Coupling molecular dynamics and machine learning to accelerate the design of bioinspired materials

Joshua K. Smith

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2019

Reading Committee:

Shaoyi Jiang, co-Chair

Jim Pfaendtner, co-Chair

Dave Beck

Program Authorized to Offer Degree:

Chemical Engineering

University of Washington

Abstract

Coupling molecular dynamics and machine learning
to accelerate the design of bioinspired materials

Joshua K. Smith

Chairs of the Supervisory Committee:
Professor Shaoyi Jiang and Associate Professor Jim Pfaendtner
Department of Chemical Engineering

Biological systems, including the human body, are directed by networks of exquisitely selective chemical interactions. Following the paradigm of bioinspired materials design, scientists and engineers aim to understand the rules of natural selectivity and to design materials that can achieve specific functions in complex biological environments without unintended consequences. Knowledge transfer from natural to synthetic design is limited by our understanding of specific and nonspecific interactions in nature at the molecular scale. Molecular dynamics (MD) simulations provide a means to study biophysical systems with atomistic detail. Machine learning (ML) techniques provide a means to synthesize meaningful insights from the deluge of high-dimensional data generated by MD. In this dissertation, I explore a range of important biophysical phenomena with MD and ML, yielding insights with unprecedented resolution and interpretability. Notably, I apply advanced simulation and unsupervised learning techniques to describe the molecular characteristics of nonfouling biomaterials and to guide the design of next generation renal replacement therapies for chronic kidney disease. This work exemplifies the potential for MD and ML in combination to accelerate bioinspired materials design.

Table of Contents

List of Figures

List of Tables

Acknowledgements

Every PhD is a group project. Enumerable people have contributed to my academic and personal journey, and I'd like to take this opportunity to inadequately thank a few of them.

Thanks to my friends in the Jiang and Pfaendtner research groups for your comradery and motivation. Your excellent work and curious minds have motivated and inspired me. Thanks to the PRG elders Vance Jaeger, Blake Hough, and Kayla Sprenger for teaching me the tools of the trade. Special thanks to Wesley Beckner and Arushi Prakash who have been my partners in science and great friends from my first day in PRG. Thanks also to Wesley Beckner, Fernando Centurion, Brian Gerwe, Kelly Carpenter, Willy Voje, and Jac Clark for sharing the cost of shelter and for remaining friends through the trials and tribulations of roommatehood. Grad school provides plenty of challenges without domestic disputes, so I was grateful to live with a group of considerate friends.

Thanks to Adam Richie-Halford, Ariel Rokem, and Jason Yeatman for allowing me to be a part of the AFQBrowser project. Your commitment to open source software development and reproducible research profoundly impacted my ethos as a researcher. Thanks to Skip Rochefort, Phillip Harding, Ariel Rokem, and Dave Beck for providing lending your recommendations in support of my numerous failed attempts to secure research funding and my successful attempt to secure employment. I appreciate your willingness to help and the valuable time you spent to do so.

Prior to graduate school, I had very little experience conducting scientific research. Over the past 5 years I have benefitted tremendously from the tutelage of two experts in the fields of biomaterials and biophysics, my advisers Shaoyi Jiang and Jim Pfaendtner. Thanks to Shaoyi for encouraging me to focus on the big picture, in spite of the constant temptation to obsess over the minutiae of atomistic simulations. Your advice has consistently improved my ability to communicate my work to scientists and engineers with diverse technical backgrounds, a skill that will undoubtedly serve me well for the duration of my scientific career. Thanks to Jim for conferring to me a small sliver of your knowledge in the mystical realm of biophysical simulations and for pushing me to pursue data science. Your expressions of pride and optimism about my work have preserved my mental health and made my grad school experience enjoyable. I am indebted to you for steering me into a field that I am truly excited about, and for playing an instrumental role in me securing an excellent job straight out of grad school. I hope that you will accept repayment by the pint in gradual installments.

Thanks to my family for nurturing and supporting me in all endeavors. Many thanks to my father, Sean, who served as a loving coach, teacher, and role model until his passing during my second year in graduate school. I am forever grateful for the time we spent together and will always strive to make you proud. Thank you to my mom, Tami, for instilling in me an appreciation for the value of education, and for continuing to support me in grad school through the most tumultuous time of her life. I am in awe of your strength and so proud to be your son. Thanks to my younger brother, Spencer, for understanding all of my movie references and reliably serving up laughs. I'm happy to consider you one of my best friends and I am so proud of the man you have become. Thank you to my grandparents, aunts, uncles, and cousins for their love and admiration. Your attendance at soccer games, birthday parties, award ceremonies, and graduations has served as a constant reminder of your support and has motivated me to do great things.

Lastly, I'd like to thank Crissy Nogoy for her love and support over the last few years. Thank you for bearing some of the emotional weight I brought home when deadlines loomed. Thank you for listening patiently as I struggled through practice talks for my exams and job interview. Thank you for being the shield that protected me from the realm of slovenly habits. The Long Night has ended, and I look forward to sharing with you my life after grad school.

Dedication

for Mom and Dad, who are more responsible than myself for the enumerable blessings in my life.

Chapter 1
Introduction

Life is the manifestation of complex chemistries, enabled by exquisitely selective biomolecular interactions. Over the course of billions of years Nature has developed chemical strategies for directing biological systems, with biomolecules that faithfully identify their native binding partners and avoid spurious interactions. The delicate balance between specific and nonspecific interactions poses a difficult challenge for human intervention in ailing biological systems, such as the diseased human body. Synthetic materials for biomedical applications must often replace the functionality of highly optimized, specific chemical interaction networks that have been damaged or eliminated by disease without triggering any number of unintended biological cascades that could result from relatively weak nonspecific interactions.

The paradigm of bioinspired materials design, where synthetic materials are engineered to mimic natural chemistries, has proven effective in developing artificial solutions to challenges in biomedicine. Without the luxury of billions of years for trial and error design, scientists and engineers have resorted to copying Nature's biochemical notes for a head start on biomaterials design. A basic understanding of specific and nonspecific interactions between proteins has fueled advances in controlled drug delivery and in the development of biocompatible medical devices.[1,2] Insights into protein interactions with small molecules have led to accelerated drug design, for ailments ranging from bacterial infection to HIV.[3,4] Designing the next generation of multifunctional biomaterials will require a detailed understanding of the mechanistic underpinnings of specific and nonspecific interactions between biomolecules, especially at the molecular scale.

Molecular dynamics (MD) is an especially useful scientific tool for understanding the relationship between the chemical properties and the behavior of biophysical systems. With MD simulations, a computational scientist can probe the molecular interactions in a biophysical system with atomistic resolution. Nanoscale phenomena observed with MD can be used to predict or explain macroscopic phenomena observed with experiments. An intimate understanding of the link between material properties and behavior in a particular biophysical system can be exploited in the rational design of functional biomaterials. Recent advances in computer hardware and simulation methodology have enabled MD simulations of important biochemical processes, such as protein folding and protein-ligand binding, that were formerly intractable due to their relative long timescales.[5–8] At the same time, the system size accessible to MD simulations has soared to the scale of billions of atoms.[9] While direct simulation of long-timescale biochemical processes can provide an intimate understanding of biological interactions, it can also generate massive amounts of computational data. Clever strategies for data analysis are necessary for extracting interpretable conclusions from mountains of noisy molecular data.

Machine learning (ML) has become an increasingly popular tool in the field of biophysics for drawing interpretable insights from the deluge of data generated by MD simulations. The ML methods with the broadest appeal to the MD community have proven to be those under the umbrellas of dimensionality reduction and unsupervised learning. Dimensionality reduction techniques can be used to project the extremely high-dimensional data from MD onto comprehensible, low dimensional spaces.[10] Information-rich features that describe important structural rearrangements in a biophysical system can be constructed using feature extraction techniques like principal component analysis (PCA).[11–13] The molecular order parameters that discriminate between important states in a biochemical process can be identified with feature selection.[7] Unsupervised learning techniques can be used to identify reoccurring molecular structures that might explain key steps in a biophysical

process.[14,15] The interpretability afforded by ML processing of MD data facilitates digestible insights that can be practically applied for rational biomaterials design.

In this dissertation, I describe how I have used MD simulations and ML analysis methods to address basic research questions about systems of biotechnological interest. In each chapter, I introduce a challenge in biomedical research and describe the molecular level insights I have contributed toward addressing that challenge. I explore the nonfouling mechanism for a new zwitterionic biomaterial, PTMAO, and describe new metrics for quantifying the hydration of biocompatible materials in Chapter 2. I investigate the delicate balance between specific and nonspecific interactions between natural biomolecules in Chapter 3, and propose design rules to improve the performance of a peptide-based drug delivery vehicle for protein therapeutics. In Chapter 4, I describe my pioneering contributions to the molecular understanding of protein bound uremic toxin (PBUT) interactions with human serum albumin (HSA). I conclude with a discussion of the broader impacts of my graduate research in the fields of biophysics and biomaterials design.

Chapter 2
Investigating the molecular origins of nonfouling for zwitterionic materials

The human body is a notoriously difficult system to direct with engineered materials. Natural protective mechanisms, such as immune clearance and the foreign body response, have limited the efficacy of drug delivery vehicles and implanted materials for decades. Bioinspired zwitterionic materials have been engineered to overcome these biological impediments. Polycarboxybetaine (PCB) and polysulfobetaine (PSB) nanocages effectively block the immune response to encapsulated proteins without generating polymer-specific antibodies.[16] Zwitterionic hydrogels and surface coatings resist the foreign body response and protein adsorption more effectively than traditional nonionic biomaterials.[2,17] Extensive modeling and physical characterization suggests the exceptional biocompatibility of zwitterionic materials is related to their extremely strong interactions with water, or superhydrophilicity.[18,19] Although all zwitterionic materials interact strongly with water, a concrete explanation for the differences in nonfouling performance between zwitterionic materials has not been established.

Developing a more robust understanding of the link between the physical characteristics and the behavior of biocompatible materials has been a primary motivation for my graduate research. In this chapter, I describe 3 projects that represent important conceptual milestones towards an improved understanding of nonfouling biomaterials. In section 2.1, I rely on the traditional explanation for the nonfouling behavior of zwitterions, that hydration begets biocompatibility, to provide a mechanism for the excellent resistance to nonspecific interactions observed for a new zwitterionic polymer. I take a more comprehensive look at the hydrophilicity of various biomaterials in section 2.2, where I explicitly account for the tendency of "hydrophilic" materials to adsorb to a hydrophobic interface from water. Lastly, I propose in section 2.3 an approach for more precisely quantifying biomaterial-water interactions using unsupervised learning.

2.1 Modeling the nonfouling mechanism of the trimethylamine N-oxide zwitterionic polymer*

Introduction

The design of synthetic biomaterials for nonfouling applications has been heavily influenced by the chemistry of naturally occurring molecules. Two of the most prevalent materials in the class of zwitterionic nonfouling polymers, PMPC and PCB, have clear biological inspirations. PMPC was inspired by the phosphorylcholine headgroups of lipids in the eukaryotic cell membrane.[20] PCB was inspired by the protective osmolyte, carboxybetaine.[21] It is supposed that incorporating a more effective protective osmolyte than carboxybetaine might produce a more effective nonfouling polymer. Recently, researchers in the Jiang group managed to synthesize a zwitterionic polymer based on TMAO, perhaps the most effective and widely studied protective osmolyte [*in press*]. Previous investigations into the physical properties of the TMAO small molecule have established that very strong interactions with water molecules are partially responsible for its protective properties.[22,23]

* Reproduced in part with permission from J. Smith, J. Pfaendtner, and S. Jiang. Trimethylamine N-oxide derived Zwitterionic Polymers: A New Class of Ultra-low Fouling Bioinspired Materials. Science Advances, In Press. Copyright 2019 AAAS.

In this section, I describe my investigation of TMAO and PTMAO hydration using classical MD simulations and traditional techniques for quantifying hydration. I observed that the headgroups of PTMAO are similar to the TMAO small molecule in terms of the number of hydrogen bonds accepted from water, although the hydrogen bond lifetimes were longer for PTMAO. The structure of the first hydration shell around PTMAO headgroups was also found to be similar to the hydration structure around the TMAO osmolyte. In general, these results suggest that PTMAO maintains the superhydrophilic properties of the TMAO small molecule and that the polymer and osmolyte act by related mechanisms.

Methods

Molecular dynamics simulations of TMAO, OEG, and PTMAO in water

Partial charges for TMAO, OEG, and PTMAO atoms were assigned using the restrained electrostatic potential (RESP) method.[24] Quantum mechanical calculations for RESP were performed in Gaussian 09 using the B3LYP hybrid functional with the 6-31G(d) basis set.[25,26] For PTMAO, the quantum calculations were performed for a monomer with a methyl-capped backbone, and the partial charge of the backbone atoms were adjusted to neutralize the net charge of the monomer. Water molecules were described by the TIP3P water model.[27]

The initial configuration for each simulation was generated with GROMACS 5.1.2 utilities.[28] For the small molecule simulations, one solute molecule (TMAO or OEG) was centered in a cubic box with 2.5 nm sides. The box was then solvated with 504 randomly placed water molecules. A two-step equilibration procedure was followed to provide reasonable starting positions and velocities for production simulations. Energy minimization was performed with the solvated configuration using the steepest descent algorithm for 10,000 steps. Energy minimization was followed by a 1 ns simulation in the NPT ensemble, with the Bussi-Donadio-Parrinello (v-rescale) thermostat and Berendsen barostat for temperature and pressure control, respectively.[29,30] The output atomic coordinates and velocities for each system were used as the initial coordinates and velocities for the production simulations.

For the production phase, each system was simulated for 4 ns in the NPT ensemble with the Bussi-Donadio-Parrinello (v-rescale) thermostat and Parrinello-Rahman barostat for temperature and pressure control, respectively.[29,31] Frames were saved every 100 fs (50 steps) during the production simulation, and trajectories were analyzed with Gromacs utilities. The hydrogen bond count for every frame was calculated with the *`gmx hbond`* utility, using the default geometric definition of the hydrogen bond: a donor-acceptor distance of less than 0.35 nm and a hydrogen-donor-acceptor angle of less than 30 degrees. The probability distribution of the number of hydrogen bonds accepted by OEG and TMAO oxygens were generated with kernel density estimation (KDE), essentially a technique to smooth the hydrogen bond histogram with Gaussians. The autocorrelation function for the hydrogen bond lifetime was calculated using the method of Luzar and Chandler, as implemented in the Gromacs utility *`gmx hbond`*.[32] Radial distribution functions were calculated using the *`gmx rdf`* utility.[32]

Results and Discussion

To study the mechanism accounting for the extraordinary hydration of PTMAO, we performed molecular dynamics (MD) simulations of a TMAO small molecule and a 10-residue PTMAO oligomer in an aqueous solution. Simulations of a 3-residue PEG oligomer (OEG), were conducted for comparison. **Figure 2.1.1** shows the number and lifetime of hydrogen bonds with water for each molecule.

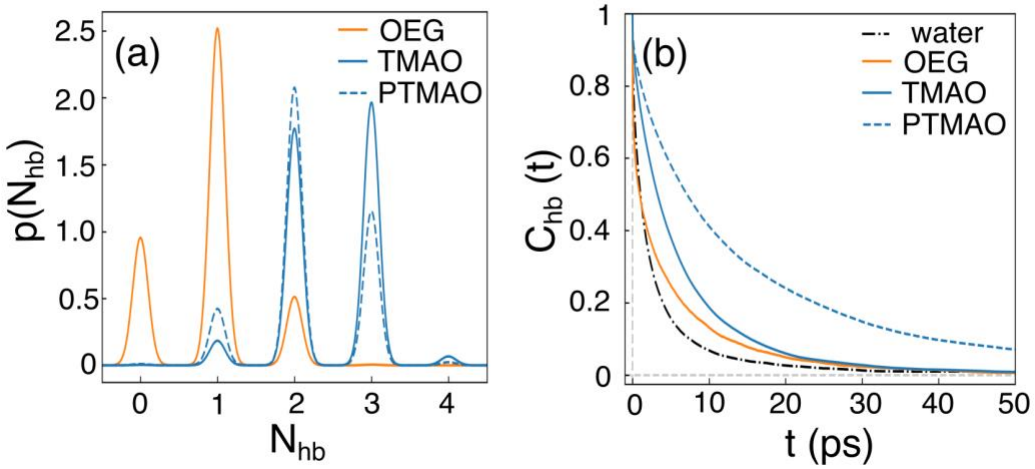

Figure 2.1.1 – Count and lifetime of biomaterial-water hydrogen bonds in MD simulations. (a) The probability distributions for the number of hydrogen bonds, N_{hb}, between water and the TMAO oxygen (blue), water and the PTMAO oxygen (blue dashed), and water and each OEG oxygen (orange). (b) Autocorrelation function for hydrogen bond lifetime, $C_{hb}(t)$, for water-TMAO oxygen (blue), for water-PTMAO oxygen (blue dashed), water-OEG oxygen (orange), and water-water hydrogen bonds (black dashed).

Simulation results showed that the TMAO oxygen accepts an average of 2.5 hydrogen bonds from water – accepting either 2 or 3 hydrogens bonds with approximately equal probability, while OEG oxygen atoms typically accept only 1 hydrogen bond from water, as shown in **Figure 2.1.1a**. In addition, the hydrogen bond lifetime (τHB) for TMAO-water interactions was observed to be longer than that for OEG-water interactions (**Figure 2.1.1b**). PTMAO formed longer lived hydrogen bonds with water than either TMAO or OEG. The persistent binding of multiple water molecules to the PTMAO headgroups suggested very strong interactions with water.

The structure of water around the TMAO and PTMAO molecules also suggests a hydration-based mechanism for their behavior. The radial distribution function (RDF) for water oxygen atoms with respect to each of the TMAO heavy atoms is pictured in **Figure 2.1.2**, along with representative snapshots of the hydration structure around TMAO and PTMAO.

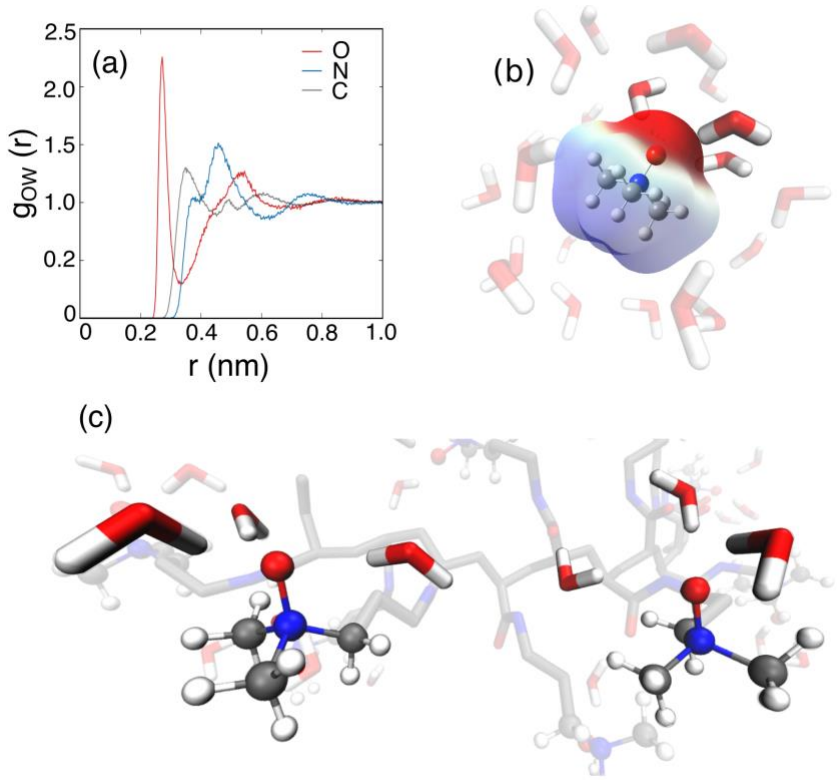

Figure 2.1.2 – Radial distribution functions and snapshots of hydration structure around TMAO and PTMAO. (a) RDF of water oxygen with respect to O (red), N (blue), and C (gray) atoms of TMAO. (b) Snapshot of aqueous TMAO with opaque and semi-transparent water corresponding to the first shoulder and main peak in the N-OW radial distribution function (RDF), respectively. (c) Snapshot of PTMAO in aqueous solution, highlighting the tightly bound waters near two of the TMAO headgroup.

The RDFs for water oxygen atoms with respect to the heavy atoms of the TMAO small molecule (O_{TMAO}, N_{TMAO}, C_{TMAO}) suggested a near-contiguous sphere of hydration, centered on the quaternary nitrogen (**Figure 2.1.2a**). The first peak of the N_{TMAO} RDF is comprised of a polar hydration peak, contributed by hydration around the quaternary amine cation, and a shoulder, contributed by the tightly bound water at the oxygen anion. **Figure 2.1.2b** shows a single frame from the TMAO monomer simulation, including the water molecules with a N_{TMAO}-water oxygen distance less than the first minimum in the N_{TMAO} RDF (0.6 nm). This snapshot reveals that the first hydration shell, with respect to N_{TMAO}, covers the whole TMAO molecule and includes the strongly hydrogen-bonded water at the TMAO oxygen (high opacity). The contiguous hydration shell observed in the simulation of TMAO small molecule was also observed in the PTMAO simulation (**Figure 2.1.2c**), which indicates that water is similarly ordered near the PTMAO headgroups.

Conclusion

Taken together, the MD simulation results suggest that PTMAO retains the superhydrophilicity observed for the TMAO small molecule. The strong hydrogen bonding with water and a contiguous hydration shell around the PTMAO headgroups could be responsible for its superhydrophilic properties. This mechanism should be revisited when more precise methods for quantifying hydration structure, such as those discussed in Sections 2.2 and 2.3, are established.

2.2 Quantifying the amphiphilicity of biomaterials with small molecule surrogates

Introduction

Explaining the differential in the resistance to nonspecific protein adsorption of various zwitterionic materials currently requires an amalgamation of several molecular-level observables. Shao et al. have explained the different properties of zwitterions based on several metrics related to their hydration properties and propensity for self-interactions.[33] Hydration free energy, which clearly distinguishes the hydrophilicity of zwitterionic from nonionic biomaterials, fails to discriminate nonfouling zwitterions from other zwitterions. For zwitterions with different charged groups, Shao et al. have suggested that the residence time of water molecules in the first hydration shell near the anion is correlated with nonfouling performance.[34] Unlike hydration free energy, this metric ranks CB as more hydrophilic than SB, in line with the relative resistance to nonspecific protein adsorption of PCB and PSB. However, adding carbon spacers between identical charged groups increases the residence times of water near the anion but leads to increased protein adsorption.[35] To account for the shortcomings of the strong anion hydration metric, Shao et al. argue zwitterions that tend to associate in aqueous solution are less hydrophilic than those which resist self-interaction.[33] While anion hydration and self-association can be qualitatively combined to rank the resistance to nonspecific interactions of the few zwitterionic materials that have already been characterized experimentally, the extensibility of this ranking system is questionable. A concise, quantitative explanation for zwitterionic resistance to nonspecific protein adsorption is required to efficiently design and screen improved nonfouling materials.

White et al. have suggested the partial desolvation energy is more relevant to biomolecular interactions than hydration free energy, because molecules can form stable intermolecular contacts in biological systems without complete desolvation.[36] The partial desolvation energy, which describes the energetic cost of stripping water molecules from the first coordination shell of an atom or molecule, can be straightforwardly related to the previously identified criteria for zwitterionic nonfouling. If the partial desolvation of the anion is high, the residence time of waters near the anion will also be high. If the partial desolvation energy of either the anion or the cation is high, the zwitterion should be resistant to self-association. Quantifying the partial desolvation energy required for different zwitterions to bind to physiologically relevant chemical moieties in aqueous solution may provide a way to predict nonfouling performance.

MD simulations of zwitterionic small molecules in water and organic solvents can be used to better understand which properties explain the behavior of zwitterionic biomaterials in biological environments. In this section, I report results for simulations of various biomolecules in a partitioned water-hexane system, which clearly distinguish hydrophilic from amphiphilic materials. I explain the relevance of partial desolvation energy in the partitioned system, highlighting the need to consider the hydration of the whole molecule. I also propose further simulations to extend the discriminative power of this approach.

Methods

PBMetaD for water-hexane transfer free energy

Parallel bias metadynamics (PBMetaD) simulations were performed to calculate the relative preference for water or hexane of various small molecule solutes.[37] These small molecules included carboxybetaine with 1, 3, and 5 carbon spacers between the charged groups (CB1, CB3, CB5), trimethylamine N-oxide (TMAO), oligo-(ethylene glycol)$_3$ with methyl and hydroxyl termination (PEG). Water and hexane were also tested as solutes for reference. The packmol software program was used to generate initial atomic configurations with a 5.0 nm x 2.5 nm x 2.5 nm system, filled half with water and half with hexane. We added 509 water molecules to one half of the box and 72 hexane

water molecules to the other half of the box, based on the bulk densities of water and hexane (1.0 g/mol and 0.66 g/mol, respectively). One solute molecule was placed in the center of the water half of the box. **Figure 2.2.1** shows the initial configuration for the water-hexane partition system CB1.

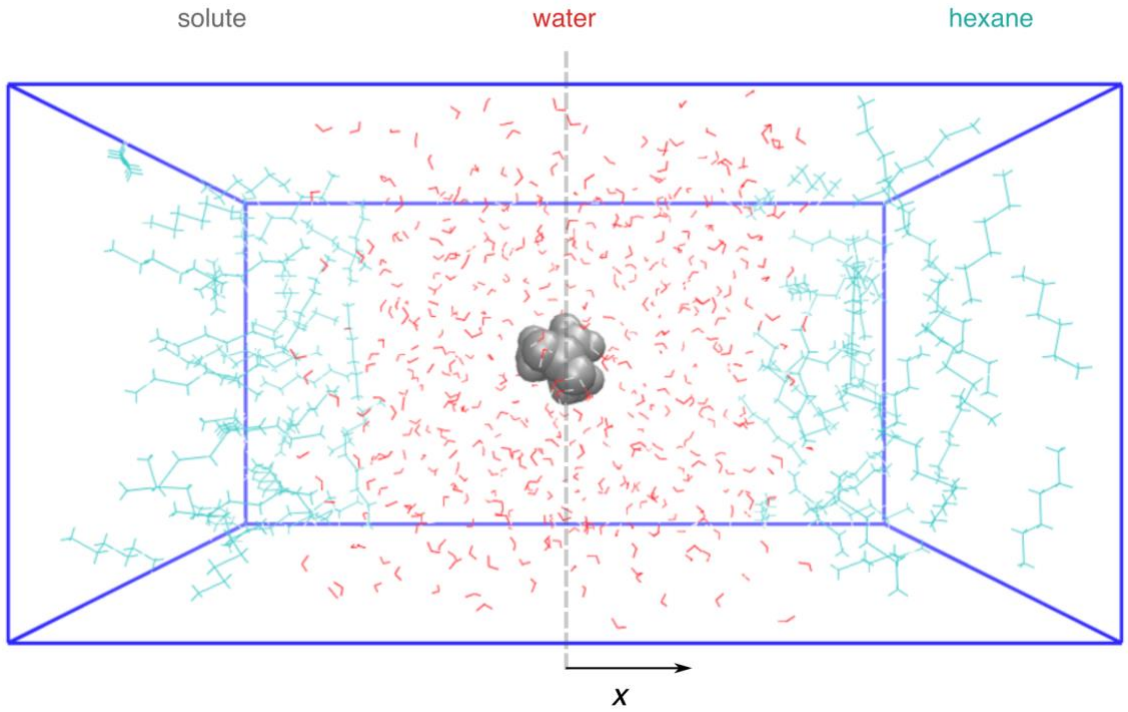

Figure 2.2.1 – Water-hexane transfer free energy configuration for CB1. Half of the box is filled with water (red), and half with hexane (cyan). The center of the water region is placed at the center of the box for ease of implementation.

For hexane and small molecule solutes, atomic charges were assigned using the restrained electrostatic potential (RESP)[24] method, based on an electronic structure calculation performed at the HF-631G level of theory. The electronic structure calculation was performed using the Gaussian 09 software package.[38] Force field parameters for the water molecules were assigned based on the TIP3P water model.[27] Each system was equilibrated with 10,000 steps of steepest descents energy minimization, a 250 ps annealing simulation from to increase the simulation temperature from 5 K to 300 K, and a 500 ps NPT simulation with the Berendsen barostat and v-rescale thermostat.[29,30]

The system is uniform in the y and z directions, so the reaction coordinate of interest is the distance in the x direction between the solute center of mass and the center of the box, x. A 1.1 ns steered MD simulation was performed to get 8 atomic configurations with different x. These 8 configurations were used as the starting configurations for PBMetaD with multiple walkers. The number of water oxygens bound to the solute, N_{OW}, was also biased. PBMetaD was performed for 100 ns (total of 800 ns of sampling for each solute) using the well-tempered variant of metadynamics to deposit bias every 2 ps to x (sigma = 0.1 nm) and N_{OW} (sigma = 0.25), with an initial height of 2.0 kJ/mol and a bias factor of 10.0. Upper and lower restraints on water oxygen positions were placed at -1.25 nm and 1.25 nm from the center of the box to prevent water molecules from being pulled into the hexane phase. Reweighting was performed with the Torrie-Valleau method[39], as described by Pfaendtner and Bonomi.[37]

Results and Discussion

Amphiphilic biomaterials preferentially accumulate at the water-hexane interface

A quantitative description of the amphiphilicity of biomaterials should help distinguish between fouling, low fouling, and ultra-low fouling materials. We generated free energy profiles for each solvent with respect to the *x* component of its center of mass position during the PBMetaD simulation, which can be naturally interpreted in terms of hydrophilicity and hydrophobicity. The hydration free energy of each solute can be approximated by the difference between the free energy in the center of the hexane (x = 2.5) and the center of the water (x = 0). Amphiphilic molecules with separate hydrophilic and hydrophobic domains may prefer to accumulate at the interface rather than either the water or the hexane phase. This preference is marked by a free energy minimum at the interface (x = 1.25), and can be quantitatively described as a relative free energy difference from either the water or hexane phase. **Figure 2.2.2** contains the free energy profiles for CB1, CB3, CB5, TMAO, and PEG.

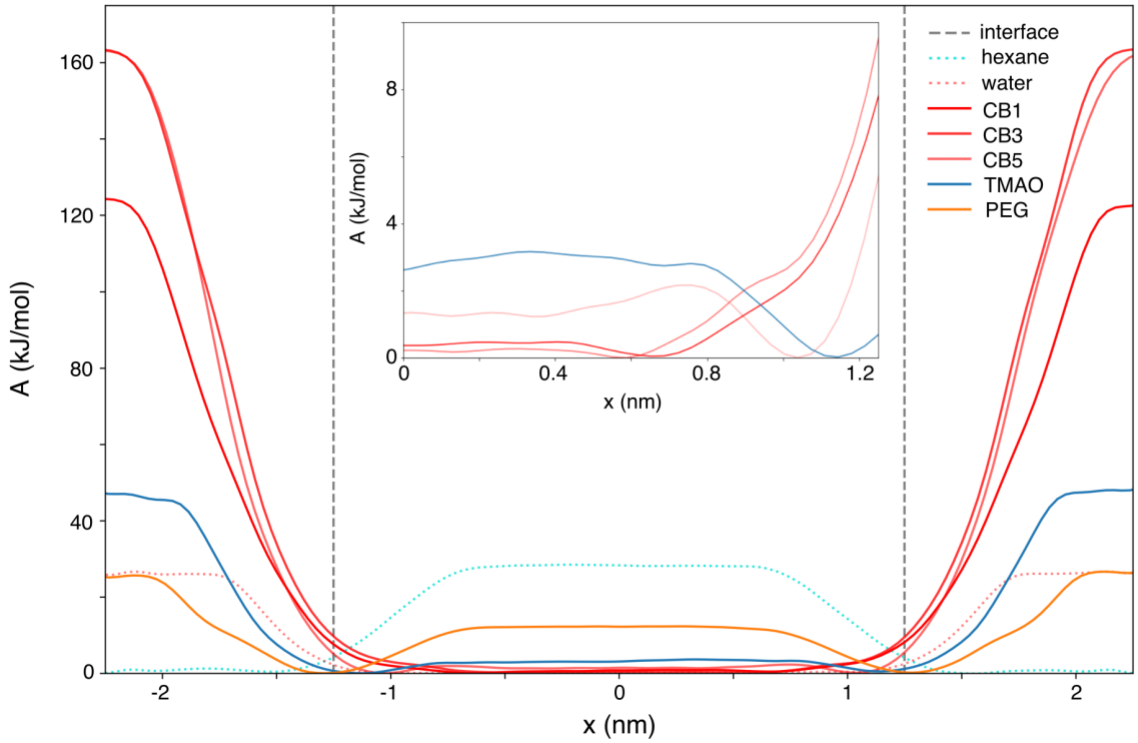

Figure 2.2.2 – Free energy profiles for small molecule center of mass position relative to the center of the simulation box. The position of the upper and lower restraints on water oxygen position are marked with vertical dashed lines, with the areas inside and outside the dashed lines corresponding to the water and hexane phases, respectively. Solid free energy profiles correspond to solutes of interest, with the dotted profiles for water and hexane included for reference. (inset) Zoomed in view of the profiles for zwitterionic materials in the water phase. TMAO and CB5 show shallow minima near the interface, suggesting a weak preference for the interface over the water phase.

PEG can be easily distinguished from zwitterionic molecules by a free energy well at the water-hexane interface that is 12 kJ/mol lower than the free energy of PEG in water. The inset in **Figure 2.2.2** shows that TMAO and CB5 also have a slight preference for the interface, although the free energy wells are less pronounced (2-3 kJ/mol) and on the order of energy fluctuations at 300 K. TMAO has been previously described as a dipolar hydrophobic osmolyte and has recently been reported to accumulate at the air-water interface, so accumulation at the interface is not surprising.[40,41] The hydration free energy for both CB3 and CB5 is -165 kJ/mol, about 40 kJ/mol lower than the

hydration energy of CB1. This finding is in line with previously reported hydration energies for these species determined by free energy perturbation.[35] The larger hydration free energy of CB3 relative to CB1 and the lack of a discernable hydrophobic signature in the CB3 free energy profile (such as the shallow interfacial minimum observed for CB5) suggests a non-monotonic disturbance in hydration as extra carbon spacers are added to CB1. The partial desolvation energy of tightly bound and weakly bound waters might offer more insight into the relative nonfouling performance of CB molecules.

The partial desolvation energy of hydrophobic hydration shell correlates with interface binding

We created two-dimensional free energy surfaces of *x* vs. water coordination for CB1, CB3, and CB5 to determine whether the partial desolvation energy can be used to distinguish between CB molecules with different carbon spacer lengths. We calculated two different coordination numbers for each molecule, one corresponding to tightly bound water (C_{tight}) and one corresponding to loosely bound water (C_{loose}). Coordination numbers were calculated by adding the number of contacts between each CB heavy atom (C, N, O) and water oxygens within a spherical shell. The radius and thickness of the C_{tight} coordination shell were defined as 0.27 nm and 0.03 nm. The radius and thickness of the C_{loose} coordination shell were defined as 0.35 nm and 0.06 nm. **Figure 2.2.3** shows the free energy surfaces for each CB molecule for C_{tight} vs. x (solute center of mass x distance from the center of the box) and C_{loose} vs. x.

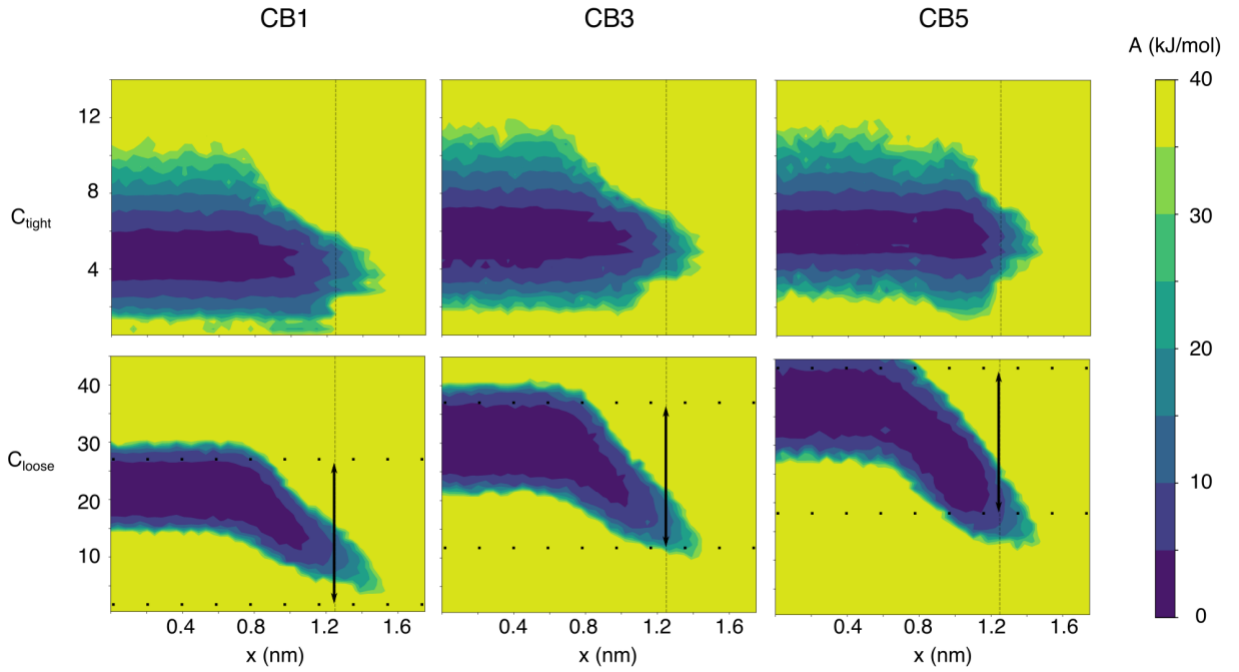

Figure 2.2.3 – Free energy surfaces for position versus coordination for tightly and loosely bound waters for CB1, CB3, and CB5. The solute center of mass is in the center of the water phase at x = 0 nm, and at the water-hexane interface at x = 1.25 nm. The putative interface is marked by the vertical dashed line. (top) Coordination number between CB heavy atoms and tightly bound water oxygens atoms. (bottom) Coordination number between CB carbons and loosely bound water oxygen atoms. Extent of the lowest energy region (within 5 kJ/mol) marked with dotted line, and arrow corresponding to this extent for CB1 transposed to CB3 and CB5 plots for comparison.

CB3 and CB5 appear to tightly bind more water molecules (C_{tight} = 6) than CB1 (C_{tight} = 5). However, the most favorable coordination state for each molecule is roughly the same in the middle of the water phase (x = 0 nm) and at the water-hexane interface (x = 1.25 nm). This suggests that the tightly bound water, most likely interacting with the carboxylate group, does not determine the

likelihood of CB molecules to interact with the hexane interface. The coordination of weakly bound water, however, changes significantly for each molecule between the middle of the water phase and the interface. This dehydration suggests that the liberation of hydrophobically bound waters dictates the preference of a CB molecule for interactions with the hydrophobic hexane interface. The coordination with loosely bound water can change from 43 (in the middle of the water) to 18 (near the interface) for CB5 with an energy change of 5 kJ/mol. The same decrease in C_{loose}, marked by the vertical arrows in **Figure 2.2.3**, would require ~30 kJ/mol for CB3 and > 40 kJ/mol for CB1. The free energy cost of removing many loosely bound waters is inversely correlated with carbon spacer length for CB molecules. Further analysis is required to establish an unambiguous criterion for nonfouling performance based the partial desolvation energy.

Conclusions

Because the current mechanistic explanation has been designed to match the experimentally observed nonfouling ranking *a posteriori*, it is tailored to explain the differences between the few zwitterionic moieties that we have already synthesized. Quantitative metrics for amphiphilicity and hydrophobic hydration, such as those developed in the present work, should provide a simpler and more extensible description of the properties of zwitterionic materials that relate to their nonfouling behavior. A straightforward, quantitative criteria for the property-function relationship of zwitterionic materials should unblock the nonfouling materials design cycle and open the possibility of *de novo* nonfouling materials design.

2.3 Identifying unique water structures at biointerphases with unsupervised learning

Introduction

The behavior of synthetic materials in complex biological environments is largely determined by their interactions with water. Hydrophobic materials initiate protein adsorption and the foreign body response, while hydrophilic materials resist protein adsorption through strong water interactions. While some materials can be easily distinguished as purely hydrophobic or hydrophilic based on chemical intuition, most biologically relevant molecules are amphiphilic, containing both polar and apolar moieties. While molecules like CB5 and PEG are obviously amphiphilic, as discussed in Section 2.2, even the cationic and anionic moieties of CB1 have significantly different water affinities. Quantifying the relative populations of loosely- and tightly-bound water molecules could provide insight into the role of mismatched water affinities in nonfouling biomaterials.

Experimentally, sum frequency generation (SFG) vibrational spectroscopy can distinguish between strongly and weakly hydrogen-bonded water near the biomaterial interface. However, there is some disagreement about how the relative amount of weakly-bound and strongly-bound water relates to nonfouling performance. For example, Leng et al. suggest that nonfouling polymer surfaces can be identified by the presence of a large population of strongly hydrogen-bonded water molecules at the interface.[18] Others have suggested that a large population of weakly bound water is a shared characteristic of most biocompatible materials.[42–44] It is difficult to settle this dispute with molecular insight because traditional metrics for water structure and dynamics are not capable of discriminating between unique structures in non-crystalline phases. **Figure 2.3.1** shows that the Q6 Steinhardt order parameter, a traditional descriptor for molecular structure, can't distinguish Lennard-Jones spheres in the liquid phase from those in the amorphous solid phase.

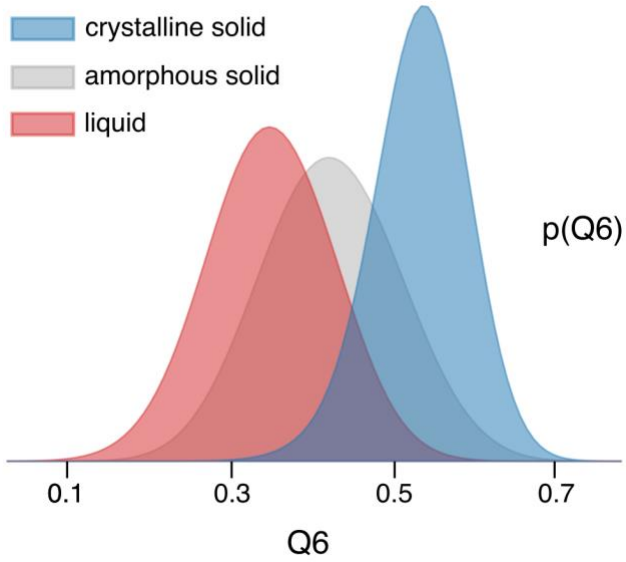

Figure 2.3.1 – Probability densities for the Q6 Steinhardt order parameter calculated for populations of Lennard-Jones spheres simulated in different phases. The mean Q6 of the populations increases from (red) liquid to (gray) amorphous solid to (blue) crystalline solid. The degree of overlap between the distributions determines the likelihood that a molecule in one phase could be improperly labeled as belonging to another phase based only on Q6.

Identifying reoccurring local structural motifs in non-crystalline environments could help to quantify the relative populations of tightly bound water, loosely bound water, and bulk water in biomolecular systems. Several groups have applied advanced statistical learning techniques to MD trajectory data to identify recurring molecular structures in aqueous systems. Soto et al. used supervised machine learning techniques to distinguish two types of local order in liquid water based on MD simulation trajectory data.[45] A major drawback of this approach is that supervised learning requires human labeling based on *a priori* knowledge of different structural regimes, so it is not well suited to finding unknown molecular patterns. Gasparotto et al. recently introduced an unsupervised learning approach to identifying structural motifs, and used it to automatically identify hydrogen bonds without an arbitrary geometric definition.[14,46] They call this approach the probabilistic analysis of molecular motifs (PAMM). The PAMM approach can be straightforwardly adapted to find recurrent structural patterns in MD systems without previous knowledge of the characteristics of those patterns. Unfortunately, the PAMM algorithm, written by Gasparotto in Fortran 90, cannot be directly trained on MD trajectories from GROMACS or other popular simulation engines.

In this section, I describe the application of PAMM to identify reoccurring structural motifs in an MD simulation of water near a model self-assembled monolayer (SAM) surface. I found that with a small list of simple order parameters, the PAMM algorithm could distinguish tightly bound water molecules from molecules in the bulk water phase. I demonstrate that improved order parameters can provide a more specific classifier, suggesting that an improved list of structural descriptors might capture the subtler disturbances in water dynamics and structure expected for weakly surface-associated phases.

Methods

Molecular dynamics simulation of a nonanol SAM in aqueous solution

The simulation system was comprised of a hydroxyl-terminated alkyl SAM surface in neat water. We built the pdb structure for nonanol in GaussView and replicated this pdb structure in the x and y directions to create a 7 molecule by 7 molecule SAM surface (49 molecules total), using an in-

house Python script. We calculated partial charges for the nonanol molecule following the RESP method, based on quantum mechanical calculations performed with the Gaussian 09 software at the HF 6-31G* level of theory.[24,38] We used bonded parameters from the general amber force field (GAFF) to complete the nonanol force field.[47] We solvated the system using GROMACS tools, adding a total 710 water molecules.[48] Finally, we equilibrated the system with 10000 steps of energy minimization following the steepest descents algorithm. We then performed a 1 ns MD simulation, with an integration timestep of 2 fs. The temperature of the system was coupled to 300 K, at a frequency of 10 ps^{-1}, with the Bussi-Donadio-Parrinello stochastic thermostat.[29] Order parameters were recorded every 2 ps for each water molecule using Plumed 2.4.[49]

Probabilistic Analysis of Molecular Motifs (PAMM)

The PAMM algorithm is a multi-step process for automatically identifying reoccurring structural patterns in a MD simulation. The input is MD trajectory data, and the output is a probabilistic model that can classify atoms as belonging to a particular free energy minima in structural descriptor space. I will describe the process step-by-step to clarify what this means and why it will be especially useful in describing biomolecule hydration.

The first step of the algorithm involves calculating structural descriptors (features) from the atomic coordinates of an MD trajectory. These descriptors are used as the structural fingerprints that distinguish one local environment from another. PAMM can be applied in a high-dimensional feature space so that any local order parameter that could provide discriminative information about the structural environment of an atom should be calculated. With the specific example of liquid water in mind, this will include calculating the local tetrahedral order, the Steinhardt order parameters,[50] and the coordination number for each water molecule in each frame of the MD trajectory.

The second step mitigates the exploding computational cost associated with calculating a high-dimensional feature vector for each atom in a large MD simulation system. Downsampling is performed for computational efficiency by using farthest point sampling.[51] This approach provides a smaller sample size with which to construct a probability density in high-dimensional space, but faithfully captures irregularities from the initial distribution. The points selected from farthest point sampling then serve as grid points for kernel density estimation. The result of this step is a probability density function, P(x), which describes for the likelihood that a molecule taken at random will have a given structural feature vector, x.

The third step is nonparametric clustering of the grid points selected by farthest point sampling using the quick-shift algorithm and P(x). This algorithm follows steepest ascent in probability from each grid point, and clusters all grid points which end up at the same stationary point. The straightforward interpretation of this clustering approach is that grid points from the same cluster belong to the same free energy well in structural descriptor space. A Gaussian mixture model (GMM) is then constructed by combining Gaussians for each cluster. The final GMM is composed of Gaussians for each cluster with the mean placed at the mode identified by quick-shift, and shape determined by the covariance of the grid points assigned to that cluster. This method is better suited to irregularly shaped data than training a GMM with the expectation maximization algorithm directly on the unclustered high-dimensional probability distribution P(x). The resultant GMM can be used to quantify the number of atoms belonging to each cluster in each frame, and to identify the regions of the simulation box where each molecular pattern is most likely.

We applied the PAMM algorithm to the data from our MD simulations to identify unique water structures in an unsupervised fashion. For each of the 710 water molecules in our simulation system, we calculated the local order parameters in 501 simulation frames, for a total of 355,710 unique observations of the structural environment around a water molecule. We developed a Python library called *nonstick* to follow the PAMM procedure with the help of mature open source libraries for

analyzing MD simulations in Python, like MDTraj.[52] The Python code is available at https://github.com/anotherjoshsmith/nonstick.

Results and Discussion

Waters bound to hydrophilic surface can be distinguished from bulk water with PAMM

The PAMM algorithm provides an automated way to distinguish unique structures sampled in an MD simulation. I simulated a simple SAM surface in water, a model system that includes 2 organic interfaces with different chemistries. A robust implementation of the PAMM procedure, trained with an appropriate list of order parameters, should be able to distinguish at least three populations of water in this simulation: tightly bound water at the hydrophilic interface, loosely bound water at the hydrophobic interface, and free water molecules in the bulk. I calculated a list of 5 order parameters to describe the local structural environment for each water molecule in each frame of the simulation: the coordination number, the third (Q3), fourth (Q4), and sixth (Q6) Steinhardt order parameters, and the instantaneous velocity. I then used the PAMM algorithm to identify distinct populations of water based on this array of descriptors. **Figure 2.3.2** shows all of the observations of local water structure projected onto 2 dimensional order parameter surfaces and colored by the population assigned by PAMM.

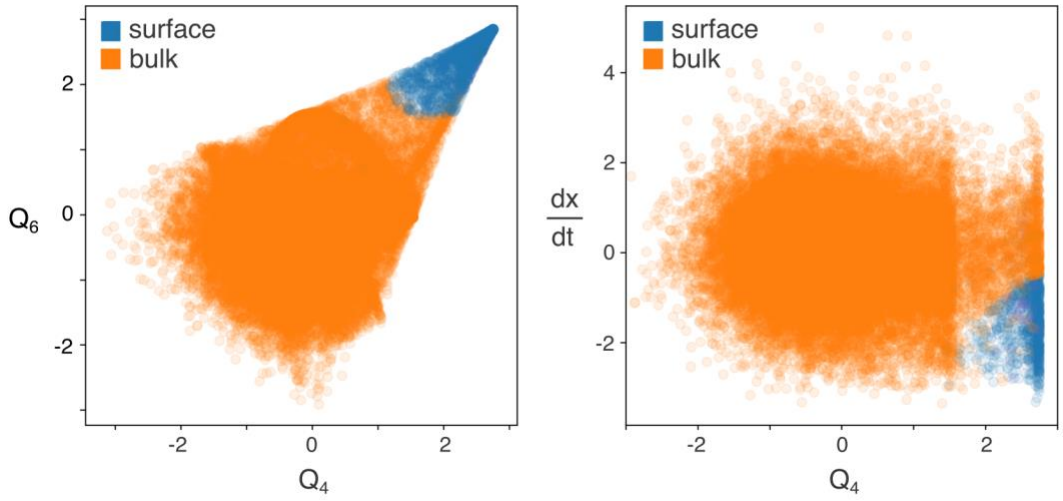

Figure 2.3.2 – Distinct populations of water molecules identified by PAMM and their projections onto 2 dimensional order parameter surfaces. (left) The primary and secondary clusters – colored orange and blue, respectively – are separated by the Steinhardt order parameters Q4 and Q6, suggesting greater spherical symmetry in the coordination shell of molecules assigned to the secondary cluster. (right) Molecules in the secondary cluster also have relatively low instantaneous velocity in the direction parallel to the SAM surfaces (dx/dt). These observations and closer inspection support the interpretation of the secondary and primary clusters as the populations of surface-bound and bulk water molecules.

Figure 2.3.2 shows that two regions of phase space were identified by the PAMM algorithm. The orange and blue clusters in **Figure 2.3.2** accounted for 0.94 and 0.06, respectively, of the water molecules observed during the simulations. The molecules in the blue cluster seem to have a more structured environment than average, as captured by the higher Q4 and Q6 Steinhardt parameters (**Figure 2.3.2** left). The blue cluster also has slower displacement in the xy plane (**Figure 2.3.2** right). The restrained dynamics and increased order associated with molecules assigned to the blue cluster are also expected to be characteristics of the population of water molecules bound to the hydrophilic surface. Visual inspection of the molecules assigned to the blue cluster in each frame confirmed that

the blue cluster describes water molecules near the hydrophilic surface, while the orange cluster encapsulates bulk water molecules.

A more rigorous way to interpret the two populations identified by PAMM is to compare the spatial distribution of water molecules assigned to either population. We constructed a probability distribution function (PDF) for water position in the z direction by calculating a finely discretized histogram of the z positions of each water molecule. We constructed weighted PDFs for the bulk and surface populations, recalculating the same histogram with each observation weighted by the cluster probability assigned by the PAMM model. **Figure 2.3.3** shows the probability density function for water position in the z direction, weighted by cluster probability calculated with the PAMM model.

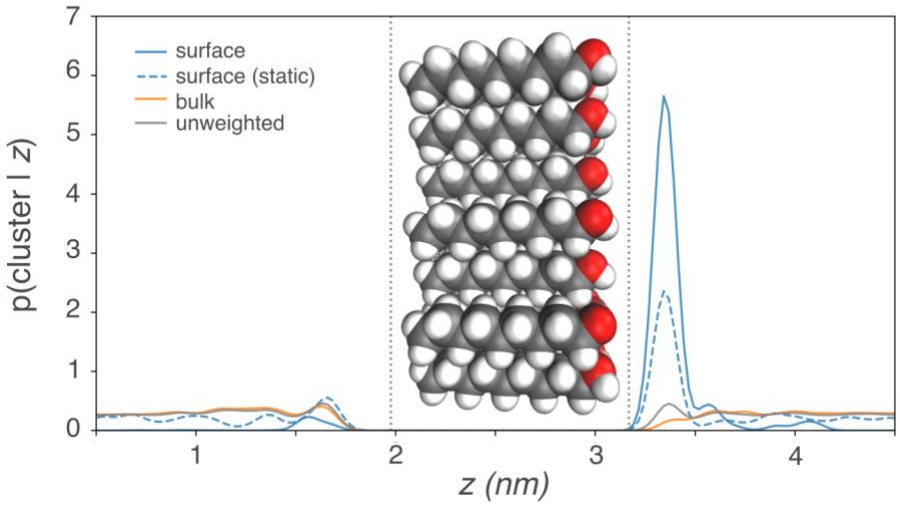

Figure 2.3.3 - Probability density functions for cluster assignment, conditional on water position in the z direction. Solid blue and orange lines represent the likelihood that a water molecule is assigned to the surface or bulk clusters, respectively, by the PAMM model trained with Steinhardt parameters and instantaneous x velocity. The gray solid line represents the baseline probability that a water in found at a given z position, regardless of cluster assignment. The dashed blue line represents the surface cluster PDF for a PAMM model trained with only the Steinhardt parameters. Horizontal, dashed gray lines mark the putative locations of methyl- and hydroxyl-terminated surfaces. A profile snapshot of hydroxylated SAM surface is embedded to show the effect of the hydroxyl headgroups.

The unweighted (baseline) PDF for water position in the z direction shows slight peaks near the hydrophobic and hydrophilic surface, which represent the locations of the weakly and strongly bound populations of water. The peak near the hydrophilic surface was enhanced in the PDF weighted by the secondary cluster probability and diminished in the PDF weighted by the primary cluster probability. The effects on peak height in the weighted PDFs supported our previous assertion that the primary and secondary clusters correspond to bulk and strongly bound water, respectively. Neither of the clusters specifically capture the population of weakly bound water molecules at the hydrophobic surface. More advanced structural/dynamical descriptors may be necessary to distinguish from bulk water the lightly perturbed dynamics and structure of water molecules at the hydrophobic surface.

We also compared a weighted PDF for the secondary cluster from a PAMM model trained without the dynamical observable. The smaller peak for the weighted PDF of the surface cluster without the dynamical observable demonstrates the specificity added by considering dynamics in addition to structure. It is intuitive that dynamics should help to distinguish bound from bulk water, given that hydrogen bonds with the hydroxyl surface are expected to reduce the mobility of bound water molecules. Interestingly, the PDF without dynamical information exhibits periodic peaks emanating from the hydrophobic surface that are not present in the unweighted PDF. This suggests

that the PAMM algorithm is identifying distinct layers of water which have "surface-like" structure. This observation provides hope that a more robust implementation of PAMM, and an improved list of order parameters, could distinguish these intermediate layers of water.

Conclusions

While it is quite difficult to define a universal hydration metric that can be applied to different heterogeneous molecules, there is clear evidence that water structure is different near hydrophobic and hydrophilic moieties in aqueous solution. We have demonstrated that the PAMM algorithm can automatically identify different types of water in an MD simulation. Adding more complicated order parameters, such as the local dipole magnitude and dipole reorientation dynamics may help to identify populations of weakly bound and intermediate water.

An improved PAMM model that can distinguish between different 'types' of hydration and pinpoint specifically where they occur in an MD simulation will provide unparalleled mechanistic insight into hydration. It will allow researchers to develop new metrics to describe hydration (fraction tight vs. loose), identify whether 'intermediate water' is necessary for biocompatible materials, and provide a clearer depiction of a 'contiguous shell' of water. Further investigation into probabilistic models for biomaterial hydration could facilitate the *de novo* design of nonfouling materials.

Chapter 3
Designing peptide-based materials for protein drug delivery applications

Protein drugs (biologics) have great potential for precision medicine, due to their high affinity and specificity for biological targets.[53] The therapeutic potential of biologics has been hindered by their instability during long-term storage and their short therapeutic half-life. The chemical conjugation of biocompatible polymers to proteins of interest has been among the most successful and widely used approaches to overcoming these issues.[53–57] Bioinert, synthetic polymers, such as polyethylene glycol (PEG) and polycarboxybetaine (PCB), have been shown to improve the pharmacokinetics and shelf-life of various biologics.[58–60] However, polymer conjugation complicates the synthesis and decreases the yield of relatively expensive-to-produce protein drugs.[61] The high cost of conjugation, coupled with questions about the safety of introducing non-biodegradable synthetic polymers into otherwise degradable protein drugs, have prompted the investigation of alternative approaches to stabilizing biologics.[62]

Peptide-based materials provide an attractive alternative to synthetic polymers for drug delivery applications. Because peptides are composed of the same natural building blocks as native proteins – the amino acids – they are inherently biodegradable. Peptides can also be synthesized more precisely than artificial polymers, so the chemistry of peptide-based materials can be strictly controlled. However, because peptide-based materials are composed of the same chemistries that mediate specific interactions between proteins, extra care must be taken in the design of peptide-based materials to avoid spurious interactions between the engineered material and natural proteins. Understanding the delicate balance between specific and nonspecific interactions is essential to designing functional peptide-based materials.

Describing the molecular mechanisms of specific and nonspecific interactions between proteins and peptides has been another major thrust of my graduate research. I have demonstrated that each amino acid contributes uniquely to specific protein-protein interactions (Section 3.1), suggesting that amino acids with "similar" chemistry cannot be used interchangeably in peptide design. Exploring this concept in more detail, I have found that the propensity for salt bridging of charged amino acid sidechains can partially explain the propensity of alternating charge peptides to promote or resist unintended biological interactions (Section 3.2). Building off these insights, I studied the physical properties of an alternating charge peptide that has already been used for stabilizing protein drugs, and proposed *in silico* designed peptide as an alternative with superior physical properties (Section 3.3). The research in this chapter provides valuable insight into design strategies for peptide-based drug delivery vehicles.

3.1 Coarse graining and combinatorics for an improved understanding of amino acid contributions to the protein–protein binding affinity*

Introduction

Protein-protein interactions (PPIs) are ubiquitous in biology. Biological systems depend at every level on proteins faithfully recognizing their natural binding partners and avoiding spurious interactions. Proteins often function in crowded cellular environments where they must bind strongly on chance encounters with their target partners in a sea of other proteins.[63] Health issues ranging from Huntington's disease[64] to certain cancers[65] have been attributed in part to proteins binding with the wrong counterpart or failing to bind altogether. The ability to prevent harmful interactions and promote healthy ones has been limited by the lack of a quantitative description of PPIs. While resistance to nonspecific interactions in proteins has been studied and replicated with some success[33,59,66,67], a fundamental understanding of the mechanisms governing specific recognition and binding of proteins has eluded extensive research efforts.

Binding affinity (BA) prediction is perhaps the most common approach to quantifying PPIs. Models are trained to predict experimentally determined BAs for protein complexes from feature vectors traditionally derived from primary or tertiary structure information. The primary sequence approach is attractive because there is an unambiguous representation (its amino acid sequence) readily available for all known proteins. However, it is difficult to generate a mechanistic understanding of specific interfacial interactions between binding partners without their bound and unbound tertiary structures. A surge in high-quality 3D protein complex structures deposited in the RCSB Protein Data Bank (PDB)[68] has recently made computational investigation of structure-affinity relationships feasible. The publication of benchmark data sets[69,70] for structure-affinity prediction has fostered the rapid development and comparison of diverse BA models.[71–74] Strikingly, no consensus has been reached as to the most useful descriptors for structure-affinity prediction.

Owing to the massive size of the theoretical number of different PPIs (e.g., consider the combinatorial explosion of two 20 residue interfaces in light of the number of possible amino acid combinations), the design space for spatiochemical PPI features must be judiciously reduced to make BA models comprehensible and computationally tractable without sacrificing physical significance. Dimensionality reduction is especially important given the few reliable training examples available for structure-affinity prediction. Advanced machine learning (ML) algorithms have been used for automated feature selection and model fitting, though this often leads to "black box" type outputs where there is not intuitive mapping between input features and calculated affinities.[71,75] Manual feature selection is commonly paired with simple ML techniques, such as least squares regression, to maintain transparency in I/O relationships. Unfortunately, pairwise intermolecular contacts (ICs) must be binned by type to avoid overfitting (even a coarse-grained interface results in 400 pairwise features if amino acids are treated uniquely). This approach leads to highly degenerate features that, while simple, obscure the unique contributions to binding of each amino acid. It has been demonstrated that the interface distance cutoff influences the performance of intermolecular contact models[24], although the mechanistic reason for this has not been explored. Given that pairwise contacts can only be reasonably defined between residues in immediate contact, this approach does not lend itself to meaningful exploration of residues deeper in the interface neighborhood. A residue level description of the interface where amino acids are simply counted without regard to specific contacts

* Reproduced in part with permission from J. Smith, J. Pfaendtner, and S. Jiang. Redefining the Protein-Protein Interface: Coarse Graining and Combinatorics for an Improved Understanding of Amino Acid Contributions to the Protein-Protein Binding Affinity. Langmuir, 33(42):11511-11517. Copyright 2017 American Chemical Society.

may be less restricted. The utility of a residue-based model in predicting binding affinity and informing our understanding of PPI specificity has not been comprehensively investigated.

In this section, we use a combinatorial approach to investigate the role of each amino acid in various positions of the binding interface. We use a modern regression algorithm, lasso, to maximize interpretability of our model and systematically avoid overfitting. We find the amino acid composition (AAC) model is more predictive than a detailed atomic contact model, especially for flexible and enzyme-containing complexes. We then demonstrate the relative robustness to interfacial distance cutoff of a combinatorial approach. Finally, we discuss the value for BA prediction of each amino acid, as evidenced by our results. The implications of these findings are compared to those of previous BA models and suggestions are made about the form of future models.

Methods

Dataset

Protein complexes were chosen from the updated Structure Affinity Benchmark (SAB 2.0) previously curated by Vreven et al.[70] The full SAB contains 179 complexes for which experimental binding affinity and 3D structure are known. While there are several hundred more complexes in the RCSB Protein Data Bank with known structure and affinity, the SAB has quickly been adopted by the community as a means of comparison between diverse approaches to structure-affinity prediction. To our knowledge this is the most reliable set of structures for rapid training and comparison of binding affinity models. All subsets considered in this work are derived from this list of 179 proteins. The relative size and approximate intersections of notable subsets are pictured in **Figure 3.1.1**. We specifically compare on the bases of flexibility (86 flexible and 93 rigid complexes), class (69 enzyme-containing complexes), and availability of PICCOLO intermolecular contacts (118 complexes, discussed below). Flexibility is recorded in the SAB as the root mean square deviation (RMSD) of atomic positions from crystal structures of the unbound components to the bound complex. PDB ID, protein class, flexibility and experimental binding affinity were taken from the SAB. Structural feature vectors were generated from PDB structures as described below.

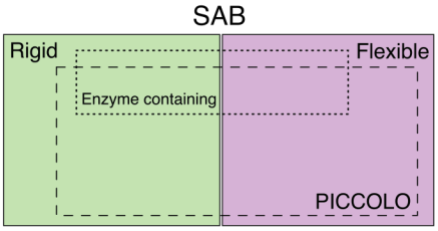

Figure 3.1.1 – Subsets of the SAB used for analysis. The number of complexes in each subset and intersections between subsets are approximated by area.

For initial comparison of IC and AAC models, we scraped data tables from the online PICCOLO database of structurally characterized protein interactions.[76] PICCOLO contains detailed information about intermolecular contacts for 118 of the 179 SAB complexes. Intermolecular contact (IC) models were trained based on 11 interaction types from PICCOLO. These interactions include van der Waals contact, van der Waals clash, hydrogen bonds, aromatic interactions, pi-cation interactions, disulphide bonds, and polar, apolar, ionic, and proximal contacts. This level of detail exceeds that commonly used (grouping IC by polarity), and is likely a better representation of the necessary detail for a highly specific IC model. Contact residues are also reported, so residue-level feature vectors were derived directly from PICCOLO-defined pairwise interactions for the initial comparison of IC and AAC models.

For further AAC analysis, we used a publicly available protein interface analysis script written by Konrad Krawczyk of the Oxford Protein Informatics Group. This script uses the Biopython library to identify contacts and "interfacial neighborhood" residues from complex PDB files based on user supplied distance cutoffs. Residues on each chain containing heavy (non-hydrogen) atoms within the specified cutoff distance of the opposing partner were counted as interfacial and others are ignored. The AAC representation of one SAB complex is illustrated for multiple cutoffs in **Figure 3.1.2**.

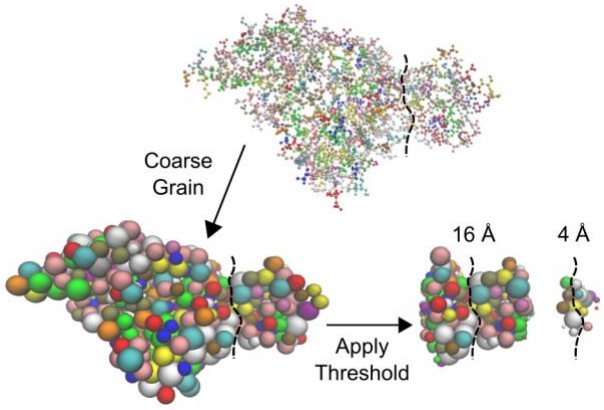

Figure 3.1.2 – AAC representation of Cysteine desulfurase IscS from atomistic structure file (PDB: 3LVK). Dashed black lines represent the enzyme-inhibitor interface. (top) Atomistic representation of the enzyme-inhibitor pair with atoms and bonds colored by amino acid residue. (bottom) Coarse-grained representation of (left to right) the full complex, and interface residues based on 16 Å and 4 Å distance cutoffs.

We generated AAC feature vectors for all 179 complexes at 13 equally spaced intervals from 4-16 Å (every 1 Å) and filtered these for analysis of the subsets mentioned above. Concretely, we generated a 179 by 20 matrix with each column corresponding to an amino acid, and the entries in that column corresponding to the number of times that amino acid is counted within the supplied distance cutoff for each complex. Each complex is represented by a single binding conformation as determined by x-ray crystallography. Although proteins are dynamic and adopt many conformations in their environment, our coarse-grained representation of the interface is less sensitive to the specific x-ray resolved sidechain positions than contacts-based approaches.

Linear Regression with LassoLarsIC

Linear regression has commonly been used for protein-protein BA models.[69,70,72,77] However, ordinary least squares regression with many features and relatively few training examples is prone to overfitting. Luckily, the statistical learning community has developed regression algorithms that preserve the interpretability of least squares regression but improve prediction accuracy for unseen data. To reduce the model variance introduced by ordinary least squares, we used the least absolute shrinkage and selection operator (lasso). Lasso regression minimizes the residual sum of squares (RSS) subject to an ℓ_1 constraint on the regression coefficients. A set of AAC feature vectors, x_i, and experimental BAs, $Gexp_i$, are inputs to the training algorithm, which adjusts regression coefficients, β_j, for each feature to minimize

$$\sum_{i=1}^{n}\left(Gexp_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j|$$

where the first term is the RSS and the second is the ℓ_1 norm of the coefficient vector times a regularization factor, λ. A major strength of the lasso is that it performs automatic feature selection. Features that don't add predictive value are dropped from the final equation, while coefficients for

meaningful parameters are scaled based on the regularization parameter. The regularization parameter, λ, may be selected via cross-validation (CV), resampling (e.g., bootstrap), or information criterion. We used the Bayesian information criterion (BIC)[78] to select λ. This criterion rewards models for good prediction, but incurs a penalty based on number of features to prevent overfitting.

This approach was implemented using the LassoLarsIC tool in the python package, scikit learn.[79] LassoLarsIC is based on lasso regression implemented with the computationally efficient Least Angle Regression (Lars) algorithm.[80] Regression models were trained using every unique combination of amino acid features for each given training set and cutoff distance. This combination space includes 1,048,575 unique feature sets, ranging from 1 to 19 features in length. The breakdown of models for each feature vector length is provided in **Table 3.1.1**.

Table 3.1.1 – Breakdown of amino acid combination space.

# features	# models
1, 19	20
2, 18	190
3, 17	1,140
4, 16	4,845
5, 15	15,504
6, 14	38,760
7, 13	77,520
8, 12	125,970
9, 11	167,960
10	184,756
Total	1,048,574

Computational Toolkit

All data wrangling and analysis was performed using the Python 2.7 programming language on a commercial strength desktop computer. The Python ecosystem is well suited to bioinformatics research due to the extensive and ongoing development of structural biology and data analysis libraries.[81] We used the pandas library for data management and algorithms in scikit-learn for model generation.[79]

For combinatorial work, we used the itertools library in python to generate all 1,048,575 combinations of amino acids. We iteratively trained models using the LassoLarsIC module from scikit-learn, and wrote to a pandas dataframe for analysis of top amino acid features. The Pearson's correlation coefficient, r, was calculated as the square root of the coefficient of determination, R^2. We also performed 5-fold cross-validation with ordinary least squares for all amino acid feature combinations at 4, 8, 12, and 16 Å cutoffs to support lasso results. The mean squared error for cross validation is included in **Figure I.4**.

Results and Discussion

AAC vs. Atomic Contacts

Before building models for various distances, we assessed the performance of a simple residue count model relative to a detailed intermolecular contacts model. To start we scraped intermolecular contact data from PICCOLO, an online database containing detailed structural characterizations for 118 of the 179 complexes in the updated SAB. We generated models from PICCOLO data based on three representations of the interface: detailed intermolecular contacts (IC), AAC (count of each

amino acid residue participating in PICCOLO-defined contacts), and lumped residue type (Res Type; count of positive, negative, polar, and apolar residues participating in PICCOLO-defined contacts). We used a combinatorial approach to scan for the best 5-feature models for IC and AAC, and were limited to one 4-feature model for Res Type. **Figure 3.1.3** shows the binding affinity parity plots for the top performing models in each case, trained on all 118 PICCOLO complexes as well as subsets with only rigid (RMSD <= 1.0 Å) or enzyme-containing complexes. Root mean squared error (RMSE) for each model is not included here because of its direct correlation with Pearson's coefficient, r.

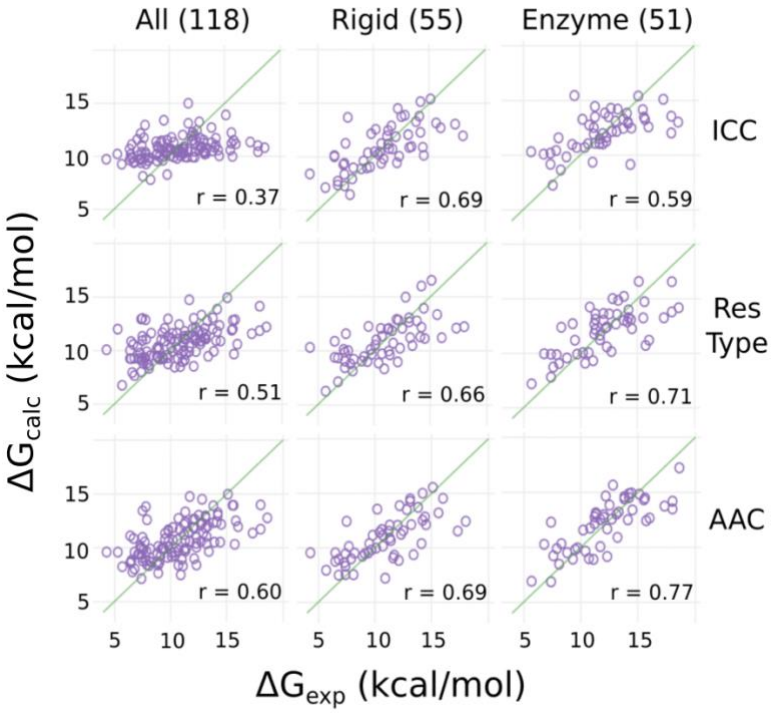

Figure 3.1.3 – Experimental binding free energy given in SAB versus binding free energy calculated with PICCOLO structural data. (top to bottom) Intermolecular contact, residue type, and amino acid composition models. (left to right) Training data set comprised of all 118 complexes intersecting PICCOLO and updated SAB, subset of those 118 with RMSD < 1.0, subset of those 118 containing enzymes.

The AAC framework attains the highest correlation coefficient for all 118 complexes (r_{IC} = 0.60). This coefficient is commensurate to those reached by previously published models on similarly large and diverse training data.[71,73] The IC model performs equally well in the rigid case (r_{IC} = r_{AAC} = 0.69), where its superior representation of geometric compatibility likely improves correlation with experiment. Notably, the IC exhibits worse performance for the enzyme containing (r_{IC} = 0.59) than the larger rigid subset. This may be due in part to the inclusion of flexible residues in the enzyme subset, which is comprised of 27 rigid complexes and 24 flexible complexes. Alternatively, the rules governing binding affinity for these complexes may be better represented in terms of amino acid composition than intermolecular contacts. The chemistry of enzyme-containing interfaces is constrained by the limited number of enzyme catalytic mechanisms. Like amino acids may be substituted to serve the same biological function (i.e. form similar interfacial contacts), but the specific amino acid will affect the strength of these interactions. The IC model is less sensitive to these subtle differences because like sidechains are considered equivalent. AAC outperforms the lumped Res Type model in all cases. Interestingly, consideration of only a few key residues leads to better prediction than counting all amino acids and binning them into polarity groups. **Table 3.1.2** contains a list of

selected features for each of the models plotted in **Figure 3.1.3**, along with their regression coefficients, r, and RMSE. The lack of statistical significance for most IC and Res Type features further supports the superiority of an AAC approach.

Table 3.1.2 – Regression coefficients for IC, AAC, and ResType models trained on PICCOLO data.

Feature	Full SAB			Rigid			Enzyme-containing		
covalent bond	0.278*			-			4.22		
vdw contact	0.002*			0.008*			-		
ionic contact	-0.000*			-			-		
aromatic intrxn	-0.022*			-0.020*			-0.027*		
proximal contact	-0.002*			-0.003*			-		
hydrogen bond	-			-0.236*			-0.340		
vdw clash	-			-0.048*			0.000*		
aploar contact				-			0.031		
MET		0.621			-			0.830	
GLY		-0.380			-0.832			-	
SER		-0.548			-			-0.769	
TYR		-0.267			-			-0.495	
HIS		-0.344*			-			-	
GLU		-			-0.487			-	
GLN		-			-0.384*			-	
ASP		-			-0.213			-	
ASN		-			-			-0.550	
LYS		-			0.367*			0.422	
apol			0.045*			-0.017*			0.144
pol			-0.287			-0.317			-0.403
pos			0.267			0.286*			0.210*
neg			-0.105*			-0.245*			0.026*
Intercept	-8.53	-7.75	-7.94	-4.77	-6.92	-6.60	-10.70	-9.42	-9.18
r	0.37	0.60	0.51	0.69	0.69	0.66	0.59	0.77	0.71
RMSE (kcal/mol)	2.77	2.39	2.56	2.21	2.22	2.30	2.42	1.93	2.13

* $p > 0.05$

These results suggest a residue-level description on interfacial contacts is more appropriate than an atomic-level description but do not demonstrate the full potential of an AAC model. In this case the residue features are derived from intermolecular contact data stored in PICCOLO so that even the AAC model relies on pairwise contacts. We also consider only the 5 most predictive residues in the choose-5 approach. Below we extend our definition of the protein-protein interface to include amino acids that improve prediction without making direct contact across the interface.

Distance Threshold Variation

It has been established that amino acids in the core and remote areas of the protein influence binding.[82] We must first determine whether AAC is robust to and informative across various interface distance cutoffs to qualitatively understand indirect contributions of residues buried deeper in the interface. To address this question, we generated AAC feature vectors for all 179 SAB 2.0 complexes at 13 equally spaced intervals from 4-16 Å (every 1 Å). This range of distances was explored to ensure a span from very short to long range intermolecular interactions and avoid an arbitrarily selected distance cutoff. We trained a LassoLarsIC model for each unique amino acid feature combination (1,048,575) on each feature matrix at each distance. For each distance, we compared the peak

performance (Pearson's correlation coefficient, r) and AAC descriptors selected in top models. The peak performance of the combinatorial approach and r values of 4 selected feature sets for each distance are illustrated in **Figure 3.1.4**. The highest correlation is achieved with slightly different feature sets at each distance since the feature vectors change with cutoff distance. The "best" line thus refers to the highest correlation of any of the 1,048,575 models trained at each distance. In contrast, the other 4 series in the graph follow the correlation coefficients of 4 individual models with distance. These sets correspond to the top performing models, or "winners", at 4, 8, 12, and 16 Å

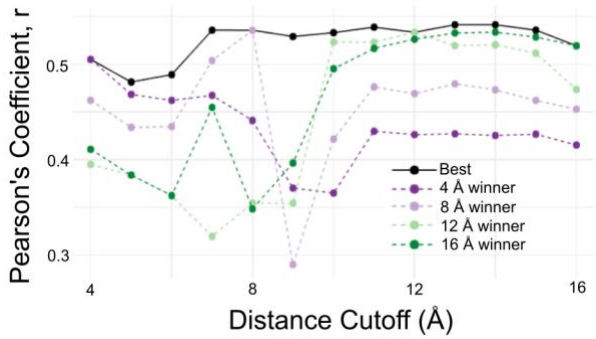

Figure 3.1.4 – Performance of various AAC models trained with different cutoff distances. Black line marks the single highest Pearson's correlation coefficient, r, achieved by any model trained at the specified distance. The other four lines track the correlation coefficients with distance of individual feature sets that were most predictive at 4, 8, 12, and 16 Å. These are the feature sets ICAGSWYH (4 Å winner), VGSWYHEQDKR (8 Å winner), VFCAGSWQN (12 Å winner), VFCMGSWQ (16 Å winner).

The combinatorial approach is robust to distance cutoff. Between 4 and 16 Å, the best AAC model reached a correlation low of 0.48 (5 Å) and high of 0.54 (14 Å), although prediction is consistent from 7 to 15 Å. At thresholds below 4 Å, too few residues were counted for acceptable prediction. Above 16 Å, correlation decreases as the count converges to that of the entire protein. The specific chemistry of winning models changes with distance, so any single AAC model is less robust to distance. This point can be seen in **Figure 3.1.4**, where the fluctuation of individual models is much greater than that of the "best" AAC model. Notably, the two models with superior correlation at short distances (4, 8 Å winners) trade rank with the high-cutoff models (12, 16 Å winners) on either side of **Figure 3.1.4**. Serine, glycine, and tryptophan appear as features in all four individual models, while glutamine and valine appear in all but the 4 Å winner.

Extension to full combination space

We extended our analysis to models at each distance separated by feature set size to better understand the influence on binding of other amino acids. At each of the distance cutoffs described in the previous section, regression models were trained on the same training set for every unique combination of amino acids from 1-20 features (1,048,575 models at each distance). By comparing the regression coefficients of amino acids, we get a qualitative understanding for the interfacial action of each. Negative binding affinities indicate stronger thermodynamic attraction, so negative regression coefficients indicate promotion of binding whereas positive coefficients indicate hindrance. The amino acid regression coefficients are mapped in **Figure 3.1.5** for the top-performing model at each distance cutoff and number of features for the full SAB, and for rigid and flexible subsets. Table 1 in the Materials and Methods section provides a breakdown of the number of models compared for each feature vector length. Residues are ordered from top left to bottom right in order of increasing Kyte-Doolittle hydropathy.[83] Here green, purple, and tan indicate binding, nonbinding, and no action, respectively. The subset heat maps are relatively sparse compared to that for the full SAB because the

BIC penalty for additional features is larger for smaller training sets, so the increased likelihood of zero coefficients being assigned by the LassoLarsIC estimator greater.

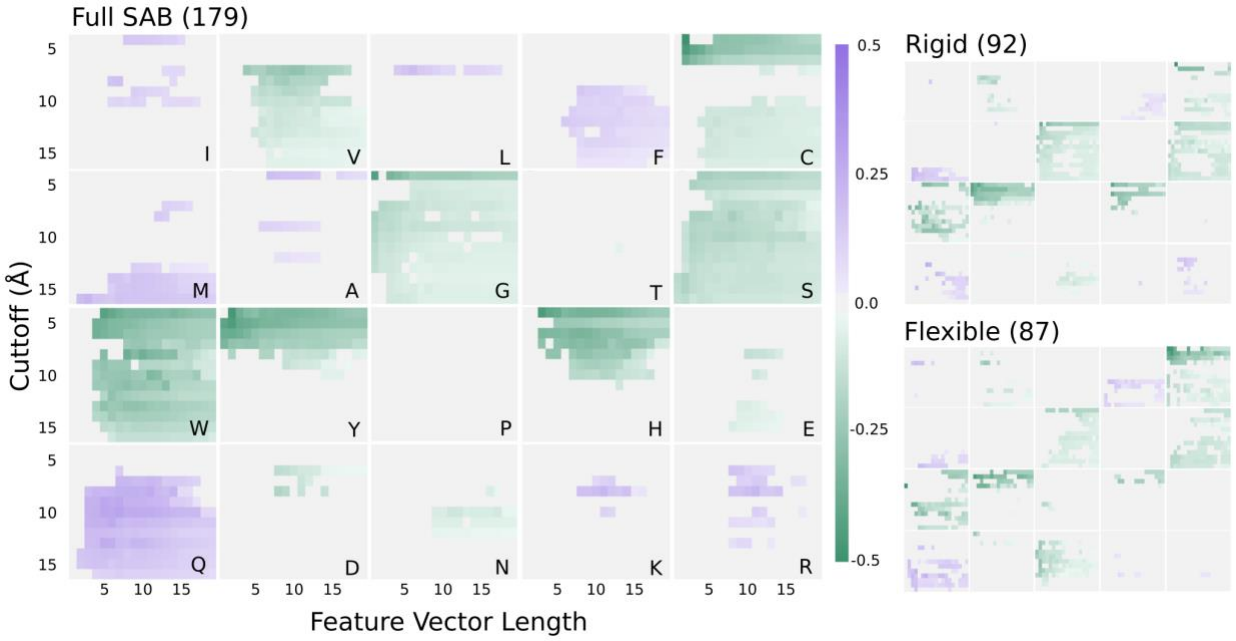

Figure 3.1.5 – Amino acid regression coefficients vs. cutoff distance and feature vector length. (left) Regression coefficients (w_i) for each amino acid feature for models trained on the full SAB (179 complexes) with distance cutoffs from 4-16 Å and 1-19 features. Purple and green are indicative of positive and negative correlation with ΔGexp, respectively. Tan squares indicate where features have been dropped from the model ($w_i = 0$). Similar plots are included for subsets (top right) rigid (RMSD <= 1.0; 92 complexes), (bottom right) flexible (RMSD > 1.0; 87 complexes).

Importantly, there is qualitative agreement for the regression coefficients between the full SAB as well as the disjoint rigid and flexible subsets. A similar plot for the enzyme-containing subset (69 complexes) is included in the supporting information (**Figure I.1**) demonstrating qualitative agreement for all amino acids with the plots included here, along with full-size plots for the rigid and flexible subsets (**Figure I.2**). The qualitative influence of each amino acid feature was confirmed by separate tests of 5-fold cross-validation and bootstrapping. This consistency allows us to draw conclusions about which amino acids actively participate in specific binding, and where in the interface they are most likely to play a significant role.

Contrary to the basis of most intermolecular contact models, "like" residues do not appear to have like influence on BA, with a few exceptions. Charged residue counts (E, D, K, and R) in the interface are not often selected as important features in the LassoLarsIC scheme, especially not in the flexible case. It can be seen in **Figure 3.1.5** that when these residues are included in a top AAC model, the negative residues (E and D) have negative regression coefficients and the positive residues (K and R) have positive regression coefficients. This may seem counterintuitive given popular sentiment that R promotes binding via hot spots.[84–86] While hot spot R residues provide stabilizing energy necessary for binding, they are on average equally prevalent in SAB poses regardless of BA and thus not useful for prediction. Further, there is no clear correlation between hydrophobicity and regression coefficients in **Figure 3.1.5**. Valine, especially 7 Å or more into the interface promotes BA while similarly small and hydrophobic residues leucine and methionine detract from BA. Serine and threonine are very similar in terms of size and polarity, yet their utility in BA prediction is vastly different. Serine is one of the most consistently selected features in top models, whereas threonine is given a very small regression coefficient only once for the full SAB (as evidenced in **Table 3.1.3**).

Table 3.1.3 – Number of models selecting each amino acid (out of 247).

Res	# models	Res	# models
SER	226	ASN	34
GLY	216	ARG	32
TRP	195	ILE	29
GLN	161	ALA	25
CYS	154	GLU	24
VAL	127	ASP	18
HIS	100	LYS	14
TYR	100	LEU	13
PHE	90	THR	1
MET	61	PRO	0

It is also apparent that selected amino acids depend significantly on the distance cutoff. Tyrosine and histidine apparently promote binding at short distance cutoffs, but become less predictive as models consider core residues. These residues may only contribute to binding via direct surface contacts so that their presence becomes less informative as we consider residues deeper in the interface. Phenylalanine and methionine are increasingly selected as the distance threshold is extended to 9 and 13 Å, respectively. Cysteine shows a unique heatmap structure with a gap between the coefficients at for low distance cutoffs and high distance cutoffs. This may be related to distinct interaction modalities, strong surface interactions via its thiol moiety and structural stabilization in the core via disulphide bonds.

Serine, glycine, tryptophan, and glutamine have the most consistent contributions to top performing models across distance and feature vector length (see **Table 3.1.3**). Glycine confers flexibility to the interface, which may be key to predicting the binding affinity of flexible complexes. Serine is a small polar amino acid that has been previously noted for promoting tripeptide aggregation and for having the lowest desolvation energy of all amino acids.[87,88] Glycine and serine are small amino acids that might be undervalued in pairwise contact models. Tryptophan is a bulky aromatic residue capable of pi-stacking, pi-cation, and hydrogen bonding, making it a versatile candidate for favorable interactions across the interface. Glutamine is the most commonly selected feature for the SAB that opposes binding, being selected 161 times out of 247 top models.

Conclusions

In this study, we have demonstrated that interfacial residue content is more useful than intermolecular contacts for understanding and predicting the strength of specific PPIs. The present comparison of AAC and IC models of the interface suggests consideration beyond surface exposed contacts is necessary for structure-affinity prediction. We've identified several regularly overlooked amino acids as important contributors to binding specificity, including serine and glycine. Hydrophobic, electrostatic, and aromatic contacts often described as the drivers of PPIs may depend on less obvious interface features. Flexibility and higher-order interactions play pivotal roles in highly specific interactions in complex media, so future studies must explore geometry more explicitly. Altogether, these results imply the need for efforts focused on more inclusive descriptions than pairwise contacts.

3.2 Conformational preferences of Pos-Gly-Neg tripeptides in aqueous solution relates to biofunctionality

Introduction

Charged amino acids have the greatest potential for strong electrostatic interactions within and between biomolecules, through hydrogen bonding and ion pairing. How this potential manifests itself can be very different in terms of biological function. For example, the RGD sequence motif is known to promote integrin binding and to play an important role in cell adhesion and cell spreading. The simplest explanation for the binding activity of RGD is that it contains two oppositely charged amino acids which can be recognized by, and form strong interactions with, protein residues in a binding pocket. Meanwhile, the ability of proteins to resist nonspecific interactions in crowded environments has been attributed to the prevalence of K and E residues randomly distributed on the protein surface.[66] The corresponding mechanistic explanation is that K and E form very strong interactions with water molecules, thereby resisting interactions with other proteins. The protective nature of strongly hydrated K and E peptides has been harnessed in the form of alternating charge peptides which resist nonspecific interactions.[59,89,90] We have previously demonstrated the functional contrast of similar amino acid moieties with self-assembled monolayer (SAM) surfaces of alternating charge peptides, where EK-based peptides resisted protein adsorption and cell adhesion, while RGD peptides promoted cell adhesion and spreading.[90] The functional duality of these chemically similar peptide motifs requires further characterization.

Two structural criteria have been established which relate to the experimental activity of RGD-containing peptides. The first is a pseudodihedral angle that describes the relative orientation of the positive and negative side chains. Experiments with precisely geometry-constrained RGD cyclic peptides show the sidechains must be oriented on the same side of the peptide to promote binding.[91] Once this criterion is met, a larger distance between the positive and negative charged groups correlates with higher activity. **Figure 3.2.1** shows a schematic representation of the structure-activity relationship for RGD peptides and potential binding scenarios for a linear Pos-Gly-Neg peptide following these criteria.

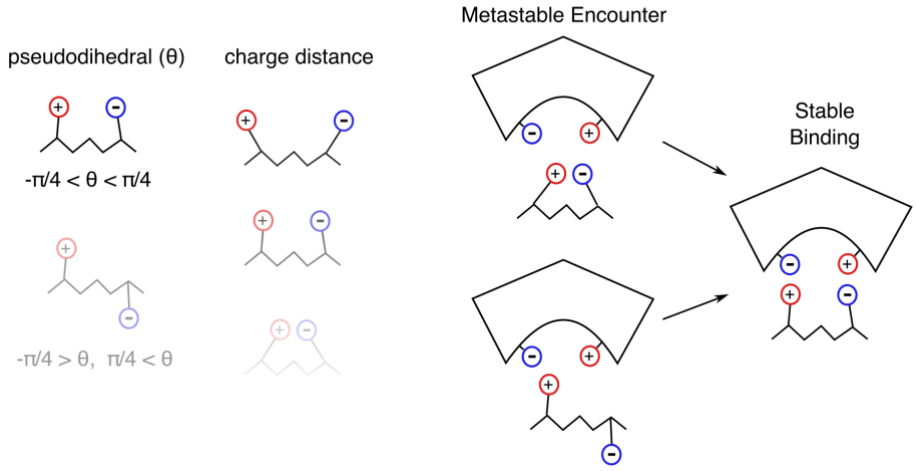

Figure 3.2.1 – Schematic representation of RGD structure affinity relationship and binding scenarios. (left) Experimentally determined structural features related to the binding activity of the RGD structural motif. Darker structures correspond to higher activity. The pseudodihedral angle describing the relative orientation of the positive and negative sidechain must fall in the range $-\pi/4$ to $\pi/4$ radians. Once this criterion is satisfied, activity increases with increasing charge separation. (right) Stable binding for linear RGD peptides may occur through a metastable encounter with a binding partner where the sidechains are properly oriented, but too close for stable binding Alternatively, the distance between sidechains could be initially large, but the sidechains oriented incorrectly.

Given that linear RGD tripeptides are active for many binding targets, the behavior of Pos-Gly-Neg tripeptides in aqueous solution may provide insight into the functional differences RD-peptides and KE-peptides. Using parallel bias metadynamics (PBMetaD) simulations, we found that RGD adopts active conformations in aqueous solution much more frequently than other Pos-Gly-Neg tripeptides. We also found that KGE had the greatest preference for inactive conformations. Our results suggest that the colocation of separable positive and negative charge groups may be responsible for the binding property of RGD, while the preference for hydration over salt bridging between K and E sidechains is indicative of its nonfouling behavior.

Methods

Parallel bias metadynamics (PBMetaD) is an enhanced sampling technique for molecular dynamics (MD) simulation which accelerates sampling by simultaneously biasing multiple slow degrees of freedom.[37] This approach is especially useful for biomolecular simulations, where the complexity of the phenomena of interest typically precludes the existence of a single collective variable that adequately describes the relevant free energy states of a system. We performed PBMetaD simulations with Pos-Gly-Neg tripeptides to bias water coordination, backbone dihedrals, and charged group separation. Here we describe in detail each step of our simulation protocol, including generating starting structures for each tripeptide, unbiased equilibration MD, steered MD, and production PBMetaD. These simulations are summarized in **Table I.5**.

Equilibration

The initial Pos-Gly-Neg peptide configurations, with N-terminal acetyl and C-terminal amide capping groups, were generated using the tleap application.[92] Each subsystem was solvated with TIP3P water.[27] All simulations were performed with the GROMACS 2016.3 simulation engine[28,48]. Energy minimization was performed with 10,000 steps of steepest descent with 1.0 nm cutoff for VDW and coulomb interactions. A 250 ps annealing simulation was then performed using the v-rescale thermostat, increasing the temperature of the system from 5 K to 300 K. Pressure equilibration was performed for 250 ps in the NPT ensemble with the Bussi-Donadio-Parrinello (v-rescale) thermostat[29] and Berendsen barostat[30]. Output configurations and velocities from NPT equilibration were used to initialize the steered MD simulation.

Steered MD

Steered MD was implemented through the Plumed 2.4 software plugin.[49,93] Steered MD was performed for 800 ps in the NPT ensemble with the v-rescale thermostat and Parrinello-Rahman barostat, starting from the output configuration and velocities from equilibration NPT.[29,31] At time zero, a harmonic restraint (k = 100.0 kJ/mol/nm) was placed at 0.3 nm on the distance between the centroid of the positive charge center (all atoms in the guanidino group of R or sidechain amino group of K) and the centroid of the negative group (sidechain carboxylate atoms for E or D). The target distance for the harmonic restraint was increased by 0.2 nm every 100 ps, to a final target distance of 1.7 nm. Eight conformations were selected for PBMetaD by taking a one frame from the steered MD simulation every 100 ps, ensuring the starting configurations for PBMetaD covered a range of charge separation distances.

PBMetaD

PBMetaD with multiple walkers (8 identical systems) was performed for 500 ns for each tripeptide (a total of 4 μs of sampling per tripeptide). The well-tempered variant of the metadynamics algorithm was used for smooth convergence.[94] Collective variables were selected to capture the slow degrees of freedom of the system and biased by depositing 1-dimensional Gaussian hills in each CV

dimension every 500 steps (1 ps) with an initial height of 2.0 kJ/mol, and biasfactor of 10. The distance between the centroids of the positive and negative sidechains was biased with a sigma of 0.02 nm. The coordination number between cation N-bound hydrogens and water oxygens (using a distance switching function to define contacts with $d_0 = 0.19$ nm and $r_0 = 0.03$ nm) was biased with a sigma of 0.25. The coordination number between anion oxygens and water oxygens (using a distance switching function to define contacts with $d_0 = 0.27$ nm and $r_0 = 0.03$ nm) was biased with a sigma of 0.25. Four backbone dihedral angles were biased with a sigma of 0.25 radians, including the psi angle for the positive amino acid, phi and psi angles for the glycine, and the phi angle for the negative amino acid. The unbiased ensemble statistics for each biased simulation were recovered using the Torrie-Valleau reweighting method, as described by Pfaendtner and Bonomi in the original PBMetaD paper.[37]

Results and Discussion

We studied the solution behavior of complementary charge tripeptides to determine whether the bioactive RGD peptide could be distinguished from the bioinert peptides RGE, KGD, and KGE. We used parallel bias metadynamics (PBMetaD) to simultaneously bias the distance between the charge centers of the positive and negative sidechains (1 distance collective variable), the number of water molecules in the first coordination shell of the positive and negative sidechains (2 coordination collective variables), and the torsional angles of all the backbone dihedrals (4 torsion collective variables). This biasing scheme provided quantitative information relevant to the experimental criterion for active RGD conformations by encouraging the system to exhaustively sample side chain orientations and charge center distances. For each peptide, we performed 500 ns of PBMetaD with 8 walkers – identical systems sampling in different parts of phase space, and sharing a bias potential for each CV – for a total of 4 μs of sampling. We used the Torrie-Valleau reweighting method[39] with a quasi-static bias potential, as described in the original PBMetaD paper by Pfaendtner and Bonomi.[37]

RGD adopts active conformation in solution

We wanted to see if the experimentally determined criteria for RGD activity could be used to distinguish RGD from the other complementary charge tripeptides. For each peptide, we calculated the pseudodihedral angle between the positive and negative sidechains (experimental criterion 1) for each frame of the biased simulation and applied the frame weights, determined by reweighting, to recover the unbiased distribution of pseudodihedral angles. We then calculated the two-dimensional free energy surface for the pseudodihedral angle and charge center distance (experimental criterion 2). The two-dimensional free energy surface for each peptide is provided in **Figure 3.2.2**, along with representative structures from the deepest free energy wells for RGD and KGE.

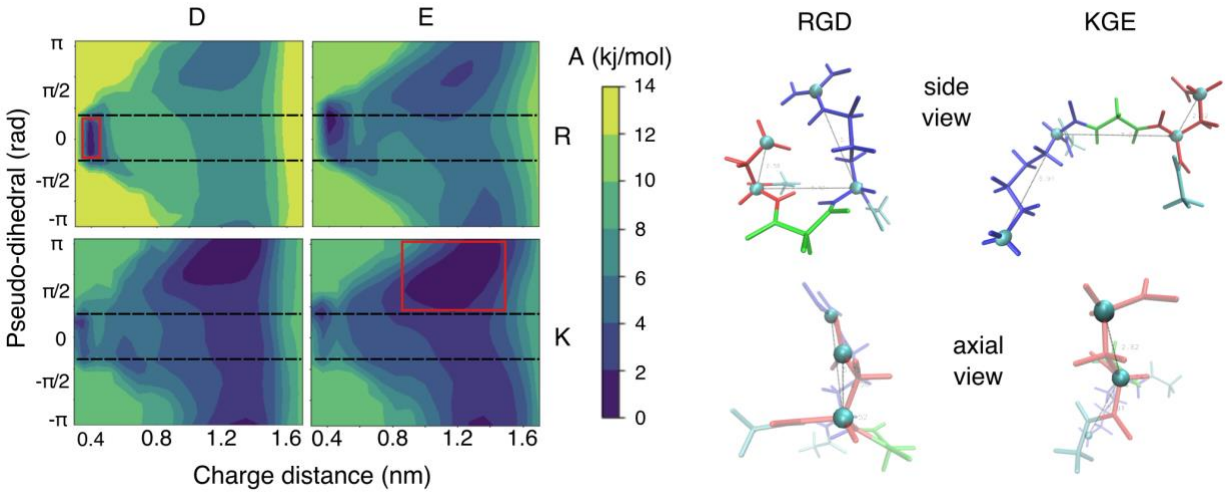

Figure 3.2.2 – RGD satisfies the experimental activity criterion by forming a stable salt-bridge in solution. (left) Two-dimensional free energy surface for RGD (top-left), RGE (top-right), KGD (bottom-left), and KGE (bottom-right). The color map increases in brightness from low to high free energy (A), with an energy shift so that the lowest well for each peptide is at 0 kJ/mol. The area of phase space that satisfies the pseudo-dihedral criterion for an active RGD conformation (-π/4 to π/4 radians) is contained between the horizontal dashed lines. Red squares highlight the deepest energy wells for RGD and KGE. (right) Representative conformations from RGE and KGE free energy wells.

The two-dimensional free energy surfaces in **Figure 3.2.2** show that the RGD peptide forms a stable salt-bridge between the R and D side chains in aqueous solution. This salt bridge ensures that the RGD peptide adopts an active conformation in solution, as defined by the relative orientation of the basic and acidic sidechains. While the lowest point on the RGE free energy surface also represents a salt-bridged conformation, the 1D free-energy surface for distance shows that the free energy of this salt bridge is half that of the RGD salt bridge (**Figure I.6**). In contrast, KGD and KGE are more stable in the unbound state, with large free energy wells corresponding to sidechains orientation outside of the active range. Interestingly, the KGE peptide, which contains the positive and negative sidechains we have deemed responsible for the resistance of proteins to nonspecific interactions, has the lowest probability of adopting a conformation with relative sidechain orientations in the active range.

Conclusions

We demonstrate that the conformational preferences of Pos-Gly-Neg tripeptides in aqueous solution are indicative of functionality by comparing bioactive and bioinert tripeptides with MD simulations. The colocation of charge groups that can bind separately to a protein provides the potential for the RGD motif to interact strongly with binding pockets. Nonfouling materials should be composed of alternating charges that prefer not to associate (K and E) or charges that cannot be separated (zwitterionic materials). RGD, the example of specific binding, thus provides insight into the relation between natural and synthetic nonfouling materials. In the next section, we describe how salt bridging in larger peptides leads to conformational preferences that interfere with functionality in drug delivery applications.

3.3 Characterizing the conformational ensembles of charge-alternating protective peptides*

Introduction

Recombinant fusion proteins provide a scalable synthesis platform for engineered biologics, whereby a polypeptide domain is appended to alter the physical characteristics of a therapeutic protein and enhance its pharmaceutical viability. Genetic fusion of an engineered polypeptide domain to a biologic eliminates the additional conjugation and purification steps required with a synthetic polymer, decreasing the time and cost of synthesis at scale. Using a polypeptide chain, composed of natural amino acids, is also inherently biodegradable. However, in dealing with biological building blocks, extra care must be taken to avoid unintended interactions with the biological machinery *in vivo*. The vast combinatorial space of protein sequences poses an additional challenge, as it provides a virtually endless design landscape for protective polypeptide domains.

Preliminary attempts to design protective polypeptides have been directly inspired by the most successful polymers in the chemical conjugation literature. "Conformationally disordered" peptides have been designed to mimic the random structure and large hydrodynamic radius of PEG.[95–97] This class of peptide aims to reduce renal clearance, via increased size, and reduce immunogenicity, via entropic repulsion of antibodies. Conformationally disordered peptides include HAP (polyG), PAS, and XTEN.[95–98] These peptides have been shown to modestly improve the circulation half-life of hGH in a mouse model.[96] Much like PEG, these peptides are amphiphilic. Alternating-charge peptides have been designed to mimic the superhydrophilicity of the zwitterionic polymers PCB. Superhydrophilicity, or extremely strong affinity for water, can also provide a large hydrodynamic radius for reduced renal clearance and provides an enthalpic repulsion of antibodies. The alternating-charge peptide, polyEK, has been shown to increase the thermal stability of beta-lactamase enzyme without decreasing its catalytic rate.[59] More recently, polyEK was shown to increase the efficiency of an organophosphate hydrolase.[60] The conformation of polyEK has not been characterized in detail. A polypeptide domain combining disorder and superhydrophilicity would have the desirable feature of entropic and enthalpic resistances to unintended biological interactions.

There is reason to believe polyEK, despite being near optimal in terms of water affinity, will not achieve the same degree of conformational disorder as other peptide domains or synthetic polymers. Conformationally disordered peptides such as HAP were intentionally designed with amphiphilic amino acid residues with low potential for sidechain-sidechain interactions. PCB (the inspiration for polyEK) is expected to assume disordered conformations due to repulsive interactions between its zwitterionic headgroups. However, salt-bridging between the positive and negative sidechains of polyEK could potentially stabilize regular secondary structures. Indeed, closely related peptide sequences have been used to create ionic self-complimentary peptides (EAK and RAD), which form β-sheet structures in water and macroscopic aggregates at physiological salt concentrations.[99,100] It is hypothesized that these structures are stabilized by extended salt-bridging. A tendency toward regular secondary structure and aggregation could be detrimental to the protective performance of a polypeptide domain, potentially inciting peptide-specific immunogenicity and decreasing the stability of concentrated fusion protein drug formulations, respectively. If polyEK assumes an ordered secondary structure, strategies should be developed to provide conformational disorder to this superhydrophilic template.

In the present work, we provide evidence from simulation and experiment that $(EK)_{15}$, a computationally tractable polyEK surrogate, forms stable secondary structure in aqueous solution. We

* Reproduced in part with permission from J. Smith, J. Pfaendtner, and S. Jiang. Molecular dynamics simulation and circular dichroism studies of the conformations of charge-alternating peptides. In preparation.

identify a high propensity for β-strand structures in $(EK)_{15}$ using enhanced sampling molecular dynamics (MD) simulations. We also find that the dilution of the superhydrophilic $(EK)_{15}$ construct with glycine (G) residues results in conformationally disordered, alternating-charge peptides. We confirm with circular dichroism spectroscopy the presence and absence of regular secondary structure in $(EK)_{15}$ and $(EKGG)_7$, respectively. We conclude that polyEK, while near optimal in terms of hydrophilicity, explores a restricted conformational space due to strong side chain interactions. Adding glycine or small polar amino acids, such as serine or asparagine, can be used as a strategy to add flexibility and prevent the formation of stable or metastable secondary structures. Further research should be performed to explore the seeming trade-off between superhydrophilicity and structural diversity.

Computational Methods

Classical molecular dynamics (MD) simulations cannot effectively capture the long time and length scales characteristic of protein conformational changes. Enhanced sampling techniques for the characterization of protein and peptide structural ensembles, however, are well-established.[101–103] In the present work, we use parallel tempering metadynamics in the well-tempered ensemble (PTMetaD-WTE), which has previously been used to characterize the kinetics of folding pathways for small proteins and the ensemble of structures adopted for inherently unstructured peptides.[103–105] Our simulation protocol included several equilibration steps to improve sampling efficiency, including selecting starting structures in different areas of conformational phase space, unbiased equilibration MD, parallel tempering in the well-tempered ensemble (PT-WTE) to improve exchange, and production parallel tempering metadynamics (PTMetaD). We separately describe each step of the enhanced sampling simulation pipeline below and summarize these simulations in **Table I.6** and **Table I.7**.

Selection and equilibration of diverse starting structures

Starting structures were selected from 30 residue segments of real protein structures deposited in the PDB.[68] Twelve structures were selected with diverse radii of gyration and secondary structure based on visual inspection in VMD.[106] Backbone coordinates were saved from VMD and residue names were changed to match the appropriate EK-based peptide sequence. The backbone conformations of the selected structures are depicted in **Figure 3.3.1**. Sidechains were added for E and K residues based on the Amber ff14SB force field[107] with the tleap application.

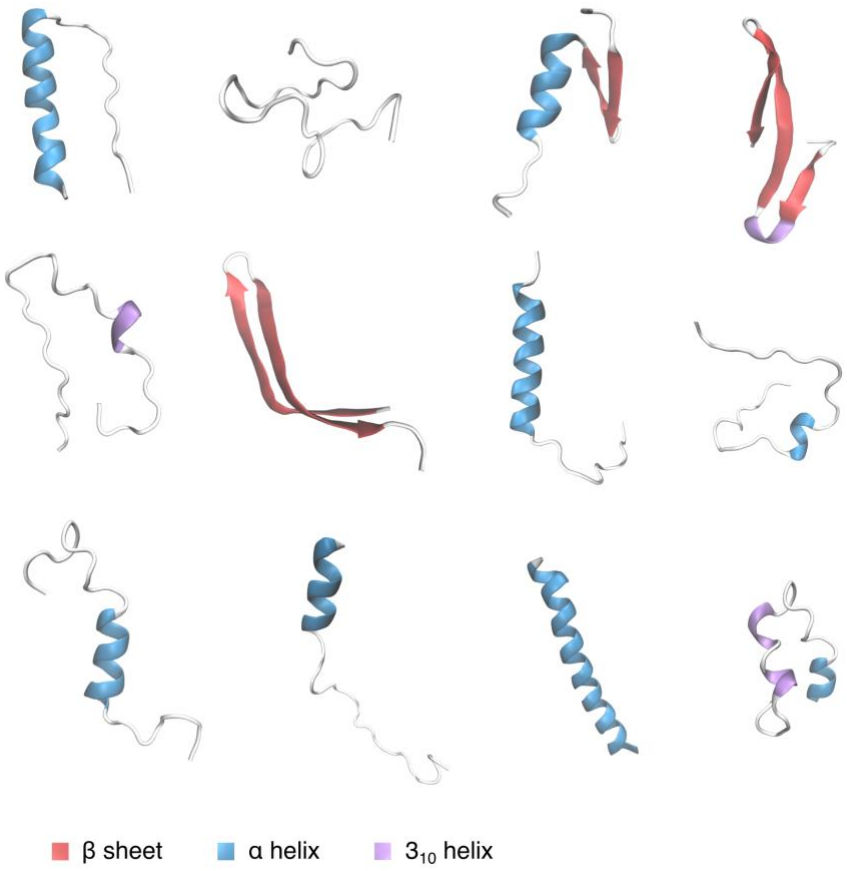

Figure 3.3.1 – Starting structures for PTMetaD-WTE. $(EK)_{15}$ and G-substituted variants started with the same 12 backbone structures. These backbone segments were selected from the protein XYZ crystal structure (protein selected arbitrarily), after being manually identified as diverse in secondary structure and radius of gyration. Secondary structures not explicitly labeled at the bottom of the figure (ie. coil, turn, bridge) are colored white.

Each subsystem was solvated with TIP3P water[27]. All simulations were performed with the GROMACS 5.1.2 simulation engine[28,48], with PT-WTE and PTMetaD implemented through the Plumed 2.3 plugin[49,93]. Energy minimization was performed with 10,000 steps of steepest descent with 1.0 nm cutoff for VDW and Coulomb interactions. Solvent equilibration was performed for 250 ps in the NPT ensemble with the Bussi-Donadio-Parrinello (v-rescale) thermostat[29] and Berendsen barostat[30]. The largest box volume of any subsystem at the end of NPT equilibration was assigned to all NPT outputs to ensure identical subsystems for replica exchange in PT-WTE. From NPT output, a 250 ps annealing simulation was performed using the v-rescale thermostat which smoothly increased the coupling temperature for each replica from 5 K to its PT-WTE starting temperature (between 300-450 K). This preliminary annealing step ensures configurations and velocities fed to PT-WTE are appropriate for initial temperatures and prevents the deposition of large potential energy bias for unequilibrated structures.

Parallel-tempering in the well-tempered ensemble (PT-WTE)

PT-WTE was performed as described by Bonomi and Parinello.[103] The 12 replicas were initialized at 12 temperatures geometrically distributed ($T_{n+1}/T_n = \text{const}$) between 300 K and 450 K. Every 250 MD steps (0.5 ps), the MD engine attempts to exchange the atomic coordinates of replicas

at neighboring temperatures based on a Boltzmann exchange probability. Gaussian bias hills with height 4.18 kJ/mol and bias factor 30 are added to the potential energy of the system every 250 steps (0.5 ps). PT-WTE runs for EK and EKGGG peptides were performed for initial exploration of accessible conformational space and extended 300 ns. Because the exchange probability for these simulations leveled off before 100 ns, PT-WTE was limited to 100 ns to build bias potentials for EKG, EGKG, EKGG, and GG. Output structures and potential energy bias for each ensemble were passed forward to PTMetaD for more efficient replica exchange. **Figure I.7** shows that all replicas in the PTMetaD simulation regularly visited each of the 12 temperature levels.

Parallel tempering metadynamics (PTMetaD)

PTMetaD was used with the same temperature distribution, and with output conformations and bias potentials from PT-WTE. An identical exchange protocol was employed for parallel-tempering, and the radius of gyration was biased to force peptides to explore extended and collapsed configurations. This approach encourages each system to explore a range of compact and extended states, without biasing towards a specific type of secondary structure. Bias was deposited with a stride of 250 steps (0.5 ps), an initial hill height of 2.0 kJ/mol, and bias factor of 15 for radius of gyration. Sprenger et al. showed infrequently updating the potential energy bias can improve sampling efficiency,[108] so we deposited potential energy bias every 250000 steps (500 ps), with an initial hill height of 4.18 kJ/mol, and a bias factor of 30.

Unbiased production MD simulations

Long unbiased production MD simulations were performed to assess the stability of a seemingly stable EK structure. Microsecond long NPT simulations using the Parrinello-Rahman barostat[31] and v-rescale thermostat[29] were performed for EK at 300 K and 450 K with 16766 water molecules. Shorter simulations (500 ns) were also conducted with identical MD inputs excepting mutations in the peptide sequence to DK and EKG. All peptides studied were net neutral and simulated in pure water. We also checked the stability effects of salt concentration by simulating EK in 0.154 M and 0.308 M solutions of NaCl (corresponding to approximately once and twice the concentration of normal saline, respectively), and found no destabilization in salt solutions.

Analysis of biased trajectories

The PTMetaD method produces biased statistical ensembles of the conformations sampled at each temperature. We used the time-independent free energy estimator and reweighting scheme proposed by Tiwary and Parrinello to recover the unbiased statistics for each conformational ensemble.[109] We applied the resultant frame weights to calculate ensemble average observables for each peptide at each temperature. All trajectory analysis (including reweighting) was performed using in-house python scripts and analysis functions in the MDTraj python library.[52] We used the DSSP algorithm for residue-wise structure assignment, as implemented in `mdtraj.compute_dssp`. We used the Shrake and Rupley algorithm to calculate solvent accessible surface area (SASA), as implemented in `mdtraj.shrake_rupley`. We used the `mdtraj.rmsd` function to calculate the Cα root mean squared deviation (RMSD) for unbiased and biased trajectories, and an in-house implementation of the clustering method introduced by Daura et al. (sometimes called the gromos method) to identify the lowest-energy structures.[110] The python scripts for performing these analyses are publicly available at https://github.com/UWPRG/ek-conformation-project.

Experimental Materials and Methods

Peptide materials and synthesis

Fmoc-Glu(tBu)-OH, H-Lys(Boc)-OH, Fmoc-Gly-OH, ethyl cyanohydroxyiminoacetate (Oxyma) and Fmoc-Rink Amide resin (0.54 mmol/g) were purchased from AAPPTec. N,N'-diisopropylcarbodiimide (DIC) and piperidine were obtained from Chemimpex. Trifluoroacetic acid (TFA) and 3,6-dioxa-1,8-octanedithiol (DODT) were purchased from TCI America. Diisopropylethylamine (DIEA), N,N-dimethylformamide (DMF) and triisopropylsilane (TIS) were obtained from Sigma–Aldrich.

An amino acid dimer (Fmoc-Glu(tBu)-Lys(Boc)-OH) was first obtained by reacting Fmoc-Glu(tBu)-OH and H-Lys(Boc)-OH and used as EK monomer during the synthesis of (EK)10 and (EKGG)7. Microwave-assisted peptide syntheses were performed using Fmoc Solid-Phase Peptide Synthesis (SPPS) strategy on a Microwave Peptide Synthesizer (CEM Corporation, Matthews, NC). Default standard 90°C deprotection and coupling methods were used for synthesis. Peptide preparation was set at scale of 0.1 mmol and starting on Fmoc-Rink Amide resin. Fmoc deprotection was achieved in 20% piperidine DMF solution and the final peptides cleavage were performed in a TFA/scavenger cocktail (TFA/TIS/H2O/DODT = 92.5/2.5/2.5/2.5) for 3 hours at room temperature.

Circular dichroism

Peptides were dissolved in neat water to a concentration of 150 μM and pH adjusted to 7.4. The resulting solutions were transferred to a 1.0 mm thickness quartz cuvette for analysis on Jasco J-720 Circular Dichroism machine. For each condition, eight replicate spectra were obtained and averaged with wavelengths ranging from 215 nm to 250 nm. The resulting spectra was smoothed using the Jasco circular dichroism software. For conditions with urea, urea was added in 2M increments to the peptide containing solutions, and circular dichroism spectra were obtained as previously described.

Results and Discussion

We used PTMetaD simulations to assess the conformational freedom of polyEK. We performed atomistic enhanced sampling simulations of an EK 30-mer, $(EK)_{15}$, as a surrogate for longer polyEK peptides because it would be computationally infeasible to simulate the 30 kDa peptide used in experiment. Our foundational assumption was that the conformational ensemble of this relatively short peptide will provide insight into the structural preferences of polyEK. We believe that this assumption is well-founded given that the chemistry of the surrogate peptide should capture local preferences for water binding or self-interaction. We performed identical PTMetaD simulations for a 30-residue glycine peptide (G_{30}) for comparison. We chose to use this polyG surrogate as a conformationally-disordered control because glycine-rich peptides are known to adopt predominantly random coil structures in aqueous solution. Also, the glycine-based homo amino acid peptide (HAP), the first reported conformationally disordered protective polypeptide domain, was designed based on this observation.[95] We routinely refer to the set of structures visited during the PTMetaD simulation for a given sequence, at a given temperature, as a "conformational ensemble". For example, we constructed 12 conformational ensembles for $(EK)_{15}$ by sampling in parallel at 12 temperatures. While the low temperature ensembles are of the most practical interest, corresponding to ambient experimental and physiological conditions, high temperature ensembles can provide additional informational about the thermal stability of secondary structure motifs.

$(EK)_{15}$ samples a restricted number of structures relative to disordered G_{30}

Our primary objective was to characterize the relative stability of collapsed versus extended and structured versus unstructured conformations of alternating-charge EK-based peptides. We

performed PTMetaD simulations for accelerated sampling of the vast conformational space, biasing the radius of gyration of the peptide backbone and exchanging replicas between 12 temperatures geometrically distributed from 300 K to 450 K. This approach provided a biased conformational ensemble at each temperature for $(EK)_{15}$ and the G_{30} control. We recovered unbiased statistics from each biased conformational ensemble by reweighting with the time-independent estimator proposed by Tiwary and Parrinello.[109] We calculated the ensemble-averaged secondary structure content, the backbone conformational entropy, the number of salt bridges, and the hydrophilic solvent accessible surface area (SASA) for each ensemble. We have tabulated these ensemble-averaged physical characteristics across the range of temperatures for $(EK)_{15}$ and for G_{30} in **Table I.8**. **Figure 3.3.2** shows the secondary structure content of the conformational ensemble generated at all 12 temperatures for $(EK)_{15}$ and G_{30}. In this case, we used the simplified DSSP structure classification scheme, so that residues with loop, turn, and irregular structures are assigned to "coil", residues participating in isolated β-bridges or extended β-strands (ladders) are assigned to "extended", and residues in α helix, π helix, or 3_{10} helix are assigned to "helix". We provide the detailed breakdown of these structural classes to their component parts in **Figure I.8**.

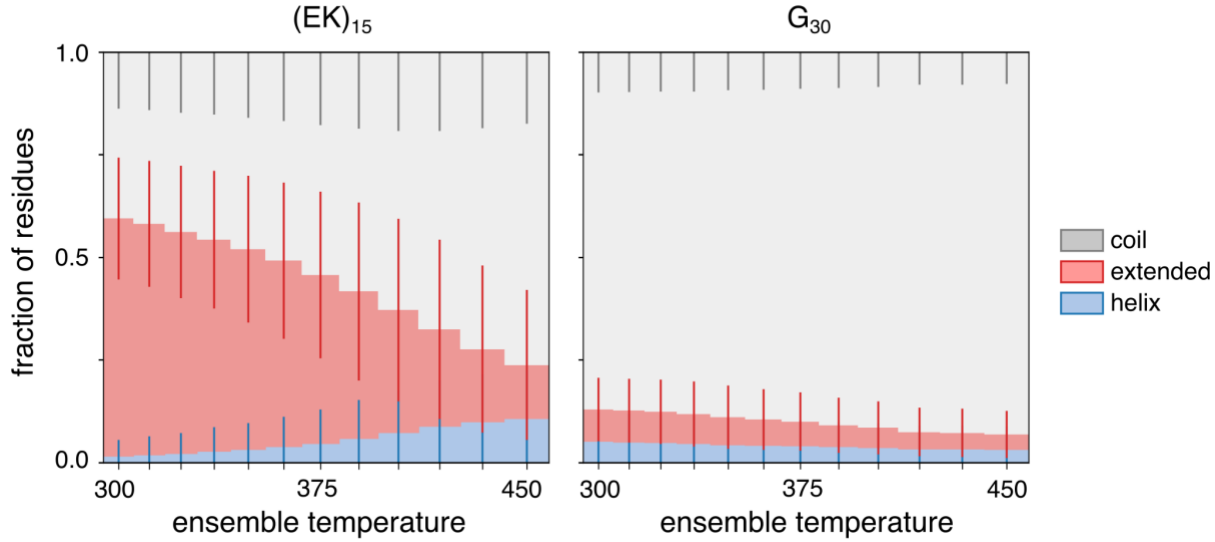

Figure 3.3.2 – Average fraction of (EK)15 and G30 residues assigned to each of the simplified structure classes using the DSSP algorithm. Stacked bars are included for the structural ensemble sampled at each of the 12 temperatures during the PTMetaD simulation. The gray, red, and blue fractions of the bar represent the fraction of residues assigned to coil, extended, and helix structure classes, respectively. For example, the average fraction of residues assigned to coil, extended, and helix for (EK)15 at 300 K were 0.41, 0.58, and 0.01, respectively. Error bars represent +/- 1 standard deviation.

In line with theoretical[111] and experimental[95,112] characterizations of the structure of glycine-rich peptides in aqueous solution, an average of 87.5% of G_{30} residues in the conformations sampled at 300 K were assigned to the coil structure class. Practically, this means that an average of only 3.75 out of 30 residues were assigned to the extended or helix class for each conformation in the 300 K conformational ensemble. The small fraction of residues assigned to extended and helix structures for G_{30} are likely due to chance alignments of a few residues in the primarily disordered peptide chain.[113] Indeed, performing DSSP with the expanded list of secondary structure types shows the "extended" fraction of G_{30} is primarily composed of β-bridges rather than β-strands (**Figure I.8**).

In contrast, only 41% of $(EK)_{15}$ residues were assigned to the coil structure class at 300 K. For $(EK)_{15}$, about 17.4 of 30 residues were assigned to the extended class, suggesting the presence of

persistent, extended β-strands. The high fraction of $(EK)_{15}$ residues participating in β-strands is seemingly at odds with empirical evidence that suggests E and K are underrepresented in β-sheets compared to most amino acids.[114–116] However, it has been observed experimentally and explained theoretically that β-strand propensities are largely context dependent.[117–119] Because long, uninterrupted runs of charged amino acids are uncommon in natural proteins, empirical estimates of the β-sheet propensities for E and K may underestimate the likelihood of β-sheet formation in the context of charge-alternating peptides. It has also been suggested that pairwise sidechain interactions are required to stabilize β-strands.[118] Since salt bridges represent the strongest non-covalent sidechain interactions, it may be reasonable to suspect that E-K sidechain interactions provide stabilizing energy for the observed β-strand frequency. The potential for $(EK)_{15}$ to form β-strand configurations would also be consistent with the observation that the ionic self-complementary peptide EAK forms β-sheets in aqueous solution, despite the high propensity of alanine for α-helix (and slight preference of E and K for α-helix) observed in globular proteins and polyalanine peptides.[100,115,120]

Hyperstable β-strand moiety dominates $(EK)_{15}$ conformation in aqueous solution

We performed a clustering analysis to identify the molecular features of stable extended structures in the conformational ensemble of $(EK)_{15}$ at 300 K. We calculated the pairwise C_α root mean squared distance (RMSD) for every pair of structures in the ensemble, resulting in an NxN distance matrix, where N is the number of structures in the ensemble. We followed the peptide structure clustering algorithm of Daura et al.[110] (sometimes referred to as the "gromos" method) with a RMSD cutoff of 0.5 nm. This algorithm identifies the structure with the greatest number of neighbors (other structures with C_α RMSD < 0.5 nm), and designates that structure as the "central structure" of a cluster comprised of all its neighbors. These structures are removed from consideration and the algorithm iteratively identifies clusters and their central structures until no neighboring structures remain. After clustering, each cluster was assigned a probability based on the number of constituent structures and their associated unbiased weights. **Figure 3.3.3** shows the cumulative distribution function of structures for $(EK)_{15}$ (at 300, 374, and 450 K) and for G_{30} (at 300 K), in terms of the number of clusters required to account for each structure given the 0.5 nm RMSD cutoff. **Figure 3.3.3** also shows representative structures for the three most populated clusters for $(EK)_{15}$ and G_{30} at 300 K.

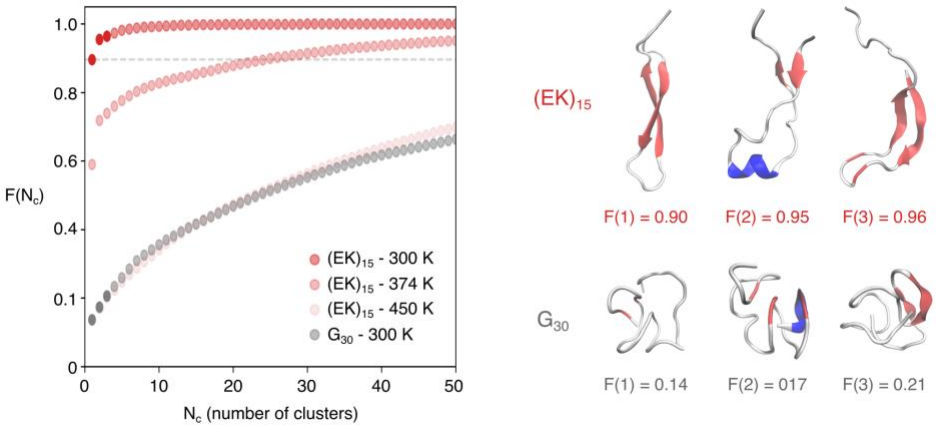

Figure 3.3.3 - Cumulative distribution function for structure clusters at each temperature from PT-MetaD simulation. Each cluster contains structures with a Cα RMSD less than 0.5 nm from the central structure of the cluster, identified using the gromos clustering method.72 The horizontal dashed line corresponds to the probability that structures in the (EK)15 300 K ensemble fall into the most likely cluster (~0.90). For the conformational ensembles of (EK)15 and G30 at 300 K, the central structures of the top three clusters are included on the right. These structures correspond to the first three points in the respective CDFs on the left, which are highlighted with increased opacity.

The clustering results show that $(EK)_{15}$ explored a very restricted set of conformations at 300 K due to the predominance of a twisted, antiparallel β-strand moiety. The likelihood that a structure selected at random from the 300 K ensemble would contain this moiety was approximately 0.9. As expected, persistent secondary structures appear to be stabilized by extensive sidechain salt-bridging. By inspecting the central structure of the most populous cluster for (EK)15 at 300 K, we found that the β-strand moiety was stabilized by alternating salt-bridges forming a zipper-like pattern on either side of the strand. The presence of antiparallel β-strand moieties in the second and third most populous clusters further supports the propensity of $(EK)_{15}$ for antiparallel β-strand structure. While transient β-bridges and short helices will unavoidably occur in polypeptide domains, the persistent secondary structure, like the β-strand moiety identified here, could be deleterious for fusion protein applications in terms of immune recognition and quaternary aggregation. Aggregation was not reported for *in vitro* studies of polyEK fusion proteins,[59,60] suggesting unintended interactions with the biological machinery *in vivo* may pose the bigger threat. Further engineering may be required to disrupt stable secondary structures and encourage a more entropically favorable disordered conformation.

Glycine mutations disrupt stable β-sheet structure in unbiased simulations

To further investigate the stability of the β-strand structure identified in PTMetaD simulations, we performed long classical MD simulations of $(EK)_{15}$ at 300 K and 450 K. We also simulated two peptides derived from $(EK)_{15}$ to assess the potential for amino acid substitutions to disrupt the stable conformation. For each of these peptides, we started with the same β-sheet backbone conformation as identified for $(EK)_{15}$, then mutated some of the sidechains. For the first mutant, we replaced all the negatively charged E sidechains to aspartate (D). We refer to this mutant as DK. We chose not to mutate positively charged K sidechains to arginine (R) sidechains considering the potential role of R residues in protein-protein binding, evidenced by their prevalence in protein-protein interfaces and their role as hot spot residues in such interfaces.[84,121] For the second peptide, we removed approximately half of the EK sidechains, resulting in the sequence $(EKGG)_4G_2(EKGG)_3$. We refer to this peptide as EKGG, although the EKGG repeat sequence is disrupted at the turn β-strand to preserve the maximum number of initial E-K salt bridges in the β-strand portion. **Figure 3.3.4** shows the C_α RMSD for the β-sheet residues over the first 500 ns of each unbiased simulation.

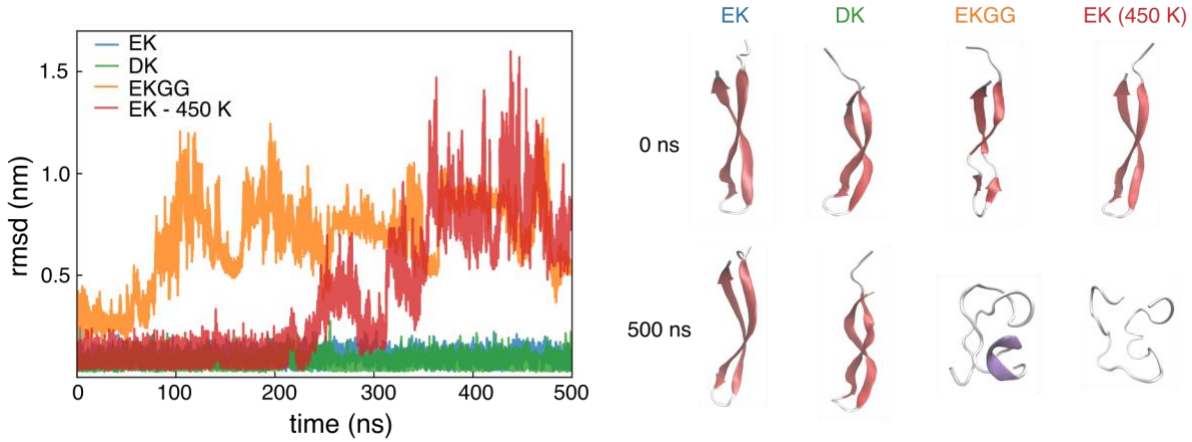

Figure 3.3.4 – C_α RMSD for $(EK)_{15}$ and DK/EKGG variants from 500 ns of unbiased simulations. Starting from the stable antiparallel β-strand conformation identified as stable in the PTMetaD simulations. The unchanged $(EK)_{15}$ peptide remained stable for the whole simulation at 300 K (green) and began to unfold after 200 ns at 450 K (red). The $(DK)_{15}$ peptide also remained folded for the entire simulation at 300 K (blue). The EKGG mutant, with half as many E-K salt bridges to start, unfolded in less than 100 ns at 300 K (orange). Conformations for each sequence at 0 ns and 500 ns are included on the right.

The β-strand conformation for $(EK)_{15}$ remained stable for over 1 μs at 300 K. **Figure 3.3.4** shows it took 200 ns to disrupt this conformation of $(EK)_{15}$ at 450 K, explaining the persistence of this structure in the 300 ns PTMetaD simulation (**Figure 3.3.2**). The DK peptide also remained folded for over 500 ns, while the EKGG peptide unfolded completely in less than 100 ns. With slightly greater than 50% of the EK pairs mutated to GG, the likelihood of cooperative salt-bridging and the formation of extended β-strands is much lower. These results suggest that dilution with G may be an effective design strategy for protective polypeptide domains that are both disordered and superhydrophilic.

To test how the dilution factor effects the conformational disorder and hydrophilicity of EK-based peptides, we performed PTMetaD simulations for several G-substituted variants of $(EK)_{15}$. These peptides were $(EKG)_{10}$, $(EKGG)_7$, $(EGKG)_7$ and $(EKGGG)_6$. For each sequence, we used the same PTMetaD protocol (and the same 12 starting structures) as described for $(EK)_{15}$ and G_{30}. We also performed identical analyses to calculate ensemble averages for secondary structure content, backbone entropy, and hydrophilic SASA. **Figure 3.3.5** shows a comparison of these physical features, at 300 K, for each sequence. Backbone entropy was scaled so that the values for G_{30} and $(EK)_{15}$ were set to 1 and 0, respectively. Hydrophilic SASA fraction was scaled so that the values for G_{30} and $(EK)_{15}$ were set to 0 and 1, respectively.

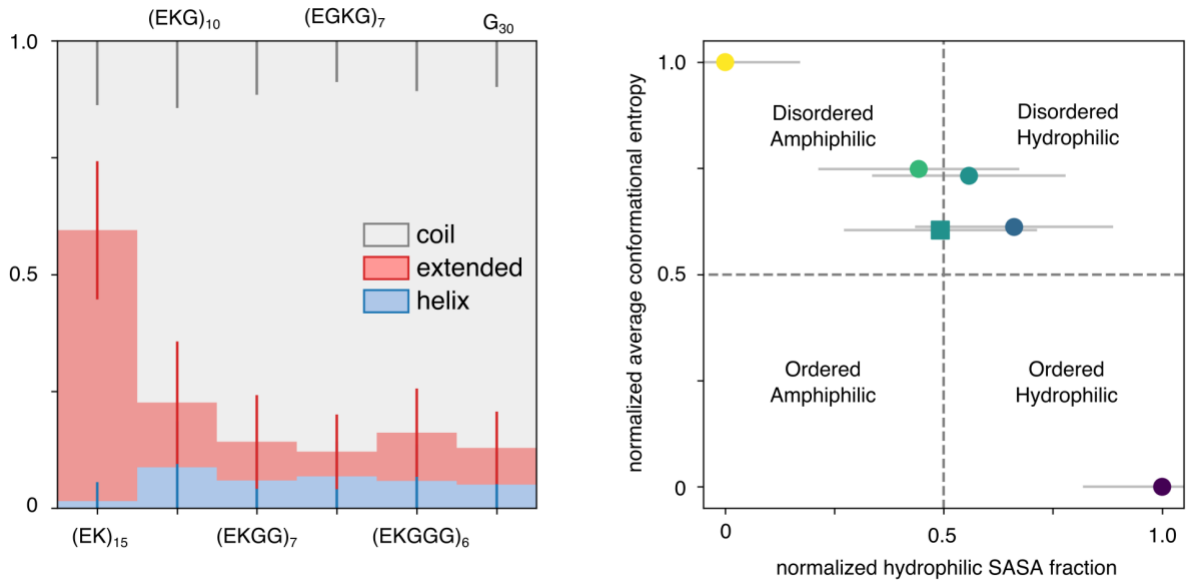

Figure 3.3.5 – Ensemble average secondary structure fractions and physical properties of $(EK)_{15}$, $(EKG)_{10}$, $(EKGG)_7$, $(EGKG)_7$, $(EKGGG)_6$, and $(G)_{30}$ at 300 K. (left) The gray, red, and blue fractions of the bar represent the fraction of residues assigned to coil, extended, and helix structure classes, respectively. (right) Average conformational entropy vs. fraction hydrophilic SASA. Conformational entropy is normalized so that the maximum and minimum values correspond to G_{30} and $(EK)_{15}$, respectively. Conversely, the fraction hydrophilic SASA is normalized so that the maximum and minimum values correspond to $(EK)_{15}$ and G_{30}, respectively. High values of disorder and hydrophilicity are desirable. All error bars represent +/- 1 standard deviation.

The 300 K conformational ensemble for each G-diluted sequence contained significantly less ordered secondary structure than for $(EK)_{15}$. $(EKGG)_7$ and $(EGKG)_7$ were especially close to G_{30} (0.87 +/- 0.10) in terms of coil structure fraction with 0.86 (+/- 0.11) and 0.88 (+/- 0.09), respectively. The drastic decrease in ordered secondary structure, even for $(EKG)_{10}$, suggests that a small dilution factor may effectively destabilize β-strands in EK-based peptides (**Figure I.9**). As expected, the increased conformational freedom came at the cost of hydrophilicity. While progressive dilution with

glycine increased the ensemble average backbone entropy relative to $(EK)_{15}$, the average hydrophilic SASA fraction was decreased. However, we also found that for $(EK)_{15}$ salt bridges generally decreased hydrophilic SASA (**Figure I.10**). It may be possible to engineer sequences with greater hydrophilicity and disorder using small polar amino acids, like serine and glutamine, or proline in place of glycine.

The slight increase in mean backbone entropy for $(EKGGG)_6$ relative to $(EKGG)_7$ suggests there may be diminishing entropic returns beyond 50 % dilution. Also noteworthy is the difference between $(EKGG)_7$ and $(EGKG)_7$ in terms of backbone entropy, which suggests that sequence and dilution factor must be considered separately even within the simplest contexts of peptide design. It is expected that the conformational ensemble of $(EGKG)_7$ is more restricted than that for $(EKGG)_7$ because the isolated E and K residues are more amenable to tightly packed random coils than the EK pair. This hypothesis is supported by the lower average radius of gyration and greater number of G-G contacts for $(EGKG)_7$ relative to $(EKGG)_7$. The mean scaled backbone entropy and scaled hydrophilic SASA fraction were both greater than 0.5 for $(EKG)_{10}$ and for $(EKGG)_7$, landing these sequences in the "Disordered Hydrophilic" quadrant of **Figure 3.3.5b**. Given that $(EKG)_{10}$ contains the repetitive X-G-Y moiety characteristic of the collagen triple helix, its potential for intermolecular interactions may be greater than predicted by single-molecule simulations.

Circular dichroism confirms presence of β sheet in $(EK)_{15}$ and structure breaking in $(EKGG)_7$

We performed circular dichroism spectroscopy (CD) to verify the structural predictions from our simulations. We synthesized $(EK)_{15}$ and $(EKGG)_7$, and measured their CD spectra in neat water, 2 M urea, 4 M urea, and 6 M urea solutions. Changes in the CD spectra with additional urea should reflect a shift toward a disordered, denatured state. We selected $(EKGG)_7$ as the diluted variant of $(EK)_{15}$ based on our observation that 50 % glycine substitutions disrupted the $(EK)_{15}$ *β*-strand in the classical MD simulations (**Figure 3.3.4**). The CD spectra for $(EK)_{15}$ and $(EKGG)_7$ are provided in **Figure 3.3.6**.

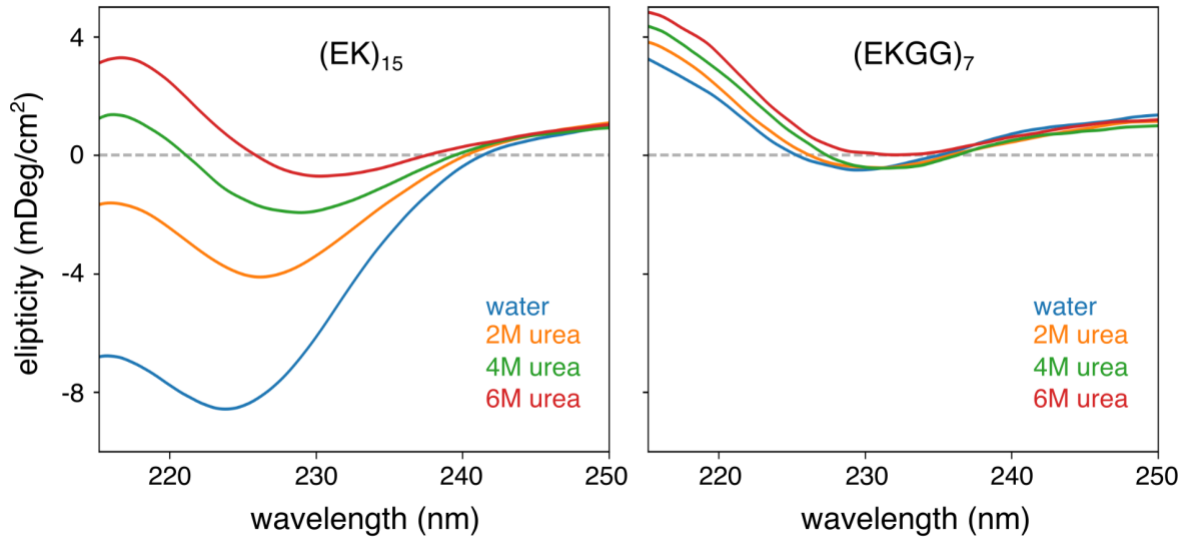

Figure 3.3.6 – CD spectra for $(EK)_{15}$ and $(EKGG)_7$ in neat water and solutions with various urea concentrations. (left) CD spectra for $(EK)_{15}$ shows the peptide is assumes some stable secondary structure in neat water, which is progressively disrupted as urea concentration is increased. (right) The spectra for $(EKGG)_7$ are relatively insensitive to urea, suggesting the peptide is completely disordered across the concentration range tested here.

Consistent with previous studies of self-complementary, EK-based peptides the *β*-strand character of $(EK)_{15}$ is apparent based on the negative band near at 225 nm. Typically, *β*-sheets exhibit a negative band near 218 nm.[122] However, $(EK)_{15}$ exhibits a red shifted negative band that is consistent with the CD spectra for *β*-turns.[123] As the denaturant urea is added to solution at increasing concentrations, the *β*-strand signature of the CD spectrum disappears. The spectrum for $(EKGG)_7$ in water, on the other hand, shows no sign of secondary structure other than random coil. The relatively small effect of urea on the CD spectrum supports this conclusion.

Conclusions

In the present work, we found the alternating-charge peptide $(EK)_{15}$ has a high propensity for self-interaction which manifests in the formation of persistent secondary structure motifs. This tendency may also be present for other purely alternating-charge sequences (such as DK investigated here), but the addition of glycine reliably disrupts secondary structure. We provide molecular level insights about the mechanism of disruption and suggest lead sequence candidates for future protective peptides.

An important consideration is that these simulations do not directly model polypeptide dynamics on the experimental scale. Protective polyEK has been reported for peptides from 10-30 kDa in weight, whereas the $(EK)_{15}$ peptide is on the order of 5 kDa. Atomistic simulation of peptides on the experimental scale are still computationally intractable, even with enhanced sampling methods. Advances in reliable coarse-grained force fields may bring this length scale into the realm of possibility in the coming years. For now, we are confident that our microscopic insight reflects true structural tendencies observable in the macroscopic world, given these aggregate microseconds of simulation time and the consistency of our G_{30} control with experimental observations for longer glycine-rich polypeptides.

While the glycine-diluted sequences studied here were intermediate to $(EK)_{15}$ and $(G)_{30}$ in terms of hydrophilicity and conformational disorder, molecular insights from our simulations may provide direction for more complicated protective peptide domains starting from the polyEK template. It may be possible to engineer diluted sequences with greater hydrophilic SASA using small polar amino acids, like serine and glutamine, in place of glycine. Hydrodynamic radius may be tuned by introducing proline, as demonstrated for PAS-based peptides. The current work shows how molecular modeling can provide useful molecular insights to support the *de novo* design of more complicated and performant protective polypeptides.

Chapter 4
Designing materials for protein bound uremic toxin capture

The removal of protein bound uremic toxins (PBUTs) from the bloodstream of chronic kidney disease (CKD) patients is an unmet challenge. PBUTs are uremic retention solutes with a significant fraction of the toxin mass in the bloodstream existing in complex with proteins – primarily human serum albumin (HSA) – which renders them unsusceptible to clearance by traditional dialysis.[124,125] Several of these toxins, including indoxyl sulfate (IS) and p-cresyl sulfate (pCS), have been correlated with poor clinical outcomes for CKD patients.[126–128] Fully functional kidneys are surprisingly effective at managing PBUT levels in the bloodstream, but the molecular mechanisms for natural PBUT clearance have not been characterized.[129] Without a detailed understanding of the protein unbinding and renal clearance process to guide design, strategies for artificial PBUT clearance in CKD have not produced comparable efficacy.

During the last year of my graduate research, I have attacked the problem of PBUT clearance with other researchers in the Center for Dialysis Innovation (CDI) at the University of Washington. Very little information has been previously reported about the interactions between PBUTs and proteins, so I used MD simulations to provide the first dynamic picture of the general and toxin-specific binding modes of IS, pCS, IAA, and HA (Section 4.1). With the goal of bridging the gap between molecular and experimental observables, I used advanced simulation and ML techniques to quantify the kinetics of transitions between metastable IS-HSA binding modes (Section 4.2) and to predict the dissociation rate of the IS-HSA complex (Section 4.3). The molecular level insights from these simulations are being considered in the design of materials for PBUT capture, and the calculated kinetics are being incorporated into ongoing device-scale modeling of a PBUT adsorption column.

4.1 Elucidating the molecular interactions between uremic toxins and the Sudlow II binding site of human serum albumin*

Introduction

Several strategies have been suggested for managing PBUT concentrations in CKD patients, aimed at displacing toxins from their protein binding site (rendering them dialyzable) and more efficiently capturing free toxin from solution. A few of the most extensively studied (and clinically relevant) PBUTs are known to bind to the major drug binding sites of HSA, called Sudlow site I and Sudlow site II.[130] Armed with this knowledge, Tao et al. recently demonstrated that introducing competitive binders for Sudlow site II (ibuprofen and tryptophan) to uremic plasma increased the unbound fraction of the uremic toxins IS, pCS, indole-3-acetic acid (IAA), and hippuric acid (HA).[131] Others have demonstrated that diluting uremic plasma with hypertonic solutions, thereby increasing the ionic strength, facilitated PBUT unbinding and increased the clearance of IS and pCS during *in vitro* and *ex vivo* hemodialysis.[132,133] The use of simple adsorbent chemistries targeting the hydrophobic and negatively charged moieties of many PBUTs has also shown promise in improving the efficiency of free PBUT capture.[134,135] The efficacy of these strategies could be greatly increased given information about the dominant stabilizing interactions between PBUTs and the protein residues in their native binding sites.

* Reproduced in part with permission from J. Smith and J. Pfaendtner. Elucidating the molecular interactions between uremic toxins and the Sudlow II binding site of human serum albumin. In preparation.

Protein data bank structures of CMPF and IS bound to Sudlow site I and Sudlow site II, respectively, are the only molecular scale data currently available to describe the interactions between PBUTs and HSA at the molecular scale.[136] While these experimental binding poses have provided valuable preliminary insights into the protein residues which stabilize CMPF and IS, they fail to capture the inherently dynamic nature of the protein-ligand complex. Also, insights derived from the experimental binding poses do not necessarily generalize to other PBUTs. Luckily, the 3-dimensional structures for CMPF and IS provide an entry point for more comprehensive characterization of PBUT-HSA interactions with molecular dynamics (MD) simulations. MD is a computational tool especially well-suited for investigating the inherently dynamic interactions of protein-ligand complexes with atomic resolution.

Sudlow site II is an attractive starting point for investigation with MD simulations because it has been established as the primary binding site of IS and pCS, the two PBUTs with the most extensive body of literature to support their deleterious physiological effects.[137] The chemical similarity between IS and other Sudlow site II binders, including pCS, IAA, and HA, can also be exploited to develop initial binding poses for MD simulations of other toxins. **Figure 4.1.1(a)** shows the chemical structures of the 4 toxins IS, pCS, IAA, and HA. **Figure 4.1.1(b)** shows a 2D PoseView representation of the only experimentally resolved crystal structure of a PBUT (IS) bound to Sudlow site II (PDB ID: 2BXH).[136]

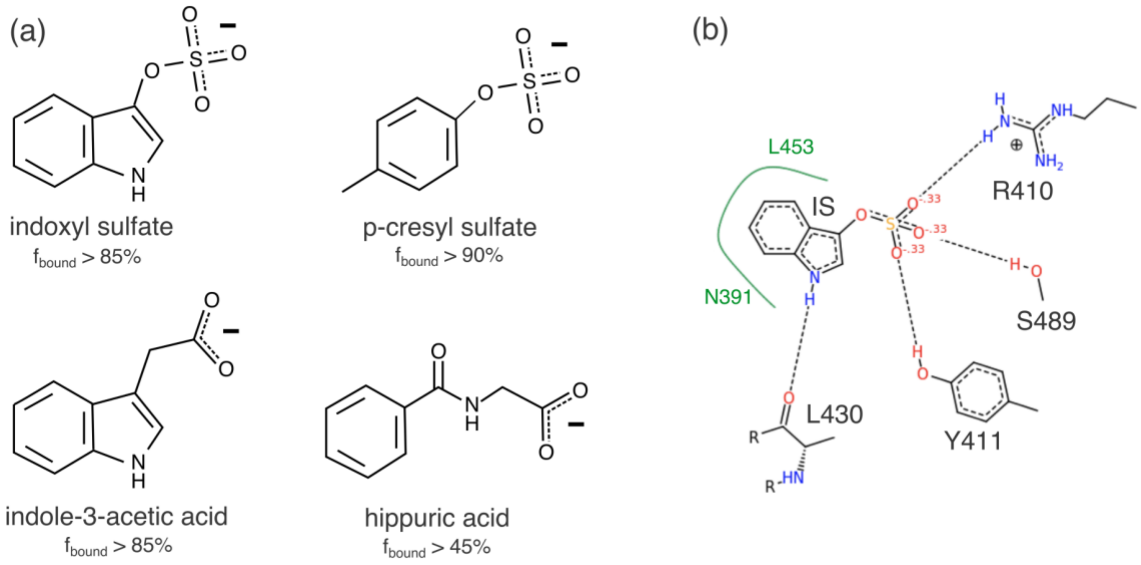

Figure 4.1.1 – Chemical similarity of PBUTs suggests similar interactions to IS in Sudlow site II. (a) Chemical structures of uremic retention solutes known to bind to the Sudlow site II of human serum albumin. (b) PoseView representation of the experimentally resolved binding structure for indoxyl sulfate in Sudlow site II (PDB: 2BXH). Putative hydrogen bonds are represented by dashed gray lines, and hydrophobic interactions by solid green lines.

In this article, we describe an extensive molecular dynamics (MD) simulation study of PBUTs in the Sudlow site II of HSA. We simulated the IS-HSA complex in water, starting from the experimental binding pose. We performed identical simulations with pCS, IAA, and HA in place of IS, starting with initial binding poses derived from the IS-HSA binding pose. The energetics of the protein-toxin interactions were used to identify the key toxin-stabilizing residues in Sudlow site II. Unique binding modes for IS, pCS, IAA, and HA were identified with unsupervised learning.

Our analysis provides insight into the general features of PBUT binding in Sudlow site II, as well as insights into toxin-specific interactions. All 4 toxin residues were found to be primarily

stabilized by electrostatic interactions with hydrophilic protein residues. The primary binding modes for IS and IAA were stabilized by a hydrogen bond with the carbonyl oxygen of L430 deep in the hydrophobic core of the binding pocket, while pCS and HA were stabilized by hydrogen bonding interactions near the mouth of the binding pocket. Excellent agreement was observed between the experimental binding pose of the IS-HSA complex and the primary binding mode identified with MD, providing external validation for our results. This work represents a pioneering contribution to the molecular level understanding of PBUT-HSA interactions and establishes a baseline for more advanced molecular modeling.

Methods

Preparing initial structures for HSA-toxin complexes

Of the uremic toxins known to bind Sudlow site II with high affinity, an experimentally resolved structure of the toxin in complex with HSA has only been reported for indoxyl sulfate (RCSB PDB ID: 2BXH).[136] The 2BXH PDB structure contains two HSA molecules, each associated with 3 IS molecules associated and crystallographic waters. We took the coordinates of the "chain A" HSA molecule (atoms 1-4263) and the IS molecule bound to its Sudlow site II (atoms 8531-8544). The other 2 IS molecules associated with chain A, overlapping reconstructions of alternative binding poses of IS to Sudlow site I, were ignored based on experimental evidence that suggests there is negligible IS binding to Sudlow site I under physiologically relevant conditions.[138] The tleap program from Amber tools was used to add hydrogens and missing atoms to the HSA molecule.[92] We built an IS molecule (complete with hydrogens) in GaussView and superimposed this upon the IS heavy atoms from the PDB using the "RMSD Calculator" extension in VMD.[38,106] Gromacs tools were used to construct a cubic box (with sides 10.8 nm in length) around the IS-HSA complex and to add TIP3P water and sodium counter ions.[27,28] The Amber 14 force field parameters were used for the HSA protein.[107] We calculated the partial charges for IS atoms using the restrained electrostatic potential (RESP) method and assigned bonded parameters according to the Generalized Amber Forcefield (GAFF).[24,47] Quantum mechanical calculations for RESP were performed with Gaussian 09 using density functional theory with the B3LYP/6-31G* level of theory.[38]

Motivated by the chemical similarity between IS and the other uremic that bind to Sudlow site II, we used the IS-HSA PDB structure as a template for putative binding poses of p-cresyl sulfate (pCS), indole-3-acetic-acid (IAA), and hippurate. We built each toxin molecule in GaussView and superimposed the structure onto the IS-HSA structure in VMD so that the anionic and hydrophobic moieties of the new toxin aligned with the sulfate and indole groups of the IS molecule, respectively. The initial superimposed structures are pictured in **Figure II.11**. Solvent and counter ions were added to each of these structures following the method outlined above for IS. An identical protocol was also used for assigning forcefield parameters for each toxin.

Molecular dynamics simulations of the toxin-HSA complexes

All molecular dynamics (MD) simulations were performed using Gromacs 2016.3.[28] A preliminary 3-step equilibration protocol was used to relax the initial solvated structure of each protein-toxin complex, which included energy minimization, annealing, and a short simulation in the NPT ensemble. Energy minimization was performed on the initial solvated structures following the steepest descent algorithm for 10000 steps. All other simulations were propagated using an integration timestep of 2 fs. A 250 ps annealing simulation was performed to gently relax the system from the energy minimized conformation to a realistic configuration at the target temperature of 298 K. This was achieved by ramping the coupling temperature for the Bussi-Donadio-Parrinello thermostat smoothly from 5 to 298 K over the first 240 ps of the simulation.[29] The output configuration and

velocities from the annealing simulation were used to initiate a 500 ps NPT simulation with the Berendsen barostat for rapid coupling to the target pressure of 1 bar.[30]

The strong binding affinity between HSA and the 4 PBUTs suggests that unbinding events are not likely to occur on the timescale currently accessible with unbiased molecular dynamics simulations. We added an additional equilibrium simulation to ensure that each protein-toxin complex (especially the pCS, IAA, and hippurate complexes) reached a metastable binding pose that would not lead to rapid unbinding in production simulations. We performed a 5 ns NPT simulation with the Bussi-Donadio-Parrinello thermostat and Parrinello-Rahman barostat.[29,31] The system temperature was coupled to 298 K with a frequency of 10 ps^{-1}, and the pressure was coupled to 1.0 bar with a frequency of 1 ps^{-1}. The potential energy of the system was extended with an upper wall potential, implemented with Plumed 2.4,[49] according to the following

$$V_{wall}(x) = \begin{cases} k(x-a)^2 & x > a \\ 0 & x \leq a \end{cases}$$

where x is equal to the root mean square deviation (RMSD) of carbon atoms of the toxin and neighboring binding pocket residues, a is the RMSD value at which the restraining potential is turned on (in this case 0.1 nm), and k is the restraining force constant (in this case 150 kJ/mol-nm). This restraint was intended to prevent rapid unbinding as might occur with a poor initial guess at the binding pose, which could result in steric clashes or atomic overlap.

Three 250 ns NPT simulations were performed for each protein-toxin complex (12 production simulations in total). Each production simulation was initiated with the output configuration from the restrained binding-pocket simulation and different randomly generated velocities. The Bussi-Donadio-Parrinello thermostat and Parrinello-Rahman barostat were used to maintain a temperature and pressure of 298 K and 1 bar, respectively.[29,31] The first 50 ns of the simulation were allowed for any further relaxation of the protein structure, and the final 200 ns of each simulation were used for all following analysis.

Energy and structural descriptor calculations

Protein-toxin interaction energy was calculated using the GROMACS tool `gmx energy`. The initial list of 37 residues to be screened for interactions with each toxin was generated by taking all protein residues with at least one atom within 6 Å of any atom of indoxyl sulfate in the experimentally resolved PDB structure (PDB: 2BXH). The number of atomic contacts and hydrophilic contacts between each of these residues and each toxin were calculated, using Plumed 2.4, using the following switching function

$$N_{contacts}(\boldsymbol{r}) = \sum_{i \in A} \sum_{j \in B} \frac{1 - \left(\frac{r_{ij} - d}{r}\right)^6}{1 - \left(\frac{r_{ij} - d}{r}\right)^{12}}$$

where A and B represent the sets of heavy atoms (nonhydrogen) atoms of the toxin and a given protein residue, r_{ij} is the intermolecular distance between atoms i and j, and d and r are switching function parameters that control the diameter and radius of a coordination shell around each atom. Atomic contacts were calculated every 10 fs (5 MD steps) of each production simulation, setting a 4 Å coordination sphere around each atom (r = 0.4 nm, d = 0). Hydrophilic contacts were calculated with a coordination annulus around each atom, 2.7 Å in diameter and 0.3 Å thick (r = 0.03 nm, d =

0.27 nm). This definition of hydrophilic contacts is consistent with the donor-acceptor distance expected for strong hydrogen bonds.

For the 9 residues selected for further analysis (selected based on atomic and hydrophilic contacts), the Lennard-Jones and Coulomb components of interaction energy with the toxin were calculated every 2 ps of each production simulation, again using `gmx energy`. The center of mass distance between each of the 9 key residues and the toxin (and the minimum distance between any R410 heavy atom and the toxin heavy atoms) were calculated every 2 ps of each simulation using Plumed 2.4. Subsequent analysis and visualization of these energetic and structural descriptors were performed in the Python programming language. Data files and analysis scripts have been made publicly available at www.github.com/UWPRG/uremic_toxins.

Dimensionality reduction and clustering with principal component analysis (PCA) and mean shift

With the large number of MD frames (300,000 for each complex) and the relatively high dimensionality of the structural descriptors for each frame (9 center of mass distances), an automated solution for pattern recognition was required to identify different binding states. We used the mean shift algorithm for clustering. Mean shift is a mode-seeking algorithm that essentially identifies peaks (modes) in a probability density function and clusters points that climb to the same peak following gradient ascent.[139] This results in a more physically meaningful clustering than other commonly used clustering algorithms such as k-means (see **Figure II.12** for further explanation). The natural fit of mean shift for processing data from biophysical simulations has been previously noted.[14,46]

The results of mean shift clustering are dependent on an initial estimate of the underlying probability density. Nonparametric techniques for density estimation suffer from the curse of dimensionality and perform better on dense, low dimensional data. A principal component analysis (PCA) was performed to project the 9 toxin-residue center of mass distances for each frame of the MD simulations onto a 2 dimensional structure space. PCA is a dimensionality reduction technique that is commonly used to project the inherently high-dimensional data from molecular dynamics simulations into an informative low-dimensional space. The coefficients for each principal component are provided in **Table II.9**. We found that considering 3 or more principal components did not qualitatively change the clustering results. Mean-shift clustering was performed on the PCA-transformed data to identify unique metastable states in this dimensionally-reduced structure space for each protein-toxin complex. The average silhouette score – a quantitative measure to assess the quality of clustering – was used to select the optimal bandwidth for density estimation in the 2D principal component space (**Table II.10**).[140]

Results and Discussion

We performed MD simulations of each PBUT-HSA complex in aqueous solution to investigate the key stabilizing interactions and structural features of the bound state. For each toxin we performed 3 classical MD simulations, each 250 ns long. The first 50 ns of each simulation was used to allow further relaxation of the toxin-protein complex. Frames were saved for analysis every 2 ps (1000 MD steps) over the final 200 ns of each replica. Unless otherwise stated, the average values for energetic and structural observables reported in this section were calculated from the aggregated frames of all 3 simulations and the standard deviations were calculated from the mean values for each replica (standard deviation over 3 means).

Uremic toxins are primarily stabilized by electrostatic interactions in Sudlow site II

Each of the PBUTs considered in this work is comprised of a bulky hydrophobic moiety and a hydrophilic, anionic moiety. As a first step in characterizing the nature of PBUT binding to Sudlow site II of HSA, we investigated the relative contributions of hydrophobic and electrostatic interactions

to the overall protein-toxin interaction energy in the bound state. We calculated the interaction energy between the toxin and the protein in each frame of the simulation and tracked the contributions due to the Lennard-Jones, E_{LJ}, and Coulomb, E_C, potentials over time. The average E_{LJ} and E_C, as well as the total interaction energy, for each complex are tabulated in **Table 4.1.1**.

Table 4.1.1 – Average interaction energy between HSA and uremic toxins bound to Sudlow site II.

	E_C (± sem) [kJ/mol]	E_{LJ} (± sem) [kJ/mol]	E_{total} (± sem) [kJ/mol]
Indoxyl sulfate	- 114 (6.12)	- 103 (1.68)	- 217 (7.38)
p-cresyl sulfate	- 136 (8.92)	- 87.2 (0.138)	- 223 (9.08)
Indole-3-acetic-acid	- 154 (8.37)	- 90.0 (0.403)	- 244 (8.21)
hippurate	- 162 (27.3)	- 84.9 (3.75)	- 247 (24.5)

For every toxin, the average magnitude of E_C is greater than that of E_{LJ}. Not surprisingly, IS (the bulkiest of the 4 toxins) has the highest average Lennard-Jones interaction energy, followed by IAA. The bulky indole group (shared also by tryptophan, an endogenous HSA ligand and metabolic precursor to IS and IAA) apparently fits more tightly in the hydrophobic pocket of Sudlow site II than the smaller phenyl groups of pCS and HA.

Although IS and pCS have the same anionic moiety (sulfate) and IS has an additional hydrogen bond donor in its indole ring, pCS was found to have a lower average E_C than IS. We also found that, on average, pCS actually forms more hydrogen bonds with HSA residues in Sudlow site II than does IS (**Table II.11**). This could be in part because the electron withdrawing N of the IS indole ring decreases the polarity of the sulfate linker oxygen atom of IS relative to the linker oxygen of pCS. The carboxylate-containing toxins, IAA and HA, have stronger electrostatic interactions with the protein binding site than the sulfate-containing toxins, IS and pCS. This is likely due to the increased charge density of the carboxylate anion of IAA and HA compared to the sulfate anion of IS and pCS (see **Figure II.13**). The geometry of the carboxylate anion also facilitates a very low-energy double hydrogen bonding conformation with the guanidinium cation of the R410 residue. A representative structure for the HA-R410 double hydrogen bond conformation is provided in **Figure II.14**.

The observation that average E_C is greater than average E_{LJ} in each case suggests that electrostatic interactions play an important role in stabilizing the protein-toxin complex. This is in good agreement with an experimental study that demonstrated the binding affinity between IS and HSA in various NaCl solutions decreased with increased ionic strength.[138] Hypertonic predilution has been demonstrated to increase HD removal of only IS and pCS (as well as the uremic toxin phenyl acetic acid, which does not bind Sudlow site I or site II). The observation that all 4 PBUTs simulated in the present work are primarily stabilized by electrostatic interactions suggests that increasing ionic strength may be a general strategy for favoring PBUT dissociation from Sudlow site II.

The shape of the probability density function (pdf) for each component of the protein-toxin interaction energy provides additional information beyond the mean. **Figure 4.1.2** shows the probability distribution of E_{LJ} and E_C from all 3 simulations for each PBUT complex. An additional explanatory figure for **Figure 4.1.2**, showing the interaction energy vs. time for each MD simulation is provided in **Figure II.15**.

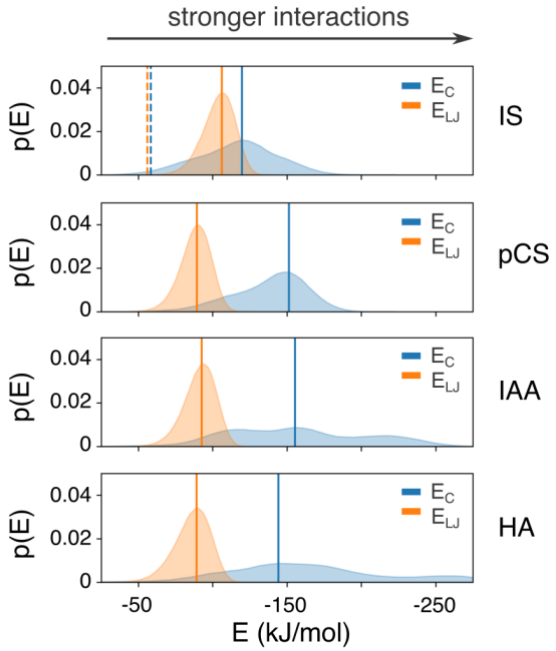

Figure 4.1.2 – Probability distribution functions for interaction energies between toxins and protein residues calculated from MD simulations. The blue and orange shaded regions represent the pdf for E_C and E_{LJ}, respectively, based on a histogram of every frame of the production MD simulations for each complex (3 x 100,000 observations). The peak of each pdf is marked by a solid, vertical line. The dashed, vertical lines in the top panel represent the E_C and E_{LJ} energies calculated from the experimental IS-HSA structure (PDB: 2BXH).

One would expect the pdf to have a single well-defined peak if fluctuations around a single low energy binding pose could account for the structural ensemble sampled with MD. Indeed, the pdf for the Lennard-Jones energy could be accounted for by a single peak. However, the broad irregular shape of the pdf for Coulomb energy for each complex suggests that each PBUT visits a number of metastable binding poses during the MD simulations (**Figure 4.1.2**). In the following results and discussion, we identify the key protein residues for stabilizing PBUTs in Sudlow site II and use structural descriptors derived from these residues to describe the metastable binding modes for each toxin.

Protein-toxin interactions are dominated by the same residues for all PBUTs tested

For further insight into the roles of specific Sudlow site II residues in stabilizing each toxin, and to identify possible structural distinctions between unique binding poses, we analyzed the interactions between toxins and binding pocket residues in more detail. We started with a list of 37 residues close to IS in the experimental X-ray structure, defining close residues as those with at least one atom (including hydrogens) within 6 Å of any atom of IS. To screen for residues of interest, we calculated the number of atomic and hydrophilic contacts between the toxin and each of these 37 residues in each frame of the MD simulations. The average number of atomic and hydrophilic contacts for each protein residue and each toxin are included in **Figure II.16**.

Taking as the pertinent residues the set that includes all residues ranking in the top 5 residues in terms of either atomic or hydrophilic contacts for each toxin, we developed a list of 9 key residues for further analysis. A 3D structure highlighting the original 37 residues and filtered list of 9 residues is provided in the supporting information (Figure S7). This list includes all residues identified as hydrogen bond partners with IS in the experimental crystal structure (R410, Y411, L430, S489).[136] The other key residues identified in this filtering process (L387, N391, K414, V433, L453) may help to

distinguish between metastable binding modes for IS, or be involved in stabilizing the primary binding modes of the other toxins. **Figure 4.1.2** shows the average Coulomb and Lennard-Jones energy contributions of each key residue.

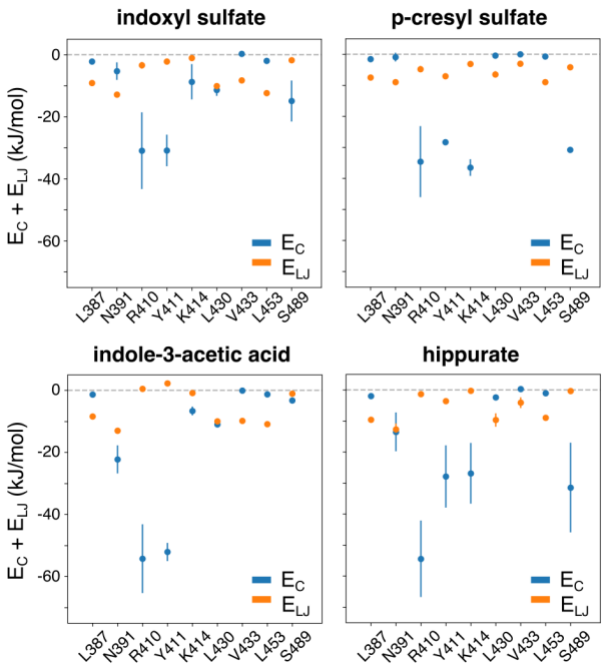

Figure 4.1.2 – The interaction energies contributed by each of the 9 key binding pocket residues to the (left) IS-HSA complex and (right) pCS-HSA complex. The average Coulomb and Lennard-Jones components of the residue-toxin interaction energy (for all 3 simulations) are represented by blue and orange circles, respectively. Error bars represent +/- 1 standard deviation from the mean.

The two residues that interact most strongly with IS in the MD simulations are R410 and Y411. This observation is in good agreement with an experimental study that showed R410A and Y411A mutations together decreased the binding capacity of HSA for IS.[141] R410 and Y411 are also the primary energetic contributors to the IAA complex. pCS and HA interact strongly with R410 and Y411, but also with K410 and S489, two other potential hydrogen bonding partners. These observations are particularly interesting because it might have been expected that the anion of the toxin would dictate its interactions with hydrophilic residues in the binding pocket.

Both IS and IAA participate in electrostatic interactions with the hydrophobic L430 residue. This interaction can be explained by a hydrogen bond between the indole NH group of each toxin and the carbonyl (backbone) oxygen of L430, pictured in the PoseView representation of the IS-HSA complex in **Figure 4.1.1(b)**. This interaction has previously been described as a polar feature inside the hydrophobic core of the binding pocket.[136] Also of note, N391 contributes to the interaction energy with IS and pCS primarily through E_{LJ}, but contributes also to the electrostatic interactions with IAA and HA.

Transient salt bridge with R410 is a primary contributor to low energy poses

The energetic contributions of R410 also have the largest standard deviation of the 9 residues. Large changes in the interaction energy between the toxin and one or a few residues may signify transitions between metastable states. To monitor the relationship between binding pocket conformation and interaction energy, we calculated the center of mass distance to each of the 9 key residues identified above and calculated their correlation with overall protein-toxin interaction energy.

Table 4.1.2 shows the Pearson correlation coefficient between the overall toxin-protein interaction energy and the center of mass distance between IS and each of the 9 residues.

Table 4.1.2 – Correlation between proposed order parameters and protein-toxin interaction energy.

indoxyl sulfate		**p-cresyl sulfate**		**IAA**		**hippurate**	
Residue (ranked)	r	Residue (ranked)	r	Residue (ranked)	r	Residue (ranked)	r
R410	0.57	R410	0.60	R410	0.41	R410	0.63
Y411	0.44	Y411	0.52	Y411	0.30	Y411	0.53
S489	0.43	S489	0.46	L387	0.19	K414	0.45
K414	0.34	K414	0.45	N391	0.10	S489	0.41
L387	0.05	L387	- 0.13	K414	0.07	L453	0.00
L453	- 0.17	N391	- 0.26	V433	0.01	L387	- 0.28
L430	- 0.19	L453	- 0.28	L453	0.00	V433	- 0.34
N391	- 0.29	L430	- 0.42	S489	- 0.07	L430	- 0.39
V433	- 0.30	V433	- 0.44	L430	- 0.15	N391	- 0.56

Interestingly, there is a strong correlation between the protein-toxin interaction energy and the center of mass distance between IS and several of the residues (**Table 4.1.2**). There is a positive correlation between interaction energy and center of mass distance to each of the hydrophilic residues at the mouth of Sudlow site II, and a relatively strong negative correlation between interaction energy and the distance to hydrophobic residues buried in the binding pocket. This suggests an enthalpy gain for the toxins upon shifting away from the hydrophobic core and toward the opening of the binding pocket.

For each toxin, the R410-toxin distance has the strongest correlation with the interaction energy of the complex. Closer inspection of R410-toxin interactions showed that this important stabilizing residue frequently breaks contact with the toxin in favor of a fully solvated state. **Figure 4.1.4** shows that the breaking and forming of a salt bridge between IS and R410 is an important process in transitioning between high and low energy states.

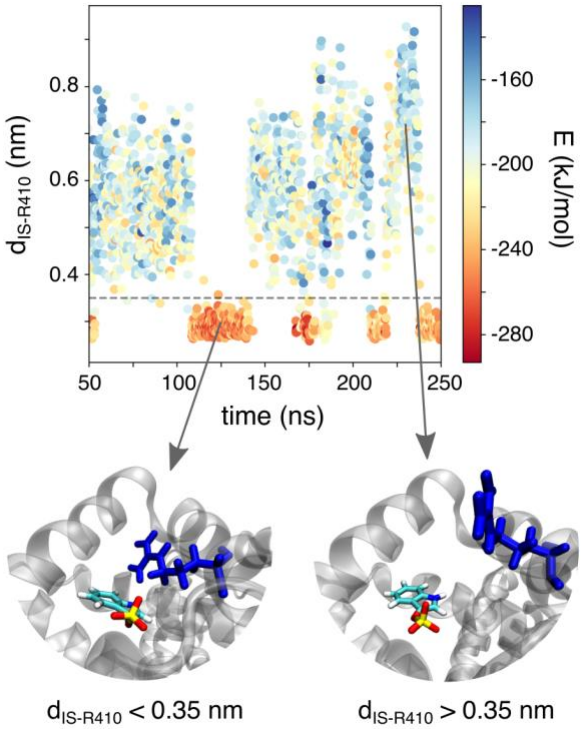

Figure 4.1.4 – The minimum distance between IS and R410 heavy atoms vs. time from IS simulation 3, and it's relation to the overall protein-toxin interaction energy. Each point is colored based on the overall protein-toxin interaction energy, ranging from dark red (lower energy) to dark blue (higher energy). The dashed horizontal line at 0.35 nm represents a putative cutoff for IS-R410 hydrogen bonding. The images below the plot show representative conformations for (left) low energy, low minimum distance and (right) high energy, high minimum distance.

Figure 4.1.4 shows clearly that the IS-HSA complex is stabilized by salt bridging with the charge-neutralizing arginine in Sudlow site II. The weakened protein-toxin interactions associated with the breaking of this salt bridge, as well as the alleviation of potential steric hindrance to unbinding – observable in the representative conformations in **Figure 4.1.4** – signal the potential importance of this molecular movement in the unbinding process. The breaking/forming of this charge pairing was observed multiple times for all toxins (**Figure II.18**). Supporting the previous hypothesis that the charge-dense carboxylate groups provide especially strong interactions with R410, we find that the mean distance of R410 hydrogen bonds with IAA and HA is lower than those with IS and pCS (**Figure II.18**).

Shifting the conformational equilibrium for R410 toward the solvated state could be a potential target for strategies to encourage toxin unbinding. The equilibrium could be shifted by adding cosolutes that lower the free energy of solvation for the R410 guanidinium side chain. R410 solvation could potentially be involved in the mechanism for IS release in hypertonic solutions. One could also target allosteric regulation of the Sudlow site II that specifically favors HSA conformations with R410 in the solvated state. Further investigation, both experimental and computational, into the effect on R410 position of ions in solution and remote binding to HSA could provide more actionable suggestions for toxin displacement strategies.

3.4 The toxin hydrophobic moiety determines the primary binding mode in Sudlow site II

The relatively fast breaking and forming of the R410 salt bridge can happen in isolation without significant movement of the bound toxin relative to other residues in the binding site. We

performed a principal component analysis (PCA) on the 9 center of mass distances to project onto a low-dimensional conformation space where identifying unique states with unsupervised learning was more tractable. We used the mean shift algorithm to assign each frame of the MD simulation to a binding mode, based on the first two principal component values for the frame. A representative structure for each mean shift mode was taken as the MD frame with principal components nearest (based on 2D Euclidian distance) the center of that mode. **Figure 4.1.5(a-b)** shows the 2D probability density function along the first two principal components for the IS-HSA complex and the results of clustering points on this pdf with the mean shift algorithm. **Figure 4.1.5(c-f)** contains a summary of the structural features and representative structures of the IS binding modes identified with mean shift. Similar figures for the other 3 toxins are provides in **Figure II.19-II.11**.

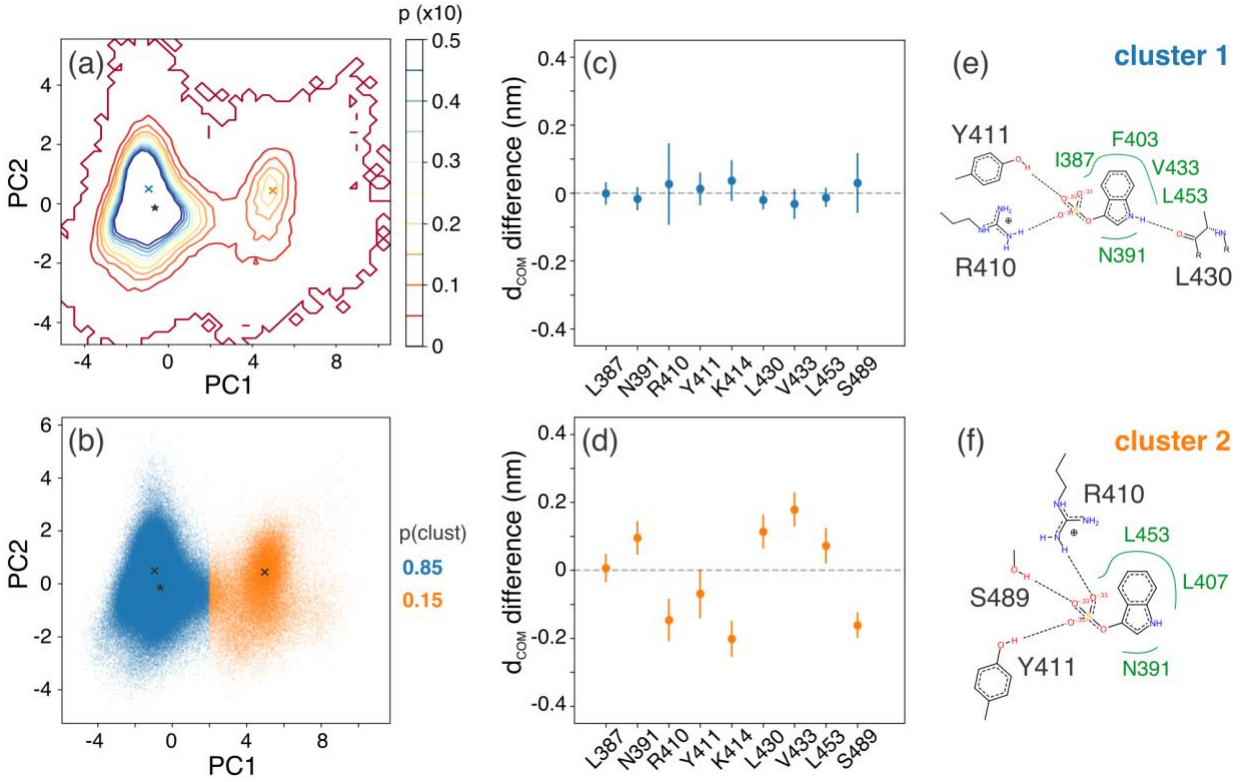

Figure 4.1.5 – PCA and clustering results for the indoxyl-sulfate-HSA complex. (a) Contour plot of the 2D pdf for the IS-HSA complex, created from a 2D histogram of the PC values for each frame of the MD simulations, with a probability step of 0.005 between contour lines. Blue and orange 'x' symbols mark the center of each mode identified with mean-shift. The star symbol represents the experimental structure for IS-HSA (PBD: 2BXH). **(b)** Scatter plot of the first 2 PC values for each MD frame, colored by the cluster assigned by mean-shift. **(c-d)** The difference between the average IS-residue center of mass distance for all points in each cluster and the overall average IS-residue center of mass distance (all points in all clusters). **(e-f)** A PoseView representation of the central structure of each cluster (the MD frame with PC values nearest the 'x' symbols in **(a-b)**.

The first two principal components of the IS-HSA structural descriptors accounted for 57% and 15%, respectively, of the variance in the data set (PC coefficients available in **Table II.9**). We identified 2 binding modes for IS in Sudlow site II, which accounted for fractions of 0.85 and 0.15, respectively, of the conformations sampled with MD. Each binding mode from MD has partial overlap with the experimental binding pose (**Figure 4.1.1(b)**). The primary binding mode for IS from MD and the experimental binding pose both have hydrogen bonding interactions with L430, R410, and Y411 and hydrophobic interactions with N391 and L453. Relative to the experimental structure, the primary MD structure is shifted deeper into the hydrophobic core of the binding pocket,

interacting with hydrophobic residues I387, F403, and V433, and preventing a hydrogen bond with S489. In the secondary binding mode of IS, losing the L430 hydrogen bond has apparently allowed IS to shift nearer to the hydrophilic mouth of the binding pocket. In the secondary binding mode, IS regains the hydrogen bond with S489, but loses hydrophobic interactions with I387, F403, and V433. Further segmentation of these clusters, using k-means clustering ($k > 2$), captures the movement of individual protein residues rather than significant changes in IS position (**Figure II.22**).

The results of PCA and clustering for the other three PBUT complexes shed interesting light on the residue-wise energetic analysis described above. **Figure 4.1.6** shows a PoseView representation of the central structure of the most populated cluster for each complex.

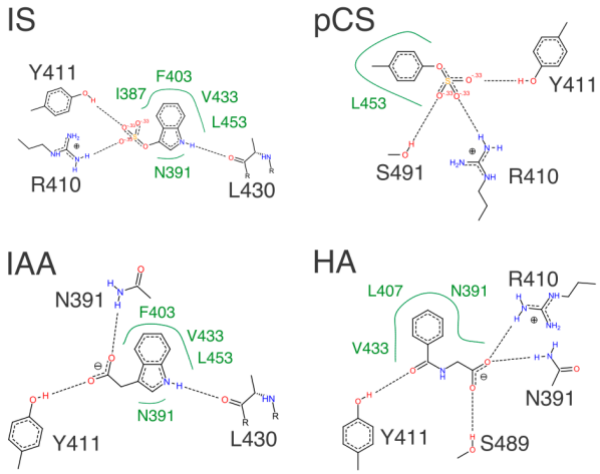

Figure 4.1.6 – PoseView representation of the most populated binding mode for each PBUT-HSA complex. Putative hydrogen bonds are represented by black dashed lines, accompanied by the name and sidechain structure for the associated protein residues. Hydrophobic interactions are represented by solid green curves, labeled with the names of the associated protein residues. Polar atoms O, N, and S, are colored red, blue, and yellow, respectively.

The primary binding mode for IAA is also stabilized by a hydrogen bond with the L430 carbonyl oxygen, accounting for the electrostatic interactions with L430 identified above. The bulky indole groups of IS and IAA are apparently pinned deep in the hydrophobic core by this hydrogen bond and extensive hydrophobic interactions with other residues. Notably, only one binding mode was identified for IAA. The representative binding pose extracted from the sole IAA cluster captures a pose without hydrogen bonds donated from R410. The R410 position is observed to fluctuate in and out of hydrogen bonding distance with one, or both, of the carboxylate oxygens of IAA. The stability of the IAA binding mode suggests that hydrogen bonds with L430, Y411, and N391, can effectively hold IAA in place when the transient salt bridge with R410 is broken.

pCS and HA assume primary binding modes similar to the secondary binding mode of IS. Lacking an appropriately positioned hydrogen bond donor to exploit the L430 "polar" moiety, pCS and HA are not pinned in the hydrophobic pocket like IS and IAA. pCS and HA are allowed to shift closer to the mouth of the binding pocket where they interact with K414 and S489, as shown in **Figure 4.1.2**. The carboxylate-containing toxins, IAA and HA, both accept hydrogen bonds from the NH_2 group of N391 in their primary binding pose. This interaction is also observed in an alternate binding mode of pCS (in **Figure II.19**).

Conclusions

The preference of each toxin for interactions with different chemical functional groups in Sudlow site II could be useful in designing toxin capture materials with high affinity and specificity

for each toxin. Generally, a positively charged moiety, such as the guanidinium moiety of R410, and hydrogen bond donors, such as the hydroxyl of Y411 and S489, could be incorporated into synthetic constructs to nonspecifically stabilize all 4 toxins. Incorporating hydrophobic constituents into a synthetic construct for PBUTs may also be generally effective. Embedding a hydrogen bond acceptor, such as the L430 carbonyl, in the hydrophobic region may increase specificity for IS and IAA. The observation that a polar moiety embedded in a hydrophobic pocket stabilizes the primary IS binding mode may serve as a preliminary mechanistic explanation for the recent finding that cyclodextrin can bind IS from solution.[142]

Future work should target a direct connection between molecular modeling and the macroscopic observables that impact CKD patient health. Advanced simulation techniques for quantifying ligand binding affinity and unbinding kinetics represent promising avenues to make this connection.[5,6] The current work provides the key ingredients necessary to use these advanced techniques, including representative structures for metastable binding poses and insight into discriminative structural features. Hopefully this contribution marks the beginning of a more precise and informed materials design process that will ultimately lead to better outcomes for CKD patients. In Sections 4.2 and Section 4.3 we provide preliminary estimates of the kinetics for transitions between metastable binding modes and for unbinding, respectively.

4.2 Molecular modeling of the kinetics of the indoxyl sulfate-HSA complex

Introduction

Recent advances in molecular dynamics (MD) simulation techniques have made it possible to accurately determine the kinetics of long-timescale biomolecular processes, such as protein folding and protein-ligand unbinding.[5,143,144] The infrequent metadynamics enhanced sampling technique has proven especially useful in the case of calculating protein-ligand dissociation rates.[145,146] However, collective variables that account for the slow degrees of freedom in the unbinding process must be identified for a given protein-ligand complex before infrequent metadynamics can be applied effectively.

A popular approach for identifying the slowest processes in a biophysical system is to build a Markov state model (MSM) to describe the exchange rates between conformational microstates. It has been demonstrated that constructing MSMs from a collection of relatively short MD simulations can be used to describe the behavior of biomolecular systems on much longer timescales.[7,8,147] Pérez-Hernández et al. showed that an MSM could be used to recreate the conformational dynamics of an intrinsically disordered peptide that had been previously measured by NMR.[7] They also showed that the MSM and the MD simulation data it was built from could be used to identify the molecular order parameters which control the conformational relaxation timescales.[7] Such an approach could be very useful in predicting which ligand-residue interactions control transitions between metastable binding modes for the IS-HSA complex, and which binding state transitions set the timescale for unbinding.

In the current work, we build upon our previous investigation of PBUT-HSA complex structures to quantify the kinetics of the IS-HSA complex. We start by identifying the kinetically separated binding modes of the complex by performing a tICA analysis on 15 classical MD trajectories. We construct a MSM to describe the transition probabilities and rates between binding modes. The eigenvectors of the MSM transition matrix, corresponding to the slowest degrees of freedom in the IS-HSA complex, are correlated with molecular order parameters to identify those descriptors that best encode binding mode transitions. The results from this work are incorporated into the infrequent metadynamics simulations in Section 4.3 to calculate IS-HSA unbinding rates.

Methods

Classical molecular dynamics simulations

We performed 15 classical MD simulations of the IS-HSA complex in aqueous solution, starting from the same equilibrated output conformation of the binding pocket restrained MD simulation described in our previous work (Section 4.1). Our simulation protocol was also identical to the protocol described for the production simulations in Section 4.1, with different random velocity seeds for each simulation. Briefly, each simulation was 250 ns long, with the final 200 ns taken for analysis. MD timesteps were 2 fs, with conformations saved for analysis every 1000 steps (2 ps). The system temperature was coupled to 300 K with the Bussi-Donadio-Parrinello stochastic thermostat, with a frequency of 10 ps^{-1}.[29] The system pressure was coupled to 1.0 bar with the Parrinello-Rahman barostat, with a frequency of 1 ps^{-1} and anisotropic compressibility $4.5e^{-5}$ bar^{-1}.[31] For each frame of each simulation, we calculated the 9 toxin-residue intermolecular distances described in Section 4.1 and 3 protein-protein order parameters, for a total of 12 structural descriptors. The protein-protein order parameters were the pairwise center of mass distances between the three helices that comprise the mouth of the Sudlow site II. These helical distances are pictured in **Figure II.23**. All simulations were performed with the GROMACS simulation engine and molecular order parameters (intermolecular distances) were calculated using the Plumed plugin, as described in Section 4.1.[49]

Dimensionality reduction, clustering, and construction of Markov state models

We took several steps to construct and cross-validate our MSM for the IS-HSA complex. First, we used a time-structure independent component analysis (tICA) to project the 12 structural descriptors for each frame onto a 2-dimensional space. This tICA transformation is very similar to the PCA dimensionality reduction described in Section 4.1, except that it aims to identify linear projections that maximize kinetic separation between conformations, rather than structural separation. We used kmeans clustering with large k values (>50) to discretize the tICA space into microstates, so each trajectory could be described based on a single number – the microstate ID – for each frame. We used a 5-fold cross validation procedure to build a reliable MSM upon the discretized trajectories, as recommended by McGibbon and Pande, splitting 3 trajectories into each fold.[148] We fixed the lag time at 1 ns and number of timescales at 4 for our MSM, because these properties cannot be optimized using the generalized matrix Rayleigh quotient (GMRQ), which we used as the scoring metric for MSMs. We performed a grid search over tICA lag times (0.002-0.2 ns), number of tICA components (2-5), and the number of clusters in our k-means discretization (50-200). For the final MSM, we used a second order tICA with a lag time of 0.2 ns and 200 k-means clusters. We used the PCCA+ spectral clustering algorithm to create a coarse-grained MSM for better interpretability.

All postprocessing and analysis of the output from our classical MD simulations was performed using in-house analysis scripts written in the Python 3 programming language. tICA dimensionality reduction, kmeans clustering, and MSM construction were performed with the MSMBuilder library.[149] The grid search and 5-fold cross validation were performed with the scikit-learn library. PCCA+ was performed with the msmtools library. The analysis scripts have been made publicly available at www.github.com/UWPRG/uremic_toxins.

Results and Discussion

Constructing a Markov state model provides an improved estimate of the IS-HSA relaxtion timescales compared to tICA

The accuracy of an MSM constructed from MD simulation data is sensitive to the selection of initial order parameters, and various hyperparameters for dimensionality reduction and clustering prior to the MSM training phase. We followed a 5-fold cross validation protocol to assess the quality of MSMs trained with different hyperparameters. Having established in our previous work the most

important order parameters for discriminating between different IS-HSA conformations, we focused on tuning the tICA lag time, the number of tICA components, and the number of clusters used in our kmeans clustering. We selected 2 ICs, 0.2 ns lag time, and 200 clusters based on 5-fold CV search for maximum GMRQ.

Another common way to quantify the accuracy of MSMs with different hyperparameters is to compare the estimated timescale of their slowest relaxation mode. Because tICA and MSM strictly underestimate the slowest process of a given system, the model with that places the highest lower bound on the first timescale is considered the most accurate. **Figure 4.2.1** illustrates the improvements in the estimate of the slowest process afforded by cross-validation and using and MSM compared to tICA alone.

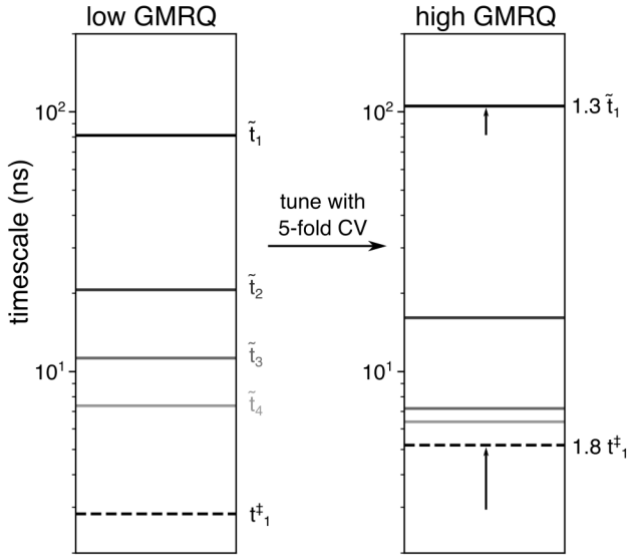

Figure 4.2.1 – Implied timescales calculated from the eigenvalues of the transition matrices from tICA and MSM, using different training hyperparameters. (left) Solid lines represent the implied timescales for the 4 slowest relaxation modes in the IS-HSA simulations, with lighter lines corresponding to faster modes, as estimated by the MSM with the lowest GMRQ score from 5-fold CV. The dashed line was the implied timescale for the slowest relaxation mode, estimated by the associated tICA projection ($d = 5$, $\tau = 0.1$ ns). (right) Solid lines represent the implied timescales for the 4 slowest relaxation modes in the IS-HSA simulations, with lighter lines corresponding to faster modes, as estimated by the MSM with the highest GMRQ score from 5-fold CV. The dashed line was the implied timescale for the slowest relaxation mode, estimated by the associated tICA projection ($d = 2$, $\tau = 0.2$ ns). The MSM with the highest GMRQ from cross-validation yields the best estimate of the slowest relaxation mode, establishing a lower bound on t_1 that is 1.3 times the t_1 estimate from the other MSM and about 20 times the t_1 estimate from tICA.

Figure 4.2.1 shows that the estimated timescale for the first relaxation, a lower bound for the slowest process of observed in the simulations, was 1.3 times higher for the MSM with the highest GMRQ ($t_1 \sim 105$ ns) compared to the MSM with the lowest GMRQ ($t_1 \sim 80$ ns). The implied timescale of the first tIC associated with the best MSM (using a lag time of 2 ns) was about 5 ns, about 20 times lower than the bound established by the MSM. These results show that an MSM trained on dimensionally reduced and finely discretized trajectories provides a much better estimate of the slow modes of the IS-HSA complex than the tICA analysis alone, and that cross-validation provides a notably better MSM. In the following results and discussion, we work to assign the implied timescales of our cross-validated MSM to specific structural changes and binding mode transitions of the IS-HSA complex.

Macrostates in coarse-grained MSM correspond to previously identified binding modes for the IS-HSA complex

It is difficult to directly connect the slow relaxation times from the MSM to conformational changes when considering the finely discretized microstates used to build the MSM. We used the robust Perron cluster analysis (PCCA+) algorithm to develop coarse-grained MSM by grouping related microstates into macrostates.[150] The number of macrostates is an input parameter to the PCCA+ algorithm, with fewer macrostates providing more interpretability at the cost of kinetic information. We performed coarse-graining over a range of granularities and found that 4 was the smallest number of macrostates we could consider without losing information key to our ultimate molecular interpretation of the MSM (**Figure II.24**). **Figure 4.2.2** illustrates the relative tractability of interrogating a coarse-grained MSM with 4 macrostates vs. a fine-grained MSM with 200 microstates.

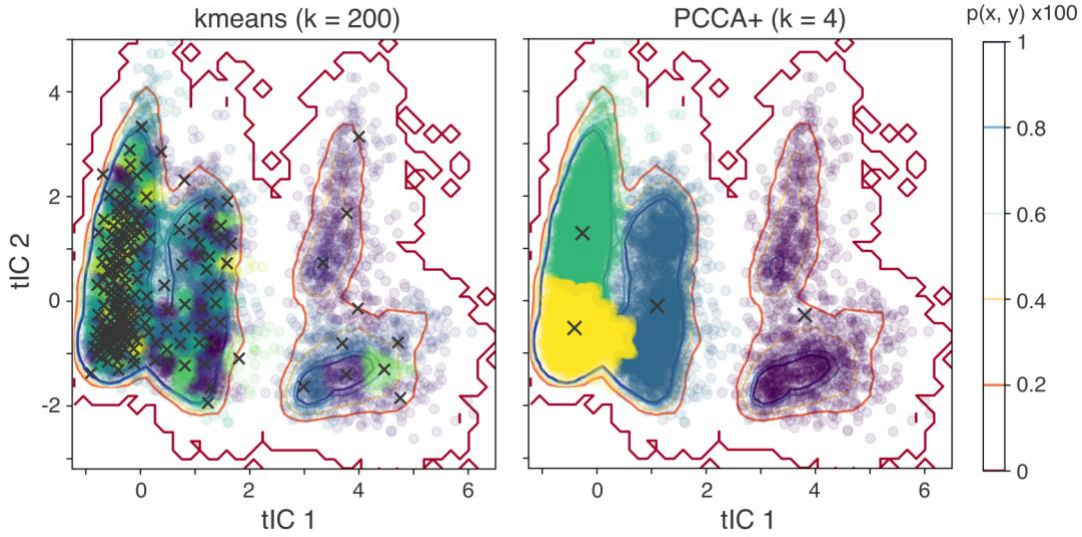

Figure 4.2.2 – Clustering results before and after coarse-graining with the PCCA+ algorithm. (left) The MD conformations from all 15 simulations, projected onto the first 2 independent components and colored by microstate from k-means clustering (k = 200). The center of each microstate is marked with an "x". (right) The same MD conformations, projected onto the first 2 tICs and colored by macrostate from PCCA+ clustering (k = 4). Macrostate centers are again marked with an "x". (both) Contours for the underlying probability density function in tIC space are marked at probability intervals of 0.002, ranging from dark red to dark blue for values from 0 to 0.01, respectively.

Whereas the 200 microstates are difficult to visually separate, the 4 macrostates identified by the PCCA+ algorithm are neatly separated and aligned with the probability contours of the underlying 2-dimensional tICA projection. Just as we identified representative structures for various PBUT binding modes in Section 4.1, we pulled representative structures from each PCCA+ macrostate to connect our kinetic model to transitions between specific molecular conformations. We also generated a coarse-grained transition probability matrix by summing the inter-macrostate transition probabilities for the microstates associated with each larger cluster. **Figure 4.2.3** shows a network representation of the resultant MSM with representative conformations for each macrostate embedded in the nodes and transition probabilities encoded by the edge widths.

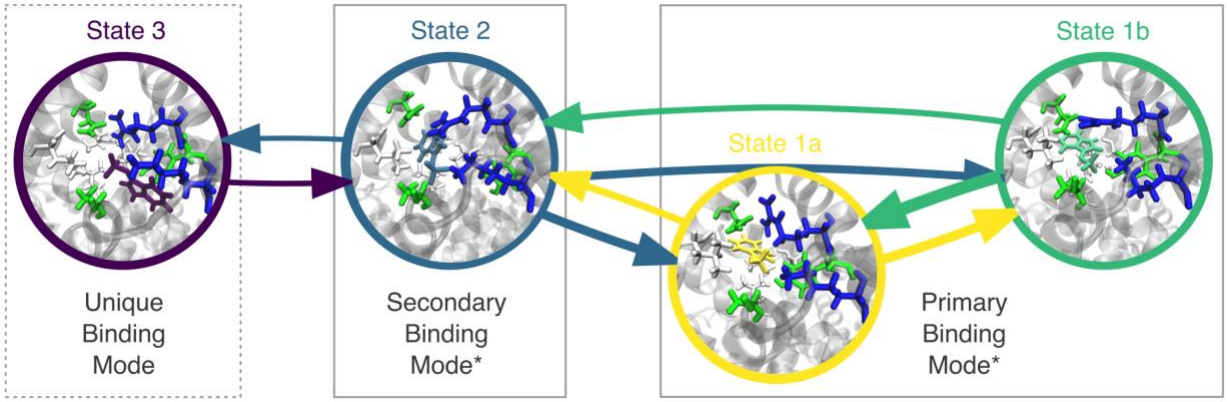

Figure 4.2.3 – Network visualization of the coarse-grained Markov State Model for the IS-HSA complex. Representative structures are embedded in the nodes for each state, taken from the MD frame with the closest tIC values to the average value of the macrostate. The thickness of the edges is proportional to the transition probability from one state to another. **macrostate corresponds to binding mode previously identified in Section 4.1.*

As noted in **Figure 4.2.3**, we found direct correspondence between the 3 of the macrostates of our MSM and the binding modes we identified for the IS-HSA in our previous work. Macrostates 1a and 1b correspond to the primary IS-HSA binding mode reported in Section 4.1. The only apparent difference between these microstates is whether the transient salt bridge between the IS sulfate and the R410 sidechain is intact (yes for microstate 1a, no for microstate 1b). These macrostates interconvert rapidly, as implied by the relatively thick lines and arrows connecting them in **Figure 4.2.3**. This is in good agreement with our previous discussion about the relatively fast movement of the R410 sidechain compared to the shifting of IS between binding modes (Section 4.1). Macrostate 2 is similar to the secondary binding mode we discussed in the previous section, although IS seems to interact more closely with K414 in the representative structure for State 2 than was observed in our previous simulations.

Macrostate 3 represents a previously unknown binding mode for IS-HSA, in which the indole group of IS has completely reoriented in the binding pocket. Interestingly, macrostate 3 was observed in only one of our 15 classical simulations (**Figure II.25**). Further investigation, discussed in the upcoming Section 4.3, revealed that this relatively rare binding mode plays an important role in the kinetics of the unbinding process. The large extent of States 2 and 3 in the tIC2 direction and the strong correlation between tIC2 and R410 center of mass distance suggest that these states include conformations with and without the R410 salt bridge. Indeed, performing PCCA+ with 6 macrostates divides each of these states neatly along the tIC2 axis (**Figure II.24**). In the remaining results and discussion, we describe the correspondence between macrostate-macrostate transitions and the relaxation timescales estimated by the original fine-grained MSM.

The slowest implied timescales of the MSM correspond to transitions between the macrostates

The relaxation timescales can be linked to specific molecular changes by taking the correlation between order parameters and the projection on the eigenvector modes of the MSM. We found the 1st and 2nd slowest relaxation modes correlated best with the IS-K414 center of mass distance, and that the 3rd and 4th slowest relaxation modes correlated best with the IS-R410 center of mass distance. **Figure 4.2.4** shows the projection of each MD frame onto the eigenvectors of the MSM vs. the order parameters with which they are most strongly correlated.

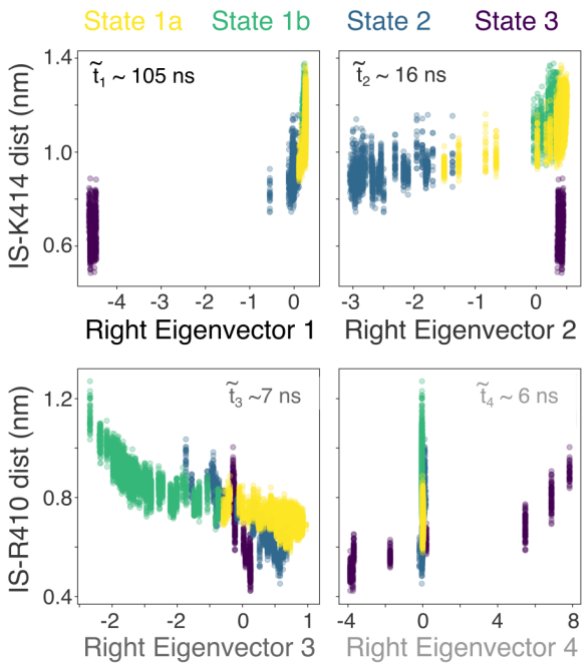

Figure 4.2.4 – Molecular order parameters vs. the projections onto the first 4 eigenvectors of the MSM transition matrix. (top) IS-K414 center of mass distance vs. projection onto the first and second eigenvectors of the MSM transition matrix, with corresponding timescale estimates of 105 ns and 16 ns, respectively. (bottom) IS-R410 center of mass distance vs. projection onto the third and fourth eigenvectors of the MSM transition matrix, with corresponding timescale estimates of 7 ns and 6 ns, respectively.

Figure 4.2.4 shows that the relaxation timescales can be straightforwardly connected to transitions between (and within) the 4 macroscopic binding modes. The slowest process, with an estimated characteristic timescale of 105 ns, is the transition from the secondary binding mode (State 2) to the unique binding mode sampled only in simulation 9 (State 3). The second slowest process – relaxation timescale of 16 ns – was similarly related to a transition between clusters, moving from the primary to the secondary binding pose. These 2 slow processes are both captured by the tIC 1, which does an excellent job of discriminating between the major binding modes of the IS-HSA complex.

The 3rd and 4th relaxation timescales, encoded by eigenvectors 3 and 4 of the MSM transition matrix, are associated with the breaking of the sulfate-R410 salt bridge within different binding modes. Eigenvector 3 distinguishes state 1a from state 1b, and eigenvector 4 distinguishes various microstates within macrostate 3. The timescale of breaking the sulfate-R410 salt bridge seems to be relatively insensitive to the position of IS, be it in the primary binding mode (t_3 ~ 7 ns) or the unique binding mode (t_4 ~ 6 ns). The tIC 2 direction, which is highly correlated with the IS-R410 distance, captures the 3rd and 4th relaxation timescales.

Conclusions

In this section, we estimated the slow relaxation timescales of the IS-HSA complex and identified collective variables that discriminate between macrostates that are kinetically separated. Our results provide more quantitative context for the potential role in unbinding of the previously-identified binding modes of the IS-HSA complex. The observation that the timescale for transitions between metastable binding modes exceeds 100 ns suggests a CV that directly accounts for these transitions should be used to efficiently calculate the unbinding time with infrequent metadynamics simulations. In Section 4.3, we extend our investigation to the dissociation of the IS-HSA complex, providing a direct link between our molecular level insights and experimental observables.

4.3 Calculating PBUT-HSA dissociation rates with infrequent metadynamics

Introduction

Synthetic materials designed to specifically bind the protein bound uremic toxins (PBUTs) from the bloodstream of patients being treated for chronic kidney disease (CKD) are in direct competition with the natural binding partner of PBUTs, human serum albumin (HSA). The efficacy of these synthetic materials is thus constrained by the binding affinity and kinetics of the PBUT-HSA binding process. In particular, the dissociation rate, k_{off}, of the toxin-protein complex will dictate the amount of free toxin that can be cleared via a PBUT separations process – such as adsorption. While equilibrium binding affinities have been measured for most PBUTs of interest,[125,138,151] PBUT-HSA binding kinetics have not been reported. The specialized equipment and experience required for acquiring k_{off} measurements from experiment may be prohibitive.[152]

The infrequent metadynamics technique has been established as a useful tool for quantifying protein-ligand unbinding.[5,145,146,153–155] Direct correspondence was recently demonstrated between the unbinding rates calculated with infrequent metadynamics and those measured experimentally for the protein FKBP and two small molecule ligands.[156] Infrequent metadynamics could be an effective computational strategy for directly calculating the IS-HSA dissociation rate. Accurately modeling the kinetics of a biomolecular system with infrequent metadynamics requires an *a priori* understanding of collective variables that capture the slowest degrees of freedom. Building upon the molecular and kinetic insights from our previous investigation of the IS-HSA structures, we are poised to exploit this more advanced simulation technique for protein-ligand unbinding.

In the previous section, we showed that the time-structure independent component analysis (tICA) could effectively distinguish between various metastable binding poses of the IS-HSA complex, and suggested the potential relationship between these poses and the unbinding process. The two key molecular process identified in our previous work, the shifting of the IS molecule from deep in the hydrophobic core of the Sudlow site II and the fluctuation of the charge-stabilizing R410 position are undoubtedly important steps on the unbinding process. In this work we demonstrate that the same tICA coordinates, learned from simulation of the bound pose only, serve as acceptable collective variables to facilitate the dissociation of the IS-HSA complex in infrequent metadynamics simulations. We report an estimated dissociation rate of 91 s^{-1} for the IS-HSA complex, based on a Poisson fit to the accelerated unbinding times from 15 unbinding trajectories.

Methods

Infrequent metadynamics with tICA collective variables

We performed 15 infrequent metadynamics simulations of the IS-HSA complex, one starting from the atomic positions and velocities in the final frame of each of the 15 classical simulations from Section 4.2. We considered as collective variables (CVs) the first 2 time-structure independent components from a tICA analysis of the 15 classical trajectories with a time lag of 10 ns, which we refer to as tIC1 and tIC2. Preliminary simulations biasing these CVs did not facilitate rapid unbinding in our simulations, likely because of the relatively fast timescale of tIC2. For production unbinding simulations we instead used tIC1 and the center of mass distance between the IS molecule and the alpha carbons of the binding pocket (BP) residues. These two collective variables should account for the potentially long timescale transitions between metastable binding modes (tIC1) and transitions from binding conformations to dissociated conformations (IS-BP distance).

The temperature and pressure coupling scheme and MD timestep were identical to those in our classical MD simulations (described in Section 4.2). A 2-dimensional well-tempered metadynamics bias was added to the system every 5 ps with a bias factor of 14, and initial hill height of 2.0 kJ/mol. These biasing parameters were identical to those reported by Tiwary in a seminal paper using

infrequent metadynamics for calculating protein-ligand dissociation kinetics.[154] The sigma values for Gaussian hills were 0.02 and 0.03 for tIC1 and IS-BP distance, respectively. The simulations were performed until IS was committed to the unbound state, as determined by the value of tIC1 exceeding a threshold of 5. Unbinding simulations took between 10 and 50 ns of simulation time. Accelerated unbinding times for the 15 unbinding trajectories were calculated to range from 22 ms to 10^4 s, using the method introduced by Tiwary and Parrinello.[5]

It has been previously established that a collection of unbinding times calculated from infrequent metadynamics simulations should be Poissonian if the biasing has not corrupted the underlying dynamics of the process at hand.[157] We used an in-house analysis script in the Python 3 programming language to construct the empirical cumulative distribution function (ECDF) for our calculated binding times and to perform a least squares fit to a Poissonian cumulative distribution function (CDF). We used the two sample Kolmogorov-Smirnoff test (KS test), as implemented in the scipy Python library, to test the null hypothesis that the ECDF and the fit CDF were generated by the same underlying distribution. Ratios of the sample mean (μ), standard deviation (σ), and median (t_m) were also calculated and compared to the expected ratios for a Poisson sample. This analysis script has also been made publicly available at www.github.com/UWPRG/uremic_toxins.

Results and Discussion

Estimating the IS-HSA dissociation rate with tICA and infrequent metadynamics

All 15 infrequent metadynamics simulations produced unbinding events in less than 60 ns of simulation time. We calculated the accelerated unbinding time for each unbinding trajectory, following the original procedure described by Tiwary and Parrinello.[5] The accelerated binding time is simply calculated by adding together the number of recorded timesteps, weighted by the saving frequency (10 fs) and a Boltzmann factor derived from the value of the bias potential experienced at each frame. We then fit a Poisson distribution to the empirical cumulative distribution function (ECDF) of the 15 unbinding times, and assessed the goodness of fit (GoF) to Poissonian statistics. Poisson The ECDF for our accelerated unbinding times is plotted in **Figure 4.3.1**, along with the CDF for the Poisson distribution of best fit and several measures of the correspondence to Poissonian statistics of the ECDF.

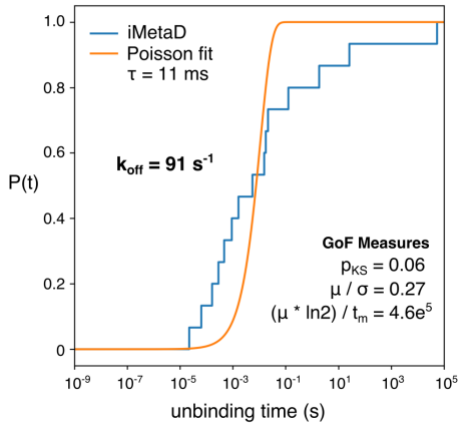

Figure 4.3.1 – Poisson fit and unbinding rate estimated from the accelerated unbinding times from infrequent metadynamics simulations. The empirical cumulative distribution function of unbinding times and the Poisson cumulative distribution function of best fit are represented by blue and orange lines, respectively. The estimated dissociation rate, k_{off}, is given to the left of the curves. To the right, various statistical measures are given for the Goodness of Fit (GoF) of the empirical times to the corresponding Poisson distribution. The p-value of the KS test (p_{KS}), the ratio of the sample mean and standard deviation (μ/σ), and the log-weighted ratio of the sample mean to the sample median ($\mu \ln(2)/t_m$) should assume values near unity for Poisson statistics.

The unbinding times calculated with infrequent metadynamics ranged from 0.022 ms to 10^5 s. This range encompasses the average dissociation time previously reported for IS in complex with bovine serum albumin (0.67 ms), which was measured using ^{1}H relaxation dispersion NMR spectroscopy.[158] The IS binding site of BSA is nearly identical to the Sudlow site II of HSA, so it is expected that the IS-BSA dissociation time serves as a good ballpark estimate of the IS-HSA dissociation time.

Fitting the unbinding times to the nearest Poisson distribution resulted in an estimated unbinding time of 11 ms, corresponding to a k_{off} of 91 s^{-1}. Performing the 2 sample KS test with the unbinding rates and 5000 draws from the nearest Poisson distribution resulted in a p value of 0.06, just above the traditionally used cutoff for rejecting the null hypothesis ($p = 0.05$). In agreement with the relatively low KS score for our unbinding rates, other goodness of fit metrics for Poisson statistics ($\mu/\sigma = 1$; $\mu \ln(2)/t_m = 1$) suggest a relatively poor fit to our unbinding rates ($\mu/\sigma = 0.27$; $\mu \ln(2)/t_m = 4.6e^5$). These metrics are much more sensitive to outliers in the long timescale tail of the ECDF, and are considered less reliable than the KS test. Departure from Poissonian statistics could be explained by a poor selection of CVs, which would corrupt the underlying dynamics of the system, or the existence of different unbinding pathways, which should have separately Poisson distributions. We explore the latter possibility below.

Key intermediate states identified along the unbinding pathway

Our previous investigations of the IS-HSA complex revealed several metastable binding modes for IS in Sudlow site II, which are all sampled in our unbinding trajectories. In Section 4.2, a lower bound of 105 ns was established for the slowest transitions observed between these metastable states. If the round trip transition times between metastable states occur on a timescale commensurate with the unbinding time, then contributions from various unbinding pathways – with IS visiting different subsets of the binding modes – could account for the wide range of calculated unbinding times. We compared trajectories with relatively fast unbinding times ($t_{acc} < \tau$) and relatively slow unbinding times ($t_{acc} > \tau$) to look for differences in the unbinding pathways. **Figure 4.3.**2 shows representative pathways for fast and slow unbinding.

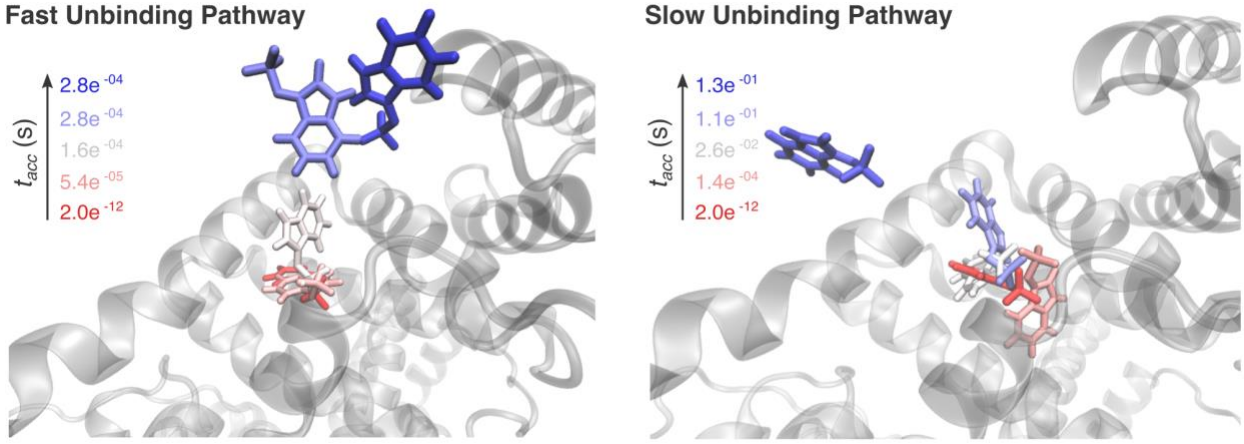

Figure 4.3.2 – Intermediate binding positions of IS along alternative binding pathways. (left) Representative IS positions for the unbinding pathway associated with low residence time trajectories ($t_{acc} < \tau$). IS shifts from the primary IS-HSA binding mode to the secondary IS-HSA binding mode (red to pink), then flips its indole group out of the binding pocket (pink to white) before quickly unbinding (white to lavender to blue). (right) Representative IS positions for the unbinding pathway associated with low residence time trajectories ($t_{acc} > \tau$). IS visits the tertiary binding mode of the IS-HSA complex (red to pink) and must shift back to the secondary binding mode (pink to white) before unbinding (white to lavender to blue).

Two unbinding pathways with vastly different timescales were identified. In the fast unbinding pathway (t_{acc} < 1 ms), IS moves from its primary binding mode to its secondary binding mode near the mouth of binding pocket, then directly transitions into a presolvated state, where the indole group is oriented toward solution. In the slow unbinding pathway (t_{acc} > 100 ms), IS takes a more circuitous route, transitioning to an alternative binding mode from the secondary binding mode rather than proceeding directly to the presolvated state. The excursion to the alternative binding mode takes tens of ms before IS moves back to the secondary binding mode and proceeds according to the rapid unbinding pathway. These example binding paths suggest that our unbinding times are actually drawn from a heterogenous distribution of processes.

Conclusions

In this work, we have started to bridge the gap between nanoscale and macroscale observables for the IS-HSA unbinding process. Our simulations have produced a collection unbinding trajectories that capture the primary unbinding pathways of IS in atomistic detail. Of the 15 unbinding trajectories, 6 produced k_{off} estimates within an order of magnitude of the experimental unbinding rates measured for IS unbinding from BSA (0.67 ms),[158] a nearly identical point of reference to IS-HSA. While our collection of unbinding times does not match Poissonian statistics to the degree of statistical certainty to confidently assert a single *in silico* unbinding time for the IS-HSA complex, we are close enough to begin incorporating our range of calculated k_{off} values into device scale models for PBUT clearance.

In ongoing work in the Pfaendtner Research Group, we have developed a numerical model for a PBUT adsorption column to quantify the implications for CKD treatment of k_{off} values across our estimated range. Preliminary results suggest that an order of magnitude change in the IS-HSA unbinding rate can dramatically impact the therapeutic lifetime of a small PBUT adsorption column. Strategies for improving the accuracy of the IS-HSA dissociation rate calculated from MD will increase the utility of computer-aided device design, and should be explored in future work. Further investigation into the selectivity of IS for various unbinding pathways should provide a better estimate of the overall IS-HSA dissociation rate. Simply increasing the number of infrequent metadynamics simulations with the same CVs and biasing parameters could also improve the accuracy of the computational k_{off} estimate.

Chapter 5
Conclusions

5.1 Impact

Most problems of scientific interest are hard. Hard scientific problems often persist longer than the timescale of a single PhD.* It typically takes a collection of researchers, separated in time and space, working tirelessly and scientific-methodically to push the boundaries of a narrow body of knowledge before consensus about a physical phenomenon can be established. The boundary of knowledge surrounding a particular problem can vary in plasticity, based largely on the amount of effort (in terms of scientist years) that has previously been exerted to extend the boundary. The primary thrusts of my graduate research have been dedicated to 3 research projects that I joined at different stages of maturity. My role as a computational scientist and the nature of my contributions differed in each case.

Researchers in the Jiang lab have investigated the molecular origins of the nonfouling properties of zwitterionic materials for almost 15 years.[33,67,159–161] Our understanding of zwitterionic materials has stagnated with a focus on "superhydrophilicity", a property which remains difficult to describe quantitatively. In the pursuit of new fundamental insights on this thoroughly trodden theoretical ground, I applied novel simulation and analysis techniques to provide a more holistic description of biomolecule hydration than did previous researchers (Chapter 2). My work has emphasized the importance of considering the asymmetry in water affinities that is inherent in heterogeneous biomaterials. I have contributed methods for quantifying the amphiphilicity and the relative populations of weakly and strongly bound water, which can be used to refine our understanding of the behavior of biomaterials in physiologically relevant environments.

Linking the physical properties of a peptide chain to its performance as a protective domain for protein drug delivery is a challenging problem that has not been previously approached with MD and ML. By developing insights into the amino acid contributions to protein-protein interactions (Section 3.1) and salt bridge propensities between charged amino acid sidechains (Section 3.2), I effectively narrowed the design space for peptide-based materials with natural resistance to nonspecific interactions. I provided reason to believe that the current state-of-art protective peptide used in the Jiang group is not optimal for drug delivery applications, and I made a concrete recommendation for amino acid mutations that will produce a peptide with more desirable physical properties (Section 3.3). If it is confirmed that superhydrophilic, conformationally disordered peptides can increase the pharmacokinetic properties of protein drugs, my simulation methodology will provide an exceptionally useful means of predicting peptide performance from simulations.

Prior to my research on protein bound uremic toxin (PBUT) interactions with HSA, there was virtually no information available about the molecular scale behavior of PBUTs. I have identified key natural chemistries for toxin stabilization (Section 4.1), which are being considered in the design of PBUT capture materials by other researchers in the CDI. My estimates for IS-HSA dissociation rate have also enabled numerical modeling of hypothetical PBUT adsorption columns, which is guiding device-scale design. By investigating PBUT interactions at the molecular scale, I have provided a new lens through which other researchers can view the problem of PBUTs in CKD.

More broadly, my graduate research has demonstrated the tremendous value of coupling MD and ML methods. Together, MD and ML can provide unparalleled resolution and interpretability for

* Markov state model to establish a lower bound on the timescale for addressing hard scientific problems is forthcoming.

investigating biophysical phenomenon. By making my analysis code available and favoring open source software, I have also provided an example for reproducible research in computational science.

5.2 Future work

The rational design of zwitterionic materials with improved nonfouling properties will remain out of reach until a means of quantitatively predicting nonfouling performance from structure or hydration can be established. Future work in this area should build upon my preliminary attempts to quantify the amphiphilicity and distinct hydration modes of zwitterionic materials. The best molecular order parameters for describing distinct hydration states should be identified, and probabilistic definitions of hydration used to distinguish between the hydration modes of various materials. While I was unable to bring this work to an emphatic conclusion, I hope the next researcher to take up the gauntlet of molecular biomaterials research will directly extend the final portion of my research (Section 2.3) to describe the complex phenomenon of water-mediated resistance to nonspecific interactions.

The search for protective peptide domains for fusion proteins is ongoing, and the need for further physical and computational characterization is growing. Physical characterization of production-length polyEK and polyEKGG should be performed to establish a connection between the predicted properties from my simulations and the longer polypeptides used in practice. A systematic study should be performed to verify whether a peptide that is both superhydrophilic and disordered provides better *in vivo* performance for drug delivery than peptides with only one property or the other. Computational studies, following the protocol I have described for EK-based peptides, should focus on the effects on hydration and conformation of various mutations to the polyEK template, using small polar amino acids other than glycine. A broader mapping of the sequence-property-function relationships for alternating charge protective peptides will provide a path toward rational peptide design.

My research into PBUTs has provided fertile ground for further investigation. Future work should follow my example of estimating the IS-HSA dissociation rate to calculate unbinding kinetics for the other Sudlow site II PBUTs. Similar methodology should also be applied to CMPF in Sudlow site I, and extended to any other PBUTs that are identified as Sudlow I binders in the future. It will also be illustrative to apply similar MD and ML techniques to interactions between toxins and synthetic materials to provide an *in silico* screening platform for high affinity PBUT binders. These research directions promise a drastically accelerated and more efficient search for materials strategies to address the accumulation of PBUTs in CKD patients.

References

1. Keefe, A. J. & Jiang, S. Poly(zwitterionic)protein conjugates offer increased stability without sacrificing binding affinity or bioactivity. *Nat. Chem.* **4,** 59–63 (2011).
2. Zhang, L. *et al.* Zwitterionic hydrogels implanted in mice resist the foreign-body reaction. *Nat. Biotechnol.* **31,** 553–556 (2013).
3. Yu, W. & Mackerell, A. D. Docking and scoring in virtual screening for drug discovery. *Methods Mol. Biol.* **1520,** 85–106 (2017).
4. Liu, X. *et al.* Molecular dynamics simulations and novel drug discovery. *Expert Opin. Drug Discov.* **13,** 23–37 (2018).
5. Tiwary, P. & Parinello, M. From metadynamics to dynamics. *Phys. Rev. Letters.* **111,** 230602 (2013).
6. Limongelli, V. *et al.* Funnel metadynamics as accurate binding free-energy method. *Proc. Natl. Acad. Sci.* **110,** 6358–6363 (2013).
7. Pérez-hernández, G. *et al.* Identification of slow molecular order parameters for Markov model construction. *J. Chem. Phys.* **139,** 015102 (2013).
8. Schwantes, C. R. & Pande, V. S. Improvements in Markov state model construction reveal many non-native interactions in the folding of NTL9. *JCTC.* **9,** 2000-2009 (2013).
9. Jung, J. *et al.* Scaling molecular dynamics beyond 100,000 processor cores for large-scale biophysical simulations. *J. Comput. Chem.* jcc.25840 (2019).
10. Sultan, M. M. & Pande, V. S. tICA-metadynamics: Accelerating metadynamics by using kinetically selected collective variables. *JCTC.* **13,** 2440-2447 (2017).
11. Signorelli, S., *et al.* Structural characterization of the intrinsically disordered protein p53 using raman spectroscopy. *Appl. Spectrosc.* **71,** 823–832 (2017).
12. Karamzadeh, R. *et al.* Machine learning and network analysis of molecular dynamics trajectories reveal two chains of red/ox-specific residue interactions in human protein disulfide isomerase. *Sci. Rep.* **7,** 1–11 (2017).
13. Maisuradze, G. G. *et al.* Principal Component Analysis for Protein Folding Dynamics. *J. Mol. Bio.* **385,** 312-329 (2009).
14. Gasparotto, P. & Ceriotti, M. Recognizing molecular patterns by machine learning: An agnostic structural definition of the hydrogen bond. *J. Chem. Phys.* **141,** 1–14 (2014).
15. Gasparotto, P. *et al.* Recognizing local and global structural motifs at the atomic scale. *JCTC.* **14,** 486-498 (2018).
16. Li, B. *et al.* Mitigation of inflammatory immune responses with hydrophilic nanoparticles. *Angew. Chemie - Int. Ed.* 4527–4531 (2018).
17. Jiang, S. & Cao, Z. Ultralow-fouling, functionalizable, and hydrolyzable zwitterionic materials and their derivatives for biological applications. *Adv. Mater.* **22,** 920–932 (2010).
18. Leng, C. *et al.* In situ probing of the surface hydration of zwitterionic polymer brushes: Structural and environmental effects. *J. Phys. Chem. C* **118,** 15840–15845 (2014).
19. Leng, C. *et al.* Molecular level studies on interfacial hydration of zwitterionic and other antifouling polymers in situ. *Acta Biomater.* **40,** 6–15 (2016).
20. Ishihara, K. Successful Development of biocompatible polymers designed by natures original inspiration. *Procedia Chem.* **4,** 34–38 (2012).
21. Zhang, Z. *et al.* Superlow fouling sulfobetaine and carboxybetaine polymers on glass slides. *Langmuir* **22,** 10072–10077 (2006).
22. Usui, K. *et al. Ab Initio* liquid water dynamics in aqueous TMAO solution. *J. Phys. Chem. B* **119,** 10597–10606 (2015).
23. Stirnemann, G. *et al.* Ab Initio simulations of water dynamics in aqueous TMAO solutions: Temperature and concentration effects. *J. Phys. Chem. B* **121,** 11189-11197 (2017).
24. Bayly, C. I. *et al.* A well-behaved electrostatic potential based method using charge restraints for deriving atomic charges: The RESP model. *J. Phys. Chem.* **97,** 10269–10280 (1993).
25. Becke, A. D. Density-functional thermochemistry. III. The role of exact exchange. *J. Chem. Phys.* **98,** 5648 (1993).
26. Petersson, G. A. *et al.* A complete basis set model chemistry. I. The total energies of closed-shell atoms and hydrides of the first-row elements. *J. Chem. Phys.* **89,** 2193 (1988).
27. Jorgensen, W. L. *et al.* Comparison of simple potential functions for simulating liquid water. *J Chem Phys* **79,** 926–935 (1983).
28. Abraham, M. J. *et al.* Gromacs: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* **1–2,** 19–25 (2015).

29. Bussi, G. *et al.* Canonical sampling through velocity rescaling. *J. Chem. Phys.* **126,** 014101 (2007).
30. Berendsen, H. J. C. *et al.* Molecular dynamics with coupling to an external bath. *J. Chem. Phys.* **81,** 3684–3690 (1984).
31. Parrinello, M. & Rahman, A. Polymorphic transitions in single crystals: A new molecular dynamics method. *J. Appl. Phys.* **52,** 7182–7190 (1981).
32. Luzar, A. & Chandler, D. Hydrogen-bond kinetics in liquid water. *Nature* **379,** 55–57 (1996).
33. Shao, Q. & Jiang, S. Influence of charged groups on the properties of zwitterionic moieties: A molecular simulation study. *J. Phys. Chem. B* **118,** 7630–7637 (2014).
34. Shao, Q. *et al.* Differences in cationic and anionic charge densities dictate zwitterionic associations and stimuli responses. *J. Phys. Chem. B* **118,** 6956–6962 (2014).
35. Shao, Q. & Jiang, S. Effect of carbon spacer length on zwitterionic carboxybetaines. *J. Phys. Chem. B* **117,** 1357–1366 (2013).
36. White, A. D. *et al.* Free energy of solvated salt bridges: A simulation and experimental study. *J. Phys. Chem. B* **117,** 7254–7259 (2013).
37. Pfaendtner, J. & Bonomi, M. Efficient Sampling of High-Dimensional Free-Energy Landscapes with Parallel Bias Metadynamics. *J. Chem. Theory Comput.* **11,** 5062–5067 (2015).
38. Frisch, M. J. *et al.* Gaussian 09. (2010).
39. Torrie, G. M. & Valleau, J. P. Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling. *J. Comput. Phys.* **23,** 187–199 (1977).
40. Schneck, E. *et al.* Insight into the molecular mechanisms of protein stabilizing osmolytes from global force-field variations. *J. Phys. Chem. B* **117,** 8310–8321 (2013).
41. Ohto, T. *et al.* Unveiling the amphiphilic nature of TMAO by vibrational sum frequency generation spectroscopy. *J. Phys. Chem. C* **120,** 17435–17443 (2016).
42. Morita, S. *et al.* Time-resolved in situ ATR-IR observations of the process of sorption of water into a poly(2-methoxyethyl acrylate) film. *Langmuir* **23,** 3750–3761 (2007).
43. Tanaka, M. *et al.* The roles of water molecules at the biointerface of medical polymers. *Polym. J.* **45,** 701–710 (2013).
44. Sato, K. *et al.* Synthesis and thrombogenicity evaluation of poly(3-methoxypropionic acid vinyl ester): A candidate for blood-compatible polymers. *Biomacromolecules* **18,** 1609–1616 (2017).
45. Soto, A. *et al.* Unveiling two types of local order in liquid water using machine learning. *arXiv* 1707.04593v1 (2017).
46. Gasparotto, P. *et al.* Recognizing local and global structural motifs at the atomic scale. *J. Chem. Theory Comput.* **14,** 486–498 (2018).
47. Wang, J. M. *et al.* Development and testing of a general amber force field. *J. Comput. Chem.* **25,** 1157–1174 (2004).
48. Van Der Spoel, D. *et al.* GROMACS: Fast, flexible, and free. *J. Comput. Chem.* **26,** 1701–1718 (2005).
49. Tribello, G. A. *et al.* PLUMED 2: New feathers for an old bird. *Comput. Phys. Commun.* **185,** 604–613 (2014).
50. Steinhardt, P. J. *et al.* Bond-orientational order in liquids and glasses. *Phys. Rev. B* **28,** 784–805 (1983).
51. Eldar, Y. *et al.* The farthest point strategy for progressive image sampling. *IEEE Trans. Image Proc.* **6,** 1305–1315 (1997).
52. McGibbon, R. T. *et al.* MDTraj: A modern open library for the analysis of molecular dynamics trajectories. *Biophys. J.* **109,** 1528–1532 (2015).
53. Pisal, D. S. *et al.* Delivery of protein therapeutics. *J Pharm Sci.* **99,** 2557–2575 (2010).
54. Turecek, P. L. *et al.* PEGylation of biopharmaceuticals: A review of chemistry and nonclinical safety information of approved drugs. *J. Pharm. Sci.* **105,** 460–475 (2016).
55. Banerjee, S. S. *et al.* Poly(ethylene glycol)-prodrug conjugates: Concept, design, and applications. *J. Drug Deliv.* **2012,** 1–17 (2012).
56. Veronese, F. M. & Mero, A. The impact of PEGylation on biological therapies. *BioDrugs* **22,** 315–329 (2008).
57. Swierczewska, M. *et al.* What is the future of PEGylated therapies? *Expert Opin. Emerg. Drugs* **20,** 531–536 (2015).
58. Kozlowski, A. & Milton Harris, J. Improvements in protein PEGylation: Pegylated interferons for treatment of hepatitis C. *J. Control. Release* **72,** 217–224 (2001).
59. Liu, E. J. *et al.* EKylation: Addition of an alternating-charge peptide stabilizes proteins. *Biomacromolecules* **16,** 3357–3361 (2015).
60. Liu, E. J. & Jiang, S. Expressing a monomeric organophosphate hydrolase as an EK fusion protein. *Bioconjug. Chem.* **29,** 3686–3690 (2018).
61. Pfister, D. & Morbidelli, M. Process for protein PEGylation. *J. Control. Release* **180,** 134–149 (2014).
62. Bendele, A. *et al.* Renal tubular vacuolation in animals treated with polyethylene-glycol-conjugated proteins. *Toxicol. Sci.* **42,** 152–157 (1998).

63. Sear, R. P. Highly specific protein-protein interactions, evolution and negative design. *Phys. Biol.* **1,** 166–172 (2004).
64. Gonzalez, M. W. & Kann, M. G. Chapter 4: Protein Interactions and Disease. *PLoS Comput. Biol.* **8,** (2012).
65. Brown, C. J. *et al.* Awakening guardian angels: drugging the p53 pathway. *Nat. Rev. Cancer* **9,** 862–873 (2009).
66. White, A. D. *et al.* Decoding nonspecific interactions from nature. *Chemical Science* **3,** 3488 (2012).
67. Shao, Q. & Jiang, S. Molecular understanding and design of zwitterionic materials. *Adv. Mater.* **27,** 15–26 (2015).
68. Berman, H. M. *et al.* The protein data bank. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **58,** 899–907 (2002).
69. Kastritis, P. L. *et al.* A structure-based benchmark for protein-protein binding affinity. *Protein Sci.* **20,** 482–491 (2011).
70. Vreven, T. *et al.* Updates to the integrated protein-protein interaction benchmarks: Docking benchmark version 5 and affinity benchmark version 2. *J. Mol. Biol.* **427,** 3031–41 (2015).
71. Moal, I. H. *et al.* Protein-protein binding affinity prediction on a diverse set of structures. *Bioinformatics* **27,** 3002–3009 (2011).
72. Janin, J. A minimal model of protein-protein binding affinities. *Protein Sci.* **23,** 1813–7 (2014).
73. Kastritis, P. L. *et al.* Proteins feel more than they see: fine-tuning of binding affinity by properties of the non-interacting surface. *J. Mol. Biol.* **426,** 2632–52 (2014).
74. Moal, I. H. *et al.* Inferring the microscopic surface energy of protein-protein interfaces from mutation data. *Proteins* **83,** 640–50 (2015).
75. Yugandhar, K. & Gromiha, M. M. Protein-protein binding affinity prediction from amino acid sequence. *Bioinformatics* **30,** 3583–3589 (2014).
76. Bickerton, G. R. *et al.* Comprehensive, atomic-level characterization of structurally characterized protein-protein interactions: the PICCOLO database. *BMC Bioinformatics* **12,** 313 (2011).
77. Vangone, A. & Bonvin, A. M. Contacts-based prediction of binding affinity in protein-protein complexes. *Elife* **4,** e07454 (2015).
78. Schwarz, G. Estimating the dimension of a model. *Ann. Stat.* **6,** 461–464 (1978).
79. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **12,** 2825–2830 (2012).
80. Efron, B. *et al.* Least angle regression. *Ann. Stat.* **32,** 407–499 (2004).
81. Ekmekci, B *et al.* An introduction to programming for bioscientists : A python-based primer. *PLOS Comput. Bio.* 1–43 (2016).
82. Moal, I. H. & Fernández-Recio, J. SKEMPI: A Structural Kinetic and Energetic database of Mutant Protein Interactions and its use in empirical models. *Bioinformatics* **28,** 2600–2607 (2012).
83. Kyte, J. & Doolittle, R. F. A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.* **157,** 105–132 (1982).
84. Bogan, A. A. & Thorn, K. S. Anatomy of hot spots in protein interfaces. *Journal of Molecular Biology* **280,** 1–9 (1998).
85. Del Sol, A. & O'Meara, P. Small-world network approach to identify key residues in protein-protein interaction. *Proteins Struct. Funct. Genet.* **58,** 672–682 (2005).
86. Cho, K. *et al.* A feature-based approach to modeling protein-protein interaction hot spots. *Nucleic Acids Res.* **37,** 2672 (2009).
87. Frederix, P. W. J. M. *et al.* Exploring the sequence space for (tri-)peptide self-assembly to design and discover new hydrogels. *Nat. Chem.* **7,** 30–37 (2014).
88. Fiorucci, S. & Zacharias, M. Prediction of protein-protein interaction sites using electrostatic desolvation profiles. *Biophys. J.* **98,** 1921–1930 (2010).
89. Nowinski, A. K *et al.* Biologically inspired stealth peptide-capped gold nanoparticles. *Langmuir* **30,** 1864–1870 (2014).
90. Nowinski, A. K. *et al.* Sequence, structure, and function of peptide self-assembled monolayers. *J. Am. Chem. Soc.* **134,** 6000–6005 (2012).
91. Kostidis, S. *et al.* The relative orientation of the Arg and Asp side chains defined by a pseudodihedral angle as a key criterion for evaluating the structure-activity relationship of RGD peptides. *J. Pept. Sci.* **10,** 494–509 (2004).
92. Case, D. A. Amber 2014. (2014).
93. Bonomi, M. *et al.* PLUMED: A portable plugin for free-energy calculations with molecular dynamics. *Comput. Phys. Commun.* **180,** 1961–1972 (2009).
94. Barducci, A. *et al.* Well-tempered metadynamics: A smoothly converging and tunable free-energy method. *Phys. Rev. Lett.* **100,** 1–4 (2008).
95. Schlapschy, M. *et al.* Fusion of a recombinant antibody fragment with a homo-amino-acid polymer: Effects on biophysical properties and prolonged plasma half-life. *Protein Eng. Des. Sel.* **20,** 273–284 (2007).
96. Schlapschy, M. *et al.* PASylation: A biological alternative to PEGylation for extending the plasma half-life of

pharmaceutically active proteins. *Protein Eng. Des. Sel.* **26,** 489–501 (2013).
97. Breibeck, J. & Skerra, A. The polypeptide biophysics of proline/alanine-rich sequences (PAS): Recombinant biopolymers with PEG-like properties. *Biopolymers* **109,** 1–12 (2018).
98. Schellenberger, V. *et al.* A recombinant polypeptide extends the in vivo half-life of peptides and proteins in a tunable manner. *Nat. Biotechnol.* **27,** 1186–1190 (2009).
99. Zhang, S., Holmes, T., Lockshin, C. & Rich, A. Spontaneous assembly of a self-complementary oligopeptide to form a stable macroscopic membrane. *Proc. Natl. Acad. Sci.* **90,** 3334–3338 (1993).
100. Zhang, S. *et al.* Self-complementary oligopeptide matrices support mammalian cell attachment. *Biomaterials* **16,** 1385–1393 (1995).
101. Sugita, Y. *et al.* Multidimensional replica-exchange method for free-energy calculations. *J. Chem. Phys.* **113,** 6042–6051 (2000).
102. Bussi, G. *et al.* Free-energy landscape for β hairpin folding from combined parallel tempering and metadynamics. *J. Am. Chem. Soc.* **128,** 13435–13441 (2006).
103. Bonomi, M. & Parrinello, M. Enhanced sampling in the well-tempered ensemble. *Phys. Rev. Lett.* **104,** 1–4 (2010).
104. Deighan, M. *et al.* Efficient simulation of explicitly solvated proteins in the well-tempered ensemble. *J. Chem. Theory Comput.* **8,** 2189–2192 (2012).
105. Bernetti, M. *et al.* Structural and kinetic characterization of the intrinsically disordered protein SeV NTAIL through enhanced sampling simulations. *J. Phys. Chem. B* **121,** 9572–9582 (2017).
106. Humphrey, W. *et al.* VMD: Visual molecular dynamics. *J. Mol. Graph.* **14,** 33–38 (1996).
107. Maier, J. A. *et al.* ff14SB: Improving the accuracy of protein side chain and backbone parameters from ff99SB. *J. Chem. Theory Comput.* **11,** 3696–3713 (2015).
108. Sprenger, K. G. & Pfaendtner, J. Strong electrostatic interactions lead to entropically favorable binding of peptides to charged surfaces. *Langmuir* **32,** 5690–5701 (2016).
109. Tiwary, P. & Parrinello, M. A time-independent free energy estimator for metadynamics. *Journal of Physical Chemistry B* **119,** 736–742 (2015).
110. Daura, X. *et al.* Peptide folding: When simulation meets experiment. *Angew. Chemie Int. Ed.* **38,** 236–240 (1999).
111. Gō, M. *et al.* Molecular theory of the helix-coil transition in polyamino acids. III. Evaluation and analysis of s and σ for polyglycine and poly-L-alanine in water. *J Chem Phys* **54,** 4489–4503 (1971).
112. Ananthanarayanan, V. S. *et al.* Helix-coil stability constants for the naturally occurring amino acids in water. III. Glycine parameters from random poly(hydroxybutylglutamine-co-glycine). *Macromolecules* **4,** 417–424 (1971).
113. Smith, L. J. *et al.* The concept of a random coil: Residual structure in peptides and denatured proteins. *Fold. Des.* **1,** R95-106 (1996).
114. Chou, P. Y. & Pasman, G. D. Conformational parameters for amino acids in helical, β-sheet, and random coil regions calculated from proteins. *Biochemistry* **13,** 211–222 (1974).
115. Costantini, S. *et al.* Amino acid propensities for secondary structures are influenced by the protein structural class. *Biochem. Biophys. Res. Commun.* **342,** 441–451 (2006).
116. Bhattacharjee, N. & Biswas, P. Position-specific propensities of amino acids in the -strand. *BMC Struct. Biol.* **10,** (2010).
117. Minor Jr., D. L. & Kim, P. S. Context is a major determinant of beta-sheet propensity. *Nature* **371,** 264-267 (1994).
118. Yang, A. S. & Honig, B. Free energy determinants of secondary structure formation: II. Antiparallel β-sheets. *J. Mol. Biol.* **252,** 366–376 (1995).
119. Fujiwara, K. *et al.* Dependence of alpha-helical and beta-sheet amino acid propensities on the overall protein fold type. *BMC Struct. Biol.* **12,** 1 (2012).
120. Yang, A. S. & Honig, B. Free energy determinants of secondary structure formation: I. A-Helices. *J. Mol. Biol.* **252,** 351–365 (1995).
121. Jones, S. *et al.* Protein domain interfaces: characterization and comparison with oligomeric protein interfaces. *Protein Eng.* **13,** 77–82 (2000).
122. Greenfield, N. J. Using circular dichroism spectra to estimate protein secondary structure. *Nat. Protoc.* **1,** 2876–2890 (2006).
123. Bush, C. A. *et al.* Circular dichroism of β-turns in peptides and proteins. *Biochemistry* **17,** 4951–4954 (1978).
124. Vanholder, R. C. *et al.* Assessment of urea and other uremic markers for quantification of dialysis efficacy. *Clin. Chem.* **38,** 1429–1436 (1992).
125. Itoh, Y. *et al.* Protein-bound uremic toxins in hemodialysis patients measured by liquid chromatography/tandem mass spectrometry and their effects on endothelial ROS production. *Anal. Bioanal. Chem.* **403,** 1841–1850 (2012).
126. Vanholder, R. *et al.* The uremic toxicity of indoxyl sulfate and p-cresyl sulfate: A systematic review. *J. Am. Soc. Nephrol.* **25,** 1897–1907 (2014).

127. Leong, S. C. & Sirich, T. L. Indoxyl sulfate-review of toxicity and therapeutic strategies. *Toxins (Basel).* **8,** (2016).
128. Gryp, T. *et al.* P-cresyl sulfate. *Toxins (Basel).* **9,** 1–24 (2017).
129. Sirich, T. L *et al.* Numerous protein-bound solutes are cleared by the kidney with high efficiency. *Kidney Int.* **84,** 585–590 (2013).
130. Sakai, T. *et al.* Characterization of binding site of uremic toxins on human serum albumin. *Biol. Pharm. Bull.* **18,** 1755–1761 (1995).
131. Tao, X. *et al.* Improved dialytic removal of protein-bound uraemic toxins with use of albumin binding competitors: An in vitro human whole blood study. *Sci. Rep.* **6,** 2–10 (2016).
132. Bohringer, F.. *et al.* Release of uremic retention solutes from protein binding by hypertonic predilution hemodiafiltration. *ASAIO J.* 55–60 (2015).
133. Krieter, D. H. *et al.* Haemodiafiltration at increased plasma ionic strength for improved protein-bound toxin removal. *Acta Physiol.* ***219,*** 510–520 (2017).
134. Sandeman, S. R. *et al.* A haemocompatible and scalable nanoporous adsorbent monolith synthesised using a novel lignin binder route to augment the adsorption of poorly removed uraemic toxins in haemodialysis. *Biomed. Mater.* **12,** 035001 (2017).
135. Sandeman, S. R. *et al.* An adsorbent monolith device to augment the removal of uraemic toxins during haemodialysis. *J. Mater. Sci: Meter. Med.* **25,** 1589–1597 (2014).
136. Ghuman, J. *et al.* Structural basis of the drug-binding specificity of human serum albumin. *J. Mol. Biol.* **353,** 38–52 (2005).
137. Florens, N. *et al.* Using binding competitors of albumin to promote the removal of protein-bound uremic toxins in hemodialysis: Hope or pipe dream? *Biochimie* **144,** 1–8 (2018).
138. Devine, E. *et al.* Binding affinity and capacity for the uremic toxin indoxyl sulfate. *Toxins (Basel).* **6,** 416–430 (2014).
139. Comaniciu, D. & Meer, P. Mean shift: A robust approach toward feature space analysis. *IEEE Trans. Pattern Anal. Mach. Intell.* **24,** 603–619 (2002).
140. Rousseeuw, P. J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. **20,** 53–65 (1987).
141. Watanabe, H. *et al.* Interaction between two sulfate-conjugated uremic toxins, p-cresyl sulfate and indoxyl sulfate, during binding with human serum albumin. *Drug Met. Disp.* **40,** 1423–1428 (2012).
142. Li, J. *et al.* Removal of indoxyl sulfate by water-soluble poly-cyclodextrins in dialysis. *Colloids Surfaces B Biointerfaces* **164,** 406–413 (2018).
143. Chodera, J. D. & Noé, F. Markov state models of biomolecular conformational dynamics. *Curr. Opin. Struct. Biol.* **25,** 135–144 (2014).
144. Tiwary, P. & Berne, B. J. Caliber based spectral gap optimization of order parameters (SGOOP) for sampling complex molecular systems. *Proc. Natl. Acad. Sci.* **113(11),** 2839-2844 (2015).
145. Tiwary, P. *et al.* Kinetics of protein–ligand unbinding: Predicting pathways, rates, and rate-limiting steps. *Proc. Natl. Acad. Sci.* **112,** 201424461 (2015).
146. Casasnovas, R. *et al.* Unbinding kinetics of a p38 MAP kinase type II inhibitor from metadynamics simulations. *J. Am. Chem. Soc.* **139,** 4780–4788 (2017).
147. Nu, F. *et al.* Variational approach to molecular kinetics. *JCTC* **10,** 1739-1752 (2014).
148. Mcgibbon, R. T. & Pande, V. S. Variational cross-validation of slow dynamical modes in molecular kinetics. *J. Chem. Phys.* ***142,*** 124105 (2016).
149. Harrigan, M. P. *et al.* MSMBuilder : Statistical models for biomolecular dynamics. *Biophys. J.* **112,** 10–15 (2017).
150. Deuflhard, P. & Weber, M. Robust Perron cluster analysis in conformation dynamics. *ZIB-Report 03-19* (2003).
151. Deltombe, O. *et al.* Exploring binding characteristics and the related competition of different protein-bound uremic toxins. *Biochimie* **139,** 20–26 (2017).
152. Pollard, T. D. A guide to simple and informative binding assays. *Mol. Biol. Cell* **21,** 4061–4067 (2010).
153. Tiwary, P. & Berne, B. J. How wet should be the reaction coordinate for ligand unbinding? *J. Chem. Phys.* ***144,*** 054113 (2016).
154. Tiwary, P. Molecular determinants and bottlenecks in the dissociation dynamics of biotin-streptavidin. *J. Phys. Chem. B* **121,** 10841–10849 (2017).
155. Smith, Z. *et al.* Multi-dimensional spectral gap optimization of order parameters (SGOOP) through conditional probability factorization. *J. Chem. Phys.* **149,** 234105 (2018).
156. Pramanik, D. *et al.* Can one trust kinetic and thermodynamic observables from biased metadynamics simulations: detailed quantitative benchmarks on millimolar drug fragment dissociation. *bioRxiv* 558601 (2019).
157. Salvalaglio, M. *et al.* Assessing the reliability of the dynamics reconstructed from metadynamics. *JCTC* **10,** 1420-1425 (2014).

158. Moschen, T. *et al.* Measurement of ligand-target residence times by 1H relaxation dispersion NMR spectroscopy. *J. Med. Chem.* **59,** 10788-10793 (2016).
159. Chen, S. *et al.* Strong resistance of phosphorylcholine self-assembled monolayers to protein adsorption: Insights into nonfouling properties of zwitterionic materials. *J. Am. Chem. Soc.* **127,** 14473–14478 (2005).
160. Chen, S. *et al.* Surface hydration: Principles and applications toward low-fouling/nonfouling biomaterials. *Polymer (Guildf).* **51,** 5283–5293 (2010).
161. Shao, Q. *et al.* Difference in hydration between carboxybetaine and sulfobetaine. *J. Phys. Chem. B* **114,** 16625–16631 (2010).

Appendix I - Supplemental materials for Chapter 3

Figures

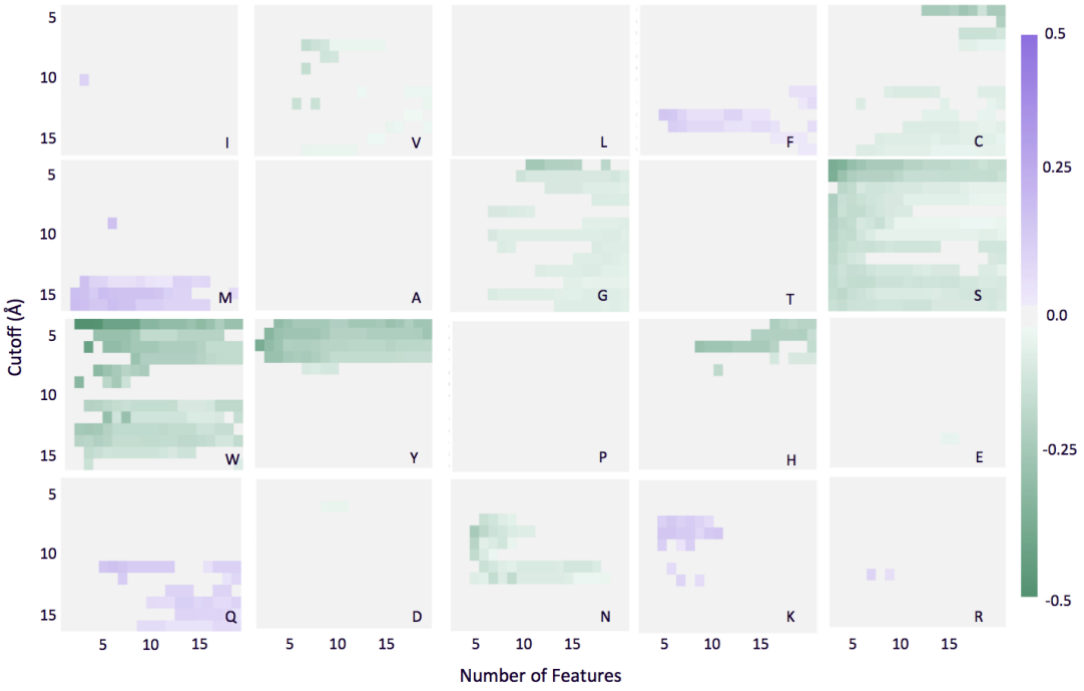

Figure I.1 – Regression coefficients plot for models trained on enzyme-containing complexes (69). Regression coefficients (w_i) for each amino acid feature for models trained on the subset with distance cutoffs from 4-16 Å and 1-19 features. Purple and green are indicative of positive and negative correlation with ΔGexp, respectively. Tan squares indicate where features have been dropped from the model ($w_i = 0$).

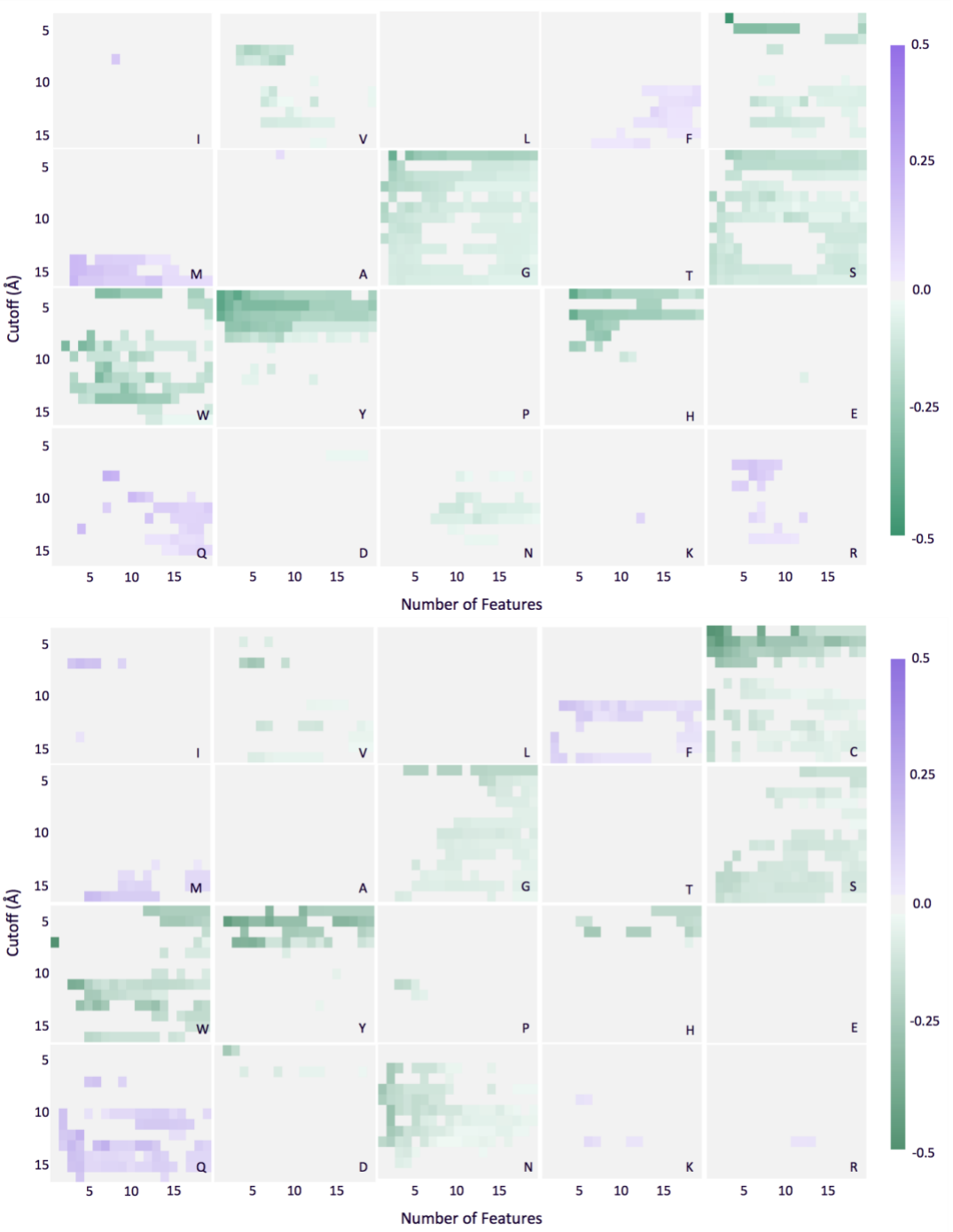

Figure I.2 – Regression coefficients plots for models trained on (top) rigid complexes (92) and (bottom) flexible complexes (87). Regression coefficients (w_i) for each amino acid feature for models trained on the subset with distance cutoffs from 4-16 Å and 1-19 features. Purple and green are indicative of positive and negative correlation with ΔGexp, respectively. Tan squares indicate where features have been dropped from the model ($w_i = 0$).

Figure I.3 – Pairplot of 5 most predictive amino acid features and experimental binding affinity (G_{exp}). Diagonal plots represent the univariate distribution of observations for each feature in the full SAB, off-diagonal plots individual observations of each feature with respect to the others. This plot demonstrates the lack of strong collinearity between important features and approximately normal distribution of binding affinity observations (bottom right corner). Points are colored by 'Class', (green) enzyme-containing (69), (red) antibody-antigen (33), and (blue) other (77).

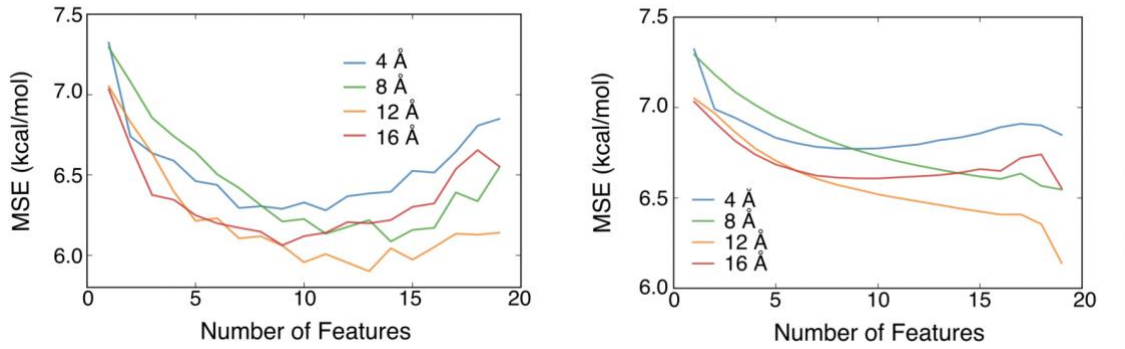

Figure I.4 – Mean squared error (MSE) for 5-fold cross validation of amino acid composition models trained with ordinary least squares regression. (left) Single lowest cross-validation MSE from any model trained with given feature number and distance cutoff. Additional features add complexity but do not improve prediction after 10 features. (right) Averaged for top 5% of models for each number of features to account for many more combinations in the middle (20 combinations of 1 amino acid; 184,756 combinations of 10 amino acid). The AAC features appearing most frequently in top 5% models are ranked below in Tables I.1-I.4.

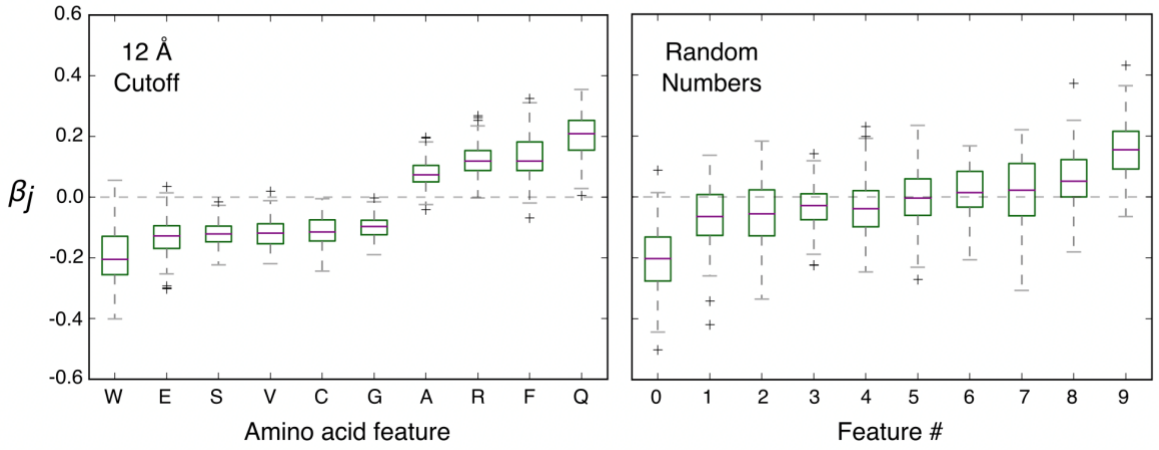

Figure I.5 – Linear regression coefficients for selected obtained from 1000 bootstrapped samples. (left) Top 10 AAC features for binding affinity prediction based on cross-validation with 12 Å distance cutoff. (right) Ten randomly generated features from the normal distribution for equally sized sample (179 examples). The interquartile range for every AAC feature trained at a 12 Å distance cutoff is either completely above or below zero, supporting the qualitative impact of each feature predicted by lasso. Eight of ten random features contain 0 in the interquartile range. Standard errors of the regression coefficients are available for 12 Å cutoff in Table I.3.

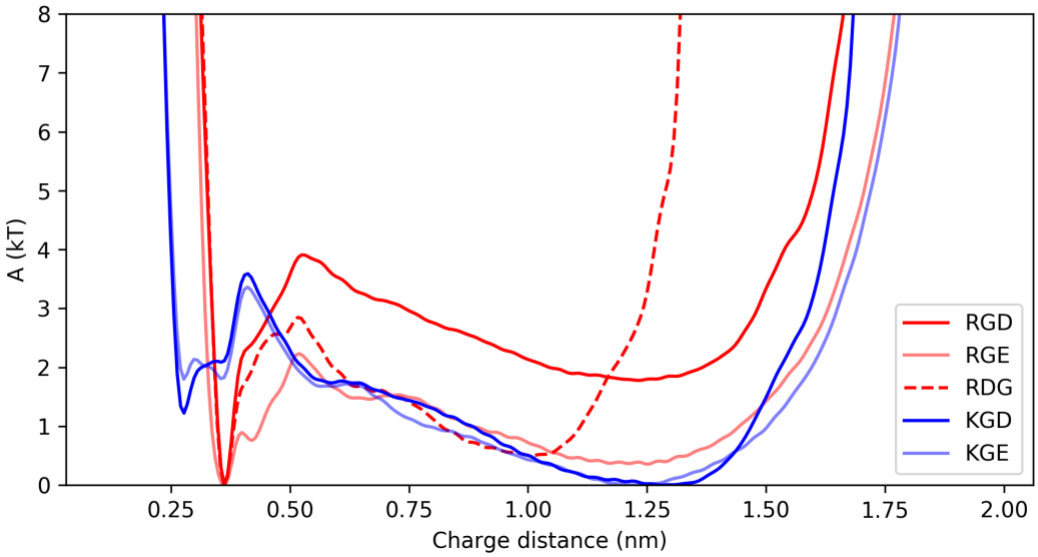

Figure I.6 – One-dimensional free energy profiles for Pos-Gly-Neg charged group distances. Helmholtz free energy (A) reported in terms of kT (T = 300 k) for straightforward comparison to energy fluctuations in the system. RGD (dark red), RGE (light red), and RDG (dashed) prefer salt-bridged conformations (minimum below 0.5 nm), but only the RGD salt-bridge is significantly stable relative to free energy fluctuations in the system (2 kT). The free energy minimum for KGD (dark blue) and for KGE (light blue) correspond to conformations with individually solvated charges.

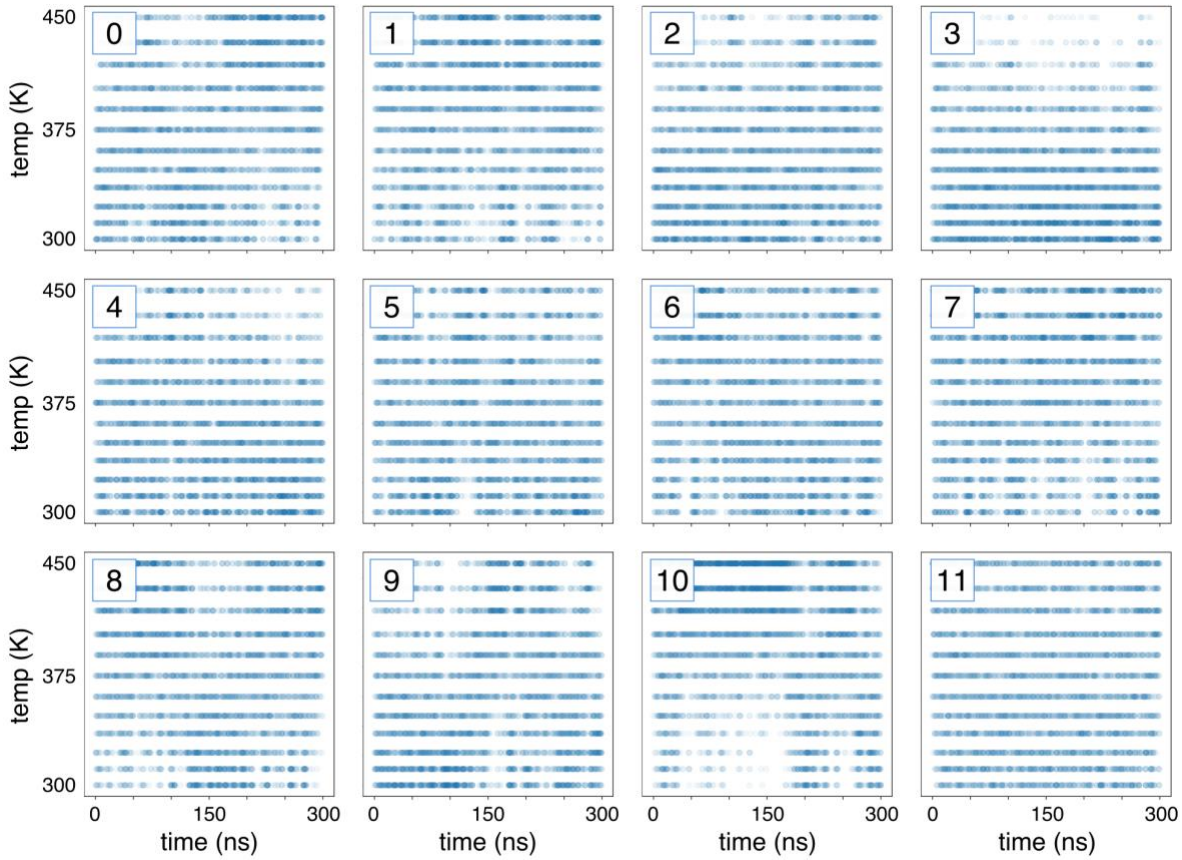

Figure I.7 – Temperature level vs. time for each replica during the PTMetaD simulation for $(EK)_{15}$. These plots show that each replica of the system visited each temperature level during the PTMetaD simulation.

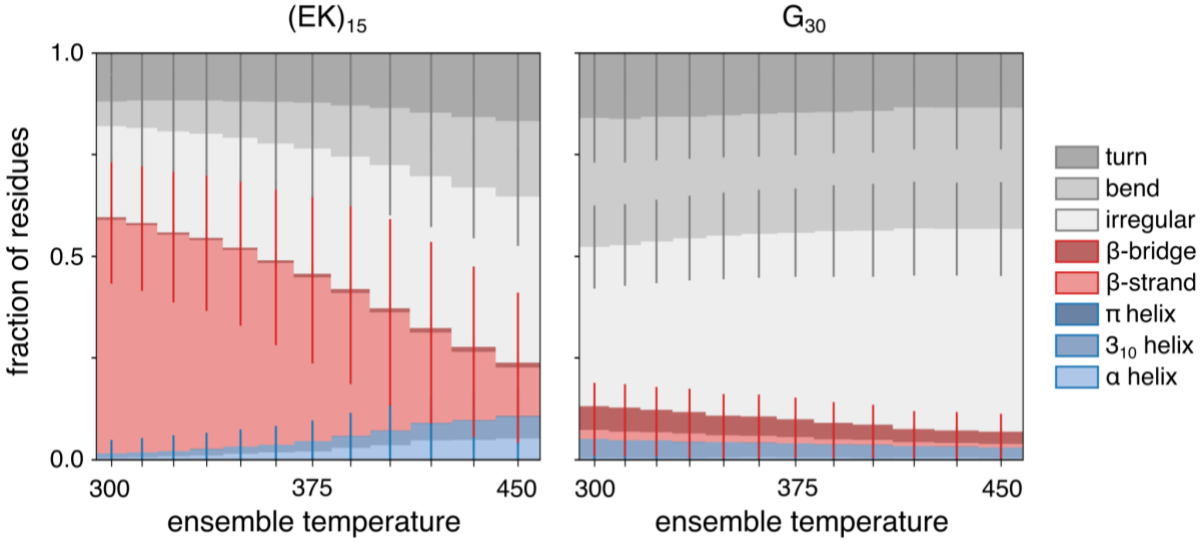

Figure I.8 – Detailed breakdown of secondary structure fractions for $(EK)_{15}$ and G_{30} conformational ensemble at each temperature. Stacked bars are included for the structural ensemble sampled at each of the 12 temperatures during the PTMetaD simulation. Ensemble averaged structure fractions are based on the 8-class structure list from DSSP, as designated in the legend. Error bars represent +/- 1 standard deviation.

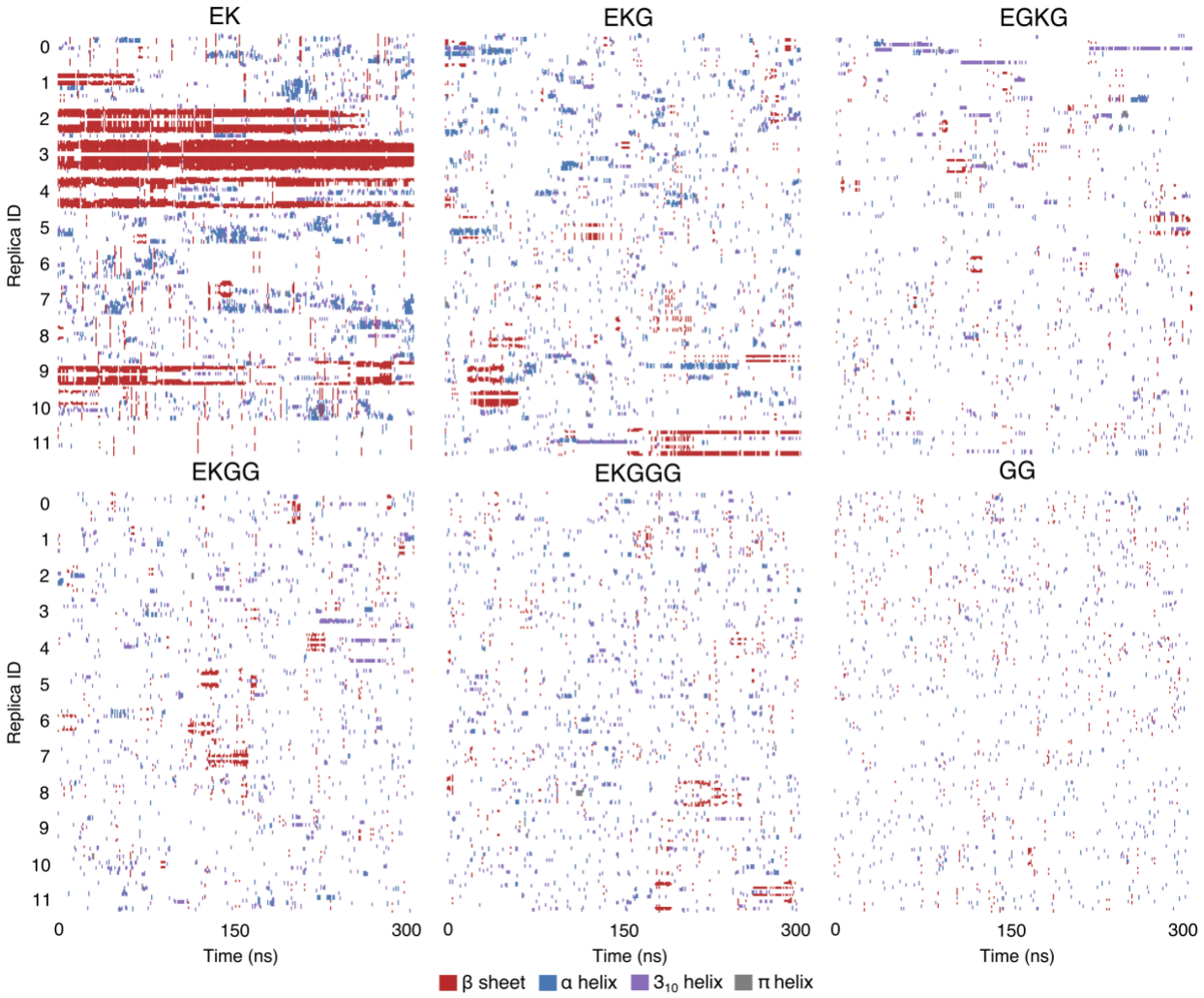

Figure I.9 – Secondary structure for demultiplexed replica trajectories over the course of the PTMetaD simulation for each sequence. Colored areas correspond to ordered secondary structures: beta sheet (red), alpha helix (blue), 3_{10} helix (purple), and pi helix (gray). White areas correspond to disordered structures: turn and random coil. The tendency of $(EK)_{15}$ to form stable secondary structure is eliminated by adding G to the repeat sequence.

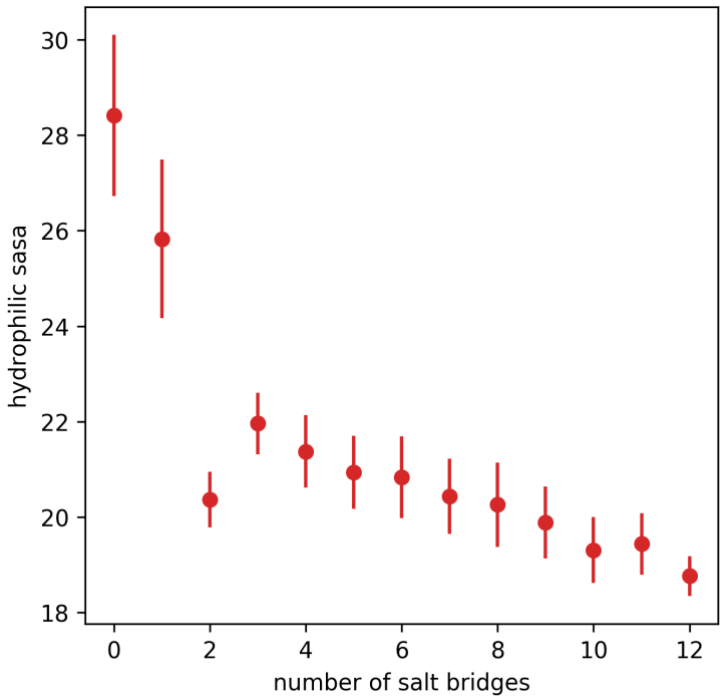

Figure I.10 – Number of salt bridges vs. hydrophilic surface area for $(EK)_{15}$ at 300 K. Sub-ensembles were created from the conformational ensemble, conditional on the number of salt bridges. Weighted average and standard deviation from each of these conditional sub-ensembles are pictured here.

Tables

Table I.1 – Feature ranking and regression coefficients for validation techniques and lasso at 4 Å.

4 Å	Feature Rank		β_j	
	lasso	CV	lasso	Bootstrap (+/- SE)
TYR	1	1	-	-0.35
GLY	2	2	-	-0.24
HIS	3	3	-	-0.36
SER	4	4	-	-0.18
CYS	6	5	-	-0.28
TRP	5	6	-	-0.32
ALA	16	7	+	0.21
ARG	14	8	+	0.07
VAL	13	9	-	-0.14
ILE	20	10	+	0.13

Table I.2 – Feature ranking and regression coefficients for validation techniques and lasso at 8 Å.

8 Å	Prediction Rank		β_j	
	lasso	CV	lasso	Bootstrap OLS
SER	1	1	-	-0.14
VAL	6	2	-	-0.21
GLN	3	3	+	0.23
LYS	11	4	+	0.15
TYR	2	5	-	-0.12
HIS	5	6	-	-0.27
GLY	8	7	-	-0.07
ASN	7	8	-	-0.10
ARG	17	9	+	0.14
TRP	4	10	-	-0.25

Table I.3 – Feature ranking and regression coefficients for validation techniques and lasso at 12 Å.

12 Å	Prediction Rank		β_j	
	lasso	CV	lasso	Bootstrap
SER	1	1	-	-0.14
GLN	2	2	+	0.18
GLY	4	3	-	-0.08
CYS	5	4	-	-0.10
TRP	3	5	-	-0.26
VAL	7	6	-	-0.11
PHE	6	7	+	0.14
GLU	12	8	-	-0.09
ALA	17	9	+	0.07
ARG	14	10	+	0.09

Table I.4 – Feature ranking and regression coefficients for validation techniques and lasso at 16 Å.

16 Å	Prediction Rank		β_j	
	lasso	CV	lasso	Bootstrap
SER	1	1	-	-0.10
GLY	3	2	-	-0.06
GLN	4	3	+	0.12
MET	2	4	+	0.17
TRP	5	5	-	-0.18
CYS	6	6	-	-0.07
VAL	8	7	-	-0.06
ARG	20	8	+	0.07
PHE	7	9	+	0.05
GLU	19	10	-	-0.06

Table I.5 – Equilibration and production simulations for each tripeptide.

Phase	Step	RGD	RGE	KGD	KGE
Equil.	em	10000 steps	10000 steps	10000 steps	10000 steps
	Anneal	250 ps	250 ps	250 ps	250 ps
	NPT (Berendsen)	500 ps	500 ps	500 ps	500 ps
Prod.	Steered MD	800 ps	800 ps	800 ps	800 ps
	PBMetaD	500 ns x 8	500 ns x 8	500 ns x 8	500 ns x 8

Table I.6 – Equilibration and production simulation time (per replica) for enhanced sampling MD.

Phase	Step	EK	EKG	EGKG	EKGG	EKGGG	GG
Equil.	em	10000 steps	10000 steps	10000 steps	10000 steps	10000 steps	10000 steps
	NPT (Berendsen)	250 ps	250 ps	250 ps	250 ps	250 ps	250 ps
	Anneal	250 ps	250 ps	250 ps	250 ps	250 ps	250 ps
Prod.	PT-WTE	350 ns	100 ns	100 ns	100 ns	100 ns	350 ns
	PT-MetaD	300 ns	300 ns	300 ns	300 ns	300 ns	300 ns

Table I.7 – Equilibration and production simulation time for unbiased MD.

Phase	Step	EK (300 K)	EK (450 K)	DK	EKGG
Equil.	em	10000 steps	10000 steps	10000 steps	10000 steps
	Anneal	250 ps	250 ps	250 ps	250 ps
	NPT (Berendsen)	250 ps	250 ps	250 ps	250 ps
Prod.	NPT (Parrinello-Rahman)	1 μs	1 μs	1 μs	500 ns

Table I.8 – Secondary structure fraction, normalized backbone entropy, and salt bridges for EK_{15} and G_{30}.

	$(EK)_{15}$			G_{30}
	300 K	374.26 K	450 K	300 K
Coil	0.41 +/- 0.14	0.54 +/- 0.18	0.76 +/- 0.18	**0.87 +/- 0.10**
Extended	**0.58 +/- 0.15**	0.41 +/- 0.20	0.13 +/- 0.18	0.08 +/- 0.08
Helix	0.01 +/- 0.04	0.05 +/- 0.08	**0.11 +/- 0.12**	0.05 +/- 0.07
Conf. entropy	0.37 +/- 0.17	0.62 +/- 0.09	0.83 +/- 0.03	**1.0 +/- 0.03**
Salt bridges	**7.1 +/- 1.9**	6.8 +/- 2.1	6.0 +/- 2.2	N/A
Hydrophilic SASA Fraction	**0.55 +/- 0.02**	0.54 +/- 0.02	0.53 +/- 0.02	0.46 +/- 0.02

***bold values correspond to the highest mean in each column**

Appendix II - Supplemental material for Chapter 4

Figures

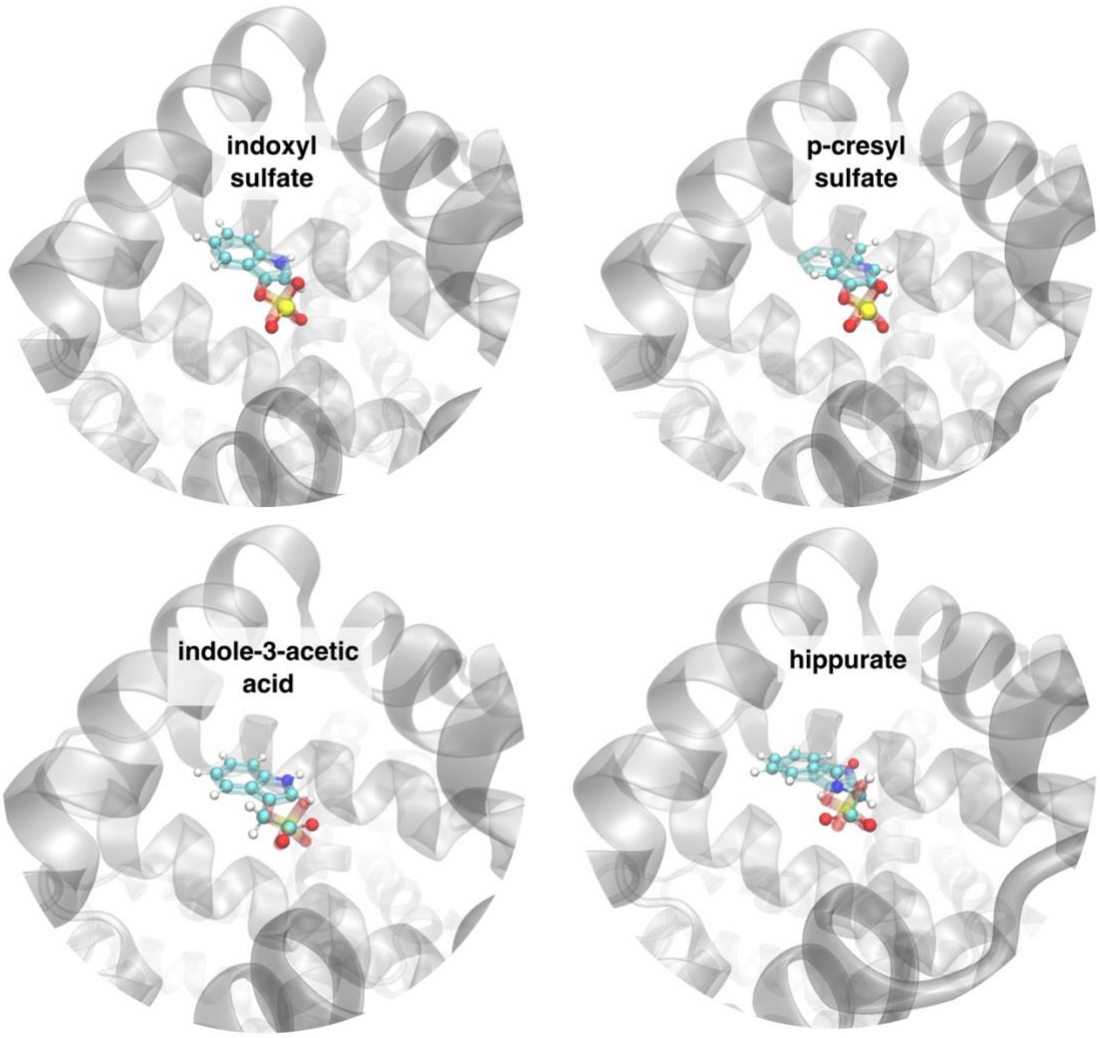

Figure II.11 – Binding poses used to initialize (top left) indoxyl sulfate-HSA, (top right) p-cresyl sulfate-HSA, (bottom left) indole-3-acetic acid-HSA, and (bottom right) hippurate-HSA simulations. The initial protein structure (transparent gray) is taken in all 4 cases from the experimental x-ray structure for IS bound to Sudlow site II (PDB ID: 2BXH), and each toxin molecule (opaque ball and stick) was superimposed in place of the experimentally resolved IS molecule (transparent, colored by atom).

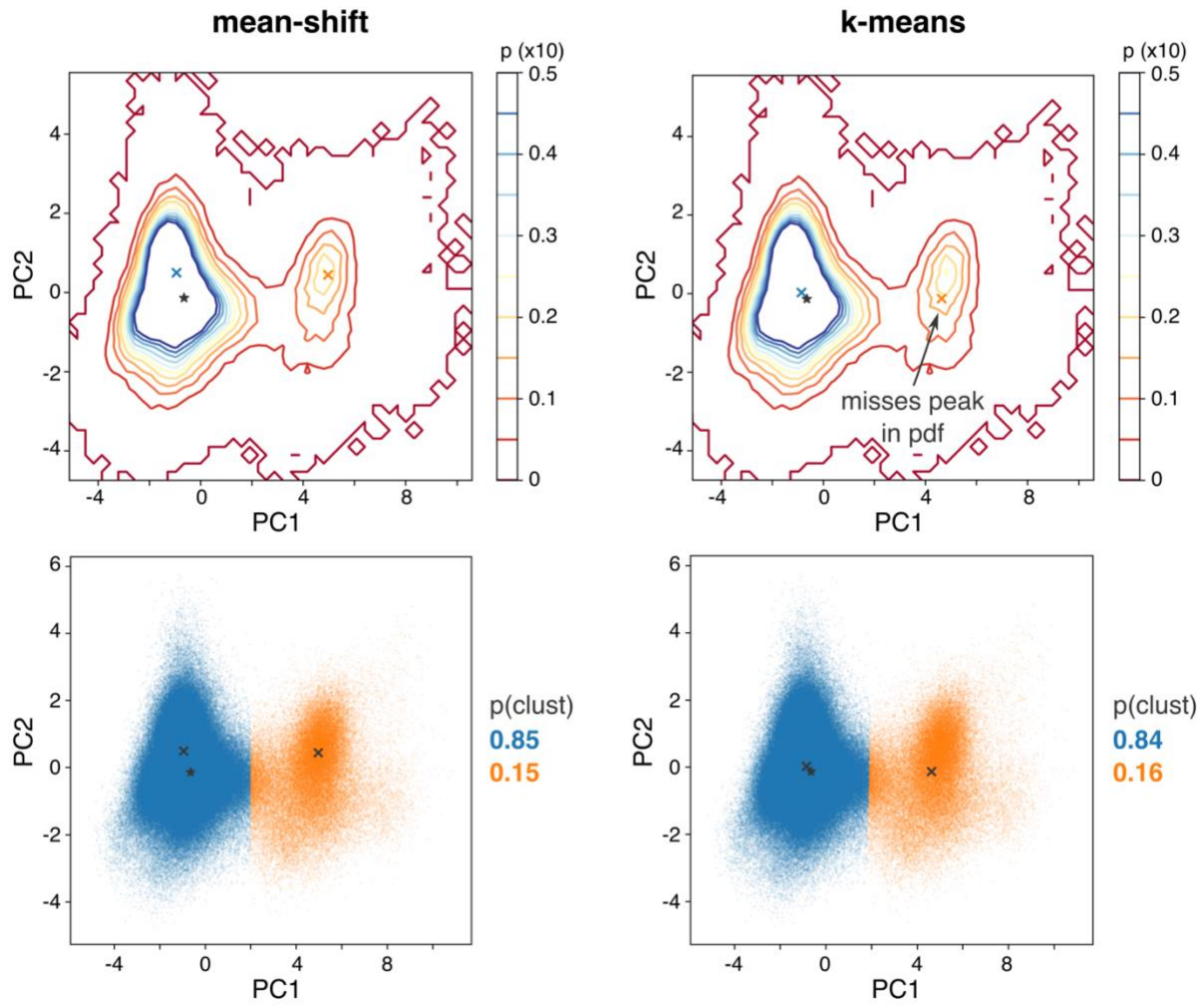

Figure II.12 – Comparison of conformation clustering results for IS-HSA complex from the mean-shift and k-means algorithms. (left) Because mean-shift is a "mode-seeking algorithm", cluster centers converge to peaks on the underlying probability density function. (right) The k-means algorithm tends to converge on clusters with similar spatial extent (here in 2D principal component space) without consideration of the underlying probability density. The centers of the k-means clusters are less representative of metastable binding modes. Ultimately, the dividing line between the two clusters for the IS-HSA complex is insensitive to the clustering method.

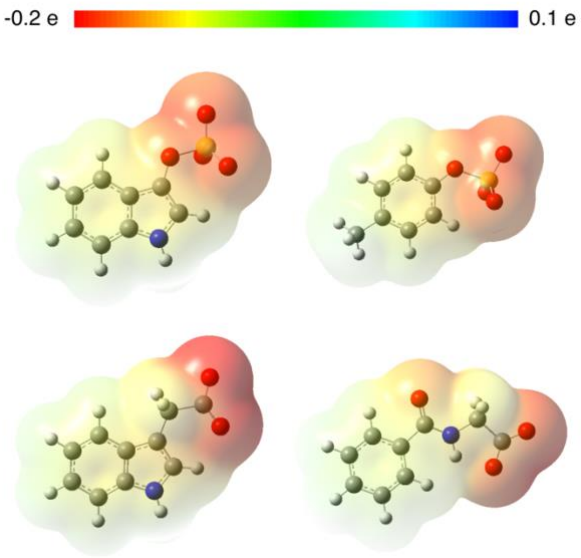

Figure II.13 – Electrostatic potential surface visualization for each PBUT, based on point charges from the RESP method. Darker red represents regions of more negative electrostatic charge, while green and dark blue represent neutral and positive charge, respectively. The electrostatic potential surface here is mapped onto an electron density isosurface at 0.0004 (number density).

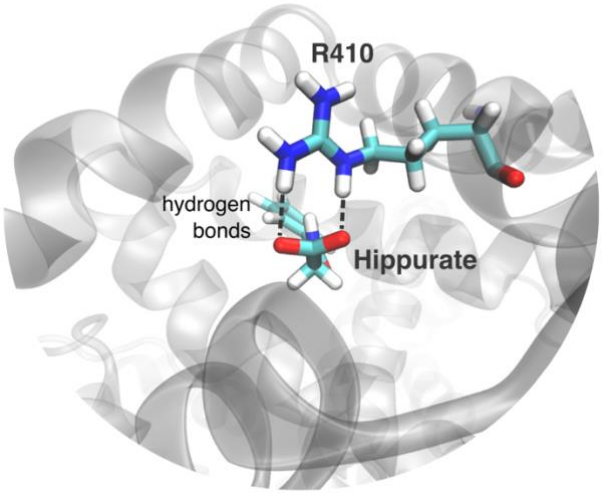

Figure II.14 – Representative image of carboxylate-R410 double hydrogen bonded conformation for HA. This double hydrogen bonded geometry was also observed for IAA-R410 interactions.

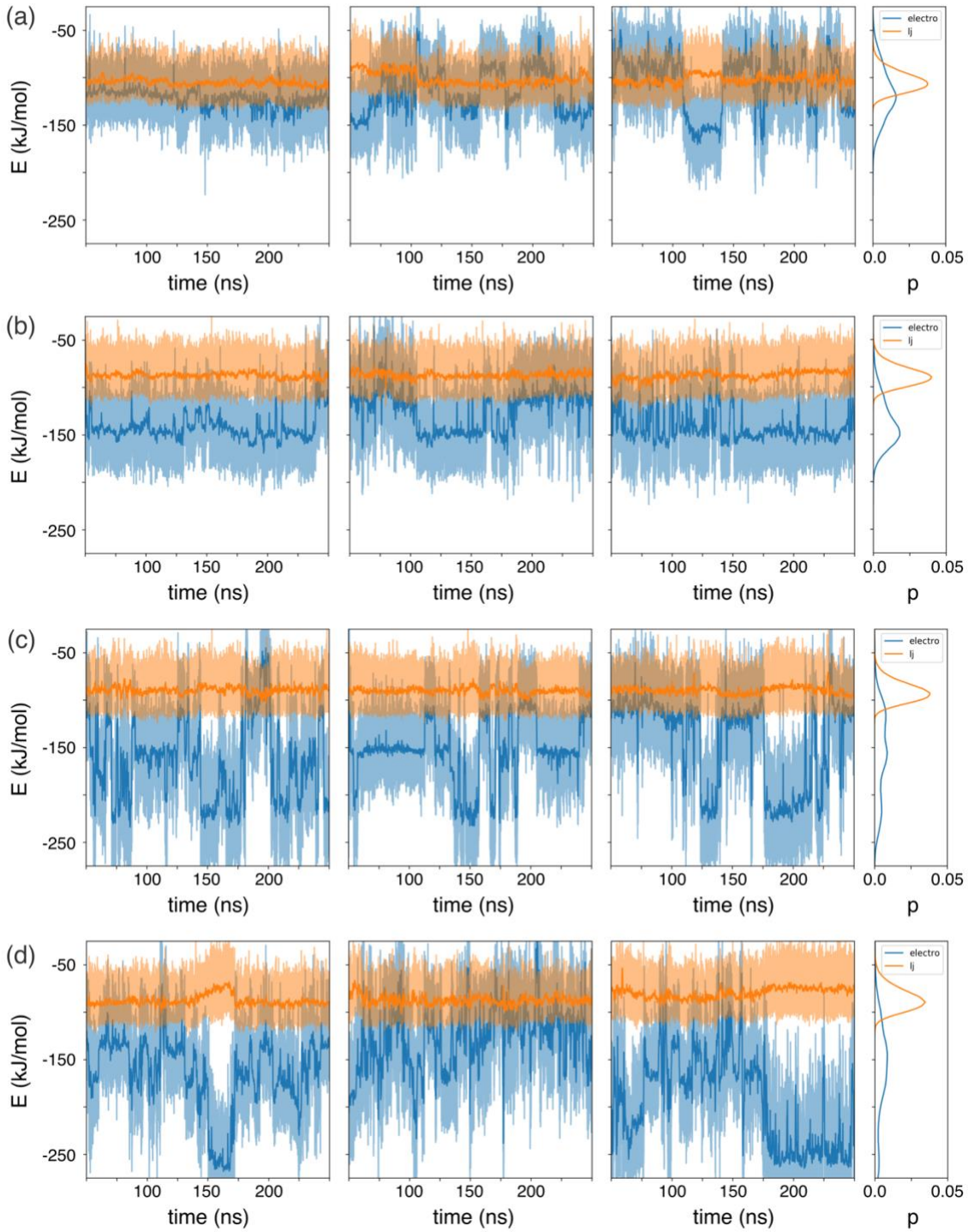

Figure II.15 – The protein-toxin interaction energy calculated from MD simulations of HSA in complex with (top) indoxyl sulfate and (bottom) p-cresyl sulfate. The light blue and light orange shaded areas represent the Coulomb and Lennard-Jones energies, respectively, calculated every 2 ps of simulation time. The dark blue and dark orange lines represent the average Coulomb and Lennard-Jones energies, respectively, over the previous 0.2 ns (i.e. the mean of the last 100 observations at 2 ps intervals). The rightmost panel of each row shows the probability distribution function of each energy for all three simulations of the given toxin based on simple histogramming.

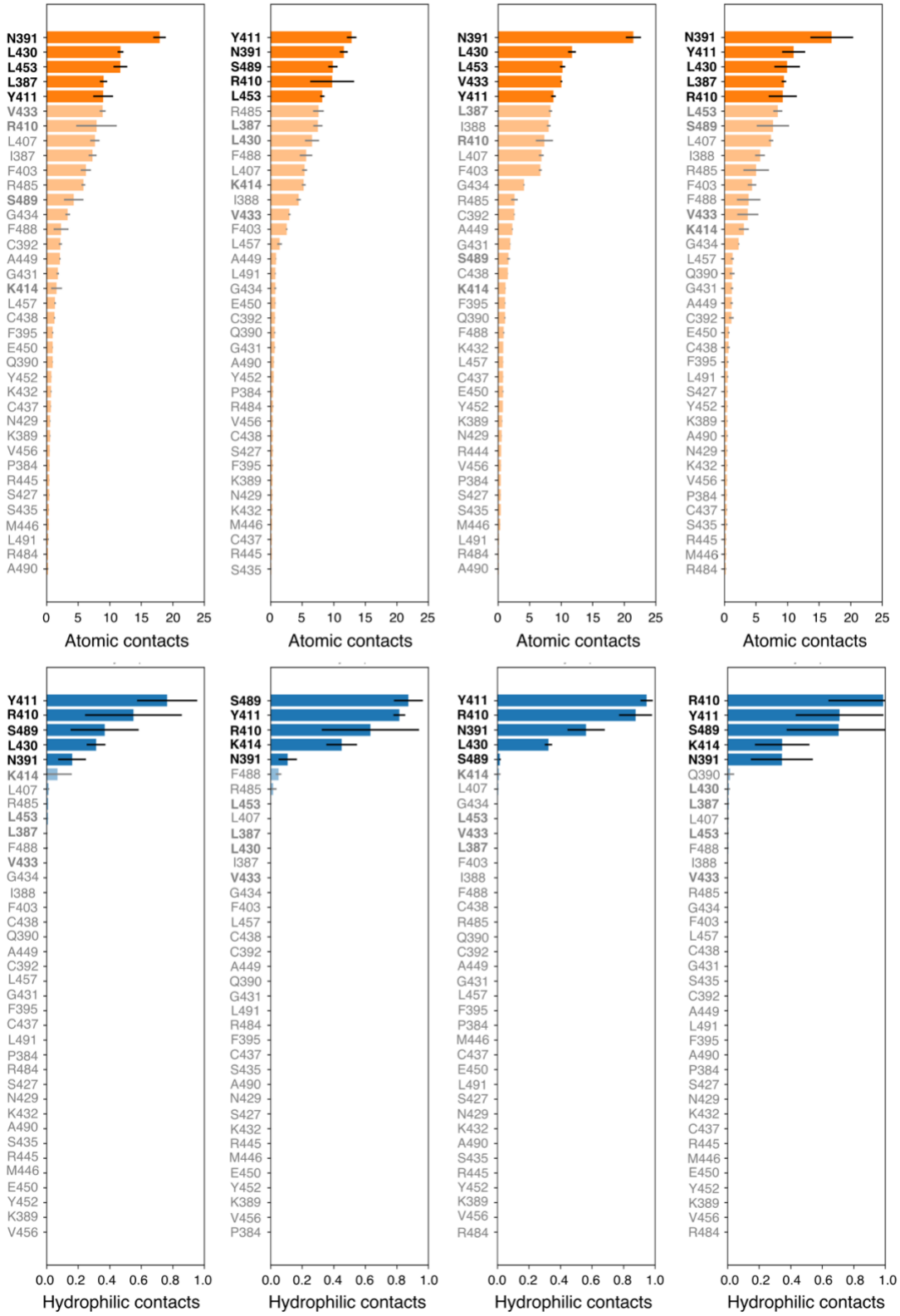

Figure II.16 – Average number of atomic and hydrophilic contacts observed in the MD simulations for the 37 residues within 6 Å of IS in the X-ray structure (pdb: 2BXH). The top 5 residues in each bar plot were added to the set of key residues, which resulted in 9 residues to be considered for further analysis.

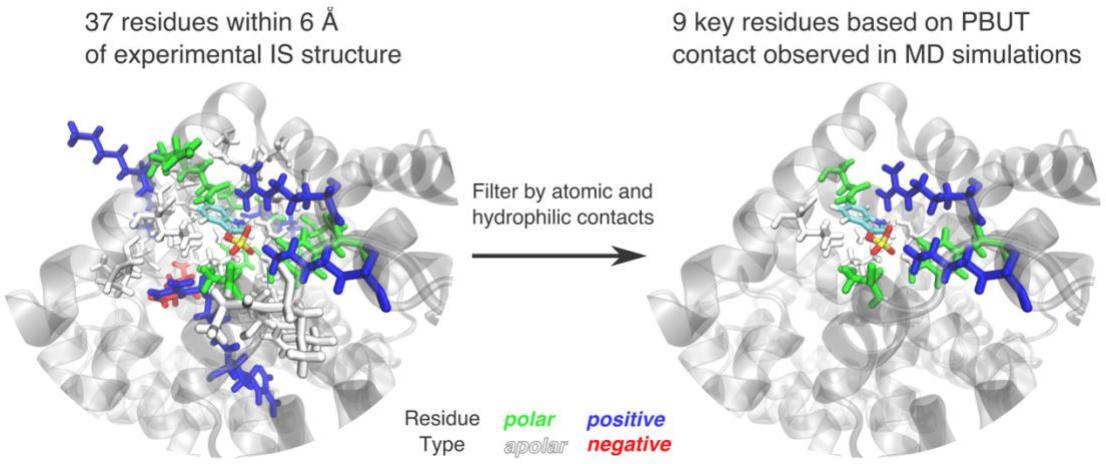

Figure II.17 – IS-HSA protein data bank structure (2BXH) with highlighted (left) the 37 residues considered before atomic contact filtering and (right) the 9 residues remaining after filtering by atomic contacts. Polar, apolar, positive, and negative amino acids are colored green, white, blue, and red, respectively. Indoxyl sulfate is colored by atom type with hydrogen, carbon, oxygen, and sulfur atoms colored white, cyan, red, and yellow, respectively.

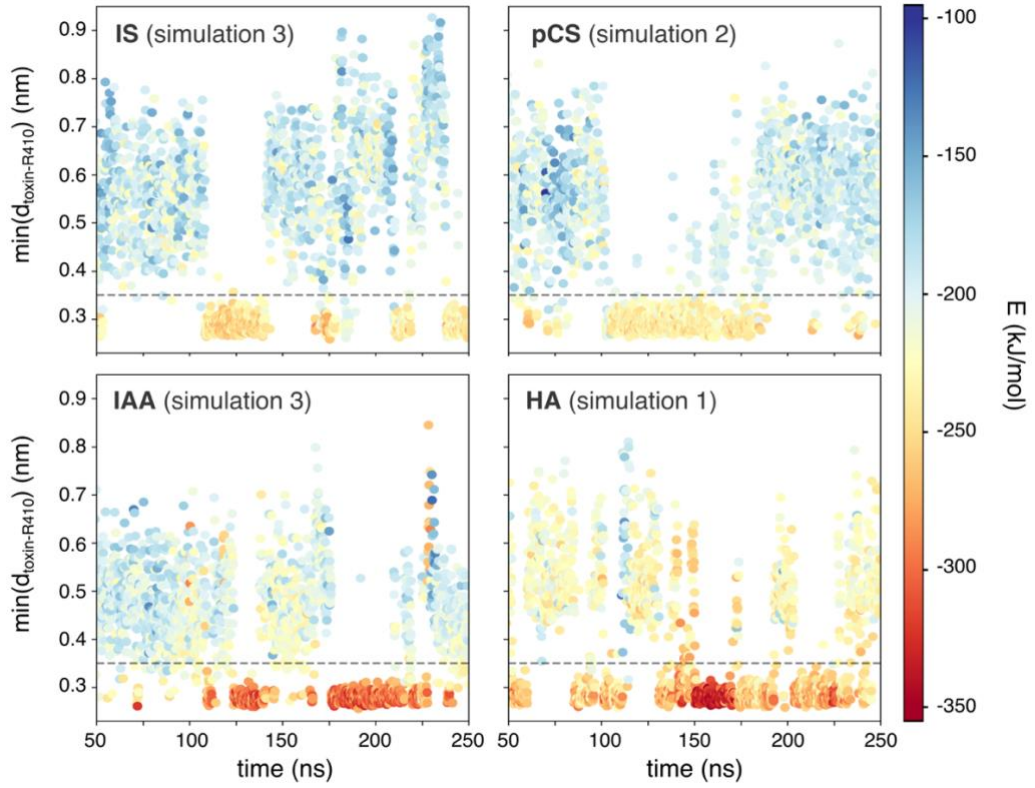

Figure II.18 – The minimum distance between each PBUT and R410 heavy atoms vs. simulation time, and it's relation to the overall protein-toxin interaction energy. Each point is colored based on the overall protein-toxin interaction energy, ranging from dark red (lower energy) to dark blue (higher energy). The dashed horizontal line at 0.35 nm represents a putative cutoff for toxin-R410 hydrogen bonds. The simulation that best captures the transient nature of the R410 complex for each PBUT is pictured, although the breaking/forming of this salt bridge was observed at least once in each production simulation.

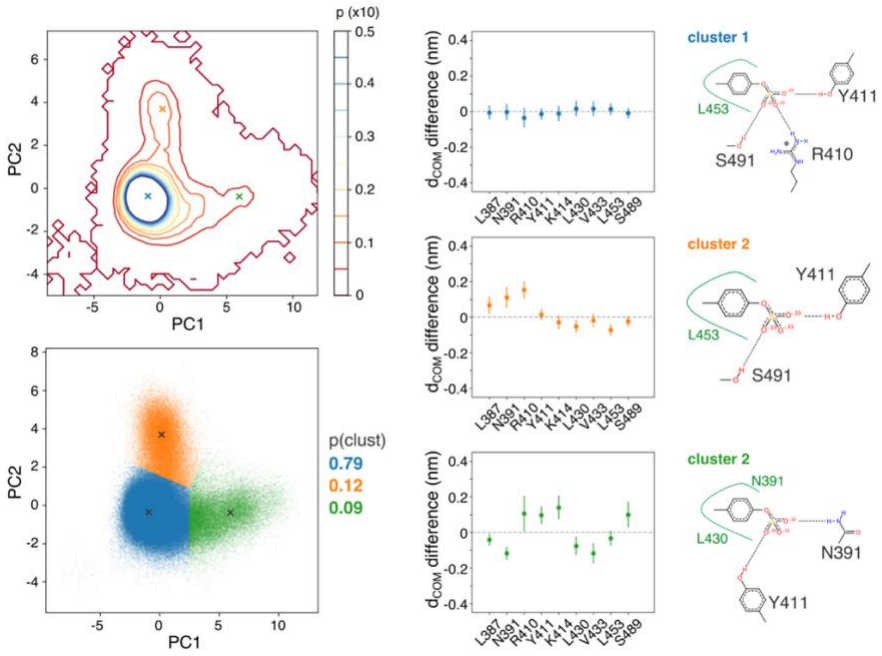

Figure II.19 – PCA and clustering results for the pCS-HSA complex. (top left) Contour plot of the 2D pdf for the pCS-HSA complex, created from a 2D histogram of the PC values for each frame of the MD simulations, with a probability step of 0.005 between contour lines. Blue, orange, and green 'x' symbols mark the center of each mode identified with mean-shift. **(bottom left)** Scatter plot of the first 2 PC values for each MD frame, colored by the cluster assigned by mean-shift. **(center column)** The difference between the average pCS-residue center of mass distance for all points in each cluster and the overall average pCS-residue center of mass distance (all points in all clusters). **(right column)** A PoseView representation of the central structure of each cluster (the MD frame with PC values nearest the 'x' symbols).

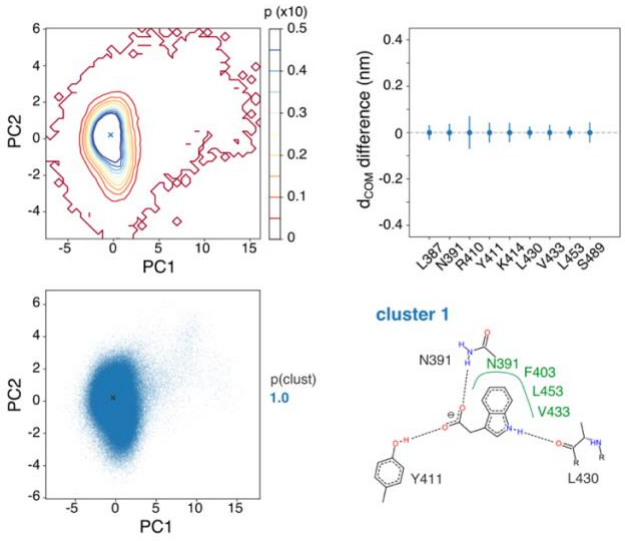

Figure II.20 – PCA and clustering results for the IAA-HSA complex. (top left) Contour plot of the 2D pdf for the IAA-HSA complex, created from a 2D histogram of the PC values for each frame of the MD simulations, with a probability step of 0.005 between contour lines. Blue 'x' symbol marks the center of the lone binding mode identified with mean-shift. **(bottom left)** Scatter plot of the first 2 PC values for each MD frame. **(top right)** The difference between the average IAA-residue center of mass distance for all points in the cluster and the overall average IAA-residue center of mass distance (all points in all clusters). **(bottom right)** A PoseView representation of the central structure of the lone cluster (the MD frame with PC values nearest the 'x' symbol).

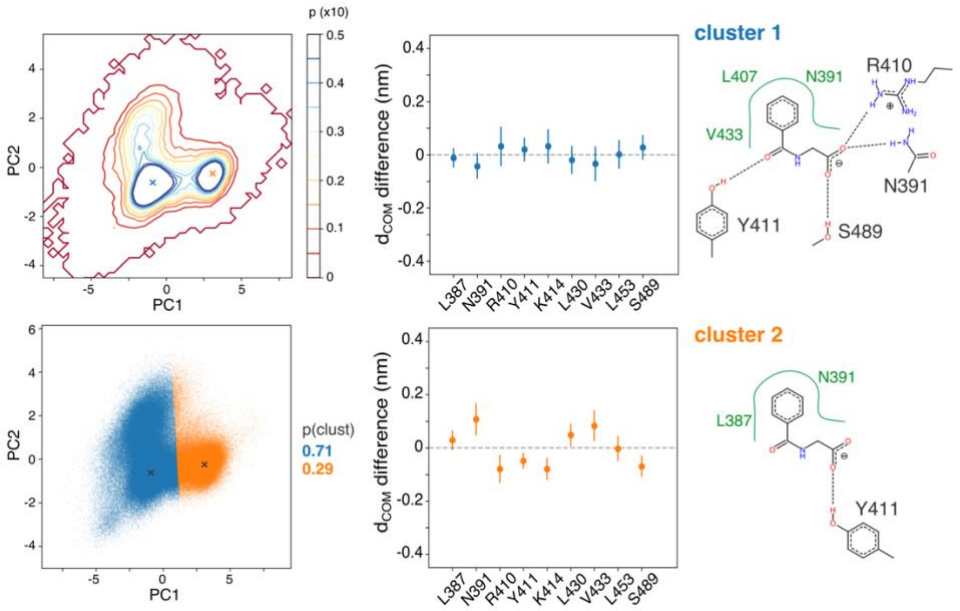

Figure II.21 – PCA and clustering results for the hippurate-HSA complex. (top left) Contour plot of the 2D pdf for the HA-HSA complex, created from a 2D histogram of the PC values for each frame of the MD simulations, with a probability step of 0.005 between contour lines. Blue and orange 'x' symbols mark the center of each mode identified with mean-shift. **(bottom left)** Scatter plot of the first 2 PC values for each MD frame, colored by the cluster assigned by mean-shift. **(center column)** The difference between the average HA-residue center of mass distance for all points in each cluster and the overall average HA-residue center of mass distance (all points in all clusters). **(right column)** A PoseView representation of the central structure of each cluster (the MD frame with PC values nearest the 'x' symbols in **(a-b)**.

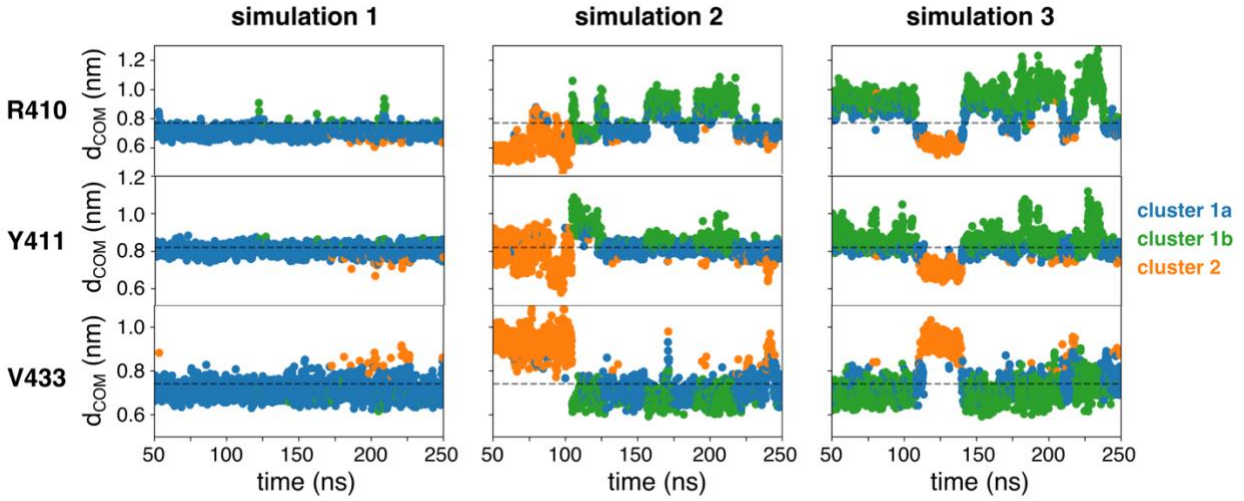

Figure II.22 – Three residue-IS distances vs. simulation time, and their relation to the structural cluster id assigned by k-means clustering (k = 3). For each of the three MD simulations of the IS-HSA complex, the (top) IS-R406, (middle) IS-Y407, and (bottom) IS-V429 center of mass distances are plotted at 0.1 ns intervals between 50 and 250 ns. Colors are assigned according to the cluster id assigned by k-means, named (blue) cluster 1a, (green) cluster 1b, and (orange) cluster 2. The horizontal dashed line in each row of plots represents the overall mean of the given residue-toxin distance.

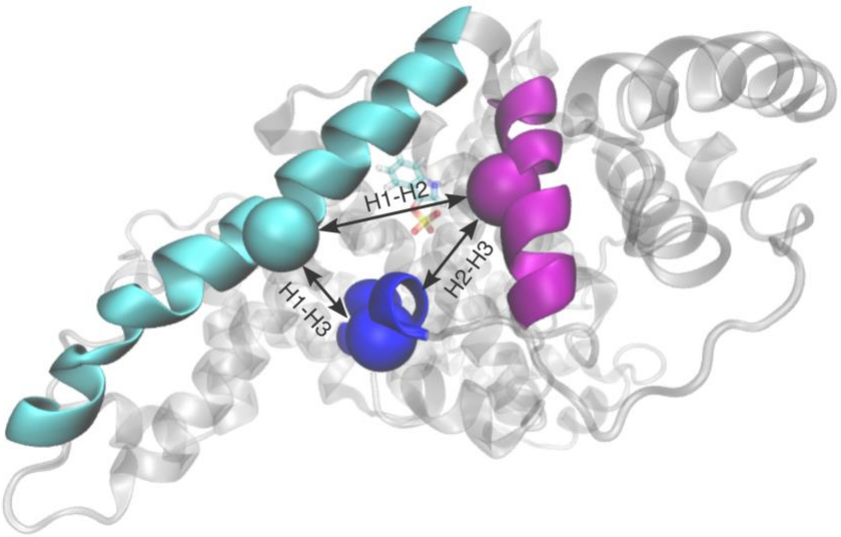

Figure II.23 – Protein-protein distances appended to the list of order parameters for the IS-HSA complex. The mouth of Sudlow site II is comprised of 3 helical domains, highlighted in cyan, magenta, and blue. The center of mass of the alpha carbon atoms in each helix, represented by spheres, were used to calculate the three intramolecular distances.

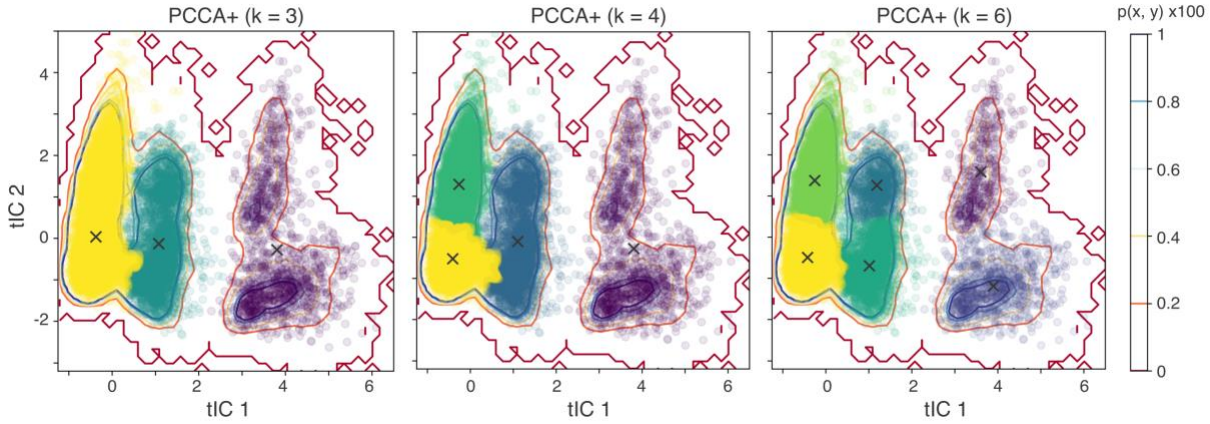

Figure II.24 – Macrostates identified with PCCA+ at various resolutions. PCCA+ with 3 states identifies the major binding modes, separated along tIC 1. Increasing to 6 states shows that each major state can be separated along tIC 2, presumably based on the position of the R410 sidechain. Using 4 states reveals that the tIC 2 is associated with R410 position, which is not captured with 3 macrostates, and results in a more interpretable network visualization (Figure 4.2.2) than when all 6 macrostates are included.

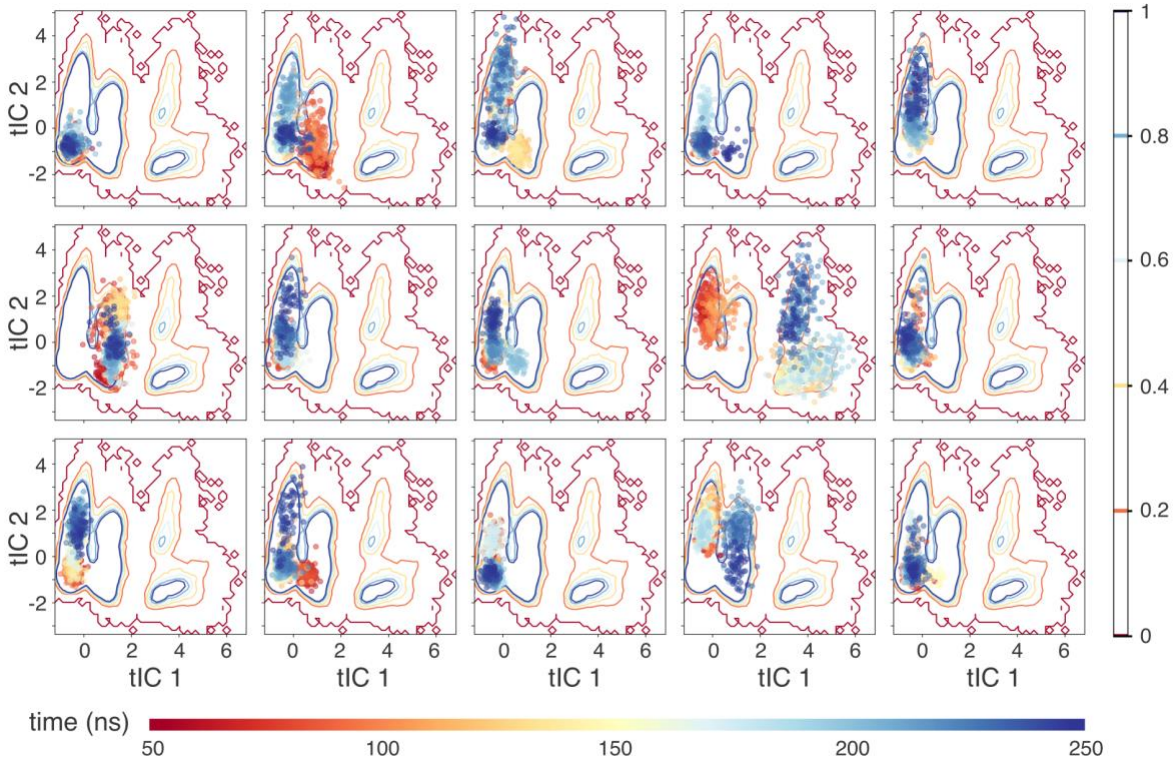

Figure II.25 – Small multiples plots of all 15 classical MD simulations projected onto the first two time-structure independent components (tIC 1 and tIC 2). Contours for the underlying probability density function in tIC space are marked at probability intervals of 0.002, ranging from dark red to dark blue for values from 0 to 0.01, respectively. Data points are colored by simulation time, with dark red points corresponding to sampling early in the trajectory and dark blue points corresponding to sampling late in the trajectory. Sampling in the unique binding pose is limited to the second half of the ninth simulation (row 2, column 4).

Tables

Table II.9 – Coefficients of the linear transformation to the first 2 principal components calculated for the IS-HSA and pCS-HSA complex.

	indoxyl sulfate			p-cresyl sulfate			indole-3-acetic acid			hippurate		
	PC1 (0.57)	**PC2 (0.15)**	PC3 (0.09)	**PC1 (0.46)**	**PC2 (0.24)**	PC3 (0.09)	**PC1 (0.28)**	**PC2 (0.17)**	PC3 (0.12)	**PC1 (0.48)**	**PC2 (0.15)**	PC3 (0.12)
L383	0.03	0.67	0.69	- 0.16	0.51	0.34	0.13	- 0.55	- 0.01	0.25	- 0.01	- 0.60
N387	0.37	0.20	0.11	- 0.29	0.47	0.05	0.41	- 0.37	- 0.06	0.41	- 0.10	- 0.25
R406	- 0.28	- 0.40	0.51	0.28	0.39	- 0.47	- 0.36	0.26	0.42	- 0.36	- 0.22	0.13
Y407	- 0.30	- 0.32	0.35	0.42	0.03	- 0.05	- 0.31	- 0.24	0.60	- 0.38	- 0.17	- 0.12
K410	- 0.40	0.11	-0.09	0.42	- 0.17	0.20	- 0.46	- 0.16	0.12	- 0.35	- 0.37	0.19
L426	0.38	- 0.02	-0.04	- 0.31	- 0.28	0.44	0.16	0.39	0.09	0.30	- 0.41	0.45
V429	0.41	- 0.05	0.05	- 0.43	- 0.06	- 0.04	0.44	0.23	0.25	0.37	- 0.04	0.41
L449	0.33	- 0.32	0.18	- 0.21	- 0.48	- 0.43	- 0.14	0.42	- 0.26	- 0.01	0.74	0.36
S485	- 0.33	0.37	-0.27	0.36	- 0.13	0.48	- 0.38	- 0.16	- 0.56	- 0.38	0.23	- 0.11

***first 2 PCs used for pdf visualizations and clustering**

Table II.10 – Silhouette scores for mean-shift clustering with various bandwidths.

	IS	pCS	IAA	HA
0.5	0.20	0.30	0.18	0.30
0.6	0.32	0.33	0.24	0.44
0.7	0.57	0.40	0.37	0.46
0.8	0.58	0.50	0.32	0.46
0.9	0.61	**0.55**	0.37	0.46
1.0	0.60	0.55	0.45	0.46
1.1	**0.64**	0.55	0.60	**0.52**
1.2	0.64	0.55	0.54	0.52
1.3	0.64	0.50	0.64	0.52
1.4	0.64	0.50	0.67	0.52
1.5	0.64	0.50	0.56	0.52
1.6	0.64	0.50	0.67	0.52
1.7	0.64	0.50	0.66	0.52
1.8	0.64	0.50	0.67	0.52
1.9	0.64	0.50	**0.69**	0.52
2.0	0.64	-	0.66	0.52

***bold values selected for clustering**

Table II.11 – Average number of hydrogen bonds for each complex.

Simulation ID	IS	pCS	IAA	HA
1	2.84	3.21	3.19	3.71
2	2.40	2.64	3.13	2.97
3	2.53	2.99	3.25	4.00
mean	2.59 (0.184)	2.95 (0.238)	3.19 (0.050)	3.56 (0.436)