COMPUTATIONAL ANALYSIS OF THERMODYNAMIC FACTORS FOR INTRINSICALLY DISORDERED LIGAND RECOGNITION

A Dissertation

by

JIE SHI

Submitted to the Graduate and Professional School of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Chair of Committee,	Wonmuk Hwang
Committee Members,	Roland Kaunas
	Alvin Yeh
	Jae-Hyun Cho
Head of Department,	Mike McShane

August 2021

Major Subject: Biomedical Engineering

ABSTRACT

The nonstructural protein 1 (NS1) of the 1918 Spanish influenza A virus hijacks N-terminal Src-homology 3 (nSH3) domain of host's cellular CrkII with exceptionally high affinity through its proline-rich motif (PRM^{NS1}). By comparison, cellular PRMs are intrinsically disordered proteins mediating protein-protein interactions only with weak binding affinities. Therefore, it's critical to elucidate the thermodynamic difference between the binding of viral PRM and cellular PRMs. For that reason, molecular dynamics (MD) simulation is used to provide insights at an atomistic level.

This dissertation aims to further the understanding of nSH3:PRM binding via: (i) studying the role of fuzzy interactions of bound PRM^{NS1}, (ii) dissecting and comparing the entropic contribution for the binding of nSH3 by PRM^{NS1} (viral) and PRM^{cAbl} (cellular, from Abl kinase), and (iii) developing a method for evaluating solvation contribution upon complexation.

For aim (i), MD simulations of nSH3:PRM^{NS1} are compared with crystal structures and fuzzy interactions are shown to enhance the long-range electrostatic interactions by reducing their average pairwise distances.

For aim (ii), the associated conformational entropy are calculated with binding of two PRMs to the nSH3 domain. Different side chain/backbone contributions are observed, with implications for structure-based entropic contribution. At residue level, entropy "hotspots" are identified, some of which locate distal to the binding interface, indicating an allosteric role of ligand for downstream regulation. As a result of forming more extensive contacts, nSH3:PRM^{NS1} shows greater entropy loss than nSH3:PRM^{cAbl}.

For aim (iii), a density-based solvation analysis is developed. The result reveals a coupling effect between protein dynamics and local solvation contribution, via which the viral ligand manages to decrease the solvation penalty to enhance its binding affinity.

In sum, this dissertation reveals PRM^{NS1}'s distinctive molecular recognition mechanisms underlying fuzzy interaction, conformational entropy, and surface hydration. And our newly developed solvation analysis is expected to be applied to the hydration effect of other biomolecules.

DEDICATION

To my precious daughter, my firstborn, the culmination of a lifetime of hopes and dreams, Katherine Qiqi, without whom this dissertation would have been completed in half the time.

For my grandma, who passed away during the COVID-19 pandemic. I am so sad that I could not be around you and say goodbye. I will always remember things you have taught me about life.

ACKNOWLEDGMENTS

First and foremost, I would like to thank my advisor, Professor Wonmuk Hwang, who has led me into the world of MD simulation. Thanks to Dr. Hwang's broad knowledge, sharp mind, and unconditional support, we can make this happen. Discussing with him is always creative and inspiring. It's great honor to learn from him.

I further thank my former lab mates, Xiaojing Teng and Andrew Liao for providing their great support and help. When I firstly came to the lab, I was way less familiar with English or MD research. They taught me a lot of basic things to help me get through the early difficult stage. Thanks also to all my current lab mates for providing a creative working environment.

The work presented here was conducted in collaboration with Professor Jae-hyun Cho's lab. The experimental work and analyses were all attributed to Dr. Cho and his students. I thank him for providing resources, support, and wonderful insights from the experimental perspectives, which crucially contribute to the direction of our research.

CONTRIBUTORS AND FUNDING SOURCES

Contributors

This work was supported by a dissertation committee consisting of Professor Wonmuk Hwang (as my advisor), Professor Roland Kaunas, and Professor Alvin Yeh of the Department of Biomedical Engineering and Professor Jae-Hyun Cho of the Department of Biochemistry & Biophysics.

Dr. Qingliang Shen processed experimental data. Professor Jae-Hyun Cho designed experimental research, performed and analyzed experiment. Professor Wonmuk Hwang in part designed simulation research, analyzed simulation.

All other work conducted for the dissertation was completed by the student under the supervision of Professor Wonmuk Hwang.

Funding Sources

Research reported in this dissertation was supported by the National Institute of General Medical Sciences of the U.S. National Institutes of Health under award number R01GM127723.

NOMENCLATURE

TAMU	Texas A&M University
SH3	Src homology 3
nSH3	N-terminal SH3
PRM	Proline-rich motif
NS1	Nonstructured protein 1
PDB	Protein Data Bank
CHARMM	Chemistry at HARvard Macromolecular Mechanics
IDP	Intrinsically disordered Protein
IDR	Intrinsically disordered Region
RMSF	Root-Mean-Square Fluctuation
MIST	Maximal Information Spanning Tree
CrkII	CT-10 regulator of kinase II
IAV	Influenza A virus
PI3K	Phosphatidylinositol 3-kinase

TABLE OF CONTENTS

Page

LIST OF FIGURES

FIGURE Page

LIST OF TABLES

TABLE Page

1. INTRODUCTION *

1.1 IDP and PRM

1.1.1 IDP

Intrinsically disordered proteins (IDP) or proteins containing intrinsically disordered region (IDR) are proteins or regions of proteins remaining structurally disordered in their native, functional states, which is to say, a single equilibrium 3D structure can not represent them for thorough study [3, 4]. IDPs and IDRs are found abundant in cells [5] and they are widely involved in recognition between proteins [6, 7, 8, 9, 10]. Many IDPs/IDRs undergo disorder-to-order transitions upon binding to their partner proteins. The high entropic penalty for disorder-to-order transition leads to a relatively low binding affinity [8, 10, 11], which may help with reversible signaling when quick turnover is desired [6, 12, 8]. In systems with a higher binding affinity, the entropic penalty is offset by favorable enthalpy and/or it is reduced by retaining the "fuzziness" in the complexed state [13, 10, 11]. Such fully or partially unstructured yet functional protein complexes have challenged the traditional notion of lock-and-key binding mechanism involving well-defined structural states [5].

IDPs/IDRs tend to expose their mobile region to the aqueous environment. In their amino acid sequence, the so-called "order-promoting" nonpolar residues are less abundant compared to polar or charged residues [14, 15]. One exception is proline, which is considered one of the most disorder-promoting residues despite being nonpolar [16].

1.1.2 Fuzzy Interaction

Fuzzy interaction refers to recent findings that some IDPs/IDRs can stay structurally disordered after binding to their partner [13]. This type of complexes are also called fuzzy complexes. Compared to IDP/IDR, this phenomenon used to be overlooked but now raises serious questions about

*Part of the contents are reprinted with permissions from 1) Shen, Q., et al. *Biophysical journal*, 114(6), pp.1313-1320, and 2) Shi, J., et al. *Biophysical journal*, 118(10), pp.2502-2512.

how to interpret it functionally and evolutionarily. The fuzzy interactions based on the disorderness can be classified into four categories.

- *Polymorphic model*. The bound molecule retains well-defined structure, but adopts more than one unrelated conformations [17].
- *Clamp model*. The bound protein contains two folded regions connected by a disordered linker [18].
- *Flanking model*. The central region of the IDPs forms stable contacts with the binding partner, thus become ordered, while the two terminal ends retain conformational flexibility [19].
- *Random model*. Basically the entire protein remains disordered [20].

Among the four models, PRMs upon binding usually fall into the flanking model, but they are also found to be able to adopt at least one opposite orientation in the bound state [21, 22].

1.1.3 PRM and SH3:PRM Binding

The proline-rich motif (PRM) is one of the most abundant linear motifs found in IDPs/IDRs [23, 24]. The cyclic structure of proline limits conformational degrees of freedom (DOFs) [15], and PRMs tend to form left-handed polyproline type II (PPII) helices upon binding to its receptor [25]. In a PPII helix, backbone carbonyl and side-chain groups are exposed and can form intermolecular contacts [15]. This makes PRMs preferred recognition motifs for many signaling proteins such as Src homology 3 (SH3) and WW domains [23, 26, 24, 27].

The SH3:PRM interaction is considered a typical IDP/IDR associated protein-protein interaction [21]. The binding generally involves negative (favorable) enthalpy and negative (unfavorable) entropy [28, 29, 11]. Several mechanisms have been proposed to explain this behavior. One proposal is that formation of PPII helix upon binding to SH3 contributes to the unfavorable binding entropy [30]. For some PRMs that possess low PPII propensity [31], the backbone conformational change is found to contribute to the entropic penalty [28, 11]. Another mechanism is based

on the finding that interfacial water molecules mediate hydrogen-bond network in the binding pocket and influence thermodynamics of binding, though details of this mechanism remain unclear [29, 32, 33, 34].

1.1.4 MD-based Studies of SH3:PRM Binding

Computer simulations have contributed greatly to understanding the binding of PRM to SH3. A main focus of these studies has been calculation of the binding free energy of different PRMs. This requires calculation of terms including: electrostatic screening and hydrophobic effect, entropic effect, and structural contribution of interfacial water. The electrostatic screening effect can be treated by numerically solving the Poisson-Boltzmann equation [35]. Hydrophobic effect is calculated by an energy term proportional to the solvent accessible surface area (ASA). These two methods in combination have been applied to calculate the binding free energy of PRM to SH3 in molecular dynamics (MD) simulations [36, 37]. They have also been applied to proteome-level screening of PRM ligands for SH3 [38, 39, 40]. While MD-based calculation performs better compared to other more empirical calculation [41], still the accuracy is limited [40]. This reflects that dynamic aspects of the interaction are important. As later addressed in this dissertation, we found that mobility ("fuzziness") of a PRM bound to SH3 enhances their electrostatic interactions [42], which suggests that consideration of entropy is necessary. Previously, entropy calculation has been done mainly via normal mode analysis (NMA) [43, 44, 36, 45]. In this method, vibrational degrees of freedom calculated around a locally energy minimized structure is used to compute conformational entropy. However, due to the computational cost of NMA, only a small number of snapshots from relatively short MD simulations (a few nanoseconds) were used. Thus, only limited volume of the conformational space is considered, which may have been responsible for a small entropic contribution to the binding free energy [45].

Another challenge in studying the SH3:PRM interaction is the role of interfacial water. Despite the increasing recognition for its importance [34], few computational studies have been performed to analyze interfacial water molecules. Palencia and co-workers performed 12-ns MD simulations of an SH3 domain from the Abl kinase complexed with several point mutants of a PRM named p41

[32]. This work demonstrated the plasticity of water-mediated interaction where alternative hydrogen bond networks form for mutant peptides. They also identified a hydration site that exhibits a large change in water occupancy between SH3 complexed with wild-type and a mutant PRM. By comparing isothermal titration calorimetry data, they estimated that the hydration site provides -20 kJ/mol (-4.78 kcal/mol) of binding enthalpy. However, since systems with two different peptides (wild type and mutant) were compared, it is unclear how much the difference is attributable to the point mutation or to changes in water coordination.

1.2 Simulated Systems and Biological Significance

1.2.1 nSH3 Domain in the CrkII Protein

CrkII protein belongs to CT-10 regulator of kinase (Crk) protein family that was originally identified in chicken retrovirus [46]. Crk protein family are a group of signaling adaptor proteins that are involved in a variety of signal transduction cascades including the modulation of cell adhesion, proliferation, migration and immune cell responses [47, 48].

CrkII is composed of one Src homology 2 (SH2), one N-terminal Src homology 3 (nSH3) domain, and one C-terminal (cSH3) domain [47]. SH2 and SH3 domains are the functional groups in CrkII that are responsible for mediating protein-protein interactions [49]. SH domains are highly conserved modules and can also be found as the functional groups in other adaptor protein families like Grb2 and Nck [48].

Among the three SH3 domains found in CrkII, the nSH3 domain, along with the SH2 domain, is known to play the central role in the mediation of a number of protein-protein interactions, by binding to some proline-rich motifs (PRMs) [47, 48, 1]. The nSH3 domains usually recognize proline-rich protein motifs containing the core PxxP motifs plus some additional amino acids nearby the motif to help determine the binding specificity. But deviations from this pattern are not rare [50, 51].

Because the SH3:PRM binding pattern is so common along the signaling pathway, there is great need to elucidate the molecular mechanism of this type of protein recognition. Two SH3:PRM

complexes are visualized in Fig. 1.1.

1.2.2 PRMs in the Protein-Tyrosine Kinase Abl

cAbl belongs to the Abl family of nonreceptor tyrosine kinases, which was firstly identified in the Abelson murine leukemia virus [52]. cAbl is found to be able to interact with a variety of proteins, thus involving a variety of biological processes like regulation of cell growth, cell survival, cytoskeleton dynamics, oxidative stress, etc [53, 54]. Among those proteins, CrkII protein turns out to be one of the major binding partners[55]. cAbl consists of SH3 domain, SH2 domain, kinase domain, followed by a long structurally disordered C-terminal extension, which is known as the last exon region. The C-terminal region contains DNA-binding region, actin-binding domain, proline-rich motifs and other protein-protein interaction sites that make the C-terminal region critical for a variety of functions of the protein [53].

cAbl binds predominantly to the nSH3 domain of CrkII via its C-terminal PRMs [56]. The structure of nSH3:PRM^{cAbl} is shown in Fig. 1.1. The binding results in the phosphorylation Y221 of CrkII, which further inhibits the CrkII signaling through inducing an intramolecular pY211-SH2 interaction that disrupts the assembled structure of CrkII thus impairing the interaction of its SH2 and SH3 domains with other proteins [57].

The effects of interaction between cAbl and CrkII turn out to be bidirectional thus attract a wide range of interest to study. For CrkII, the inhibition of CrkII signaling leads to decreasing the rate of focal adhesion and lamellipodium formation but increasing the rate of filopodium formation [56]. While for cAbl, the phosphorylated CrkII in turn trans-activates cAbl kinase via its cSH3 domain which binds to the PRMs of cAbl, inducing its activation in the absence of phosphorylation [58].

Clinically, the study of interaction between cAbl and CrkII also helps develop anticancer drugs that target Bcr-Abl— fusion protein made of ABL gene on chromosome 9 and BCR gene on chromosome 22 — which causes chronic myeloid leukemia [59, 60].

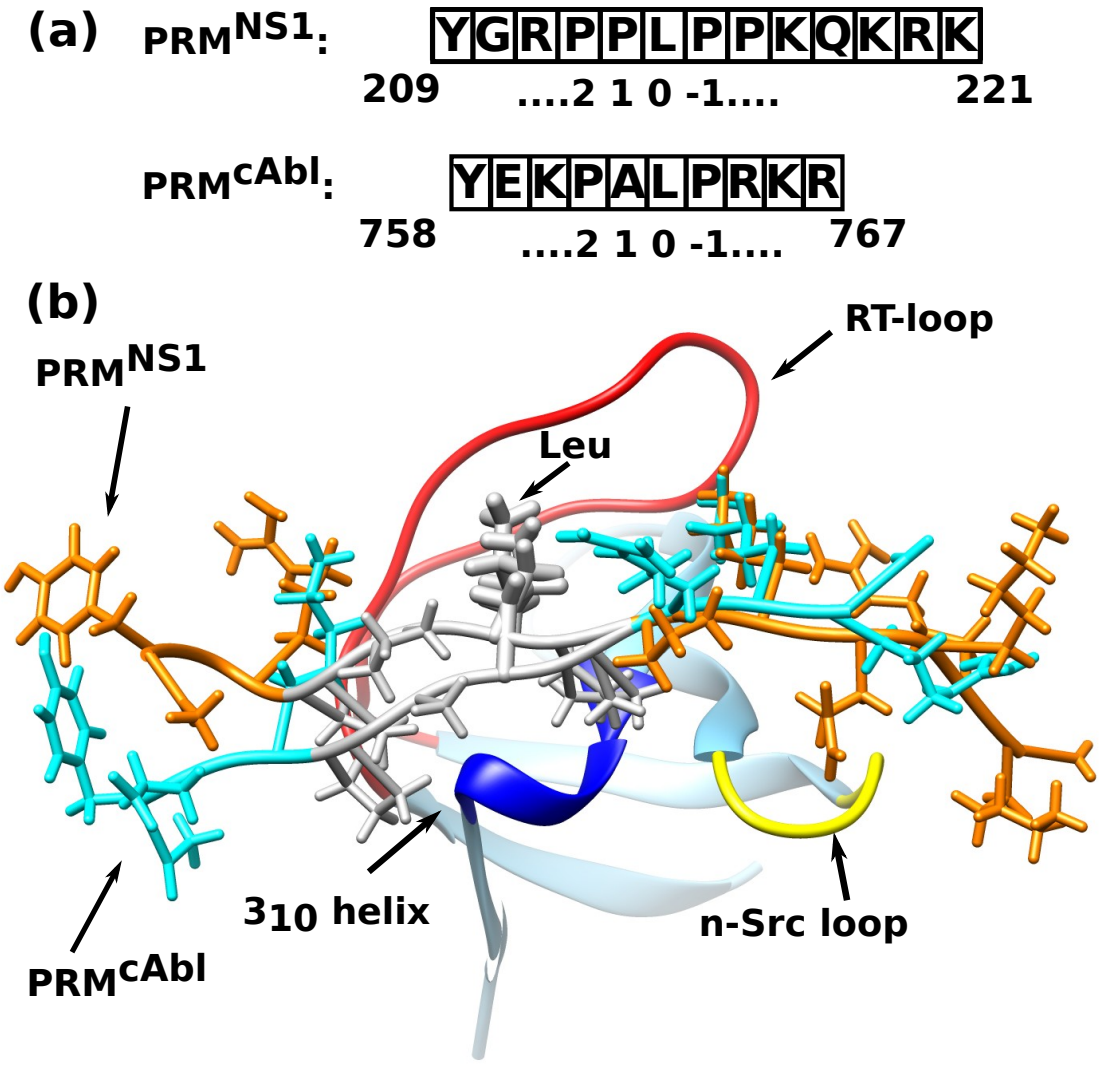

Figure 1.1: Structural overview of nSH3:PRM^{cAbl} and nSH3:PRM^{NS1}. (a) Sequences of the two ligands. Relative positions are marked below, where the central Leu is at position 0. The position index increases N-terminally [1, 2]. (b) Two structures superimposed relative to the nSH3 domain. PRM^{cAbl} and PRM^{NS1} are colored differently, and the central PxxP motif is white. Major subdomains of nSH3 (n-Src and RT loops, and 3_{10}-helix) that interact with the PRM are shown in solid colors. The rest of nSH3 is semi-transparent.

1.2.3 PRMs in the NS1 Protein of 1918 Flu Virus

1918 Spanish flu belongs to influenza A virus (IAV) and has caused one of the most severe pandemics in the history. It is estimated to infect $\sim$ 500 million people out of which the number of deaths was $\sim$ 40 million [61]. Living under the shadow of COVID-19 pandemic, this should bring

more concern that another pandemic may occur caused by a new flu strain.

The genome of IAV is found to code for 11 viral proteins. Among them, nonstructural protein 1 (NS1) is found to be multifunctional and is able to promote IAV replication and counteract antiviral activities by hijacking the host's protein–protein interactions along the immune response pathway [62, 63].

Unlike common seasonal flu, the C-terminal end of NS1 in 1918 Spanish flu contains a very rare PxxP motif that can bind to the nSH3 domain in the Crk proteins resulting in enhanced phosphatidylinositol 3-kinase (PI3K) signaling [64, 65]. The NS1-activated PI3K signaling has been suggested to promote the IAV replication [66, 64]. Recently, it has been shown that the PRM in the NS1 does bind to the nSH3 of CrkII with an exceptionally high binding affinity (Kd $\sim$10 nM) [2]. The structure of nSH3:PRM$^{\mathrm{NS1}}$ complex is shown in Fig. 1.1.

The extremely high affinity found in nSH3:PRM$^{\mathrm{NS1}}$ allows the virus to hijack the CrkII thus in turn disrupting the CrkII associated cellular signaling pathway. Therefore, the molecular mechanism underlying the high binding affinity of the viral ligand may be critical to help design the antiviral drug and prevent the future pandemics threatened by deadly flu viruses like 1918 Spanish flu. It may also shed light on rational ligand design that can strongly bind to SH3 domains involving diseases like tumor growth.

1.2.4 Thermodynamic Factors in Ligand Binding

As a promising therapeutic strategy, a rational design of potent drugs that can inhibit the SH3 domains and their corresponding binding ligands requires an in-depth understanding of the binding free energy.

The binding of nSH3 domain by PRM$^{\mathrm{cAbl}}$ is like other IDP/IDR associated SH3:PRM binding, involving negative (favorable) enthalpy and negative (unfavorable) entropy [1]. But a quantitative accounting of those factors in a molecular basis is still poorly achieved. Some groups have estimated the conformational entropy in SH3 domain based on the backbone N-H bond order parameters, suggesting the ordering backbone N-H bonds in the SH3 domain contributes significantly to the entropy decrease [28, 1]. Some found the formation of PPII helices by PRMs is also an in-

dispensable part in terms of the entropic change [30]. However, a more accurate estimation of backbone and side chain entropy change, taking into account the correlated motions from both SH3 domain and PRM is still needed. It's also important for ligand design to elucidate the entropic contribution on a residue-level. As mentioned in the previous section, water molecules at the binding interface may also play an important role in the thermodynamic changes in the PRM binding, but detailed evaluation of the hydration contribution remains elusive.

Comparing PRM^{cAbl} with viral ligand PRM^{NS1}, the exceptionally high binding affinity found in PRM^{NS1} further raises questions about how the viral ligand regulates its thermodynamic factors to achieve such low binding free energy, considering most cellular PRMs only have moderate binding affinities towards SH3 domains. To answer this, a detailed evaluation of entropic and solvation contribution is critical. Besides, the structurally disordered ends of PRM^{NS1} contain a total of 5 positively charged residues, which has been suggested to contribute to the high binding affinity via long-range electrostatic interactions [2]. According to a polyelectrostatic model proposed by Borg [67], increased charges in IDP can actually increase the binding affinity via long-range electrostatic interactions. Thus, it remains to be determined if the fuzziness found in the bound PRM^{NS1} contributes to the long-range electrostatic interactions as assumed in the polyelectrostatic model.

1.3 Conformational Entropy and Protein

1.3.1 Traditional Approaches for Entropy Calculation

Despite the importance of conformational entropy for binding, residue-level contribution has been difficult to quantify. Previous analysis considered mainly changes in the backbone conformation while less is known about specific contributions of side chains [11]. Experimentally, the "entropy meter" approach that empirically relates nuclear magnetic resonance (NMR) derived order parameters of side-chain methyl groups to entropy has been effective [68, 69, 70, 71, 72], but it is difficult to apply to proteins with low content of methyl-bearing amino acids such as SH3 and its ligands. On the other hand, various computational methods have been developed for calculating conformational entropy. Early efforts included normal mode analysis (NMA) [43, 73, 74] and

direct enumeration of side chain rotamers [75, 76]. However, NMA cannot account for transitions between energy minima since it relies on harmonic approximation of the potential energy surface about a single energy minimum. The direct enumeration approach becomes prohibitive for large proteins. NMA has also been applied to different snapshots of a molecular dynamics (MD) simulation trajectory. This mitigates the problem of working only around a single energy minimum, and it has been used for the study of ligand binding to SH3 domains [36, 45]. However, the computational cost of NMA limited the number of coordinate frames used, which likely resulted in only a small contribution of entropy to the binding specificity of different ligands [45]. These limitations were partly addressed by a Monte Carlo simulation of side chains while the protein backbone was held fixed [77]. It yielded total side-chain entropies among various systems correlating well with experimental values. Despite much insight this study provided, due to the simplified interaction potential that was introduced to allow Monte Carlo moves, and due to the backbone being fixed to the crystallographic conformation, the accuracy and applicability of the method were limited when comparing entropy changes in binding of different ligands, or for finding residue-level contributions to entropy.

1.3.2 MIST for Calculating Entropy

For traditional approaches, detailed residue-level analysis of the entropy of ligand binding based on unbiased atomistic MD simulation has been difficult. To this end, the maximum information spanning tree (MIST) approach is chosen for the study of conformational entropy in SH3:PRM binding [78]. MIST is an information theory based approximation method that overcomes the difficulty of direct enumeration arising from the large number of DOFs in typical biomolecular systems. It has also been shown to overcome the poor performance in practical as the result of limited sampling size while dealing with high-dimensional information theoretic statistics.

It first calculates entropies of individual DOFs, then it adds corrections arising from higher order correlated motions among DOFs in such a way that the estimated entropy monotonically approaches the exact value as higher order correlations are accounted for. By increasing the order of mutual information, the estimated entropy gets closer and closer to the exact value. This allows for

the optimization of the order to balance the limited sampling size and the needed accuracy. MIST yielded results that agree well with experimental estimates based on the entropy meter approach for several different systems [68]. Compared with other lower-order approximation approaches like Mutual Information Expansion (MIE), MIST has demonstrated faster convergence as well as improved accuracy, particular for larger biological systems [79]. The details of applying MIST can be found in the Method section.

1.4 Surface Hydration and Protein

1.4.1 Critical Role of Water for Protein

Surface hydration water around biomolecules plays a critical role for understanding their structure, folding and biomolecular recognition processes [80, 81, 82]. Water molecules in the vicinity of the protein surface may form a hydration shell that works more than as lubricant. It has been found that the hydration shell facilitates certain structural/dynamical aspects in protein and also changes the protein free energy landscape via water-mediated interactions to control the folding and recognition process of proteins [83, 84, 85, 86]. This is due to the fact that water molecules at hydration shell may experience distinctive dynamics and behave in concert with protein motion. Some study found that the rotation and diffusion of water in the vicinity of protein may be 3-7 times slower than that in the bulk [87]. The density of the hydration shell usually is also higher than the bulk [88]. The competition of water-mediated interactions can facilitate the protein native states dynamics [89]. On the other hand, water close the surface of biomolecules also shows different dynamics from the bulk water, suggesting a mutual "slaving" or coupling mechanism [90, 91, 86].

In the case of ligand design, some studies found the final effect for a ligand modification is strongly correlated with the displacement of the ordered water molecule nearby the protein binding site [92]. This is because the process of displacing water away from the protein surface may cost energy as long as the water is held somehow. This is critical for the ligand design as over 80% of the protein-ligand complexes show at least one water molecule present at the protein-ligand interface [93]. Therefore, there is a strong need for accurately estimating the solvation pattern as

well as the thermodynamics effect of the change in the hydration shell to help predict the binding affinity change upon ligand modifications [94, 95]. But it does not mean the displacement of water always costs energy. It's been found that the displacement of ordered cluster of multiple water molecules at once from a mainly hydrophobic cavity can be energetically favorable due to its huge entropic contribution via water liberation.

Another important contributions from water is the hydrophobic effect where the hydrophobic units in water tend to aggregate [96]. The earliest and also most pervasive explanation for this effect is Kauzmann's "iceberg" model which assumes water near the hydrophobe forms highly ordered icelike structure that can make great entropic contribution when two hydrophobes bound [97]. Following Kauzmann's model, Chandler further suggested that for larger hydrophobes—length scale above 1 nm—the surface water can "dewet" the surface, forming a vapor-liquid interface [98]. Although their models seem sound in principal, a lot of experimental and simulational results turn out to contradict their theories [99, 100, 101, 102, 80]. When it comes to the protein surface, it is thought to be more complicated and different behaviors of water were observed as compared to other idealized hydrophobic/hydrophilic surfaces [103, 104]. The current consensus is that the behavior/contribution of water molecules around protein surface is highly context-dependent and can not be determined by a single factor [105, 106, 107].

1.4.2 Existing Approaches for Calculating Solvation Contributions

Molecular dynamics (MD) simulation has proved to be useful and promising in unraveling the solvation contribution in thermodynamics, as it provides atomistic details in protein surface water, and protein-water dynamic interactions.

Classical methods—free energy perturbation (FEP) [108] and thermodynamic integration (TI) [109] —allow for the calculation of total solvation contribution by decoupling the solvent from the solute gradually. But they are mainly applicable to small solutes. When applied to protein, its usage is mainly limited to a small number of bound water molecules [110, 111, 112].

Another classical method is to apply continuum models, as compared with explicit water models, to the protein so that the electrostatic contribution of solvation can be calculated by solving

Poisson-Boltzmann equation and the nonpolar contribution by a simple scaling of the surface area or the volume of the solute [113, 114, 115, 116]. However, continuum models, due to its very nature, does not account for some essential features of liquid water, e.g., the ability of water molecules to form hydrogen bonds near the protein surface [117]. Therefore, this method has greatly reduced accuracy in the near-protein/solute surface, especially for the residues that form hydrogen bonds with the water molecules [113, 118].

Inhomogeneous fluid solvation theory (IFST) developed by Lazaridis estimates the thermodynamics of water molecules through solute-solvent and solvent-solvent two-particle correlations [119, 120]. This method, along with its later development, grid inhomogeneous solvation theory (GIST) have been applied to small solutes, single water at bound state, and some hydration sites [121, 122, 123, 124, 125]. However, applying IFST and GIST requires fixing the solutes that sacrifices the dynamic aspect of water-protein interactions, and has not been extended to provide the full picture of the hydration shell yet. The GIST has also received criticism for causing severe sampling problems for higher order correlations [126].

Henchman and McCammon combined methods of time averaged position (TAP) and averaged residue coordinate (ARC) that allows for the study of water density and structure around the flexible region in the protein. But their method loses accuracy due to the increasing sensitivity as the distance of water-protein increases and therefore is limited to the study of water molecules very close to the surface of the protein [127].

Hagai Meirovitch developed a method called hypothetical scanning, where each conformation is reconstructed step-by-step through transitions probabilities (TPs), thus can be used to calculate the absolute entropy. However, this method currently has mainly been applied to small biomolecules and pure water models [128, 129].

Grubmüller's lab developed a new approach to calculate the hydration rotational entropy via parameterizing the orientation of water molecules with quaternions, analyzed by a k-nearest-neighbor (KNN) density estimator combined with mutual information expansion (MIE). It's been reported to calculate the hydration rotational entropy with improved accuracy and spatial resolution [126].

But currently it's only been applied to bulk water simulation and only the rotational entropy has been calculated.

Our lab has previously developed an approach to construct the water hydration map which divides the space around the protein into cubic cells and extract hydration properties in a cell-based manner [130, 131]. In this dissertation, we furthered the method to calculate the solvation free energy upon the binding of SH3 domain to two ligands. Our cell-based calculation also provides details in the local region so that it may further elucidate the protein surface contribution in a residue-level.

2. MATERIALS AND METHODS *

2.1 Structures Preparations

Among the below, nSH3:PRM^{NS1A} and nSH3:PRM^{NS1B} are two different crystal structures of the nSH3:PRM^{NS1} complex, labeled in the same way as in my previous publication [42].

2.1.1 nSH3:PRM^{cAbl}

nSH3:PRM^{cAbl} is nSH3 domain of the CrkII domain bound by PRM^{cAbl} from the C-terminal disordered region of the cAbl protein [48, 132, 1]. The complex structure is from PDB code 5IH2 (1.8-Å resolution) [1]. Hydrogens were added using the HBUILD module in CHARMM [133].

2.1.2 nSH3:PRM^{NS1}

nSH3:PRM^{NS1} is nSH3 domain of the CrkII domain bound by PRM^{NS1} from the C-terminal tail of the (NS1) of the 1918 Spanish influenza A virus (IAV) [62, 134]. nSH3:PRM^{NS1A} is from PDB 5UL6 (1.45 Å) [2]. nSH3:PRM^{NS1B} is from PDB 6ATV (1.75 Å) [2].

2.1.3 Free nSH3

Free nSH3 is the unbound state of nSH3 domain of the CrkII, isolated from nSH3:PRM^{cAbl}.

2.1.4 Free PRM^{cAbl}

Free PRM^{cAbl} is the unbound state of PRM^{cAbl}, isolated from nSH3:PRM^{cAbl}.

2.1.5 Free PRM^{NS1}

Free PRM^{NS1} is the unbound state of PRM^{NS1}, isolated from nSH3:PRM^{NS1B}.

2.1.6 Bulk Water

The bulk water simulation is prepared in cubic water box from TIP3P water model [135]. The bulk water density is maintained by adding 1728 water molecules in a cubic box of side length 37.712 Å.

*Part of the contents are reprinted with permissions from 1) Shen, Q., et al. *Biophysical journal*, 114(6), pp.1313-1320, and 2) Shi, J., et al. *Biophysical journal*, 118(10), pp.2502-2512.

2.2 MD Simulations

All simulations were performed on machines at the Texas A&M High-Performance Research Computing Facility.

2.2.1 Unrestrained Simulation

For unrestrained simulations, we used CHARMM [136, 133] with the param36 all-atom force field [137]. Prior to solvation, a 4-stage energy minimization was carried out. In each stage, backbone heavy atoms were harmonically restrained, and 100 steps of steepest descent (SD) followed by 300 steps of the adopted basis Newton-Raphson (ABNR) minimization were performed. The stiffness of the harmonic restraint was, in kcal/mol·Å^2 units, 5 (stage 1), 1 (stage 2), 0.1 (stage 3), and 0 (stage 4; no restraint). The system was solvated in a cubic TIP3P water box [135] of side length of about 60 Å (nSH3:PRM), 56 Å (nSH3-only), and 54 Å (PRM-only). Ions (Cl^- and Na^+) were added to neutralize the system at about 150-mM concentration. Simulation systems had about 21,000 (nSH3:PRM), 16,000 (nSH3-only), and 15,000 (PRM-only) atoms.

After solvation, the 4-stage energy minimization explained above was applied again for the protein. The system was heated from 0 to 300 K over 100 ps followed by equilibration for 160 ps at 300 K. These were performed under a constant temperature and pressure (NPT) condition at 1 atm without any restraints. Each production run was under the constant volume and temperature (NVT) condition at 300 K for 500 ns. The integration step size was 2 fs. Lengths of covalent bonds involving hydrogen atoms were fixed by applying the SHAKE algorithm [138]. The non-bonded interaction had a 12-Å cutoff. The particle-mesh Ewald summation method [139] was used to account for long-range electrostatic interactions under a periodic boundary condition. The Domain Decomposition (DOMDEC) module [140] of CHARMM was used to enhance parallel performance. Coordinates were saved every 5 ps. Visualization of structures were done using VMD [141] or UCSF Chimera [142]. Calculations of root-mean-square fluctuation (RMSF), order parameter, entropy, and contact occupancy were all based on the 100–500 ns interval, to avoid the effect of the initial state. The 100-ns cutoff was determined by the relaxation times calculated for

the angular reorientational correlation functions of the backbone amide N-H bonds, as explained below.

2.2.2 Restrained Simulation

The majority of the preparatory procedures are the same as unrestrained simulation. Unlike unrestrained simulations, after the equilibration, production run in restrained simulation was performed with all protein atoms constrained under harmonic restraint of 50 kcal/mol·Å^2, in a similar way as previously published work [124, 125]. Since the unbound peptides undergo drastic conformational change during simulations when not constrained, it is critical to determine which peptide conformation to constrain as different choices may result in substantial differences. Here, the same orientation of the ligand as in the complex state is chosen to match the same treatment used in binding free energy calculation, the details of which can be found below. Production run for each constrained system was 50 ns long under NVT condition.

2.2.3 Bulk Water Simulation

The bulk water simulation was performed with TIP3P water model [135] using CHARMM [136, 133] with the param36 all-atom force field [137]. The simulated water box contains 1728 water molecules with a side length of 37.712 Å. Since it is water-only simulation, no restraints were needed. The production run was 10 ns long with coordinates saved every 0.5 ps.

2.3 Analysis of MD Simulations

Calculations of root-mean-square fluctuation (RMSF), order parameter, entropy, and contact occupancy were all based on the 100–500 ns interval, to avoid the effect of the initial state. The 100-ns cutoff was determined by the relaxation times calculated for the angular reorientational correlation functions of the backbone amide N-H bonds, as explained below.

2.3.1 Cluster Analysis

With backbone heavy atoms in the nSH3 domain as a positional reference, PRM^{NS1} conformations from the MD trajectories were analyzed using a cluster analysis [143]. Agglomerative

hierarchical clustering was used based on the similarity between different conformations, which is measured as the squared Euclidean distance of the backbone heavy atoms in PRM^{NS1}. The average linkage was chosen to measure inter-cluster similarity, which is the average of all pairwise similarities between observations in two clusters.The representative conformation of each cluster was chosen to have the smallest root-mean-square deviation from the calculated average structure for the cluster.

2.3.2 Order Parameters

After aligning coordinate frames to the initial frame using backbone heavy atoms, we calculated the Lipari-Szabo squared generalized order parameter S^2 [144]. Let $\vec{\mu}$ be the unit vector for the amide N-H bond with Cartesian components μ_i ($i = 1, 2, 3$). Then

$$S^2 = \frac{1}{2}\left(3\sum_{i=1}^{3}\sum_{j=1}^{3}\left\langle \mu_i\mu_j\right\rangle^2 - 1\right) \tag{2.1}$$

where $\langle\cdot\rangle$ indicates average over coordinate frames [145, 146]. To calculate S^2 for nSH3 or PRM within an nSH3:PRM complex, we aligned the backbone heavy atoms of the respective part, to be consistent with the uncomplexed case. Statistical uncertainty was estimated by subsampling the trajectory every 8th-frame and calculating the standard deviation in S^2 from the 8 subsamples. This method was more effective than using 8 consecutive 50-ns intervals, since relaxation times differ across the protein.

To assess the convergence of our trajectory, we calculated the angular reorientational correlation function $C_I(t)$ [145]:

$$C_I(t) = \left\langle P_2(\vec{\mu}(\tau)\cdot\vec{\mu}(\tau+t))\right\rangle_\tau \tag{2.2}$$

where $P_2(x) = \frac{1}{2}(3x^2 - 1)$ is the second Legendre polynomial. For each t, the average was taken over τ ranging from 100 ns to $(500 - t)$ ns.

2.3.3 Conformational Entropy

We used backbone and side-chain rotation angles as DOFs: For backbone, ϕ and ψ angles [147], and for the side chain, the dihedral χ angles [148]. Compared to bond lengths and bond angles, χ angles are the main contributors to the side-chain entropy [73, 76, 68]. Angles were measured using the CORREL facility of CHARMM. For a coordinate trajectory, histograms of angles were constructed with a bin size $\Delta = 10^\circ$. It is the smallest among bin sizes used previously, ranging from 10° [69] to 120° [68]. Other choices, $\Delta = 8^\circ$ and 15° were also used as a test, but our main results did not depend on precise choice of Δ (see Fig. 3.10 in the Appendix). For statistical uncertainty estimate, we subsampled the 100–500-ns interval every 4 frames and used the 4 sets, each containing the number of frames equal to 100 ns.

2.3.3.1 Maximum Information Spanning Tree (MIST)

To calculate the conformational entropy using MIST [78, 79], we started with the first order where DOFs are treated independently. From the histogram of a given DOF, we built the normalized probability distribution function (PDF) $P(\alpha_m^i)$ (m: DOF index, i: bin index). The first-order MIST (MIST1) entropy is:

$$S_1^{\text{MIST}} = -k_B \Delta \sum_{m=1}^{\text{DOF}} \sum_{i}^{\text{Bin}} P(\alpha_m^i) \ln P(\alpha_m^i), \tag{2.3}$$

where k_B is the Boltzmann constant.

We used the joint PDF $P(\alpha_m^i, \alpha_n^j)$ to calculate the second-order mutual information between DOF m and n:

$$I_2(m;n) = \Delta^2 \sum_{i,j}^{\text{Bin}} P(\alpha_m^i, \alpha_n^j) \ln \frac{P(\alpha_m^i, \alpha_n^j)}{P(\alpha_m^i) P(\alpha_n^j)} \tag{2.4}$$

The second-order MIST (MIST2) entropy is calculated by including I_2 for one DOF at a time, where only the maximum I_2 between the current DOF and all previously added DOFs is included:

$$S_2^{\text{MIST}} = S_1^{\text{MIST}} - k_B \sum_{p=2}^{\text{DOF}} \max_{q \in \{1,\ldots,(p-1)\}} [I_2(p;q)] \tag{2.5}$$

The third-order MIST (MIST3) is calculated similarly, by evaluating the third-order mutual information between DOF p and a pair of DOFs $\{q, r\}$:

$$I_3(p;q,r) = \Delta^3 \sum_{i,j,k}^{\text{Bin}} P(\alpha_p^i, \alpha_q^j, \alpha_r^k) \ln \frac{P(\alpha_p^i, \alpha_q^j, \alpha_r^k)}{P(\alpha_p^i) P(\alpha_q^j, \alpha_r^k)} \tag{2.6}$$

so that

$$S_3^{\text{MIST}} = S_1^{\text{MIST}} - k_B \sum_{p=3}^{\text{DOF}} \max_{\{q,r\}<p,\, q\neq r} [I_3(p;q,r)] \tag{2.7}$$

For a given p, only the maximum I_3 from all pairs of DOFs with indices less than p is included in the sum. If a certain DOF r is statistically independent of p and q, Eq 2.6 becomes identical to Eq 2.4. So Eq 2.7 includes pairwise correlations for MIST2. In principle, higher-order MIST can yield a more accurate (and smaller) estimation of entropy. But without a strong multivariate correlation, the improvement is marginal while the computational cost steeply rises and the high-dimensional joint PDF becomes less accurate due to the limited sample size. Thus, calculations up to the second or third order are practical [68, 70].

For nSH3, we excluded the first A134 and the last R191 from entropy calculation, as done previously [1]. We also excluded the terminal residues of PRMs [11]. The number of DOFs were 235 for nSH3, 59 for PRM$^{\text{NS1}}$, and 45 for PRM$^{\text{cAbl}}$. Entropies were calculated using custom-written R language scripts and also using FORTRAN 95 codes parallelized with OpenMP.

2.3.3.2 Entropy Per-Residue

The backbone entropy of each residue was calculated using the joint PDF,

$$S_{\text{bb}} = -k_B \Delta^2 \sum_i \sum_j P(\phi^i, \psi^j) \ln P(\phi^i, \psi^j), \tag{2.8}$$

which is identical to the two-DOF version of Eq 2.5. For side-chains, the number of DOFs (χ angles) varies between 1 and 5, where we used up to MIST3.

2.3.4 Solvation Free Energy

The approach of calculating solvation free energy of protein surface hydration is a further development of water density map approach previously developed by Wonmuk Hwang's lab that has already been implemented in CHARMM [130].

2.3.4.1 Construction of Water Density Map Around Protein

Firstly, the space around the protein was divided into cubic cells of side length 0.7 Å, which is about half of the radius of a water molecule. For each cell, the times that it's visited by a water oxygen atom are counted and then divided by the total frame number to get the fraction. It is then divided by the volume of the cell to get the number density of the water for that cell.

To remove the translational and rotational effect of the protein motions during the simulation, a reference structure for superimposition is needed. To further minimize the influence of internal motion of proteins, the reference structure is constructed as follows. Firstly, an average structure for each system is calculated via CHARMM (COOR DYNA command). To get rid of nonphysical contacts in the average structure, a 3-stage energy minimization was carried out. At each stage, 50 steps of steepest descent (SD) followed by 50 steps of the adopted basis Newton-Raphson (ABNR) minimization were performed. In stage 1, backbone heavy atoms were harmonically restrained. In stage 2, only backbone C_α atoms were restrained and no restraint was applied to stage 3. The spring constant for the harmonic restraint was 1 kcal/mol·Å^2 (for both stage 1 & 2). Before hydration analysis begins, all the coordinate frames were superimposed to the reference structure using only stably folded region i.e., C_α atoms with low RMSD values. The constructed hydration map and the molecular structures are either rendered in UCSF Chimera [142] or the Virtual Molecular Dynamics (VMD) [141].

2.3.4.2 Density-Based Solvation Free Energy Calculation

Once the density map is calculated, ideally the total solvation free energy is given by:

$$E_{solv} = -k_B T \sum_{i}^{N_{cube}} \ln \rho_i, \tag{2.9}$$

where k_B is the Boltzmann constant, T is the temperature for the system, ρ_i is the density for $i-th$ cubic cell, and N_{cube} is the total number of cubes selected through the procedures described below for calculation. And the summation is taken over all the N_{cube} cells.

However, for cells that are near the protein surface, especially around the mobile region, they have the chance to be occupied by protein atoms during the simulation. Since our cell size is set as small as 0.7^3 $\mathring{A}^3$, water molecules are blocked from entering the cell when cells are visited by any of the protein atom due to the excluded volume effect. Depending on how often a cell is occupied by a protein atom, the density of it may be systematically underestimated to different extents.

To fix this, one way is to simply exclude all the cubes that have ever been visited by protein atoms. Another way is to count how many times each cube has ever been visited by any protein atoms. While calculating the fraction, we exclude those protein-occupied frames out of the total frame numbers. This means the actual total frame numbers vary across the cubes. From eq 2.9, the density of each cube is further weighted by the fraction that the cube is not visited by protein. Therefore the weighted solvation free energy is given by:

$$E_{solv} = -k_B T \sum_{i}^{N_{cube}} \ln \rho_i', \tag{2.10}$$

$$\rho_i' = \rho_i \times \frac{N_{frame}}{N_{frame} - V_{frame,i}}, \tag{2.11}$$

where N_{frame} is the total frame number, and $V_{frame,i}$ is the total counts that $i-th$ cube has been visited by a protein atom.

In order to estimate the effect of the higher water density around the protein surface, the average solvation free energy per cubic cell for the bulk water is used as a reference value, given by:

$$\overline{E}_{bulk} = -k_B T \ln \overline{\rho}_{bulk}, \tag{2.12}$$

where $\overline{\rho}_{bulk}$ is estimated from MD simulation of bulk water.

For each system, the total contribution of water to the energetic change is measured as the

difference between total solvation energy from cubic cells in the system and total solvation energy of bulk water cells with the same N_{cube} as in the system, given by:

$$\begin{aligned} \Delta G_{solv} &= E_{solv} - N_{cube} \times \overline{E}_{bulk} \\ &= -k_B T \sum_{i}^{N_{cube}} (\ln \rho_i' - \ln \overline{\rho}_{bulk}) \end{aligned} \tag{2.13}$$

The desolvation penalty (ΔG_{penal}) upon binding of PRM is calculated as the difference in ΔG_{solv} between bound complex and the unbound components (nSH3 as well as the ligand).

2.3.4.3 Optimization of Parameters in Cube Selections

We firstly need to find the optimal density cutoff for our systems to exclude cells with density below the cutoff. Using equations described in the previous sections, the desolvation energy as a function of the density cutoff was calculated (see Figure 2.1). We calculated the desolvation energy for density starting from 0.026 Å^{-3}, slightly lower than the bulk water, in increment 0.002 Å^{-3}. It turns out for both complexes, the peak of the desolvation penalty appears at 0.034 Å^{-3}, right above the bulk water density. This is reasonable because as density cutoff becomes higher, less cubes will be included. The peak at 0.034 Å^{-3} suggests that after the density passes the bulk water state, the increase in density cutoff has more influence on the surface hydration structure of unbound nSH3 than those in the bound states. Interestingly, the difference between the two complexes also peaks at the same position. Based on these findings, we used density cutoff 0.034 Å^{-3} for later analysis.

Next, the distinctive hydration layer around the protein surface needs to be identified. It is known that the hydration layer is of higher density than bulk water [88, 82]. However, it appears that the criteria for defining and probing the hydration layer is still debated [91]. For our purpose, we measured the average density of the cubic cells as a function of distance to the protein surface to identify it (see Figure 2.2). As Figure 2.2 shows, when cells are too close to the protein surface, the water density would tend to be low due to the excluded volume effect. The surface water organization reaches its peak density at the distance around 3 Å to 3.5 Å. After the peak, our results agree with the continuous density distribution model that the density monotonically decreases until

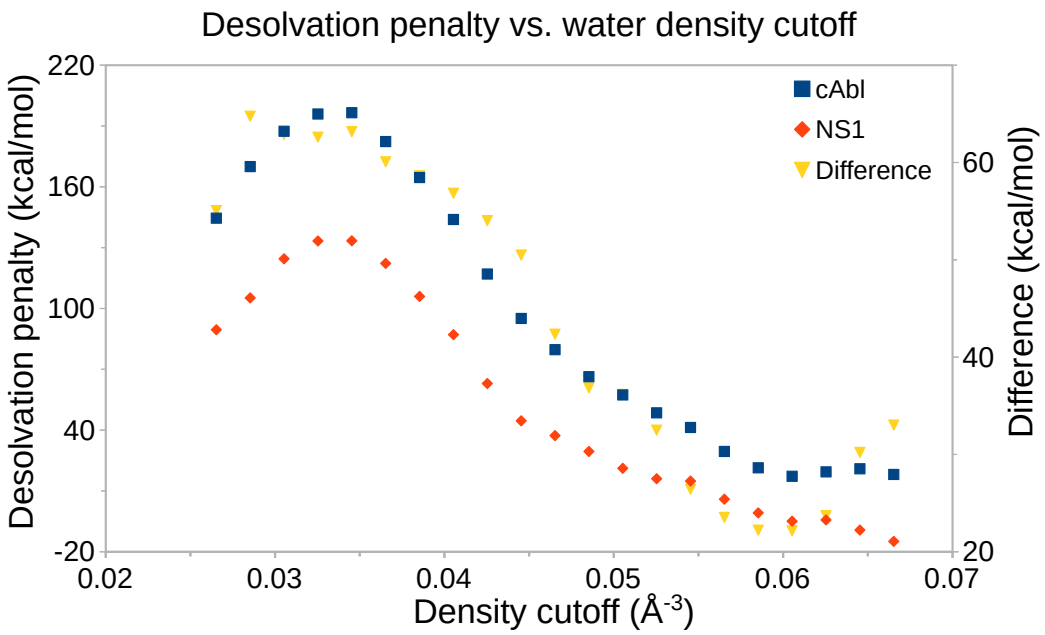

Figure 2.1: Desolvation penalty for the formation of two complexes as well as the difference between them, calculated with varying cutoff densities. cAbl: nSH3:PRM^{cAbl}. NS1: nSH3:PRM^{NS1}. Difference is calculated as cAbl minus NS1. A positive value indicates the former experiences greater desolvation penalty.

it reaches the bulk water level [80]. For our systems, we use 4.5 Å as distance cutoff in order to include all the high-density cubes in the hydration shell. This is within the typical hydration shell thickness 4-8 Å [149].

A schematic illustration of the applied exclusion criteria can be found in Figure 2.3. Since the chemical groups distribution as well as the topography of protein surface vary across different proteins, it is expected to find different optimized parameters for cube selections.

2.3.5 MM/PBSA Method

The binding free energy between nSH3 and a PRM was calculated using molecular mechanics/Poisson Boltzmann surface area (MM/PBSA), combining the thermodynamic cycling method

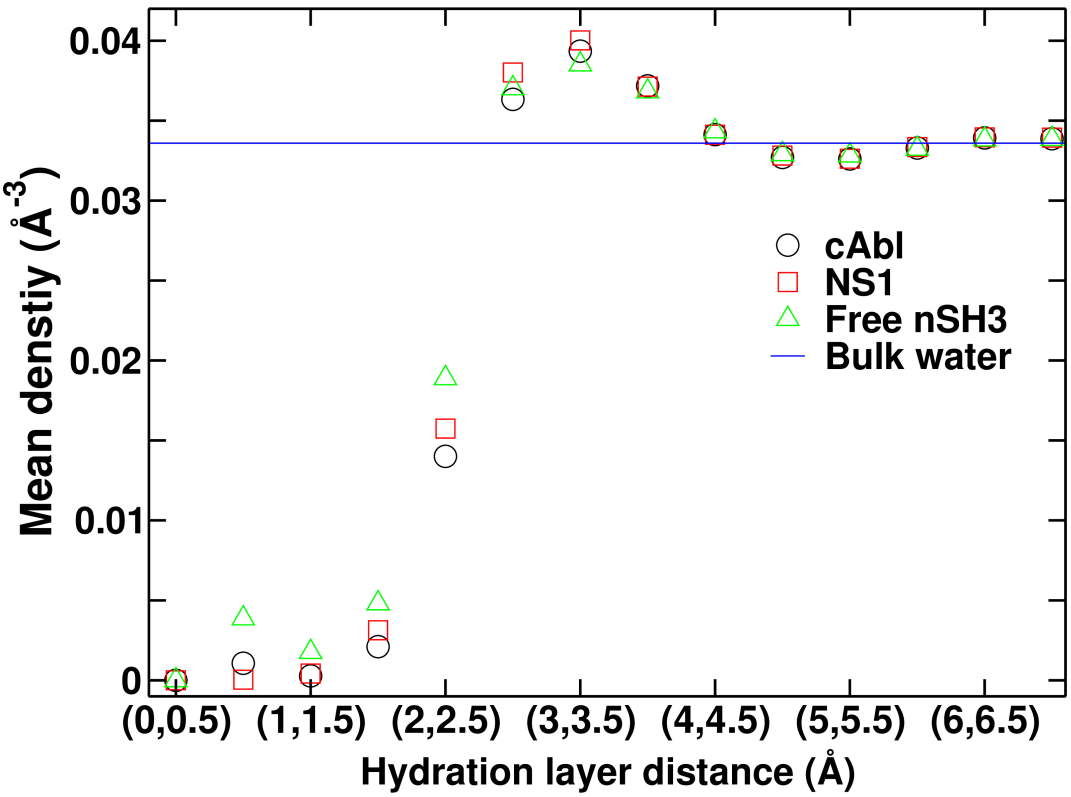

Figure 2.2: Average density of water hydration shell at each distance range. X-axis: the hydration layer distance is measured as the distance from the center of each cubic cell to the closest heavy protein atoms in the reference structure. cAbl: nSH3:PRM$^{\text{cAbl}}$. NS1: nSH3:PRM$^{\text{NS1}}$. Free nSH3: unbound nSH3 domain. Bulk water: bulk water density for reference measured from bulk water simulation.

based on the following scheme [115, 116]:

$$\begin{array}{lccccc} \text{(Vacuum)} & \text{nSH3} & + & \text{PRM} & \xrightarrow{\Delta G^{\circ}_{\text{bind}}} & \text{nSH3:PRM} \\ & \downarrow \Delta G^{nSH3}_{\text{solv}} & & \downarrow \Delta G^{\text{PRM}}_{\text{solv}} & & \downarrow \Delta G^{\text{complex}}_{\text{solv}} \\ \text{(Solvated)} & \text{nSH3} & + & \text{PRM} & \xrightarrow{\Delta G_{\text{bind}}} & \text{nSH3:PRM} \end{array}$$

In the above, $\Delta G^{\circ}_{\text{bind}}$ is the binding free energy in vacuum, given by:

$$\Delta G^{\circ}_{\text{bind}} = \Delta E_{\text{intra}} + \Delta E_{\text{vdW}} + \Delta E_{\text{elec}} \tag{2.14}$$

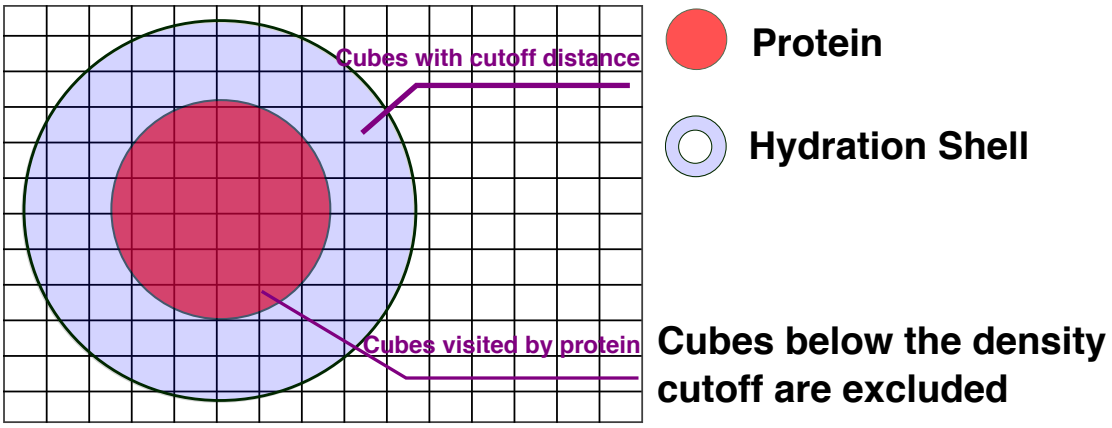

Figure 2.3: A schematic illustration of selection criteria applied. Cubes within the cutoff distance is considered as belonging to the hydration shell. Cubes visited by the protein during the simulation will be either excluded or weighted by the actual frame numbers that are not visited by any protein atoms (more details can be found in Method section). Cubes belong to the hydration shell but have density below the cutoff will not be selected into our hydration calculation.

where E_{intra} is the intramolecular energy associated with covalent bond lengths and bond angles. E_{vdW} is the van der Waals energy. E_{elec} is the electrostatic energy in vacuum. These energy terms are given by the CHARMM force field. The solvation free energy $\Delta G^{X}_{\text{solv}}$ (X: PRM, nSH3, or complex) is:

$$\Delta G^{X}_{\text{solv}} = \Delta G^{X}_{\text{np}} + \Delta G^{X}_{\text{elec}} \tag{2.15}$$

where G^{X}_{np} accounts for the non-polar contribution proportional to the ASA. G^{X}_{elec} is the generalized Born polar solvation free energy, calculated using the GBSW module of CHARMM [150].

The binding free energy ΔG_{bind} for the nSH3:PRM complex in solution is given by

$$\begin{aligned}\Delta G_{\text{bind}} &= \Delta G^{\circ}_{\text{bind}} + \Delta G^{\text{complex}}_{\text{solv}} - \Delta G^{nSH3}_{\text{solv}} - \Delta G^{PRM}_{\text{solv}} - T\Delta S \\ &= \Delta G^{\circ}_{\text{bind}} + \Delta G^{\text{complex}}_{\text{solv}} - \Delta G^{nSH3}_{\text{solv}} - \Delta G^{PRM}_{\text{solv}} - T\Delta S_{conf} - T\Delta S_{rot} - T\Delta S_{trans}\end{aligned} \tag{2.16}$$

$$S_{rot} = \frac{k_B}{2}\left(3 + \ln(\pi I_A I_B I_C) + 3\ln\left(\frac{8\pi^2}{\beta h^2}\right) - 2\ln\sigma\right) \tag{2.17}$$

$$S_{trans} = \frac{k_B}{2}\left(5 + 3\ln\left(\frac{2\pi m}{\beta h^2}\right) - 2\ln\rho\right), \tag{2.18}$$

where entropy S_{conf} is the conformational entropy calculated using the 3rd-order maximum information spanning tree (MIST) approach [78, 79] applied to the distributions of backbone and side-chain rotational angles measured during the simulations of isolated nSH3 and PRM, and their complex. Details about the conformational entropy calculation can be found in the previous section. $I_A I_B I_C$ are the three rotational moments of inertia. h is the Planck constant. σ is the symmetry factor of the molecule, which is 1 for non-symmetric molecule. m is the mass of the protein. ρ is the number density, set to 1 M for convenience.

Except for the entropy term, all other energy terms were calculated solely from the trajectories of the complexes [115], which is known as the same trajectory method (STM). STM has been proved to yield more stable results than different trajectory method (DTM) [115].

3. RESULTS AND DISCUSSIONS *

3.1 Analysis of Protein Dynamics and Fuzzy Interactions

Our comparative study of protein dynamics combining the computational and experimental methods not only helps validate 500 ns simulation time's adequacy for our purpose, but also help elucidate the mechanism of enhanced binding affinity through fuzzy interaction as observed in PRM^{NS1}.

3.1.1 Analysis of Order Parameters and Backbone Fluctuations

The order parameter ranges between 0 (flexible) and 1 (rigid) [151, 152]. To ensure convergence of calculated values, we calculated the angular reorientational correlation function $C_I(t)$[145] (Eq. 2.2; Fig. 3.1). The resulting relaxation time is generally longer in flexible regions, the longest being 46.6 ns, for K189 of nSH3 in the nSH3:PRM^{NS1B} complex (Fig. 3.2a). To minimize any potential influence of the initially prepared state, we excluded the first 100 ns from our analysis.

For both nSH3:PRM^{cAbl} and nSH3:PRM^{NS1}, the calculated and experimentally measured S^2 for the nSH3 domain overall agree (Fig. 3.3). Regions with low S^2 are in flexible loops and terminal ends that also show high root-mean-square fluctuation (RMSF) of C_α atoms, though the correlation between S^2 and C_α RMSF is weak (Fig. 3.3 *vs.* 3.4). The standard deviation in calculated S^2 is generally larger for flexible regions (lower S^2), but overall it is much smaller than S^2 itself (Fig. 3.3c,d). For most residues in nSH3, both S^2 and C_α RMSF do not change significantly upon ligand binding, which suggests that there is no major change in the flexibility of the backbone [1]. Except for terminal residues, the C_α RMSF is less than 1 Å, which also shows that nSH3 is stably folded and does not undergo any major conformational change when a PRM binds (Fig. 3.4).

nSH3:PRM^{NS1A} and nSH3:PRM^{NS1B} are two different structures of the same complex. Our previous 100 ns MD simulation did not find convergence in their conformational behaviors [42]. With a 500 ns simulation time in the present study, S^2 as well as RMSF of nSH3 in the two

*Part of the contents are reprinted with permissions from 1) Shen, Q., et al. *Biophysical journal*, 114(6), pp.1313-1320, and 2) Shi, J., et al. *Biophysical journal*, 118(10), pp.2502-2512.

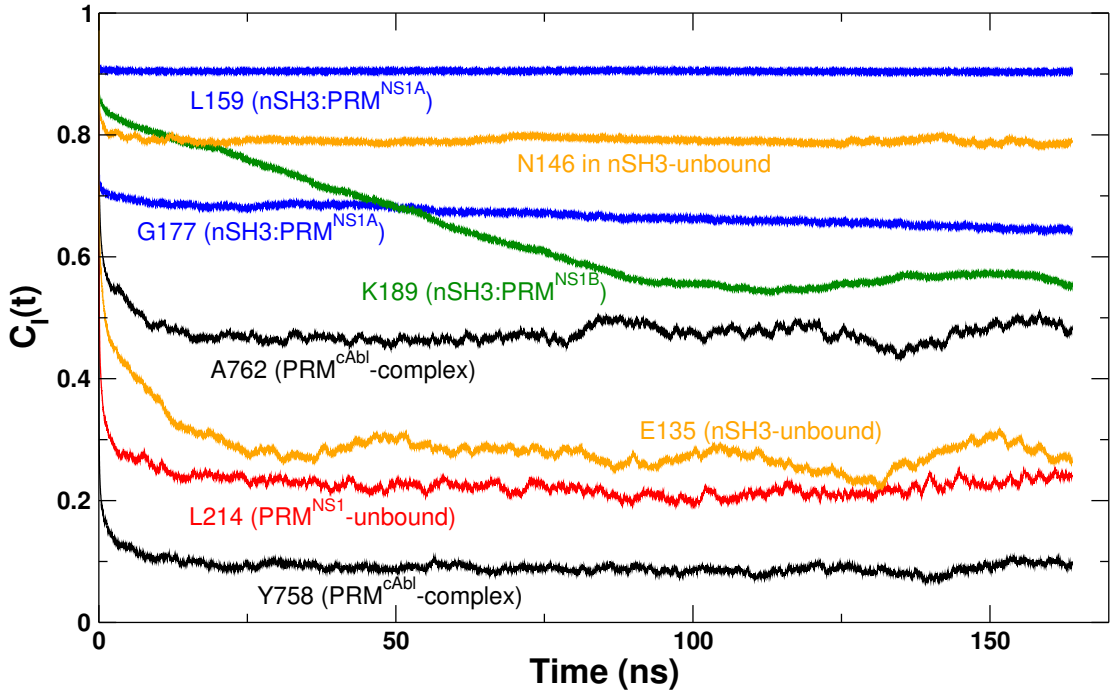

Figure 3.1: Examples of the angular reorientational correlation function, $C_I(t)$. Residues were selected to illustrate various behaviors (from flexible and rigid domains, and long and short relaxation times). Calculations were done in 0.1-ns increments. After the initial decay, $C_I(t)$ approaches the corresponding order parameter S^2 (Fig. 3.3).

structures follow similar profiles (Fig. 3.3b and Fig. 3.4b).

In contrast to nSH3, S^2 for PRM increases significantly upon binding (Fig. 3.5) [1]. The largest increase was around the PxxP motif (P761–P764 in PRM$^{\text{cAbl}}$ and P212–P215 in PRM$^{\text{NS1}}$) that forms hydrophobic contacts with nSH3 [23]. K217–R220 in PRM$^{\text{NS1}}$ also showed a large increase in S^2 as they form electrostatic contacts with negatively charged residues in nSH3 (Fig. 3.5b) [42]. Corresponding to K217 of PRM$^{\text{NS1}}$ is K766 of PRM$^{\text{cAbl}}$ (Fig. 1.1a), which did not change significantly in S^2 since it forms only low-occupancy contacts, the highest being 29% with W169 of nSH3 (Fig. 3.5a). Also note that, S^2 of PRMs in the complex is overall lower than that of nSH3 (Fig 3.3 *vs.* 3.5), indicating that PRMs remain comparatively flexible in the complex. The C_α RMSF of PRMs is mostly over 2 Å when unbound, and the value decreases for all residues upon binding to nSH3, which is another indicator of their conformational changes (Fig. 3.6).

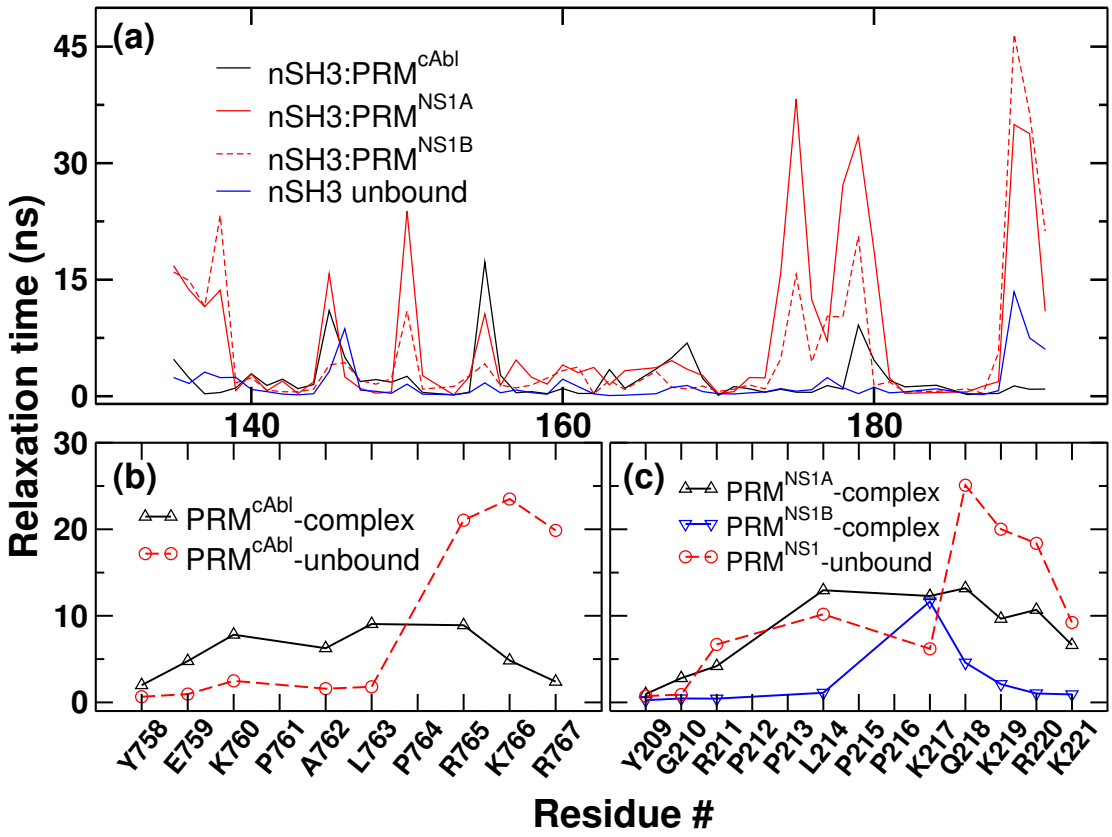

Figure 3.2: Relaxation times obtained from exponential fits of $C_I(t)$. (a) nSH3 domain in all four models. (b) PRM^{cAbl} and (c) PRM^{NS1} in the bound and unbound states.

3.1.2 Fuzzy Interaction

To further elucidate the conformational dynamics of PRM^{NS1} in the bound state, we compared the first 100 ns MD simulations of nSH3:PRM^{NS1A} and nSH3:PRM^{NS1B} (Fig. 3.7). The two simulations did not converge well with each other during the first 100 ns simulation time, indicating that the conversion between the two PRM^{NS1} conformations is slower than 100 ns. This is because the conversion requires large changes in the ϕ/ψ angles of Q218 associated with breakage and reformation of multiple interactions, such as E166(nSH3)-Q218(PRM^{NS1}) and E149(nSH3)-R220(PRM^{NS1}). Consistent with this, R220 formed a stable hydrogen bond with E149(nSH3) with 89% occupancy in the simulation of PRM^{NS1B}, whereas it did not form in the simulation of PRM^{NS1A}. The occupancy of E166(nSH3)-Q218(PRM^{NS1}) interaction was 17% and 64% in the

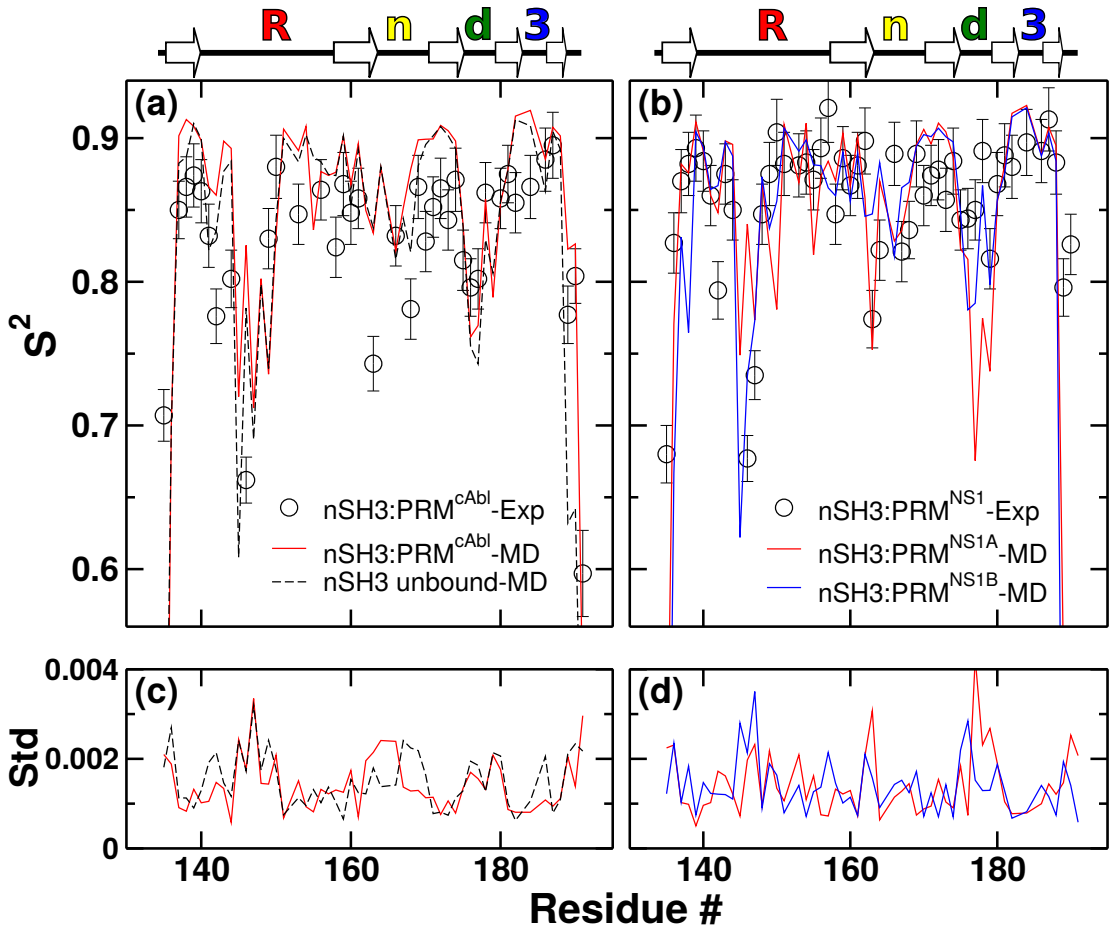

Figure 3.3: Order parameters (a,b) and the corresponding standard deviations (c,d) for nSH3. Standard deviations of experimental S^2 are given as error bars in panels (a) and (b). (a,c) nSH3:PRM^{cAbl} and unbound nSH3. (b,d) nSH3:PRM^{NS1}. Circle: experiment; Lines: simulation. Major subdomains (Fig. 1) are marked above each panel, from left to right: RT-loop ("R"), n-Src loop ("n"), distal loop ("d"), and 3_{10} helix ("3"). Experimental order parameters were estimated from NMR relaxation experiments, provided by Professor Jae-Hyun Cho.

simulation of PRM^{NS1A} and PRM^{NS1B}, respectively.

In order to explore how this fuzzy interaction found in both ends of PRMs can affect the long-range electrostatic interactions, we compared the number of long-range electrostatic interactions between MD trajectories and crystal structures that represent mobile and static states of the bound PRM^{NS1}, respectively. We measured the mean pairwise distances between the acidic residues in the nSH3 domain and positively charged residues in PRM^{NS1} in our MD simulations and in the two crystal structures. In this analysis, K217 in PRM^{NS1} was excluded because it is involved

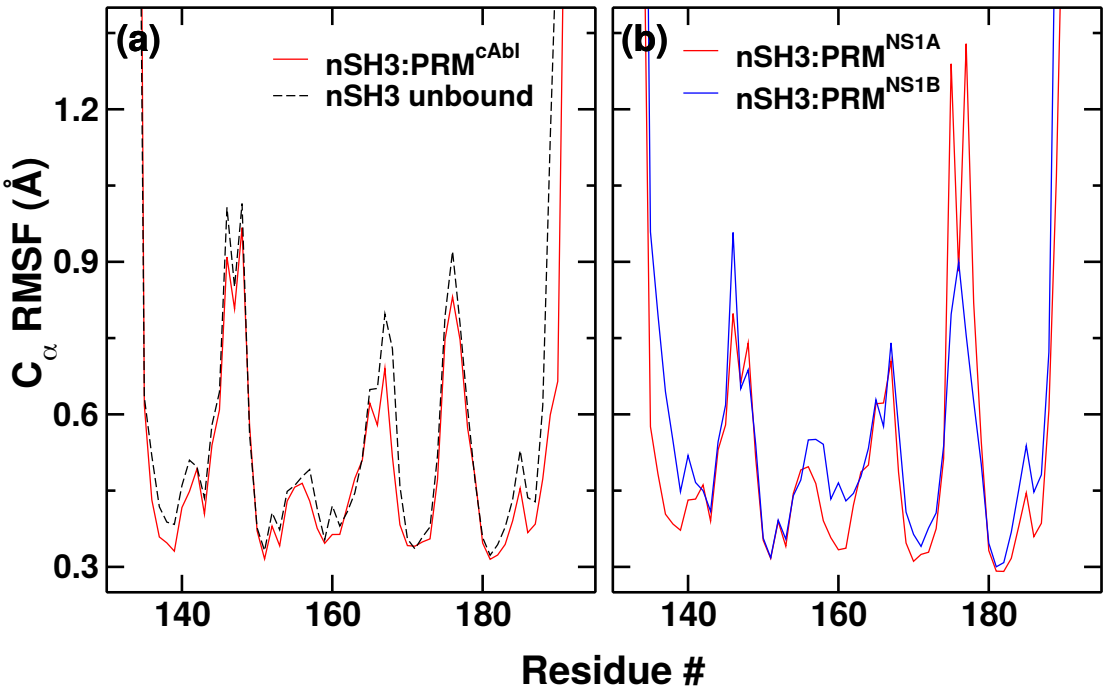

Figure 3.4: RMSF of C_α atoms for the nSH3 domain. (a) nSH3:PRM^{cAbl} and (b) nSH3:PRM^{NS1}.

in well-defined short-range electrostatics, which are also present in other nSH3:PRM complexes. Interestingly, the mean pairwise distances calculated from the MD trajectories were considerably shorter than those calculated from the crystal structures (Table 3.1). One obvious exception in Table 3.1 was the distance between K221 in PRM^{NS1} and D163 in the nSH3 domain, which is shorter in the crystal structures. This is suspected to be related to the lattice contacts around K221. These results indicate that partial disorder of the bound PRM^{NS1} increases the number of long-range electrostatic interactions, which was predicted by Borg's polyelectrostatic model, where multiple charges in an IDP increase its binding affinity to a rigid partner through nonspecific, long-range electrostatic interactions as a binding mechanism of the fuzzy interaction [67].

3.2 Conformational Entropy Change Upon Complexation

We calculated the backbone and side-chain conformational entropies separately or together. For higher accuracy, we used up to MIST3 approximation. The MIST1 entropy (Eq. 2.3) is the sum of entropies calculated for individual DOFs. Correlations among DOFs make the actual entropy

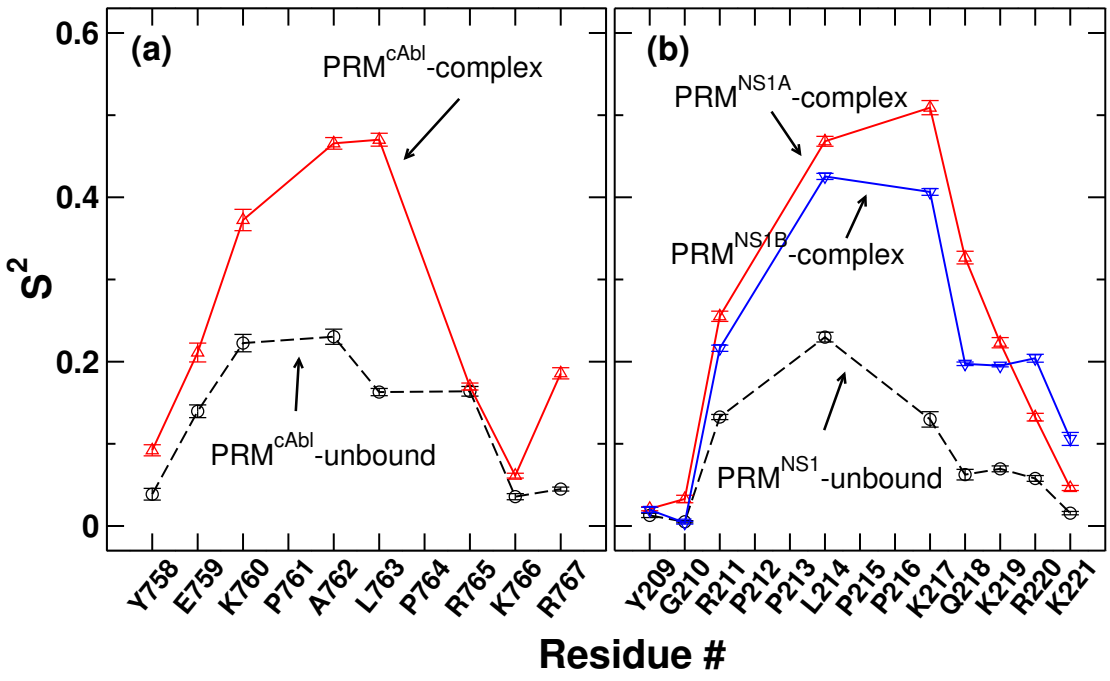

Figure 3.5: Order parameters of ligands. (a) PRM^{cAbl} and (b) PRM^{NS1}. No order parameter was assigned to proline since it does not have a backbone amide hydrogen atom.

Table 3.1: Differences in Average Pairwise Distances between MD and Crystal Structures. Values are calculated as (MD-Crystal). Each pairwise distance is measured between the side chain terminal charged heavy atoms (e.g., charged nitrogen of guanidinium group in arginine and carboxyl oxygen in a glutamate).

		Acidic Residues in the nSH3 Domain					
		D142	D147	E149	D150	D163	E166
PRM^{NS1}	R211	-0.9	—	—	—	—	—
	K219	—	-2.3	-1.3	-2.2	0.9	-0.5
	R220	—	-1.0	-0.4	-0.8	-1.1	-2.2
	K221	—	-3.4	-0.9	-1.5	3.8	-0.4

less than the MIST1 entropy (Eqs. 2.5 and 2.7). In systems studied here, the MIST1 entropy takes up the dominant portion, while MIST2 or MIST3 entropies are slightly reduced (Fig. 3.8). This indicates that correlations among DOFs are not strong. To examine whether using the 100–500 ns interval was adequate for entropy calculation, we calculated the MIST2 entropy for time intervals starting from 100-ns in 20-ns increments. Plateauing of calculated entropies suggests that the 400-

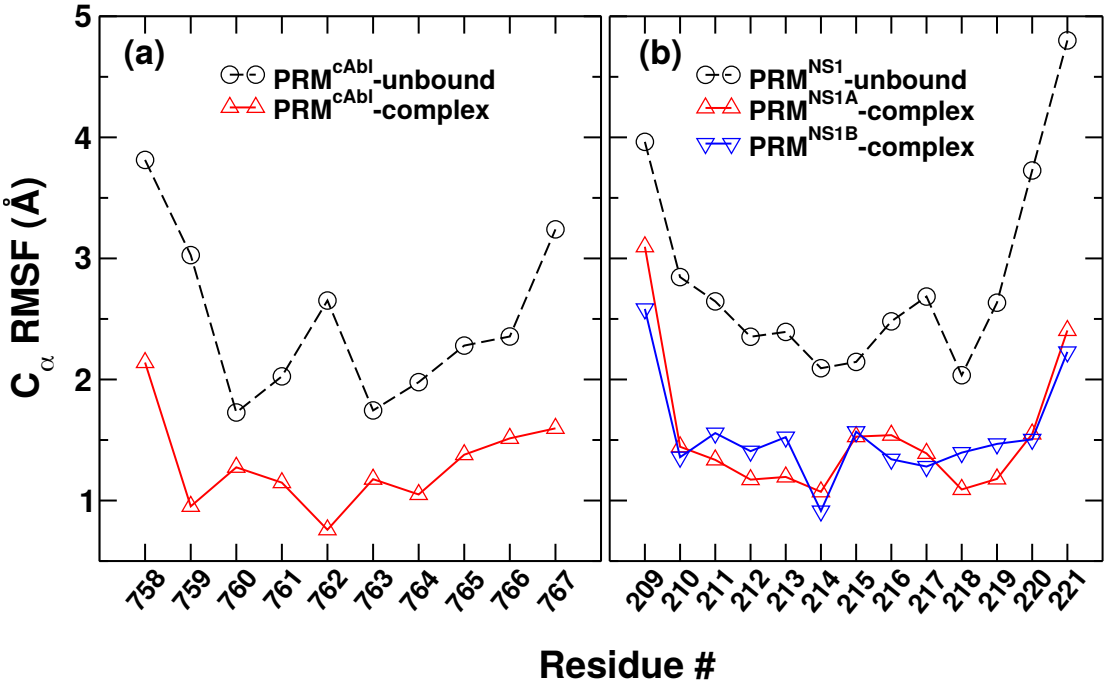

Figure 3.6: RMSF of C_α atoms for PRMs. (a) nSH3:PRM^{cAbl} and (b) nSH3:PRM^{NS1}.

ns duration was reasonable (Fig. 3.9). Dependence of the calculated entropy on the bin size Δ is in Fig. 3.10, which has little impact on relative magnitudes of different entropy terms.

3.2.1 Backbone Entropy Change of nSH3

Previous NMR-based estimates of backbone entropy changes associated with the PRM-SH3 complex formation were -7.0 ± 4.3 cal/(mol·K) [11] and -5 ± 2 cal/(mol·K) [28]. Although direct comparison is difficult since these values are for different SH3:PRM systems, our calculated values (top row of Table 3.2) are similar in magnitude. Comparing between the two complexes in the present study, the backbone entropy change of nSH3 is smaller for the PRM^{cAbl}-bound form than the PRM^{NS1}-bound form.

As previously reported, ligand binding decreases the backbone entropy of the SH3 domain [28, 11]. However, it is the smallest among the entropy changes listed in Table 3.2. This reflects that the stable nSH3 fold is little affected by the binding of a PRM, which was also seen in the analysis of the order parameter and the backbone C_α RMSF (Further discussed in the later section).

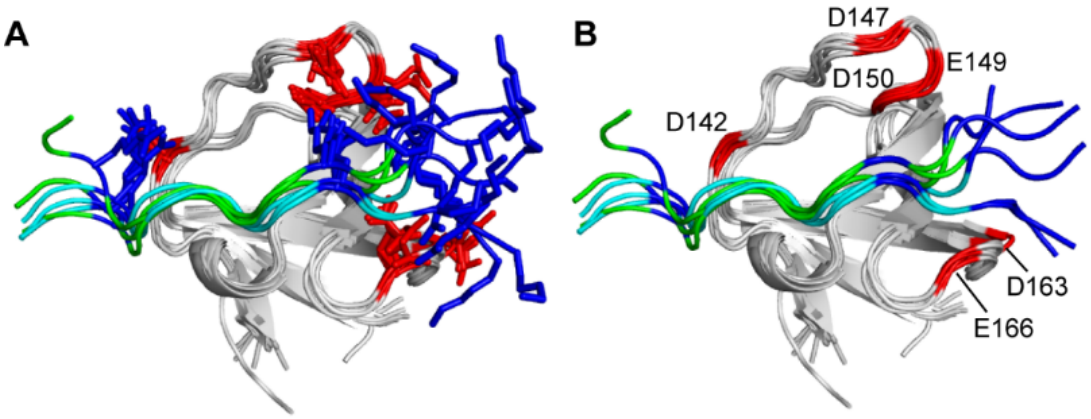

Figure 3.7: MD simulations of the nSH3:PRM^{NS1} complexes. (A) Representative PRM^{NS1} structures were selected from cluster analysis (see Materials and Methods) of the MD trajectories using PRM^{NS1A} (green) and PRM^{NS1B} (cyan). The positively and negatively charged residues are shown in blue and red, respectively. Side chains were omitted in (B) for clarity. Structures were visualized by Professor Jae-Hyun Cho.

Table 3.2: Entropy changes upon complex formation. Numbers are in units of cal/(mol·K). MIST3 was used. Since correlations among DOFs are more extensive in the complex, the net change in entropy (bottom row) is greater in magnitude than the sum of backbone and side-chain entropy changes.

		nSH3:PRM^{cAbl}	nSH3:PRM^{NS1A}	nSH3:PRM^{NS1B}
Backbone	nSH3	-1.80 ± 0.215	-2.38 ± 0.204	-4.47 ± 0.184
	PRM	-8.99 ± 0.041	-6.74 ± 0.057	-7.43 ± 0.088
Side chain	nSH3	-3.83 ± 0.187	-9.13 ± 0.171	-9.82 ± 0.202
	PRM	-3.75 ± 0.114	-7.44 ± 0.161	-8.83 ± 0.154
Net Change		-28.64 ± 0.336	-36.56 ± 0.310	-38.99 ± 0.317

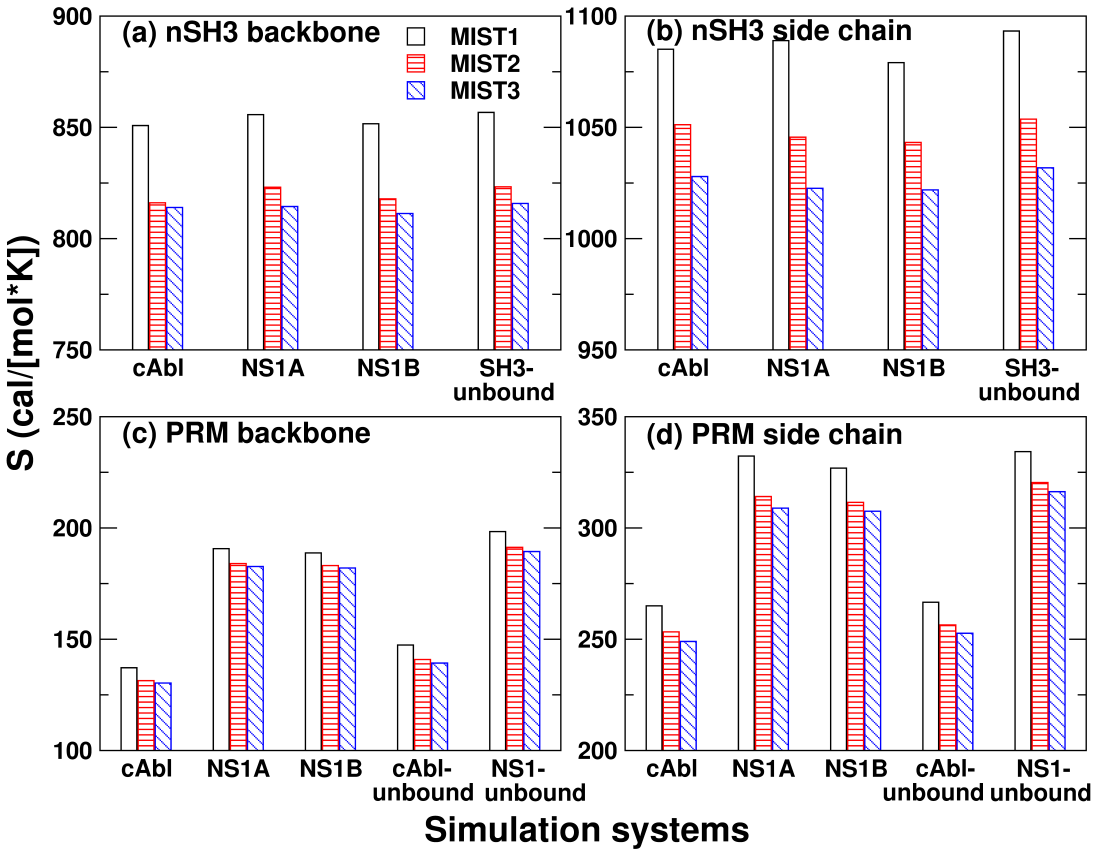

Figure 3.8: Comparison of MIST entropies. (a,b) nSH3 and (c,d) PRM. Respective MIST calculations were performed separately for (a,c) the backbone and (b,d) side chains. Vertical axes in all panels are in the same range, 150 cal/(mol·K), to highlight relative differences. Standard deviations in calculated entropies are at most 0.15 cal/(mol·K).

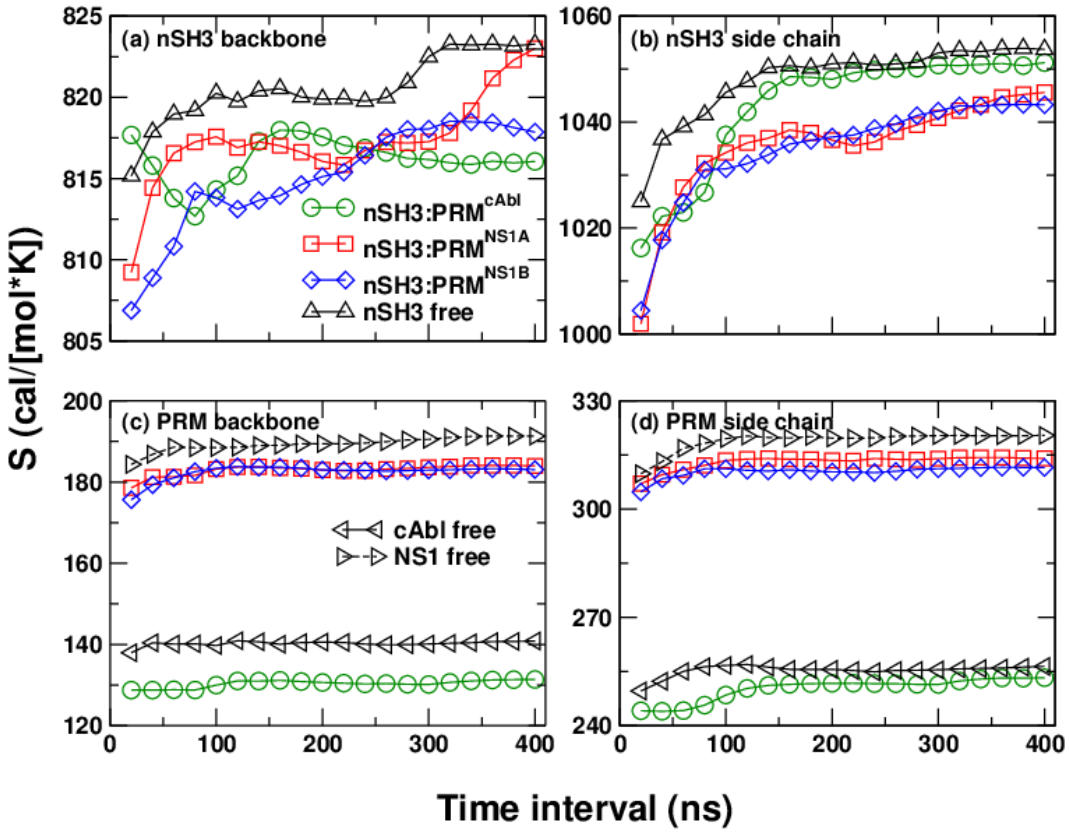

Figure 3.9: Convergence test using the MIST2 entropy. Horizontal axis is the length of the time interval used for entropy calculation. The first data point used 100–120 ns, the second point used 100–140 ns, *etc.* The last point used the whole 100–500-ns interval. (a) Backbone and (b) side-chain entropy of nSH3. (c) Backbone and (d) side-chain entropy of PRMs. For a given category, the relative difference in entropy between systems involving PRM^{cAbl} and PRM^{NS1} is established for time intervals much shorter than 400 ns, except for the backbone entropy of nSH3 where the relative difference starts to emerge for time intervals longer than 300 ns. However, the backbone of nSH3 contributes the least to the entropy change upon ligand binding (Table 3.2). Since the backbone entropy of nSH3 is similar among different systems, the vertical scale in panel (a) is narrower than those in other panels.

3.2.2 Backbone Entropy Change of PRMs

PRMs undergo disorder-order transition upon binding to nSH3, so that their backbone conformational entropy changes are larger compared to nSH3 (Table 3.2, first vs. second rows). Between the two ligands, although PRM^{cAbl} is shorter than PRM^{NS1}, it involves a greater backbone entropy loss (Table 3.2, second row). At the level of individual residues, PRM^{cAbl} shows large backbone

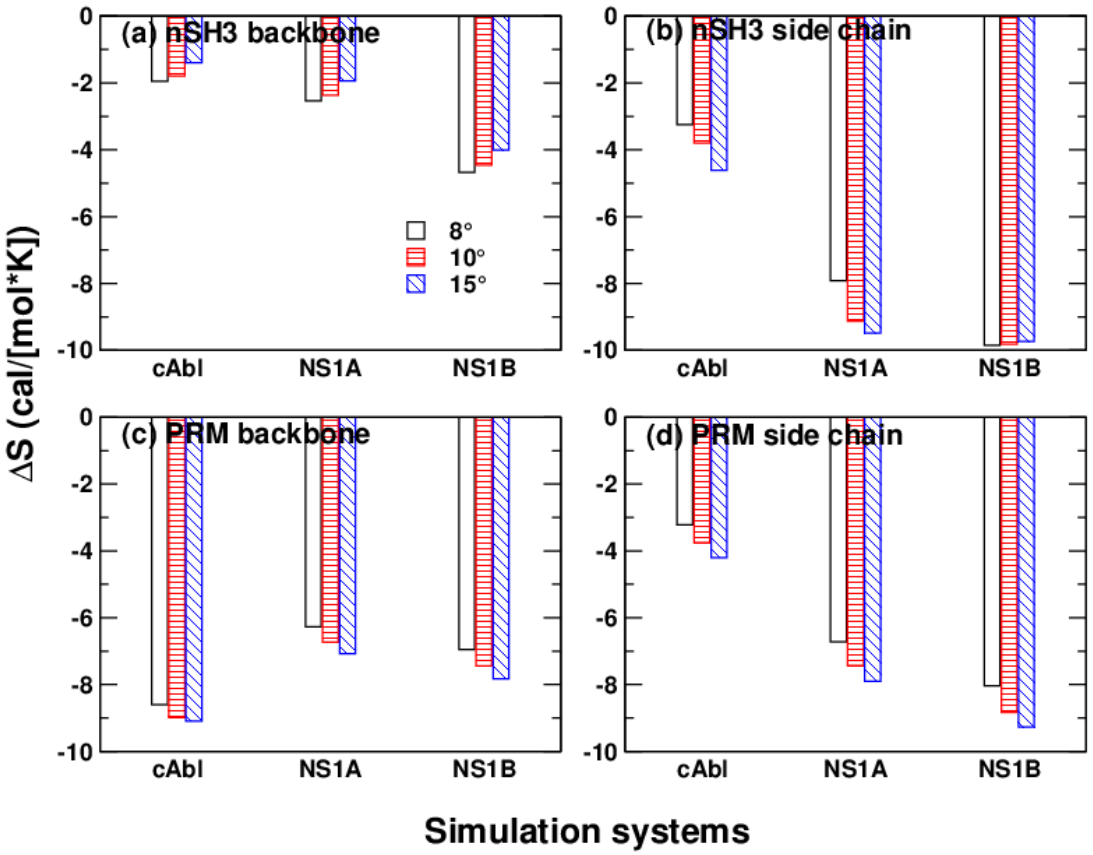

Figure 3.10: Comparison of MIST3 entropies calculated using different bin size Δ. (a,b) nSH3 and (c,d) PRM. MIST calculations were applied separately for (a,c) the backbone and (b,d) side chains. A larger bin size improves statistics but at the expense of losing information. In an extreme case of a single bin for the entire angular range, PDFs will be identical between the bound and unbound states, so that ΔS will be 0. Thus, reduction in the magnitude of ΔS in the nSH3 backbone (panel a) and to a lesser extent for its side chain in nSH3:PRM^{NS1B} (panel b) indicates that PDFs for the corresponding angles are relatively narrowly distributed, which is another indicator that nSH3 is stably folded and does not undergo any major conformational change upon PRM binding. Note that relative magnitude of entropies in each panel does not depend on Δ.

entropy loss in the PxxP motif, especially A762, and also in the last two residues at the C-terminus (Fig. 3.11a). The corresponding residue in PRM^{NS1} is P213 (Fig. 1.1), whose backbone is less flexible. The PxxP motif of PRM^{NS1} is surrounded by additional prolines, which impose a stronger conformational restraint on the backbone. This contributes to the small backbone entropy change upon binding. Furthermore, the two C-terminal residues of PRM^{NS1} show little changes in the backbone entropy upon binding (Fig. 3.11b), suggesting that they remain mobile. The first two

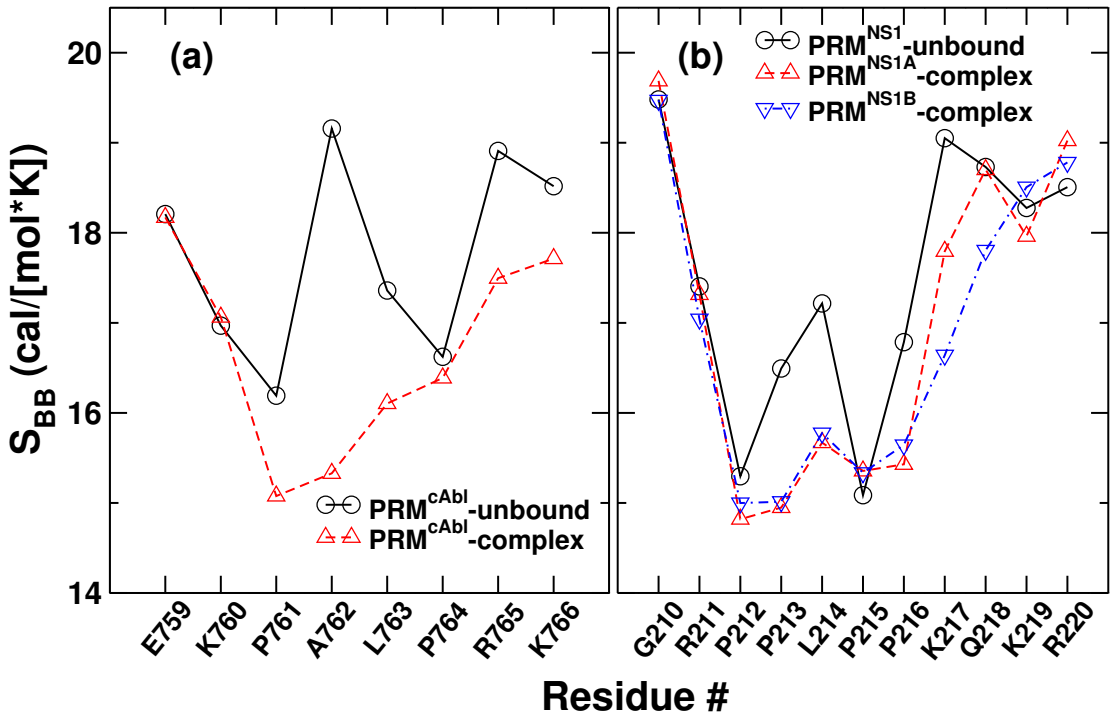

Figure 3.11: Per-residue backbone entropy of PRM in the complex and unbound states. (a) PRM^{cAbl}. (b) PRM^{NS1}.

N-terminal residues in both PRMs also remain mobile.

3.2.3 Side-chain Entropy Change in nSH3

For nSH3, side-chain entropy changes are 2.1 to 3.8 times greater than backbone entropy changes (first vs. third rows in Table 3.2). This is due to the restriction of the side chain motion upon binding of a PRM. However, residue-level analysis revealed that certain side chains gain entropy upon binding (Fig. 3.12). Side-chain entropies of R138, K155, and R162 in nSH3 increased substantially when complexed with PRM^{cAbl} and to a lesser extent with PRM^{NS1} (Fig. 3.12a,b). Y190 at the C-terminus of nSH3 is located next to R138 and its side-chain entropy increased in both of the two SH3:PRM^{NS1} systems (Fig. 3.12b,c). These residues are bulky and surface-exposed, and none of them interacts directly with PRMs. Only K155 formed contact with the N-terminal tyrosine of the two PRMs (Fig. 1.1a), but with less than 1% occupancy. Increase in the side-chain entropy of these residues should thus be an indirect effect of PRM binding. Supporting

this idea is a previous study reporting strain propagation across the entire c-Src SH3 domain as a compensatory response to ligand binding [153]. A detailed analysis of the intra-nSH3 contacts revealed that the side-chain entropy increase in remote residues is mostly a result of propagation of changes in lateral contacts of surface residues upon ligand binding (see next section for details). Breakage of hydrogen bonds plays a significant role in increasing the side chain entropy, although rearrangements of lateral nonpolar contacts are also involved. This explains why side-chain entropy increase occurs for large charged or polar residues on the surface of the protein. An additional possibility is the perturbation of the surface hydration structure upon ligand binding that in turn can affect the mobility of surface-exposed side chains of nSH3 [34]. Further studies are needed to elucidate the allosteric effect of ligand binding on the side chain motion.

Residues that showed substantial decrease in side-chain entropy are mainly in the ligand-binding pocket. The largest side-chain entropy loss in nSH3 among all three systems was in two nonpolar residues, F141 and W169 (Fig. 3.12). Both are highly conserved across SH3 domains [154, 155]. They form high-occupancy nonpolar contacts with the PRM (Table 3.3), which restricts their side chain motion. E149 in the RT-loop and E166 in the n-Src loop that form hydrogen bonds with the C-terminal positively charged residues of PRM (Fig. 3.12d) also lose side-chain entropy substantially (for simplicity, we call a salt bridge involving hydrogen atom also as a hydrogen bond; Table 3.3) . Other residues that form high-occupancy contacts with the PRM (Table 3.3) also contribute to the side-chain entropy loss but to a less extent.

Away from the binding interface, D174 in nSH3:PRM$^{\mathrm{NS1A}}$ and K178 in nSH3:PRM$^{\mathrm{NS1B}}$ in the distal loop also showed notable side-chain entropy loss (Fig. 3.12b–d). This may be a long-range effect, similar to the entropy increase in distal residues upon ligand binding. A previous NMR study of ligand binding to c-Src SH3 domain showed that changes in motion occurs both near and away from the ligand-binding interface [28]. Both D174 and K178 form hydrogen bonds with the nearby E176 in the distal loop with different contact occupancies. In isolated nSH3, their contact occupancies are 50.7% (D174-E176) and 78.7% (K178–E176). In nSH3:PRM$^{\mathrm{cAbl}}$ these occupancies increase by less than 4.6% upon complex formation, which is consistent with their moderate

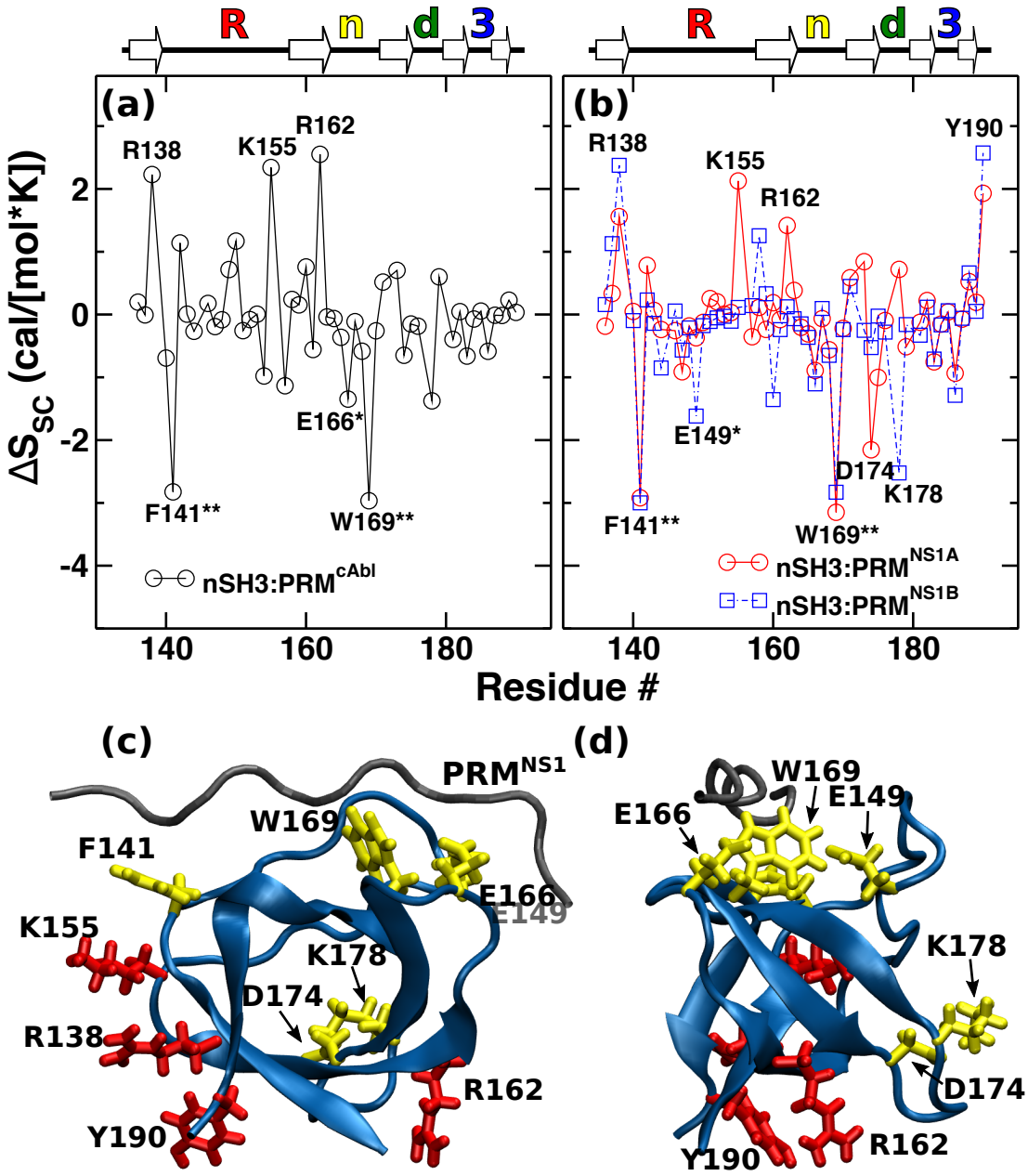

Figure 3.12: Per-residue side-chain entropy change of nSH3 upon ligand binding. Positive value means higher entropy in the complexed state. (a) nSH3:PRM^{cAbl} and (b) nSH3:PRM^{NS1}. Major subdomains (Fig. 1.1) are marked above each panel: RT-loop ("R"), n-Src loop ("n"), distal loop ("d"), and 3_{10} helix ("3"). (c,d) Locations of residues in nSH3 marked in panels (a) and (b). Red/yellow: per-residue side-chain entropy increased/decreased. (c) Side view, viewed from below in Fig. 1.1b. (d) View from the right of panel (c).

side-chain entropy changes (Fig. 3.12a). For nSH3:PRM^{NS1A}, the occupancy of D174-E176 increases to 90.0% whereas for K178-E176 it decreases to 8.7% (*cf.*, Table 3.5), which contributes

Table 3.3: List of high-occupancy contacts and side-chain entropy changes. Hydrogen bond (H-bond) includes salt bridges. Residues of nSH3 are marked with superscripts with the corresponding subdomains (^{3}: 3_{10} helix, ^{R}: RT-loop, ^{n}: n-Src loop). Side-chain entropy change of each residue upon complex formation (in cal/[mol·K]) is in parentheses.

	H-bond			Nonpolar contact		
	nSH3	PRM	Occupancy	nSH3	PRM	Occupancy
nSH3:PRM^{cAbl}	E166^{n} (−1.35)	R767 (−3.48)	0.987	F141^{R}(−2.82)	K760 (−0.45)	0.930
	Y186^{3} (−0.59)	P761 (+0.02)	0.939	F141^{R}(−2.82)	P761 (+0.02)	0.769
				F143^{R}(+0.00)	L763 (−0.63)	0.928
				Q168^{n} (−0.58)	P764 (+0.24)	0.949
				W169^{n} (−2.97)	L763 (−0.63)	0.840
				P183^{3} (−0.66)	L763 (−0.63)	0.876
				P185^{3} (+0.05)	P764 (+0.24)	0.772
				Y186^{3} (−0.59)	P761 (+0.02)	0.876
nSH3:PRM^{NS1A}	D142^{R}(+0.78)	R211 (−1.26)	0.839	F141^{R}(−2.92)	R211 (−1.26)	0.900
	D147^{R}(−0.91)	K217 (−1.41)	0.839	F141^{R}(−2.92)	P212 (−0.03)	0.785
	D150^{R}(−0.12)	K217 (−1.41)	0.943	F143^{R}(+0.06)	L214 (−0.84)	0.955
	E166^{n} (−0.89)	R220 (−4.50)	0.968	Q168^{n} (−0.56)	P215 (+0.93)	0.970
	Y186^{3} (−0.94)	P212 (−0.03)	0.978	W169^{n} (−3.15)	L214 (−0.84)	0.897
				W169^{n} (−3.15)	K217 (−1.41)	0.951
				P183^{3} (−0.75)	L214 (−0.84)	0.886
				P185^{3} (+0.04)	P215 (+0.93)	0.870
				Y186^{3} (−0.94)	R211 (−1.26)	0.831
				Y186^{3} (−0.94)	P212 (−0.03)	0.926
				Y186^{3} (−0.94)	L214 (−0.84)	0.754
nSH3:PRM^{NS1B}	D142^{R}(+0.23)	R211 (−2.11)	0.721	F141^{R}(−3.00)	R211 (−2.11)	0.928
	D147^{R}(−1.57)	K217 (−4.47)	0.933	F141^{R}(−3.00)	P212 (+0.08)	0.746
	E149^{R}(−0.37)	K217 (−4.47)	0.825	F143^{R}(−0.14)	L214 (−1.00)	0.949
	E149^{R}(−0.37)	R220 (−1.71)	0.799	Q168^{n} (−0.65)	P215 (+0.94)	0.939
	D150^{R}(−0.18)	K217 (−4.47)	0.985	W169^{n} (−2.83)	L214 (−1.00)	0.872
	E166^{n} (−1.10)	R220 (−1.71)	0.783	W169^{n} (−2.83)	K217 (−4.47)	0.920
	Y186^{3} (−1.29)	P212 (+0.08)	0.979	P183^{3} (−0.71)	L214 (−1.00)	0.914
				P185^{3} (+0.06)	P215 (+0.94)	0.826
				Y186^{3} (−1.29)	R211 (−2.11)	0.858
				Y186^{3} (−1.29)	P212 (+0.08)	0.926
				Y186^{3} (−1.29)	L214 (−1.00)	0.741

to the relatively large decrease in side-chain entropy of D174 (Fig. 3.12b). However, E176 did not experience any corresponding side-chain entropy loss. For nSH3:PRM^{NS1B}, the D174-E176 bond had nearly the same occupancy in the complex, 55.5%, and the occupancy of K178-E176 increased to 88.1%, again in line with the side-chain entropy loss of K178. Thus, the contact occupancy and side-chain entropy changes upon complex formation are somewhat correlated when the occupancy increases to a high value, which may be due to a restriction in side chain motion. However, two residues forming a high-occupancy contact do not experience comparable side-chain entropy loss, as seen for E176.

3.2.4 Allosteric Increase in Side-Chain Entropy

To elucidate how ligand binding leads to increase in the side-chain entropy of certain residues distal to the ligand-binding interface, we examined differences in intra-nSH3 contacts between the liganded and unliganded states. In Fig. 3.12, two distal residues, D174 and K178 lose side-chain entropy in nSH3:PRM$^{\mathrm{NS1A}}$ and nSH3:PRM$^{\mathrm{NS1B}}$, which are explained in the previous section in terms of changes in contacts with other residues. Here we consider residues that experience increase in side-chain entropy.

We focused on intra-nSH3 contacts with occupancy greater than 50% in either unliganded or liganded states, and the difference in occupancy between the two states larger than 10%. Among the selected contacts, we found that hydrogen bond (H-bond) occupancy and side-chain entropy change oppositely, where increase (decrease) in H-bond occupancy corresponds to decrease (increase) in side-chain entropy. (Tables 3.4–3.6). Though there is little quantitative correlation, this trend is more evident for residues with relatively large side-chain entropy changes ($|\Delta S_{\mathrm{SC}}| > 1$ cal/[mol·K]). For a given residue pair, side-chain entropy changes are not symmetric since the two residues may respectively form additional contacts with other residues.

Table 3.4: Major intra-nSH3 H-bond occupancy differences between unliganded nSH3 and nSH3:PRM$^{\mathrm{cAbl}}$. Only H-bond pairs with average occupancy of the unliganded or the liganded system greater than 50%, and the occupancy difference greater than 10% are shown. Bottom row: per-residue side-chain entropy change upon complex formation (cal/[mol·K]).

	Y136	D142	K155	R162	E173	R179	Y190
Y136	-	-	-	-	-	-	0.165
D142	-	-	-0.377	-	-	-	-
K155	-	-0.377	-	-	-	-	-
R162	-	-	-	-	-0.288	-	-
E173	-	-	-	-0.288	-	-0.127	-
R179	-	-	-	-	-0.127	-	-
Y190	0.165	-	-	-	-	-	-
ΔS_{SC}	0.198	1.139	2.341	2.552	0.703	0.600	0.035

Table 3.5: Major intra-nSH3 H-bond occupancy differences between unliganded nSH3 and nSH3:PRM^{NS1A}. See Table 3.4 for explanation.

	Y136	D142	E149	K155	E173	D174	E176	K178	R179	P183	Y186	Y190
Y136	-	-	-	-	-	-	-	-	-	-	-	-0.121
D142	-	-	-	-0.429	-	-	-	-	-	-	-	-
E149	-	-	-	-	-	-	-	-	0.595	-	-	-
K155	-	-0.429	-	-	-	-	-	-	-	-	-	-
E173	-	-	-	-	-	-	-	-	-0.740	-	-	-
D174	-	-	-	-	-	-	0.393	-	-	-	-	-
E176	-	-	-	-	-	0.393	-	-0.700	-	-	-	-
K178	-	-	-	-	-	-	-0.700	-	-	-	-	-
R179	-	-	0.595	-	-0.740	-	-	-	-	-	-	-
P183	-	-	-	-	-	-	-	-	-	-	0.156	-
Y186	-	-	-	-	-	-	-	-	-	0.156	-	-
Y190	-0.121	-	-	-	-	-	-	-	-	-	-	-
ΔS_{SC}	-0.185	0.783	-1.618	2.126	0.841	-2.154	-0.095	0.718	-0.513	-0.754	-0.935	1.931

Table 3.6: Major intra-nSH3 H-bond occupancy differences between unliganded nSH3 and nSH3:PRM^{NS1B}. See Table 3.4 for explanation.

	Y136	P183	Y186	Y190
Y136	-	-	-	-0.305
P183	-	-	0.144	-
Y186	-	0.144	-	-
Y190	-0.305	-	-	-
ΔS_{SC}	0.158	-0.708	-1.287	2.573

In comparison to H-bonds, there were more intra-nSH3 nonpolar contacts (23 to 31) that met our selection criteria. However, we did not find any clear relation between the nonpolar contact occupancy and the side-chain entropy changes. When a nonpolar contact breaks, the residues may form different nonpolar contacts due to the hydrophobic effect. In contrast, polar or charged groups on the surface of the protein may associate with water molecules after breaking H-bonds, thereby increasing side-chain entropy. Indeed, all the distal residues that had large side-chain entropy changes were surface-exposed polar or charged residues (Fig. 3.12c,d). Yet, some of them may also involve nonpolar contacts via the nonpolar part of their side chains, as explained below.

Among the residues in Tables 3.4–3.6, we focus on those away from the ligand-binding interface and also had large side-chain entropy increase: K155 and R162 in nSH3:PRM^{cAbl} (Table 3.4), K155 and Y190 in nSH3:PRM^{NS1A} (Table 3.5), and Y190 in nSH3:PRM^{NS1B} (Table 3.6). In an iso-

lated nSH3, both the backbone and the side chain of K155 form stable H-bonds with the side chain of D142 (Fig. 3.13a). Upon ligand binding, PRM residues K760 in nSH3:PRM^{cAbl} and R211 in nSH3:PRM^{NS1A} form H-bonds with D142, which impedes its H-bond with K155 (Fig. 3.13b,c). In contrast, the D142–K155 H-bond in nSH3:PRM^{NS1B} is less perturbed as R211 of the PRM mostly forms a H-bond with the backbone carbonyl oxygen of nSH3 D142 (Fig. 3.13c). Thus, increase in the side-chain entropy of K155 in nSH3:PRM^{cAbl} and nSH3:PRM^{NS1A} is due to the loss of its H-bond with D142 upon ligand binding.

For the Y136–Y190 H-bond that had occupancy decrease in nSH3:PRM^{NS1A} and nSH3:PRM^{NS1B} (Tables 3.5 and 3.6), the residues are respectively at the N- and C-termini of nSH3 (Fig. 3.13f). Between the two, Y136 also formed high-occupancy nonpolar contacts with other residues, *i.e.*, R160 (100% occupancy in both nSH3:PRM^{NS1A} and nSH3:PRM^{NS1B}) and I158 (90.7% in nSH3:PRM^{NS1A} and 93.6% in nSH3:PRM^{NS1B}). In comparison, other contacts that Y190 formed were lower in occupancy. They were also nonpolar, with the highest occupancy being 72.3% (R138 in nSH3:PRM^{NS1A}; the contact broke at 413 ns and did not form again) and 52.9% (V137 in nSH3:PRM^{NS1B}). This suggests that, when the Y136–Y190 H-bond breaks, Y190 experiences a greater increase in side-chain entropy.

For R162, the increase in its side-chain entropy upon ligand binding is the largest in nSH3:PRM^{cAbl} (Fig. 3.12a,b). In the unliganded nSH3, we found that R162 is sandwiched between R160 and R179 (Fig. 3.13e). In the liganded state, R179 occasionally detaches and points to the negatively charged residues in the RT-loop (Fig. 3.13e, arrow). This makes R162 more mobile. Among the liganded states, the R160-R162-R179 sandwich was the most preserved in nSH3:PRM^{NS1B}, nearly the same as in the unliganded nSH3 (we measured the sandwich state by the distances between the guanidinium carbon atoms in R160–R162 and R162–R179; Fig. 3.13e). This is consistent with the negligible side-chain entropy change of R162 in nSH3:PRM^{NS1B} (Fig. 3.12b). For nSH3:PRM^{cAbl}, R179 flips less compared to nSH3:PRM^{NS1A} but the greater R162–E173 H-bond occupancy decrease (Table 3.4) may have contributed to the larger side-chain entropy increase of R162.

Other than K155, R162, and Y190 explained above, R138 had side-chain entropy increase in

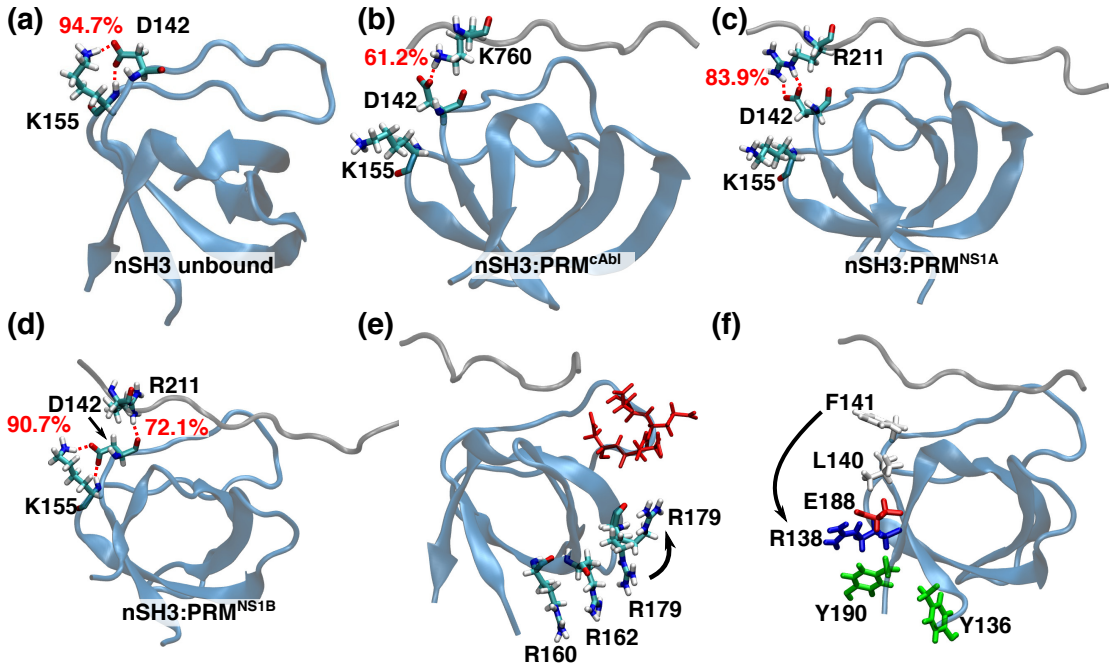

Figure 3.13: Illustration of the allosteric effect of PRM binding (*cf.*, Fig. 3.12). Orientations of panels are similar to that of Fig. 3.12c. (a–d) Increase in side-chain entropy of K155. H-bond (red dashed line) and the corresponding occupancy are shown. Occupancy is residue-to-residue based, regardless of the number of H-bonds between the two. (a) Unliganded nSH3. (b–d) Liganded cases where the K155–D142 H-bond is broken in (b,c). (e) Increase in side-chain entropy of R162. In unliganded nSH3, it is sandwiched between R160 and R179. R179 occasionally flips towards the negatively charged residues in the RT-loop (red sticks), which occurs more often in nSH3:PRM$^{\rm cAbl}$ and nSH3:PRM$^{\rm NS1A}$. The distance between the guanidinium carbon atoms of R160 and R162 is 4.20 Å–5.06 Å in all cases. Between R162 and R179, it is 4.89±0.56 Å (unliganded nSH3; avg±std), 5.26±2.11 Å (nSH3:PRM$^{\rm cAbl}$), 8.23±2.72 Å (nSH3:PRM$^{\rm NS1A}$), and 4.82±0.62 Å (nSH3:PRM$^{\rm NS1B}$). In nSH3:PRM$^{\rm cAbl}$ and nSH3:PRM$^{\rm NS1A}$ the distance is larger and also fluctuates more. In nSH3:PRM$^{\rm NS1B}$, the distance is nearly the same as for the unliganded nSH3. (f) Possible pathway of the propagation of changes upon ligand binding, starting from the entropy hotspot F141, to R138 (arrow). The Y136–Y190 pair flanking the R138–E188 pair is also shown, suggesting that ligand binding may have an allosteric effect on the mobility of residues in these terminal β-strands.

all complexes but it is absent in Tables 3.4–3.6. R138 forms H-bond with E188 with greater than 99.9% occupancy in all cases (unliganded and liganded). Hence, its increase in side-chain entropy is not due to breakage of any contact upon ligand binding. A closer examination of surface contacts shows that E188 forms nonpolar contact with L140 (over 98% occupancy in all cases), which in turn forms a nonpolar contact with F141 in the ligand binding pocket of nSH3 Fig. 3.13f). Note

that F141 is one of the two main entropy hotspots of nSH3 that experience the largest side-chain entropy decrease upon ligand binding (Fig. 3.12). In the nSH3:PRM complex the L140–F141 nonpolar contact occupancy also increases by 14%–18%. Akin to a domino effect, this in turn may have altered the positioning of E188, eventually leading to increase in the side-chain entropy of R138. (Fig. 3.13f, arrow) We also note that R138–E188 and Y136–Y190 pairs are located at the N- and C-terminal β-sheet of nSH3 (Fig. 3.13f). This suggests a possibility that the increase in side-chain entropies of R138 and Y190 in the terminal region may have originated from the change in F141 in the ligand-binding pocket. Taken together, side-chain entropy changes in remote residues appear to be caused by propagation of arrangements in lateral contacts between surface residues upon ligand binding. However, since these residues are either charged or polar, changes in surface hydration may also play an important role for the allosteric effect. Further studies are needed to elucidate the robustness of the allosteric information propagation upon ligand binding.

3.2.5 Side-chain entropy change in PRMs

Similar to nSH3, residues in PRMs that showed large side-chain entropy loss (marked by stars in Fig. 3.14) formed high-occupancy contacts with nSH3 (Table 3.3), though the converse does not necessarily hold. Contact analysis suggests a different mechanism underlying the entropy changes between nSH3 and PRMs. For nSH3, the largest side-chain entropy losses were by residues that form nonpolar contacts, *i.e.*, F141 and W169. But their binding partners in PRM do not experience correspondingly large entropy loss. For example, L763 in PRM^{cAbl} and L214 in PRM^{NS1} contact W169, but their side-chain entropy changes are not significant (Table 3.3). When a PRM binds, F141 and W169 of nSH3 become buried in the interface (Fig. 3.12c) whereas their nonpolar contact residues in PRM are still partially exposed, which allows more room for motion. For PRM, on the other hand, all residues that had more than 2 cal/(mol·K) side-chain entropy loss formed high-occupancy hydrogen bonds with nSH3 (R767 in PRM^{cAbl}, R220 in PRM^{NS1A}, and K217 in PRM^{NS1B}; single star in Fig. 3.14, and Table 3.3). Compared to nonpolar contacts, hydrogen bonds are more directional, where the acceptor-hydrogen-donor angle is greater than 120° [156, 157]. The directional constraint can suppress side chain motion more effectively than nonpolar contacts

do.

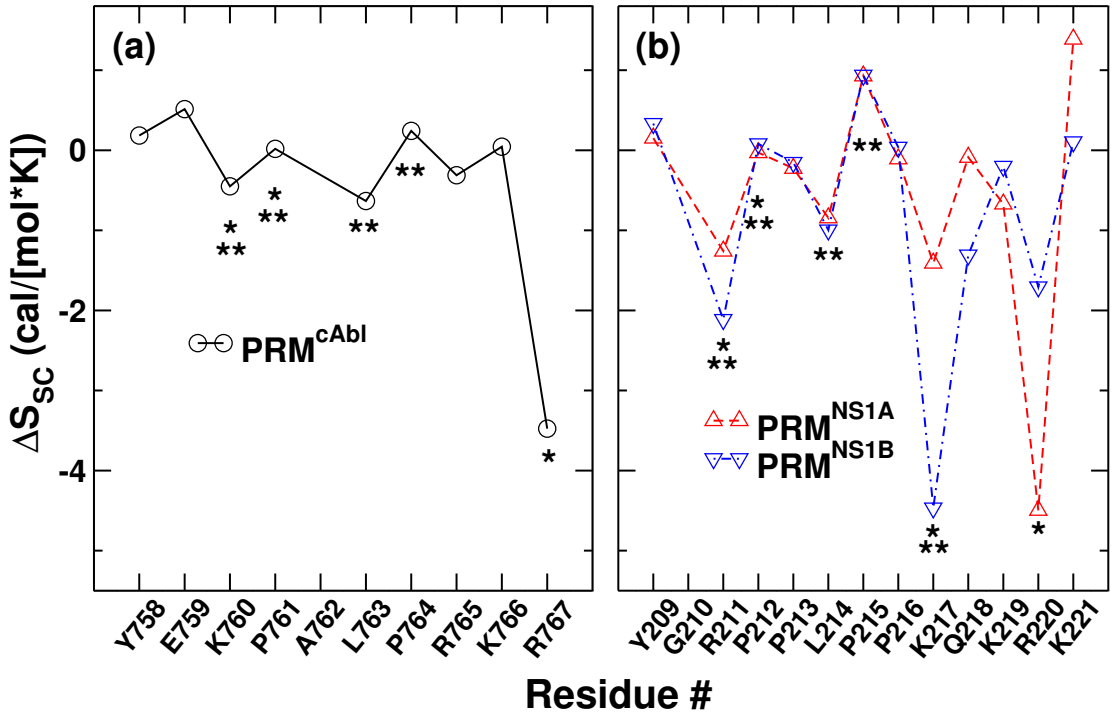

Figure 3.14: Per-residue side-chain entropy change of PRMs upon binding to nSH3. (a) nSH3:PRM^{cAbl} and (b) nSH3:PRM^{NS1}. Residues that form high-occupancy contacts with nSH3 (Table 3.3) are marked by stars. Hydrogen bond: one star. Nonpolar: two stars.

3.2.6 Net Entropy Change

When entropy change upon ligand binding is calculated for the whole system, the correlation effect between nSH3 and PRM additionally lower the entropy of the complex relative to the unbound states. This causes the net entropy change of the whole system to be greater than the sum of individual entropy changes (Table 3.2, last vs. other rows). Between nSH3:PRM^{cAbl} and nSH3:PRM^{NS1}, the latter involves a greater change in side-chain entropy. Although this is partly due to PRM^{NS1} being longer than PRM^{cAbl}, the former also involves a larger per-residue entropy loss (Fig. 3.14).

The net entropy change in ligand binding in nSH3:PRM^{NS1} is 7.9–10.4 cal/(mol·K) greater than

that of nSH3:PRM^{cAbl} (Table 3.2, last row). At 300 K, this corresponds to 2.4–3.1 kcal/mol, which is comparable to the free energy of a hydrogen bond [158]. With nSH3:PRM^{NS1} also forming more contacts (Table 3.3), the emerging picture is that PRM^{NS1} achieves higher binding affinity to nSH3 where the favorable enthalpy change afforded by the extensive contact formation is greater than the entropy loss, which is consistent with its role as a viral peptide [2]. The weaker binding and relatively smaller changes of entropy in PRM^{cAbl} is also consistent with with experiment [1], and it enables easier regulation of the interaction with nSH3 of CrkII.

It is notable that the side chain dynamics is different from the backbone dynamics. The backbone entropy loss of PRM^{cAbl} is larger than that of PRM^{NS1} while the converse holds for the side-chain entropy loss. If only the backbone order parameter or backbone entropy were considered, which has been easier to measure experimentally, one might conclude that PRM^{NS1} is more flexible than PRM^{cAbl} in the bound state. It is thus essential to also consider the side chain dynamics to establish a more complete understanding of the entropy contribution.

3.3 Surface Hydration Contribution of Protein

3.3.1 Hydration Map and Protein Dynamics

We firstly constructed the hydration map for unbound nSH3, nSH3:PRM^{cAbl}, and nSH3:PRM^{NS1} with density cutoff 0.034 Å^{-3} and distance cutoff 4.5 Å. Since the hydration map is built around the reference structure, it allows us to visualize the hydration map and directly compare the differences (Figure 3.15). Here for better illustration, we simply neglect all the cells that have ever been visited by the protein atoms, as discussed in the Method Section.

It turns out, nSH3:PRM^{NS1} elicits lower solvation free energy (ΔG_{solv}) as compared to the nSH3:PRM^{cAbl} by 63.17 kcal/mol, which is consistent with the higher binding affinity of the viral NS1 ligand. Upon ligand binding, both complexes show a slight increase in mean density. But the change is very small (only $\sim$2% difference). In the same time, the number of total cubes (N_{cube} decreases a lot in both cases, which seems to be the major cause of the differences in solvation contribution. The change in N_{cube} is probably caused by losing those high-density cubes in binding

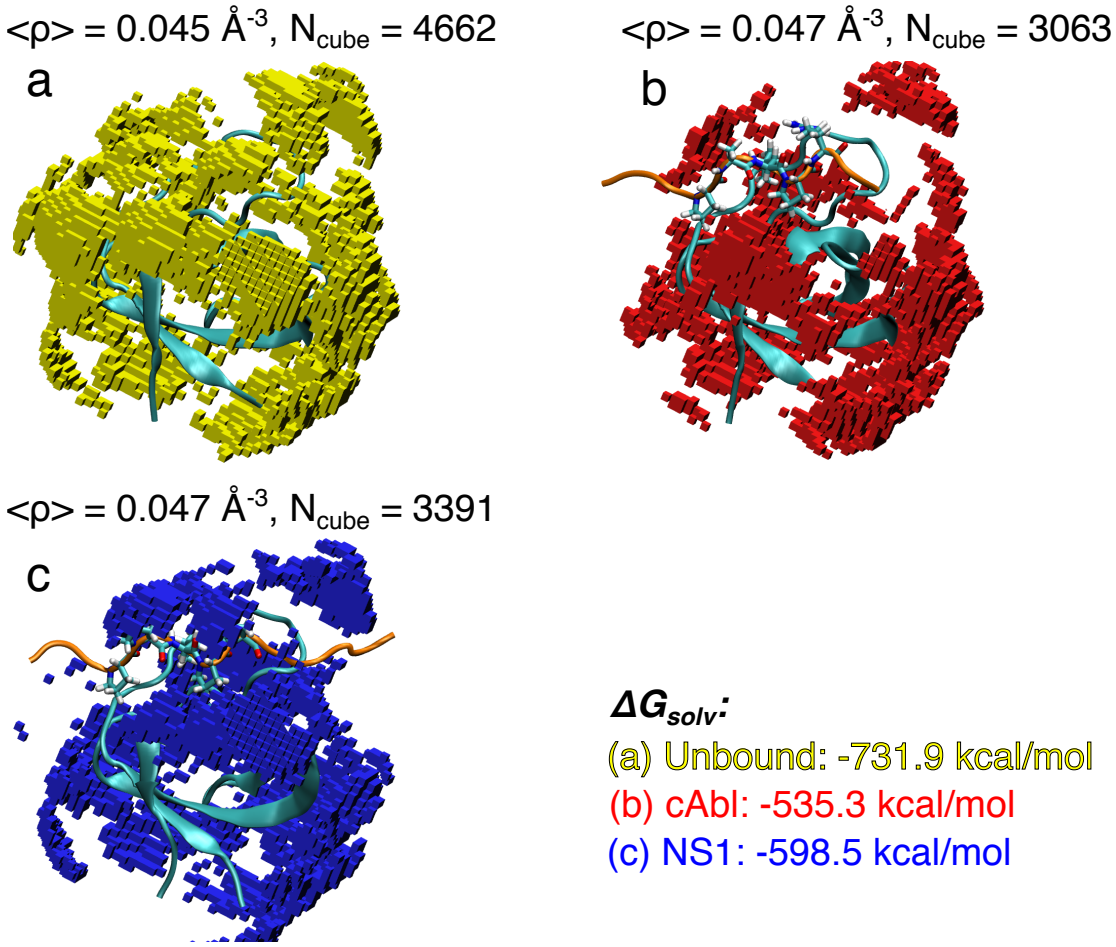

Figure 3.15: Water cubic cells of side length 0.7 Å around the three simulated proteins. Only high-density cubes that meet our criteria are shown. In all three systems, the nSH3 domains are colored with cyan. Both ligands are orange. The central regions are shown in sticks. Yellow, red, and blue cubes are the water hydration cubic cells around the three systems. All three are aligned in the same orientation. (a) Unbound nSH3. (b) nSH3:PRM^{cAbl}. (c) nSH3:PRM^{NS1}.

pocket when the ligands land on the pocket (see Figure 3.15, between a and b/c). Some additional high-density cubes may form around the ligands after binding. However, the additional cubes is much fewer due to the fact that the ligands remain partially intrinsically disordered in the bound state, which makes it difficult to form high-density water cubes around them. This is consistent with our previous study [42]. Comparing SH3:PRM^{NS1} and SH3:PRM^{cAbl}, the lower solvation free energy in PRM^{NS1}-binding comes from the difference in total number of high-density cubes. In fact there are 10% more high-density cubes formed in SH3:PRM^{NS1} than in SH3:PRM^{cAbl}. Inter-

estingly, PRM^{NS1} is a 13-amino acids long ligand (Sequence: YGRPPLPPKQKRK), and PRM^{cAbl} is only 3-amino acid shorter (Sequence: YEKPALPRKR) than PRM^{NS1}.

We further studied what causes the difference upon the binding of two ligands. Since both ligands bind to the same nSH3 domain, the two nSH3 domains in the unbound states are the same. Therefore, our comparison is focused on the bound states of the two complexes. With visualized hydration map, we are able to probe the local difference between the two. Comparing Figure 3.15b and c, more high-density cubes can be found nearby the central region (PPLPP) in PRM^{NS1} than the central region (PALPR) in PRM^{cAbl}. The central region in PRM^{NS1} can form more stable nonpolar contacts with the hydrophobic residues inside the binding pocket of nSH3 (A more detailed contact analysis has been done in our previous work [159] and can also be found in the second study). As a result, it makes it easier for water molecules to pack around the central region in the PRM^{NS1} in the bound state, even though it mainly consists of hydrophobic residues. This packing effect regardless of the particular protein surface groups has already been reported by our previous study [130]. Thus those high-density, undisturbed cubes around the complex surface results in less desolvation penalty.

Figure 3.16a and b show that there are additional cube blocks near the R138 in nSH3:PRM^{NS1}. As our previous work reported [159], the binding of the ligands induced a propagation of changes in side-chain motions, starting from an entropy hotspot F141 to R138. In the case of nSH3:PRM^{cAbl}, more back-and-forth waving of the R138 side-chain is observed. This may be the reason less high-density cubic cells are formed around R138 of nSH3:PRM^{cAbl}.

Figure 3.16c and d show another additional cube blocks near the region between R179 and the RT loop in nSH3:PRM^{NS1}. Similarly, our previous study also found that upon binding of the ligands, the R179 tends to be attracted by the negatively charged residues in the RT-loop, which cause R179 to do occasional flipping between its original orientation and the new orientation that points its side chain towards the RT-loop. Our previous work found that the the R179 in nSH3:PRM^{cAbl} does flipping motion more often, which may yield less high-density cubic cells around.

The two additional water blocks shown in Figure 3.16 suggest that the local solvation contribution differs as a result of side-chain motion difference caused by the long-range effect of binding by different ligands. For the two nSH3:PRM complexes, the solvation free energy is coupled by the side-chain motion of the surface residue in nSH3 domain. As our previous study suggested, the viral ligand can optimize its binding by regulation of long-range effect on the entropy hotspot which locates far away from the binding pocket. Our finding here seems to further suggest that the viral ligand may be able to minimize the desolvation penalty through the side-chain motion of the residues on the receptor protein via propagation of effect from binding site to remote region. The protein surface is known to affect the hydration structure around. Previous works mostly focus on the influence of surface topology, hydrophobicity (chemical groups), or secondary structure [105, 104, 107]. And some previous studies suggested that the flexibility of protein residues can impart its disorderness on nearby water molecules and affect the formation of nearby hydration shell [160, 161]. To our best knowledge, this is the first report of the coupling effect of water hydration structure and the side-chain dynamics of the nearby protein surface.

3.3.2 Contribution of Surface Hydration to the Binding Free Energy

We tested if our solvation free energy calculation can contribute to the total ligand binding free energy. Firstly, we calculated the binding free energy using MM/PBSA. This method has been proved to be capable of distinguishing binders from non-binders among potential SH3-binding ligands [28, 45]. Combining conformational entropic change calculated from our previous work [159], the binding free energy and separate energetic components can be found in Table 3.7.

Our MM/PBSA results show some patterns similar to previous MM/PBSA studies. The electrostatic components (ΔE_{elec} and ΔG_{elec}) are much larger than the non-polar part, but they turn out to be mostly canceled out by each other. As a result, the sum of non-polar contribution (ΔE_{vdW}+ ΔG_{np}) ends up playing a more significant role in determining the bind free energy. Our calculated binding free energies, ΔG_{bind} are -16.01 kcal/mol for cAbl and -20.90 kcal/mol for NS1. Both of them are significantly lower than the experimental value, which is -7.63 kcal/mol for cAbl and -11.36 for NS1 [1, 2]. This level of difference is common among studies applying MM/PBSA to the

Table 3.7: ΔE_{vdW} and ΔE_{elec} are van der Waals and electrostatic interactions differences between complex state and unbound states; ΔG_{np} and ΔG_{elec} are non-polar and electrostatic contribution to the solvation free energy change upon complex formation; -TΔS_{conf} is the conformational entropic contribution to the energetic change, where T is the temperature 300 K, and ΔS_{conf} is entropic difference calculated from backbone and side-chain angular distribution; -TΔS_{rot} is the rotational entropic contribution; -TΔS_{trans} is the translational entropic contribution. This term doesn't contain uncertainty because it is just a function of mass and thus stay constant during simulations. See the Method section for more details; ΔG_{penal}^{expo} is the density-based desolvation penalty calculated from exponential fitting; ΔG_{bind} is the binding free energy change upon ligand binding, calculated for each peptide based on MM/PBSA method; $\Delta G'_{bind}$ is the sum of ΔG_{bind} and ΔG_{penal}^{expo}.

	ΔE_{vdW}	ΔE_{elec}	ΔG_{np}	ΔG_{elec}	-TΔS_{conf}
cAbl	-32.18 ± 6.56	-541.22 ± 162.58	-4.39 ± 0.55	528.17 ± 156.60	8.59 ± 0.10
NS1	-29.79 ± 6.24	-898.85 ± 117.87	-4.93 ± 0.69	875.37 ± 115.28	11.70 ± 0.09
	-TΔS_{rot}	-TΔS_{trans}	ΔG_{penal}^{expo}	ΔG_{bind}	$\Delta G'_{bind}$
cAbl	12.86 ± 0.20	12.16	5.03	-16.01	-10.98
NS1	13.30 ± 0.16	12.32	4.59	-20.90	-16.31

SH3:PRM binding [28, 45]. Some believed the inaccuracy largely arises from the continuum solvent approximation and is difficult to improve [162]. Another possible cause is that current entropy calculation does not cover all the configurational entropy as our conformation entropy (-TΔS_{conf}) only accounts for backbone and side-chain rotational angles. But since the conformational entropy as calculated via rotational angles plays a more important role in protein-protein recognition [71], this term should suffice helping distinguish binding affinities differences among SH3 binders. The calculated binding free energy difference between cAbl and NS1 is $\sim$ 4.89 kcal/mol, which is very comparable to the experimentally determined value 3.73 kcal/mol. It also correctly reflects the much higher binding affinity found for NS1 as the 1918 flu virus uses it to hijack the nSH3 domain in order to suppress the host's immune response [2].

As shown in Table 3.8, the solvation energetic change (ΔG_{solv}) calculated from our density-based water map analysis is quite large, and they change drastically as the resolutions increase from 0.5 Å to 0.9 Å. The change rises from N_{cube} as larger resolutions always result in fewer numbers of cubes. As the result of changing resolution, the desolvation penalty ΔG_{penal} also changes greatly. The exceptionally small N_{cube} in unbound cAbl reveals another issue in our method. The short

ligand cAbl in the unbound state remains intrinsically disorder throughout the simulation while our method requires a reference structure around which the water map can be constructed. Although in our previous section, we showed that our method can reflect the mutual influence between water map and protein dynamics, however, for the sake of building reference structure, a stably folded core in our target protein is needed. Due to the disorderness of unbound cAbl, no stable core can be used for building reference structure. Therefore, the water map around the reference structure can not faithfully reflect the surface hydration structure. This explains why so few cubes is selected for unbound cAbl. Similarly, N_{cube} in the unbound NS1 is probably also underestimated.

So we tentatively fit the ΔG_{solv} as the function of resolution into the exponential equation, $y = a \times e^{-bx} + c$, and find the asymptotic value as x (resolution) $\rightarrow \infty$ (see Fig. 3.17) . The limiting value c is used as the converged energetic change in solvation and can be found in Table 3.8 as ΔG^{expo}_{solv}. The differences in those limiting values of ΔG^{expo}_{solv} between complex and both unbound states are further calculated as limiting desolvation penalty, which are shown in Table 3.7 (ΔG^{expo}_{penal}). Comparing ΔG_{bind} and $\Delta G'_{bind}$ in Table 3.7, it turns out the additional desolvation penalty slightly improved the absolute binding free energy as the updated values are closer to the experimental counterparts. And relatively small desolvation penalty found in NS1 suggests the water surface hydration around NS1 complex contributes to the higher binding affinity found in NS1 as compared with cAbl.

3.3.3 The Downside of Retaining Proteins in Hydration Analysis

A large portion of currently developing methods that try to evaluate solvation contribution in protein folding and recognition requires retaining the protein atoms during the simulation and only samples the water molecules around the region of interest [119, 92, 125]. This may provide insights in those stable hydration sites or the ordered water inside the less mobile binding pockets. However, as shown in our results, for a more general evaluation of thermodynamic contribution from the hydration shell, the protein dynamics may play a role too paramount to be neglected by the simulation with retained protein. Therefore, we tested how retaining protein atoms may affect hydration analysis in our work.

Following the same simulation procedures right before production run starts, we restrained all the protein atoms with a harmonic restraint weight of 50 kcal/mol·Å^2. The stiffness of restraint is in keeping with previous work [125]. Since the proteins are retained, we don't need to worry about sampling of protein conformational motions. So we let each of the five systems undergo 50 ns long simulation. Then our analysis is based on the behavior of hydration water during the 50 ns trajectory.

The result of hydration analysis based on 0.7 Å^{-3} grids is shown in Table 3.9. In all five cases, the mean density within the 4.5 Å thick hydration shell is much higher than unconstrained simulation, which is ∼0.044 Å^{-3}. Three out of the five have a mean density that is more than 2 times higher than the bulk water density (∼0.033 Å^{-3}). This is much higher than the experimental estimation of the water density in the hydration shell [88]. Therefore, it's very likely that the surface water network with retained protein fails to model the hydration shell network in the real world. This is a reasonable observation as protein atoms can not move much, the water molecules around the protein surface can not involve in a dynamic, mutual interaction with the surface protein atoms. Here the chemical groups and the geometry became the only two factors that affect the surrounding water. As the result, the packing effect of surface water became more enhanced than usual [130]. Compared to the results in Table 3.8, the higher density in retained cases yields lower solvation free energy and higher desolvation penalty. Similarly, the energy terms in Table 3.9 also seem unrealistic. Due to the much higher density of hydration shell around the retained the protein, the average solvation energy per cube ($\overline{E}_{solv}$) is much lower compared to those in Table 3.8, calculated using 0.7 Å^{-3} grids. The total number of cubes (N_{cube}) also increased as less disturbance experienced by water in the restrained simulations. With those factors altogether, the final desolvation penalty becomes more than 12 times larger than those displayed in Table 3.8.

We further visualized the water map for nSH3:PRM$^{\text{cAbl}}$ in the format of Medical Research Council (MRC) electron-density file and rendered in Chimera (see Figure 3.18). It seems that due to the strong packing effect, the water map turns out to be quite homogeneously distributed around protein surface, even under relatively high water density cutoff. This makes it hard to detect any

surface local effect upon the surface water around. Potentially, this can cause another problem in the case of protein-ligand complex where the ligand is partially disordered in the bound state. Just as the two complexes in our work, the ligands in nSH3:PRM^{cAbl} and nSH3:PRM^{NS1} remain disordered in both C-terminal and N-terminal ends. As shown in our previous work, there is no single preferred orientation for the two ends to adopt after binding to the nSH3 domain [42]. For the two protruding ends in the ligands, a more extended conformation may gather more water cubes around than other conformations with large bending angles, which are not rare in its bound form since it's highly dynamic. Therefore, for methods require a fixed protein structure, the choice of a predetermined orientation may affect the final evaluation of solvation free energy, thus inducing unpredictable biases in calculation. This may raise another challenge for those methods while dealing with proteins that contain intrinsically disordered region (IDR).

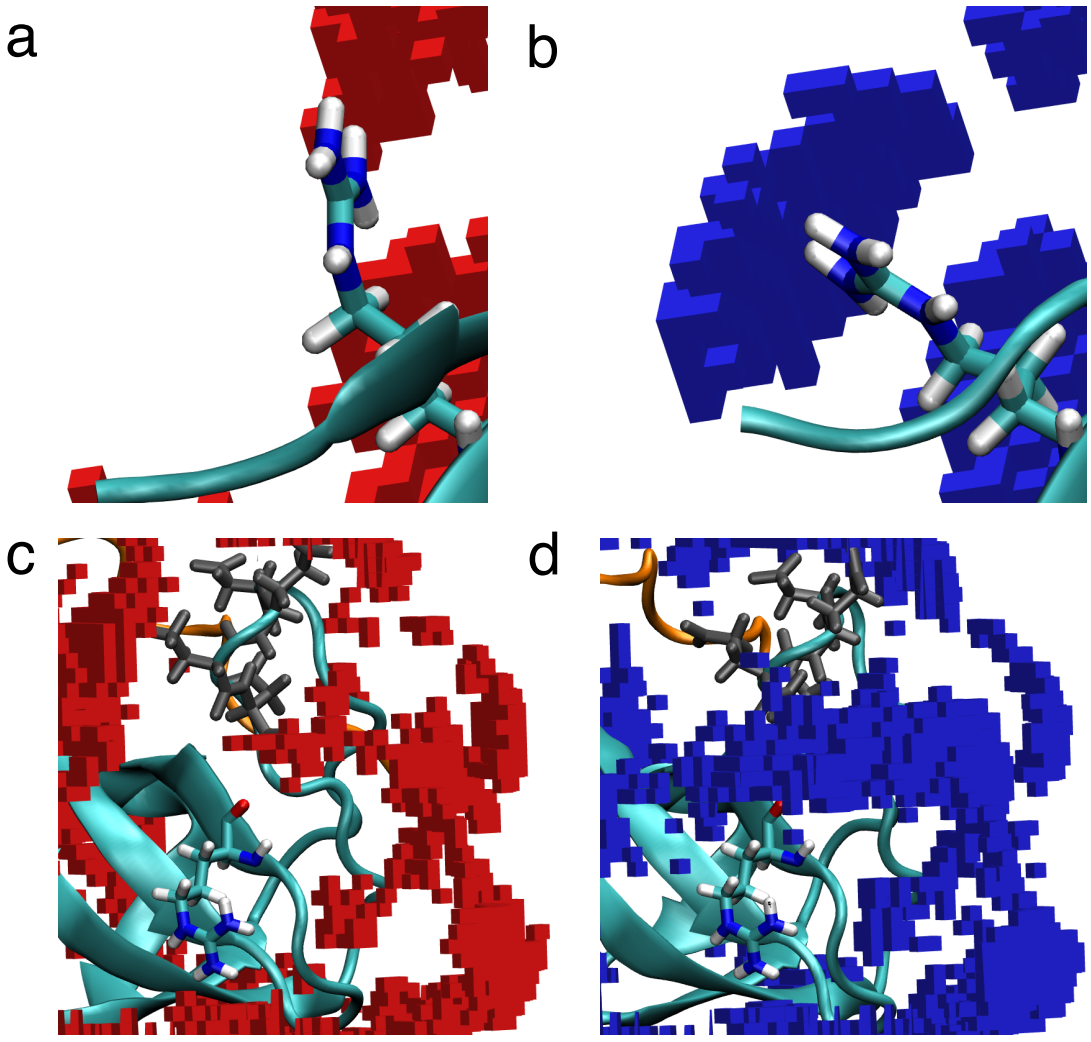

Figure 3.16: Water cubic cells around R138 and R179 in nSH3:PRM^{cAbl} (a, c) and nSH3:PRM^{NS1} (b, d). Protein and water cells represented by the same color as previous figures. R138 and R179 are shown in sticks. (a, b) Water cells around the R138 of nSH3. (c, d) Water cells around the R179 of nSH3. For comparison, nSH3:PRM^{cAbl} and nSH3:PRM^{NS1} are aligned in the same orientation. Gray sticks are the negatively charged residues in RT-loop that can attract the R179.

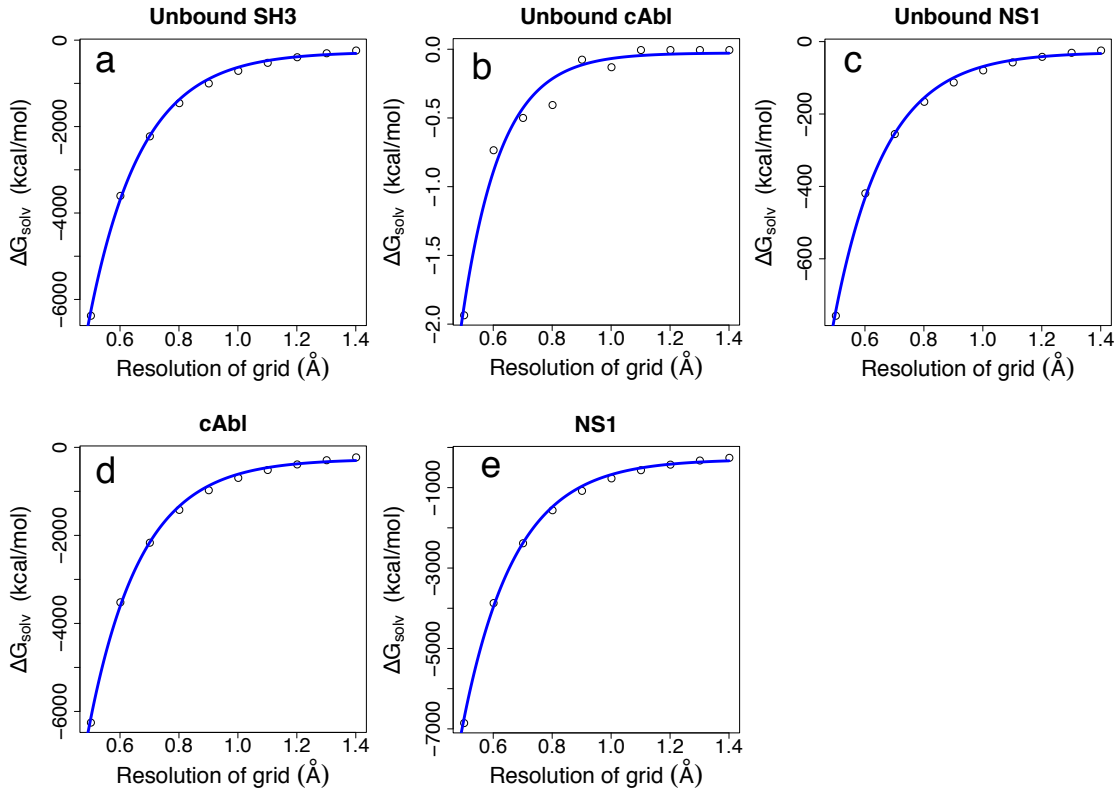

Figure 3.17: Exponential fitting of ΔG_{solv} as the function of resolution. Circle: calculated ΔG_{solv} . Blue line: fitted exponential equation ($y = a \times e^{-bx} + c$) for each model.

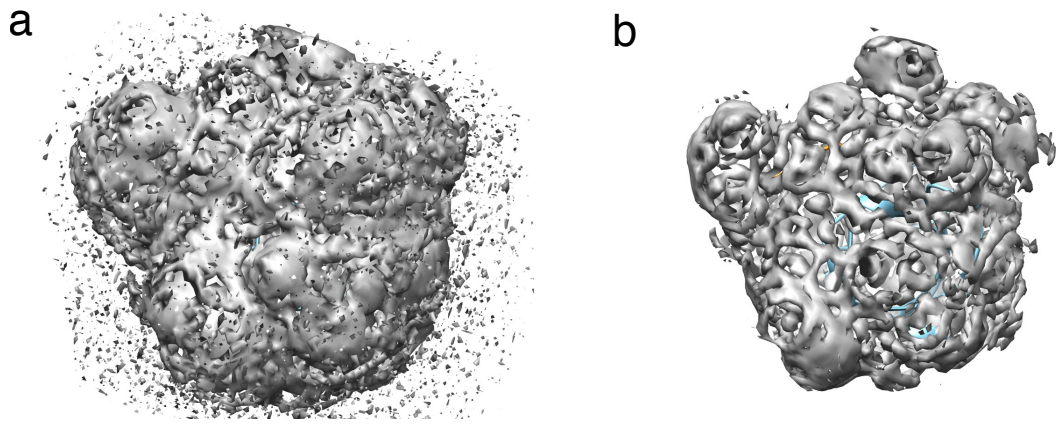

Figure 3.18: Hydration map of nSH3:PRM$^{\text{cAbl}}$ under retained simulation, visualized in the form of MRC electron-density map format via Chimera. The protein structure was rendered in the same color and aligned in the same orientation as those in Figure 3.15. (a) Hydration map visualized with density cutoff 0.036 Å^{-3} (slightly larger than the bulk water density). (b) Hydration map visualized with density cutoff 0.049 Å^{-3} ($\sim$ 1.5 times the bulk water density).

Table 3.8: Calculation of solvation free energy as well as desolvation penalty under varying resolution (unit: Å). All energy terms are presented in unit kcal/mol. N_{cube}: total number of cubes included calculation. $\overline{E}_{solv}$: average density-based solvation energy per cube. ΔG_{solv}: the energetic change calculated as the difference between mean energy and bulk water energy using equation 2.13. ΔG_{penal}: the difference in energetic changes between complex and each unbound component. . ΔG_{solv}^{expo}: the calculated asymptotic value for ΔG_{solv} via exponential fitting.

	Resolution	N_{cube}	$\overline{E}_{solv}$	ΔG_{solv}	ΔG_{penal}	ΔG_{solv}^{expo}
Unbound SH3	0.5	32199	1.83	-6358.72	—	-269.20
	0.6	18761	1.83	-3584.34	—	
	0.7	11889	1.84	-2207.39	—	
	0.8	7895	1.84	-1439.38	—	
	0.9	5532	1.85	-982.98	—	
Unbound cAbl	0.5	15	1.90	-1.93	—	-0.03
	0.6	5	1.88	-0.73	—	
	0.7	5	1.92	-0.49	—	
	0.8	4	1.92	-0.40	—	
	0.9	1	1.95	-0.07	—	
Unbound NS1	0.5	4984	1.87	-755.06	—	-28.89
	0.6	2803	1.88	-416.90	—	
	0.7	1778	1.88	-253.17	—	
	0.8	1168	1.88	-164.06	—	
	0.9	808	1.89	-110.56	—	
cAbl	0.5	32247	1.83	-6235.44	125.21	-264.20
	0.6	18565	1.84	-3502.51	82.56	
	0.7	11629	1.84	-2148.07	59.81	
	0.8	7768	1.84	-1402.87	36.91	
	0.9	5461	1.85	-956.18	26.87	
NS1	0.5	33836	1.82	-2274.337	278.00	-293.50
	0.6	19505	1.83	-1121.171	149.55	
	0.7	12236	1.83	-598.467	97.32	
	0.8	8150	1.83	-350.594	61.04	
	0.9	5741	1.84	-212.656	33.42	

Table 3.9: Calculation of density-based solvation free energy with retained protein atoms. Resolution 0.7 Å is used. Mean density: average number density of cubes within the protein surface hydration shell of 4.5Å thickness. Unit: number per $Å^{-3}$. Other terms follow the same definition as Table 3.8.

	N_{cube}	Mean density	$\overline{E}_{solv}$	ΔG_{solv}	ΔG_{penal}
Unbound SH3	11861	0.074	1.67	-4206.112	—
Unbound cAbl	5932	0.064	1.71	-1849.746	—
Unbound NS1	6672	0.064	1.71	-2078.628	—
cAbl	13869	0.074	1.67	-4917.864	1137.99
NS1	14059	0.076	1.66	-5089.189	1195.55

4. CONCLUSIONS AND FUTURE DIRECTIONS *

4.1 Fuzzy Interactions Enhance the Nonspecific Long-Range Electrostatic Interaction

In my first study, we tested whether the 500-ns simulation time was adequate for analyzing conformational motion by calculating and comparing the backbone amide bond order parameters. Based on the relaxation time of the angular reorientational correlation function $C_I(t)$ (Eq. 2.2; Fig. 3.1 and Fig. 3.2, we found that 500 ns simulation provides good sampling results that reach good convergence among the systems we are studying. Our computational and experimental order parameters agreed reasonably well for nSH3 (see Fig. 3.3). Upon complex formation, PRMs undergo larger changes in order parameter and RMSF of the backbone C_α atoms (Fig. 3.3–Fig. 3.6). This reflects that nSH3 remains stably folded while PRMs undergo disorder-order transition as they bind. Similar behaviors were observed in conformational entropy changes, as discussed in the second study. Based these findings, simulational trajectories across 100–500-ns interval seem to be adequate for the purposes of our analysis.

Both order parameters and RMSF suggest the viral ligand PRM^{NS1} in the bound states still remains partially disordered in both terminal ends. And our cluster analysis also confirms this finding. To examine if this terminal fuzziness contributes to the long-range electrostatic interactions with the nSH3 domain, we calculated the average pairwise distances between acidic residues in the nSH3 domain and the positively charged terminal residues in PRM^{NS1}. Using MD trajectories and crystal structures to represent mobile and static states of the bound PRM^{NS1}, we found the mean pairwise distances calculated from the MD trajectories were considerably shorter, suggesting the electrostatic interactions are enhanced by the disordered peripheral region of the viral PRM.

Not long ago, another study also indicated that the PRM in NS3 of the Chikungunya virus mediated the binding of SH3 domain via polyelectrostatic interactions [163]. Therefore, our results may shed lights on the role of nonspecific, long-range electrostatic interactions in other partially

*Part of the contents are reprinted with permissions from 1) Shen, Q., et al. *Biophysical journal*, 114(6), pp.1313-1320, and 2) Shi, J., et al. *Biophysical journal*, 118(10), pp.2502-2512.

fuzzy complexes as well.

4.2 Entropy Hotspots and Dynamic Allostery for the Binding of Intrinsically Disordered Ligands

The second study aims at dissecting the relative contributions of backbone and side-chain entropies, for individual residues or together, provide important details about the binding of PRMs to nSH3. Present results suggest that side chains play a critical role for determining entropy changes associated with the binding. For the stably folded nSH3, side-chain entropy change is larger than that of the backbone. For PRMs, as they belong to IDRs, both the backbone and side chains contribute comparably to the entropy change. As a result, considerable fraction of the net entropy change is due to PRMs, even though they are much smaller than nSH3 in size. We also note that, since entropy is a state function, only separate simulations of the uncomplexed PRMs and nSH3, and nSH3:PRM complexes were needed for calculation.

We found that ligand binding induces per-residue side-chain entropy changes both at the nSH3:PRM interface and across nSH3. For residues of nSH3 that form high-occupancy contacts with PRM, side-chain entropy decreases to varying extents, the largest of which being the highly conserved and nonpolar F141 and W169. Analogously, for an SH2 domain, ligand binding did not incur any significant changes in the backbone motion, while a few "hot-spot" residues at the interface made large contribution to the binding free energy [164]. Unlike nSH3, the side-chain entropy loss of PRMs is large for residues that form high-occupancy hydrogen bonds. Consistent with the non-reciprocal changes in side-chain entropy at the interface, a previous NMR study of a calmodulin domain and its target domains noted "surprisingly noncomplementary" distribution of motion at the interface [165]. Side-chain entropy changes in remote surface-exposed residues likely arises from propagation of the rearrangements in the contacts across surface residues upon ligand binding. Long-range effect of ligand binding has also been observed in other systems including SH3 domains [153, 28] and an MDM2 domain with p53-derived peptides [166]. Entropy changes in charged or polar residues in distal regions may play an allosteric role for electrostatic interactions with other domains [167, 168].

Between nSH3:PRM^{cAbl} and nSH3:PRM^{NS1}, the higher binding affinity of the latter [1, 2] appears to be driven by the favorable enthalpy change that compensates for the unfavorable entropy loss. The smaller backbone entropy change in PRM^{NS1} is due to its central region that remains relatively rigid in the unbound state and also due to its C-terminal region that form "fuzzy" contacts so that the region stays flexible in the bound state while its side chains lose entropy substantially. Taken together, the present results elucidate the role of entropy hotspots in influencing the conformational flexibility of the complex.

Moreover, the sites that are distant from the binding interface but undergo considerable changes in conformational entropy may conversely regulate protein-protein interactions, called dynamic allostery [169, 170]. In this regard, the allosteric hotspots may serve as viable targets for the rational development of protein-protein interaction modulators.

The second study combining entropy and contact analysis, together with experimental studies using various NMR relaxation methods [171, 172], will facilitate the understanding of the side chain dynamics and its role in molecular recognition. Furthermore, interfacial water molecules likely play a significant role as well [29, 32, 33, 34], which is a subject of my next study.

4.3 Coupling Effect Between Local Hydration and Protein Motion in Determining the Free Energy

In my third study, we developed our grid-based density water map analysis to further calculate the energetic contribution of surface water around the protein. We tested our approach with two SH3:PRM complexes.

As shown in the results, our approach can provide details of hydration contribution in the local surface region. Comparing the two complexes— nSH3:PRM^{cAbl} and nSH3:PRM^{NS1}— the larger desolvation penalty rises from the loss of high-density hydration region. And those regions seems to be related with the more frequent side-chain dynamic motions of protein residues nearby. More specifically, the R138 and R179 in nSH3:PRM^{cAbl} became more mobile in terms of side-chain motion upon the binding of PRM^{cAbl}, resulting in a higher desolvation penalty in the complex state. This is consistent with the experimental finding that PRM^{cAbl} has a much lower binding

affinity than the viral ligand PRM^{NS1}. This coordinated coupling effect between protein dynamics and the local solvation contribution indicates the importance of maintaining the free motion of protein atoms for the study of solvation energetic contribution.

It's worth noting that both R138 and R179 are distant from the binding pocket, only affected by the binding of ligand via a long-distance allosteric manner. This long-range effect upon the ligand binding was also found in the previous entropy study where propagation initiates from the binding pocket to remote region where some residues experience great entropic change (thus so-called "entropy hotspot") [159]. Combining the two findings together, the long-range allosteric effect that can influence both remote entropy as well as remote surface hydration is potentially one mechanism regulated by viral ligand in order to optimize its binding to the host's protein along the signaling pathway.

It has long been found that the solvent viscosity is sometimes correlated with protein dynamics [173]. Therefore a slaving role of water model then has been further suggested that the protein external and internal motions are slaved by hydration shell fluctuation, bulk water fluctuation, and water viscosity to different extents [174]. Moreover, the local hydration shell is also found to influence the protein dynamics, energy landscape, thus determining the functionality of the protein [175, 84, 89, 174, 86]. However, there is also continuing doubt about the ruling power of solvent since a large portion of the evidences supporting the slaving role model are extrapolated from the "glass transition" observed at around 200 K [176, 177, 178]. Marques's recent work showed that the fast internal side chain motion is largely independent of either bulk or local hydration layer water dynamics, indicating the protein conformational entropy is unaffected by the surrounding water [179]. Based on Marques's finding, our work further suggests a contrary scenario, where the protein conformational entropy may influence the hydration shell energy thus affect the protein-ligand binding process.

We also calculated the global solvation free energy change. A dependence of resolution of cube size is observed, so we exponentially fitted the result to find the asymptotic value in solvation energy. When adding the asymptotic values into the MM/PBSA calculation, the final binding

free energy got slightly improved. Our method used a stably folded core of the protein as our reference structure to build the cubic grid of cells. Unsurprisingly, when it is applied to intrinsically disordered ligand in the unbound state, it does not provide too much useful details since the surface hydration structure can not be correctly calculated. Further improvements are needed to better construct the water density map around the intrinsically disordered proteins.

Since our results emphasize the importance of unconstrained protein during simulation, we tested protein systems with constraints throughout the production run. An unrealistic high-density hydration shell is found in the retained simulations. And the over-homogeneous surface hydration structure may lose some details in the local hydration regions. This also suggests another weakness in retaining protein atoms if the protein structure involves any IDR. Because the orientation of IDR may affect the hydration contribution thus including unpredictable perturbation into the calculation.

In sum, our work provides new insights into the role of water in protein-protein interactions that could find use in drug design process.

4.4 Overall Summary and Outlook

The first study shows the partial fuzziness in the viral ligand enhances the long-distance non-specific electrostatic interactions by decreasing the average pairwise distances, thus enhancing the binding affinity of the PRM^{NS1}, which is consistent with what is assumed in the polyelectrostatic model. A further study may be worthwhile to examine if other viral ligands with similar charged IDR share a similar strategy to enhance the long-distance electrostatic interactions.

The second study shows for unstructured ligand, both backbone and side chain contribute significantly to the conformational entropy, but for well-structured nSH3 domain, the backbone conformational entropy change is small, due to the fact that before and after ligation, the backbone motion of the nSH3 domain is not deeply changed. This finding suggests a structural basis for predicting backbone entropy contribution, which sheds light on the recent discussion about whether backbone or side chain entropy alone is dominant in the protein-protein recognition [69, 70, 71]. We also found the decrease in entropy is not evenly distributed among all the residues, but mainly

from some entropic "hot-spot" residues that contribute to the entropic change much more than the average residues. What may be more surprising is our identification of a few residues in the remote region in nSH3 domain that end up gaining higher entropy upon ligation, which we showed may be the result of propagation of the rearrangements originating from the binding pocket. nSH3:PRM^{cAbl} and nSH3:PRM^{NS1} both have favorable enthalpy changes that compensate for the unfavorable entropy losses. Comparing to PRM^{cAbl}, PRM^{NS1} has greater entropic loss only equivalent to the free energy of one hydrogen bond but manages to form many more high-occupancy contacts, indicating the gains in enthalpy far exceed the losses in entropy. A further decomposition of the free energy contributions of hotspot residues [180] into enthalpic and entropic contributions may provide new insight into how specificity is tuned for proteins with multiple binding partners [165]. To further study the kinetic aspects of the binding process, additional simulations of the binding and unbinding processes of the PRMs are needed.

The third study reveals the local solvation free energy is significantly impacted by the protein sidechain motion. Compared with nSH3:PRM^{NS1}, nSH3:PRM^{cAbl} is found to have a greater desolvation penalty, as the result of the loss of high-density water blocks around the surface hydration nearby R138 and R179. As reported in the second study, R138 and R179 in nSH3:PRM^{cAbl} experience higher mobility as the result of propagation of effects of ligand binding from the binding pocket. Therefore, the long range allosteric effect of PRM binding is not just based on entropic factor, but also on the local solvation energy. Our findings also highlight the importance of taking into account of the protein dynamics for the calculation of hydration contribution, which is often omitted by other hydration analysis frameworks. Our newly developed method may also improve the classic binding free energy calculation, MM/PBSA, by improving the solvation free energy estimations. Further developments of the current approach, like building water map based on multiple locally aligned structures will be useful in elucidating the solvation energy around unstructured proteins. Our method may be used in combination with other entropic calculation and/or interaction energy calculation, which may have potentials for revealing a wider perspective on the role of surface hydration structure for proteins.

Combining study 1 to 3, our results reveal a unified picture of how different thermodynamic factors coordinate with one another to facilitate the exceptionally high binding affinity in PRM^{NS1}. And we look forward to applying our established analysis procedures to other viral/cellular PRMs to reveal more general principals in the SH3:RPM binding.

REFERENCES

[1] V. S. Bhatt, D. Zeng, I. Krieger, J. C. Sacchettini, and J.-H. Cho, "Binding mechanism of the n-terminal sh3 domain of crkii and proline-rich motifs in cabl," *Biophys. J.*, vol. 110, pp. 2630–2641, 2016.

[2] Q. Shen, D. Zeng, B. Zhao, V. S. Bhatt, P. Li, and J.-H. Cho, "The molecular mechanisms underlying the hijack of host proteins by the 1918 spanish influenza virus," *ACS Chem. Biol.*, vol. 12, pp. 1199–1203, 2017.

[3] P. Tompa, "Intrinsically unstructured proteins," *Trends in biochemical sciences*, vol. 27, no. 10, pp. 527–533, 2002.

[4] V. N. Uversky, C. J. Oldfield, and A. K. Dunker, "Showing your id: intrinsic disorder as an id for recognition, regulation and cell signaling," *Journal of Molecular Recognition: An Interdisciplinary Journal*, vol. 18, no. 5, pp. 343–384, 2005.

[5] V. N. Uversky and A. K. Dunker, "Controlled chaos," *Science.*, vol. 322, pp. 1340–1341, 2008.

[6] A. K. Dunker, C. J. Brown, J. D. Lawson, L. M. Iakoucheva, and Z. Obradovic, "Intrinsic disorder and protein function," *Biochemistry.*, vol. 41, pp. 6573–6582, 2002.

[7] T. Mittag, L. E. Kay, and J. D. Forman-Kay, "Protein dynamics and conformational disorder in molecular recognition," *J. Mol. Recognit.*, vol. 23, pp. 105–116, 2010.

[8] V. N. Uversky, "A decade and a half of protein intrinsic disorder: biology still waits for physics.," *Protein Sci.*, vol. 22, pp. 693–724, 2013.

[9] P. E. Wright and H. J. Dyson, "Intrinsically disordered proteins in cellular signalling and regulation," *Nat. Rev. Mol. Cell Biol.*, vol. 16, pp. 18–29, 2015.

[10] S. L. Shammas, M. D. Crabtree, L. Dahal, B. I. M. Wicky, and J. Clarke, "Insights into coupled folding and binding mechanisms from kinetic studies," *J. Biol. Chem.*, vol. 291, pp. 6689–6695, 2016.

[11] D. Zeng, Q. Shen, and J.-H. Cho, "Thermodynamic contribution of backbone conformational entropy in the binding between sh3 domain and proline-rich motif," *Biochem. Biophys. Res. Commun.*, vol. 484, pp. 21–26, 2017.

[12] H. J. Dyson and P. E. Wright, "Intrinsically unstructured proteins and their functions," *Nat. Rev. Mol. Cell Biol.*, vol. 6, pp. 197–208, 2005.

[13] P. Tompa and M. Fuxreiter, "Fuzzy complexes: polymorphism and structural disorder in protein–protein interactions," *Trends in biochemical sciences*, vol. 33, no. 1, pp. 2–8, 2008.

[14] R. Williams, Z. Obradovi, V. Mathura, W. Braun, E. Garner, J. Young, S. Takayama, C. Brown, and A. Dunker, "The protein non-folding problem: amino acid determinants of intrinsic order and disorder," *Pac. Symp. Biocomput.*, pp. 89–100, 2001.

[15] F.-X. Theillet, L. Kalmar, P. Tompa, K.-H. Han, P. Selenko, A. K. Dunker, G. W. Daughdrill, and V. N. Uversky, "The alphabet of intrinsic disorder: I. act like a pro: On the abundance and roles of proline residues in intrinsically disordered proteins," *Intrinsically Disord. Proteins.*, vol. 1, p. e24360, 2013.

[16] A. Campen, R. M. Williams, C. J. Brown, J. Meng, V. N. Uversky, and A. K. Dunker, "Top-idp-scale: A new amino acid scale measuring propensity for intrinsic disorder," *Protein Pept. Lett.*, vol. 15, pp. 956–963, 2009.

[17] T. A. Graham, D. M. Ferkey, F. Mao, D. Kimelman, and W. Xu, "Tcf4 can specifically recognize β-catenin using alternative conformations," *Nature structural biology*, vol. 8, no. 12, pp. 1048–1052, 2001.

[18] R. S. Rock, B. Ramamurthy, A. R. Dunn, S. Beccafico, B. R. Rami, C. Morris, B. J. Spink, C. Franzini-Armstrong, J. A. Spudich, and H. L. Sweeney, "A flexible domain is essential

for the large step size and processivity of myosin vi," *Molecular cell*, vol. 17, no. 4, pp. 603–609, 2005.

[19] T. Zor, B. M. Mayr, H. J. Dyson, M. R. Montminy, and P. E. Wright, "Roles of phosphorylation and helix propensity in the binding of the kix domain of creb-binding protein by constitutive (c-myb) and inducible (creb) activators," *Journal of Biological Chemistry*, vol. 277, no. 44, pp. 42241–42248, 2002.

[20] A. Sigalov, D. Aivazian, and L. Stern, "Homooligomerization of the cytoplasmic domain of the t cell receptor ζ chain and of other proteins containing the immunoreceptor tyrosine-based activation motif," *Biochemistry*, vol. 43, no. 7, pp. 2049–2061, 2004.

[21] H. Yu, J. K. Chen, S. Feng, D. C. Dalgarno, A. W. Brauer, and S. L. Schrelber, "Structural basis for the binding of proline-rich peptides to sh3 domains," *Cell*, vol. 76, no. 5, pp. 933–945, 1994.

[22] S. Feng, J. K. Chen, H. Yu, J. A. Simon, and S. L. Schreiber, "Two binding orientations for peptides to the src sh3 domain: development of a general model for sh3-ligand interactions," *Science*, vol. 266, no. 5188, pp. 1241–1247, 1994.

[23] B. Kay, M. Williamson, and M. Sudol, "The importance of being proline: the interaction of proline-rich motifs in signaling proteins with their cognate domains," *FASEB J.*, vol. 14, pp. 231–241, 2000.

[24] M. J. Macias, S. Wiesner, and M. Sudol, "Ww and sh3 domains, two different scaffolds to recognize proline-rich ligands," *FEBS Lett.*, vol. 513, pp. 30–37, 2002.

[25] A. Rath, A. R. Davidson, and C. M. Deber, "The structure of "unstructured" regions in peptides and proteins: Role of the polyproline ii helix in protein folding and recognition," *Biopolymers.*, vol. 80, pp. 179–185, 2005.

[26] B. J. Mayer, "SH3 domains: complexity in moderation," *Journal of cell science*, vol. 114, no. 7, pp. 1253–1263, 2001.

[27] S. S.-C. Li, "Specificity and versatility of sh3 and other proline-recognition domains: structural basis and implications for cellular signal transduction.," *Biochem. J.*, vol. 390, pp. 641–653, 2005.

[28] C. Wang, N. H. Pawley, and L. K. Nicholson, "The role of backbone motions in ligand binding to the c-Src SH3 domain," *J. Mol. Biol.*, vol. 313, pp. 873–887, 2001.

[29] A. Palencia, E. S. Cobos, P. L. Mateo, J. C. Martinez, and I. Luque, "Thermodynamic dissection of the binding energetics of proline-rich peptides to the abl-sh3 domain: Implications for rational ligand design," *J. Mol. Biol.*, vol. 336, pp. 527–537, 2004.

[30] J. C. Ferreon and V. J. Hilser, "Thermodynamics of binding to sh3 domains: The energetic impact of polyproline ii (pii) helix formation," *Biochemistry.*, vol. 43, pp. 7787–7797, 2004.

[31] T. P. Creamer, "Left-handed polyproline ii helix formation is (very) locally driven," *Proteins.*, vol. 33, pp. 218–226, 1998.

[32] A. Palencia, A. Camara-Artigas, M. T. Pisabarro, J. C. Martinez, and I. Luque, "Role of interfacial water molecules in proline-rich ligand recognition by the src homology 3 domain of abl," *J. Biol. Chem.*, vol. 285, pp. 2823–2833, 2010.

[33] J. Martin-Garcia, J. Ruiz-Sanz, and I. Luque, "Interfacial water molecules in sh3 interactions: a revised paradigm for polyproline recognition.," *Biochem. J.*, vol. 442, pp. 443–451, 2012.

[34] A. Zafra-Ruano and I. Luque, "Interfacial water molecules in SH3 interactions: Getting the full picture on polyproline recognition by protein-protein interaction domains," *FEBS Lett.*, vol. 586, pp. 2619–2630, 2012.

[35] W. C. Still, A. Tempczyk, R. C. Hawley, and T. Hendrickson, "Semianalytical treatment of solvation for molecular mechanics and dynamics," *Journal of the American Chemical Society*, vol. 112, pp. 6127–6129, 1990.

[36] W. Wang, W. A. Lim, A. Jakalian, J. Wang, J. Wang, R. Luo, C. I. Bayly, and P. A. Kollman, "An analysis of the interactions between the sem- 5 sh3 domain and its ligands using

molecular dynamics, free energy calculations, and sequence analysis," *J. Am. Chem. Soc.*, vol. 123, pp. 3986–3994, 2001.

[37] K. Bessonov, K. A. Vassall, and G. Harauz, "Docking and molecular dynamics simulations of the fyn-sh3 domain with free and phospholipid bilayer-associated 18.5-kda myelin basic protein (mbp)-insights into a noncanonical and fuzzy interaction," *Proteins.*, vol. 85, pp. 1336–1350, 2017.

[38] P. Zhou, F. Tian, X. Chen, and Z. Shang, "Modeling and prediction of binding affinities between the human amphiphysin sh3 domain and its peptide ligands using genetic algorithm-gaussian processes," *Peptide Science*, vol. 90, pp. 792–802, 2008.

[39] Z. Xu, T. Hou, N. Li, Y. Xu, and W. Wang, "Proteome-wide detection of abl1 sh3-binding peptides by integrating computational prediction and peptide microarray," *Molecular & Cellular Proteomics*, vol. 11, pp. O111–010389, 2012.

[40] T. Hou, N. Li, Y. Li, and W. Wang, "Characterization of domain–peptide interaction interface: prediction of sh3 domain-mediated protein–protein interaction network in yeast by generic structure-based models," *Journal of proteome research*, vol. 11, pp. 2982–2995, 2012.

[41] S. Hong, T. Chung, and D. Kim, "Sh3 domain-peptide binding energy calculations based on structural ensemble and multiple peptide templates," *PloS one*, vol. 5, p. e12654, 2010.

[42] Q. Shen, J. Shi, D. Zeng, B. Zhao, P. Li, W. Hwang, and J.-H. Cho, "Molecular mechanisms of tight binding through fuzzy interactions," *Biophys. J.*, vol. 114, pp. 1313–1320, 2018.

[43] M. Karplus and J. N. Kushick, "Method for estimating the configurational entropy of macromolecules," *Macromolecules*, vol. 14, pp. 325–332, 1981.

[44] R. M. Levy, M. Karplus, J. Kushick, and D. Perahia, "Evaluation of the configurational entropy for proteins: application to molecular dynamics simulations of an α-helix," *Macromolecules*, vol. 17, pp. 1370–1374, 1984.

[45] T. Hou, K. Chen, W. A. McLaughlin, B. Lu, and W. Wang, "Computational analysis and prediction of the binding motif and protein interacting partners of the abl sh3 domain," *PLoS computational biology*, vol. 2, p. e1, 2006.

[46] B. J. Mayer, M. Hamaguchi, and H. Hanafusa, "A novel viral oncogene with structural similarity to phospholipase c," *Nature*, vol. 332, no. 6161, pp. 272–275, 1988.

[47] M. Matsuda, C. Reichman, and H. Hanafusa, "Biological and biochemical activity of v-crk chimeras containing the sh2/sh3 regions of phosphatidylinositol-specific phospholipase c-gamma and src.," *Journal of virology*, vol. 66, no. 1, pp. 115–121, 1992.

[48] S. M. Feller, "Crk family adaptors-signalling complex formation and biological roles," *Oncogene.*, vol. 20, pp. 6348–6371, 2001.

[49] R. B. Birge, C. Kalodimos, F. Inagaki, and S. Tanaka, "Crk and crkl adaptor proteins: networks for physiological and pathological signaling," *Cell Communication and Signaling*, vol. 7, no. 1, pp. 1–23, 2009.

[50] S. Gorina and N. P. Pavletich, "Structure of the p53 tumor suppressor bound to the ankyrin and sh3 domains of 53bp2," *Science*, vol. 274, no. 5289, pp. 1001–1005, 1996.

[51] H. Murakami, T. Iwashita, N. Asai, Y. Iwata, S. Narumiya, and M. Takahashi, "Rho-dependent and-independent tyrosine phosphorylation of focal adhesion kinase, paxillin and p130 cas mediated by ret kinase," *Oncogene*, vol. 18, no. 11, pp. 1975–1982, 1999.

[52] J. Y. Wang, "Abl tyrosine kinase in signal transduction and cell-cycle regulation," *Current opinion in genetics & development*, vol. 3, no. 1, pp. 35–43, 1993.

[53] O. Hantschel and G. Superti-Furga, "Regulation of the c-abl and bcr–abl tyrosine kinases," *Nature reviews Molecular cell biology*, vol. 5, no. 1, pp. 33–44, 2004.

[54] D. Lindholm, D. D. Pham, A. Cascone, O. Eriksson, K. Wennerberg, and M. Saarma, "c-abl inhibitors enable insights into the pathophysiology and neuroprotection in parkinson's disease," *Frontiers in aging neuroscience*, vol. 8, p. 254, 2016.

[55] R. Ren, Z.-S. Ye, and D. Baltimore, "Abl protein-tyrosine kinase selects the crk adapter as a substrate using sh3-binding sites.," *Genes & development*, vol. 8, no. 7, pp. 783–795, 1994.

[56] S. Antoku, K. Saksela, G. M. Rivera, and B. J. Mayer, "A crucial role in cell spreading for the interaction of abl pxxp motifs with crk and nck adaptors," *Journal of cell science*, vol. 121, no. 18, pp. 3071–3082, 2008.

[57] Y. Kobashigawa, M. Sakai, M. Naito, M. Yokochi, H. Kumeta, Y. Makino, K. Ogura, S. Tanaka, and F. Inagaki, "Structural basis for the transforming activity of human cancer-related signaling adaptor protein crk," *Nature structural & molecular biology*, vol. 14, no. 6, pp. 503–510, 2007.

[58] T. Shishido, T. Akagi, A. Chalmers, M. Maeda, T. Terada, M.-M. Georgescu, and H. Hanafusa, "Crk family adaptor proteins trans-activate c-abl kinase," *Genes to Cells*, vol. 6, no. 5, pp. 431–440, 2001.

[59] M. Brehme, O. Hantschel, J. Colinge, I. Kaupe, M. Planyavsky, T. Köcher, K. Mechtler, K. L. Bennett, and G. Superti-Furga, "Charting the molecular network of the drug target bcr-abl," *Proceedings of the National Academy of Sciences*, vol. 106, no. 18, pp. 7414–7419, 2009.

[60] M. M. Savitski, F. B. Reinhard, H. Franken, T. Werner, M. F. Savitski, D. Eberhard, D. M. Molina, R. Jafari, R. B. Dovega, S. Klaeger, *et al.*, "Tracking cancer drugs in living cells by thermal profiling of the proteome," *Science*, vol. 346, no. 6205, 2014.

[61] K. Subbarao, D. Swayne, and C. Olsen, "Influenza virology: current topics," 2005.

[62] T. Horimoto and Y. Kawaoka, "Influenza: lessons from past pandemics, warnings from current incidents," *Nature Reviews Microbiology*, vol. 3, no. 8, pp. 591–600, 2005.

[63] R. M. Krug, "Functions of the influenza a virus ns1 protein in antiviral defense," *Current opinion in virology*, vol. 12, pp. 1–6, 2015.

[64] Y.-K. Shin, Q. Liu, S. K. Tikoo, L. A. Babiuk, and Y. Zhou, "Influenza a virus ns1 protein activates the phosphatidylinositol 3-kinase (pi3k)/akt pathway by direct interaction with the p85 subunit of pi3k," *Journal of General Virology*, vol. 88, no. 1, pp. 13–18, 2007.

[65] L. S. Heikkinen, A. Kazlauskas, K. Melén, R. Wagner, T. Ziegler, I. Julkunen, and K. Saksela, "Avian and 1918 spanish influenza a virus ns1 proteins bind to crk/crkl src homology 3 domains to activate host cell signaling," *Journal of Biological Chemistry*, vol. 283, no. 9, pp. 5719–5727, 2008.

[66] B. G. Hale, D. Jackson, Y.-H. Chen, R. A. Lamb, and R. E. Randall, "Influenza a virus ns1 protein binds p85β and activates phosphatidylinositol-3-kinase signaling," *Proceedings of the National Academy of Sciences*, vol. 103, no. 38, pp. 14194–14199, 2006.

[67] M. Borg, T. Mittag, T. Pawson, M. Tyers, J. D. Forman-Kay, and H. S. Chan, "Polyelectrostatic interactions of disordered ligands suggest a physical basis for ultrasensitivity," *Proceedings of the National Academy of Sciences*, vol. 104, no. 23, pp. 9650–9655, 2007.

[68] V. Kasinath, K. A. Sharp, and A. J. Wand, "Microscopic insights into the nmr relaxation-based protein conformational entropy meter," *J. Am. Chem. Soc.*, vol. 135, pp. 15092–15100, 2013.

[69] M. C. Baxa, E. J. Haddadian, J. M. Jumper, K. F. Freed, and T. R. Sosnick, "Loss of conformational entropy in protein folding calculated using realistic ensembles and its implications for nmr-based calculations," *Proc. Natl. Acad. Sci. USA*, vol. 111, pp. 15396–15401, 2014.

[70] K. Sharp, E. O'Brien, V. Kasinath, and A. Wand, "On the relationship between nmr-derived amide order parameters and protein backbone entropy changes.," *Proteins.*, vol. 83, pp. 922–930, 2015.

[71] J. A. Caro, K. W. Harpole, V. Kasinath, J. Lim, J. Granja, K. G. Valentine, K. A. Sharp, and A. J. Wand, "Entropy in molecular recognition by proteins," *Proc. Natl. Acad. Sci. USA*, vol. 114, pp. 6563–6568, 2017.

[72] A. J. Wand and K. A. Sharp, "Measuring entropy in molecular recognition by proteins," *Annu. Rev. Biophys.*, vol. 47, pp. 41–61, 2018.

[73] M. Karplus, T. Ichiye, and B. Pettitt, "Configurational entropy of native proteins," *Biophys. J.*, vol. 52, pp. 1083–1085, 1987.

[74] R. Grünberg, M. Nilges, and J. Leckner, "Flexibility and conformational entropy in protein-protein binding," *Structure*, vol. 14, pp. 683–693, 2006.

[75] K. H. Lee, D. Xie, E. Freire, and L. M. Amzel, "Estimation of changes in side chain configurational entropy in binding and folding: general methods and application to helix formation," *Proteins.*, vol. 20, pp. 68–84, 1994.

[76] A. Doig and M. j. Sternberg, "Side-chain conformational entropy in protein folding.," *Protein Sci.*, vol. 4, pp. 2247–2251, 1995.

[77] K. H. DuBay and P. L. Geissler, "Calculation of proteins' total side-chain torsional entropy and its influence on protein–ligand interactions," *J. Mol. Biol.*, vol. 391, pp. 484–497, 2009.

[78] B. King and B. Tidor, "Mist: Maximum information spanning trees for dimension reduction of biological data sets," *Bioinformatics.*, vol. 25, pp. 1165–1172, 2009.

[79] B. King, N. Silver, and B. Tidor, "Efficient calculation of molecular configurational entropies using an information theoretic approximation," *J. Phys. Chem. B.*, vol. 2012, pp. 2891–2904, 2012.

[80] P. Ball, "Water as an active constituent in cell biology," *Chemical reviews*, vol. 108, no. 1, pp. 74–108, 2008.

[81] E. Duboué-Dijon, A. C. Fogarty, J. T. Hynes, and D. Laage, "Dynamical disorder in the dna hydration shell," *J. Am. Chem. Soc.*, vol. 138, no. 24, pp. 7610–7620, 2016.

[82] M.-C. Bellissent-Funel, A. Hassanali, M. Havenith, R. Henchman, P. Pohl, F. Sterpone, D. van der Spoel, Y. Xu, and A. E. Garcia, "Water determines the structure and dynamics of proteins," *Chemical Reviews*, vol. 116, no. 13, pp. 7673–7697, 2016.

[83] J. D. Dunitz, "The entropic cost of bound water in crystals and biomolecules," *Science*, vol. 264, no. 5159, pp. 670–671, 1994.

[84] M. Tarek and D. Tobias, "Role of protein-water hydrogen bond dynamics in the protein dynamical transition," *Physical review letters*, vol. 88, no. 13, p. 138101, 2002.

[85] G. A. Papoian, J. Ulander, M. P. Eastwood, Z. Luthey-Schulten, and P. G. Wolynes, "Water in protein structure prediction," *Proceedings of the National Academy of Sciences*, vol. 101, no. 10, pp. 3352–3357, 2004.

[86] M. Maurer and C. Oostenbrink, "Water in protein hydration and ligand recognition," *Journal of Molecular Recognition*, vol. 32, no. 12, p. e2810, 2019.

[87] M. Marchi, F. Sterpone, and M. Ceccarelli, "Water rotational relaxation and diffusion in hydrated lysozyme," *Journal of the American Chemical Society*, vol. 124, no. 23, pp. 6787–6791, 2002.

[88] F. Merzel and J. C. Smith, "Is the first hydration shell of lysozyme of higher density than bulk water?," *Proceedings of the National Academy of Sciences*, vol. 99, no. 8, pp. 5378–5383, 2002.

[89] C. C. Goldmon and G. A. Papoian, "Hierarchical organization of eglin c native state dynamics is shaped by competing direct and water-mediated interactions," *Proceedings of the National Academy of Sciences*, vol. 105, no. 31, pp. 10659–10664, 2008.

[90] A. Oleinikova, N. Smolin, and I. Brovchenko, "Influence of water clustering on the dynamics of hydration water at the surface of a lysozyme," *Biophysical Journal*, vol. 93, no. 9, pp. 2986–3000, 2007.

[91] D. Laage, T. Elsaesser, and J. T. Hynes, "Water dynamics in the hydration shells of biomolecules," *Chemical Reviews*, vol. 117, no. 16, pp. 10694–10725, 2017.

[92] J. Michel, J. Tirado-Rives, and W. L. Jorgensen, "Energetics of displacing water molecules from protein binding sites: consequences for ligand optimization," *Journal of the American Chemical Society*, vol. 131, no. 42, pp. 15403–15411, 2009.

[93] Y. Lu, R. Wang, C.-Y. Yang, and S. Wang, "Analysis of ligand-bound water molecules in high-resolution crystal structures of protein- ligand complexes," *Journal of chemical information and modeling*, vol. 47, no. 2, pp. 668–675, 2007.

[94] S. de Beer, N. P. Vermeulen, and C. Oostenbrink, "The role of water molecules in computational drug design," *Current topics in medicinal chemistry*, vol. 10, no. 1, pp. 55–66, 2010.

[95] J. Wahl and M. Smieško, "Thermodynamic insight into the effects of water displacement and rearrangement upon ligand modifications using molecular dynamics simulations," *ChemMedChem*, vol. 13, no. 13, pp. 1325–1335, 2018.

[96] P. W. Snyder, M. R. Lockett, D. T. Moustakas, and G. M. Whitesides, "Is it the shape of the cavity, or the shape of the water in the cavity?," *The European Physical Journal Special Topics*, vol. 223, no. 5, pp. 853–891, 2014.

[97] W. Kauzmann, "Some factors in the interpretation of protein denaturation," in *Advances in protein chemistry*, vol. 14, pp. 1–63, Elsevier, 1959.

[98] D. Chandler, "Interfaces and the driving force of hydrophobic assembly," *Nature*, vol. 437, no. 7059, pp. 640–647, 2005.

[99] W. Blokzijl and J. B. Engberts, "Hydrophobic effects. opinions and facts," *Angewandte Chemie International Edition in English*, vol. 32, no. 11, pp. 1545–1579, 1993.

[100] D. T. Bowron, A. Filipponi, M. A. Roberts, and J. L. Finney, "Hydrophobic hydration and the formation of a clathrate hydrate," *Physical review letters*, vol. 81, no. 19, p. 4164, 1998.

[101] H. S. Ashbaugh, D. Asthagiri, L. R. Pratt, and S. B. Rempe, "Hydration of krypton and consideration of clathrate models of hydrophobic effects from the perspective of quasi-chemical theory," *Biophysical chemistry*, vol. 105, no. 2-3, pp. 323–338, 2003.

[102] R. Zhou, X. Huang, C. J. Margulis, and B. J. Berne, "Hydrophobic collapse in multidomain protein folding," *Science*, vol. 305, no. 5690, pp. 1605–1609, 2004.

[103] L. Hua, X. Huang, P. Liu, R. Zhou, and B. J. Berne, "Nanoscale dewetting transition in protein complex folding," *The Journal of Physical Chemistry B*, vol. 111, no. 30, pp. 9069–9077, 2007.

[104] N. Giovambattista, C. F. Lopez, P. J. Rossky, and P. G. Debenedetti, "Hydrophobicity of protein surfaces: Separating geometry from chemistry," *Proceedings of the National Academy of Sciences*, vol. 105, no. 7, pp. 2274–2279, 2008.

[105] Y.-K. Cheng and P. J. Rossky, "Surface topography dependence of biomolecular hydrophobic hydration," *Nature*, vol. 392, no. 6677, pp. 696–699, 1998.

[106] E. Barratt, R. J. Bingham, D. J. Warner, C. A. Laughton, S. E. Phillips, and S. W. Homans, "Van der waals interactions dominate ligand- protein association in a protein binding site occluded from solvent water," *Journal of the American Chemical Society*, vol. 127, no. 33, pp. 11827–11834, 2005.

[107] P. W. Snyder, J. Mecinović, D. T. Moustakas, S. W. Thomas, M. Harder, E. T. Mack, M. R. Lockett, A. Héroux, W. Sherman, and G. M. Whitesides, "Mechanism of the hydrophobic effect in the biomolecular recognition of arylsulfonamides by carbonic anhydrase," *Proceedings of the National Academy of Sciences*, vol. 108, no. 44, pp. 17889–17894, 2011.

[108] R. W. Zwanzig, "High-temperature equation of state by a perturbation method. i. nonpolar gases," *J. Chem. Phys.*, vol. 22, no. 8, pp. 1420–1426, 1954.

[109] J. G. Kirkwood, "Statistical mechanics of fluid mixtures," *J. Chem. Phys.*, vol. 3, no. 5, pp. 300–313, 1935.

[110] C. Peter, C. Oostenbrink, A. Van Dorp, and W. F. Van Gunsteren, "Estimating entropies from molecular dynamics simulations," *J. Chem. Phys.*, vol. 120, no. 6, pp. 2652–2661, 2004.

[111] D. J. Huggins, "Quantifying the entropy of binding for water molecules in protein cavities by computing correlations," *Biophys. J.*, vol. 108, no. 4, pp. 928–936, 2015.

[112] V. Khanna, J. I. Monroe, M. F. Doherty, and B. Peters, "Performing solvation free energy calculations in lammps using the decoupling approach," *Journal of Computer-Aided Molecular Design*, pp. 1–6, 2020.

[113] P. A. Kollman, I. Massova, C. Reyes, B. Kuhn, S. Huo, L. Chong, M. Lee, T. Lee, Y. Duan, W. Wang, *et al.*, "Calculating structures and free energies of complex molecules: combining molecular mechanics and continuum models," *Accounts of chemical research*, vol. 33, no. 12, pp. 889–897, 2000.

[114] M. S. Lee, F. R. Salsbury Jr, and C. L. Brooks III, "Novel generalized born methods," *The Journal of chemical physics*, vol. 116, no. 24, pp. 10606–10614, 2002.

[115] V. Zoete, M. Meuwly, and M. Karplus, "Study of the insulin dimerization: binding free energy calculations and per-residue free energy decomposition.," *Proteins.*, vol. 61, pp. 79–93, 2005.

[116] J. Park, B. Kahng, and W. Hwang, "Thermodynamic selection of steric zipper patterns in the amyloid cross-β spine," *PLoS Comput. Biol.*, vol. 5, p. e1000492, 2009.

[117] B. Roux, "Continuum electrostatic behavior of a 3d-rism theory," *The Journal of Physical Chemistry B*, vol. 124, no. 34, pp. 7444–7451, 2020.

[118] P.-K. Yang, "Modifying poisson equation for near-solute dielectric polarization and solvation free energy," *Chemical Physics*, vol. 472, pp. 229–240, 2016.

[119] T. Lazaridis, "Inhomogeneous fluid approach to solvation thermodynamics. 1. theory," *The Journal of Physical Chemistry B*, vol. 102, no. 18, pp. 3531–3541, 1998.

[120] T. Lazaridis, "Inhomogeneous fluid approach to solvation thermodynamics. 2. applications to simple fluids," *The Journal of Physical Chemistry B*, vol. 102, no. 18, pp. 3542–3550, 1998.

[121] Z. Li and T. Lazaridis, "Thermodynamic contributions of the ordered water molecule in hiv-1 protease," *Journal of the American Chemical Society*, vol. 125, no. 22, pp. 6636–6637, 2003.

[122] D. J. Huggins, M. Marsh, and M. C. Payne, "Thermodynamic properties of water molecules at a protein–protein interaction surface," *J. Chem. Theory Comp.*, vol. 7, no. 11, pp. 3514–3522, 2011.

[123] C. N. Nguyen, T. Kurtzman Young, and M. K. Gilson, "Grid inhomogeneous solvation theory: hydration structure and thermodynamics of the miniature receptor cucurbit [7] uril," *The Journal of chemical physics*, vol. 137, no. 4, p. 044101, 2012.

[124] C. N. Nguyen, T. Kurtzman, and M. K. Gilson, "Spatial decomposition of translational water–water correlation entropy in binding pockets," *Journal of chemical theory and computation*, vol. 12, no. 1, pp. 414–429, 2016.

[125] J. Kraml, A. S. Kamenik, F. Waibl, M. Schauperl, and K. R. Liedl, "Solvation free energy as a measure of hydrophobicity: application to serine protease binding interfaces," *Journal of chemical theory and computation*, vol. 15, no. 11, pp. 5872–5882, 2019.

[126] L. P. Heinz and H. Grubmüller, "Computing spatially resolved rotational hydration entropies from atomistic simulations," *J. Chem. Theory Comp.*, vol. 16, no. 1, pp. 108–118, 2019.

[127] R. H. Henchman and J. A. McCammon, "Extracting hydration sites around proteins from explicit water simulations," *Journal of computational chemistry*, vol. 23, no. 9, pp. 861–869, 2002.

[128] R. P. White and H. Meirovitch, "A simulation method for calculating the absolute entropy and free energy of fluids: Application to liquid argon and water," *Proc. Natl. Acad. Sci. USA*, vol. 101, no. 25, pp. 9235–9240, 2004.

[129] I. J. General, R. Dragomirova, and H. Meirovitch, "New method for calculating the absolute free energy of binding: The effect of a mobile loop on the avidin/biotin complex," *The Journal of Physical Chemistry B*, vol. 115, no. 1, pp. 168–175, 2011.

[130] K. M. Ravikumar and W. Hwang, "Role of hydration force in the self-assembly of collagens and amyloid steric zipper filaments," *J. Am. Chem. Soc.*, vol. 133, pp. 11766–11773, 2011.

[131] X. Teng and W. Hwang, "Effect of methylation on local mechanics and hydration structure of dna," *Biophys. J.*, vol. 114, no. 8, pp. 1791–1803, 2018.

[132] T. Takino, M. Tamura, H. Miyamori, M. Araki, K. Matsumoto, H. Sato, and K. M. Yamada, "Tyrosine phosphorylation of the crkii adaptor protein modulates cell migration," *J. Cell Sci.*, vol. 116, pp. 3145–3155, 2003.

[133] B. R. Brooks, C. L. Brooks III, A. D. Mackerell Jr., L. Nilsson, R. J. Petrella, B. Roux, Y. Won, G. Archontis, C. Bartels, S. Boresch, A. Caflisch, L. Caves, Q. Cui, A. R. Dinner, M. Feig, S. Fischer, J. Gao, M. Hodoscek, W. Im, K. Kuczera, T. Lazaridis, J. Ma, V. Ovchinnikov, E. Paci, R. W. Pastor, C. B. Post, J. Z. Pu, M. Schaefer, B. Tidor, R. M. Venable, H. L. Woodcock, X. Wu, W. Yang, D. M. York, and M. Karplus, "CHARMM: the biomolecular simulation program," *J. Comput. Chem.*, vol. 30, no. 10, pp. 1545–1614, 2009.

[134] K. Melen, L. Kinnunen, R. Fagerlund, N. Ikonen, K. Y. Twu, R. M. Krug, , and I. Julkunen, "Nuclear and nucleolar targeting of influenza a virus ns1 protein: Striking differences between different virus subtypes," *J. Virol.*, vol. 81, pp. 5995–6006, 2007.

[135] W. Jorgensen, J. Chandrasekhar, and J. D. Madura, "Comparison of simple potential functions for simulating liquid water," *J. Chem. Phys.*, vol. 79, pp. 926–935, 1983.

[136] B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan, and M. Karplus, "CHARMM: A program for macromolecular energy, minimization, and dynamics calculations," *J. Comp. Chem*, vol. 4, pp. 187–217, 1983.

[137] K. Hart, N. Foloppe, C. M. Baker, E. J. Denning, L. Nilsson, and A. D. MacKerell, "Optimization of the charmm additive force field for dna: Improved treatment of the bi/bii conformational equilibrium," *J. Chem. Theory Comp.*, vol. 8, pp. 348–362, 2012.

[138] J.-P. Ryckaert, G. Ciccotti, and H. J. Berendsen, "Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes," *J. Comput. Phys.*, vol. 23, pp. 327–341, 1977.

[139] S. E. Feller and R. W. Pastor, "Effect of electrostatic force truncation on interfacial and transport properties of water," *J. Phys. Chem.*, vol. 100, pp. 17011–17020, 1996.

[140] A.-P. Hynninen and M. F. Crowley, "New faster charmm molecular dynamics engine," *J. Comput. Chem.*, vol. 35, pp. 406–413, 2014.

[141] W. Humphrey, A. Dalke, and K. Schulten, "Vmd: Visual molecular dynamics," *J. Mol. Graphics.*, vol. 14, no. 1, pp. 33–38, 1996.

[142] E. F. Pettersen, T. D. Goddard, C. C. Huang, G. S. Couch, D. M. Greenblatt, E. C. Meng, and T. E. Ferrin, "Ucsf chimera-a visualization system for exploratory research and analysis," *J. Comput. Chem.*, vol. 25, pp. 1605–1612, 2004.

[143] J. Shao, S. Tanner, N. Thompson, and T. Cheatham, "Clustering molecular dynamics trajectories: 1. characterizing the performance of different clustering algorithms," *J. Chem. Theory Comp.*, vol. 6, pp. 2312–2334, 2007.

[144] G. Lipari and A. Szabo, "Model-free approach to the interpretation of nuclear magnetic resonance relaxation in macromolecules. 1. theory and range of validity," *J. Am. Chem. Soc.*, vol. 104, pp. 4546–4559, 1982.

[145] I. Chandrasekhar, G. Clore, A. Szabo, A. Gronenborn, and B. Brooks, "A 500 ps molecular dynamics simulation study of interleukin-1β in water: Correlation with nuclear magnetic resonance spectroscopy and crystallography," *J. Mol. Biol.*, vol. 226, pp. 239–250, 1992.

[146] N. Trbovic, J.-H. Cho, R. Abel, R. A. Friesner, M. Rance, and A. G. PalmerIII, "Protein side-chain dynamics and residual conformational entropy," *J. Am. Chem. Soc.*, vol. 131, pp. 615–622, 2009.

[147] W. E. Stites and J. Pranata, "Empirical evaluation of the influence of side chains on the conformational entropy of the polypeptide backbone," *Proteins*, vol. 22, pp. 132–140, 1995.

[148] IUPAC-IUB, "Abbreviations and symbols for the description of the conformation of polypeptide chains," *J. Biol. Chem.*, vol. 245, pp. 6489–6497, 1970.

[149] B. Bagchi, "Water dynamics in the hydration layer around proteins and micelles," *Chemical Reviews*, vol. 105, no. 9, pp. 3197–3219, 2005.

[150] W. Im, M. S. Lee, and C. L. Brooks, "Generalized born model with a simple smoothing function," *J. Comput. Chem.*, vol. 24, pp. 1691–1702, 2003.

[151] V. A. Jarymowycz and M. J. Stone, "Fast time scale dynamics of protein backbones: Nmr relaxation methods, applications, and functional consequences," *Chem. Rev.*, vol. 106, pp. 1624–1671, 2006.

[152] P. Maragakis, K. Lindorff-Larsen, M. P. Eastwood, R. O. Dror, J. L. Klepeis, I. T. Arkin, M. ø. Jensen, H. Xu, N. Trbovic, R. A. Friesner, A. G. P. III, and D. E. Shaw, "Microsecond molecular dynamics simulation shows effect of slow loop dynamics on backbone amide order parameters of proteins," *J. Phys. Chem. B.*, vol. 112, pp. 6155–6158, 2008.

[153] F. Cordier, C. Wang, S. Grzesiek, and L. K. Nicholson, "Ligand-induced strain in hydrogen bonds of the c-Src SH3 domain detected by NMR," *J. Mol. Biol.*, vol. 304, pp. 497–505, 2000.

[154] S. M. Larson, A. A. Di Nardo, and A. R. Davidson, "Analysis of covariation in an sh3 domain sequence alignment: applications in tertiary contact prediction and the design of compensating hydrophobic core substitutions," *J. Mol. Biol.*, vol. 303, pp. 433–446, 2000.

[155] M. Harkiolaki, M. Lewitzky, R. J. Gilbert, E. Jones, R. P. Bourette, G. Mouchiroud, H. Sondermann, I. Moarefi, and S. M. Feller, "Structural basis for sh3 domain-mediated high-affinity binding between mona/gads and slp-76," *The EMBO journal*, vol. 22, pp. 2571–2582, 2003.

[156] H. De Loof, L. Nilsson, and R. Rigler, "Molecular dynamics simulation of galanin in aqueous and nonaqueous solution," *J. Am. Chem. Soc.*, vol. 114, pp. 4028–4035, 1992.

[157] P. A. Wood, F. H. Allen, and E. Pidcock, "Hydrogen-bond directionality at the donor h atom – analysis of interaction energies and database statistics," *CrystEngComm*, vol. 11, pp. 1563–1571, 2009.

[158] A. R. Fersht, "The hydrogen bond in molecular recognition," *Trends Biochem. Sci.*, vol. 12, pp. 301–304, 1987.

[159] J. Shi, Q. Shen, J.-H. Cho, and W. Hwang, "Entropy hotspots for the binding of intrinsically disordered ligands to a receptor domain," *Biophysical Journal*, vol. 118, no. 10, pp. 2502–2512, 2020.

[160] M. Levitt and B. H. Park, "Water: now you see it, now you don't," *Structure (London, England)*, vol. 1, no. 4, pp. 223–226, 1993.

[161] A. Damjanović, J. L. Schlessman, C. A. Fitch, A. E. García, *et al.*, "Role of flexibility and polarity as determinants of the hydration of internal cavities and pockets in proteins," *Biophysical Journal*, vol. 93, no. 8, pp. 2791–2804, 2007.

[162] H.-J. Woo and B. Roux, "Calculation of absolute protein–ligand binding free energy from computer simulations," *Proceedings of the National Academy of Sciences*, vol. 102, no. 19, pp. 6825–6830, 2005.

[163] H. Tossavainen, O. Aitio, M. Hellman, K. Saksela, and P. Permi, "Structural basis of the high affinity interaction between the alphavirus nonstructural protein-3 (nsp3) and the sh3 domain of amphiphysin-2," *Journal of Biological Chemistry*, vol. 291, no. 31, pp. 16307–16317, 2016.

[164] J. M. Ward, N. M. Gorenstein, J. Tian, S. F. Martin, and C. B. Post, "Constraining binding hot spots: NMR and molecular dynamics simulations provide a structural explanation for enthalpy- entropy compensation in sh2-ligand binding," *J. Am. Chem. Soc.*, vol. 132, pp. 11058–11070, 2010.

[165] M. S. Marlow, J. Dogan, K. K. Frederick, K. G. Valentine, and A. J. Wand, "The role of conformational entropy in molecular recognition by calmodulin," *Nat. Chem. Biol.*, vol. 6, no. 5, pp. 352–358, 2010.

[166] O. Schon, A. Friedler, S. Freund, and A. R. Fersht, "Binding of p53-derived ligands to MDM2 induces a variety of long range conformational changes," *J. Mol. Biol.*, vol. 336, pp. 197–202, 2004.

[167] E. Ozkirimli, S. S. Yadav, W. T. Miller, and C. B. Post, "An electrostatic network and long-range regulation of Src kinases," *Protein Sci.*, vol. 17, no. 11, pp. 1871–1880, 2008.

[168] K. H. DuBay, G. R. Bowman, and P. L. Geissler, "Fluctuations within folded proteins: implications for thermodynamic and allosteric regulation," *Acc. Chem. Res.*, vol. 48, no. 4, pp. 1098–1105, 2015.

[169] S.-R. Tzeng and C. G. Kalodimos, "Protein activity regulation by conformational entropy," *Nature*, vol. 488, no. 7410, pp. 236–240, 2012.

[170] H. N. Motlagh, J. O. Wrabl, J. Li, and V. J. Hilser, "The ensemble nature of allostery," *Nature*, vol. 508, no. 7496, pp. 331–339, 2014.

[171] K. A. Stafford, F. Ferrage, J.-H. Cho, and A. G. Palmer III, "Side chain dynamics of carboxyl and carbonyl groups in the catalytic function of *escherichia coli* ribonuclease H," *J. Am. Chem. Soc.*, vol. 135, pp. 18024–18027, 2013.

[172] U. Weininger, "Optimal isotope labeling of aromatic amino acid side chains for nmr studies of protein dynamics," *Methods Enzymol.*, vol. 614, pp. 67–86, 2019.

[173] D. Beece, L. Eisenstein, H. Frauenfelder, D. Good, M. Marden, L. Reinisch, A. Reynolds, L. Sorensen, and K. Yue, "Solvent viscosity and protein dynamics," *Biochemistry*, vol. 19, no. 23, pp. 5147–5157, 1980.

[174] H. Frauenfelder, G. Chen, J. Berendzen, P. W. Fenimore, H. Jansson, B. H. McMahon, I. R. Stroe, J. Swenson, and R. D. Young, "A unified model of protein dynamics," *Proceedings of the National Academy of Sciences*, vol. 106, no. 13, pp. 5129–5134, 2009.

[175] P. W. Fenimore, H. Frauenfelder, B. H. McMahon, and F. G. Parak, "Slaving: solvent fluctuations dominate protein dynamics and functions," *Proceedings of the National Academy of Sciences*, vol. 99, no. 25, pp. 16047–16051, 2002.

[176] A. Ansari, C. M. Jones, E. R. Henry, J. Hofrichter, and W. A. Eaton, "The role of solvent viscosity in the dynamics of protein conformational changes," *Science*, vol. 256, no. 5065, pp. 1796–1798, 1992.

[177] V. Réat, R. Dunn, M. Ferrand, J. L. Finney, R. M. Daniel, and J. C. Smith, "Solvent dependence of dynamic transitions in protein solutions," *Proceedings of the National Academy of Sciences*, vol. 97, no. 18, pp. 9961–9966, 2000.

[178] D. Vitkup, D. Ringe, G. A. Petsko, and M. Karplus, "Solvent mobility and the protein'glass' transition," *Nature structural biology*, vol. 7, no. 1, pp. 34–38, 2000.

[179] B. S. Marques, M. A. Stetz, C. Jorge, K. G. Valentine, A. J. Wand, and N. V. Nucci, "Protein conformational entropy is not slaved to water," *Scientific reports*, vol. 10, no. 1, pp. 1–8, 2020.

[180] T. Clackson and J. A. Wells, "A hot spot of binding energy in a hormone-receptor interface," *Science*, vol. 267, no. 5196, pp. 383–386, 1995.

APPENDIX A

Examples of CHARMM and R scripts.

```
## function for I3 of MIST
## define the dir for angles carefully!!!!

get.I3 <- function(i,j,k,dat_dir,bsize) { ##  j,k angles are together; i is the new angle dat_dir data dir, bsize: bin size


    ## dir angle of i,j,k
    dat_names_I3=dir(dat_dir)

    # i,j,k as the angle index
    x=read.table(paste(dat_dir,dat_names_I3[i],sep=""),header=F)
    y=read.table(paste(dat_dir,dat_names_I3[j],sep=""),header=F)
    z=read.table(paste(dat_dir,dat_names_I3[k],sep=""),header=F)

    ## i angle
    df <- data.frame(x)
    n=nrow(df)
    #range of angles
    x0=-180
    x1=180

    # find angle distribution
    x.nbin= ceiling((x1-x0)/bsize) + 1
    x11=x0+bsize*(x.nbin-1)
    x.bin = seq(x0,x11,length=x.nbin)
    fI.x <- findInterval(df[,1], x.bin)
```

```
27
28    # find frequency and probability with table
29    freq <-  as.data.frame(table(fI.x))
30    freq[,1] <- as.numeric(freq[,1])
31    freq[,2]=freq[,2]/n
32    m=nrow(freq)
33
34
35    ## j angle
36
37    df2 <- data.frame(y)
38    n2=nrow(df2)
39    y0=-180
40    y1=180
41
42    # find angle distribution
43    y.nbin= ceiling((y1-y0)/bsize) + 1
44    y11=y0+bsize*(y.nbin-1)
45    y.bin = seq(y0,y11,length=y.nbin)
46    fI.y <- findInterval(df2[,1], y.bin)
47
48    # find frequency and probability with table
49    freq2 <-  as.data.frame(table(fI.y))
50    freq2[,1] <- as.numeric(freq2[,1])
51    freq2[,2]=freq2[,2]/n2
52    m2=nrow(freq2)
53
54    ## then k angle
55    df3 <- data.frame(z)
56    n3=nrow(df3)
57    z0=-180
58    z1=180
59
```

```
    z.nbin= ceiling((z1-z0)/bsize) + 1
    z11=z0+bsize*(z.nbin-1)
    z.bin = seq(z0,z11,length=z.nbin)
    fI.z <- findInterval(df3[,1], z.bin)

    # find frequency and probability with table
    freq3 <-  as.data.frame(table(fI.z))
    freq3[,1] <- as.numeric(freq3[,1])
    freq3[,2]=freq3[,2]/n3
    m3=nrow(freq3)

    ############################
    ###################
    # freq4 & freq5
    ## freq4 in I3: P(x_j,x_k)
    freq4 <-  as.data.frame(table(fI.y,fI.z))
    freq4[,1] <- as.numeric(freq4[,1])
    freq4[,2] <- as.numeric(freq4[,2])
    freq4[,3]=freq4[,3]/n
    m4=nrow(freq4)


    ## freq5 in I3: P(x_i,x_j,x_k)
    freq5 <-  as.data.frame(table(fI.x,fI.y, fI.z))
    freq5[,1] <- as.numeric(freq5[,1])
    freq5[,2] <- as.numeric(freq5[,2])
    freq5[,3] <- as.numeric(freq5[,3])
    freq5[,4]=freq5[,4]/n
    m5=nrow(freq5)


    # build the matrix to avoid looping
    rep_freq <- rep(freq[,2],m2*m3)
```

```
93      rep_freq4 <- rep(freq4[,3],each=m) # notice the "each" in the second rep,
        copy each element m times
94
95      freq_matrix <- cbind(i=rep_freq,jk=rep_freq4,ijk=freq5[,4])
96
97      #change all the zero to NA
98      freq_matrix[freq_matrix==0]=NA
99      # remove rows containing NA
100     freq_matrix <- na.omit(freq_matrix)
101
102     # calculation of I3
103     I3=apply(freq_matrix, 1 ,function(row) row[3]*log(row[3]/row[1]/row[2]) )
104
105     I3.tot <- sum(I3)
106     return(I3.tot)
107 }
```

Listing A.1: R code for calculating third order mutual information in Eq. 2.6

```
* cube_exclude.inp: find xyz coord and identify cube id that are visited by
   protein atoms
* by Jie 04/14/2020
*

bomlev -1

! stream forcefield files
stream ~/include/include.str

! set cube info based on dat file from SMAP
set reso 0.6
set xmin  -23.1576
set ymin  59.8932
set zmin  -57.4977
! calc the negative value for later "add"
calc nxmin -1 * @xmin
calc nymin -1 * @ymin
calc nzmin -1 * @zmin

! read reference structures and dcd traj files
! ....

! define the same region as in smap.inp to superimpose!!
define stable sele resi 136:143 .or. resi 150:154 .or. resi 157:162 .or. resi
   169:173 .or. resi 180:187 end
define base sele type CA end
define a1 sele (segi SH3 ) .and. stable  .and. base end


! start the frame searching
traj query unit 11
traj first 11 nunit @{ndcd} begin ?START skip ?SKIP
```

```
32
33 ! supress output to reduce the file size
34 prnlev 1 node 0
35
36      ! j: frame #
37      set j 1
38      label L1
39      traj read
40
41      ! orient the frame as in smap
42      coor orie rms sele a1 end
43
44      ! x-dir
45      scalar x add @nxmin
46      scalar x divi @reso
47      scalar x INTE
48      scalar x add 1
49
50      ! y-dir
51      scalar y add @nymin
52      scalar y divi @reso
53      scalar y INTE
54      scalar y add 1
55
56      ! z-dir
57      scalar z add @nzmin
58      scalar z divi @reso
59      scalar z INTE
60      scalar z add 1
61
62      ! print coor, should be faster, use awk to extact coord later
63      prnlev 2 node 0
64      print coor sele segi SH3 .or. segi LIG end
```

```
    prnlev 1 node 0

    incr j by 1
    if @j .le. @ntot goto L1
stop
```

Listing A.2: CHARMM code for finding cube ID that's visited by any of the protein atom at each frame. The results can be later extracted for calculating occupancy.

```
# get the shell out of solmap dat, and calculate energy G=-kT*log(rho)

# grid resolution
reso = 0.7

## gas constant (cal/mol/K)
r=1.987
## Temperature (K)
T <- 300
ptm <- proc.time() # system starting timing

# read in smap results as well as reference structure coordinate information.
shell_raw <- read.table(paste("ns1b_smap_",reso,"_stable_orie.dat",sep=""),
    header=F)
prot_raw <- read.table("../../../../../../coor_dyna/ns1b_mean_struc_min_200_
    orie_heavy.dat",header=F) # protein file only contains coord
shell_raw <- data.frame(shell_raw)
n_raw <- nrow(shell_raw)


#extract cube coord
shell <- shell_raw[,c(2,4,6)] # shell only contains coord
shell <- as.matrix(shell)
prot_raw <- as.matrix(prot_raw)

cutoff <- 4.5 # cutoff for determining hydration shell by distance
flag_cut4.5 <- rep(FALSE,nrow(shell)) # by default all are out of real shell

# find cubes that has min disance < 4.5
for(i in 1:nrow(shell)) {
    tmp <- shell[i,]
    dist_t <- sweep(prot_raw,2,tmp) # subtract x-x0,y-y,z-z0
    dd_t <- apply(dist_t,1,function(x) {sqrt(sum(x^2))} )
```

```
    if(min(dd_t) < cutoff ) {flag_cut4.5[i] <- TRUE}
}


#########################
#Calculating the energy excluding the cubes visited by protein

# import info about cubes that has ever been visited by any protein atoms
flag_v <-  read.table(paste("flag_",reso,"_cubes_rcut15_visited_stable.dat",
    sep=""),header=F)
flag_v <-  as.logical(flag_v[,1])

# density flag
rho_min.0.034 <- 0.0345306
flag_rho.0.034 <- shell_raw[,7] > rho_min.0.034

# flag for cubes within distance cutoff, density cutoff and unvisited
flag_vc.cut4.5_0.034 <- (!flag_v) & flag_cut4.5 & flag_rho.0.034

n_vc.cut4.5_0.034 <- sum(flag_vc.cut4.5_0.034)
rho_vc.cut4.5_0.034 <- shell_raw[flag_vc.cut4.5_0.034,7] # density of each
    cube

# calculating energy, unit kcal/mol
E_vc.cut4.5_0.034 <- -r*T*log(rho_vc.cut4.5_0.034)/1000
E_vc_tot.cut4.5_0.034 <- sum(E_vc.cut4.5_0.034)


ftm <- proc.time() -ptm  # total wall time
```

Listing A.3: R code for calculating density-based energy in Eq. 2.9