1O43, 1O44, 1O45, 1O46, 1O47, 1O48, 1O49) were placed back into the active site of the protein holo conformation, and the de-solvation free energies of different ligands were estimated using the standard rigid-protein method as well as flexible-protein method as presented in the current chapter. In short, ΔG_{desolv} using the rigid protein method is estimated by adding the de-solvation free energies of the holoHS that are displaced by the bound ligand (within the cutoff distance of 2.24 Å to any heavy atom of the ligand). The flexible protein method utilized all hydration-site pairs information from method IIb, and the contribution from three types of hydration sites were computed

Table 3.5.
Comparison of matched hydration sites between method I and II for the SH2 system.



Method I				Method II				
Index of apoHS	ΔG (kcal/mol)	Index of holoHS	ΔG (kcal/mol)	Index of apoHS	ΔG (kcal/mol)	Index of holoHS	ΔG (kcal/mol)	Percent similarity score
5	0.5	2	0.7	5	0.5	2	0.7	26.9%
30	2.3	4	1.4	30	2.3	4	1.4	6.2%
31	2.7	5	0.4					
26	1.7	6	1.8					
13	3.7	7	2.7	13	3.7	7	2.7	76.0%
24	1.8	8	0.0	24	1.8	8	0.0	30.3%
2	1.7	9	2.0	2	1.7	9	2.0	22.3%
4	2.1	11	2.4	4	2.1	11	2.4	46.0%
				29	1.9	13	1.4	6.8%
10	2.6	14	2.8	10	2.6	14	2.8	33.7%
12	4.1	15	4.1	12	4.1	15	4.1	42.2%
8	1.6	16	0.5	34	1.9	16	0.5	20.4%
14	0.0	17	-1.9	14	0.0	17	-1.9	27.4%
6	0.9	18	1.0	6	0.9	18	1.0	69.8%
				40	3.2	19	0.0	5.1%
11	1.3	22	2.4	11	1.3	22	2.4	27.2%
25	3.0	24	-0.1	25	3.0	24	-0.1	27.6%
17	3.0	25	3.9	27	3.4	25	3.9	19.9%
				22	3.6	27	1.1	56.1%
				48	2.2	28	1.3	6.0%
20	-0.4	29	1.5	20	-0.4	29	1.5	48.4%
46	1.2	30	1.1	46	1.2	30	1.1	25.2%
39	4.1	31	2.3	39	4.1	31	2.3	79.2%
				49	2.7	32	3.1	21.2%
27	3.4	33	5.7	23	1.4	33	5.7	7.4%
42	2.2	36	3.7	45	4.9	36	3.7	35.1%
28	2.2	38	2.7	28	2.2	38	2.7	22.2%
22	3.6	39	0.8					
33	2.4	40	2.6	33	2.4	40	2.6	44.4%
				44	1.8	42	2.4	22.7%
45	4.9	44	1.6	1	0.8	44	1.6	2.9%
32	1.9	45	2.6	32	1.9	45	2.6	45.8%
				51	1.8	46	1.5	30.9%
3	1.9	47	4.0	3	1.9	47	4.0	32.5%
				52	3.1	48	1.8	3.4%
				41	2.9	49	2.7	39.1%
				38	1.9	50	1.6	32.4%
				19	-0.4	52	1.4	23.9%
				31	2.7	53	3.1	37.2%
				26	1.7	54	3.0	19.4%
				35	2.3	56	3.1	41.0%
				42	2.2	57	3.6	2.5%
				8	1.6	58	6.5	4.9%

Table 3.6.
Heat map showing all hydration-site pairs identified using method IIb allowing for multiple pairing with the same hydration site (SH2 system).



	1	2	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	22	24	25	26	27	28	29	30	31	32	33	35	36	38	39	40	41	42	44	45	46	47	48	49	50	51	52	53	54	55	56	57	58
1			2.9%																																		2.9%														
2	2.3%				1.1%			22.3%																											1.9%																
3																																								32.5%											
4										46.0%																																									
5		26.9%																																					28.1%					1.3%							
6																	69.8%																									2.3%	1.1%								
7																															32.6%																				
8															10.4%																																				4.9%
10													33.7%																															32.4%							
11																			27.2%																																
12														42.2%																																					14.4%
13						76.0%																																													
14																27.4%																																			
15																													3.3%																						
17								2.3%													9.1%																										12.2%				1.2%
19																13.2%																													23.9%						
20																									48.4%																										
21															2.2%																																				
22																							56.1%										14.6%																		
23																					18.0%								7.4%																						
24							30.3%								1.1%								1.7%										3.6%			5.1%															
25			1.1%						11.7%											27.6%										3.8%																					
26					10.4%																			1.5%																							19.4%				
27																					19.9%								1.2%																						14.5%
28																																22.2%											17.6%								
29												6.8%										1.7%																													
30			6.2%																																														3.9%		
31				8.3%																																				5.9%						37.2%					
32										4.3%																												45.8%													
33																																		44.4%																	
34															20.4%						8.6%								1.2%																			3.7%			
35																																																	41.0%		
36																										14.1%																								1.1%	
37																												2.1%																							
38																	6.3%																										32.4%								
39																											79.2%																			2.2%					
40							7.3%								13.6%			5.1%																														2.8%			
41											1.9%								1.0%																							39.1%									
42																															4.3%							31.9%												2.5%	
43																																													8.1%						
44															7.9%																					22.7%															
45			1.3%						24.5%																						35.1%						1.6%														
46																										25.2%																									
48																								6.0%																							11.1%				
49											17.0%																	21.2%																							
51		12.3%																																					30.9%					14.1%							
52																																																			
53												1.9%																				19.8%									3.4%										

*A percent similarity score of each hydration-site pair is shown in the table and color coded from white (0%) to blue (100%).

Hydration-site pairs that identified differently by method I and II are circled in red (method I) and green (original method II).

using equations 3.3 and 3.4. The predicted ΔG_{desolv} are then correlated with the experimentally measured binding affinities (pIC$_{50}$ values).

In addition to the protein de-solvation free energy, we also computed the correlation between pIC$_{50}$ versus buried apolar surface area which is frequently used as an estimate for dehydration costs (Figure 3.11 A). A favorable correlation was observed, although the correlation coefficient is lower in absolute magnitude compared to using explicit hydration sites and protein flexibility in method II. A more negative ΔG_{desolv} means a more favorable contribution to the free energy of binding, and therefore should be associated with a higher pIC$_{50}$ value. Whereas ΔG_{desolv} estimated from the flexible-protein method shows a meaningful correlation with the experimental affinity data, the predicted de-solvation free energies using the rigid-protein method actually displays an inverse correlation to the experimental affinity (Figure 3.11).

Clearly, the protein de-solvation free energy is only one but important contribution to the free energy of binding, and future studies will aim to improve the predictive capacities of the flexible protein method by combining it with other energy terms such as the molecular mechanics direct protein-ligand interaction energy and terms from the generalized Born surface area (MM-GB/SA) method. In the current study, we just wanted to demonstrate the importance of including protein flexibility in hydration site prediction for the more accurate estimation of protein de-solvation and its contribution to the free energy of binding.

3.4 Conclusion

In this chapter, we presented the development of two methods to incorporate protein flexibility in the prediction of hydration site location and thermodynamic profile. Method I requires time-consuming MTMD simulations, but provides a detailed picture on the hydration site changes during the apo to holo transition. This method is useful for obtaining a detailed understanding of the changes in water density and thermodynamic properties of localized water molecules during the conformational change.

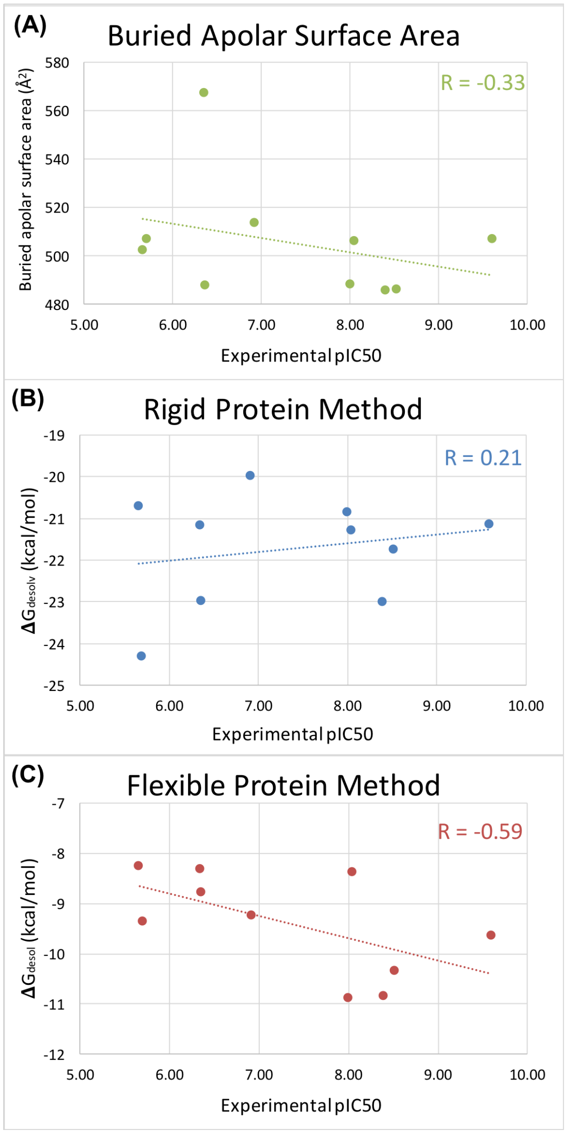


Fig. 3.11. The correlation between buried apolar surface area and predicted protein de-solvation free energies (ΔG_{desolv}) versus the experimentally measured pIC$_{50}$ values. (A) The buried apolar surface area is estimated using PyMol as described in the method section. (B) ΔG_{desolv} using the rigid protein method is estimated with the hydration sites predicted using the protein holo form. (C) ΔG_{desolv} from the flexible protein method is estimated using all identified hydration-site pairs between apoHS and holoHS using the equations 3.3 and 3.4.

It should, however, be noted that the identification of hydration-site transition pathways using method I is sensitive to the clustering algorithm and to changes in water density, so that temporarily low occupancy regions during the conformational transition may lead to the disappearance of hydration sites and therefore a loss in transition pathways. Method II is computationally more efficient compared to method I. It can be easily applied to a large library of compounds that bind to an ensemble of different holo structures of the same protein, and thus may be useful for the calculation of protein de-solvation free energy in context of docking or post-processing methods such as MM-GB/SA.

Our study also highlights the large difference in the predicted de-solvation free energy of the same hydration site between the apo and the holo protein conformation, which can be as large as 3.5 kcal/mol. Thus, using the hydration site information without inclusion of protein flexibility may lead to the wrong estimation of protein de-solvation free energy. With the methods presented in this chapter, we are able to incorporate protein flexibility into the estimation of the de-solvation free energy. The implicit protein de-solvation term used in the MM-GB/SA method can be replaced by the estimated de-solvation free energy using hydration site replacement. Guimarães et.al. have demonstrated an improved correlation between the MM-GB/SA results and experimental data when replacing the implicit de-solvation term with explicit hydration site free energies. Thus, a potential new strategy which we will investigate in future studies is to first associate the hydration sites in the apo and holo forms using our new flexible-protein method, estimate the de-solvation free energy for each ligand in the protein system, and use this term in the context of MM-GB/SA replacing the implicit protein de-solvation term.

4. GPU-ACCELERATION OF HYDRATION SITE PREDICTION PROGRAM WATSITE

4.1 Introduction

We have shown in Chapter 4 that the incorporation of protein flexibility improves the accuracy of predicting the protein de-solvation free energy. Different ligands binding to the same target protein can induce different conformational adaptations. Thus, for applying hydration-site profiling for a large library of compounds binding to a variety of different holo structures of the same protein, WATsite needs to be applied to an ensemble of different protein conformations. With the current implementation of WATsite the MD-based hydration site profiling is rather time consuming. This motivated a more efficient implementation of WATsite for hydration site analysis and binding free energy prediction.

Graphical processing units (GPUs) have been designed with a large number (thousands) of simple processors that will work in parallel. Some of the calculation during a MD simulation such as computing nonbonded pair interactions between a large number of atoms can be accelerated with GPU [72]. Thus, most widely used MD packages such as OpenMM [48], Amber [73], CHARMM [74], NAMD [75], and GROMACS [76] have been adapted to take advantages of GPU architectures. We have implemented a GPU-accelerated version of WATsite within the OpenMM software library.

A description of OpenMM-WATsite, which includes an on-the-fly calculation of interaction energy of each water molecule with its surrounding throughout the MD simulation, will be provided in this chapter. System truncation, another method to speed up the MD simulations for hydration site prediction, has been implemented in OpenMM-WATsite and will be discussed at the end of this chapter.

4.2 Materials and Methods

4.2.1 Implementation of Water Interaction-Energy Computation in OpenMM

In order to speed up hydration-site profiling using WATsite, calculation and report of water interaction-energy with its surrounding was implemented in OpenMM using CUDA for GPU architecture. For each water molecule in the system, the interaction energy between this water molecule and all other protein atoms, water molecules and potentially ligand atoms was calculated. The interaction energy consists of short range Lennard-Jones interactions and the short range (or direct) part of the electrostatic interactions calculated with the particle-mesh Ewald (PME) method.

$$E = \frac{1}{2} \quad \epsilon\left[\Bigg((\frac{\sigma}{r_{ij}})^{12} - (\frac{\sigma}{r_{ij}})^{6}\right]\Bigg(+ q_i q_j \frac{erfc(\alpha r_{ij})}{r_{ij}}\Bigg)\Bigg(\tag{4.1}$$

The long-range part of PME cannot be broken down into individual atomic interaction energy terms and was therefore neglected in the analysis. Water-interaction energies are subsequently used to calculate the de-solvation enthalpy in WATsite. Thus, for each saved MD snapshot, a text file with the above computed water interaction energies was generated asynchronously via std::async. The asynchronous output processing permits the MD calculation to continue without waiting for the write-out of the water energies. This makes the data out-write non-blocking and saves considerable amount of time.

4.2.2 Preparation of Protein Systems

HIV-1 protease (PDB: 1PHX) has been used in our previous study with recorded computation time, thus it was also used in this study for to compare with the GPU-acceleration. The ligand was removed from the HIV-1 protease binding site in order to predict the protein de-solvation free energy of bound ligands. Protein structures were first prepared using Schrodinger's Protein Preparation Wizard [77]. In short, hydrogen positions, bond orders, and protonation states of histidines, glutamic and aspartic acids, and conformational flips of asparagine and glutamine side chains were optimized using the default protocol.

Additionally for testing the speed of the new implementation, the dihydrofolate reductase (DHFR) system was used. This system allows for direct comparison with other MD programs as it used as standard benchmark for Amber, Charmm, OpenMM, etc.

4.2.3 System Truncation

For further speed-up of the MD simulation protocol, the protein was truncated at different cutoffs (12, 15, 17, 20, 25, and 30 Å) from the center of the binding site. This truncation is motivated by the fact that the protein is restraint in all WATsite simulations to achieve convergence in water occupancy and free-energy profiling. We analyzed the impact of such truncations by comparing the WATsite predictions to those of the full protein system. For all truncated systems, a python script is used to add capping acetyl (ACE) and amide (NME) groups to the break points in the protein sequence. The protein structures were then solvated in an orthorhombic box of water molecules. A minimum distance of 10 Å between any protein atom and the faces of the box is chosen.

4.2.4 MD Simulations

MD simulations were performed using the modified GPU-accelerated OpenMM-WATsite package with the AMBER14SB force field [49] and SPC/E water model [78, 79]. The SHAKE algorithm [50] was applied to constrain bonds including hydrogen atoms to their equilibrium lengths and maintain rigid water geometries. Long-range electrostatic interactions were treated with the Particle Mesh Ewald method [51] with a cutoff of 10 Å for the direct interactions. The Lennard-Jones interactions were truncated at a distance of 10 Å, and a long-range isotropic correction was applied to the pressure representing Lennard-Jones interactions beyond the cutoff. A Langevin integrator with a time step of 2 fs was used together with a stochastic thermostat collision frequency of 1 ps^{-1}. The pressure control was implemented via isotropic box edge adjusting by MC moves every 25 time steps simulating the effect of constant pressure.

The system is first energy minimized and then heated to 298 K over 50 ps of MD simulations, followed by 1 ns of equilibration MD simulations at 298 K and 1 bar with periodic boundary conditions in all three dimensions. During the minimization and equilibration process, all protein heavy atoms were harmonically restrained with a spring constant of 4.8 kcal mol^{-1} Å^{-2}.

4.2.5 Hydration Site Analysis

The theory of hydration site identification has been described in detail in Chapter 2.2.2 and Chapter 2.2.3.

4.2.6 Grid-based Water Analysis

A potential limitation of hydration site approaches is the sensitivity of the hydration site identification and profiling due to the clustering algorithms, in particular for diffused water-occupancy regions. Inspired by the grid inhomogeneous solvation theory (GIST), an alternative grid-based analysis was performed in this study.

A 3D grid with spacing of 0.5 Å was placed over the user-defined binding site. Following the same protocol as hydration site analysis, occupancy of water molecules is distributed onto the 3D grid with a Gaussian distribution function centered on each water's oxygen atom. In contrast to standard WATsite, the occupancy is not clustered into hydration sites, but every grid point with larger than twice the bulk occupancy is considered as a 'pseudo-hydration site' and any water molecule within 1 Å radius throughout the MD trajectory is considered to contribute to it. The desolvation enthalpy and entropy of the 'pseudo-hydration site' is calculated similarly as in the original hydration site analysis using the contribution of any water molecule within 1 Å radius throughout the MD trajectory.

4.2.7 Convergence

Grid energy

For grid-based water energies analysis, it is important to confirm that energy convergence was achieved throughout the MD simulation of a given duration. 100 ns MD simulations were performed for the production run for the full protein system, and snapshots were saved every picosecond in NetCDF format, generating 100,000 frames. To test the convergence of water-energy calculation, we perform hydration site and grid-based analysis for the first 1, 2, 3, 4, 5, 10, 20, and 50 ns from the 100 ns simulation.

Energy grids predicted from shorter simulations (1) were compared to that from 100 ns (2) by calculating the overlap coefficient (OC) (Equation 4.2).

$$OC = \sum_{i=1}^{N} \Bigg(\min \left(\frac{Q_i^1}{\sum_{j=1}^{N} Q_j^1};\ \frac{Q_i^2}{\sum_{j=1}^{N} Q_j^2} \right) \tag{4.2}$$

OC values range from 0 to 1 with the latter expressing full convergence/reproducibility between the two sets of energy calculations.

Hydration site energy

For hydration site predictions, the Pearson correlation coefficients R^2 between two sets of energy calculations to the regression line with slope = 1 and zero point = 0, i.e. y = x were then calculated. The geometric distances between paired hydration sites were color coded, ranging from red (=identical position) to blue (=1 Å distance).

4.3 Results

4.3.1 Validation of Hydration Site Prediction

For validation that the GPU-accelerated WATsite3.0 can reproduce the results of the previous WATsite2.0 version which was using GROMACS, we used the same MD trajectory as input for hydration sites prediction. Figure 4.1 shows the consistency between the two sets of hydration sites predictions. Both positions and predicted thermodynamic properties are almost identical. There is a slight difference in the predicted ΔH values, which might be explained by the different implementation of complementary function(*erfc*) in OpenMM and GROMACS.

4.3.2 Grid-based Analysis of Water Energy

As shown in Figure 4.2, hydration sites with most favorable (blue) and most unfavorable (red) free-energy values overlap with grid energies at high absolute free-energy cutoff (4.2, left). At lower absolute free-energy cutoff (4.2, right), energy grids reveal all predicted hydration sites within the range of free-energy cutoff. There is a clear correspondence from hydration sites to energy grids, whereas energy grids can also be found at positions where hydration site is not identified. Thus, grid-based analysis is in particular advantageous in regions with less pronounced, diffusive water-density peaks or irregularly shaped density distributions .

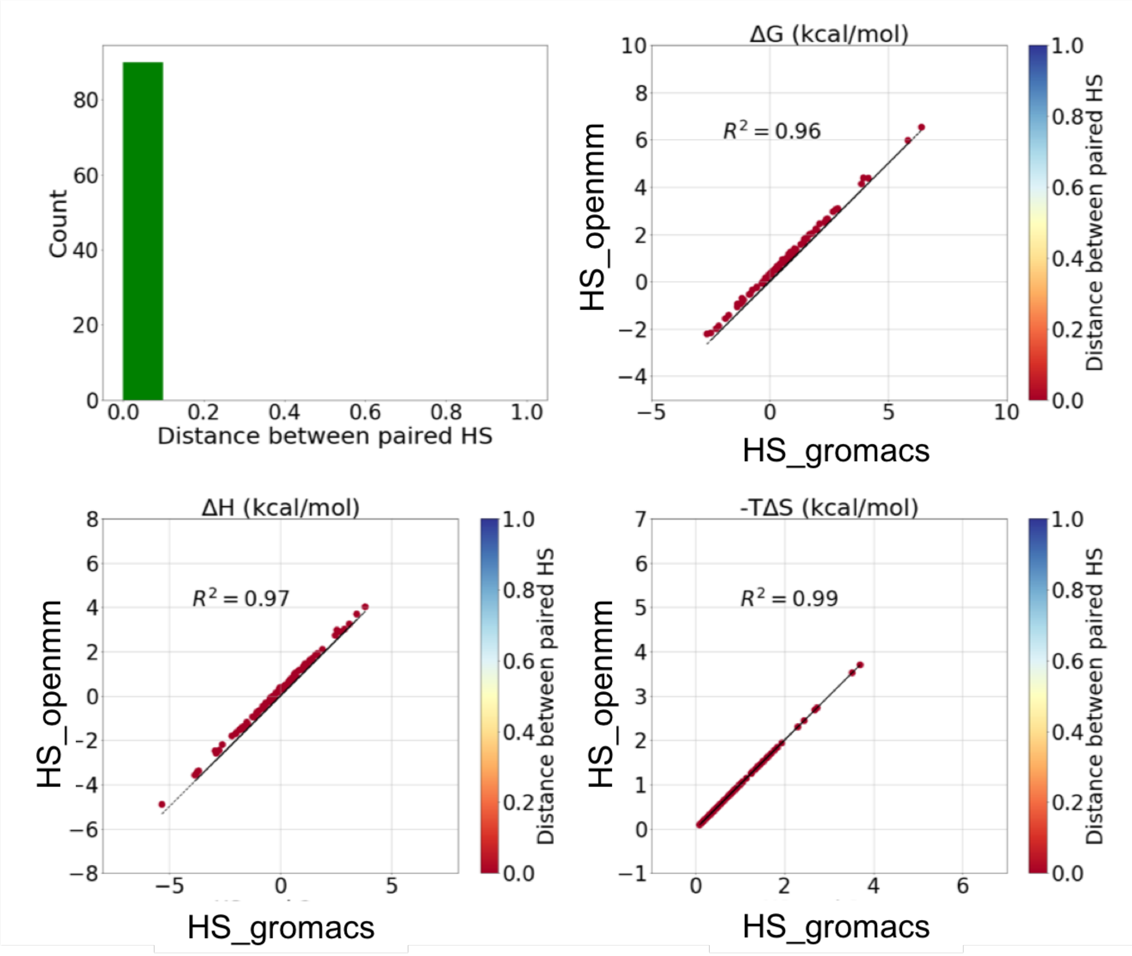


Fig. 4.1. Comparison of hydration sites prediction from WATsite2.0 based on GROMACS and WATsite3.0 based on OpenMM.

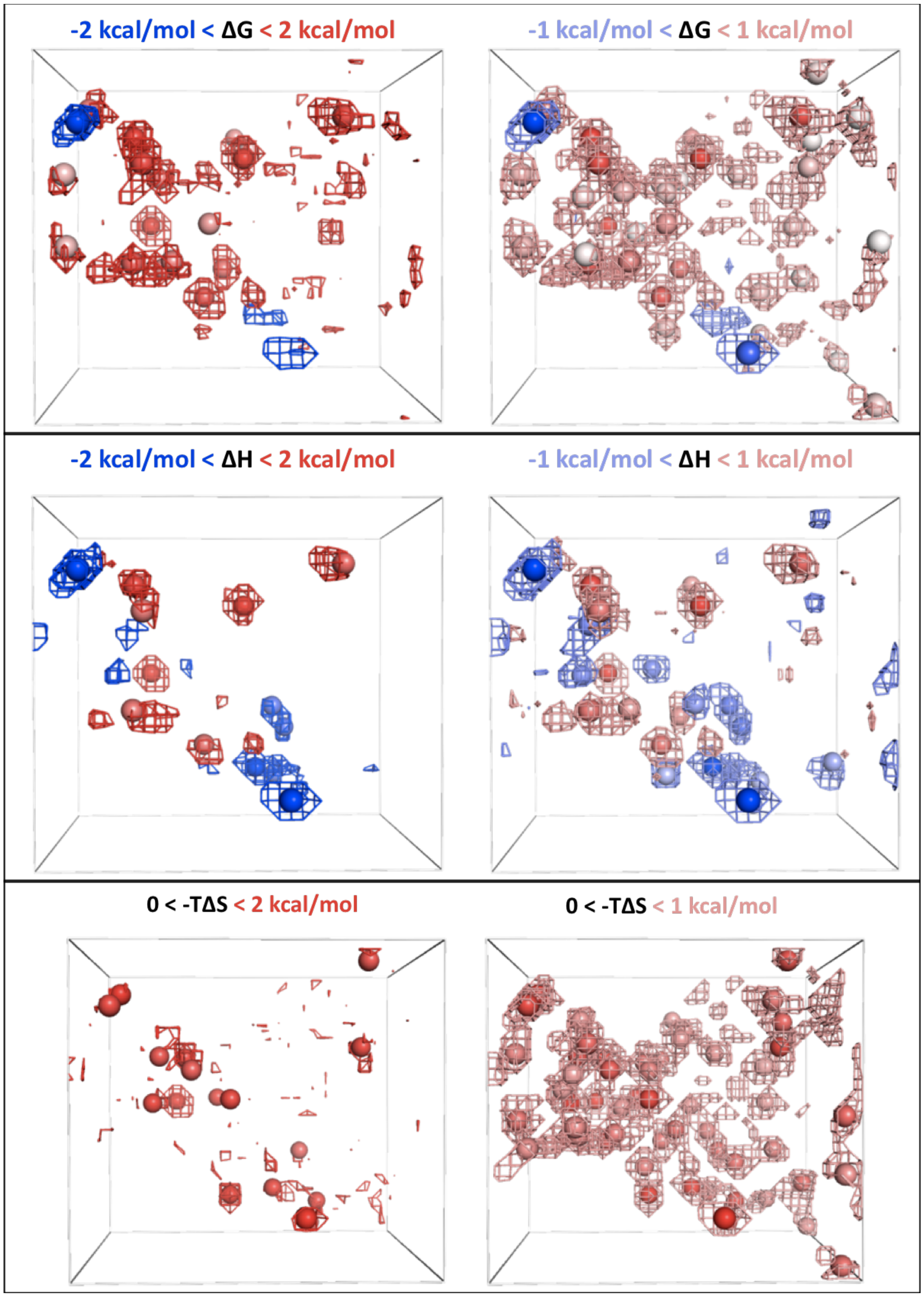

-2 kcal/mol < ΔG < 2 kcal/mol
-1 kcal/mol < ΔG < 1 kcal/mol
-2 kcal/mol < ΔH < 2 kcal/mol
-1 kcal/mol < ΔH < 1 kcal/mol
0 < -TΔS < 2 kcal/mol
0 < -TΔS < 1 kcal/mol

Fig. 4.2. Overlay of hydration sites and energy grids at two absolute energy cutoffs.

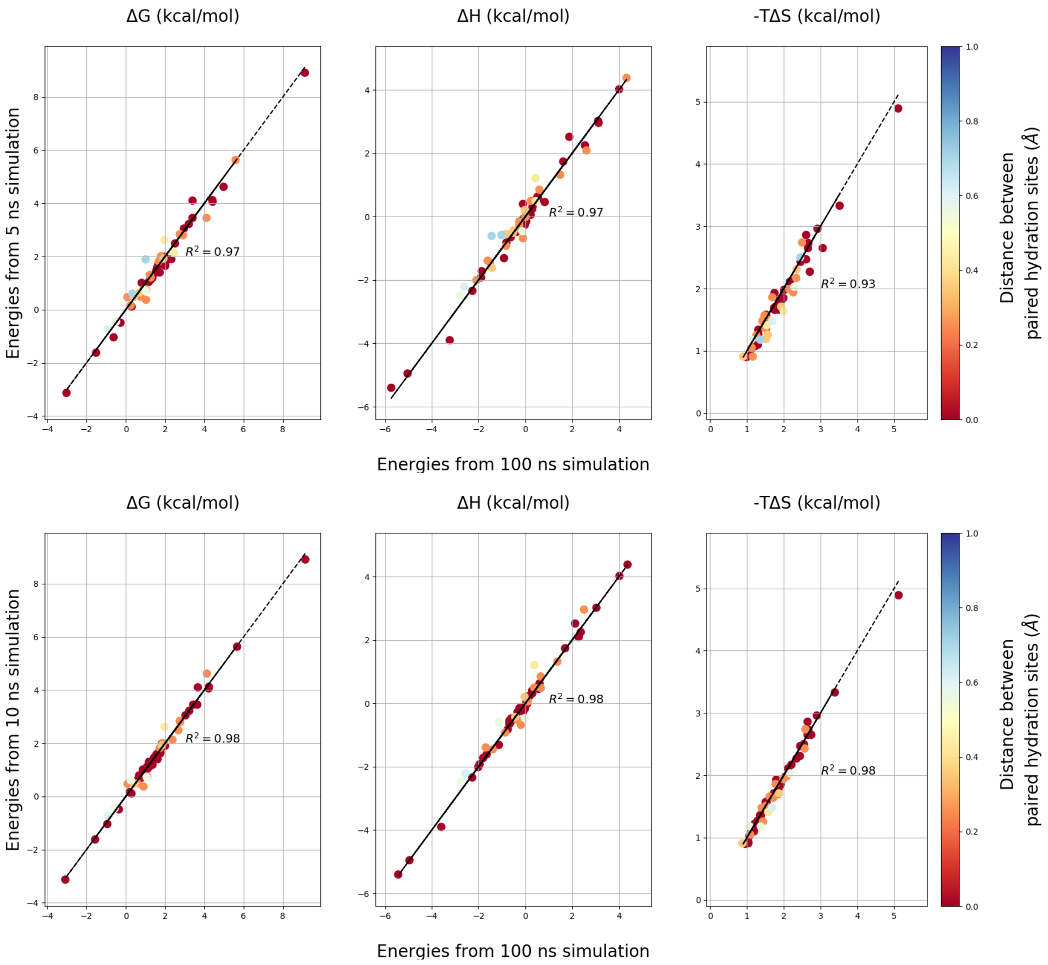


Fig. 4.3. Convergence of hydration site energies for HIV-1 protease at 5 ns and 10 ns comparing to 100 ns.

4.3.3 Energy Convergence of Hydration Site and Grid-based Analysis

We performed convergence analysis for the free energies, enthalpies and entropies of the predicted hydration sites obtained from the first 5 ns and 10 ns of the 100 ns MD simulation. The R^2 value of 0.97 for 5 ns versus 100 ns indicates that 5 ns is sufficient to generate converged thermodynamic profiles for hydration sites (Figure 4.3).

Checkpoint (ns)	Occupancy	ΔG	ΔH	-TΔS
1	0.92	0.65	0.78	0.70
2	0.93	0.55	0.80	0.66
3	0.95	0.62	0.85	0.54
4	0.95	0.68	0.85	0.72
5	0.95	0.74	0.87	0.80
10	0.97	0.86	0.92	0.89
20	0.98	0.91	0.94	0.93
50	0.99	0.97	0.98	0.97

Table 4.1.
Overlap coefficients (OC) for occupancy and energy grids of HIV-1 protease with a total of 100 ns simulation time.

The convergence study on the grid-based de-solvation free energy analysis revealed OC values for ΔG grids (when compared to the full 100 ns simulation results) increasing from 0.74 over 0.86 to 0.91 when extending the analysis from the first 5 ns over 10 ns to 20 ns of the MD simulation (Table 4.1). For reasonable convergence of grid energies, we suggest a simulation length of at least 10 ns is required, with 20 ns being preferable.

4.3.4 System Truncation

System truncations in MD simulations have a long history. In the early days of MD, such truncations were often necessary to make the calculation feasible. In order to further speed up WATsite prediction, we also tested the impact of system truncation on the accuracy and speed of grid-based and hydration site analysis.

Hydration site and grid-based analysis are performed using 20 ns MD simulations of systems truncated at different cutoff distances. For the grid-based analysis, the OC values of energy grids when compared to the full system are reported in Table 4.2. With OC value of 0.91, 0.91, 0.93, truncation at 12 Å seems to be able to reproduce the results of the full system with an additional 80 ns/day speed increase for the test system. It should be noted that the full HIV-1 protease system has 198 protein residues and 3130 atoms in total. The 30 Å cutoff consisting of the entire protein is used as the reference of full system. The speed increase is not obvious going from 17 to 25 Å truncation. The advantage of truncation will be more pronounced for larger systems.

We also observe that the accuracy of truncation may be system dependent. Charged residues may have a strong influence on the water interaction energies due to the electrostatic interactions. Study on additional protein systems will be required to fully understand this observation.

The comparison of hydration sites obtained with 12 Å system truncation and the full system also shows excellent agreement (Figure 4.4).

4.3.5 Speed Increase with GPU-Accelerated Implementation of WATsite

Finally, we compared the simulation performance of our new GPU-accelerated WATsite3.0 implementation in OpenMM with the previous WATsite2.0 version using GROMACS. When considering just the MD simulation portion itself, a three time speed increase is achieved using a single GPU with OpenMM comparing to using 16 CPU cores with GROMACS (Figure 4.5). In addition, the total computation time

	OC				
Truncation (Å)	Occupancy	ΔG	ΔH	-TΔS	Speed (ns/day)
12	0.97	0.91	0.91	0.93	255.0
15	0.97	0.91	0.91	0.93	213.5
17	0.97	0.92	0.91	0.94	189.0
20	0.97	0.93	0.93	0.95	172.8
25	0.97	0.93	0.93	0.95	173.5
30	1.0	1.0	1.0	1.0	172.7

Table 4.2.
Effects of truncation on the accuracy of grid energies and the speed of simulation.

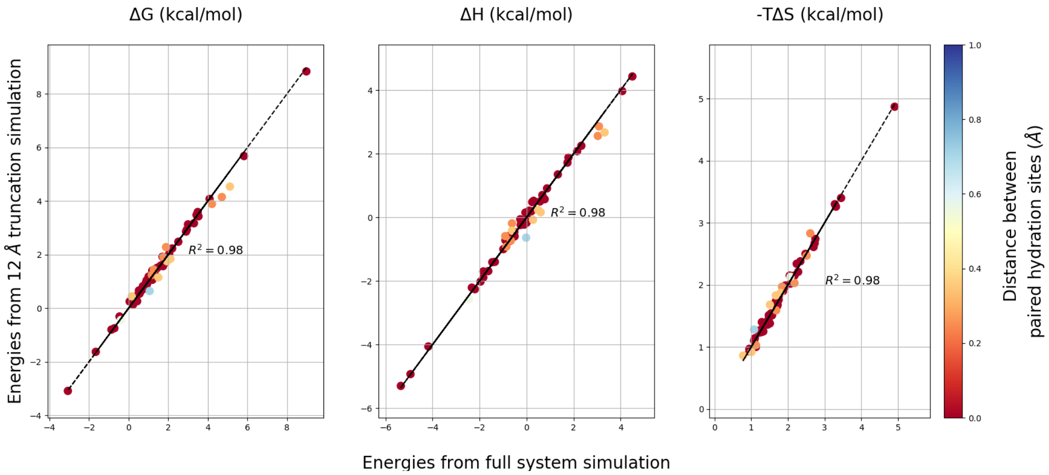


Fig. 4.4. Comparison of hydration sites predicted with 12 Å system truncation versus full system.

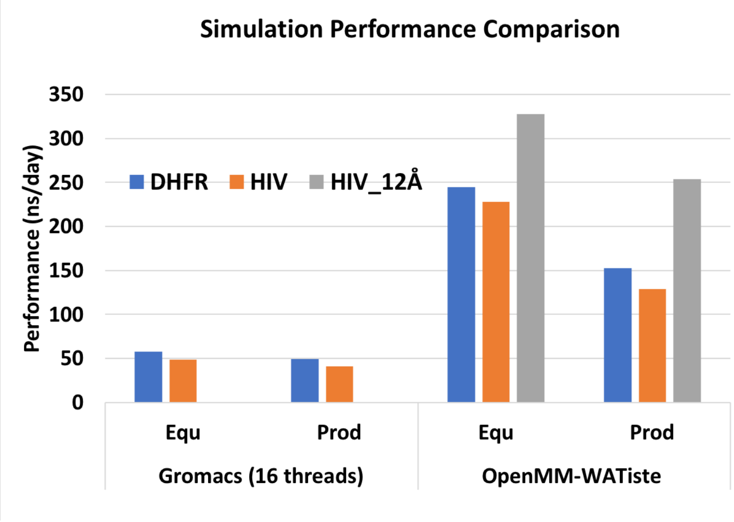


Fig. 4.5. Simulation performance of WATsite with and without GPU-acceleration.

includes post-analysis for hydration site prediction. The previous WATsite2.0 implementation based on GROMACS performs free energy calculation for each individual hydration site afterwards as a energy rerun. In contrast, the new implementation of WATsite3.0 based on OpenMM generates water energies during the MD simulation. Whereas this implementation slightly reduces the performance of the OpenMM-Watsite simulation compared to standard OpenMM (cf. Figure 4.5, Equ vs Prod), it accelerates the energy analysis significantly. A speed-up of 15-18 can therefore be obtained for the whole hydration site profiling procedure using WATsite3.0 compared to WATsite2.0 (Figure 4.6).

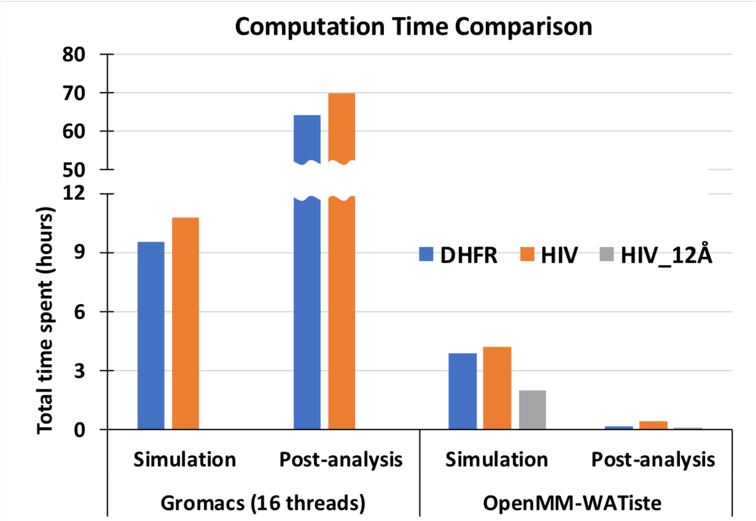


Fig. 4.6. Total time cost of MD simulation and post hydration site analysis.

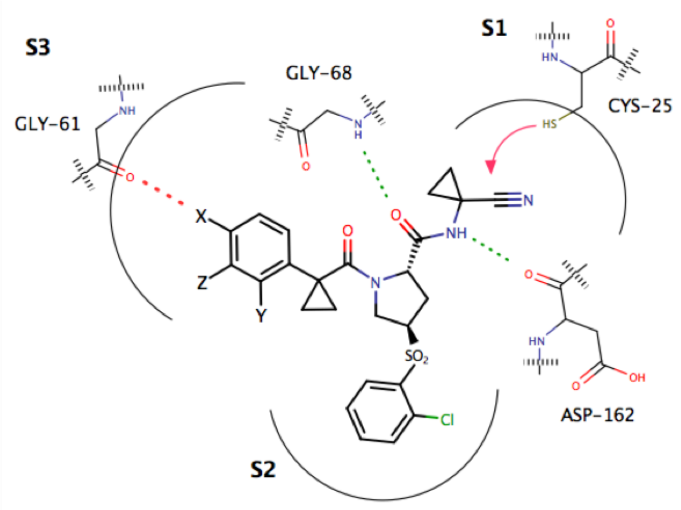
5. MODELING OF HALOGEN-PROTEIN INTERACTIONS IN CO-SOLVENT MOLECULAR DYNAMICS SIMULATIONS

5.1 Introduction

Co-solvent molecular dynamics (MD) simulations [80] have recently become important tools in structure-based drug design, for example, for identifying binding hotspots [81–84], assessing druggability of binding site [85], identifying allosteric or cryptic sites [15, 82, 86], and assisting the scoring and ranking of ligands [83, 87–89]. Commonly, a small set of probes is used to represent aromatic, aliphatic, hydrogen-bond donor, hydrogen-bond acceptor, and charged functional groups of potentially interacting ligands. Other important interactions, such as halogen-bonding, are not incorporated in standard co-solvent simulations.

Halogen substituents, however, are frequently used in pharmaceutics to increase binding affinity via halogen bonding (XB) [90] or improve pharmacokinetic properties such as oral bioavailability [91] and blood-brain barrier permeability [92]. Halogen bonding is a noncovalent interaction between the electrophilic region on the halogen atom (also called σ-hole) and a nucleophilic region of an acceptor group such as a Lewis base or a π system [90]. Halogen bond strength increases with the magnitude of the σ-hole in the order Cl<Br<I. Fluorine, which is lacking any σ-hole, does not form halogen bonds.

The impact of halogen bonding on binding affinity has been demonstrated in several protein-ligand systems [93–97]. One well-documented study contains a series of halogenated, methylated, and unsubstituted analogs binding to human Cathepsin L (hCatL). As shown in Figure 5.1, increase in affinity was observed upon changing the *para*-aryl-X substituent from F, over H, CH_3, Cl, Br, to I. Co-crystal structures



	X	Y	Z	PDB	IC_{50} (µM)
IA1	H	H	H		0.29
IA2	CH_3	H	H	2XU5	0.13
IA3	F	H	H	2XU4	0.34
IA4	Cl	H	H	2YJC	0.022
IA5	Br	H	H	2YJ2	0.012
IA6	I	H	H	2YJ8	0.0065
IA7	CF_3	H	H	2YJ9	0.095
IB1	H	F	H		0.32
IB2	F	F	H		0.35
IB3	Cl	F	H		0.03
IB4	Br	F	H		0.0065
IB5	I	F	H		0.0043
IC1	H	H	F		0.52
IC2	F	H	F		0.93
IC3	Cl	H	F		0.022
IC4	Br	H	F		0.03

Fig. 5.1. Structures of the human cathepsin L (hCatL) inhibitors and their interaction scheme. Inhibitors of hCatL undergo reversible covalent binding to the thiol group of CYS-25 in the S1 pocket under formation of thioimidates. In addition, hydrogen bonds are also formed to the backbone NH group of GLY-68, and C=O group of ASP-162 (green dashed lines). The XB interaction to GLY-61 is highlighted in red. (adapted from Hardegger et al. [98])

(Figure 5.2) further support the formation of halogen bonds showing close proximities of Cl, Br, and I atoms of around 3.1 Å to the GLY-61 backbone oxygen while the C-X$\cdots$O angles are around 175°.

To incorporate halogen-bonding interactions into co-solvent MD simulation, we propose to add halogenated probes (fluoro-benzene, chloro-benzene, bromo-benzene, and iodo-benzene) to the arsenal of standard co-solvent probes. The approach will be tested by investigating its potential to differentiate the binding affinities of a series of inhibitors to hCatL (IA1-7, IB1-5, IC1-4 in Fig 5.1).

In order to investigate the impact of halogenated probes in co-solvent simulations, unmodified and halogenated benzene (PhX) probes (fluoro-benzene, chloro-benzene, bromo-benzene, iodo-benzene) were employed in the study. The set of probes was completed by the addition of commonly used co-solvent probes propane (aliphatic),

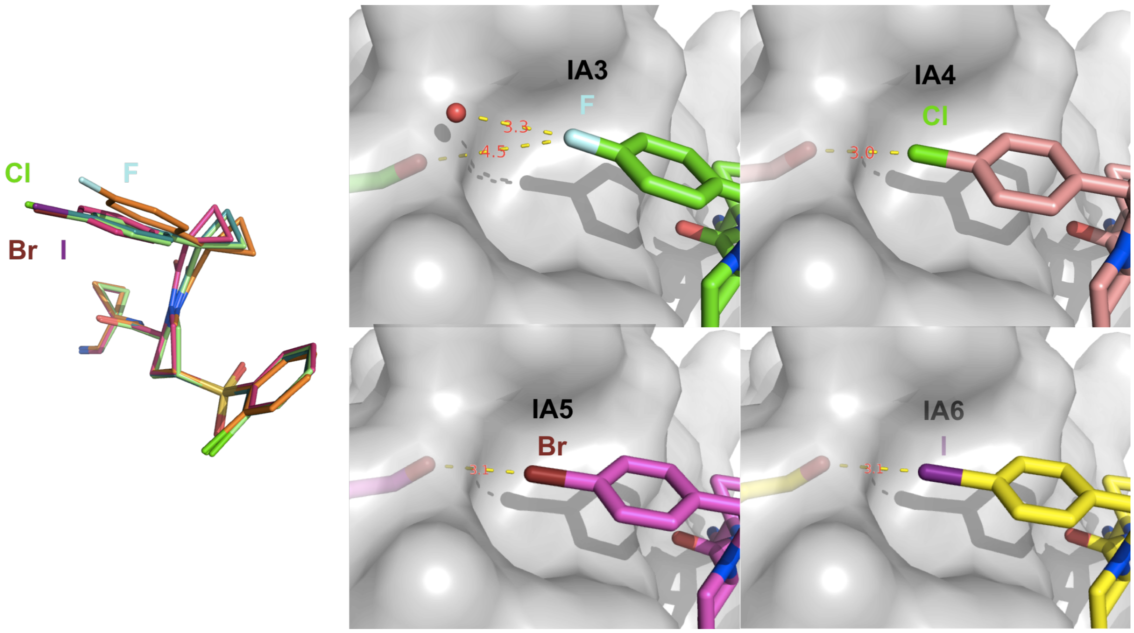


Fig. 5.2. Crystal structures showing the interaction between the backbone carbonyl of GLY-61 in the S3 pocket of hCatL with F, Cl, Br, and I. PDB ID: F 2XU4, Cl 2YJC, Br 2YJ2, I 2YJ8. Color: F turquoise, Cl green, Br brown, I purple. Figure was generated using PyMOL [99].

formamide (hydrogen-bond donor and acceptor), and acetaldehyde (hydrogen-bond acceptor). The full set of probes and the functional group atoms represented by the probes is summarized in Table 5.1.

5.2 Materials and Methods

The purpose of this study is to investigate whether the inclusion of explicit halogen probes is useful for distinguishing the variant strengths of halogenated, methylated and unsubstituted inhibitors of hCatL.

5.2.1 Preparation of Protein Systems and Co-solvent Probes

Crystal structures (PDB ID: 2XU4, 2XU5, 2YJC, 2YJ2, 2YJ8, 2YJ9) of human Cathepsin L were obtained from the Protein Data Bank [98] (Figure 5.2). The protein structure of 2YJ8 was prepared using Schrodinger's Protein Preparation Wizard [77]. In short, hydrogen positions, bond orders, protonation states (HIS, ASP, GLU), and conformational flips of ASN and GLN side chains were optimized using the default protocol.

A total of 11 co-solvent probes were prepared and used in this study, including propane, formamide, acetaldehyde, benzene, fluoro-benzene, chloro-benzne with and without extra-particle (EP), bromo-benzene with and without EP, iodo-benzene with and without EP.

Gaussian16 [100] was used to generate the electrostatic potential for each probe molecule at the HF/6-31G* level with iodine treated with the aug-cc-pVDZ-PP basis set. The ESPGEN program in Amber [101] was subsequently used to extract the RESP charges. Additionally for chloro-, bromo-, and iodo-benzene with EP, a mass-less dummy atom was placed along the C-X bond at a 1.6 Å, 1.6 Å, 1.8 Å distance [102] from the halogen atom, respectively (Figure 5.3). Next, the two-step restrained electrostatic potential (RESP) procedures were carried out in order to assign the partial charge to all atoms including the extra-particle. The electrostatic potential surfaces

Table 5.1.
Fragment map type and correspondence to co-solvent probe atoms

	Fragment map type	Description of fragment map	Co-solvent probe atoms
1	HPHOB, ARO	Aromatic hydrophobic	Benzene carbon
2	HPHOB, ALI	Aliphatic hydrophobic	Propane carbon
3	ACC	Acceptor	Formamide & Acetaldehyde oxygen
4	DON	Donor	Formamide nitrogen
5	FBZ	Fluorine	Fluoro-benzene fluorine
6	CBZ	Chlorine	Chloro-benzene chlorine
7	BBZ	Bromine	Bromo-benzene bromine
8	IBZ	Iodine	Iodo-benzene iodine
9	EXCLUSION	Steric overlap with protein	Protein heavy atoms

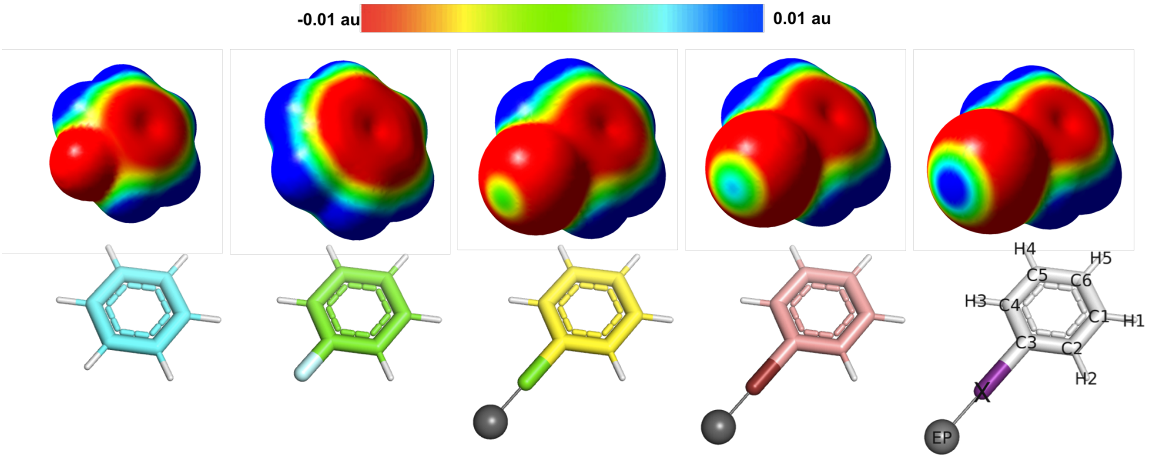


Fig. 5.3. Electrostatic-potential surfaces of five PhX probes used in the co-solvent simulation. The electrostatic potential is mapped onto the isosurface at an electron density value of 0.0004 e au^{-3} .

of the different phenylhalide (PhX) probes generated with GaussView [103] are shown in Figure 5.3, and their RESP fitted partial charges are listed in Table 5.2. Atom types and other parameters (e.g., LJ parameters) of the probes were obtained from the general AMBER force field (GAFF).

Eight sets of co-solvent systems were generated all including propane, formamide, and acetaldehyde, but each system containing a different PhX probe (Table 5.3). For each system, GROMACS insert-molecules utility [76] was used to randomly place the probes around the protein system. Ten different simulation systems with varying initial probe locations were generated. Water molecules were added to obtain a final concentration of 0.25 M for each probe molecule.

5.2.2 MD simulations

MD simulations were performed using the GPU-accelerated OpenMM [48] package with the AMBER14SB force field and SPC/E water model.

Table 5.2.
RESP fitted partial charges of PhX probes used in the co-solvent simulation.

Atom name	EP	X	C3	C2/C4	C1/C5	C6	H2/H3	H1/H4	H5
Benzene	N/A	0.13	-0.13	-0.13	-0.13	-0.13	0.13	0.13	0.13
F-benzene	N/A	-0.257	0.475	-0.338	-0.070	-0.235	0.190	0.147	0.158
Cl-benzene	N/A	-0.132	0.003	-0.047	-0.190	-0.111	0.131	0.156	0.142
Cl-benzene EP	0.064	-0.313	0.324	-0.236	-0.126	-0.156	0.180	0.149	0.147
Br-benzene	N/A	-0.085	-0.140	0.015	-0.189	-0.104	0.118	0.150	0.140
Br-benzene EP	0.103	-0.359	0.309	-0.218	-0.137	-0.145	0.178	0.150	0.146
I-benzene	N/A	-0.058	-0.318	0.192	-0.309	0.005	0.072	0.168	0.123
I-benzene EP	0.113	-0.360	0.270	-0.197	-0.157	-0.118	0.177	0.153	0.142

Table 5.3.

Co-solvent System Setup.

Set	Probes in the system	Fragment Map
1	Benzene, Propane, Formamide, Acetaldehyde	HPHOB, ARO HPHOB, ALI ACC & DON EXCLUSION
2	Fluoro-benzene Propane, Formamide, Acetaldehyde	FBZ
3	Chloro-benzene **with EP** Propane, Formamide, Acetaldehyde	CBZ
4	Bromo-benzene **with EP** Propane, Formamide, Acetaldehyde	BBZ
5	Iodo-benzene **with EP** Propane, Formamide, Acetaldehyde	IBZ
6	Chloro-benzene without EP Propane, Formamide, Acetaldehyde	CBZb
7	Bromo-benzene without EP Propane, Formamide, Acetaldehyde	BBZb
8	Iodo-benzene without EP Propane, Formamide, Acetaldehyde	IBZb

The SHAKE algorithm [50] was applied to constrain bonds containing hydrogen atoms to their equilibrium length and to maintain rigid water geometries. Long-range electrostatic interactions were handled with the Particle Mesh Ewald method [51] with a cutoff of 10 Å for the direct interactions. The Lennard-Jones interactions were truncated at a distance of 10 Å, and a long-range isotropic correction was applied for Lennard-Jones interactions beyond the cutoff. A Langevin integrator with a time step of 2 fs was used together with a stochastic thermostat collision frequency of 1 ps^{-1}. The pressure control was done by adjusting the size of the periodic box, simulating the effect of constant pressure.

With all heavy atoms harmonically restrained (spring constants of 1 kcal mol^{-1} Å^{-2}), the system was first energy minimized and then heated to 298 K over the 50 ps length of an MD simulation, followed by 1 ns of equilibration at a temperature of 298 K and pressure of 1 bar with periodic boundary conditions in all three dimensions. A weak restraint on the backbone heavy atoms with a force constant of 0.1 kcal mol^{-1} Å^{-2} was applied in the production run, to prevent potential protein denaturation in the presence of highly concentrated co-solvent solution. Each of the protein-co-solvent systems was simulated for 50 ns, resulting in a total simulation length of 4 μs (8 different co-solvens x 10 different initial probe locations). Snapshots were saved every 10 picoseconds in NetCDF format, generating 10 x 5,000 frames for each PhX co-solvent system.

In order to prevent the aggregation between hydrophobic probes, an artificial repulsive force (Equation 5.1) between the center of an aromatic ring and/or the centroid of propane was added defined by a CustomNonbondedForce in OpenMM [48].

$$V = \frac{1}{2}(r-7)^2\Theta(7-r) \tag{5.1}$$

with the step function Θ(x) = 1 for x > 0 and Θ(x) = 0 for x ≤ 0. We decided to not mix different halogenated aromatic probes during each co-solvent MD simulation. Such a mixing would require much longer co-solvent MD simulations due to the relatively large number of probes and the artificial repulsive forces added between

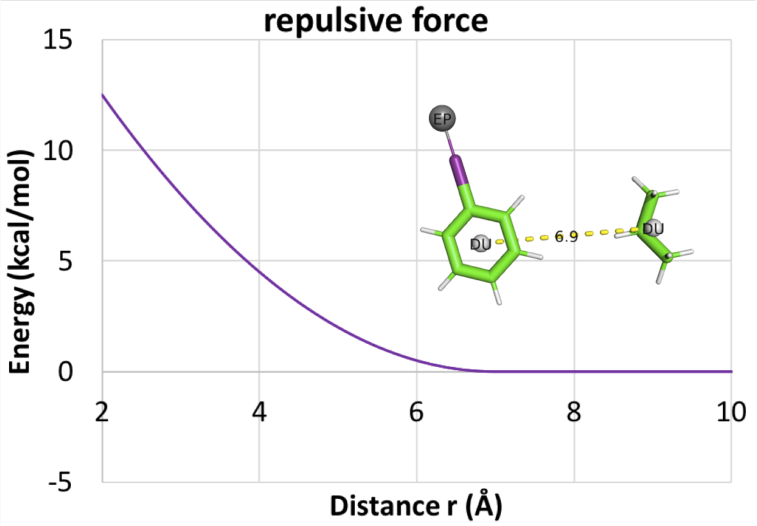


Fig. 5.4. Artificial repulsive force between all hydrophobic probes. One example shown for propane and iodo-benzene.

hydrophobic probes, thus reducing the exchange and therefore sampling of those probes in the binding site. It should, however, be noted that competition between aromatic probes to other probes is still present in all simulations.

5.2.3 Fragment Map and Free Energy Map Generation

The trajectories of the co-solvent MD simulations were analyzed using cpptraj from AmberTools [101]. 3D histograms of the occupancies of probe atoms were generated: In short, a 3D grid is placed over the entire volume of the simulation system with a grid spacing of 1 Å. Occupancy of each type of probe atom of interest throughout the simulations was computed in the 3D grid, generating a total of nine occupancy grids.

To test convergence of the co-solvent simulations, we followed the procedure of Rraman et al. [87]. In short, the ten independent simulations are split into two groups of five and the density on each grid point for all atom types is compared by calculating the overlap coefficient (OC) (Equation 5.2).

$$OC = \sum_{i=1}^{N} \Bigg(\min \left(\frac{Q_i^1}{\sum_{j=1}^{N} Q_j^1};\ \frac{Q_i^2}{\sum_{j=1}^{N} Q_j^2} \right) \tag{5.2}$$

OC values range from 0 to 1 with the latter expressing full convergence/reproducibility between the two sets of simulations. Fragment occupancy maps were converted to free energy maps (FEMap) using Equation 5.3,

$$\Delta G_{i,j,k} = -k_B T \ln\left(\frac{voxel\ occupancy\ at\ grid\ point\ i,\ j,\ k}{average\ bulk\ occupancy}\right) \tag{5.3}$$

For each fragment map, the grid occupancies obtained from the co-solvent simulation were normalized by the bulk occupancy which was obtained from a bulk simulation with only co-solvent probes in water.

5.2.4 Ligand Scoring and Local MC Sampling

To investigate the usefulness of the FEMaps for the various halogen atoms for ranking and scoring the compound series, local Monte Carlo (MC) sampling was performed starting from the initial X-ray structure of the compounds. Compounds without X-ray structure were first aligned to existing co-crystallized ligands based on their common scaffold.

Force field parameters for the inhibitors were obtained by the ANTECHAMBER program [101]. A Python script based on Siremol [104] was developed to read the ligand parameters, and perform the MC sampling. At each step S, the energy of the current configuration was computed by Equation 5.4,

$$E_S = LGFE + E_{intra\_vdW} + E_{intra\_el.st.} + E_{intra\_dihedral} + E_{constraint} \tag{5.4}$$

where LGFE is ligand grid free energy summed over all of ligand heavy atoms, where the energy for each atom is calculated by trilinear interpolation in the corresponding FEMaps. E_{intra_vdW}, $E_{intra\_el.st.}$, $E_{intra\_dihedral}$ are the intra-ligand van der Waals, electrostatic and torsion energies. A constraint energy $E_{constraint}$ was added between the position of sulfur atom of CYS-25 and the carbon atom of cyano group in the ligand in order to mimic the covalent interaction of the inhibitors with the protein.

During the MC sampling, only acyclic and single bonds were included for torsional rotation, and random translation and rotation at each step of up to 0.2 Å and 9° was allowed, respectively. MC sampling was repeated for 20 runs. For each run, 1000 steps of standard MC at a temperature of 300 K followed by 4000 steps of simulated annealing were performed, and the minimum LGFE was reported. The MC sampling was performed against each of the three collections of FEMaps (Table 5.5), and the minimum LGFE from each FEMap collection was used for predicting the free energy which was correlateded with the experiment ΔG values.

Table 5.4.
Overlap coefficient (OC) calculated for all fragment FEMaps.

Fragment Map	Overlap Coefficient
HPHOB,ARO	0.918
HPHOB,ALI	0.905
ACC	0.869
DON	0.790
FBZ	0.811
CBZ	0.769
BBZ	0.804
IBZ	0.793

5.3 Results and Discussion

5.3.1 Convergence of Fragment FEMap

For each type of FEMap, we checked if convergence has been achieved within the duration of the simulations. Ten trajectories (50 ns each) were divided into two groups (250 ns each group). Comparison of fragment FEMaps were performed qualitatively (Figure 5.5) and quantitatively by calculating OCs (Table 5.4).

Good convergence (OC values close to the maximum of one) is obtained for all FEMaps, with higher OC for the hydrophobic and acceptor maps as a larger number of atoms is used to generate these fragment maps (six benzene, three propane, two oxygen acceptor atoms).

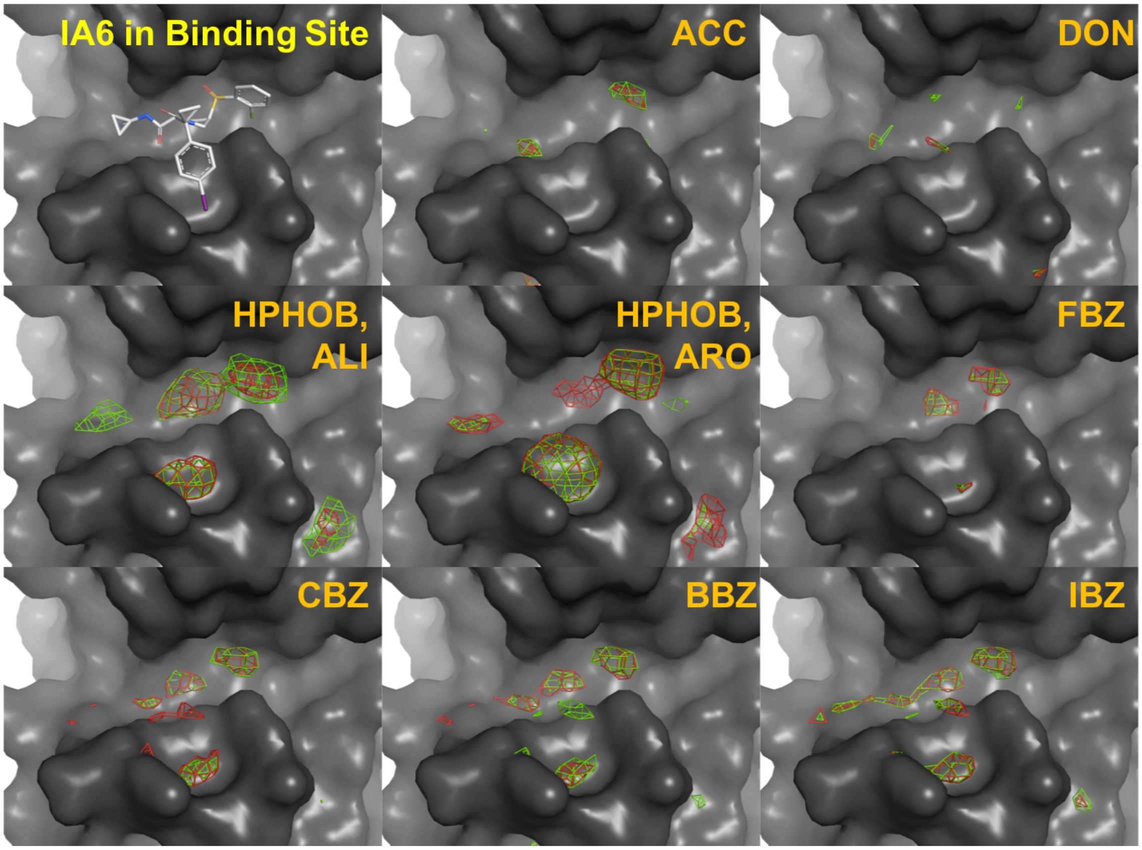


Fig. 5.5. Qualitative comparison of fragment maps. Fragment FEMaps from trajectories 1-5 shown in green, and 6-10 shown in red. FEMaps are shown at free energy values of -1.5 kcal/mol for ACC, DON, HPHOB,ALI, HPHOB,ARO, and -2.5 kcal/mol for the remaining atom types.

5.3.2 Inhibitor Halogen Atom Location Revealed by Fragment Map

To validate the utility of the halogen FEMaps in identifying halogen-bonding interactions, the FEMaps of different halogen maps together with the co-crystallized ligands are shown in Figure 5.6 (free energy isovalues of -2.6 kcal/mol and -3.2 kcal/mol are displayed on the left and right side, respectively). At the free energy level of -2.6 kcal/mol, there is no density for fluorine (FBZ) where the halogen atoms are located in the x-ray structure consistent with the unfavorable fluorine-oxygen interactions. The minimum free energy values for the different halogen FEMaps in this particular region are -2.59, -3.28, -3.56, -4.13 kcal/mol for F, Cl, Br and I, respectively. This trend corresponds qualitatively with the observed ranking of the compounds (cf. Fig 5.1).

5.3.3 Correspondence between LGFE and Experimental Affinity

To test the scoring and ranking power of the FEMaps with explicit modeling of halogen-bonding interactions (termed EP-FEMaps), FEMaps for the same halogenated probes without extra-particle (termed noEP-FEMaps) were generated using different sets of co-solvent simulations removing the EPs for treating the σ-holes in the force field (Table 5.5). To compare our more advanced treatment of halogen-containing compounds with standard co-solvent approaches, all halogen FEMaps were ignored throughout the MC sampling and scoring. The atomistic score of halogen atoms in these so-called standard-FEMaps was obtained by mapping halogen atoms to the aliphatic hydrophobic FEMap obtained from the propane probe.

Figure 5.7 summarizes the scoring and ranking quality of the different FEMaps used during MC sampling and free energy prediction. Excellent correlation (R^2=0.85) and ranking (σ=0.96) was obtained for the full set of compounds using the EP-FEMaps, demonstrating the potential usefulness of the halogenated benzene probes for co-solvent simulations. A similar good correlation and ranking was observed (R^2=0.84, σ=0.90) for the full ligand set using the corresponding FEMaps of halo-

Table 5.5.
Three collections of FEMaps used in the MC sampling. Exclusion map are always included.

Naming	FEMaps included
Standard-FEMaps	HPHOB,ARO HPHOB,ALI ACC DON
noEP-FEMaps	HPHOB,ARO HPHOB,ALI ACC DON FBZ CBZb BBZb IBZb
EP-FEMaps	HPHOB,ARO HPHOB,ALI ACC DON FBZ CBZ BBZ IBZ

genated probes without EP (noEP-FEMaps). Surprisingly, a decent, although significant lower correlation and ranking quality was found (R^2=0.46, σ=0.64) when ignoring any information from the halogenated benzene probes (standard-FEMaps), representing the information of standard co-solvent simulations.

To understand these observations, it should be noted that the overall affinity trend cannot be solely explained by the strength of halogen-bonding. Besides the backbone carbonyl group, the para-substituent X on the aromatic ring is partially surrounded by hydrophobic amino-acid side-chains perpendicular to the C-X axis. Thus, hydrophobic contacts and van-der-Waals interactions play an important role for the observed differences in binding affinity, preferring larger and more hydrophobic groups such as Cl, Br and I over less hydrophobic elements such as H or F. Therefore, the pure hydrophobic treatment of the halogen atoms in the standard-FEMaps treatment is sufficient to separate the high affinity from low affinity group of compounds.

Removing the low affinity compounds with X=H, F from the analysis, however, demonstrates the advantage of explicit treatment of halogen atoms in the co-solvent sampling (Figure 5.8). Whereas, the correlation and ranking using EP-FEMaps remained high over this much narrower range of affinities, the standard-FEMaps completely fail in predicting the experimental binding affinities. Ignoring the explicit modeling of σ-hole in the noEP-FEMaps also lowered the correlation and ranking quality more significantly compared to the EP-FEMaps, but still allows for a relatively good prediction of binding affinities.

Focus on the most affine compounds (X=Cl, Br, I) whose affinity gain is due to halogen bonding, reveals the full advantage of the explicit representation of the σ-hole by EPs (EP-FEMaps) compared to standard single point-charge representation of the halogen atoms (noEP-FEMaps) (Figure 5.9). However, as still high ranking quality was observed when using EP-FEMaps, noEP-FEMaps lacks similar qualities. This observation is consistent with the experimental findings, which suggest that the affinity differences between IA4 (X=Cl; IC_{50}=22nM), IA5 (X=Br; IC_{50}=12nM) and IA6 (X=I; IC_{50}=6.5nM), for example, are due to increasing halogen-bonding strength from Cl, over Br, to I [90].

5.4 Conclusion

We have described the first attempt to model halogen-bonding interactions within co-solvent simulations. Our study shows that the inclusion of those probes allows for the accurate scoring and ranking of compound libraries containing halogenated ligands. It should be noted that binding affinity increases due to halogen-substitutions are not always driven by direct halogen-bonding interactions but the hydrophobic effect. In those cases, simple hydrophobic probes may be sufficient to identify those hydrophobic subpockets in the binding site during co-solvent simulations. Even if halogen-bonding interactions are involved in the protein-ligand complexes of interest, hydrophobic contacts are likely to be contributing factors to the observed binding affinity trends. For compound classifications, e.g. in actives or non-actives, neglecting halogen-bonding interactions during co-solvent interactions may be acceptable. Nevertheless, our study demonstrated that explicit modeling of the heterogeneous electron density of halogen atoms with extra particle point charges provides advantages for accurate ranking of different halogen-containing compounds binding to the same target. Increasing the number of co-solvent probes by the addition of halogen-benzene fragments exerts a challenge to achieve sufficient sampling. Here, we addressed the issue by increasing the number of simulation systems, not mixing the different halogenated probes. With the current computational resources using GPU architecture, this may not be a significant limitation as the simulation need to be run only once per target protein.

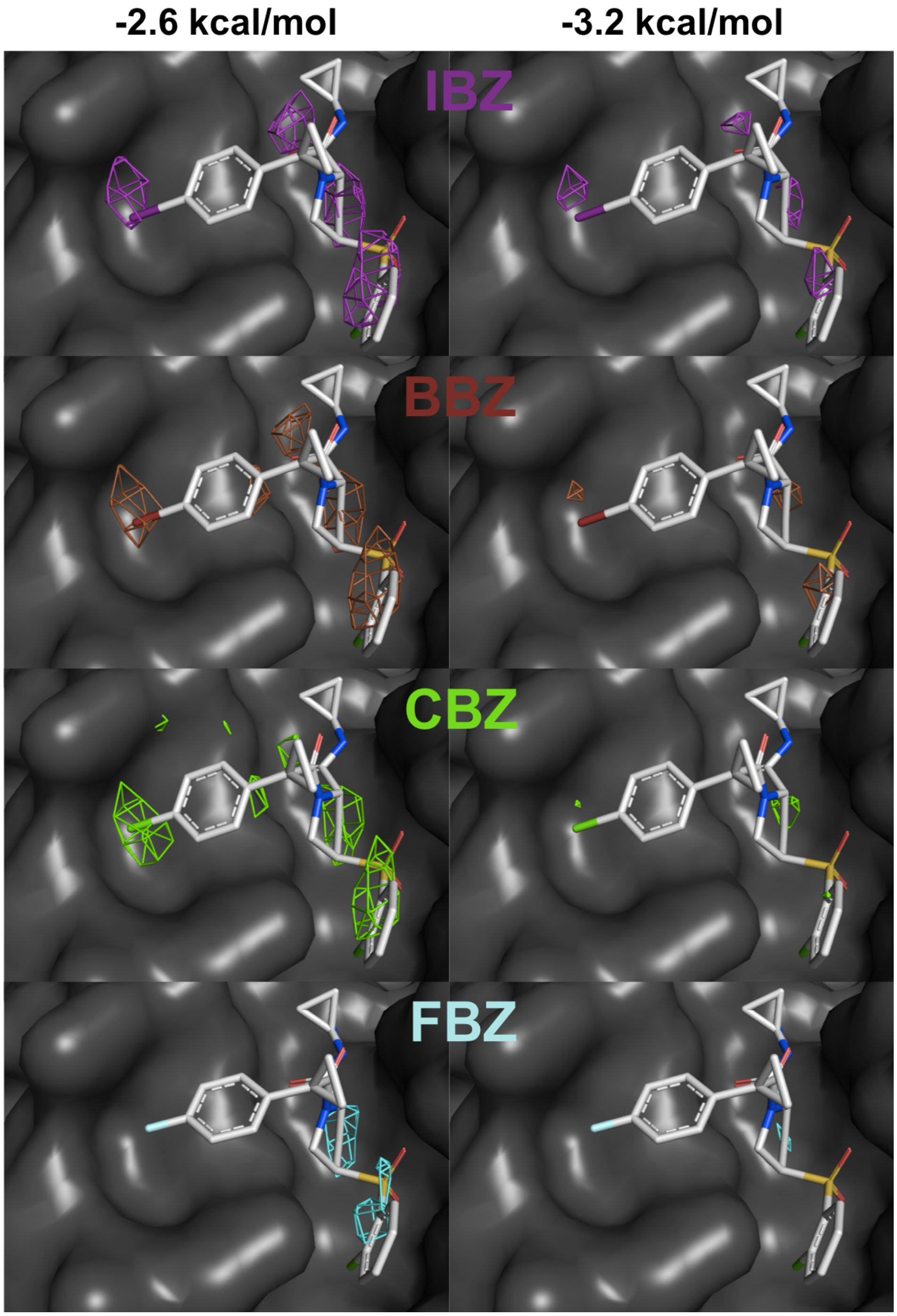
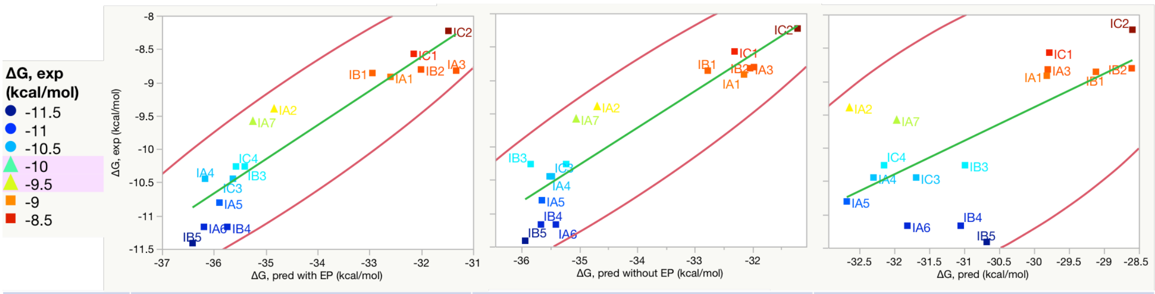
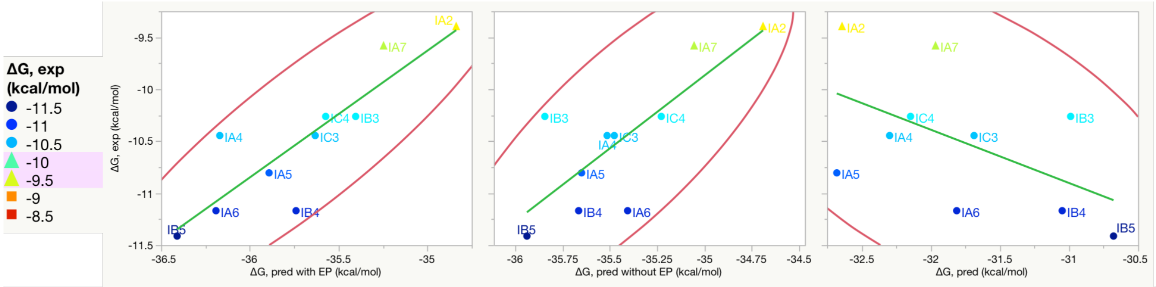


Fig. 5.6. Individual halogen FEMaps at two free energy isolevels. Left: -2.6 kcal/mol; Right: -3.2 kcal/mol. Color: F turquoise, Cl green, Br brown, I purple.



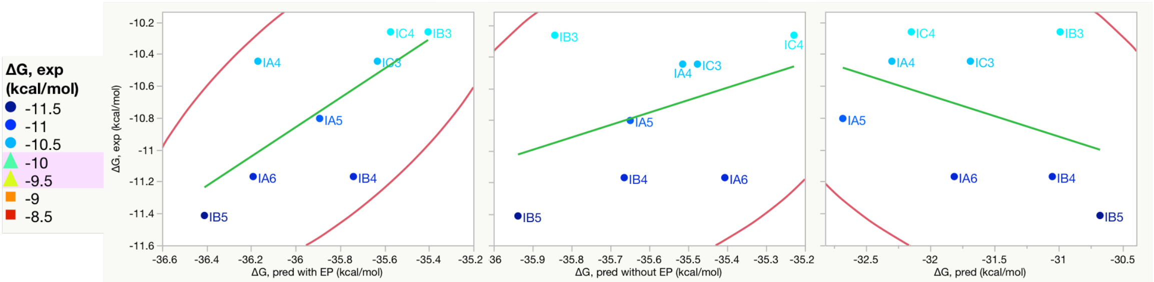
FEMaps	EP-FEMaps	noEP-FEMaps	Standard-FEMaps
R^2 of linear fit	0.85	0.84	0.46
Spearman's ρ	0.96	0.90	0.64
Prob>|ρ|	<0.0001*	<0.0001*	0.0078*

Fig. 5.7. Bivariate fit of experiment ΔG by the predicted LGFE for all ligands. LGFE predicted with (Left) explicit halogen probes with EP, (Middle) explicit halogen probes without EP, (Right) only general FEMaps.



FEMaps	EP-FEMaps	noEP-FEMaps	Standard-FEMaps
R^2 of linear fit	0.76	0.62	0.29
Spearman's ρ	0.92	0.69	-0.44
Prob>|ρ|	0.0002*	0.0288*	0.2028

Fig. 5.8. Bivariate fit of experiment ΔG by the predicted LGFE for Methyl, Cl, Br, I ligands. LGFE predicted with (Left) explicit halogen probes with EP, (Middle) explicit halogen probes without EP, (Right) only general FEMaps.



FEMaps	EP-FEMaps	noEP-FEMaps	Standard-FEMaps
R^2 of linear fit	0.50	0.16	0.16
Spearman's ρ	**0.84**	0.38	-0.28
Prob>|ρ|	0.0096*	0.3589	0.5037

Fig. 5.9. Bivariate fit of experiment ΔG by the predicted LGFE for only Cl, Br, I ligands. LGFE predicted with (Left) explicit halogen probes with EP, (Middle) explicit halogen probes without EP, (Right) only general FEMaps.

6. FUTURE DIRECTIONS

6.1 Research Summary

Calculation of thermodynamic properties associated with protein-ligand binding has been a grand challenge in computational chemistry with particular importance to drug discovery. The overall goal of this thesis was to address how free energies of individual water molecules under consideration of protein flexibility can be incorporated into the prediction of thermodynamic profiles of protein-ligand binding. Chapter 1 detailed the computational method of hydration site analysis and presented two types of scenarios where water prediction can be useful. Chapter 2 discussed the influence of the simulation protocol on hydration site prediction. Chapter 3 incorporated protein flexibility into the prediction of protein de-solvation free energies which is a significant contribution to the free energy of protein-ligand binding. Chapter 4 developed and validated two methods to speed up hydration site analysis: GPU-acceleration and system truncation. Chapter 5 extended the simulation protocol from pure water to mixed water-organic probes simulations with particular emphasis on the accurate modeling of halogen atom-protein interactions. Whereas there is still a lack in routine inclusion of water analysis in drug discovery projects, this thesis conceptually proved the significance of incorporating water free energies and protein flexibility into structure-based drug design. The remainder of this chapter will highlight some potential future directions resulting from the work presented here.

6.2 Methodological Improvements for Hydration Site Prediction

For some systems, the prediction of water locations and energies remains challenging/unsuccessful due to the nature of the binding site and/or the current force field parameters used for the MD simulations.

6.2.1 Occluded Binding Sites

The protein systems we used to validate our hydration site prediction methodology possess solvent accessible binding sites. Analyzing the convergence behavior of hydration site prediction for different proteins in chapter 3 indicated that for more buried binding sites or cavities much longer simulations are required to reach convergence.

For example, human Interleukin-1β (IL-1β, PDB: 2NVH) contains five cavities with four of them containing one to two water molecules and a central buried pocket with a volume of 40 Å^3 which is nonpolar. The prediction of hydration sites in the latter pocket largely depends on the initial placement of water molecules in the cavity as water exchange between pocket and bulk solvent is prevented or highly limited. In such instances, methods such as grand canonical Monte Carlo (GCMC) may be useful for inserting and deleting water molecules. Thus, combination of CGMC with WATSite could allow the prediction of hydration sites for occluded binding sites.

6.2.2 Protonation states of binding site residues

Another issue is that changes in protonation states of amino acids can significantly influence hydrogen-bond networks involving water molecules and therefore hydration site prediction. A preliminary test on HIV-1 protease shows significant changes in hydration site profiling dependent on the chosen protonation states of residue ASP-25 and ASP-125. It is not obvious to assign unique protonation states to binding site residues as changes can occur during ligand binding or protein conformational changes or may depend on the formed water-mediated hydrogen-bond network. Utilization of constant pH simulations to probe flexible protonation states of binding site residues may be a future direction to explore.

6.2.3 Influence of Polarization

Using classic force fields, the electrostatic interactions within the protein is treated by placing fixed point charges on the atomic centers. While classic force fields have been successful for many biochemical and pharmacological applications, one of the major approximation is the omission of polarizability, i.e. changes in charge distribution in a molecules or chemical group in response to the environment. Polarization is expected to contribute 10-20% of the total interaction energy of a protein-ligand complex and even more for charged systems [105]. With advances in computer hardware, polarizable force fields have been developed to explicitly address polarization [106–108].

Conformational changes in proteins may trigger changes in charge distribution, with locations and thermodynamic profiles of water molecules being significantly influenced. Modeling of explicit polarization may improve hydration site prediction in both highly charged and highly hydrophobic binding site of proteins where the "average" charge distribution of classical force fields may provide imprecise representations of the electrostatic interactions. Additional exploration of the influence of polarization on hydration site prediction is of particular interest.

6.3 Routine Consideration of Explicit Water Molecules in Drug Discovery Projects

Reliability of predicted locations of hydration sites might be confirmed by high-resolution X-ray crystallographic structure. Thermodynamic properties of hydration sites, however, especially entropy, can not be directly validated by experiments. As noted by the Roche Pharmaceutical Research and Early Development group, the predicted water positions and energies can not be easily translated into hypotheses driving drug discovery [109]. Therefore, in order to routinely include water information in drug discovery projects, water locations and energies need to be combined with other approaches to provide a full picture of the thermodynamics of ligand binding including de-solvation.

6.3.1 MM-GB/SA

Prior studies by Guimarães et al. presented success stories of replacing the implicit de-solvation term in MM-GB/SA (Molecular Mechanics with Generalized Born and a hydrophobic Solvent Accessible surface area) by explicit water displacement obtained from hydration site analysis [110]. However, such studies did not consider the influence of protein flexibility on hydration site prediction as we discussed in Chapter 4. The accelerated hydration site prediction implemented and validated in Chapter 5 will make it possible to apply hydration site analysis under the influence of protein flexibility to a large library of compounds that bind to an ensemble of different holo structures of the same protein. Incorporation of protein flexibility and other energy terms in MM-GB/SA may further improve the accuracy of free energies estimation especially for protein systems that undergo large conformational change upon ligand binding.

6.3.2 Docking and Scoring

Efficient approaches like docking-based Virtual screening (VS) for a library of compounds are routinely used in drug discovery research. Multiple studies have included water locations and energies in docking [ref], however, the improvement is modest [111, 112] or system dependent [113, 114]. One possible reason arises from neglecting and/or not optimizing mediating water molecules between protein and ligand. Water molecules are incorporated as rigid entity, while slight movement and optimization of water molecules in the binding site may be beneficial. In addition, a hard distance cutoff between ligand atoms and predicted water molecule/grid is typically used to determine whether or not a water molecules is replaced by the bound ligand. The predicted protein de-solvation free energy, however, is rather sensitive to such a cutoff. Future research needs to address those issues for hydration site prediction to become a routine concepts for drug discovery projects.

REFERENCES

REFERENCES

[1] G. A. V. Norman, "Drugs, devices, and the FDA: Part 1," *JACC: Basic to Translational Science*, vol. 1, no. 3, pp. 170–179, apr 2016. [Online]. Available: https://doi.org/10.1016/j.jacbts.2016.03.002

[2] J. Hughes, S. Rees, S. Kalindjian, and K. Philpott, "Principles of early drug discovery," *British Journal of Pharmacology*, vol. 162, no. 6, pp. 1239–1249, feb 2011. [Online]. Available: https://doi.org/10.1111/j.1476-5381.2010.01127.x

[3] J. Bajorath, "Computer-aided drug discovery," *F1000Research*, vol. 4, p. 630, aug 2015. [Online]. Available: https://doi.org/10.12688/f1000research.6653.1

[4] W. Yu and A. D. MacKerell, "Computer-aided drug design methods," in *Methods in Molecular Biology.* Springer New York, nov 2016, pp. 85–106. [Online]. Available: https://doi.org/10.1007/978-1-4939-6634-9_5

[5] R. C. Godwin, R. Melvin, and F. R. Salsbury, "Molecular dynamics simulations and computer-aided drug discovery," in *Methods in Pharmacology and Toxicology.* Springer New York, 2015, pp. 1–30. [Online]. Available: https://doi.org/10.1007/7653_2015_41

[6] M. S. Bodnarchuk, "Water, water, everywhere... it's time to stop and think," *Drug Discovery Today*, vol. 21, no. 7, pp. 1139–1146, jul 2016. [Online]. Available: https://doi.org/10.1016/j.drudis.2016.05.009

[7] J. E. Ladbury, "Just add water! the effect of water on the specificity of protein-ligand binding sites and its potential application to drug design," *Chemistry & Biology*, vol. 3, no. 12, pp. 973–980, dec 1996. [Online]. Available: https://doi.org/10.1016%2Fs1074-5521%2896%2990164-7

[8] R. Baron, P. Setny, and J. A. McCammon, "Water in cavity-ligand recognition," *Journal of the American Chemical Society*, vol. 132, no. 34, pp. 12 091–12 097, aug 2010. [Online]. Available: https://doi.org/10.1021%2Fja1050082

[9] G. Hummer, "Under water's influence," *Nature Chemistry*, vol. 2, no. 11, pp. 906–907, nov 2010. [Online]. Available: https://doi.org/10.1038%2Fnchem.885

[10] P. W. Snyder, J. Mecinovic, D. T. Moustakas, S. W. Thomas, M. Harder, E. T. Mack, M. R. Lockett, A. Heroux, W. Sherman, and G. M. Whitesides, "Mechanism of the hydrophobic effect in the biomolecular recognition of arylsulfonamides by carbonic anhydrase," *Proceedings of the National Academy of Sciences*, vol. 108, no. 44, pp. 17 889–17 894, oct 2011. [Online]. Available: https://doi.org/10.1073%2Fpnas.1114107108

[11] R. Baron, P. Setny, and F. Paesani, “Water structure, dynamics, and spectral signatures: Changes upon model cavity–ligand recognition,” *The Journal of Physical Chemistry B*, vol. 116, no. 46, pp. 13 774–13 780, nov 2012. [Online]. Available: https://doi.org/10.1021%2Fjp309373q

[12] B. Breiten, M. R. Lockett, W. Sherman, S. Fujita, M. Al-Sayah, H. Lange, C. M. Bowers, A. Heroux, G. Krilov, and G. M. Whitesides, “Water networks contribute to enthalpy/entropy compensation in protein–ligand binding,” *Journal of the American Chemical Society*, vol. 135, no. 41, pp. 15 579–15 584, oct 2013. [Online]. Available: https://doi.org/10.1021%2Fja4075776

[13] F. Spyrakis, M. H. Ahmed, A. S. Bayden, P. Cozzini, A. Mozzarelli, and G. E. Kellogg, “The roles of water in the protein matrix: A largely untapped resource for drug discovery,” *Journal of Medicinal Chemistry*, vol. 60, no. 16, pp. 6781–6827, may 2017. [Online]. Available: https://doi.org/10.1021/acs.jmedchem.7b00057

[14] E. Fischer, “Einfluss der configuration auf die wirkung der enzyme,” *Berichte der deutschen chemischen Gesellschaft*, vol. 27, no. 3, pp. 2985–2993, oct 1894. [Online]. Available: https://doi.org/10.1002/cber.18940270364

[15] D. Alvarez-Garcia and X. Barril, “Relationship between protein flexibility and binding: Lessons for structure-based drug design,” *J. Chem. Theory Comput.*, vol. 10, no. 6, pp. 2608–2614, 2014.

[16] D. E. Koshland, “Application of a theory of enzyme specificity to protein synthesis,” *Proc. Natl. Acad. Sci. U.S.A.*, vol. 44, no. 2, pp. 98–104, Feb 1958.

[17] S. Kumar, B. Ma, C.-J. Tsai, N. Sinha, and R. Nussinov, “Folding and binding cascades: Dynamic landscapes and population shifts,” *Protein Science*, vol. 9, no. 1, pp. 10–19, dec 2000. [Online]. Available: https://doi.org/10.1110/ps.9.1.10

[18] D. Bucher, B. J. Grant, and J. A. McCammon, “Induced fit or conformational selection? the role of the semi-closed state in the maltose binding protein,” *Biochemistry*, vol. 50, no. 48, pp. 10 530–10 539, dec 2011. [Online]. Available: https://doi.org/10.1021/bi201481a

[19] R. D. Skeel, “What makes molecular dynamics work?” *SIAM Journal on Scientific Computing*, vol. 31, no. 2, pp. 1363–1378, jan 2009. [Online]. Available: https://doi.org/10.1137/070683660

[20] W. F. V. GUNSTEREN and H. J. C. BERENDSEN, “Molecular dynamics: perspective for complex systems,” *Biochemical Society Transactions*, vol. 10, no. 5, pp. 301–305, oct 1982. [Online]. Available: https://doi.org/10.1042/bst0100301

[21] W. F. van Gunsteren and H. J. C. Berendsen, “Computer simulation of molecular dynamics: Methodology, applications, and perspectives in chemistry,” *Angewandte Chemie International Edition in English*, vol. 29, no. 9, pp. 992–1023, sep 1990. [Online]. Available: https://doi.org/10.1002/anie.199009921

[22] D. Frenkel and B. Smit, *Understanding Molecular Simulation, Second Edition: From Algorithms to Applications (Computational Science Series, Vol 1).* Academic Press, 2001.

[23] J. Gelpi, A. Hospital, R. Goñi, and M. Orozco, "Molecular dynamics simulations: advances and applications," *Advances and Applications in Bioinformatics and Chemistry*, p. 37, nov 2015. [Online]. Available: https://doi.org/10.2147/aabc.s70333

[24] J. A. McCammon, B. R. Gelin, and M. Karplus, "Dynamics of folded proteins," *Nature*, vol. 267, no. 5612, pp. 585–590, jun 1977. [Online]. Available: https://doi.org/10.1038/267585a0

[25] T. Mashimo, Y. Fukunishi, N. Kamiya, Y. Takano, I. Fukuda, and H. Nakamura, "Molecular dynamics simulations accelerated by GPU for biological macromolecules with a non-ewald scheme for electrostatic interactions," *Journal of Chemical Theory and Computation*, vol. 9, no. 12, pp. 5599–5609, nov 2013. [Online]. Available: https://doi.org/10.1021/ct400342e

[26] J. Michel, J. Tirado-Rives, and W. L. Jorgensen, "Energetics of displacing water molecules from protein binding sites: Consequences for ligand optimization," *Journal of the American Chemical Society*, vol. 131, no. 42, pp. 15 403–15 411, oct 2009. [Online]. Available: https://doi.org/10.1021%2Fja906058w

[27] J. M. Chen, S. L. Xu, Z. Wawrzak, G. S. Basarab, and D. B. Jordan, "Structure-based design of potent inhibitors of scytalone dehydratase: displacement of a water molecule from the active site‡," *Biochemistry*, vol. 37, no. 51, pp. 17 735–17 744, dec 1998. [Online]. Available: https://doi.org/10.1021%2Fbi981848r

[28] A. Wissner, D. M. Berger, D. H. Boschelli, M. B. Floyd, L. M. Greenberger, B. C. Gruber, B. D. Johnson, N. Mamuya, R. Nilakantan, M. F. Reich, R. Shen, H.-R. Tsou, E. Upeslacis, Y. F. Wang, B. Wu, F. Ye, and N. Zhang, "4-anilino-6,7-dialkoxyquinoline-3-carbonitrile inhibitors of epidermal growth factor receptor kinase and their bioisosteric relationship to the 4-anilino-6,7-dialkoxyquinazoline inhibitors," *Journal of Medicinal Chemistry*, vol. 43, no. 17, pp. 3244–3256, aug 2000. [Online]. Available: https://doi.org/10.1021%2Fjm000206a

[29] W. R. Pitt and J. M. Goodfellow, "Modelling of solvent positions around polar groups in proteins," *"Protein Engineering, Design and Selection"*, vol. 4, no. 5, pp. 531–537, 1991. [Online]. Available: https://doi.org/10.1093%2Fprotein%2F4.5.531

[30] M. L. Verdonk, J. C. Cole, and R. Taylor, "SuperStar: A knowledge-based approach for identifying interaction sites in proteins," *Journal of Molecular Biology*, vol. 289, no. 4, pp. 1093–1108, jun 1999. [Online]. Available: https://doi.org/10.1006%2Fjmbi.1999.2809

[31] G. Rossato, B. Ernst, A. Vedani, and M. Smiesvko, "AcquaAlta: A directional approach to the solvation of ligand-protein complexes," *Journal of Chemical Information and Modeling*, vol. 51, no. 8, pp. 1867–1881, aug 2011. [Online]. Available: https://doi.org/10.1021%2Fci200150p

[32] M. Zheng, Y. Li, B. Xiong, H. Jiang, and J. Shen, "Water PMF for predicting the properties of water molecules in protein binding site," *Journal of Computational Chemistry*, vol. 34, no. 7, pp. 583–592, nov 2012. [Online]. Available: https://doi.org/10.1002/jcc.23170

[33] P. J. Goodford, "A computational procedure for determining energetically favorable binding sites on biologically important macromolecules," *Journal of Medicinal Chemistry*, vol. 28, no. 7, pp. 849–857, jul 1985. [Online]. Available: https://doi.org/10.1021%2Fjm00145a002

[34] J. Goodfellow and F. Vovelle, "Biomolecular energy calculations using transputer technology," *European Biophysics Journal*, vol. 17, no. 3, sep 1989. [Online]. Available: https://doi.org/10.1007%2Fbf00254771

[35] M. Baroni, G. Cruciani, S. Sciabola, F. Perruccio, and J. S. Mason, "A common reference framework for analyzing/comparing proteins and ligands. fingerprints for ligands and proteins (FLAP): theory and application," *Journal of Chemical Information and Modeling*, vol. 47, no. 2, pp. 279–294, mar 2007. [Online]. Available: https://doi.org/10.1021/ci600253e

[36] G. A. Ross, G. M. Morris, and P. C. Biggin, "Rapid and accurate prediction and scoring of water molecules in protein binding sites," *PLoS ONE*, vol. 7, no. 3, p. e32036, mar 2012. [Online]. Available: https://doi.org/10.1371%2Fjournal.pone.0032036

[37] N. OpenEye Scientific Software, Santa Fe, "Szmap 1.2.1.4," 2015.

[38] T. Imai, R. Hiraoka, A. Kovalenko, and F. Hirata, "Locating missing water molecules in protein cavities by the three-dimensional reference interaction site model theory of molecular solvation," *Proteins: Structure, Function, and Bioinformatics*, vol. 66, no. 4, pp. 804–813, dec 2006. [Online]. Available: https://doi.org/10.1002%2Fprot.21311

[39] G. A. Ross, M. S. Bodnarchuk, and J. W. Essex, "Water sites, networks, and free energies with grand canonical monte carlo," *Journal of the American Chemical Society*, vol. 137, no. 47, pp. 14 930–14 943, nov 2015. [Online]. Available: https://doi.org/10.1021/jacs.5b07940

[40] J. Michel, J. Tirado-Rives, and W. L. Jorgensen, "Prediction of the water content in protein binding sites," *The Journal of Physical Chemistry B*, vol. 113, no. 40, pp. 13 337–13 346, oct 2009. [Online]. Available: https://doi.org/10.1021/jp9047456

[41] Z. Li and T. Lazaridis, "Thermodynamics of buried water clusters at a protein-ligand binding interface," *The Journal of Physical Chemistry B*, vol. 110, no. 3, pp. 1464–1475, jan 2006. [Online]. Available: https://doi.org/10.1021%2Fjp056020a

[42] T. Lazaridis, "Inhomogeneous fluid approach to solvation thermodynamics. 1. theory," *The Journal of Physical Chemistry B*, vol. 102, no. 18, pp. 3531–3541, apr 1998. [Online]. Available: https://doi.org/10.1021%2Fjp9723574

[43] ——, "Inhomogeneous fluid approach to solvation thermodynamics. 2. applications to simple fluids," *The Journal of Physical Chemistry B*, vol. 102, no. 18, pp. 3542–3550, apr 1998. [Online]. Available: https://doi.org/10.1021%2Fjp972358w

[44] R. Abel, T. Young, R. Farid, B. J. Berne, and R. A. Friesner, "Role of the active-site solvent in the thermodynamics of factor xa ligand binding," *Journal of the American Chemical Society*, vol. 130, no. 9, pp. 2817–2831, mar 2008. [Online]. Available: https://doi.org/10.1021%2Fja0771033

[45] C. N. Nguyen, T. K. Young, and M. K. Gilson, "Grid inhomogeneous solvation theory: Hydration structure and thermodynamics of the miniature receptor cucurbit[7]uril," *The Journal of Chemical Physics*, vol. 137, no. 14, p. 149901, oct 2012. [Online]. Available: https://doi.org/10.1063%2F1.4751113

[46] B. Hu and M. A. Lill, "Protein pharmacophore selection using hydration-site analysis," *Journal of Chemical Information and Modeling*, vol. 52, no. 4, pp. 1046–1060, mar 2012. [Online]. Available: https://doi.org/10.1021%2Fci200620h

[47] ——, "WATsite: Hydration site prediction program with PyMOL interface," *Journal of Computational Chemistry*, vol. 35, no. 16, pp. 1255–1260, apr 2014. [Online]. Available: https://doi.org/10.1002%2Fjcc.23616

[48] P. Eastman, M. S. Friedrichs, J. D. Chodera, R. J. Radmer, C. M. Bruns, J. P. Ku, K. A. Beauchamp, T. J. Lane, L. P. Wang, D. Shukla, T. Tye, M. Houston, T. Stich, C. Klein, M. R. Shirts, and V. S. Pande, "OpenMM 4: A reusable, extensible, hardware independent library for high performance molecular simulation," *J. Chem. Theory Comput.*, vol. 9, no. 1, pp. 461–469, 2013.

[49] J. A. Maier, C. Martinez, K. Kasavajhala, L. Wickstrom, K. E. Hauser, and C. Simmerling, "ff14sb: Improving the accuracy of protein side chain and backbone parameters from ff99sb," *Journal of Chemical Theory and Computation*, vol. 11, no. 8, pp. 3696–3713, jul 2015. [Online]. Available: https://doi.org/10.1021/acs.jctc.5b00255

[50] J. paul Ryckaert, G. Ciccotti, and H. J. C. Berendsen, "Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes," *J. Comput. Phys*, pp. 327–341, 1977.

[51] T. Darden, D. York, and L. Pedersen, "Particle mesh Ewald: An Nṡlog(N) method for Ewald sums in large systems," *J. Chem. Phys.*, vol. 98, pp. 10 089–10 092, 1993.

[52] K. Haider and D. J. Huggins, "Combining solvent thermodynamic profiles with functionality maps of the hsp90 binding site to predict the displacement of water molecules," *Journal of Chemical Information and Modeling*, vol. 53, no. 10, pp. 2571–2586, oct 2013. [Online]. Available: https://doi.org/10.1021%2Fci4003409

[53] Z. Zhou and B. Joós, "Convergence issues in molecular dynamics simulations of highly entropic materials," *Modelling and Simulation in Materials Science and Engineering*, vol. 7, no. 3, pp. 383–395, jan 1999. [Online]. Available: https://doi.org/10.1088%2F0965-0393%2F7%2F3%2F307

[54] S. Genheden, M. Akke, and U. Ryde, "Conformational entropies and order parameters: Convergence, reproducibility, and transferability," *Journal of Chemical Theory and Computation*, vol. 10, no. 1, pp. 432–438, dec 2013. [Online]. Available: https://doi.org/10.1021%2Fct400747s

[55] L. Weaver, M. Grütter, and B. Matthews, "The refined structures of goose lysozyme and its complex with a bound trisaccharide show that the goose-type lysozymes lack a catalytic aspartate residue," *Journal of Molecular Biology*, vol. 245, no. 1, pp. 54–68, jan 1995. [Online]. Available: https://doi.org/10.1016%2Fs0022-2836%2895%2980038-7

[56] J.-D. Pédelacq, B.-S. Rho, C.-Y. Kim, G. S. Waldo, T. P. Lekin, B. W. Segelke, B. Rupp, L.-W. Hung, S.-I. Kim, and T. C. Terwilliger, "Crystal structure of a putative pyridoxine 5/-phosphate oxidase (rv2607) from mycobacterium tuberculosis," *Proteins: Structure, Function, and Bioinformatics*, vol. 62, no. 3, pp. 563–569, dec 2005. [Online]. Available: https://doi.org/10.1002%2Fprot.20824

[57] J. Word, S. C. Lovell, J. S. Richardson, and D. C. Richardson, "Asparagine and glutamine: using hydrogen atom contacts in the choice of side-chain amide orientation 1 1edited by j. thornton," *Journal of Molecular Biology*, vol. 285, no. 4, pp. 1735–1747, jan 1999. [Online]. Available: https://doi.org/10.1006%2Fjmbi.1998.2401

[58] H. J. C. Berendsen, J. P. M. Postma, W. F. van Gunsteren, and J. Hermans, "Interaction models for water in relation to protein hydration," in *The Jerusalem Symposia on Quantum Chemistry and Biochemistry*. Springer Netherlands, 1981, pp. 331–342. [Online]. Available: https://doi.org/10.1007%2F978-94-015-7658-1_21

[59] J. Blaszczyk, Y. Li, H. Yan, and X. Ji, "Crystal structure of unligated guanylate kinase from yeast reveals GMP-induced conformational changes," *Journal of Molecular Biology*, vol. 307, no. 1, pp. 247–257, mar 2001. [Online]. Available: https://doi.org/10.1006%2Fjmbi.2000.4427

[60] D. Seeliger and B. L. de Groot, "Conformational transitions upon ligand binding: Holo-structure prediction from apo conformations," *PLoS Computational Biology*, vol. 6, no. 1, p. e1000634, jan 2010. [Online]. Available: https://doi.org/10.1371%2Fjournal.pcbi.1000634

[61] G. Lange, D. Lesuisse, P. Deprez, B. Schoot, P. Loenze, D. Bénard, J.-P. Marquette, P. Broto, E. Sarubbi, and E. Mandine, "Requirements for specific binding of low affinity inhibitor fragments to the SH2 domain ofpp60src are identical to those for high affinity binding of full length inhibitors," *Journal of Medicinal Chemistry*, vol. 46, no. 24, pp. 5184–5195, nov 2003. [Online]. Available: https://doi.org/10.1021%2Fjm020970s

[62] D. A. Case, T. E. Cheatham, T. Darden, H. Gohlke, R. Luo, K. M. Merz, A. Onufriev, C. Simmerling, B. Wang, and R. J. Woods, "The amber biomolecular simulation programs," *Journal of Computational Chemistry*, vol. 26, no. 16, pp. 1668–1688, 2005. [Online]. Available: https://doi.org/10.1002%2Fjcc.20290

[63] K. Lindorff-Larsen, S. Piana, K. Palmo, P. Maragakis, J. L. Klepeis, R. O. Dror, and D. E. Shaw, "Improved side-chain torsion potentials for the amber ff99sb protein force field," *Proteins: Structure, Function, and Bioinformatics*, pp. NA–NA, 2010. [Online]. Available: https://doi.org/10.1002%2Fprot.22711

[64] T. Darden, D. York, and L. Pedersen, “Particle mesh ewald: An n·log(n) method for ewald sums in large systems,” *The Journal of Chemical Physics*, vol. 98, no. 12, pp. 10 089–10 092, jun 1993. [Online]. Available: https://doi.org/10.1063%2F1.464397

[65] U. Essmann, L. Perera, M. L. Berkowitz, T. Darden, H. Lee, and L. G. Pedersen, “A smooth particle mesh ewald method,” *The Journal of Chemical Physics*, vol. 103, no. 19, pp. 8577–8593, nov 1995. [Online]. Available: https://doi.org/10.1063%2F1.470117

[66] R. J. Loncharich, B. R. Brooks, and R. W. Pastor, “Langevin dynamics of peptides: The frictional dependence of isomerization rates ofN-acetylalanyl-n?-methylamide,” *Biopolymers*, vol. 32, no. 5, pp. 523–535, 1992. [Online]. Available: https://doi.org/10.1002%2Fbip.360320508

[67] Y. Yang, B. Hu, and M. A. Lill, “Analysis of factors influencing hydration site prediction based on molecular dynamics simulations,” *Journal of Chemical Information and Modeling*, vol. 54, no. 10, pp. 2987–2995, oct 2014. [Online]. Available: https://doi.org/10.1021%2Fci500426q

[68] “The pymol molecular graphics system, version 1.8.” [Online]. Available: https://www.schrodinger.com

[69] R. Abel, N. K. Salam, J. Shelley, R. Farid, R. A. Friesner, and W. Sherman, “Contribution of explicit solvent effects to the binding affinity of small-molecule inhibitors in blood coagulation factor serine proteases,” *ChemMedChem*, vol. 6, no. 6, pp. 1049–1066, apr 2011. [Online]. Available: https://doi.org/10.1002%2Fcmdc.201000533

[70] C. R. W. Guimarães and M. Cardozo, “MM-GB/SA rescoring of docking poses in structure-based lead optimization,” *Journal of Chemical Information and Modeling*, vol. 48, no. 5, pp. 958–970, 2008. [Online]. Available: https://doi.org/10.1021%2Fci800004w

[71] A. Kohlmann, X. Zhu, and D. Dalgarno, “Application of MM-GB/SA and WaterMap to SRC kinase inhibitor potency prediction,” *ACS Medicinal Chemistry Letters*, vol. 3, no. 2, pp. 94–99, jan 2012. [Online]. Available: https://doi.org/10.1021%2Fml200222u

[72] D. Xu, M. J. Williamson, and R. C. Walker, “Advancements in molecular dynamics simulations of biomolecules on graphical processing units,” pp. 2–19, 2010. [Online]. Available: https://doi.org/10.1016/s1574-1400(10)06001-9

[73] R. Salomon-Ferrer, A. W. Götz, D. Poole, S. L. Grand, and R. C. Walker, “Routine microsecond molecular dynamics simulations with AMBER on GPUs. 2. explicit solvent particle mesh ewald,” *Journal of Chemical Theory and Computation*, vol. 9, no. 9, pp. 3878–3888, aug 2013. [Online]. Available: https://doi.org/10.1021/ct400314y

[74] B. R. Brooks, C. L. Brooks, A. D. Mackerell, L. Nilsson, R. J. Petrella, B. Roux, Y. Won, G. Archontis, C. Bartels, S. Boresch, A. Caflisch, L. Caves, Q. Cui, A. R. Dinner, M. Feig, S. Fischer, J. Gao, M. Hodoscek, W. Im, K. Kuczera, T. Lazaridis, J. Ma, V. Ovchinnikov, E. Paci, R. W. Pastor, C. B. Post, J. Z. Pu, M. Schaefer, B. Tidor, R. M. Venable, H. L. Woodcock,

X. Wu, W. Yang, D. M. York, and M. Karplus, “CHARMM: The biomolecular simulation program,” *Journal of Computational Chemistry*, vol. 30, no. 10, pp. 1545–1614, jul 2009. [Online]. Available: https://doi.org/10.1002/jcc.21287

[75] J. C. Phillips, R. Braun, W. Wang, J. Gumbart, E. Tajkhorshid, E. Villa, C. Chipot, R. D. Skeel, L. Kalé, and K. Schulten, “Scalable molecular dynamics with NAMD,” *Journal of Computational Chemistry*, vol. 26, no. 16, pp. 1781–1802, 2005. [Online]. Available: https://doi.org/10.1002/jcc.20289

[76] M. J. Abraham, T. Murtola, R. Schulz, S. Páll, J. C. Smith, B. Hess, and E. Lindah, “Gromacs: High performance molecular simulations through multi-level parallelism from laptops to supercomputers,” *SoftwareX*, vol. 1-2, pp. 19–25, 2015.

[77] G. Madhavi Sastry, M. Adzhigirey, T. Day, R. Annabhimoju, and W. Sherman, “Protein and ligand preparation: parameters, protocols, and influence on virtual screening enrichments,” *Journal of Computer-Aided Molecular Design*, vol. 27, no. 3, pp. 221–234, Mar 2013. [Online]. Available: https://doi.org/10.1007/s10822-013-9644-8

[78] H. J. C. Berendsen, J. R. Grigera, and T. P. Straatsma, “The missing term in effective pair potentials,” *The Journal of Physical Chemistry*, vol. 91, no. 24, pp. 6269–6271, nov 1987. [Online]. Available: https://doi.org/10.1021/j100308a038

[79] S. Chatterjee, P. G. Debenedetti, F. H. Stillinger, and R. M. Lynden-Bell, “A computational investigation of thermodynamics, structure, dynamics and solvation behavior in modified water models,” *The Journal of Chemical Physics*, vol. 128, no. 12, p. 124511, mar 2008. [Online]. Available: https://doi.org/10.1063/1.2841127

[80] P. Ghanakota and H. A. Carlson, “Driving Structure-Based Drug Discovery through Cosolvent Molecular Dynamics,” *J. Med. Chem.*, vol. 59, no. 23, pp. 10 383–10 399, 2016.

[81] J. Seco, F. J. Luque, and X. Barril, “Binding site detection and druggability index from first principles,” *J. Med. Chem.*, vol. 52, no. 8, pp. 2363–2371, 2009.

[82] O. Guvench and A. D. MacKerell, “Computational fragment-based binding site identification by ligand competitive saturation,” *PLoS Comput. Biol.*, vol. 5, no. 7, 2009.

[83] C. Y. Yang and S. Wang, “Hydrophobic binding hot spots of Bcl-xL protein-protein interfaces by cosolvent molecular dynamics simulation,” *ACS Med. Chem. Lett.*, vol. 2, no. 4, pp. 280–284, 2011.

[84] Y. S. Tan, D. R. Spring, C. Abell, and C. Verma, “The use of chlorobenzene as a probe molecule in molecular dynamics simulations,” *J. Chem. Inf. Model.*, vol. 54, no. 7, pp. 1821–1827, 2014.

[85] A. Bakan, N. Nevins, A. S. Lakdawala, and I. Bahar, “Druggability Assessment of Allosteric Proteins by Dynamics Simulations in the Presence of Probe Molecules,” *J. Chem. Theory Comput.*, vol. 8, no. 7, pp. 2435–2447, 2012. [Online]. Available: https://doi.org/10.1021/ct300117j

[86] K. W. Lexa and H. A. Carlson, “Full protein flexibility is essential for proper hot-spot mapping,” *J. Am. Chem. Soc.*, vol. 133, no. 2, pp. 200–202, 2011.

[87] E. P. Raman, W. Yu, S. K. Lakkaraju, and A. D. Mackerell, “Inclusion of multiple fragment types in the site identification by ligand competitive saturation (SILCS) approach,” *J. Chem. Inf. Model.*, 2013.

[88] E. P. Raman, S. K. Lakkaraju, R. A. Denny, and A. D. MacKerell, “Estimation of relative free energies of binding using pre-computed ensembles based on the single-step free energy perturbation and the site-identification by Ligand competitive saturation approaches,” *J. Comput. Chem.*, vol. 38, no. 15, pp. 1238–1251, 2017.

[89] M. Xu and M. A. Lill, “Significant enhancement of docking sensitivity using implicit ligand sampling,” *J. Chem. Inf. Model.*, vol. 51, no. 3, pp. 693–706, 2011.

[90] G. Cavallo, P. Metrangolo, R. Milani, T. Pilati, A. Priimagi, G. Resnati, and G. Terraneo, “The halogen bond,” *Chem. Rev.*, vol. 116, no. 4, pp. 2478–2601, 2016.

[91] G. Gerebtzoff, X. Li-Blatter, H. Fischer, A. Frentzel, and A. Seelig, “Halogenation of drugs enhances membrane binding and permeation,” *ChemBioChem*, vol. 5, pp. 676–684, 2004.

[92] C. L. Gentry, R. D. Egleton, T. Gillespie, T. J. Abbruscato, H. B. Bechowski, V. J. Hruby, and T. P. Davis, “The effect of halogenation on blood-brain barrier permeability of a novel peptide drug,” *Peptides*, vol. 20, pp. 1229–1238, 1999.

[93] V. Andrea and H. P., “The Role of Halogen Bonding in Inhibitor Recognition and Binding by Protein Kinases,” *Curr. Top. Med. Chem.*, vol. 7, no. 14, pp. 1336–1348, jan 2007.

[94] Y. Lu, Y. Liu, Z. Xu, H. Li, H. Liu, and W. Zhu, “Halogen bonding for rational drug design and new drug discovery,” *Expert Opin. Drug Discov.*, vol. 7, no. 5, pp. 375–383, 2012.

[95] R. Wilcken, X. Liu, M. O. Zimmermann, T. J. Rutherford, A. R. Fersht, A. C. Joerger, and F. M. Boeckler, “Halogen-Enriched Fragment Libraries as Leads for Drug Rescue of Mutant p53,” *J. Am. Chem. Soc.*, vol. 134, no. 15, pp. 6810–6818, 2012. [Online]. Available: https://doi.org/10.1021/ja301056a

[96] R. Wilcken, M. O. Zimmermann, A. Lange, A. C. Joerger, and F. M. Boeckler, “Principles and applications of halogen bonding in medicinal chemistry and chemical biology,” *J. Med. Chem.*, vol. 56, no. 4, pp. 1363–1388, 2013.

[97] P. Dobeš, J. Řezáč, J. Fanfrlík, M. Otyepka, and P. Hobza, “Semiempirical quantum mechanical method PM6-DH2X describes the geometry and energetics of CK2-inhibitor complexes involving halogen bonds well, while the empirical potential fails,” *J. Phys. Chem. B*, vol. 115, no. 26, pp. 8581–8589, 2011.

[98] L. A. Hardegger, B. Kuhn, B. Spinnler, L. Anselm, R. Ecabert, M. Stihle, B. Gsell, R. Thoma, J. Diez, J. Benz, J. M. Plancher, G. Hartmann, D. W. Banner, W. Haap, and F. Diederich, “Systematic investigation of halogen bonding in protein-ligand interactions,” *Angew. Chemie - Int. Ed.*, 2011.

[99] Schrödinger, LLC, “The {PyMOL} Molecular Graphics System, Version˜1.8,” nov 2015.

[100] M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, G. A. Petersson, H. Nakatsuji, X. Li, M. Caricato, A. V. Marenich, J. Bloino, B. G. Janesko, R. Gomperts, B. Mennucci, H. P. Hratchian, J. V. Ortiz, A. F. Izmaylov, J. L. Sonnenberg, D. Williams-Young, F. Ding, F. Lipparini, F. Egidi, J. Goings, B. Peng, A. Petrone, T. Henderson, D. Ranasinghe, V. G. Zakrzewski, J. Gao, N. Rega, G. Zheng, W. Liang, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, K. Throssell, J. A. Montgomery Jr., J. E. Peralta, F. Ogliaro, M. J. Bearpark, J. J. Heyd, E. N. Brothers, K. N. Kudin, V. N. Staroverov, T. A. Keith, R. Kobayashi, J. Normand, K. Raghavachari, A. P. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, J. M. Millam, M. Klene, C. Adamo, R. Cammi, J. W. Ochterski, R. L. Martin, K. Morokuma, O. Farkas, J. B. Foresman, and D. J. Fox, “Gaussian˜16 {R}evision {B}.01,” 2016.

[101] D. Case, R. Betz, D. Cerutti, I. T.E. Cheatham, T. Darden, R. Duke, T. Giese, H. Gohlke, A. Goetz, N. Homeyer, S. Izadi, P. Janowski, J. Kaus, A. Kovalenko, T. Lee, S. LeGrand, P. Li, C.Lin, T. Luchko, R. Luo, B. Madej, D. Mermelstein, K. Merz, G. Monard, H. Nguyen, H. Nguyen, I.Omelyan, A. Onufriev, D. Roe, A. Roitberg, C. Sagui, C. Simmerling, W. Botello-Smith, J. Swails, R. Walker, J. Wang, R. Wolf, X. Wu, L. Xiao, and P. Kollman, “AMBER 2016,” San Francisco., 2016.

[102] W. L. Jorgensen and P. Schyman, “Treatment of Halogen Bonding in the OPLS-AA Force Field: Application to Potent Anti-HIV Agents,” *J. Chem. Theory Comput.*, vol. 8, no. 10, pp. 3895–3901, 2012.

[103] R. Dennington, T. A. Keith, and J. M. Millam, “Gaussview Version 6,” 2016, semichem Inc. Shawnee Mission KS.

[104] C. J. Woods and J. M. Michel, “Sire Molecular Simulation Framework,” 2017. [Online]. Available: http://siremol.org

[105] C. M. Baker, “Polarizable force fields for molecular dynamics simulations of biomolecules,” *Wiley Interdisciplinary Reviews: Computational Molecular Science*, vol. 5, no. 2, pp. 241–254, jan 2015. [Online]. Available: https://doi.org/10.1002/wcms.1215

[106] J. W. Ponder, C. Wu, P. Ren, V. S. Pande, J. D. Chodera, M. J. Schnieders, I. Haque, D. L. Mobley, D. S. Lambrecht, R. A. DiStasio, M. Head-Gordon, G. N. I. Clark, M. E. Johnson, and T. Head-Gordon, “Current status of the AMOEBA polarizable force field,” *The Journal of Physical Chemistry B*, vol. 114, no. 8, pp. 2549–2564, mar 2010. [Online]. Available: https://doi.org/10.1021/jp910674d

[107] P. Cieplak, J. Caldwell, and P. Kollman, “Molecular mechanical models for organic and biological systems going beyond the atom centered two body additive approximation: aqueous solution free energies of methanol and n-methyl acetamide, nucleic acid base, and amide hydrogen bonding and chloroform/water partition coefficients of the nucleic acid bases,” *Journal of Computational Chemistry*, vol. 22, no. 10, pp. 1048–1057, jul 2001. [Online]. Available: https://doi.org/10.1002/jcc.1065

[108] C. M. Baker, V. M. Anisimov, and A. D. MacKerell, "Development of CHARMM polarizable force field for nucleic acid bases based on the classical drude oscillator model," *The Journal of Physical Chemistry B*, vol. 115, no. 3, pp. 580–596, jan 2011. [Online]. Available: https://doi.org/10.1021/jp1092338

[109] B. Kuhn, W. Guba, J. Hert, D. Banner, C. Bissantz, S. Ceccarelli, W. Haap, M. Körner, A. Kuglstatter, C. Lerner, P. Mattei, W. Neidhart, E. Pinard, M. G. Rudolph, T. Schulz-Gasch, T. Woltering, and M. Stahl, "A real-world perspective on molecular design," *Journal of Medicinal Chemistry*, vol. 59, no. 9, pp. 4087–4102, feb 2016. [Online]. Available: https://doi.org/10.1021/acs.jmedchem.5b01875

[110] C. R. W. Guimarães and A. M. Mathiowetz, "Addressing limitations with the MM-GB/SA scoring procedure using the WaterMap method and free energy perturbation calculations," *Journal of Chemical Information and Modeling*, vol. 50, no. 4, pp. 547–559, apr 2010. [Online]. Available: https://doi.org/10.1021%2Fci900497d

[111] M. A. Lie, R. Thomsen, C. N. S. Pedersen, B. Schiøtt, and M. H. Christensen, "Molecular docking with ligand attached water molecules," *Journal of Chemical Information and Modeling*, vol. 51, no. 4, pp. 909–917, apr 2011. [Online]. Available: https://doi.org/10.1021/ci100510m

[112] T. E. Balius, M. Fischer, R. M. Stein, T. B. Adler, C. N. Nguyen, A. Cruz, M. K. Gilson, T. Kurtzman, and B. K. Shoichet, "Testing inhomogeneous solvation theory in structure-based ligand discovery," *Proceedings of the National Academy of Sciences*, vol. 114, no. 33, pp. E6839–E6846, jul 2017. [Online]. Available: https://doi.org/10.1073/pnas.1703287114

[113] G. Lemmon and J. Meiler, "Towards ligand docking including explicit interface water molecules," *PLoS ONE*, vol. 8, no. 6, p. e67536, jun 2013. [Online]. Available: https://doi.org/10.1371/journal.pone.0067536

[114] S. Uehara and S. Tanaka, "AutoDock-GIST: Incorporating thermodynamics of active-site water into scoring function for accurate protein-ligand docking," *Molecules*, vol. 21, no. 11, p. 1604, nov 2016. [Online]. Available: https://doi.org/10.3390/molecules21111604

APPENDIX

A. WATSITE USER GUIDE WITH PYMOL INTERFACE

WATsite3.0: A Hydration Site Prediction Program with PyMOL Plugin.

User Guide

Ying Yang, Matt Masters, Amr Abdallah, Bingjie Hu, Markus Lill.

Department of Medicinal Chemistry and Molecular Pharmacology
College of Pharmacy, Purdue University
575 Stadium Mall Drive
West Lafayette, IN 47907
Email: mlill@purdue.edu
http://people.pharmacy.purdue.edu/~mlill

Jan 29, 2018

WATsite is a hydration site analysis program developed together with an easy-to-use graphical user interface (GUI) based on PyMOL. WATsite identifies hydration sites from a molecular dynamics simulation trajectory with explicit water molecules. The thermodynamic profile of each hydration site is estimated by computing the enthalpy and entropy of the water molecule occupying a hydration site throughout the simulation. WATsite is available for download at http://people.pharmacy.purdue.edu/~mlill/software/watsite/.
The latest version of WATsite requires a VNIDIA GPU workstation which utilizes the OpenMM [ref] toolkit for GPU-accelerated molecular dynamics simulation. We assume NVIDIA driver and CUDA toolkit have been pre-installed. WATsite2.0, based on Gromacs simulations is available in case a GPU workstation is not available.
The current version of WATsite and the plugin are designed for Linux OS (Redhat / Ubuntu).

When using WATsite please cite the following references:

1. Hu, B.; Lill, M. A., Watsite: Hydration Site Prediction Program with Pymol Interface. J Comput Chem 2014, 35, 1255-60.
2. Yang, Y.; Hu, B.; Lill, M.A., Analysis of Factors Influencing Hydration Site Prediction based on Molecular Dynamics Simulations. J Chem Inf Model 2014, 54, 2987-95.

Contents

1 Changes and updates that have been made to WATsite3.0 3

2 Installation with WATsite docker image 4

2.1 Install docker and nvidia-docker 4

2.2 Create a local watsite container from the image . . 4

3 Installation without docker 6

3.1 Install prerequisites 6

3.2 Install openmm-watsite 6

3.3 Install WATsite3.0 7

4 Running WATsite Analysis with PyMOL plugin 8

4.1 Modify settings based on installation 9

4.2 Step 1: Prepare protein/ligand system 9

4.3 Step 2: Set parameters for MD simulation 12

4.4 Step 3: Perform WATsite analysis using MD trajectory 13

4.5 Step 4: Import WATsite results 14

4.6 Step 5: Estimate protein desolvation free energy for ligand library 16

1 Changes and updates that have been made to WATsite3.0

- GPU-acceleration is implement in OpenMM package [ref].
- Discretized grid for water energy has been implemented.
- Available force fields: Amber99SB, Amber99SBildn, Amber14SB.
- Available water models: SPC/E, TIP3p, TIP4P, TIP4pEW, OPC.

2 Installation with WATsite docker image

A docker image has been created by our group for the convenience of fast installation. Please check the installation of docker and nvidia-docker at this tutorial and nvidia-docker. Here we will only put a list of commands for installation of docker-ce and nvidia-docker2 under ubuntun.

2.1 Install docker and nvidia-docker

1. Install docker-ce

```
curl -fsSL https://download.docker.com/linux/ubuntu/gpg | sudo apt-key add -
sudo add-apt-repository "deb [arch=amd64] \
    https://download.docker.com/linux/ubuntu $(lsb_release -cs) stable"
sudo apt-get update
sudo apt-get install -y docker-ce
```

2. Install nvidia-docker2

```
curl -s -L https://nvidia.github.io/nvidia-docker/gpgkey | sudo apt-key add -
curl -s -L https://nvidia.github.io/nvidia-docker/ubuntu16.04/amd64/nvidia-
                 docker.list | sudo tee /etc/apt/sources.list.d/nvidia-docker.list
sudo apt-get update
sudo apt-get install -y nvidia-docker2
sudo pkill -SIGHUP dockerd
```

2.2 Create a local watsite container from the image

1. Obtain watsite docker image from docker hub
2. Create a local container from the watsite image

```
sudo docker load < watsite_docker.tar.gz
```

3. Get X authentication in order to use display for PyMOL plugin

```
xauth list
```

4. Run WATsite within the container

```
sudo nvidia-docker run -it –net=host -e DISPLAY –runtime=nvidia -v /tmp/.x11-
        unix -v /scratch/yang570:/opt/data -it watsite3.0 bash
xauth add <output from xauth list>
```

3 Installation without docker

Here we descibe the installation steps for all required programs.

3.1 Install prerequisites

1. Install anaconda

```
wget https://repo.continuum.io/archive/Anaconda2-5.0.1-Linux-x86_64.sh
bash Anaconda2-5.0.1-Linux-x86_64.sh -b -p /usr/local/anaconda
```

2. Install ambertools with conda

```
conda install ambertools=17 -c http://ambermd.org/downloads/ambertools/conda/
```

3. Install PyMOL

```
apt-get install pymol
```

3.2 Install openmm-watsite

The WATsite-compatible OpenMM has to be compiled in order to perform MD simulation and generate water interaction energies for later analysis. The source code is located in the openmm-watsite folder, and the user can follow the standard OpenMM compilation steps.

1. Download openmm-watsite from github

```
git clone https://github.com/mlill/openmm-watsite-siamang.git
```

2. Go to *openmm-watsite* directory, and make a new directory *build*

```
cd openmm-watsite
mkdir build && cd build
```

3. Set environment variable to the correct CUDA toolkit path:

```
export PATH=$PATH:/usr/local/cuda-8.0/bin/
export LD_LIBRARY_PATH=/usr/local/cuda-8.0/lib64:$LD_LIBRARY_PATH
export CUDA_HOME=/usr/local/cuda-8.0/
export OPENMM_CUDA_COMPILER=/usr/local/cuda-8.0/bin/nvcc
```

4. Configure with ccmake:

```
ccmake ../
press “c”
```

5. Set the variable CMAKE_INSTALL_PREFIX to the location where you want to install OpenMM
6. Set the variable PYTHON_EXECUTABLE to the Python interpreter you plan to use OpenMM with. (In case other version of OpenMM has been installed, create a new python environment following this link.)
7. Configure (press “c”) again.
8. Generate the Makefile (press “g”).
9. Start the installation (if the location of installation is not a system area, sudo is not required)

```
(sudo) make install
(sudo) make PythonInstall
```

10. Verify your installation

```
python -m simtk.testInstallation
```

3.3 Install WATsite3.0

1. Download WATsite3.0 from github

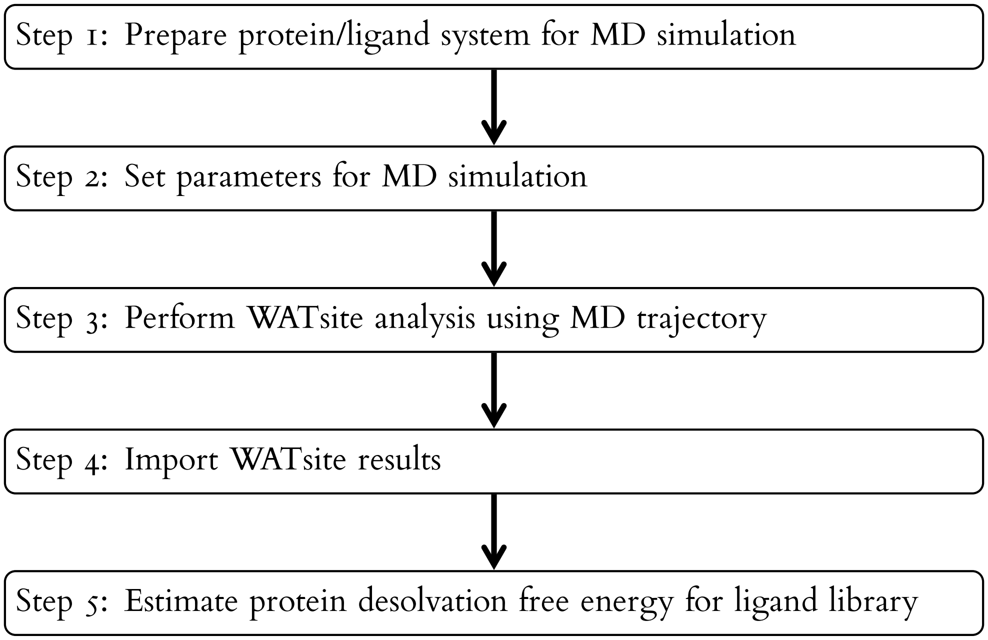
```
git clone https://github.com/mlill/watsite_collaboration.git
```

2. Go to *WATsite3.0* directory and compile with make

```
cd WATsite3.0
make -f makefile
```

4 Running WATsite Analysis with PyMOL plugin

Workflow of WATsite



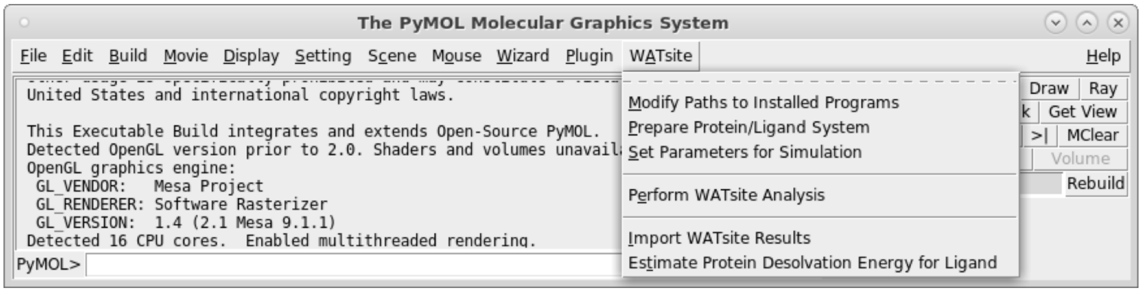
Step 1: Prepare protein/ligand system for MD simulation

Step 2: Set parameters for MD simulation

Step 3: Perform WATsite analysis using MD trajectory

Step 4: Import WATsite results

Step 5: Estimate protein desolvation free energy for ligand library



The PyMOL Molecular Graphics System
File
Edit
Build
Movie
Display
Setting
Scene
Mouse
Wizard
Plugin
WATsite
Help
Modify Paths to Installed Programs
Prepare Protein/Ligand System
Set Parameters for Simulation
Perform WATsite Analysis
Import WATsite Results
Estimate Protein Desolvation Energy for Ligand
Draw
Ray
Get View
>|
MClear
Volume
Rebuild

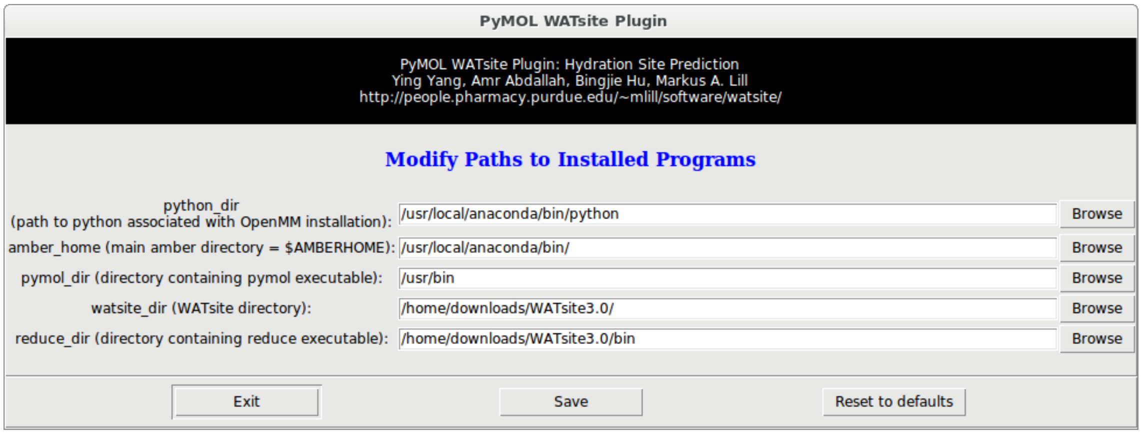
```
United States and international copyright laws.

This Executable Build integrates and extends Open-Source PyMOL.
Detected OpenGL version prior to 2.0. Shaders and volumes unavail
OpenGL graphics engine:
 GL_VENDOR:   Mesa Project
 GL_RENDERER: Software Rasterizer
 GL_VERSION:  1.4 (2.1 Mesa 9.1.1)
Detected 16 CPU cores.  Enabled multithreaded rendering.
PyMOL>
```

Figure 1: WATsite PyMOL plugin menu.

4.1 Modify settings based on installation

This step can be **omitted** if the user choose the installation with docker.
The location paths to the required molecular modeling programs need to be correctly specified.The settings only need to be modified for the first time, and will be automatically read in subsequent sessions.

1. Select the menu item "Modify Paths to Installed Programs" from the WATsite menu (Figure 1).
2. Specify or modify the paths to the programs according to their installation (Figure 2).
3. The location of WATsite directory should be defined for watsite_home. Similarly the python path associated with OpenMM, the path to AmberTools, PyMOL, and Reduce installation need to be specified.



PyMOL WATsite Plugin

PyMOL WATsite Plugin: Hydration Site Prediction
Ying Yang, Amr Abdallah, Bingjie Hu, Markus A. Lill
http://people.pharmacy.purdue.edu/~mlill/software/watsite/

Modify Paths to Installed Programs

python_dir (path to python associated with OpenMM installation): /usr/local/anaconda/bin/python
amber_home (main amber directory = $AMBERHOME): /usr/local/anaconda/bin/
pymol_dir (directory containing pymol executable): /usr/bin
watsite_dir (WATsite directory): /home/downloads/WATsite3.0/
reduce_dir (directory containing reduce executable): /home/downloads/WATsite3.0/bin

Figure 2: specify the correct paths to the installed program

4.2 Step 1: Prepare protein/ligand system

1. The user can specify a protein structure from a file or a structure already displayed in the current PyMOL session (Figure 3). The user can choose to perform protonation site analysis using Reduce[ref], or use a structure with previously predicted protonation states.
2. To define the protein binding site, the user needs to provide a ligand molecule positioned within the binding site or a "pseudo-ligand" using binding site residues. A margin (in Å) will be

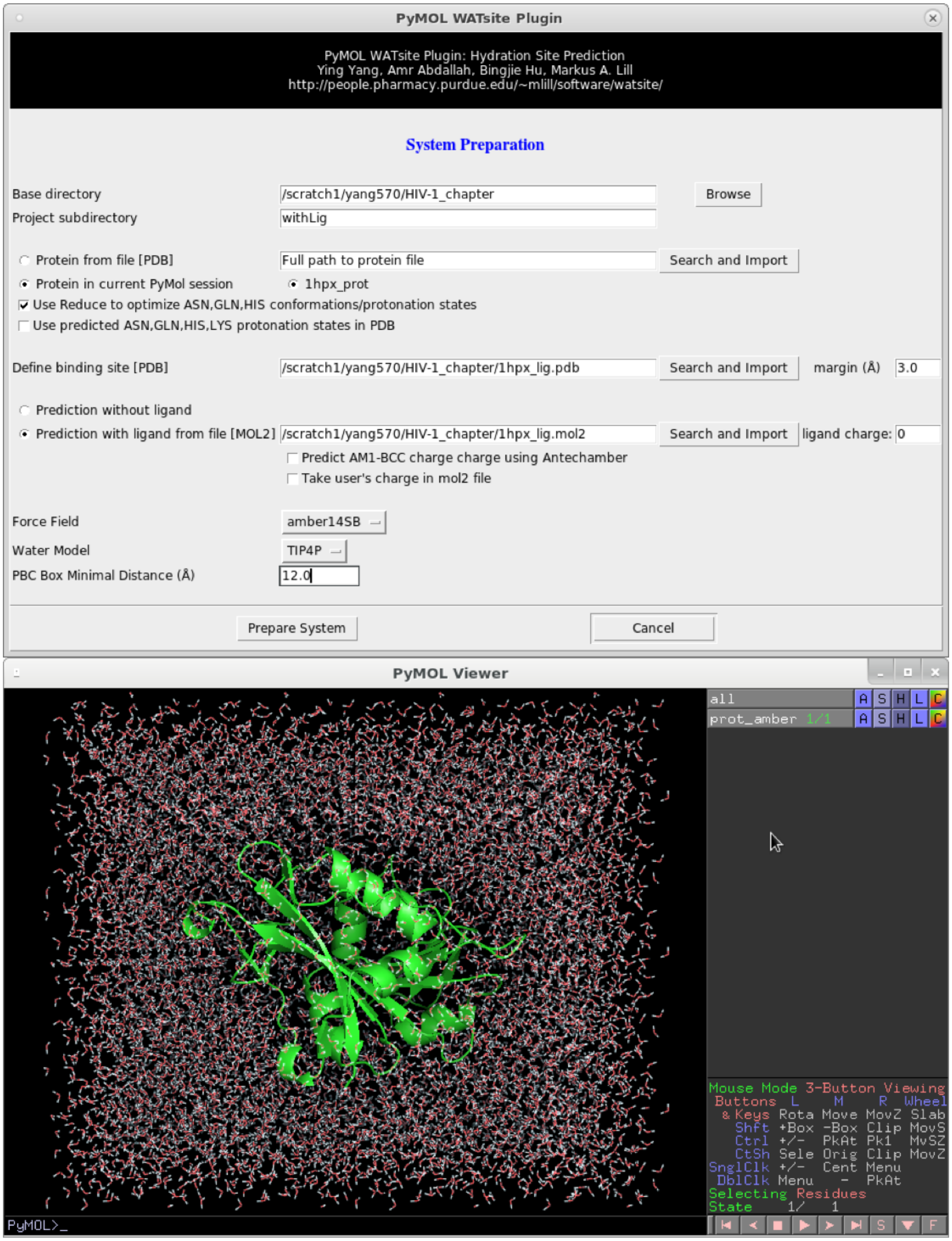
used to define the binding pocket specifying a box surrounding the ligand/pseudo-ligand. The minimum distance between any ligand heavy atom to the edge of the box equals to the margin value.

3. Hydration sites can be predicted for both ligand-free (apo) and ligand-bound (holo) protein structures. If the user intends to predict hydration sites at the interface between the protein and the ligand, the file location of the bound ligand is specified within the PyMOL plugin. The user also needs to specify the net charge of the ligand, and choose the partial charge method. The atom types[ref] for the specified ligand will be assigned by antechamber[ref] and will be included in the MD simulation and the following hydration site identification process. The current version does not have docking service for the user-specified ligand. Therefore, the provided ligand conformation needs to be a meaningful binding pose for the protein.

4. The force field and water models for the system preparation need to be chosen. Currently, three choices of different amber force fields and five water models have been tested.

force fields:	amber99SB, amber99SBildn, amber14SB
water models:	SPC\E, TIP3P, TIP4P, TIP4pEW, OPC

5. The system will be solvated in an orthorhombic water box. The user can also control the box size by specifying the minimum distance between any protein atom to the edge of the box. Lastly chloride and sodium ions will be added to neutralize the system. The prepared protein system will then be loaded into PyMOL (Figure 3).



PyMOL WATsite Plugin: Hydration Site Prediction
Ying Yang, Amr Abdallah, Bingjie Hu, Markus A. Lill
http://people.pharmacy.purdue.edu/~mlill/software/watsite/

System Preparation

Base directory: /scratch1/yang570/HIV-1_chapter [Browse]
Project subdirectory: withLig

○ Protein from file [PDB]: Full path to protein file [Search and Import]
● Protein in current PyMol session: ● 1hpx_prot
☑ Use Reduce to optimize ASN,GLN,HIS conformations/protonation states
□ Use predicted ASN,GLN,HIS,LYS protonation states in PDB

Define binding site [PDB]: /scratch1/yang570/HIV-1_chapter/1hpx_lig.pdb [Search and Import] margin (Å) 3.0

○ Prediction without ligand
● Prediction with ligand from file [MOL2]: /scratch1/yang570/HIV-1_chapter/1hpx_lig.mol2 [Search and Import] ligand charge: 0
□ Predict AM1-BCC charge charge using Antechamber
□ Take user's charge in mol2 file

Force Field: amber14SB
Water Model: TIP4P
PBC Box Minimal Distance (Å): 12.0

[Prepare System] [Cancel]

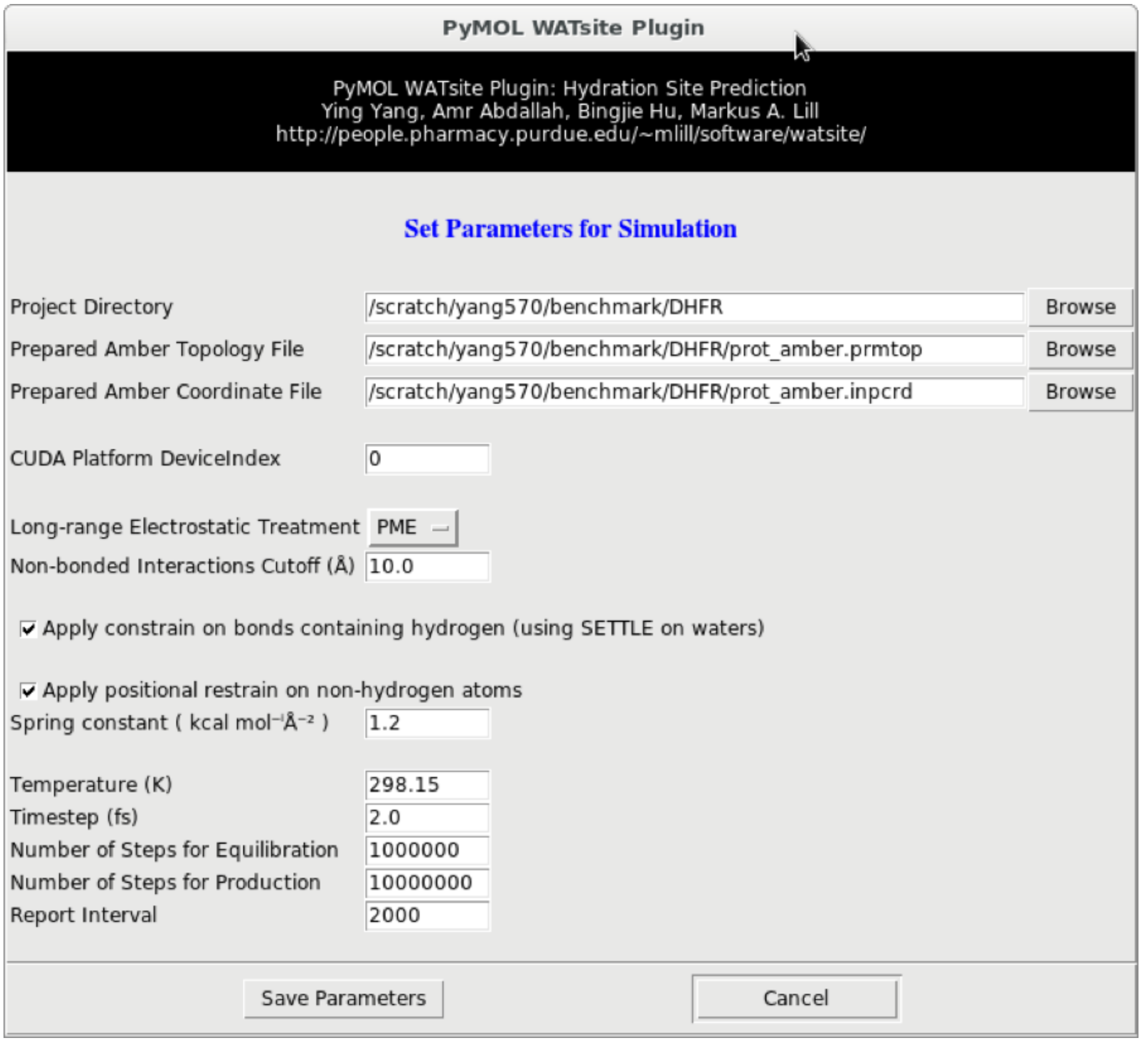
Figure 3: Prepare the protein or protein-ligand system.

4.3 Step 2: Set parameters for MD simulation

In this step, users can change parameters for the MD simulation which subsequently will be used for hydration site analysis (Figure 4).

1. The default amber topology (prot_amber.prmtop) and coordinate (prot_amber.inpcrd) files are generated from the system preparation step. The users may also choose the amber files from their own preparation.
2. The CUDA device index needs to be entered based on the user's workstation.
3. Atom pairs within the user specified cutoff will be used to calculate the direct/exact non-bonded interactions. Long-range electrostatic interactions from atom pairs beyond the cutoff can be treated using one of the three methods (PME, Ewald, or NoCutoff). If NoCutoff is chosen, the cutoff distance will be ignored.
4. By default, we apply constraints on the length of all bonds involving a hydrogen atom, and make water molecules rigid. This will allow us to run simulations with an integration timestep of 2 fs. However, the user can disable the constrain and rigid water by uncheck the box, and change to a smaller timestep accordingly.
5. User can also change the strength of positional restraint on the non-hydrogen atoms during the simulation.
6. Temperature and number of steps for the equilibration and production can also be modified. For the choice of equilibration and production simulation length, our previous studies[ref] showed that at least a 4 ns production simulation is required to obtain reliable prediction of the locations and thermodynamic properties of hydration sites.
7. Once the preparation has been finished, the user needs to change into the project directory and start the OpenMM simulation:

```
nohup ./run_omm_watsite.sh &
```



PyMOL WATsite Plugin

PyMOL WATsite Plugin: Hydration Site Prediction
Ying Yang, Amr Abdallah, Bingjie Hu, Markus A. Lill
http://people.pharmacy.purdue.edu/~mlill/software/watsite/

Set Parameters for Simulation

Project Directory	/scratch/yang570/benchmark/DHFR
Prepared Amber Topology File	/scratch/yang570/benchmark/DHFR/prot_amber.prmtop
Prepared Amber Coordinate File	/scratch/yang570/benchmark/DHFR/prot_amber.inpcrd
CUDA Platform DeviceIndex	0
Long-range Electrostatic Treatment	PME
Non-bonded Interactions Cutoff (Å)	10.0

☑ Apply constrain on bonds containing hydrogen (using SETTLE on waters)

☑ Apply positional restrain on non-hydrogen atoms

Spring constant (kcal mol^{-1}Å^{-2})	1.2
Temperature (K)	298.15
Timestep (fs)	2.0
Number of Steps for Equilibration	1000000
Number of Steps for Production	10000000
Report Interval	2000

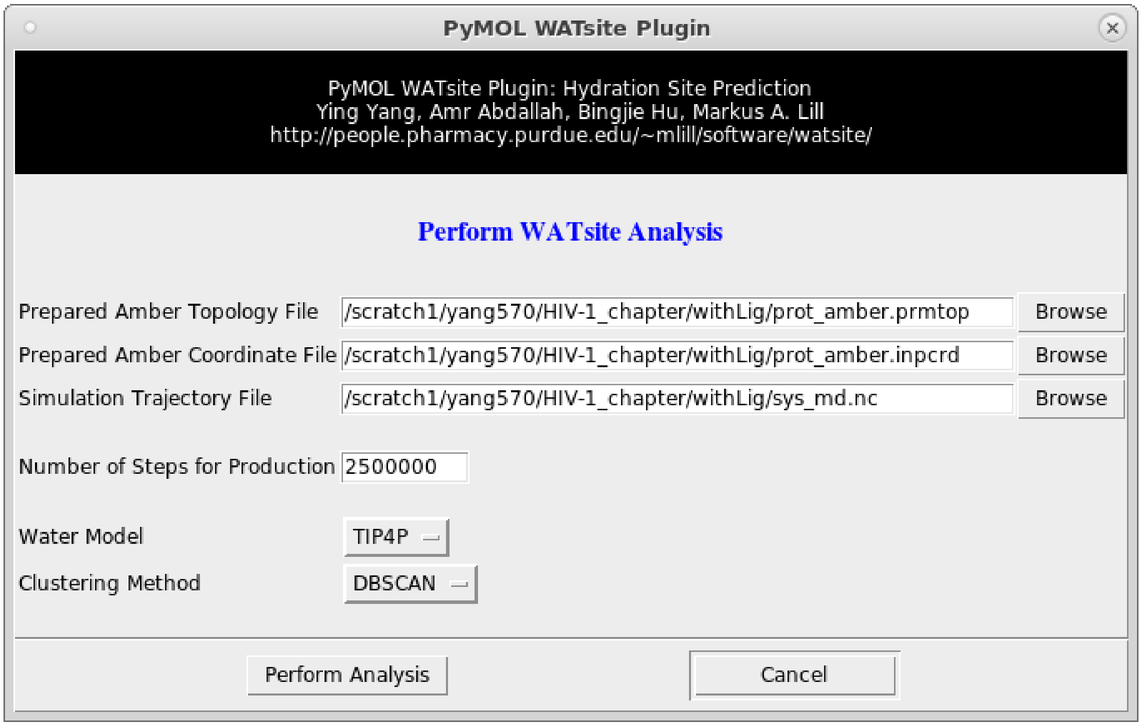
Figure 4: Set parameters for running the MD simulation via OpenMM.

4.4 Step 3: Perform WATsite analysis using MD trajectory

1. In step 3, we will perform WATsite analysis on the trajectory file generated in the previous step (Figure 5). The amber topology (prot_amber.prmtop) and coordinate (prot_amber.inpcrd) files generated in 4.2, as well as the trajectory (sys_md.nc) file generated in 4.3 should be specified correctly.

2. The number of steps and water model used for WATsite analysis need to be identical to those used during the production simulation. Clustering algorithm used to predict hydration site locations from water density can also be chosen.

3. During the WATsite analysis step, the production trajectory will first be aligned to the reference which is the user input protein structure, and saved into pdb format. Then, WATsite analysis will be performed for predicting hydration site location based on water density analysis, calculating entropy and enthalpy.



PyMOL WATsite Plugin: Hydration Site Prediction
Ying Yang, Amr Abdallah, Bingjie Hu, Markus A. Lill
http://people.pharmacy.purdue.edu/~mlill/software/watsite/

Perform WATsite Analysis

Prepared Amber Topology File /scratch1/yang570/HIV-1_chapter/withLig/prot_amber.prmtop

Prepared Amber Coordinate File /scratch1/yang570/HIV-1_chapter/withLig/prot_amber.inpcrd

Simulation Trajectory File /scratch1/yang570/HIV-1_chapter/withLig/sys_md.nc

Number of Steps for Production 2500000

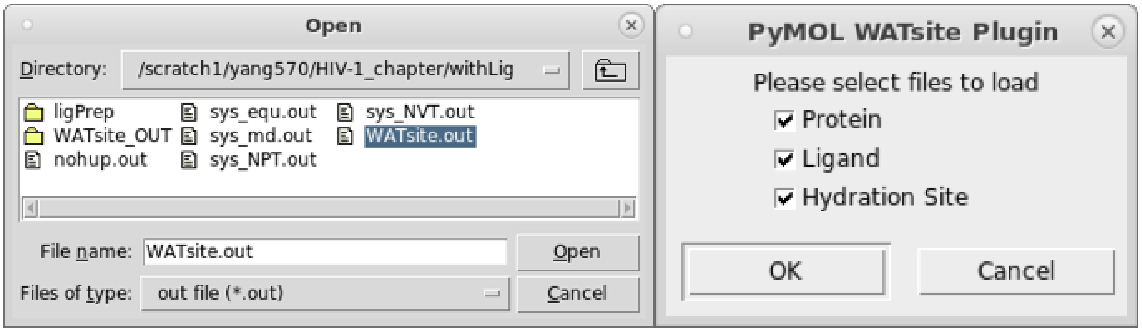
Water Model TIP4P

Clustering Method DBSCAN

Figure 5: Perform WATsite analysis.

4.5 Step 4: Import WATsite results

- After completion of WATsite analysis, we can import the results through the "Import WATsite Results" command under the WATsite menu, and select the "WATsite.out" file which stores the directory to the location of the prediction results (Figure 6).
- Here, we want to investigate water molecules at the binding interface between protein and ligand, so we select 'Protein', 'Ligand', and 'Hydration Site' to load into PyMOL.



Open

Directory: /scratch1/yang570/HIV-1_chapter/withLig

File name: WATsite.out

Files of type: out file (*.out)

PyMOL WATsite Plugin

Please select files to load

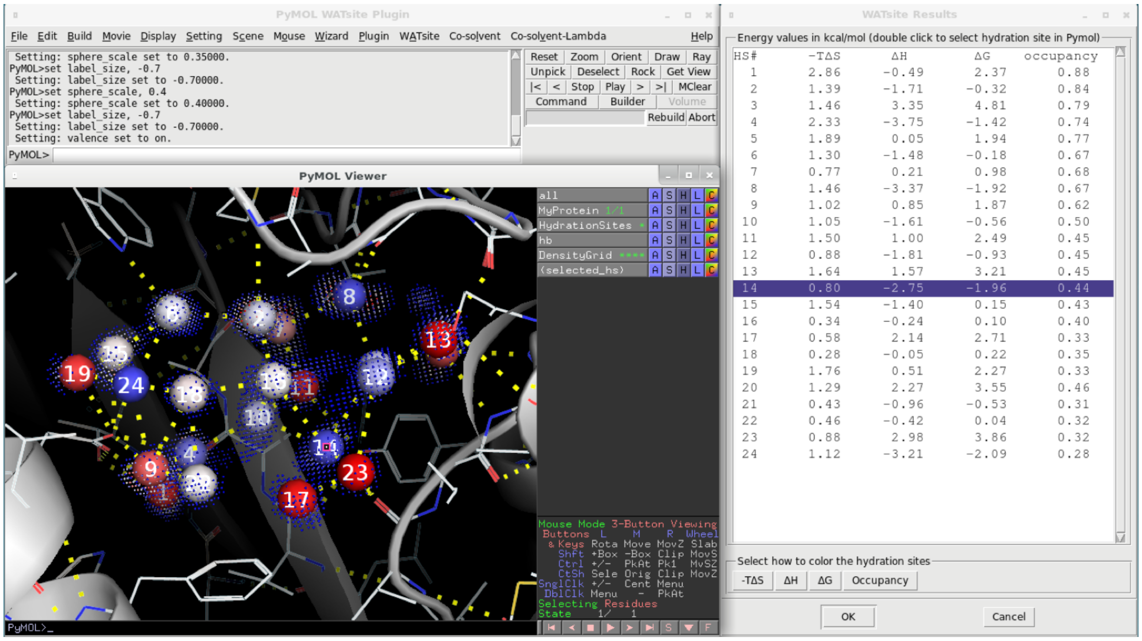
☑ Protein

☑ Ligand

☑ Hydration Site

Figure 6: Import WATsite results.

- The result of the example case of HIV-1 protease are shown in Figure 7



PyMOL WATsite Plugin

File Edit Build Movie Display Setting Scene Mouse Wizard Plugin WATsite Co-solvent Co-solvent-Lambda Help
Reset Zoom Orient Draw Ray
Unpick Deselect Rock Get View
|< < Stop Play > >| MClear
Command Builder Volume
Rebuild Abort

```
 Setting: sphere_scale set to 0.35000.
PyMOL>set label_size, -0.7
 Setting: label_size set to -0.70000.
PyMOL>set sphere_scale, 0.4
 Setting: sphere_scale set to 0.40000.
PyMOL>set label_size, -0.7
 Setting: label_size set to -0.70000.
 Setting: valence set to on.
PyMOL>
```

PyMOL Viewer

all
MyProtein 1/1
HydrationSites
hb
DensityGrid
(selected_hs)
8
13
19
24
11
12
10
14
9
23
17
PyMOL>_

WATsite Results

Energy values in kcal/mol (double click to select hydration site in Pymol)

HS#	−TΔS	ΔH	ΔG	occupancy
1	2.86	−0.49	2.37	0.88
2	1.39	−1.71	−0.32	0.84
3	1.46	3.35	4.81	0.79
4	2.33	−3.75	−1.42	0.74
5	1.89	0.05	1.94	0.77
6	1.30	−1.48	−0.18	0.67
7	0.77	0.21	0.98	0.68
8	1.46	−3.37	−1.92	0.67
9	1.02	0.85	1.87	0.62
10	1.05	−1.61	−0.56	0.50
11	1.50	1.00	2.49	0.45
12	0.88	−1.81	−0.93	0.45
13	1.64	1.57	3.21	0.45
14	0.80	−2.75	−1.96	0.44
15	1.54	−1.40	0.15	0.43
16	0.34	−0.24	0.10	0.40
17	0.58	2.14	2.71	0.33
18	0.28	−0.05	0.22	0.35
19	1.76	0.51	2.27	0.33
20	1.29	2.27	3.55	0.46
21	0.43	−0.96	−0.53	0.31
22	0.46	−0.42	0.04	0.32
23	0.88	2.98	3.86	0.32
24	1.12	−3.21	−2.09	0.28

Select how to color the hydration sites

Figure 7: Hydration site result predicted with the presence of ligand.

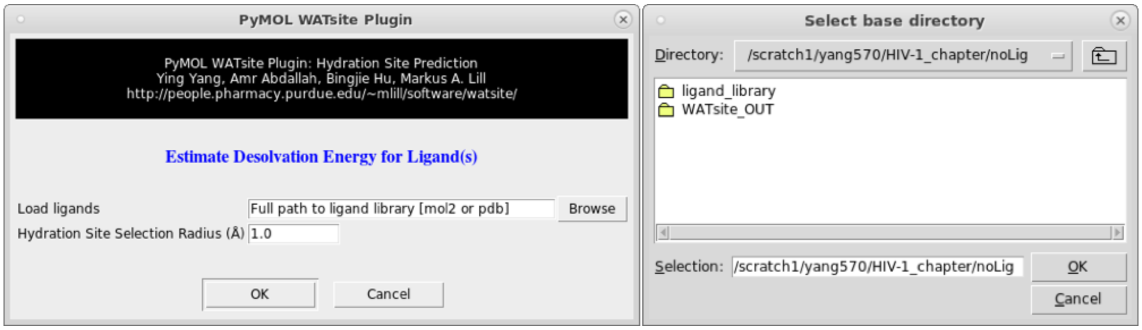
- The PyMOL viewer window shows the predicted hydration sites in the protein binding site. The hydration sites are shown as spheres and colored in this example based on their ΔG values in a blue-white-red spectrum where blue indicates relatively low ΔG values and red indicates relatively high ΔG values.
- A hydration site with a more positive ΔG value (darker red) indicates an unfavorable environment of the water molecule in

the binding site. Therefore, a gain in free energy of binding can be expected if the water in that hydration site is replaced by a ligand.

- The "occupancy" values indicate the probability a water molecule is observed in the given hydration site during the MD simulation.
- The "WATsite results" window listing the estimated desolvation free energy (ΔG), enthalpy (ΔH), entropy ($-T\Delta S$), and occupancy for each hydration site. The user can also choose according to which descriptor the hydration sites are colored by clicking the corresponding ΔG, $-T\Delta S$, and ΔH, or "Occupancy" button.

4.6 Step 5: Estimate protein desolvation free energy for ligand library

- The user can perform hydration site prediction with the ligand removed from the protein binding site. This method can be useful to compare and evaluate the different protein desolvation free energies from a congeneric series of ligands .
- the directory containing all ligands of interest as well as the radius/cutoff used to select the displaced hydration sites need to be specified (Figure 8).



PyMOL WATsite Plugin

PyMOL WATsite Plugin: Hydration Site Prediction
Ying Yang, Amr Abdallah, Bingjie Hu, Markus A. Lill
http://people.pharmacy.purdue.edu/~mlill/software/watsite/

Estimate Desolvation Energy for Ligand(s)

Load ligands: Full path to ligand library [mol2 or pdb]

Hydration Site Selection Radius (Å): 1.0

Select base directory

Directory: /scratch1/yang570/HIV-1_chapter/noLig

ligand_library
WATsite_OUT

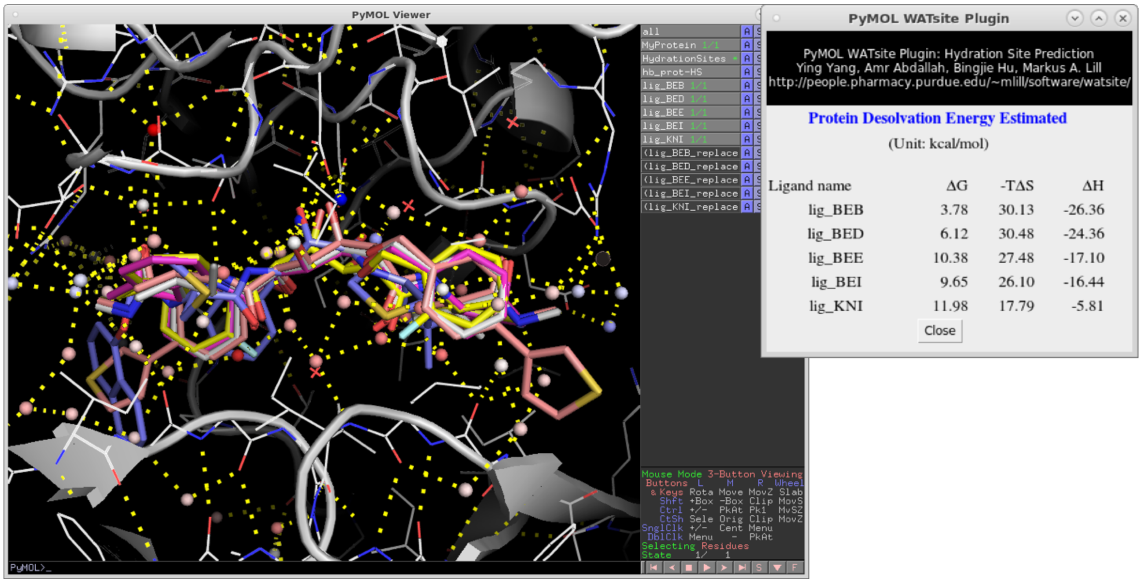
Selection: /scratch1/yang570/HIV-1_chapter/noLig

Figure 8: Estimate protein desolvation free energy for ligand library.

- For each ligand in the directory, the free energies of hydration sites that are within the user-specified distance to any of the ligand's heavy atoms are added up to estimate its protein desolvation free energy. A more positive value means a more favorable contribution to the protein-ligand binding free energy.
- The predicted desolvation energies (ΔG, $-T\Delta S$, and ΔH) are

displayed in a new window, and the selection of displaced hydration sites is highlighted in the PyMOL viewer (Figure 9).

- The result of the example case of HIV-1 protease are shown in Figure 9



PyMOL WATsite Plugin: Hydration Site Prediction
Ying Yang, Amr Abdallah, Bingjie Hu, Markus A. Lill
http://people.pharmacy.purdue.edu/~mlill/software/watsite/

Protein Desolvation Energy Estimated

(Unit: kcal/mol)

Ligand name	ΔG	-TΔS	ΔH
lig_BEB	3.78	30.13	-26.36
lig_BED	6.12	30.48	-24.36
lig_BEE	10.38	27.48	-17.10
lig_BEI	9.65	26.10	-16.44
lig_KNI	11.98	17.79	-5.81

Figure 9: Hydration site result predicted without the ligand.

VITA

VITA

Ying Yang was born in Chengdu, Sichuan, China on September 25th, 1990. She is the only daughter of Yuan Yang and Min Li.

After graduating from Chengdu Experimental Foreign Language School (CEFLS) in 2008, Ying went to China Agricultural University in Beijing where she joined a ”2+2” program with Colorado State University. She earned dual bachelor’s degree in Biomedical Sciences at Colorado State University and Veterinary Medicine at China Agricultural University with honor.

In 2012, Ying moved to Purdue University pursue her doctoral degree in the Department of Medicinal Chemistry and Molecular Pharmacology in Markus A. Lill’s laboratory. Her research involves computational modeling of protein-ligand binding with a focus on explicit de-solvation and protein flexibility. During her graduate study, she has received the Ross Fellowship, PRF research grant support from Purdue, and the Paget Travel Award from the department of MCMP. She has published two journal articles, two book chapters, and another two manuscripts in preparation.

After her PhD studies, Ying looks forward to continuing working in computational drug discovery field. She has accepted a postdoctoral position under the supervision of Dr. Brian Shoichet at University of California at San Francisco. There she will work on a mix of methods development and testing in model systems, and application to a important pharmacological target, like a GPCR.