Conformational ensemble modeling of multi-domain proteins with experimental data

Zhu, Guanhua

2017

Zhu, G. (2017). Conformational ensemble modeling of multi-domain proteins with experimental data. Doctoral thesis, Nanyang Technological University, Singapore.

<http://hdl.handle.net/10356/72807>

<https://doi.org/10.32657/10356/72807>

**NANYANG
TECHNOLOGICAL
UNIVERSITY**

SINGAPORE

CONFORMATIONAL ENSEMBLE MODELING OF MULTI-
DOMAIN PROTEINS WITH EXPERIMENTAL DATA

ZHU GUANHUA

SCHOOL OF BIOLOGICAL SCIENCES

2017

CONFORMATIONAL ENSEMBLE MODELING OF MULTI-
DOMAIN PROTEINS WITH EXPERIMENTAL DATA

ZHU GUANHUA

SCHOOL OF BIOLOGICAL SCIENCES

A thesis submitted to the Nanyang Technological University

in partial fulfilment of the requirement for the degree of

Doctor of Philosophy

2017

Acknowledgements

First and foremost, I would like to express my deepest gratitude to my supervisor Asst. Prof. Lu Lanyuan, for offering me the opportunity to pursue my Ph.D. degree and guiding me throughout my Ph.D. journey. His valuable advice, professional guidance and tremendous support have been of great help in my Ph.D. research and future career.

I would also like to express my sincere thanks to my thesis advisory committee members Assoc. Prof. Jaume Torres and Asst. Prof. Yang Liang for their insightful suggestions on my research work.

I would like to express my gratitude to Prof. Gerhard Grüber and Dr. Saw Wuan Geok for the experimental X-ray scattering data and their great efforts in the research work reported in Chapter 2 of this thesis. I also appreciate the help from Prof. Yang Daiwen and Dr. Liu Wei for providing the experimental NMR data, as well as Prof. Shen Zuowei, Assoc. Prof. Ji Hui and Dr. Bao Chenglong for writing the mathematic programs, which contributes to the research work reported in Chapter 3 of this thesis.

I am grateful to my current group members Dr. Anjaiah Nalaparaju, Dr. Chen Wenwen, Mr. Christian Bope and Mr. Zhang Haiping as well as the former members Dr. Tong Dudu, Dr. Xia Fei and Dr. Li Xiuting, for their continuous support and

numerous help in my research. My sincere appreciations are also extended to Assoc. Prof. Mu Yuguang and his group members for the inspiring discussions.

I would like to thank Nanyang Technological University for the financial support and the School of Biological Sciences for the academic opportunities offered for my Ph.D. study.

Finally, I would like to express my special thanks to my families for their unconditional love and support.

Table of Contents

Acknowledgements.....	i
Table of Contents.....	iii
List of Figures	vii
List of Tables	xvi
Abbreviations.....	xviii
Abstract.....	xxi
1 Introduction	1
1.1 Simulation and experiment.....	1
1.2 Experiments studying multi-protein/multi-domain complexes.....	4
1.3 Computation models studying multi-protein/multi-domain complexes.....	5
1.3.1 Atomistic model.....	5
1.3.2 Coarse-grained model	8
1.4 Ensemble construction with experimental data.....	11
1.5 Aims and outlines.....	15
2 Coarse-grained molecular modeling of solution structure ensemble of dengue virus NS5 with SAXS intensity	19

2.1	Background	19
2.2	Methods.....	21
2.2.1	Coarse-grained model of NS5.....	21
2.2.2	MD simulations.....	25
2.2.3	Pool structure characterization.....	26
2.2.4	Ensemble construction with SAXS data	27
2.3	Coarse-grained MD simulations.....	29
2.4	Structure pool and energy landscape.....	30
2.5	Ensemble construction with SAXS data	32
2.6	Incorporation with Ensemble Optimization Method.....	42
2.7	Validation of low-energy conformations in all-atom simulations	49
2.8	Theoretical investigation of the noise in data fitting.....	53
2.9	Energy characterization of a structure pool from REMD simulations.....	55
2.10	Conclusion.....	59
3	Structure ensemble modeling of PubRRM12 with energy-based pool structures and experimental distance constraints	61
3.1	Background	61

3.2	Methods	64
3.2.1	Atomistic model of PubRRM12	64
3.2.2	Sampling and simulation details	68
3.2.3	Pool generation and characterization	70
3.2.4	Ensemble construction with experimental PRE rates	72
3.3	Model validation and sampling efficiency	74
3.4	Energy landscape of inter-domain interactions	79
3.5	Ensemble construction with experimental data.....	82
3.6	Conformation analysis on constructed ensembles	87
3.7	The effect of regularizations	91
3.8	Ensemble construction with a topology-based pool.....	96
3.9	The role of linker	99
3.10	Conclusion.....	102
4	Discussion.....	104
4.1	Structural information from experimental constraints	104
4.2	Over-fitting challenge	105
5	Concluding remarks.....	108

References	110
Publications	129

List of Figures

Figure 1.1 A schematic illustration of the pool-based methods for ensemble construction with experimental data. Candidate structures generated by either topology-based or physics-based methods constitute a pool that would be subsequently subjected to experimental data fitting to construct the structure ensemble. 13

Figure 2.1 Atomistic structure and coarse-grained representation of the multi-domain protein DENV2 NS5. Each amino acid is coarse-grained as one bead in the structure model. RdRp domain (green) and MTase domain (cyan) are connected via a ten-residue flexible linker (orange). 22

Figure 2.2 Domain-domain orientation of DENV2 NS5. The orientation is defined by two spherical angles φ and θ . MTase domain (purple) and RdRp domain (green) are connected via a flexible linker (orange). 27

Figure 2.3 The domain-domain orientations of DENV2 NS5 structures from the simulated pool and the energy landscape. (a) The 1015 pool structures (red points) of DENV2 NS5 from MD simulations are projected to domain-domain orientation map on the basis of their φ and θ angles. (b) It shows the energy landscape of DENV2 NS5 based on the inter-domain energies of simulated pool structures. The landscape

is gridded with a $12^\circ \times 6^\circ$ grid size and energy value is defined in the color bar (kcal/mol)..... 31

Figure 2.4 The evolution of χ^2 with a loose energy constraint on the candidate pool. The χ^2 values are illustrated for the top 1%, 5%, 10%, 15%, 20%, 25%, 30%, 40%, 50% low-energy and the 100% ensembles. The inset figure is a comparison between the SAXS profile calculated from the top 5% low-energy ensemble and that from SAXS experiments..... 36

Figure 2.5 Domain-domain orientations on the energy landscape and domain contacting segments on the RdRp domain of the 5%, 15%, 30% low-energy and the 100% ensembles. (a) The conformations (blue triangle dots) in the constructed ensembles are projected on the energy landscape of domain-domain orientation. The energy values can be referred to Figure 2.3b. (b) RdRp domain of DENV2 NS5 is shown as ribbon, underlining its domain-contacting residues (red). 38

Figure 2.6 The structural characterization of the top 5% low-energy ensemble. MTase, RdRp and linker are colored in purple, green and orange, with the domain-contacting residues on the RdRp domain highlighted in red. The zoom-in figures illustrate the conserved electrostatic interactions between the linker residues E268/E270 (blue) and the RdRp residues R595/R596 (gray)..... 40

Figure 2.7 The structure ensembles constructed by NNLS and EOM without the energy constraint by employing the same simulated pool. (a) It shows the 100% NNLS-based ensemble consisting of some high-energy conformations. ΔE is the difference of the structure energy with respect to the average energy of all the 1015 simulated pool structures (kcal/mol). ΔE is decomposed into its two components of hydrophobic energy ΔE_H and electrostatic energy ΔE_{ele} . The two structures of large θ angles are circled with the dashed boxes in the energy landscape. Protein structure is colored as that in Figure 2.2. (b) It shows the conformations and their energetic profiles of the EOM-based ensemble without the energy constraint. 42

Figure 2.8 The EOM-based ensembles from the simulated pool with and without the energy constraint. The 5%, 15%, 30% low-energy and the 100% ensembles (blue triangle dots) recovered from experimental data with GA-based EOM approach are projected on the domain-domain orientation landscape. The energy values can be referred to Figure 2.3b. 45

Figure 2.9 A comparison between the standard EOM ensemble and the simulated optimal ensemble of DENV2 NS5. (a) The standard EOM scheme with its 10000-structure pool produced an ensemble of four structures (deep green triangle dots), which is depicted on the energy landscape of domain-domain orientation. The two structures of the top 5% low-energy simulated ensemble are shown (blue triangle dots) in the landscape. (b) The structure comparison is illustrated between the

standard EOM ensemble and the top 5% low-energy simulated ensemble. All the structures have their RdRp domain structure aligned (translucent green envelop) to reveal the MTase-binding position on RdRp domain. The most and second most populated structures (deep pink and yellow respectively) from the standard EOM ensemble have their MTase domain (shown as ribbon) binding to a close RdRp surface as the two structures (green and cyan) in the top 5% low-energy simulated ensemble. The two less-populated EOM structures (blue and purple) have their MTase domain binding to a very different surface on RdRp domain..... 48

Figure 2.10 The RMSD evolution with simulation time in the conventional all-atom MD simulations starting from the simulated ensemble structures. The backbone RMSD for MTase, RdRp and the complex are shown as blue, orange and red lines respectively. The cases of (a) and (b) are the all-atom simulations with the two 5% low-energy ensemble structures as the initial structures. The structure comparison was shown between the initial structure and the representative structure of the all-atom simulation (green and red ribbon respectively), with their RdRp structure well-aligned. (c) and (d) are the all-atom MD simulations starting from two high-energy structures that exist in both NNLS and EOM-based structure ensembles without the energy constraint. The structure alignment was performed between the initial structure and the last-frame structure of the all-atom simulation trajectory. 52

Figure 2.11 The ensembles constructed from a pseudo-experimental SAXS profile projected on the domain-domain orientation map. The 5%, 15% and 30% low-energy as well as the 100% simulated pools were used for ensemble construction with synthetic SAXS data. The qualities of fit are $\chi^2 = 0.87, 0.86, 0.86$ and 0.84 respectively. 55

Figure 2.12 Domain-domain orientation distribution of a structure pool resulted from REMD simulations. 24 replica simulations were run on DENV2 NS5 for 400 ns each, ranging from 300 to 613 K. The structures from the ensemble at 300 K were taken as a pool for further characterization. 57

Figure 3.1 Structure of the multi-domain protein PubRRM12 and the domain-domain orientation. (a) RRM1 (left) and RRM2 (right) domains have stable conformations (green) individually, while linker and some loop segments are flexible (red) with their residue number indicated. (b) The RRM1-RRM2 orientation is defined by two spherical angles φ and θ . The self-rotation (blue arrow) of RRM2 is not included in the spherical coordinates. (c) RRM2 is pulled away from RRM1 (center) to 160 uniformly-distributed positions (red beads)..... 65

Figure 3.2 Backbone RMSD of the structures from the MD simulations with respect to the experimental structures of individual domains. (a) Small RMSD values are found for the simulated structures with the structure-based potential with respect to the experimental RRM1 and RRM2 structures. (b) Without the structure-based

potential, the simulated structures have much larger RMSD values against the experimental domain structures. 75

Figure 3.3 The initial structures in the production MD simulations and the sampled conformational space in the production MD simulations. Left: The 661 initial structures cover all the available orientations between RRM1 (green surface) and RRM2 (ribbon representations). Right: Starting from the 661 structures, the MD simulations generated a very complete sampling of domain-domain orientation space. 77

Figure 3.4 A comparison of the sampling efficiency between the simulations with and without the enhanced sampling methods. (a) and (b) are the domain-domain orientation space sampled by the 400-ns MD simulation without any sampling method, starting from two different structures respectively. (c) is the orientation space sampled by a 40-ns REMD simulation from 300K to 720K. RRM1 and linker are colored in green and RRM2 is colored in blue. (d) is the case for our sampling method, same as Figure 3.3..... 78

Figure 3.5 A comparison of the sampling efficiency between the plain and the enhanced sampling simulations with respect to the conformational space depicted by RMSD and domain-domain distance. The plain MD simulations (a and b), the REMD simulation (c) and the enhanced sampling simulations (d) have been described in Figure 3.4. The two-dimensional conformational space here is the

backbone RMSD against the same reference structure and the domain-domain COM distance. 79

Figure 3.6 Energy landscape of inter-domain interactions of PubRRM12. The energy of each domain-domain orientation represented in an $18^\circ \times 9^\circ$ grid is indicated in colors defined by the color bar (kcal/mol). 82

Figure 3.7 PRE Q-factor evolution with ensemble size for top 5%, 10%, 20%, 30%, 40% and 100% ensembles. 85

Figure 3.8 Conserved domain-domain orientation and domain contacting interface identified by the constructed low-energy ensembles. (a). It shows the domain-domain orientation distribution of ensemble component structures of top 5%, 20%, 30% and 100% ensembles, shown as blue points. The red area indicates low-energy surface on the energy landscape with a cutoff at top 5% low energy. (b). Ribbon representation of RRM1 is shown with the contacting residues highlighted in red whose contact numbers are ranked within top 20% among all RRM1 residues. 90

Figure 3.9 Three conformations of minor domain-domain orientations in the constructed 100% ensemble and their energetic profiles. ΔE (kcal/mol) is the difference between the conformation energy and the average energy over all 4003 representative structures. ΔE_{ele} , ΔE_{vdw} , and ΔE_{sol} are the electrostatic, van der Waals

and solvation components of energy difference. RRM1 structures are colored on the basis of its secondary structure and RRM2 structures are shown in green..... 93

Figure 3.10 Domain-domain orientations of the constructed NNLS ensembles of different energy levels. Domain-domain orientations of ensemble components from reweighted 5%, 20%, 30% low-energy and 100% NNLS ensembles are shown as blue points for major conformations and green points for minor conformations. The red area indicates low-energy surface on the energy landscape with a cutoff at top 5% low energy..... 96

Figure 3.11 PRE Q-factor evolution with ensemble size for the ensemble construction with a topology-based structure pool sized 5000..... 98

Figure 3.12 The structure ensemble recovered by MQP from a topology-based pool. Domain-domain orientation distribution (blue points) of 14-structure ensemble recovered from the topology-based pool, two of whose component structures are located away from the low-energy orientation (light red) corresponding to RRM1 beta-sheets..... 99

Figure 3.13 Conformations (blue points) whose linker (red part in PubRRM12 structure) cannot connect two domains caused by the linker length limit. Structure (a) is the situation that certain domain-domain orientations are unachievable when the two linker-attached residues are in far-end position. For structure (b), domain

self-rotation is restricted by the linker despite of the availability of the domain-
domain orientation. 101

Figure 3.14 Potential energy decomposition of representative structures. Potential
energy of representative structures is decomposed into RRM1-RRM1, RRM1-linker
and RRM2-linker components (kcal/mol). 102

List of Tables

Table 1.1 Popular computational approaches to construct structure ensembles with experimental data.	15
Table 2.1 Structure ensemble of DENV2 NS5 recovered from experimental SAXS data with energy constraints. The four columns in the table refer to the energy-level of the pool/ensemble, the number of candidate pool structures, the quality of fit χ^2 , and the size of the recovered ensemble, respectively.	35
Table 2.2 The profiles of the sub-pools based on energy levels. The structures from the ensemble at 300 K in REMD simulations were ranked on the basis of the inter-domain energy (kcal/mol) from low to high, followed by dividing the pool into 10 energy-ranked sub-pools.	58
Table 3.1 Energy difference between different energy-ranked sub-pool of PubRRM12 structures from REMD simulations. 200 structures from a 40-ns REMD simulation were collected as a structure pool and ranked according to the MMPBSA energy (kcal/mol). This pool was further divided in to 5 sub-pools on the basis of structure energy, with the average energy of the highest-energy sub-pool shifting to zero energy.	87

Table 3.2 Profiles of the reweighted NNLS ensembles. Columns correspond to low-energy ensembles, Q-factors, the number of non-zero conformations in the reweighed ensemble, the number of major conformations in the reweighed ensemble, and the cumulative population occupied by the major conformations. 95

Abbreviations

CG	Coarse-grained
COM	Center of mass
DENV	Dengue virus
EM	Electron microscopy
EOM	Ensemble Optimization Method
IDP	Intrinsically disordered proteins
MD	Molecular dynamics
MES	Minimal Ensemble Search
MJ	Miyazawa-Jernigan
MMPBSA	Molecular mechanics Poisson-Boltzmann/surface area
MQP	Multiple Q-factor Pursuit
MTase	Methyltransferase

MTSL	S-(2, 2,5,5-tetramethyl-2,5-dihydro-1H-pyrrol-3-yl)methyl methanesulfonylthioate
NMR	Nuclear magnetic resonance
NNLS	Non-negativity least squares
NOE	Nuclear Overhauser effect
NS5	Non-structural protein 5
ns	nanosecond
PDB	Protein Data Bank
PRE	Paramagnetic relaxation enhancement
ps	picosecond
Pub1	Poly(U)-binding protein
RdRp	RNA-dependent-RNA polymerase
REMC	Replica-exchange Monte Carlo
REMD	Replica-exchange molecular dynamics
RMSD	Root-mean-square deviation

SAS	Sample and Select
SAXS	Small-angle X-ray scattering
SES	Sparse Ensemble Selection

Abstract

The structural variation of multi-domain proteins with flexible components mediates many biological processes. Their structure determination is crucial for understanding the mechanism of their interactions and functions, but it remains a challenge to derive the structure ensemble of flexible biomolecules solely from experimental data. Computational methods assist in the interpretation of the time- and spatial-averaged experimental data for a correct structural characterization. A conformational ensemble can be determined by selecting a weighted combination of representative structures from a given structure pool, yielding the best fit to experimental measurements. However, a raw numerical data fitting causes the over-interpretation of the experimental data, which would recover the structure ensemble of wrong conformations.

In our study, we have employed the idea of hybrid models to simulate multi-domain proteins. The intra-domain structures are modeled with the structure-based potential and the inter-domain interactions are characterized with the physics-based energies. For ensemble construction with experimental data, structure energy is proposed as a physics-based regularization to minimize the problem of data over-fitting. The absence of energy regularization exposes ensemble construction to the noise from high-energy structures. We have demonstrated the importance of the physics-based

regularization to interpret experimental data to avoid arbitrary and spurious conformational representations.

We have implemented our computational model and ensemble optimization to two systems of multi-domain proteins on the coarse-grained and atomistic levels respectively. Firstly, we simulate the dengue virus 2 (DENV2) non-structural protein 5 (NS5) at the coarse-grained level and construct its conformational ensemble with experimental small-angle X-ray scattering (SAXS) data. The energy regularization assists to recover a conformational ensemble from the SAXS profile, which successfully reveals the domain-domain orientation and domain contacting interface of NS5. Moreover, we applied the computational modeling scheme to a di-domain complex of a poly(U)-binding protein (Pub1), followed by ensemble construction with its experimental paramagnetic relaxation enhancement (PRE) data that correspond to inter-atomic distance constraints. A conformational ensemble of Pub1 has been recovered from the low-energy structures and with the ensemble-size constraint. This ensemble is in good agreement with experimental PRE rates and also supported by the experimental data of chemical shift perturbations.

1 Introduction

1.1 Simulation and experiment

Molecular dynamics (MD) simulation has served as a crucial role in the studies of biological systems including protein, nucleic acids, carbohydrates and lipids, since it was firstly employed to model the dynamics of a folded protein in 1977 [1]. In an MD simulation, the forces acting on atoms and molecules drive their movement, which is solved by numerical integration of the Newton's equation of motion $F = ma$. The forces between atoms are typically determined by a potential energy function (potential), which is also referred as an empirical all-atom force field. The commonly used atomistic force fields include Amber [2], Charmm [3], Gromos [4] and OPLS-AA [5]. Thanks to the enhanced computer power in recent decades, all-atom models based on force fields have received great popularity for computational simulations on biomolecules generally in practice on nanosecond timescales [6]. However, many biological processes occur on the timescale of microseconds, milliseconds or even seconds, such as protein major conformational change, ligand binding, protein folding and protein-protein association [6-8]. Therefore, all-atom models are limited for the simulations on large length scales and long time scales due to high computational costs. In contrast with atomistic models, coarse-grained (CG) models simplify biomolecules by representing a number of atoms as one bead and accelerate molecular dynamics simulations. Moreover, an increasing amount of

large biomolecules are investigated by medium or low resolution methods such as cryo-electron microscopy (cryo-EM) and small-angle X-ray scattering (SAXS). In those cases, CG models may be more appropriate considering its cost-effective feature and the given low-resolution experimental structural information [7].

Comparing to experimental data is an important aspect in the validation of molecular dynamics simulations. Simulation requires experimental data as criteria to check its accuracy and improve its methodology [9]. Meanwhile, simulation serves as an aid to experiments. Experimental measurements offer people the access to the understanding of structural characteristics of biomolecules, while these measurements are time- and spatial-averages of particular measurable quantities. The structure can be directly obtained from experimental measurements only when there exists one predominant conformation in the ensemble, such as in X-ray crystallography; otherwise the structural characterization is highly challenging from ensemble average of experimental measurements [10]. Therefore, experimental data generally reflects the average measurements instead of “real structures” in view of the dynamic nature of biomolecules [9-11]. For example, a T4 lysozyme mutant has two states in solution, but the experimental observables that result from the average over the ensemble correspond to neither of the two physically meaningful states [12]. In comparison, molecular dynamics simulation produces an ensemble or a trajectory of molecules that provides the details of particle motions, thus aiding the interpretation of experimental measurements. For example, Nuclear Overhauser

effect (NOE) intensities from Nuclear magnetic resonance (NMR) experiments in some cases are highly possible from multiple molecular conformations [13]. SAXS scattering profiles calculated from simulated ensembles showed a better consistency with the experimental data rather than those calculated from single crystal structures [14].

A molecular dynamics simulation would be considered as realistic and reliable if it is able to generate a structure ensemble in good agreement with experimental data, but this scenario does hardly happen on large biomolecules. van Gunsteren et al. summarized the factors causing the failure of observation of an agreement between experiment and simulation [10, 11, 15]. The primary factor is the insufficiently accurate simulation mainly including inaccurate force field, insufficient sampling, crude approximations for solving the equations of motion, omitted relevant degrees of freedom, and inappropriate boundary conditions. Another possible reason is the inaccurate experimental measurements. Besides, the experimental measurement and the simulated quantity might be subjected to different systems (e.g. crystal or solution environment) or under different conditions (e.g. temperature, pH, etc.) or averaged differently (e.g. regarding time or spatial extent).

1.2 Experiments studying multi-protein/multi-domain complexes

Multi-protein complexes mediate many biological functions. For instance, the dimerization of the inhibitor of κ B kinase subunit β (IKK β) is required for the activation of IKK β , which would be involved in the mediation of immune and inflammatory responses [16, 17]. Another example is the endosomal sorting complex required for transport (ESCRT), a complex of multi-domain proteins, which catalyzes membrane remodeling reactions [18]. Two mechanism models, “lock and key” and “induced-fit”, have been proposed for protein-protein interactions. The former model takes individual proteins as rigid bodies with almost no conformational flexibility. However, the “lock and key” model cannot explain the case that proteins bind various substrates of different shapes [19]. In the “induced-fit” model [20], binding would induce conformational changes of proteins, thus driving interacting proteins to fit each other.

The structures of multi-protein and multi-domain complexes can be solved or characterized by a number of experimental techniques. X-ray crystallography offers high-resolution structures of tightly bound complexes, but crystal structures are static snapshots of certain conformations in the condition of crystal. Techniques subject to aqueous environment like SAXS and NMR spectroscopy provide dynamic structural information of biomolecules that are close to their biologically functional conditions than crystal. SAXS experiments yield a one-dimensional curve of scattering

intensities from the biomolecules, from which structural features on medium and low resolutions can be derived such as the size and shape of the biomolecules [21]. NMR spectroscopy methods provide distance and angular information on atomistic resolution, including chemical shifts, scalar couplings, residual dipolar couplings and paramagnetic relaxation enhancement (PRE) effects, which assists the structural characterization of biomolecules [22]. The experimental methods of SAXS and PRE are elaborated in details in the following chapters in this thesis, because the experimental data from these two methods were used in structure ensemble construction in this work.

1.3 Computation models studying multi-protein/multi-domain complexes

1.3.1 Atomistic model

Theoretically, a brute-force MD simulation with an empirical all-atom force field is able to simulate the structures of protein-protein or domain-domain complexes. However, it rarely works for large biomolecular systems in practice. One bottleneck is the insufficient sampling. All-atom MD simulations are restricted by the limited simulation timescale that is dramatically shorter than biomolecular processes ranging from microseconds to seconds like protein conformation change and protein-protein binding [23-29]. Besides, many biomolecules have rugged energy landscape

consisting of multiple minima, and the energy barrier between minima traps an MD simulation locally and prevents it from jumping out to explore other important states [30]. For example, Jones et al. performed brute-force atomistic simulations on protein-protein interactions in an inhibitor of κ B kinase- β oligomer [31]. The root-mean-square deviation (RMSD) values for the dimers showed a continuous change against the starting structures throughout the 10 ns simulations, indicating the atomistic simulation does not achieve convergence. In recent decades, a number of enhanced-sampling methods have been developed, such as metadynamics [32] and replica-exchange molecular dynamics (REMD) [33]. Notably, despite of the development of the intelligent methods for improving sampling, sufficient sampling remains highly challenging for large macromolecular complexes, especially for the flexible molecules of multiple conformation states. The other bottleneck is the inherent weaknesses in force-field parameterizations [15]. For protein-protein and domain-domain complexes, empirical all-atom force field does not distinguish intra-unit from inter-unit interactions. A well-folded monomer unit, like an individual protein or domain in complexes, typically preserves its conformation in protein-protein associations, and conformational flexibility majorly comes from the inter-monomer dynamic motions [34, 35]. In the computational modeling of domain-domain or protein-protein interactions, large deformation of the individual unit probably results from the inaccuracy of the force field. To preclude the deformation of the individual well-structural unit in MD simulations, the structure-based models

are widely adopted to simulate the intra-unit structure. The general idea of structure-based models is to take the native structure of biomolecules from experiment as the global energy minimum. A popular structure-based model is Gō model or Gō-type model [35, 36], whose theoretical foundation is the energy landscape theory of protein folding [37, 38]. In energy landscape theory, protein has a funnel-like energy landscape with its folded/native state at the bottom of the funnel corresponding to the global minimum in the energy landscape. Gō model utilizes a structure-based potential by adopting the native structure as the global energy minimum and simulates the stable conformation of individual well-structured unit in protein complexes on both coarse-grained and all-atom levels. Atomistic Gō model has also been combined with an atomistic force field to simulate proteins. For instance, one hybrid model used atomistic force field to describe short-range interactions (bonds, angles, dihedrals, and 1-4 short-range van der Waals potentials) and Gō model for long-range interactions to study protein folding and large scale conformational transitions [39]. This hybrid strategy can be extended to characterize inter-unit association with an atomistic force field and model intra-unit structure with a Gō model.

An alternative approach of protein-protein docking, driven by the “lock and key” mechanism, is originally developed for the determination of the native structure of a complex from its unbound components [40, 41]. Generally, docking approaches take individual folded structural unit as rigid body and carry out a comprehensive search

of the conformations of protein complexes. The search of conformations has also been implemented on coarse-grained levels [42, 43] despite of the typical docking methods based on the atomistic resolution. Each conformation is assessed on the basis of a scoring function, and a conformation with a good score is supposed to have a higher binding affinity. Many researchers have developed their own scoring functions [44-52], while they typically include effects of solvation, hydrophobicity, and electrostatics [53]. However, the simplified rigid-body representation of a single protein is not convinced physically in light of the dynamic nature of biomolecules. Importantly, flexible loops and linkers in protein complexes typically mediate protein-protein or domain-domain associations, and they are highly flexible structures in solution [53, 54]. The modeling of those flexible segments in protein complexes is beyond conventional docking programs. Compared with protein-protein docking approaches, MD simulations adopt a more sophisticated potential for molecular interactions and incorporate more exhaustive flexibility of molecular structures.

1.3.2 Coarse-grained model

The recent decade has witnessed a significant increase in the development and application of CG models, especially those on large biomolecular complexes [55]. The simplification of several atoms represented by one coarse-grained bead

decreases the number of degrees of freedom of the biomolecules, which consequently smoothens the free energy landscape to accelerate the molecular dynamics simulations [56, 57]. Atomistic structures can be reduced to different levels of resolutions, ranging from the most popular representation of one amino acid per bead to much cruder representations. For example, Voth et al. developed ultra-coarse-grained model where the entire protein domains and subdomains in actin filament are represented as single coarse-grained particles [58].

CG models on the residue-based resolution are widely adopted to study protein-protein and domain-domain complexes. Takada categorized CG models into structure-based and structure-independent models [55]. The structure-based one, like Gō model, has been described in Chapter 1.3.1. Structure-based models assume that protein folding is guided by interactions that stabilize its native structure, and thus the native structure is taken as the global energy minimum in the energy landscape. As a result, structure-based models are popular for near-native protein dynamics, such as the well-folded unit in protein complexes. However, to characterize the inter-protein or inter-domain associations that are frequently mobile without native structures, structure-independent models are required as a complementary use. On the residue-resolution level, a number of CG force fields were developed to model the interactions between the residue types [59]. Here, we highlight Miyazawa-Jernigan (MJ) potential [60, 61] that determines the heterogeneous interaction strengths for different types of residue pairs. MJ potential is a statistical potential for

non-bonded interactions of different residue pairs characterized from a huge amount of experimentally determined protein structures.

Kim et al. developed coarse-grained models to simulate multi-protein complexes and validated the models by performing replica-exchange Monte Carlo (REMC) simulations on various protein complexes [42]. In the model, folded protein domains are treated as rigid bodies and represented by coarse-grained residue-based beads centered at the corresponding Ca atoms. The protein-protein interactions are described by a Lennard-Jones-like pair-wise potential and a Debye–Hückel-type electrostatic potential. Notably, the Lennard-Jones-like pair-wise potential is adopted from MJ statistical potential and parameterized by fitting against experimental data. Ravikumar et al. incorporated a similar strategy to study protein-protein interactions but adopted Gō model instead of rigid bodies on individual protein components in order to explore the induced-fit mechanism of protein-protein association [62]. They hybridized the inter-protein potential (hydrophobic and electrostatic terms) with the structure-based potential on intra-protein structure. Their CG MD simulations produced a conformational ensemble of the protein complexes, followed by projecting the conformations on a potential energy landscape.

1.4 Ensemble construction with experimental data

Computational approaches assist the structural characterization of flexible biomolecules whose conformational ensemble is difficult to be determined by experiments solely. Molecular dynamics simulations are extensively used with experimental observables to simulate the structures of multi-protein/domain complexes and intrinsically disordered proteins (IDP). There exist two categories of optimization methods in the research field of structure ensemble modeling. One category is to represent the structure ensemble with an MD-simulated ensemble. For instance, Terakawa et al. performed replica-exchange MD simulations with a coarse-grained model on a highly flexible protein and generated a structure ensemble that was subsequently found to agree with experimental SAXS intensities [63]. However, this scenario rarely happens between simulation and experiment on large biomolecules for some reasons [10, 11, 15]. Besides, experimental data can be incorporated as a constraint in MD simulations [64]. The bias from experiments drives the simulated structures to be more accordant with the experimental measurements. The strategy that data drives simulation is efficient for the cases when there is a single dominant conformation in the structure ensemble, but might fail to guide the simulations if multiple competing conformations exist there.

The other category of ensemble construction approaches follows a sampling-and-selection optimization scheme. As shown in Figure 1.1, a pool of candidate structures

is generated, followed by structure ensemble selection by fitting to experimental data. The pool structures can be produced from either physics-based methods such as molecular dynamics simulations [65, 66], or topology-based methods [67, 68] that divide conformational space without molecular interactions. Kozakov et al. improved their topology-based candidate pool for ensemble construction by prioritizing the structures on the basis of a molecular docking score [68]. A raw random structure pool consists of unphysical conformations that would disrupt the subsequent ensemble construction. In comparison, molecular simulations produce the structures with molecular interactions and preclude the unphysical conformations from the pool.

Figure 1.1 A schematic illustration of the pool-based methods for ensemble construction with experimental data. Candidate structures generated by either topology-based or physics-based methods constitute a pool that would be subsequently subjected to experimental data fitting to construct the structure ensemble.

Most works adopted the sampling-and-selection strategy to recover the structure ensemble from experimental observables. However, ensemble construction encounters with a severe problem of data over-fitting. It is commonly found that multiple solutions of different ensembles yield an equal or similar quality of fit with respect to experimental data, which is also known as an ill-posed problem. Solving this problem requires additional constraints or assumptions to push the solution

towards some desired properties, and the process is called regularization. One type of regularization is ensemble size, that is, the number of structures in the ensemble. Usually, a small-size ensemble is preferred, because the representation as a large ensemble would cause a more confusing interpretation of the data [69]. This type of methods includes Minimal Ensemble Search (MES) [70], Sparse Ensemble Selection (SES) [67], Sample and Select (SAS) [71] and other works [65, 72]. In addition, Ensemble Optimization Method (EOM) [73] optimizes the structure ensemble with SAXS intensities, and it predefines the ensemble size as 20 structures by default. Despite EOM does not pursue a small ensemble, it still incorporates a reasonable ensemble size in the optimization. Apart from ensemble size, another type of regularization method is maximum entropy [74]. The methods ENSEMBLE [75] and EROS [66] take maximum entropy in their ensemble optimization to prevent the constructed ensemble from spurious biases. Popular computational approaches for ensemble recovery from experimental data are listed in Table 1.1.

Table 1.1 Popular computational approaches to construct structure ensembles with experimental data.

Methods	Optimization algorithm	Regularization
MES	Genetic algorithm	Size
EOM	Genetic algorithm	Size
SAS	Simulated annealing	Size
SES	Multi-orthogonal matching pursuit	Size
ENSEMBLE	Simulated annealing	Maximum entropy
EROS	Simulated annealing	Maximum entropy

1.5 Aims and outlines

Many protein complexes consist of multiple domains whose communications are fundamental for their biological functions. The structure determination of multi-domain proteins is essential for understanding the mechanism of molecular interactions and functions, but it remains a challenge to derive structure ensemble of flexible biomolecules from experiments solely. The experimental measurements are

from time- and spatial-average, so there might be multiple conformations responsible to the averaged experimental observables. In the field of ensemble construction, most works focus on selecting structures from a candidate structure pool and fitting to experimental data numerically. We proposed structure energy as a physics-based regularization in the ensemble optimization to minimize the data over-fitting problem. This ensemble modeling strategy can be implemented on both all-atom and coarse-grained levels. We demonstrated the importance of the physics-based regularization on experimental data interpretation to avoid ambiguous and spurious conformational representations. Besides, other factors in structure ensemble construction have also been examined, for example different sources of pool structures, different optimization algorithms, structure models of multi-domain proteins, and so on. The research work in this thesis is organized as follows.

In Chapter 2, the dengue virus 2 (DENV2) non-structural protein 5 (NS5) was modeled with its experimental SAXS intensities. DENV2 NS5 has a methyltransferase (MTase) domain and a RNA-dependent-RNA polymerase (RdRp) domain connected by a ten-residue linker. We adopted a hybrid strategy to model NS5, with a structure-based intra-domain potential and a physics-based inter-domain potential. The NS5 is modeled on the residue-based coarse-grained level, because NS5 is a large protein of 900 amino acids, whose atomistic simulations are too expensive for the heavy computational burden. Besides, the SAXS intensities provide low and medium structural information, so the structure models on residue

level suffice the given experimental data. The structure ensemble of NS5 was determined by fitting to SAXS data with the low-energy structures from CG MD simulations. The simulated structures were investigated in an energy perspective on their domain-domain orientations. The physics-based regularization has been demonstrated as a necessary constraint to avoid a spurious interpretation of experimental data.

In Chapter 3, we simulated the structure ensemble of a multi-domain protein with its published experimental PRE data. The protein is the first two RNA recognition motif domains of poly(U)-binding protein, referred as PubRRM12, from *Saccharomyces cerevisiae*. We adopted a structure-based model to simulate intra-domain structure and an all-atom force field to study inter-domain interactions. This hybrid strategy is similar as the NS5 case. Atomistic MD simulations were employed to generate a structure pool of PubRRM12, followed by structure energy analysis and ensemble construction with experimental PRE rates that correspond to atomistic pairwise distance constraints. Atomistic structure model was applied to PubRRM12 instead of a coarse-grained one, considering the high sensitivity of PRE rates proportional to $\langle r^{-6} \rangle$ where r is inter-atomic distance. We used an energy landscape view to characterize the conformational space. Energy as the physics-based constraint was incorporated with the ensemble-size constraint together to determine the structure

ensemble. The crucial effect of the two regularization terms has been clearly demonstrated.

Chapter 4 is the discussion. The final part Chapter 5 is the concluding remarks.

2 Coarse-grained molecular modeling of solution structure ensemble of dengue virus NS5 with SAXS intensity

2.1 Background

Dengue virus is a continuing threat globally and exposes approximately 2.5 billion people to the danger of virus infection [76]. The virus has four serotypes DENV1, DENV2, DENV3 and DENV4 that share approximately 65-70% similarity of their sequences [77]. DENV virus replication process takes place in replication complex, and DENV non-structural proteins are important components in that microenvironment [78-80]. Among the non-structural proteins, NS5 possesses RNA synthesis and capping activities [81-83], whose crucial role in virus replication suggests its potential target for inhibitor development and drug discovery [84]. NS5 consists of an N-terminal methyltransferase (MTase) domain and a C-terminal RNA-dependent RNA polymerase (RdRp) domain whose commutations mediate its functions in virus replication, therefore a comprehensive characterization of MTase-RdRp interactions is important.

The structure of full-length NS5 has been solved crystallographically in two independent works [85, 86]. Interestingly, the NS5 structure from Japanese encephalitis virus has hydrophobic-driven MTase-RdRp contacting interface [85], but the NS5 crystal structure from DENV3 is a different conformation where MTase

domain contacts RdRP majorly via polar interactions [86]. It is evidenced that NS5 exists in solution with multiple conformations [87, 88], which indicates the high structural flexibility of MTase-RdRP associations. Actually, this conformational flexibility is reasonable in terms of its biological function, because NS5 has the dynamical interplay with virus RNA and other proteins in the virus replication process. Neither of the two solved crystal structures above is able to interpret the experimental SAXS data from the solution structure of the full-length NS5. A earlier small-angle X-ray scattering study suggested a structure ensemble of DENV3 NS5 that consists of the conformations from compact to extended shapes [87]. The representative ensemble from another SAXS work on NS5 proposed a structure ensemble where the extended conformations were absent [88]. Compared with the crystallographic structures, solution structures of NS5 from DENV1 to 4 adopted a more lengthened conformation. Importantly, flexible segments such as inter-domain linker and mobile loops are usually not well-structured and their flexibility mediates the communications between different well-structured units. The linker in DENV NS5 has been demonstrated to affect protein conformation in the linker-swapping experiments [88, 89].

It is an increasingly popular strategy to combine computational approaches with SAXS measurements for the structure characterization of multi-domain proteins on the coarse-grained level [63, 66, 90]. The coarse-grained structure modeling speeds up the MD simulations significantly and its resolution suffices the structural

information from experimental SAXS profiles. In this work, we carried out a CG modeling to characterize the structure ensemble of DENV2 NS5 with experimental SAXS intensities from a published work [88]. MTase and RdRp domains were simulated with a structure-based potential [35, 91, 92], and their inter-domain communications were modeled with hydrophobic and electrostatic energies [62]. The structures of DENV2 NS5 were produced by enhanced-sampling CG MD simulations. A candidate pool was obtained after structure clustering and further investigated in an energy view. Ensemble of NS5 was recovered from experimental SAXS profiles by data fitting with the regularization of structure energy. We successfully proved that physics-based constraint of structure energy is important to prevent spurious conformational representations from a raw numerical fitting scheme. The random pool structures might expose the SAXS-derived ensemble construction to the danger of over-fitting without the biophysical constraint.

2.2 Methods

2.2.1 Coarse-grained model of NS5

As shown in Figure 2.1, each amino acid in DENV2 NS5 was coarse-grained as one dummy bead located at its C α position. The hybrid model consists of intra-domain and inter-domain potential energies, termed as V_{intra} and V_{inter} respectively.

Figure 2.1 Atomistic structure and coarse-grained representation of the multi-domain protein DENV2 NS5. Each amino acid is coarse-grained as one bead in the structure model. RdRp domain (green) and MTase domain (cyan) are connected via a ten-residue flexible linker (orange).

V_{intra} is a structure-based potential [35, 91, 92], formulated in Equation 2.1, which includes bond, angle, dihedral, native and non-native interactions. Structure-based models use the native structure as the global energy minimum. Therefore, in Equation 2.1, the parameters of the equilibrium positions b_0 , θ_0 , φ_0 and r_{0ij} were derived from the native structures of MTase and RdRp domains. Importantly, the native interaction term stabilizes the domain conformation via a Lennard-Jones-like function where the native pairs of residues are biased to the pairwise distances in the native structure (r_{0ij}). Native pairs refer to the pairs of contacting residues that are spatially adjacent. A method of shadow map [93] was employed to generate the

native pairs for each domain, and the non-native pairs were modeled with an exclusion potential to avoid steric clash with the parameter $C = 4 \text{ \AA}$. The interaction strengths of the intra-domain potential terms are $k_b = 100 \text{ kcal}/(\text{mol} \cdot \text{\AA}^2)$, $k_\theta = 20 \text{ kcal}/(\text{mol} \cdot \text{rad}^2)$, $k_\varphi^1 = 1 \text{ kcal/mol}$, $k_\varphi^3 = 0.5 \text{ kcal/mol}$, $\varepsilon_1 = 4 \text{ kcal/mol}$ and $\varepsilon_2 = 1 \text{ kcal/mol}$ respectively.

$$\begin{aligned}
V_{\text{intra}} = & \sum_{\text{bond}} k_b (b - b_0)^2 \\
& + \sum_{\text{angle}} k_\theta (\theta - \theta_0)^2 \\
& + \sum_{\text{dihedral}} k_\varphi^{n=1,3} [1 + \cos(n \times (\varphi - \varphi_0))] \\
& + \sum_{\text{native}}^{i>j-3} \varepsilon_1 \left[5 \left(\frac{r_{0ij}}{r_{ij}} \right)^{12} - 6 \left(\frac{r_{0ij}}{r_{ij}} \right)^{10} \right] \\
& + \sum_{\text{non-native}}^{i>j-3} \varepsilon_2 \left(\frac{C}{r_{ij}} \right)^{12}
\end{aligned} \tag{2.1}$$

The native structure of MTase domain is a crystal structure (PDB entry 1L9K) of NS5 MTase domain from DENV2 [82]. The native structure of NS5 RdRp domain is constructed using homology modeling [94] with a template of the crystal structure (PDB entry 4V0R) of NS5 RdRp domain from DENV3 [86]. Some flexible segments are missing in the experimental crystal structures, including two loops in RdRp domain (residue 407-418 and 456-469) and two terminal loops. The linker (residues 264-273) and these flexible loops without native structures are simulated as chains of harmonic springs whose strength is a constant value of $100 \text{ kcal}/(\text{mol} \cdot \text{\AA}^2)$.

The inter-domain potential V_{inter} is an established coarse-grained potential [62, 95] that includes electrostatic and hydrophobic interactions referred as V_{ele} and V_H respectively, given in Equation 2.2.

$$V_{inter} = V_{ele} + V_H \quad (2.2)$$

The electrostatics V_{ele} adopts a Debye-Hückel-type potential shown in Equation 2.3. The charged residues i and j are interacted on the basis of their distance r_{ij} . To reflect the screening effect from salt in experiment, the parameter Debye length ξ is set to 5.6 Å corresponding to the buffer condition of 300 mM NaCl. At the PH condition of 7.5, the residue charges q are -e for the residue types of Glu and Asp and +e for the types of Arg and Lys. D in Equation 2.3 is a dielectric constant to describe the inter-domain environment and set to 10 in this coarse-grained model [62].

$$V_{ele} = \sum_{ij} q_i q_j \exp(-r_{ij}/\xi) / 4\pi r_{ij} D \quad (2.3)$$

The hydrophobic term V_H is defined in Equation 2.4. The residue pairs are categorized into attractive type of pairs ($\varepsilon_{ij} < 0$) and repulsive type ($\varepsilon_{ij} \geq 0$). If two residues attract each other, their interaction is formulated as a Lennard-Jones-like potential, otherwise a repulsive potential with the parameter $d = 3.8$ Å is assigned to the repulsive pairs. The residue-pair strengths ε_{ij} and equilibrium distances σ_{ij} have been published [95]. Notably, the solvation effect is implicitly incorporated in the

residue-pair strengths ε_{ij} that were developed on the basis of Miyazawa-Jernigan statistical potential [60]. Besides, the inter-domain potential V_{inter} only considers the surface residues that are exposed to domain-domain contact. A residue is determined as the exposed one if its solvent accessible surface area is over the threshold value of $10 \text{ \AA}^2$.

$$\begin{aligned} V_H(i, j) &= |\varepsilon_{ij}| \left[5(\sigma_{ij}/r_{ij})^{12} - 6(\sigma_{ij}/r_{ij})^{10} \right], (\varepsilon_{ij} < 0) \\ V_H(i, j) &= \varepsilon_{ij} \left[5(\sigma_{ij}/r_{ij})^{12} \left(1 - \exp\left(-((r_{ij} - \sigma_{ij})/d)^2\right) \right) \right], (\varepsilon_{ij} \geq 0) \end{aligned} \quad (2.4)$$

2.2.2 MD simulations

MD simulations are used to generate candidate structures for ensemble construction of DENV2 NS5. MD simulations take the advantage of precluding unphysical conformations but encounter with the problem of insufficient sampling. In this study, we performed multiple sets of independent simulations starting from different initial structures for a complete search of the conformation space. Firstly, a total of 10000 random structures of DENV2 NS5 were produced by the software RANCH [73], and a clustering scheme [96] was subsequently performed on them with a RMSD cutoff at $9 \text{ \AA}$. Finally, 102 structure clusters were grouped from the 10000 random structures, and the middle structure of each cluster was taken as the initial structure

in a MD simulation using the hybrid CG model we described above. A total of 102 MD simulations were run independently using the package GROMACS 4.6 [97]. The simulations were performed with a time step of 0.01 ps at the temperature of 300 K for 100 ns each. After MD simulations, all the simulated structures were pooled together, summing up to 204000 structures.

2.2.3 Pool structure characterization

The huge amounts of structures from MD simulations were clustered on the basis of their structure topology in two steps. The first clustering step sorted all the simulated structures into domain-domain orientation sub-groups. As shown in Figure 2.2, φ and θ angles depict domain-domain orientation in the spherical coordinate system. The entire θ - φ orientation space was uniformly divided into sub-groups sized as $\Delta\varphi = 18^\circ$ and $\Delta\theta = 9^\circ$. The θ and φ angles of each structure were calculated, followed by assigning the structure into its corresponding orientation sub-group. In the second step, a RMSD-based clustering [96] was performed on the structures within the same orientation sub-group, and a cutoff of 4 Å was used. A total of 1015 clusters have been identified after the two-step structure clustering. We took the structure of the most favorable inter-domain energy V_{inter} (Equation 2.2) within each cluster as the representative. The resultant 1015 representative structures are the candidate pool for ensemble construction with SAXS data.

Figure 2.2 Domain-domain orientation of DENV2 NS5. The orientation is defined by two spherical angles ϕ and θ . MTase domain (purple) and RdRp domain (green) are connected via a flexible linker (orange).

2.2.4 Ensemble construction with SAXS data

To fit an ensemble model to experimental SAXS profile, the theoretical scattering intensity was computed for each of the 1015 candidate structures with the software CRY SOL [98]. As CRY SOL requires the input structures at the atomistic level, the representative CG structures were firstly reconstructed back to all-atom structures with PULCHRA [99] prior to their SAXS profile computation. To construct the ensemble, the structures were selected from the pool for a best fit of experimental

data. In our study, we recovered the structure ensemble from experimental SAXS intensities using non-negativity least squares (NNLS) method [100] and a constraint was added that the total weight of the selected structures is 100% in the optimization. Our work is featured for the physics-based regularization in ensemble construction, that is, the low-energy structures were taken to construct ensemble to prevent the fitting noise from high-energy conformations. Here, the energy refers to the CG inter-domain energy V_{inter} (Equation 2.2). Apart from NNLS optimization, we also adopted an alternative approach EOM 2.0 [101] that used a GA algorithm to construct the ensemble with the simulated pool of 1015 representative structures. NNLS and EOM are two independent data-fitting approaches employed in this work for a comparison for the ensemble construction between different optimization methods. NNLS does not restrict the ensemble size, whose solution directs an optimal numerical fitting for a given pool. By contrast, EOM incorporates ensemble size (20 by default) in the optimization, though it does not directly pursue a small ensemble. In the following result section, the optimization results from NNLS and EOM are found being consistent with the energy regularization.

The consistency between experimental SAXS data and the calculated SAXS profile of the simulated ensemble is quantified by χ^2 in Equation 2.5. $I(s)$ is the calculated SAXS intensities of the constructed ensemble, $I_{exp}(s)$ is the SAXS intensities from experiments, and $\sigma(s)$ is the standard deviation of experimental data. The number of

SAXS data points K is 813 ranging up to $s = 0.35 \text{ \AA}^{-1}$ in our case, and μ is a scaling factor. The details of the scattering experiments are available in the literature [88].

$$\chi^2 = \frac{1}{K-1} \sum_{j=1}^K \left[\frac{\mu I(s_j) - I_{\text{exp}}(s_j)}{\sigma(s_j)} \right]^2 \quad (2.5)$$

2.3 Coarse-grained MD simulations

A hybrid model was applied to DENV2 NS5 at the residue-based coarse-grained resolution. The well-structured MTase and RdRp domains were subjected to a structure-based model and their inter-domain associations were simulated with a coarse-grained potential of electrostatic and hydrophobic energies. This idea of the hybrid CG model has been successfully applied to such biomolecules that have several well-structured domains linked via flexible polypeptide segments. However in most cases, the flexible segments were simplified as a chain of spring without physical interactions [95] or simulated to contact to other structure units only with electrostatics [63]. Our hybrid model is featured that the disordered segments of the linker and flexible loops contact to domains and within themselves with a more rigorous physical potential of hydrophobic and electrostatic interactions. Considering the significant role of the linker in mediating domain-domain

communications in NS5 [88, 89], an accurate potential is necessary to model these flexible segments.

2.4 Structure pool and energy landscape

MD simulations produced a huge number of structures of NS5. A structure analysis was performed on individual MTase and RdRp domains respectively. The RMSD values between the simulated structures and the native structures are smaller than 1 Å, demonstrating the structure-based potential stabilized the intra-domain conformation as expected. MD simulations sampled a great variety of protein conformations that were clustered into a pool of 1015 structures (details in Chapter 2.2.3). For multi-domain protein complexes, the protein flexibility arises from the high domain-domain mobility, thus the domain-domain orientation (Figure 2.2) is an important dimension of conformational space. As the φ and θ angles are defined as MTase rolling around RdRp in Figure 2.2, each φ - θ pair also targets one domain-domain contacting interface on RdRp domain surface. As shown in Figure 2.3a, the 1015 representative structures were projected on the φ - θ orientation map to depict the protein conformational space. Because of the limited length of the inter-domain linker of ten residues, MTase interacts with RdRp within a certain range of domain-domain orientations, which leaves a large area of orientation as empty surface of physically inaccessible region as shown in Figure 2.3a.

Figure 2.3 The domain-domain orientations of DENV2 NS5 structures from the simulated pool and the energy landscape. (a) The 1015 pool structures (red points) of DENV2 NS5 from MD simulations are projected to domain-domain orientation map on the basis of their φ and θ angles. (b) It shows the energy landscape of DENV2 NS5 based on the inter-domain energies of simulated pool structures. The landscape is gridded with a $12^\circ \times 6^\circ$ grid size and energy value is defined in the color bar (kcal/mol).

To draw an energy landscape of the domain-domain orientation from the pool structures, the orientation space was divided into small grids as conformational sub-spaces. A grid size of $12^\circ \times 6^\circ$ was adopted, and each grid contained multiple structures in the accessible orientation area. The structure of the lowest inter-domain energy V_{inter} (Equation 2.2) within a grid was chosen to represent the energy of the

grid, thus the conformational space was represented as the sub-spaces of different energy values, referred as energy landscape. As revealed in Figure 2.3b, the energy landscape of DENV2 NS5 shows one broad energy minimum consisting of contiguous low-energy grids. It is found that the energy difference within the minimum is subtle in contrast with the dramatic energy difference from the outside-minimum space. The edge between low-energy and higher-energy orientation surfaces is noticeable with large energy gap values. It is noted that the inaccessible orientation space restricted by the limited linker length is also depicted as high-energy surface in the energy landscape in Figure 2.3. Importantly, energy landscape instructs energy-favorable conformational space of NS5 that assists the ensemble construction from an energy perspective. The bad-energy conformation should be negligible in the constructed ensemble of NS5 according to the Boltzmann distribution. If an ensemble successfully recovers the experimental SAXS data but incorporates the structures from high-energy conformational space, the recovered conformational ensemble might be argued as an over-fitting result.

2.5 Ensemble construction with SAXS data

The candidate pool contains 1015 representative structures that were projected on domain-domain orientation landscape in an energy perspective in Figure 2.3. Each structure was subjected to its SAXS profile computation by using the program

CRY SOL [98] after mapped to atomistic structure with the program PULCHRA [99]. The reconstruction from CG to atomistic structure preserves the protein backbone fold and SAXS profiles are insensitive to the local conformations of side chains due to the low-resolution structural information from SAXS data. The structures were selected from the pool for a best fit to experimental SAXS data by using NNLS method [100]. In this optimization scheme, we employed structure energy as a regularization term to regularize the ill-posed problem. Specifically, the sub-pools at different energy levels were extracted from the entire 1015-structure pool. All the pool structures were ranked on the basis of their inter-domain energy and the top 1% low-energy structures were taken to constitute a top 1% low-energy pool. The ensemble resulted from fitting experimental SAXS data with the top 1% low-energy pool was called top 1% low-energy ensemble. Similarly, the ensembles at different energy thresholds were constructed, including the 100% ensemble without any energy constraint as well as the regularized top 5%, 10%, 15%, 20%, 25%, 30%, 40% and 50% low-energy ensembles.

Table 2.1 shows the constructed ensembles at different energy levels and their fitting qualities. The top 1% low-energy pool consists of the top ten low-energy structures. The data fitting resulted in the top 1% low-energy ensemble of two conformations with a χ^2 of 0.58. The fitting quality was improved to a χ^2 of 0.39 for the top 5% low-energy ensemble. Interestingly, from 5% onwards, an increasing size of candidate pool with a loose energy constraint failed to significantly improve the fitting quality.

The top 10% ~ 50% low-energy ensembles produced χ^2 values ranging from 0.35 to 0.39, and the 100% ensemble without any energy constraint can only improve the fit to 0.34. We found that the low-energy ensembles recovered by NNLS required a small number (2 ~ 5) of structures to reproduce the experimental SAXS data, as shown in Table 2.1. Importantly, an optimal structure ensemble should not only agree well with experimental observables, but also consist of physically sound conformations to prevent spurious conformational representations. High-energy conformations are of little probability to exist in structure ensemble according to the Boltzmann distribution, and later in next sections we prove that the existence of bad-energy conformation in the constructed ensemble turns out as noise of data over-fitting. In our case, the low-energy ensembles from 5% onwards show an almost equal quality of agreement with experimental data. Figure 2.4 clearly reveals the “L-shape-like” evolution of χ^2 with a loose energy constraint on the candidate pool, with a turning point at 5%. Therefore, the top 5% low-energy ensemble is recommended as an optimal representative, whose consistency with experimental SAXS profile is very high (Figure 2.4, inset). The top 5% low-energy ensemble is further supported by the experimental R_g approximated from Guinier. The R_g value averaged over the 5% ensemble is 3.49 nm that highly accords with 3.56 ± 0.05 nm from experiment. It is noted that reproducing an exactly same value as experimental R_g is challenging, because the flexible terminal loops might adopt many conformations in solution that are not able to reproduced by only two structures in our top 5% low-energy ensemble.

Table 2.1 Structure ensemble of DENV2 NS5 recovered from experimental SAXS data with energy constraints. The four columns in the table refer to the energy-level of the pool/ensemble, the number of candidate pool structures, the quality of fit χ^2 , and the size of the recovered ensemble, respectively.

ensemble	candidates	χ^2	conformations
1%	10	0.58	2
5%	51	0.39	2
10%	102	0.39	4
15%	152	0.37	3
20%	203	0.37	4
25%	254	0.37	5
30%	305	0.37	5
40%	406	0.35	4
50%	508	0.35	4
100%	1015	0.34	8

Figure 2.4 The evolution of χ^2 with a loose energy constraint on the candidate pool. The χ^2 values are illustrated for the top 1%, 5%, 10%, 15%, 20%, 25%, 30%, 40%, 50% low-energy and the 100% ensembles. The inset figure is a comparison between the SAXS profile calculated from the top 5% low-energy ensemble and that from SAXS experiments.

The optimal ensemble and other low-energy ensembles were further subjected to the conformational characterizations. The constructed ensembles were projected on the energy landscape of domain-domain orientation. Noticeably, despite of the different ensemble sizes, energy levels and structure components, the low-energy ensembles (top 5%, 10%, 15%, 20%, 25% and 30%) reproduced the experimental data well with a conserved domain-domain orientation space. As shown in Figure 2.5a, the 5%, 15% and 30% low-energy ensembles have their component structures of similar

domain-domain orientations in the energy landscape. The optimal ensemble recovered from top 5% low-energy pool has two component structures that are located in the low-energy surface in the energy landscape of domain-domain orientation. In contrast, the 100% ensemble lacking of the energy constraint has eight component structures whose orientation distribution shows a much more dispersed and arbitrary pattern. The domain-domain orientation structurally corresponds the domain-contacting interface on RdRp domain (Figure 2.2). The domain-contacting residues are identified if an RdRp residue contacts to an MTase residue with their $\text{C}\alpha$ beads separated by a distance less than 10 Å. The domain-contacting interfaces on RdRp domain are shown in Figure 2.5b, and the low-energy ensembles share a highly consistent pattern. In the low-energy ensembles, RdRp domain binds to MTase domain with one loop segment (residues 362-368) and another loop region with partial β -strand (residues 294-307). In contrast again, the 100% ensemble shows a much more disperse and noisy contacting interfaces on RdRp domain. The consistency of the conformational space among the low-energy ensembles supports the optimal (5% low-energy) ensemble we recommended which agrees well with the experimental observables and accords well with the energy profiles.

Figure 2.5 Domain-domain orientations on the energy landscape and domain contacting segments on the RdRp domain of the 5%, 15%, 30% low-energy and the 100% ensembles. (a) The conformations (blue triangle dots) in the constructed ensembles are projected on the energy landscape of domain-domain orientation. The energy values can be referred to Figure 2.3b. (b) RdRp domain of DENV2 NS5 is shown as ribbon, underlining its domain-contacting residues (red).

We performed the structural characterization of the top 5% low-energy ensemble. The two component structures have very close orientations ($\phi \approx -3^\circ$, $\theta \approx 65^\circ$ & $\phi \approx -1^\circ$, $\theta \approx 60^\circ$) within the low-energy surface, but they adopted various MTase-rotations. The structural detail is shown in Figure 2.6 where RdRp domain contacts to MTase domain with the same interface in the two structures. The linker lies between RdRp and MTase domains and mediates their associations. The linker in DENV NS5 has been proved to affect protein conformation by the linker-swapping experiments [88].

A sequence analysis of NS5 revealed that the ten-residue linker is basically not conserved among the four DENV serotypes, however except two residues. These two residues are the charged E268 and E270 in the linker, conserved from DENV1 to DENV4, which suggests they might preserve some mediating function in MTase-RdRp associations. In the two structures of the optimal ensemble, the conserved electrostatic contacts are found between the linker residues E268/E270 and the RdRp residues R595/R596, especially the specific charge-charge interaction between the linker E270 and the RdRp R596 (Figure 2.6). This electrostatics signature has also been identified in DENV3 NS5 whose linker residue E270 forms a close contact with the RdRp residue K596 in the crystallographic structure. K596 in DENV3 NS5 is a conserved substitute of the basic property of R596 in DENV2.

Figure 2.6 The structural characterization of the top 5% low-energy ensemble. MTase, RdRp and linker are colored in purple, green and orange, with the domain-contacting residues on the RdRp domain highlighted in red. The zoom-in figures illustrate the conserved electrostatic interactions between the linker residues E268/E270 (blue) and the RdRp residues R595/R596 (gray).

The ensemble constructed without the energy regularization appears an arbitrary conformational space, in contrast with its low-energy counterparts of consistent domain-domain orientations and domain-contacting interface (Figure 2.5). The 100% ensemble fails to pass the energy landscape. Most of the component structures deviate away from the low-energy surface, and some have bad-energy orientations.

A close energetic investigation on the ensemble structures was shown in Figure 2.7a. Three structures located around $\varphi \approx -100^\circ$ and $\theta \approx 85^\circ$ turn out high-energy conformations whose energy rankings are worse than 90% in the pool. The energy decomposition shows that these three structures are repulsed by both electrostatic and hydrophobic interactions. In the case of the structure at $\varphi \approx 27^\circ$ and $\theta \approx 73^\circ$, its MTase-RdRp associations are rejected by the unfavorable electrostatic energy. These conformations of bad energies interpret the protein conformation in a noisy pattern and severely undermine the reliability of the constructed ensemble in physics. The ensemble construction without the energy constraint can produce an average quantity numerically consistent with experiments, but the data might be over-interpreted to represent a wrong conformational space of protein.

Figure 2.7 The structure ensembles constructed by NNLS and EOM without the energy constraint by employing the same simulated pool. (a) It shows the 100% NNLS-based ensemble consisting of some high-energy conformations. ΔE is the difference of the structure energy with respect to the average energy of all the 1015 simulated pool structures (kcal/mol). ΔE is decomposed into its two components of hydrophobic energy ΔE_H and electrostatic energy ΔE_{ele} . The two structures of large θ angles are circled with the dashed boxes in the energy landscape. Protein structure is colored as that in Figure 2.2. (b) It shows the conformations and their energetic profiles of the EOM-based ensemble without the energy constraint.

2.6 Incorporation with Ensemble Optimization Method

Apart from the NNLS approach employed in ensemble optimization, the GA-based approach EOM [73, 101] was also adopted to construct ensemble of DENV2 NS5 with the same experimental SAXS data and the same simulated 1015-structure pool. We recovered the conformational ensemble from experimental SAXS data with

EOM to compare the ensembles constructed from different optimization algorithms and investigate the influence from the energy regularization. Firstly, EOM fitted the experimental SAXS data without the energy constraint, producing an ensemble of five structures. As shown in Figure 2.7b, these five structures do not agree with the energy landscape, and some structures are located in the high-energy orientation space. For example, the conformation at $\varphi \approx -105^\circ$ and $\theta \approx 79^\circ$ in the landscape is of high energy ranked worse than 99%, where the two domains strongly reject against each other by both unfavorable hydrophobic and electrostatic energies.

A comparison between the ensembles recovered by NNLS and EOM without the energy constraint is shown in Figure 2.7. The two methods achieve a basically same quality of fit with $\chi^2 = 0.34$ and 0.35 respectively. The EOM-based ensemble without the energy constraint (called EOM-based 100% ensemble) was identified as half conserved as the 100% NNLS-based ensemble, and they share three common structures ($\varphi \approx -17^\circ$, $\theta \approx 73^\circ$; $\varphi \approx -105^\circ$, $\theta \approx 79^\circ$ and $\varphi \approx -92^\circ$, $\theta \approx 36^\circ$). However, EOM-based 100% ensemble failed to recover a consistent protein conformational space as its NNLS-based counterpart. EOM-based 100% ensemble does not include the structures of large θ angle ($\theta > 100^\circ$) as the conformational representations. But the 100% NNLS-based ensemble has two large- θ -angle structures that constitute an important component of the conformational ensemble with a total proportion of 29% (Figure 2.7a). The ill-posed problem has been clearly revealed here that a good fit

with target biological data by the two methods results in inconsistent protein conformations.

The EOM optimization is further regularized with the structure energy. We used the same energy regularization scheme here as the NNLS case above. The structure ensembles were constructed with the experimental SAXS data by using the simulated pools at different energy thresholds. Interestingly, when the energy constraint was performed, EOM constructed the highly consistent ensembles of NS5 as NNLS. The EOM-based 5% low-energy ensemble (Figure 2.8) is found exactly identical as the NNLS-based 5% low-energy ensemble (Figure 2.5a and Figure 2.6). It can be also concluded that the proposed optimal ensemble that is recovered from NNLS with the 5% low-energy constraint is successfully reproduced by the energy-regularized EOM optimization. The consistency between EOM and NNLS has also been found in other low-energy ensembles. The EOM-based 15% low-energy ensemble consists of two structures that both exist in the ensemble recovered by NNLS method with the 15% low-energy constraint. The EOM-based ensemble from the 30% low-energy pool has a total of four component structures and all of them are found in the NNLS-based 30% low-energy ensemble. A comparison between the EOM-based ensembles (Figure 2.8) and its NNLS counterparts (Figure 2.5a) with the energy regularizations demonstrates the highly consistent interpretation of protein conformations. In contrast with the cases without the energy constraint, the energy regularizations prevent the ensemble construction from arbitrary representations of conformations

caused by the noise from bad-energy conformations. In summary, the energy regularization of the 5% low-energy threshold drives NNLS and EOM to construct the same ensemble that agrees well with experimental quantities and the energy landscape. This optimal ensemble is also supported by other NNLS and EOM low-energy ensembles in terms of the conformational space.

Figure 2.8 The EOM-based ensembles from the simulated pool with and without the energy constraint. The 5%, 15%, 30% low-energy and the 100% ensembles (blue triangle dots) recovered from experimental data with GA-based EOM approach are projected on the domain-domain orientation landscape. The energy values can be referred to Figure 2.3b.

The “true” ensemble of flexible biomolecules contains many structures that are distributed in the energy basins. People usually use a limited number of representatives to represent the biomolecular conformations from the averaged experimental quantities. The popular SAXS-data modeling software EOM was also

used for DENV2 NS5 in its standard fashion, where its pool of 10000 random structures was employed instead of our simulated pool. EOM produced 10000 structures with the rigid MTsae and RdRp structures and optimized the structure ensemble with genetic algorithm. The standard EOM approach constructed an ensemble of four structures with a quality of fit $\chi^2 = 0.36$. We compared the standard EOM result and the simulated optimal ensemble in Figure 2.9. The two most populated structures in the four-structure EOM ensemble show an adjacent domain-domain orientation with the simulated optimal ensemble we proposed with the 5% low-energy constraint (Figure 2.9a). A structure examination reveals these two most populated EOM structures bind their MTase domain to a close surface on RdRp domain compared with the two structures of the simulated 5% low-energy ensemble (Figure 2.9b). However, as shown in Figure 2.9, the other two structures in the standard EOM ensemble significantly deviate away from the optimal domain-domain orientation space with a totally different MTase-RdRp contacting pattern. The two minor structures are located in the bad-energy surface of the domain-domain orientation landscape. Interestingly, if attempting ensemble construction with EOM for multiple times, the two minor structures are hard to be reproduced. Thus, the standard EOM constructed structure ensemble of NS5 whose major components cover the conformational space suggested by our energy-regularized ensembles. But the other two states in the standard EOM result appear highly suspicious because of their bad-energy orientation and poor reproducibility. The bad-energy noise from the

pool affects the conformational representation of flexible biomolecular systems, and the physics-based regularization is able to preclude questionable conformations from ensemble fitting.

It is clarified that the previously published paper [88] used EOM to model DENV2 NS5 ensemble with MTase and RdRp structures from DENV3 NS5 because of the high sequence conservation among the DENV serotypes. In our work on DENV2 NS5, the MTase structure came from DENV2 NS5, and RdRp structure was a is homology model that was constructed with the DENV3 NS5 crystal structure as a template. Our physics-based modeling scheme requires the correct sequence of DENV2 NS5 to simulate the residue-specific molecular interactions. We are able to reproduce an EOM ensemble consistent with the one from the publication above if inputting the same MTase and RdRp structures from NS3 (data not shown).

Figure 2.9 A comparison between the standard EOM ensemble and the simulated optimal ensemble of DENV2 NS5. (a) The standard EOM scheme with its 10000-structure pool produced an ensemble of four structures (deep green triangle dots), which is depicted on the energy landscape of domain-domain orientation. The two structures of the top 5% low-energy simulated ensemble are shown (blue triangle dots) in the landscape. (b) The structure comparison is illustrated between the standard EOM ensemble and the top 5% low-energy simulated ensemble. All the structures have their RdRp domain structure aligned (translucent green envelop) to reveal the MTase-binding position on RdRp domain. The most and second most populated structures (deep pink and yellow respectively) from the standard EOM ensemble have their MTase domain (shown as ribbon) binding to a close RdRp surface as the two structures (green and cyan) in the top 5% low-energy simulated ensemble. The two less-populated EOM structures (blue and purple) have their MTase domain binding to a very different surface on RdRp domain.

2.7 Validation of low-energy conformations in all-atom simulations

Coarse-grained models have been successfully applied molecular modeling of large biomolecular systems. The enhanced efficiency of CG models is at the cost of accuracy, as the structural information at atomistic level is averaged. CG models are expected to be able to target the free energy basins of the conformational space and reflect the biophysical basis [8, 102, 103]. In our case, the low-energy ensemble has successfully reproduced the experimental SAXS profiles. Here, we validated the proposed optimal low-energy ensemble from the CG modeling scheme with the conventional atomistic MD simulations. The two CG structures of the top 5% low-energy ensemble were constructed to full-atom representations by using PULCHRA [99], followed by being inputted as starting structures in the all-atom MD simulations with the package GROMACS 4.6 [97]. The simulation was run with the force field Amber99SB [104] in a cubic box with TIP3P water model. The time step is 2 fs and the simulation temperature is 300 K. The simulation length is 20 ns, which suffices the purpose of identifying the local minimum and examining the conformational stability of the initial conformations. It is noted that the structure-based potential was added to the native pairs of heavy atoms in the individual MTase and RdRp domains respectively to stabilize their intra-domain structure.

The 5% low-energy ensemble structures are able to preserve their conformations in the conventional all-atom simulations (Figure 2.10a and b). Specifically, one low-

energy conformation ($\varphi \approx -3^\circ$, $\theta \approx 65^\circ$ in the 5% case of Figure 2.5a) stayed stable in the all-atom MD simulation without dramatic change in its backbone RMSD. Its equilibrium trajectory was subjected to a clustering scheme [96] with the RMSD cutoff at 5 Å, and only one cluster was generated. A structure comparison was performed between the initial conformation and the middle structure of the cluster. As shown in Figure 2.10a, the middle structure, also referred as the representative structure of the simulation, shares the same MTase-RdRp orientation as its initial structure, with an RMSD value of 4.9 Å. This slight structure change occurred on the minor self-rotation of MTase domain, while this low-energy structure from CG modeling scheme preserved its global conformation and the basis of molecular interactions in the atomistic simulation. In other words, this CG low-energy conformation captures one energy basin cluster. Besides, the other structure in the 5% low-energy ensemble ($\varphi \approx -1^\circ$, $\theta \approx 60^\circ$ in the 5% case of Figure 2.5a) was highly stable in its atomistic simulation. The structure clustering on its equilibrium trajectory resulted in only one cluster whose middle structure was very close to the initial conformation with a RMSD of 2.8 Å (Figure 2.10b). Therefore, the structures in the low-energy ensemble based on our CG force field preserve their conformations in the atomistic force field, which supports our CG model being capable of capturing the biophysical basis of the molecular interactions.

Without the energy constraint, NNLS and EOM recovered the structure ensembles with bad-energy conformations from data fitting, among which two structures exist

in the both NNLS and EOM 100% ensembles. These two high-energy structures were subjected to a conformational stability examination in the all-atom simulations. The simulations were set up as the same as the two low-energy cases above. The two conformations ($\varphi \approx -105^\circ$, $\theta \approx 79^\circ$ and $\varphi \approx -92^\circ$, $\theta \approx 36^\circ$ respectively in Figure 2.7) deviated against the initial structures dramatically from the start of the simulations and kept a continuously changing trend over the simulations (Figure 2.10c and d). The backbone RMSD values between the initial conformations and the last-frame structures of the 20-ns MD trajectories are 11.4 Å and 10.4 Å respectively. As shown in Figure 2.10c and d, the simulations evolved to form new MTase-RdRd binding interfaces for the two high-energy cases, which suggests the high-energy conformations in the CG force field are also strongly repulsed in the conventional atomistic force field.

Figure 2.10 The RMSD evolution with simulation time in the conventional all-atom MD simulations starting from the simulated ensemble structures. The backbone RMSD for MTase, RdRp and the complex are shown as blue, orange and red lines respectively. The cases of (a) and (b) are the all-atom simulations with the two 5% low-energy ensemble structures as the initial structures. The structure comparison was shown between the initial structure and the representative structure of the all-atom simulation (green and red ribbon respectively), with their RdRp structure well-aligned. (c) and (d) are the all-atom MD simulations starting from two high-energy structures that exist in both NNLS and EOM-based structure ensembles without the energy constraint. The structure alignment was performed between the initial structure and the last-frame structure of the all-atom simulation trajectory.

2.8 Theoretical investigation of the noise in data fitting

The ensemble construction without the energy constraint produced a noisy conformational ensemble that contains suspicious bad-energy components. The high-energy conformational space might result from data over-fitting instead of the “true” conformations, because 1) they deviate against the protein energy profiles and 2) they fail to be consistently reproduced by different optimization methods or with different candidate pools. To prove this hypothesis, we performed a theoretical investigation on the fitting noise with a synthesized SAXS profile. The pseudo-experimental SAXS data was the fitting target in this theoretical study, which was synthesized from a pseudo ensemble of DENV2 NS5. The solution pseudo ensemble consists of the top ten low-energy structures (10% proportion for each) from our 1015-structure simulated pool. To make synthesized SAXS data more realistic to the experimental condition, random noise was added in the synthesized data in computational works [105], as the experimental observables always contain noise in the measured signal. We added 5% Gaussian noise in the pseudo-experimental SAXS intensity, and the choice of 5% noise level would contribute to comparable χ^2 values with those from fitting experimental data.

The ensemble construction for pseudo-experimental data recovery was performed using EOM program with the simulated pools of different energy constraints. The fitting results are shown in Figure 2.11, where the low-energy ensembles from 5%

to 30% ($\chi^2 = 0.86 \sim 0.87$) consistently reproduce the conformational space of the predefined solution ensemble. Interestingly, the constructed ensemble without the energy constraint has its component structures with a more disperse and noisy distribution of domain-domain orientation. The absence of the energy regularization slightly improves the consistency ($\chi^2 = 0.84$) with the pseudo-experimental data, but this improvement is a result of data over-fitting because the recovered conformational space is wrong. It is clearly demonstrated in Figure 2.11 that three ensemble structures are located outside the correct region of the solution conformations. We noted that one structure ($\varphi \approx 45^\circ$, $\theta \approx 104^\circ$) has a novel domain-domain orientation remarkably deviating against correct orientation space. This structure turns out lacking favorable energy and has been precluded from the low-energy structure pools for ensemble construction. Our investigation demonstrates the important role of the physics-based energy regularization in ensemble construction with SAXS data fitting from a theoretical perspective. A raw numerical fitting without physics is highly likely to over-interpret the data with a spurious representation of conformational space.

Figure 2.11 The ensembles constructed from a pseudo-experimental SAXS profile projected on the domain-domain orientation map. The 5%, 15% and 30% low-energy as well as the 100% simulated pools were used for ensemble construction with synthetic SAXS data. The qualities of fit are $\chi^2 = 0.87, 0.86, 0.86$ and 0.84 respectively.

2.9 Energy characterization of a structure pool from REMD simulations

This section is to illustrate the importance of energy on the simulated structure pool, because structure energy is usually neglected when the simulated structures are used to fit experimental data.

The sampling-and-selection optimization scheme is the most popular strategy to identify structure ensemble with experimental data. However, the reasonableness of the over-sampled pool structures cannot be guaranteed, which imposes severe uncertainty to the reliability of the constructed ensemble. The structures from physics-based molecular modeling, such as MD and MC simulations, are more robust in physics than the structure pool that is produced purely on the basis of topology and geometry. Previous studies adopted REMD simulation to generate an enhanced-sampling ensemble and compared it with experimental observables [63, 106]. Here, we performed REMD simulations on DENV2 NS5 and characterized the simulated structures in an energy perspective. A total of 24 replica MD simulations were run from 300 to 613 K using the same CG model as described above. Each replica simulation was run for 400 ns, and finally the simulated ensemble at 300 K was taken as a pool consisting of 1000 structure. As shown in Figure 2.12, the domain-domain orientation distribution from this REMD-based pool shows a much narrower space compared with that from our method (Figure 2.3) despite the total simulation time is comparable. The sampled space is around the initial structure, indicating the REMD simulation is still trapped around the conformation near the starting point. Notably, even the REMD-based pool is much limited than our simulated pool, large energy difference among the REMD-based structures has also been found (Table 2.2). This observation suggests the crucial role of energy

characterization of the candidate structures for any given simulated pool in structure ensemble construction.

Figure 2.12 Domain-domain orientation distribution of a structure pool resulted from REMD simulations. 24 replica simulations were run on DENV2 NS5 for 400 ns each, ranging from 300 to 613 K. The structures from the ensemble at 300 K were taken as a pool for further characterization.

Table 2.2 The profiles of the sub-pools based on energy levels. The structures from the ensemble at 300 K in REMD simulations were ranked on the basis of the inter-domain energy (kcal/mol) from low to high, followed by dividing the pool into 10 energy-ranked sub-pools.

energy-ranked sub-pool of REMD structures	average energy
0% ~ 10%	-144
10% ~ 20%	-138
20% ~ 30%	-136
30% ~ 40%	-134
40% ~ 50%	-132
50% ~ 60%	-130
60% ~ 70%	-128
70% ~ 80%	-126
80% ~ 90%	-123
90% ~ 100%	-117

2.10 Conclusion

The large protein complexes were modeled in our hybrid coarse-grained model, which takes the advantage of the greatly enhanced computational efficiency for a more complete explore on protein conformations. Moreover, the CG structures suffice the medium/low-resolution structural information of SAXS data. The model characterizes inter-domain interactions including flexible unstructured segments with hydrophobic and electrostatic energies. Flexible segments have been demonstrated to mediate molecular interactions and the linker critically affects the global conformational arrangement of NS5 [89]. The model is potential to apply to all complicate biomolecules with well-structural units connected via intrinsically disordered regions that has been received an increasing attention for their critical roles in molecular functions like signaling [107].

Deriving an ensemble representative from the noisy experimental observables is a hot research topic. The key problem is data over-interpretation, or called over-fitting. Despite many works have been published on ensemble modeling with SAXS intensities, a systematic examination of over-fitting from a random structure pool is absent. Importantly, the representative conformations of the recovered ensemble should be able to reflect the true conformational space and molecular interactions, more than a numerical agreement with the given averaged experimental quantities. We found the raw fitting to experimental SAXS profile caused the arbitrary

conformational representations of NS5, which has been illustrated as a result of data over-fitting from the noise of the high-energy pool structures. The usage of two independent optimization methods also illustrated the instability of the ensemble solution without the energy constraint. Our results show the significant role of the physics-based regularization on the prevention of the spurious conformational representations from a fitting scheme. The idea of physics-based constraints can be integrated with different ensemble optimization methods and simulation protocols.

3 Structure ensemble modeling of PubRRM12 with energy-based pool structures and experimental distance constraints

3.1 Background

NMR methods provide distance and angular information on atomistic resolution to assist the structural and dynamic studies of biomolecules. NMR spectroscopy techniques can measure chemical shifts, scalar couplings, residual dipolar couplings and PRE effects, among which PRE method [108, 109] is able to offer valuable long-range structural information as long as 10 ~ 35 Å thanks to the strong PRE effects contributed by the large magnetic moment of unpaired electron. The PRE effects refer to the enhancement in the relaxation rate of nuclear magnetization caused by the magnetic interactions between a proton and an unpaired electron in the spin label. The magnetic interactions between the electron and proton extends to a long range, and this strong effect provides long-range distance information in biomolecular structures [69]. Therefore, PRE method has been widely employed to study large biomolecules including multi-protein complexes [110-114] and protein-DNA interactions [115, 116].

The challenge of PRE-based structure determination is the lack of a single structure model to satisfy all PRE constraints. Even PRE fitting can identify a single structure that agrees well with PRE rates, this structure is not a correct conformational

representative [69]. This scenario is attributed to the conformational flexibility of the studied biomolecules. Thus, structure ensemble instead of a single structure model is used as the conformational representation of flexible biomolecules with their PRE measurements [117]. Many works on PRE-based structure optimization employed the known experimental structures [112, 114] or simulated structures [65] as the major components of the ensemble, followed by ensemble refinement with some minor states added in. However, the prior-knowledge of ensemble structures is absent in most applications, which requires a common strategy for structure ensemble determination with PRE data. The over-fitting problem and its regularization methods have been comprehensively discussed in the previous chapters in this thesis.

In this chapter, the protein system is a multi-domain Poly(U)-binding protein (Pub1) from *Saccharomyces cerevisiae*, which is a RNA binding protein serving as a critical role in the regulation of cellular mRNA decay [118, 119]. We performed structure ensemble modeling of its two RNA recognition motif domains connected via a ten-residue linker, referred as PubRRM12. The individual domain structures have been solved in NMR and crystallography, and the crystal structure [120] of the di-domain PubRRM12 shows no interactions between the two domains. However, the NMR experiments [121] demonstrated the existence of weak interactions between the two domains, which cannot be explained by the crystal structure. PRE experiments provided the long-range distance structural information that is able to be utilized as

the distance constraints to construct the structure ensemble of PubRRM12. Like the hybrid model on DENV2 NS5 above, we adopted a structure-based potential to model intra-domain structure and a physics-based potential for inter-domain associations. Structure-based models have also been implemented on the all-atom resolution in the previous works [39, 122], despite they were mostly adopted with CG structure models. Considering the high sensitivity of PRE rate to inter-atomic distance, we simulated PubRRM12 on the all-atom resolution. We run restrained MD simulations with the enhanced sampling techniques and then characterized the simulated structures in an energy perspective. The clustered structures from MD simulations were subjected to ensemble construction of PubRRM12 with the published PRE data in literature [121].

We highlight the PRE-based ensemble construction in our work in several aspects as follows. Firstly, we depicted the energy landscape of a multi-domain protein PubRRM12 based on an all-atom force field and validated it with experimental PRE measurements. Besides, we constructed the ensemble with both structure energy and ensemble size as the regularization terms to prevent the data over-fitting problem and systematically examined their regularization effects on PRE-based ensemble optimization. Lastly, apart from the structure pool from MD simulations on the atomistic level, we also recovered the ensemble from PRE rates using a topology-based pool and investigated the effect of the pool sources on the results of ensemble construction.

3.2 Methods

3.2.1 Atomistic model of PubRRM12

The di-domain protein PubRRM12 was modeled in a hybrid fashion. The inter-domain interactions were simulated with an atomistic force field. The two individual domains, referred as RRM1 and RRM2, were subjected to a structure-based model. The structure-based model was applied to RRM1 and RRM2 individually, because their solution structures are very stable and have been solved in NMR experiments [121]. But the solution NMR experiments also identified some loops in PubRRM12 that are flexible and disordered in solution. These flexible segments are one RRM2 loop (residue 195-202), the inter-domain linker (residue 149-160) and two terminal loops (Figure 3.1a), which were not modeled with the structure-based potential. Thus, PubRRM12 structure was decomposed into stable part and flexible part.

Figure 3.1 Structure of the multi-domain protein PubRRM12 and the domain-domain orientation. (a) RRM1 (left) and RRM2 (right) domains have stable conformations (green) individually, while linker and some loop segments are flexible (red) with their residue number indicated. (b) The RRM1-RRM2 orientation is defined by two spherical angles φ and θ . The self-rotation (blue arrow) of RRM2 is not included in the spherical coordinates. (c) RRM2 is pulled away from RRM1 (center) to 160 uniformly-distributed positions (red beads).

The overall potential energy function is formulated in Equation 3.1, including bonded term V_{bonded} and non-bonded term $V_{nonbonded}$.

$$V = V_{bonded} + V_{nonbonded} \quad (3.1)$$

The bonded term is formulated in Equation 3.2, including bond, angle, improper and proper dihedrals. The rigid parts of RRM1 and RRM2 adopted their individual experimental structures as the global energy minima in the structure-based potential. The values of equilibrium positions b_0 , θ_0 , χ_0 and φ_0 were taken from their respective experimental structures. The energy strengths of these bonded terms are 2000 kJ/(mol · Å²), 150 kJ/(mol · rad²), 5 kJ/(mol · rad²), and 8 kJ/mol for k_b , k_θ , k_χ , and k_φ respectively, which are comparable to their counterparts in the atomistic force field Amber99SB [104]. For the flexible structure parts, the equilibrium positions and interaction strengths were directly derived from Amber99SB.

$$\begin{aligned} V_{bonded} = & \sum_{bond} k_b (b - b_0)^2 \\ & + \sum_{angle} k_\theta (\theta - \theta_0)^2 \\ & + \sum_{\substack{improper/ \\ planar}} k_\chi (\chi - \chi_0)^2 \\ & + \sum_{dihedral} k_\varphi \left\{ [1 - \cos(\varphi - \varphi_0)] + \frac{1}{2} [1 - \cos 3(\varphi - \varphi_0)] \right\} \end{aligned} \quad (3.2)$$

The non-bonded term is shown in Equation 3.3.

$$V_{nonbonded} = V_{ff} + \sum_{native} \varepsilon \left[\left(\frac{r_{0ij}}{r_{ij}} \right)^{12} - 2 \left(\frac{r_{0ij}}{r_{ij}} \right)^6 \right] \quad (3.3)$$

The term V_{ff} in Equation 3.3 is the non-bonded potential energy from the atomistic force field Amber99SB, consisting of V_{ff-LJ} and V_{ff-ele} terms. V_{ff-LJ} is a Lennard-Jones formulated function $V_{ff-LJ} = \sum_{ij} (A_{ij}/r_{ij}^{12} - C_{ij}/r_{ij}^6)$, where the parameters A_{ij} and C_{ij} are taken from Amber99SB for the pair of atom type i and j . r_{ij} is the distance between the two atoms. The term V_{ff-ele} is electrostatic potential, which is formulated as a Debye–Hückel-type function $V_{ff-ele} = \sum_{ij} q_i q_j \exp(-r_{ij}/\xi) / 4\pi r_{ij} D$. The electrostatic energy between atom i and j is dependent on their inter-atomic distance r_{ij} and respective charge q_i and q_j . The Debye length ξ is equal to 14 Å to reflect the screening effect of the salt environment of 20mM sodium phosphate and 20mM NaCl in the experiment. We did not have explicit water in the MD simulations for a greatly enhanced sampling. The purpose of the MD simulations here is to generate a structure pool for further structure selection with experimental data. Therefore, the solvation is modeled with a dielectric constant, which suffices the purpose of simulating the basis of molecular interaction and precluding unphysical conformations from the pool. In our study, a choice of dielectric constant D is 4 to model the dielectric environment of protein contacting interfaces, which is consistent with many previous works [123-128].

Besides V_{ff} , the other component term of the non-bonded energy $V_{nonbonded}$ is a structure-based potential to stabilize the individual domains. As shown in Equation 3.3, the intra-domain native pairs are subjected to a Lennard-Jones potential, where r_{0ij} is the pairwise distance between the heavy atom i and j from the native experimental structures. The native pairs are the pairs of the spatially adjacent heavy atoms in the same domain excluding the flexible loops, determined by shadow map [93]. Notably, the native pairs are exempted from the computation of V_{ff} potential energy. The contacting strength of the structure-based potential ε is parameterized to 50 kJ/mol.

3.2.2 Sampling and simulation details

One severe challenge of MD simulations is the insufficient sampling. The simulation is usually trapped in some local minima and fails to explore the distant conformational space. In this work, we employed the pull-push plus high-temperature simulation strategies to generate the initial structures for further production MD simulations. These initial structures basically cover all the available conformational space and serve as good starting points in the production MD simulations.

A three-step sampling scheme was carried out to generate initial structures. Firstly, centered at the center of mass (COM) of RRM1, 160 directions were generated by a published method [129] to evenly divide the whole domain-domain orientation space (Figure 3.1c). RRM2 was pulled towards the positions that the 160 direction vectors point at, in 160 independent pulling MD simulations. This pulling step assisted to sample all the domain-domain orientations. Then, the 160 structures after pulling simulations were taken for independent MD simulations at 1000K, where RRM1 was kept rigid and RRM2 had a harmonic restraint on its center of mass. Each high-temperature simulation was run for 10 ns with a time step of 0.001 ps, and each resultant trajectory at 1000K was subjected to the structure clustering [96] of RRM2 using a backbone RMSD cutoff at 3 Å. The high-temperature step is to sample the domain self-rotations, and each simulation usually resulted in 2 ~ 6 structure clusters after clustering. Next, the middle structure of each cluster was taken out followed by being pushed back to RRM1 along the RRM1-RRM2 COM direction, which finally resulted in a total of 611 structures. In the whole sampling process of pull-push and high-temperature MD simulations, RRM1 was fixed as rigid body, and RRM2 was simulated with the potential in Equation 3.1 to generate the initial structures for production MD simulations.

The 611 structures were subjected to independent production MD simulations at 300K for 20 ns each with a time step 0.002 ps using the package GROMACS 4.5 [130]. To guarantee a more complete sampling of domain-domain orientations,

position restraints were imposed on the production MD simulations. RRM1 was subjected to harmonic restraints on its heavy atoms. In addition, RRM2 can move only along the direction connecting the domain-domain COMs by adding weak harmonic restraints along the other two perpendicular directions. If no restraint is imposed, some replicas might deviate from their domain-domain orientations due to the energetically unfavorable initial conformations, consequently causing the loss of orientation sampling. Furthermore, the weak strength of restraints permits the flexibility of domain-domain interactions to cover adjacent conformational space.

3.2.3 Pool generation and characterization

A total of 152750 structures were generated after the production MD simulations. The huge amounts of the simulated structures were grouped according to their structural similarities. The clustering approach is as same as that in the previous NS5 chapter (see Chapter 2.2.3). Firstly, all the structures were grouped into sub-groups of domain-domain orientations based on φ and θ angles as shown in Figure 3.1b. The orientation sub-groups are sized as $\Delta\varphi = 30^\circ$ and $\Delta\theta = 30^\circ$. Then, within each orientation sub-groups, a RMSD-based clustering [96] was implemented with a backbone RMSD cutoff of 1 Å, finally producing a total of 4003 clusters. A small value of RMSD cutoff was chosen because of the high sensitivity of PRE rates to inter-atomic distances. The middle structure of each cluster was extracted as the

representative structure, which constituted a 4003-structure pool for the further ensemble construction with experimental PRE data.

The pool was characterized in an energy perspective. The molecular mechanics Poisson-Boltzmann/surface area (MMPBSA) method [131] was employed to compute the structure energy. The MMPBSA calculations were carried out on the 4003 pool structures using the program MMPBSA.py [132] with the package Amber 14 and AmberTools 14. The MMPBSA energy, referred as E_{MMPBSA} , contains the potential energy term E_{MM} and the solvation free energy term $G_{solvation}$. Specifically, E_{MM} is the non-bonded potential energy based on the molecular mechanics (MM) Amber99SB. E_{MM} does not incorporate the intra-domain interactions, because intra-domain structures were simulated with a structure-based potential. $G_{solvation}$ consists of polar solvation free energy and nonpolar solvation free energy. The polar solvation energy is determined with the Poisson-Boltzmann method [133, 134], and the nonpolar component contains attractive dispersion and repulsive cavitation [135]. The interior environment of the solute has a dielectric constant equal to 1, and the exterior environment is set to 80. The energy calculations were performed on each structure in the candidate pool, followed by pool characterization based on structure energies E_{MMPBSA} . The low-energy (E_{MMPBSA}) structures were extracted out to constitute the low-energy candidate pools for the further low-energy ensemble construction with the experimental PRE data.

3.2.4 Ensemble construction with experimental PRE rates

In PRE experiments of PubRRM12, its amino acids M107, H123, N148, S190 and N218 were subjected to mutation to Cys, followed by attaching a label molecule MTSL to the mutated Cys connected by a disulfide bridge. The atomic partial charges were calculated over the MTSL molecule by Gaussian [136]. The MTSL molecules were treated as flexible structures and excluded from the structure-based potential as they did not have their local conformations determined experimentally.

The distance between MTSL oxygen atom and protein backbone amide hydrogen atom is correlated with their PRE rates. As shown in Equation 3.4, PRE rate Γ_2^{cal} is proportional to the average distance $\langle r_i^{-6} \rangle$, where K is equal to $1.23 \times 10^{-44} \text{ m}^6/\text{s}^2$, the correlation time τ_c^{app} is equal to $6 \times 10^{-9} \text{ s}$, and Larmor frequency ω_H is equal to $5.03 \times 10^9 \text{ rad/s}$ for 800 MHz spectrometers. Each pool structure was taken to calculate its PRE rates and a combination of multiple structures produced an ensemble average of the calculated PRE rates.

$$\Gamma_2^{\text{cal}}(i) = K \langle r_i^{-6} \rangle \left[4\tau_c^{\text{app}} + \frac{3\tau_c^{\text{app}}}{1 + (\omega_H \tau_c^{\text{app}})^2} \right] \quad (3.4)$$

The agreement between the selected ensemble and the experimentally measured PRE rates for residue i is evaluated by the Q-factor as formulated in Equation 3.5 [117].

When the values of Γ_2^{obs} and Γ_2^{cal} are larger than 80 s^{-1} , they would be set to 80 for the calculation of Q-factor. The details of Equation 3.4 and 3.5 have been discussed in literature [121].

$$Q = \left[\frac{\sum_i \{\Gamma_2^{\text{obs}}(i) - \Gamma_2^{\text{cal}}(i)\}^2}{\sum_i \{\Gamma_2^{\text{obs}}(i)\}^2} \right]^{1/2} \quad (3.5)$$

In this work, we recovered the structure ensemble from experimental PRE data with the regularizations of ensemble size and structure energy. A small-sized ensemble is popularly employed to prevent an ambiguous data interpretation. Following this strategy, SES [67] seeks to recover an ensemble for a best fit to experimental data by increasing the ensemble size until a converged plateau. SES selects top K best solutions instead of one when increasing the ensemble size by adding a new structure, where a large value of the parameter K leads to a more global optimization but a heavier computation burden. Inspired by the idea of the multi-path optimization, we developed a greedy pursuit method for PRE-based ensemble construction with the ensemble size constraint, named Multiple Q-factor Pursuit (MQP).

The MQP optimization scheme is outlined here. 1). In the initial step, all the combinations $N(N-1)/2$ between two structures are taken into consideration where N is the total number of candidate structures. The top K combinations that result in the smallest Q-factor values are selected as top K two-structure ensembles for next step.

This ergodic-like search of all possible two-structure ensembles guarantees a good starting point of optimization. 2). Then, for each of the top K ensembles sized two (consisting of two structures) returned from the previous step, a new structure from the remaining pool of $(N-2)$ structures is added to minimize Q-factor. There are $K*(N-2)$ combinations of the three-structure ensembles, among which the top K solutions of the lowest Q-factors are selected for future steps. The weights of ensemble structures are determined by performing NNLS with the constraint that the summation of all conformation populations is 100%. 3). The top K four-structure ensembles are determined with the same method as described in the previous step. 4). This procedure evolves to some desired steps.

3.3 Model validation and sampling efficiency

The individual conformations of RRM1 and RRM2 are stable and the conformational flexibility arises from the inter-domain associations. In our simulations, we imposed a structure-based potential on the individual domains, which stabilized the conformations of individual RRM1 and RRM2 in the MD simulations (Figure 3.2a). Without the structure-based potential, the individual domains would deviate from their native structures (Figure 3.2b).

Figure 3.2 Backbone RMSD of the structures from the MD simulations with respect to the experimental structures of individual domains. (a) Small RMSD values are found for the simulated structures with the structure-based potential with respect to the experimental RRM1 and RRM2 structures. (b) Without the structure-based potential, the simulated structures have much larger RMSD values against the experimental domain structures.

The MD simulations generated the structures of PubRRM12 for further candidate pool construction and ensemble fitting. It is well known that a single MD simulation usually encounters with the problem of insufficient sampling, especially for flexible biomolecules. For a complete explore on the conformational space, multiple simulation trajectories starting from various conformations have been employed in the previous works [62, 137]. In our study, a similar strategy was adopted. The production MD simulations started from 661 different structures (Figure 3.3) that

basically cover the entire domain-domain orientation space via the pull-push sampling technique. The sampling of the domain self-rotation is also improved by the high-temperature simulations (see Chapter 3.2.2). The sampling space is revealed on the domain-domain orientation map. As shown in Figure 3.3, the simulations are able to achieve a complete sampling of the orientation space despite of some inaccessible orientations caused by the length constraint of the linker. Compared with the plain MD simulations and REMD simulations, the MD simulations with our sampling methods explored a much more extensive conformational space (Figure 3.4). The plain MD simulations without enhanced sampling methods show a narrow orientation space around their initial structures (Figure 3.4a and b), indicating that the plain MD simulations might fail to jump out from some local conformations. Besides, we also demonstrated the two-dimensional conformational space of the PubRRM12 backbone RMSD and the RRM1-RRM2 COM distance. As shown in Figure 3.5, our enhanced sampling simulations significantly outperform the plain MD and REMD simulations in terms of the extensive search of the conformational space.

Figure 3.3 The initial structures in the production MD simulations and the sampled conformational space in the production MD simulations. Left: The 661 initial structures cover all the available orientations between RRM1 (green surface) and RRM2 (ribbon representations). Right: Starting from the 661 structures, the MD simulations generated a very complete sampling of domain-domain orientation space.

Figure 3.4 A comparison of the sampling efficiency between the simulations with and without the enhanced sampling methods. (a) and (b) are the domain-domain orientation space sampled by the 400-ns MD simulation without any sampling method, starting from two different structures respectively. (c) is the orientation space sampled by a 40-ns REMD simulation from 300K to 720K. RRM1 and linker are colored in green and RRM2 is colored in blue. (d) is the case for our sampling method, same as Figure 3.3.

Figure 3.5 A comparison of the sampling efficiency between the plain and the enhanced sampling simulations with respect to the conformational space depicted by RMSD and domain-domain distance. The plain MD simulations (a and b), the REMD simulation (c) and the enhanced sampling simulations (d) have been described in Figure 3.4. The two-dimensional conformational space here is the backbone RMSD against the same reference structure and the domain-domain COM distance.

3.4 Energy landscape of inter-domain interactions

An energy landscape of domain-domain interactions is a useful tool to reveal the conformational distribution of a multi-domain protein in an energy perspective [62]. The two-dimensional energy landscape based on the data describes the domain-

domain interactions of PubRRM12, where the depth of a minimum and the number of clusters in the minimum indicate the energetic stabilization and the conformational entropy respectively [37]. In this work, domain-domain orientation defined by angles φ and θ in Figure 3.1b is selected to present the conformational space of energy landscape. The φ versus θ space depicting domain-domain orientations was divided uniformly into the grids sized $18^\circ \times 9^\circ$. To organize the great amount of simulated structures, a two-step clustering procedure was performed, consequently leading to a total of 4003 representative conformations. Each representative structure fell into its corresponding grid on the conformational map according to its φ and θ values. In each grid, the structure of the lowest energy E_{MMPBSA} was extracted to represent the energy of that grid. To explicitly display the energy contrast among most areas on the landscape, we performed a search on resetting the zero energy point, consequently leading to the new zero point that is shifted from the grid energy value ranked at the lowest ~46% among all non-empty grids on the landscape.

The energy landscape (Figure 3.6) is found as a good illustration to suggest the distribution of energetically favorable domain-domain orientations. Generally, a number of areas of energy minima have been identified on the landscape and each minimum contains multiple low-energy structures. The most striking feature is that these low-energy areas are contiguous on the landscape, constituting a wide energy minimum whose energy level is dramatically lower than other conformational areas.

Typically, a relatively smooth “funnel-like” energy landscape for multi-protein complexes is considered as a reflection of guiding individual components to specific binding [42, 138]. In our case, the observation of multiple low-energy areas reflects the rugged energy landscape of our multi-domain protein, suggesting the absence of strong binding between the two domains. The chemical shift perturbations from NMR experiments demonstrated the existence of weak interactions between the two domains [121]. Besides, several experimental works on nucleolin [139, 140] and sex lethal [141] suggested that RRM domains adopt high conformational flexibility without specific binding in solution and RNA-binding mediates domain-domain interactions. Our energy landscape is further validated by experimental observables in the chapters below. Note that some structures located in low-energy areas of domain-domain orientations might still be high-energy structures because of the self-rotation of individual domains.

Figure 3.6 Energy landscape of inter-domain interactions of PubRRM12. The energy of each domain-domain orientation represented in an $18^\circ \times 9^\circ$ grid is indicated in colors defined by the color bar (kcal/mol).

3.5 Ensemble construction with experimental data

The production simulations achieve a better sampling efficiency at the expense of abandoning the rigorous canonical distribution, thus the populations of conformations require a reweighting procedure to fit the experimental PRE rates. Nevertheless, it is not a drawback of our sampling approach since the direct computation of PRE rates from an equilibrated MD trajectory is typically formidable

in terms of computational expense and force field accuracy. Here, we aim to explore the conformational arrangement of the two domains, therefore only inter-domain PRE rates, referring to the MTSL and amide hydrogen located in different domains, are used for data fitting. The simulated PRE rates are calculated from pair distances between backbone amide hydrogens and MTSL oxygens (see Equation 3.4). The PRE fitting against experimental data was performed by using a set of structure pools based on energy. Specifically, we applied our MQP method to construct ensembles from top 5%, 10%, 20%, 30%, 40%, 50% and 100% low-energy structure pools, to fit the experimental PRE rates. In the rest of this article, a constructed ensemble optimized from the pool of top a% low-energy structures is called a top a% ensemble.

The solution sparsity of the recovered ensemble is determined by the nature of the biological system. A smooth energy landscape with one or several significant energy minima reflects the specific-binding conformations of multi-protein complexes, while a rugged landscape suggests the absence of specific binding. Without the prior knowledge about the “real” solution such as pre-determined conformations from experiments, a small-sized ensemble, empirically no more than 20 structures [69], was recommended, because larger-sized solutions would cause an ambiguous over-interpretation of experimental data. Besides, another popular approach Ensemble Optimization Method [73] performs ensemble fitting of the SAXS scattering profile from flexible biomolecules and it predefines 20 conformations by default to

represent flexible systems. In our study, we adopted MQP approach to construct ensemble sized up to 20 structures, with different energy-level pools.

Figure 3.7 shows the Q-factor evolution as a function of ensemble size. Q-factors are minimized monotonically for all energy-based ensembles, and the low-energy ensembles that were constructed from low-energy structure pools exhibit a plateau-like pattern within 20 structures. Specifically, the top 5% ensemble produces a converged Q-factor of 0.47 at 11 structures whose fitting quality fails in further improvement with more structures introduced. The enlarged pool of top 10% minimizes the Q-factor to a plateau value of 0.44 using 10 structures. For top 20% and 30% low-energy ensembles, 14 structures are able to lower the Q-factors to 0.39 and 0.35 respectively, an equal Q-factor as their respective 20-structures counterpart. However, this plateau-like pattern disappears from top 40% ensemble onward. As shown in Figure 3.7, the Q-factors of top 40% and 100% ensembles display a continuously decreasing trend within ensemble size up to 20, which imposes a problem of arbitrary ensemble selection without a plateau Q-factor. A looser energy control is likely to expose the constructed ensemble to the noise from high-energy structures, consequently failing to construct a structure ensemble with the small-size constraint. Ensemble-size regularization discourages larger ensembles consisting of over 20 structures as it would over-interpret the ambiguity of PRE data. Besides ensemble size, energy is proposed as another control factor to prevent over-fitting, since energetically unfavorable conformations are of little probability to exist even

though they might improve the quality of fit. High-energy structures cause the constructed ensembles doubtful in physics. Therefore, the cross-regularization of ensemble size and energy suggests the use of 14-structure top 30% ensemble to represent the conformation ensemble because of its good performance in both fitting quality and the low-energy control.

Figure 3.7 PRE Q-factor evolution with ensemble size for top 5%, 10%, 20%, 30%, 40% and 100% ensembles.

Here, the role of energy exclusion in ensemble construction from a given structure pool was emphasized. A structures pool from any computational methods is likely to encounter with significant energy contrast between different sampled poses, while

the high-energy one might be selected to represent the conformational space without the energy regularization. For example, MD [62, 63, 106] and Monte Carlo [65, 106] simulations, sometimes with replica exchange that is a popular way to enhance sampling, were adopted to generate candidate structures for a subsequent comparison with experiments. A 200-structure pool from the simulated ensemble at 300K of a 40 ns REMD simulation was examined using the same method of energetic analyses above. Large energy differences were observed between structures of PubRRM12 (Table 3.1), especially considering its conformation sampling is much less comprehensive (Figure 3.4c and Figure 3.5c) than our method. Therefore, energy aspect should be investigated in ensemble construction from a simulated pool to prevent data over-fitting with bad-energy structures.

Table 3.1 Energy difference between different energy-ranked sub-pool of PubRRM12 structures from REMD simulations. 200 structures from a 40-ns REMD simulation were collected as a structure pool and ranked according to the MMPBSA energy (kcal/mol). This pool was further divided into 5 sub-pools on the basis of structure energy, with the average energy of the highest-energy sub-pool shifting to zero energy.

Energy-ranked Sub-pool of REMD Structures	Average Energy
0% ~ 20%	-221
20% ~ 40%	-135
40% ~ 60%	-94
60% ~ 80%	-52
80% ~ 100%	0

3.6 Conformation analysis on constructed ensembles

We investigated the distribution patterns of domain-domain orientation for the constructed low-energy ensembles. The conformational space of domain-domain orientation is depicted in an energy perspective, where a wide major minimum was

identified merged by multiple low-energy areas (Figure 3.6). Figure 3.8a compares the orientation distributions of the structures from top 5%, 20%, 30% and 100% ensembles. It is noted that the low-energy 5%, 20% and 30% ensembles are the smallest-sized ensembles with the plateau Q-factors as shown in Figure 3.7. The 100% ensemble is the 20-structure one as it lacks the plateau pattern of the Q-factor evolution with ensemble size. Despite of their different energetic levels, Q-factors and ensemble sizes, their component structures show a highly similar preference on the domain-domain orientation that is located in the energy minimum surface. This conserved conformational space is supported by both simulated energy landscape and experimental PRE measurements. Interestingly, the 100% ensemble without the energy constraint adopts most of its component structures located in or adjacent to the energy minimum, but three structures ($\varphi < -60^\circ$) of very different orientations located in high-energy area in the energy landscape. It suggests that the major conformational space can be roughly identified from a raw mathematic fit of the experimental PRE observables, but the absence of energy regularization would recover experimental data to a noisier conformational representation caused by high-energy conformations. To further investigate the ensemble structures corresponding to the conserved orientation, the inter-domain residue contacts are counted when the residue in RRM1 domain forms a pair-wised contact with the RRM2 residue with their C α distance less than 10 Å. The inter-domain contacting residues that have most contact numbers are spotted for the contacting interface analysis (Figure 3.8b). In

NMR experiments [121], some residues with significant chemical shift perturbations were identified indicating these residues are involved in the domain-domain interaction interface. In RRM1, all the residues with significant chemical shift perturbations are located in its β -sheet and adjacent loops, indicating this RRM1 β -sheet face dominates the contacting interface. In RRM2, these residues are from its β -sheet face and also other faces (V171, N172 and Q226) including Q226 in the opposite α -helix face. Compared to the experimental suggestions above, our low-energy ensembles also target the RRM1 β -sheet surface as the predominant domain-domain interacting area (Figure 3.8b). Actually, the RRM1 β -sheet face does correspond to the major minimum area on the domain-domain orientation map (Figure 3.8a) because the nature of this φ - θ conformational space is the interaction interface of RRM1. However, the 100% ensemble without the energy constraint also targets another novel binding face located at the end of one helix segment in RRM1 domain shown in Figure 3.8b. This new contacting segment in RRM1 is not supported by chemical shift perturbation experiment, indicating the 100% ensemble without the energy constraint may not reflect the real conformational space and molecular interaction. Besides RRM1, The contacting interfaces on RRM2 from all the constructed ensembles are more dispersed than RRM1, just as the chemical shift perturbations indicated.

Figure 3.8 Conserved domain-domain orientation and domain contacting interface identified by the constructed low-energy ensembles. (a). It shows the domain-domain orientation distribution of ensemble component structures of top 5%, 20%, 30% and 100% ensembles, shown as blue points. The red area indicates low-energy surface on the energy landscape with a cutoff at top 5% low energy. (b). Ribbon representation of RRM1 is shown with the contacting residues highlighted in red whose contact numbers are ranked within top 20% among all RRM1 residues.

For comparison, the previous work [121] on PubRRM12 ensemble determination with a genetic algorithm approach used six orthogonal parameters to describe domain-domain relative positions assuming that each domain is a pure rigid body. Aiming for using a minimal-size ensemble to recover experimental measurements, four representative conformations were proposed to reproduce PREs of PubRRM12, and this ensemble also had majorly RRM1 β -sheet facing towards RRM2 domain.

Apart from the cross-validation from chemical shift perturbations above, the small-angle X-ray scattering experiments [121] also indicated the existence of multiple conformations of PubRRM12 in solution and derived an approximation of a radius of gyration (R_g) of 18.66 ± 0.53 Å. The calculated R_g of the top 30% low-energy ensemble is 17.44 Å, consistent with the experimental approximation. Thus the constructed low-energy ensembles are further supported by SAXS data in terms of the radius of gyration.

3.7 The effect of regularizations

In this study, ensemble construction was performed with a combined use of energy-based physics regularization and L0-norm based mathematic regularization. The 20-structure ensemble constructed from the 100% pool does not take low-energy control in PRE fitting. Although the 100% ensemble can target the major conformational space of interest that is governed by the experimental distance constraints, three component structures of high-energy orientations ($\varphi < -60^\circ$) cause confusion in the conformational representation of PubRRM12. These three structures of novel domain-domain orientations were subjected to energetic analysis to investigate whether there might exist such minor states. A previous work [142] conducted a theoretical investigation and found that a raw numerical fitting without physics-based constraint may cause the wrong conformational representations of the

investigated protein. As shown in Figure 3.9, these three structures turn out high-energy structures and their binding modes are highly rejected by electrostatics. Besides, their domain contacting interfaces in RRM1 are novel (Figure 3.9), totally deviating from RRM1 beta-sheet face that was suggested by chemical shift perturbations. Therefore, the three minor states may be an over-fitting result of experimental data without the energy regularization because of its violation against both experimental observables and energy landscape.

Figure 3.9 Three conformations of minor domain-domain orientations in the constructed 100% ensemble and their energetic profiles. ΔE (kcal/mol) is the difference between the conformation energy and the average energy over all 4003 representative structures. ΔE_{ele} , ΔE_{vdw} , and ΔE_{sol} are the electrostatic, van der Waals and solvation components of energy difference. RRM1 structures are colored on the basis of its secondary structure and RRM2 structures are shown in green.

The constraint type of ensemble size prevents an ambiguous interpretation of experimental data and has been implemented in ensemble recovery in many studies [65, 67, 70-72, 121, 143, 144]. To examine the effect of ensemble size on ensemble construction of PubRRM12, the method of non-negativity least squares [100] was adopted to reweight candidate structures to fit experimental data best, by assuming

the total population to be 100%. With the NNLS optimization scheme, we did not limit the number of structures with non-zero populations, which enables a direct comparison between the most populated conformations suggested by PRE fitting and simulated energetic data. The detailed descriptions of NNLS ensembles of different energy levels are illustrated in Table 3.2. The constructed ensembles consist of many dozens of structures after NNLS reweighting scheme, for example top 5% NNLS ensemble of 58 structures, 20% NNLS ensemble of 108 structures, and 30% NNLS ensemble of 121 structures. Despite of different ensemble sizes, these NNLS ensembles have similar numbers of the major conformations (~20) and major conformations constitute a predominant proportion in the constructed ensembles. A major conformation is one with its reweighted population larger than 1% in the constructed NNLS ensemble. However, these NNLS ensembles without size regularization show a highly arbitrary distribution of domain-domain orientations shown in Figure 3.10. This ambiguous interpretation of experimental data causes a spurious representation of PubRRM1 conformations. Interestingly, the major conformations from different NNLS ensembles show a highly similar preference on the domain-domain orientation that is located in the energy minimum (Figure 3.10). This is not a surprise because the rough orientation of the two domains is governed by the experimental distance constraints. NNLS fitting without size regularization recovers the major conformational space but introduces too much noise of minor states to interpret the protein conformations.

Table 3.2 Profiles of the reweighted NNLS ensembles. Columns correspond to low-energy ensembles, Q-factors, the number of non-zero conformations in the reweighed ensemble, the number of major conformations in the reweighed ensemble, and the cumulative population occupied by the major conformations.

Energy cut-off	Q-factor	Number of non-zero structures	Number of major structures	Total population of major structures
1%	0.67	24	11	0.97
5%	0.46	58	16	0.91
10%	0.36	81	21	0.86
20%	0.30	108	21	0.82
30%	0.22	121	24	0.76
40%	0.19	132	20	0.80
50%	0.16	155	19	0.79
100%	0.12	170	16	0.80

Figure 3.10 Domain-domain orientations of the constructed NNLS ensembles of different energy levels. Domain-domain orientations of ensemble components from reweighted 5%, 20%, 30% low-energy and 100% NNLS ensembles are shown as blue points for major conformations and green points for minor conformations. The red area indicates low-energy surface on the energy landscape with a cutoff at top 5% low energy.

3.8 Ensemble construction with a topology-based pool

Pool structure generation and selection are a common strategy in ensemble determination with experimental data, and the candidate pool comes from computational methods based on either molecular interactions [65, 66] or molecular topology [67, 68, 73]. To investigate the potential effect of the pre-generated structure pool on ensemble construction of PubRRM12, we also employed a different source of candidate structures to fit experimental PRE profiles. Unlike the 4003 pool conformations above produced on the basis of atomistic interactions, RANCH [21, 73] with a topology-based sampling method for multi-domain proteins was adopted to generate a pool of 5000 structures, followed by adding MTSL

molecules to PubRRM12 structures using XPLORE-NIH [145, 146]. The 5000-structure pool size is comparable with our simulated pool. The same MQP optimization scheme above was applied to the topology-based pool. Q-factor evolved with ensemble size up to 20 structures to examine whether the use of a topology-based pool is able to recover experimental observables by a small-sized ensemble. The trend of Q-factor evolution along with structure number has been found as converged at 14 structure of a Q-factor 0.39, equal to its 20-structure counterpart (Figure 3.11). It is noted that this Q-factor value is worse than the 100% simulated ensemble of 0.27 and even the 30% low-energy simulated ensemble of 0.35 whose candidate pool is much smaller. The reason is our simulated pool employed molecular simulations to preclude bad-energy structures from pool structures and resulted in a more complete sampling in the important conformational space, such as the energy-favored orientations of PubRRM12.

Figure 3.11 PRE Q-factor evolution with ensemble size for the ensemble construction with a topology-based structure pool sized 5000.

The 14-structure ensemble recovered from the topology-based pool was taken to investigate their conformational space. This topology-based ensemble produces an ensemble average in agreement with experimental PRE measurements, but its most component structures are found being located at the edge of or adjacent to the orientation area of interest (Figure 3.12). Compared with our low-energy simulated ensembles, the topology-based ensemble has a remoter and more dispersed distribution of domain-domain orientations in the experiment-suggested area. Besides, two structures in the topology-based ensemble have different domain contacting interfaces (Figure 3.12) deviating from the RRM1 beta-sheet, which undermines its interpretation of the conformation space of PubRRM12.

Figure 3.12 The structure ensemble recovered by MQP from a topology-based pool. Domain-domain orientation distribution (blue points) of 14-structure ensemble recovered from the topology-based pool, two of whose component structures are located away from the low-energy orientation (light red) corresponding to RRM1 beta-sheets.

3.9 The role of linker

Linkers connecting domains usually serve as an important role in protein-protein interaction interface [53, 54, 147, 148], thus we treated them as flexible structures using force field interactions instead of the structure-based potential. One role of linkers is to preclude unphysical conformations in the generated structure pool due to the length constraint of the linker. The empty space in Figure 3.13a depicts the

unachievable domain-domain orientations where the two linker-attached residues are in far-end position. Apart from domain-domain orientation, linker also restricts domain self-rotation (Figure 3.14b). Most studies investigate conformational ensemble by generating a pool of protein conformations followed structure ensemble optimization. Pool generation should carefully treat the flexible linker considering its significant effect on sampling. To further investigate the linker contribution in domain-domain interactions, the decomposition of inter-domain potential energy is examined (Figure 3.15). The linker interacts with both domains at the same strength level as the direct interactions between RRM1 and RRM2, implying its important energetic role in domain-domain associations. Linkers are considered as a “scaffold” to precluded unfavorable inter-domain interactions [148]. Therefore, a multi-domain protein might not be simply treated as two proteins by ignoring the contribution from the conformational arrangement of the linker.

Figure 3.13 Conformations (blue points) whose linker (red part in PubRRM12 structure) cannot connect two domains caused by the linker length limit. Structure (a) is the situation that certain domain-domain orientations are unachievable when the two linker-attached residues are in far-end position. For structure (b), domain self-rotation is restricted by the linker despite of the availability of the domain-domain orientation.

Figure 3.14 Potential energy decomposition of representative structures. Potential energy of representative structures is decomposed into RRM1-RRM1, RRM1-linker and RRM2-linker components (kcal/mol).

3.10 Conclusion

In this work, we have developed a theoretical pipeline to study multi-domain proteins via hybrid structure-based/physics-based atomistic model and pull-push plus high temperature sampling methods. Firstly, we highlight that our hybrid model is a physically meaningful treatment. It employs a structure-based potential to stabilize the experimentally identified rigid structures and meanwhile ensures the conformational flexibility involving protein-protein interactions. An atomistic force field characterizes inter-domain interactions, which permits the more accurate

structure analysis with energy calculations using MMPBSA on the atomistic level. Moreover, our MD-based sampling scheme not only precludes bad-energy conformations in contrast with the sampling techniques without molecular interactions, but also conducts a more complete search in the important conformational space.

One wide energy minimum was detected on the energy landscape of PubRRM12 that shows a high compatibility with experimental PRE observables and chemical shift perturbations, indicating RRM1 β -sheet face dominates the interaction interface. Our regularization idea features favorable structures in physics, in a combined use of L0-norm constraint to avoid the over-interpretation of experimental data. Our simulation approach can in principle be applied to any multi-domain proteins and multi-protein complexes with dispersed binding states. It would be of particular interest to model a wider variety of experimental measurements, such as data from combined NMR and SAXS studies.

4 Discussion

4.1 Structural information from experimental constraints

Our ensemble constructions were performed to simulate solution experimental data of SAXS-based NS5 and PRE-based PubRRM12. The experimental observable from SAXS experiments is a one-dimensional scattering curve, where the structural information is much degenerated. As shown in Figure 2.7, the ensembles recovered from SAXS data show a highly noisy conformational space, which suggests the structural “signal” from the SAXS data is not sufficient enough for deriving a correct structural representation. In comparison, NMR offers inter-atomic distances and its high-resolution constraints identify the major protein conformations of PubRRM12, as shown in Figure 3.9. However, SAXS experiments provide the advantage of handling large biomolecules without size limitation and offer molecular shape information. In the PRE-based ensemble construction of PubRRM12, SAXS profile was also utilized to validate the PRE-based ensemble. Recently, the integration of different structural information has become a trend in this field. Many works [149-155] have been carried out with a combination of NMR and SAXS experiments, because the atomistic distance/angle constraints from NMR and the shape information from SAXS are complementary structural information for a more comprehensive structural characterization.

4.2 Over-fitting challenge

The ensemble determination from given experimental measurements might be an ill-posed problem, as there exist multiple solutions that agree equally well with the data. Especially, the data is a measured quantity averaging over the heterogeneous conformations with the errors. Over-fitting might result in unphysical structures, wrong conformational representatives and the loss of structural heterogeneity [156]. The way to regularizing this over-fitting problem is to add constraints to drive the solution to some correct ensemble properties.

As we discussed in Chapter 4.1, independent experimental methods offer the constraints from different structural properties, which drives the constructed ensemble close to the “true” conformational space. The constraints from experiments are expected to be 1) sensitive, which is a small change in the structural quantities would lead to detectable change in the experimental measurements, and 2) unique, which is various conformations have distinguishable structural quantities [67]. Besides, the random and systematic errors in the observed experimental data may cause the ambiguity of the conformational interpretations, thus the incorporation with additional sources of experimental data would assist to strengthen the structural signal from the correct conformations [157].

Ensemble size is a common regularization term. We demonstrated a raw numerical fitting resulted in dozens of structures in the ensemble of PubRRM12 and numerous

random conformations (Figure 3.10). It is noted that a small-sized ensemble is not always optimal, because some systems are highly flexible, in which cases an ensemble representation of several structures failed to sufficiently describe the molecular conformations and inherent heterogeneity. The ensemble sparsity depends on the true dynamics nature of the molecule's flexibility. As a result, the ensemble size should be investigated with the flexibility of the ensemble sparsity, instead of a pre-defined fixed size. For example, the quality of fit can be investigated as a function or evolution with ensemble size to determine a stable and converged solution.

Apart from the constraints from experiments and ensemble size above, we proposed to use the physics-based constraints in ensemble construction. Previous works recovered structure ensembles using MD/MC simulated structures [65, 66] or the docked structures with high-rank scores [68], and we systematically investigated the effect of structure energy on ensemble optimization. It has been demonstrated in both cases of DENV2 NS5 and PubRRM12 that the energy regularization is able to prevent the ensemble construction from spurious conformational representations caused by the noise from high-energy structures. The theoretical mechanics model of the biophysical interactions is an important complementary to other types of regularizations and various optimization algorithms. Our optimization scheme employed structure energy to exclude high-energy and unphysical structures from the candidate pool, which requires a sufficient accuracy of the force field. If the force

field is not sufficiently accurate, large errors would occur in the simulated molecular properties [158-160]. Therefore, one goal in this field is the improvement of force fields in molecular modeling [161].

Our computational strategy on ensemble construction with experimental data emphasizes the physical interpretation of conformational ensembles, which could be a complementary with other constraints like L0-norm regularization. To further optimize the current computational strategy, one potential direction is to combine all the constraints with the data fitting quality in one formula that would serve as the target for minimization. Therefore, the optimization target becomes the formulated term incorporating fitting quality and other constraints like ensemble size and conformational energy, instead of the quality of experimental data fitting alone. Especially, more and more studies [149-155] have characterized biomolecular structures with a combined use of different experiments, thus extra experimental data could be also added as constraint terms for ensemble optimization.

5 Concluding remarks

The structure ensemble of a flexible biomolecule cannot be derived from noisy experimental data alone, and ever-improving computational power is a strong complement. We developed a computational pipeline to simulate the structure ensemble of multi-domain proteins by using hybrid models that employ structure-based potential for intra-domain structures and physics-based potential for inter-domain interactions. This hybrid model is able to be applied to PubRRM12 and NS5 on the atomistic and coarse-grained levels respectively, depending on the various resolutions of the structural information from PRE and SAXS experiments. Besides, our model highlights the importance of the flexibility from the inter-domain linker. The linker of NS5 mediates the conformational arrangement of the complex, and the case of PubRRM12 also demonstrates that the contribution to the domain-domain associations from its linker is not negligible. It is suggested that the structure modeling should consider the full flexibility of the unstructured regions when applying to multi-domain proteins and other complexes with their well-structural units connected via intrinsically disordered regions.

For the critical ensemble-recovery problem, energy is proposed as a physics-based constraint that can be taken with different optimization methods or as a complementary to other regularization terms like ensemble size to minimize the problem of data over-fitting. The absence of energy constraint would cause a raw

numerical fitting and expose the structure ensemble to the high-energy noise. Fitting without energy would cause a spurious conformation representation and undermine the reliability of the constructed ensemble. Given the recent trend to construct ensemble representations of protein structures by using mixed experimental and simulation methods, our results shed light on the physical interpretation of experimentally derived conformation ensembles.

References

1. McCammon, J.A., Gelin, B.R., and Karplus, M., *Dynamics of folded proteins*. Nature, 1977. **267**(5612): p. 585-90.
2. Cornell, W.D., Cieplak, P., Bayly, C.I., Gould, I.R., et al., *A second generation force field for the simulation of proteins, nucleic acids, and organic molecules*. J Am Chem Soc, 1995. **117**(19): p. 5179-5197.
3. Brooks, B.R., Bruccoleri, R.E., Olafson, B.D., States, D.J., et al., *Charmm - a Program for Macromolecular Energy, Minimization, and Dynamics Calculations*. J Comput Chem, 1983. **4**(2): p. 187-217.
4. Scott, W.R.P., Hunenberger, P.H., Tironi, I.G., Mark, A.E., et al., *The GROMOS biomolecular simulation program package*. J Phys Chem A, 1999. **103**(19): p. 3596-3607.
5. Jorgensen, W.L., Maxwell, D.S., and TiradoRives, J., *Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids*. J Am Chem Soc, 1996. **118**(45): p. 11225-11236.
6. Dror, R.O., Dirks, R.M., Grossman, J.P., Xu, H., et al., *Biomolecular simulation: a computational microscope for molecular biology*. Annu Rev Biophys, 2012. **41**: p. 429-52.
7. Tozzini, V., *Coarse-grained models for proteins*. Curr Opin Struct Biol, 2005. **15**(2): p. 144-50.
8. Riniker, S., Allison, J.R., and van Gunsteren, W.F., *On developing coarse-grained models for biomolecular simulation: a review*. Phys Chem Chem Phys, 2012. **14**(36): p. 12423-30.

9. Karplus, M. and McCammon, J.A., *Molecular dynamics simulations of biomolecules*. Nat Struct Mol Biol, 2002. **9**(9): p. 646-652.
10. Keller, B., Christen, M., Oostenbrink, C., and van Gunsteren, W.F., *On using oscillating time-dependent restraints in MD simulation*. J Biomol NMR, 2007. **37**(1): p. 1-14.
11. van Gunsteren, W.F., Dolenc, J., and Mark, A.E., *Molecular simulation as an aid to experimentalists*. Curr Opin Struct Biol, 2008. **18**(2): p. 149-153.
12. Bouvignies, G., Vallurupalli, P., Hansen, D.F., Correia, B.E., et al., *Solution structure of a minor and transiently formed state of a T4 lysozyme mutant*. Nature, 2011. **477**(7362): p. 111-4.
13. Trzesniak, D., Glatli, A., Jaun, B., and van Gunsteren, W.F., *Interpreting NMR data for beta-peptides using molecular dynamics simulations*. J Am Chem Soc, 2005. **127**(41): p. 14320-9.
14. Lau, A.Y. and Roux, B., *The free energy landscapes governing conformational changes in a glutamate receptor ligand-binding domain*. Structure, 2007. **15**(10): p. 1203-14.
15. van Gunsteren, W.F., Bakowies, D., Baron, R., Chandrasekhar, I., et al., *Biomolecular modeling: Goals, problems, perspectives*. Angew Chem Int Ed Engl, 2006. **45**(25): p. 4064-92.
16. Xu, G., Lo, Y.-C., Li, Q., Napolitano, G., et al., *Crystal structure of inhibitor of [kgr] B kinase [bgr]*. Nature, 2011. **472**(7343): p. 325-330.
17. de Martin, R., Schmid, J.A., and Hofer-Warbinek, R., *The NF- κ B/Rel family of transcription factors in oncogenic transformation and apoptosis*. Mutat Res, 1999. **437**(3): p. 231-243.

18. Hurley, J.H. and Hanson, P.I., *Membrane budding and scission by the ESCRT machinery: it's all in the neck*. Nat Rev Mol Cell Biol, 2010. **11**(8): p. 556-566.
19. Goh, C.S., Milburn, D., and Gerstein, M., *Conformational changes associated with protein-protein interactions*. Curr Opin Struct Biol, 2004. **14**(1): p. 104-9.
20. Koshland, D.E., *Application of a Theory of Enzyme Specificity to Protein Synthesis*. Proc Natl Acad Sci U S A, 1958. **44**(2): p. 98-104.
21. Petoukhov, M.V. and Svergun, D.I., *Global rigid body modeling of macromolecular complexes against small-angle scattering data*. Biophys J, 2005. **89**(2): p. 1237-50.
22. Fisher, C.K. and Stultz, C.M., *Constructing ensembles for intrinsically disordered proteins*. Curr Opin Struct Biol, 2011. **21**(3): p. 426-31.
23. Klepeis, J.L., Lindorff-Larsen, K., Dror, R.O., and Shaw, D.E., *Long-timescale molecular dynamics simulations of protein structure and function*. Curr Opin Struct Biol, 2009. **19**(2): p. 120-7.
24. Fersht, A.R., *On the simulation of protein folding by short time scale molecular dynamics and distributed computing*. Proc Natl Acad Sci U S A, 2002. **99**(22): p. 14122-5.
25. Pande, V.S., Baker, I., Chapman, J., Elmer, S.P., et al., *Atomistic protein folding simulations on the submillisecond time scale using worldwide distributed computing*. Biopolymers, 2003. **68**(1): p. 91-109.
26. Lindorff-Larsen, K., Piana, S., Dror, R.O., and Shaw, D.E., *How Fast-Folding Proteins Fold*. Science, 2011. **334**(6055): p. 517-520.
27. Klein, M.L. and Shinoda, W., *Large-scale molecular dynamics simulations of self-assembling systems*. Science, 2008. **321**(5890): p. 798-800.

28. Freddolino, P.L., Harrison, C.B., Liu, Y., and Schulten, K., *Challenges in protein folding simulations: Timescale, representation, and analysis*. Nat Phys, 2010. **6**(10): p. 751-758.
29. Maragakis, P., Lindorff-Larsen, K., Eastwood, M.P., Dror, R.O., et al., *Microsecond molecular dynamics simulation shows effect of slow loop dynamics on backbone amide order parameters of proteins*. J Phys Chem B, 2008. **112**(19): p. 6155-6158.
30. Bernardi, R.C., Melo, M.C., and Schulten, K., *Enhanced sampling techniques in molecular dynamics simulations of biological systems*. Biochim Biophys Acta, 2015. **1850**(5): p. 872-7.
31. Jones, M.R., Liu, C., and Wilson, A.K., *Molecular Dynamics Studies of the Protein-Protein Interactions in Inhibitor of κ B Kinase- β* . J Chem Inf Model, 2014. **54**(2): p. 562-572.
32. Laio, A. and Parrinello, M., *Escaping free-energy minima*. Proc Natl Acad Sci U S A, 2002. **99**(20): p. 12562-6.
33. Sugita, Y. and Okamoto, Y., *Replica-exchange molecular dynamics method for protein folding*. Chem Phys Lett, 1999. **314**(1-2): p. 141-151.
34. Lo Conte, L., Chothia, C., and Janin, J., *The atomic structure of protein-protein recognition sites*. J Mol Biol, 1999. **285**(5): p. 2177-98.
35. Clementi, C., Nymeyer, H., and Onuchic, J.N., *Topological and energetic factors: what determines the structural details of the transition state ensemble and "en-route" intermediates for protein folding? An investigation for small globular proteins*. J Mol Biol, 2000. **298**(5): p. 937-953.
36. Whitford, P.C., Noel, J.K., Gosavi, S., Schug, A., et al., *An all-atom structure-based potential for proteins: bridging minimal models with all-atom empirical forcefields*. Proteins, 2009. **75**(2): p. 430-41.

37. Onuchic, J.N. and Wolynes, P.G., *Theory of protein folding*. Curr Opin Struct Biol, 2004. **14**(1): p. 70-5.
38. Bryngelson, J.D., Onuchic, J.N., Socci, N.D., and Wolynes, P.G., *Funnels, pathways, and the energy landscape of protein folding: a synthesis*. Proteins, 1995. **21**(3): p. 167-95.
39. Sutto, L., Mereu, I., and Gervasio, F.L., *A Hybrid All-Atom Structure-Based Model for Protein Folding and Large Scale Conformational Transitions*. J Chem Theory Comput, 2011. **7**(12): p. 4208-17.
40. Wodak, S.J. and Janin, J., *Computer analysis of protein-protein interaction*. J Mol Biol, 1978. **124**(2): p. 323-342.
41. Warwicker, J., *Investigating Protein Protein-Interaction Surfaces Using a Reduced Stereochemical and Electrostatic Model*. J Mol Biol, 1989. **206**(2): p. 381-395.
42. Kim, Y.C. and Hummer, G., *Coarse-grained models for simulations of multiprotein complexes: application to ubiquitin binding*. J Mol Biol, 2008. **375**(5): p. 1416-33.
43. Solernou, A. and Fernandez-Recio, J., *pyDockCG: New Coarse-Grained Potential for Protein Protein Docking*. J Phys Chem B, 2011. **115**(19): p. 6032-6039.
44. Vangone, A., Spinelli, R., Scarano, V., Cavallo, L., et al., *COCOMAPS: a web application to analyze and visualize contacts at the interface of biomolecular complexes*. Bioinformatics, 2011. **27**(20): p. 2915-2916.
45. Chen, R., Li, L., and Weng, Z., *ZDOCK: an initial-stage protein-docking algorithm*. Proteins, 2003. **52**(1): p. 80-7.

46. Comeau, S.R., Gatchell, D.W., Vajda, S., and Camacho, C.J., *ClusPro: an automated docking and discrimination method for the prediction of protein complexes*. Bioinformatics, 2004. **20**(1): p. 45-50.
47. Dominguez, C., Boelens, R., and Bonvin, A.M., *HADDOCK: a protein-protein docking approach based on biochemical or biophysical information*. J Am Chem Soc, 2003. **125**(7): p. 1731-7.
48. Mashiach, E., Schneidman-Duhovny, D., Andrusier, N., Nussinov, R., et al., *FireDock: a web server for fast interaction refinement in molecular docking*. Nucleic Acids Res, 2008. **36**(Web Server issue): p. W229-32.
49. Kozakov, D., Brenke, R., Comeau, S.R., and Vajda, S., *PIPER: an FFT-based protein docking program with pairwise potentials*. Proteins, 2006. **65**(2): p. 392-406.
50. Torchala, M., Moal, I.H., Chaleil, R.A., Fernandez-Recio, J., et al., *SwarmDock: a server for flexible protein-protein docking*. Bioinformatics, 2013. **29**(6): p. 807-9.
51. Lyskov, S. and Gray, J.J., *The RosettaDock server for local protein-protein docking*. Nucleic Acids Res, 2008. **36**(Web Server issue): p. W233-8.
52. Klett, J., Nunez-Salgado, A., Dos Santos, H.G., Cortes-Cabrera, A., et al., *MM-ISMSA: An Ultrafast and Accurate Scoring Function for Protein-Protein Docking*. J Chem Theory Comput, 2012. **8**(9): p. 3395-408.
53. Ritchie, D.W., *Recent progress and future directions in protein-protein docking*. Curr Protein Pept Sci, 2008. **9**(1): p. 1-15.
54. Swain, J.F., Dinler, G., Sivendran, R., Montgomery, D.L., et al., *Hsp70 chaperone ligands control domain association via an allosteric mechanism mediated by the interdomain linker*. Mol Cell, 2007. **26**(1): p. 27-39.

55. Takada, S., *Coarse-grained molecular simulations of large biomolecules*. Curr Opin Struct Biol, 2012. **22**(2): p. 130-7.
56. Izvekov, S. and Voth, G.A., *A multiscale coarse-graining method for biomolecular systems*. J Phys Chem B, 2005. **109**(7): p. 2469-73.
57. Lu, L. and Voth, G.A., *The Multiscale Coarse-Graining Method*. Adv Chem Phys, 2012. **149**: p. 47-81.
58. Dama, J.F., Sinitskiy, A.V., McCullagh, M., Weare, J., et al., *The Theory of Ultra-Coarse-Graining. 1. General Principles*. J Chem Theory Comput, 2013. **9**(5): p. 2466-80.
59. Kmiecik, S., Gront, D., Kolinski, M., Wieteska, L., et al., *Coarse-Grained Protein Models and Their Applications*. Chem Rev, 2016. **116**(14): p. 7898-7936.
60. Miyazawa, S. and Jernigan, R.L., *Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading*. J Mol Biol, 1996. **256**(3): p. 623-44.
61. Miyazawa, S. and Jernigan, R.L., *Estimation of Effective Interresidue Contact Energies from Protein Crystal-Structures - Quasi-Chemical Approximation*. Macromolecules, 1985. **18**(3): p. 534-552.
62. Ravikumar, K.M., Huang, W., and Yang, S., *Coarse-grained simulations of protein-protein association: an energy landscape perspective*. Biophys J, 2012. **103**(4): p. 837-45.
63. Terakawa, T., Higo, J., and Takada, S., *Multi-scale ensemble modeling of modular proteins with intrinsically disordered linker regions: application to p53*. Biophys J, 2014. **107**(3): p. 721-9.
64. Zheng, W.J. and Tekpinar, M., *Accurate Flexible Fitting of High-Resolution Protein Structures to Small-Angle X-Ray Scattering Data Using a Coarse-*

- Grained Model with Implicit Hydration Shell*. Biophys J, 2011. **101**(12): p. 2981-2991.
65. Kim, Y.C., Tang, C., Clore, G.M., and Hummer, G., *Replica exchange simulations of transient encounter complexes in protein-protein association*. Proc Natl Acad Sci U S A, 2008. **105**(35): p. 12855-60.
 66. Rozycki, B., Kim, Y.C., and Hummer, G., *SAXS Ensemble Refinement of ESCRT-III CHMP3 Conformational Transitions*. Structure, 2011. **19**(1): p. 109-116.
 67. Berlin, K., Castaneda, C.A., Schneidman-Duhovny, D., Sali, A., et al., *Recovering a representative conformational ensemble from underdetermined macromolecular structural data*. J Am Chem Soc, 2013. **135**(44): p. 16595-609.
 68. Kozakov, D., Li, K., Hall, D.R., Beglov, D., et al., *Encounter complexes and dimensionality reduction in protein-protein association*. Elife, 2014. **3**: p. e01370.
 69. Clore, G.M., *Generating accurate contact maps of transient long-range interactions in intrinsically disordered proteins by paramagnetic relaxation enhancement*. Biophys J, 2013. **104**(8): p. 1635-6.
 70. Pelikan, M., Hura, G.L., and Hammel, M., *Structure and flexibility within proteins as identified through small angle X-ray scattering*. Gen Physiol Biophys, 2009. **28**(2): p. 174-189.
 71. Chen, Y., Campbell, S.L., and Dokholyan, N.V., *Deciphering protein dynamics from NMR data using explicit structure sampling and selection*. Biophys J, 2007. **93**(7): p. 2300-2306.
 72. Huang, J.-r. and Grzesiek, S., *Ensemble calculations of unstructured proteins constrained by RDC and PRE data: a case study of urea-denatured ubiquitin*. J Am Chem Soc, 2009. **132**(2): p. 694-705.

73. Bernado, P., Mylonas, E., Petoukhov, M.V., Blackledge, M., et al., *Structural characterization of flexible proteins using small-angle X-ray scattering*. J Am Chem Soc, 2007. **129**(17): p. 5656-5664.
74. Jaynes, E.T., *Information theory and statistical mechanics*. Phys Rev, 1957. **106**(4): p. 620.
75. Choy, W.Y. and Forman-Kay, J.D., *Calculation of ensembles of structures representing the unfolded state of an SH3 domain*. J Mol Biol, 2001. **308**(5): p. 1011-32.
76. Guzman, M.G., Halstead, S.B., Artsob, H., Buchy, P., et al., *Dengue: a continuing global threat*. Nat Rev Microbiol, 2010. **8**(12 Suppl): p. S7-16.
77. Azhar, E.I., Hashem, A.M., El-Kafrawy, S.A., Abol-Ela, S., et al., *Complete genome sequencing and phylogenetic analysis of dengue type 1 virus isolated from Jeddah, Saudi Arabia*. Virol J, 2015. **12**(1): p. 1.
78. Mackenzie, J., *Wrapping things up about virus RNA replication*. Traffic, 2005. **6**(11): p. 967-977.
79. Welsch, S., Miller, S., Romero-Brey, I., Merz, A., et al., *Composition and three-dimensional architecture of the dengue virus replication and assembly sites*. Cell Host Microbe, 2009. **5**(4): p. 365-75.
80. Fernandez-Garcia, M.D., Meertens, L., Bonazzi, M., Cossart, P., et al., *Appraising the roles of CBLL1 and the ubiquitin/proteasome system for flavivirus entry and replication*. J Virol, 2011. **85**(6): p. 2980-9.
81. Bollati, M., Alvarez, K., Assenberg, R., Baronti, C., et al., *Structure and functionality in flavivirus NS-proteins: perspectives for drug design*. Antiviral Res, 2010. **87**(2): p. 125-48.
82. Egloff, M.P., Benarroch, D., Selisko, B., Romette, J.L., et al., *An RNA cap (nucleoside-2'-O-)-methyltransferase in the flavivirus RNA polymerase NS5*:

- crystal structure and functional characterization*. EMBO J, 2002. **21**(11): p. 2757-68.
83. Issur, M., Geiss, B.J., Bougie, I., Picard-Jean, F., et al., *The flavivirus NS5 protein is a true RNA guanylyltransferase that catalyzes a two-step reaction to form the RNA cap structure*. RNA, 2009. **15**(12): p. 2340-2350.
 84. Lim, S.P., Wang, Q.Y., Noble, C.G., Chen, Y.L., et al., *Ten years of dengue drug discovery: progress and prospects*. Antiviral Res, 2013. **100**(2): p. 500-19.
 85. Lu, G. and Gong, P., *Crystal Structure of the full-length Japanese encephalitis virus NS5 reveals a conserved methyltransferase-polymerase interface*. PLoS Pathog, 2013. **9**(8): p. e1003549.
 86. Zhao, Y., Soh, T.S., Zheng, J., Chan, K.W., et al., *A crystal structure of the Dengue virus NS5 protein reveals a novel inter-domain interface essential for protein flexibility and virus replication*. PLoS Pathog, 2015. **11**(3): p. e1004682.
 87. Bussetta, C. and Choi, K.H., *Dengue virus nonstructural protein 5 adopts multiple conformations in solution*. Biochemistry, 2012. **51**(30): p. 5921-31.
 88. Saw, W.G., Tria, G., Gruber, A., Manimekalai, M.S.S., et al., *Structural insight and flexible features of NS5 proteins from all four serotypes of Dengue virus in solution*. Acta Crystallogr Sect D Biol Crystallogr, 2015. **71**(11): p. 2309-2327.
 89. Manimekalai, M.S.S., Saw, W.G., Pan, A., Gruber, A., et al., *Identification of the critical linker residues conferring differences in the compactness of NS5 from Dengue virus serotype 4 and NS5 from Dengue virus serotypes 1-3*. Acta Crystallogr Sect D Biol Crystallogr, 2016. **72**(6): p. 795-807.

90. Yang, S., Blachowicz, L., Makowski, L., and Roux, B., *Multidomain assembled states of Hck tyrosine kinase in solution*. Proc Natl Acad Sci U S A, 2010. **107**(36): p. 15757-62.
91. Shea, J.E., Onuchic, J.N., and Brooks, C.L., 3rd, *Exploring the origins of topological frustration: design of a minimally frustrated model of fragment B of protein A*. Proc Natl Acad Sci U S A, 1999. **96**(22): p. 12512-7.
92. Koga, N. and Takada, S., *Roles of native topology and chain-length scaling in protein folding: A simulation study with a Go-like model*. J Mol Biol, 2001. **313**(1): p. 171-180.
93. Noel, J.K., Whitford, P.C., and Onuchic, J.N., *The shadow map: a general contact definition for capturing the dynamics of biomolecular folding and function*. J Phys Chem B, 2012. **116**(29): p. 8692-702.
94. Webb, B. and Sali, A., *Comparative Protein Structure Modeling Using MODELLER*. Curr Protoc Protein Sci: p. 2.9. 1-2.9. 37.
95. Huang, W., Ravikumar, K.M., and Yang, S., *A Newfound Cancer-Activating Mutation Reshapes the Energy Landscape of Estrogen-Binding Domain*. J Chem Theory Comput, 2014. **10**(8): p. 2897-900.
96. Daura, X., Gademann, K., Jaun, B., Seebach, D., et al., *Peptide folding: When simulation meets experiment*. Angew Chem Int Ed, 1999. **38**(1-2): p. 236-240.
97. Hess, B., Kutzner, C., van der Spoel, D., and Lindahl, E., *GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation*. J Chem Theory Comput, 2008. **4**(3): p. 435-47.
98. Svergun, D., Barberato, C., and Koch, M.H.J., *CRY SOL - A program to evaluate x-ray solution scattering of biological macromolecules from atomic coordinates*. J Appl Crystallogr, 1995. **28**(6): p. 768-773.

99. Rotkiewicz, P. and Skolnick, J., *Fast procedure for reconstruction of full-atom protein models from reduced representations*. J Comput Chem, 2008. **29**(9): p. 1460-5.
100. Press, W.H., Teukolsky, S.A., Vetterling, W.T., and Flannery, B.P., *Numerical recipes: the art of scientific computing*. 2007, Cambridge University Press, New York.
101. Tria, G., Mertens, H.D.T., Kachala, M., and Svergun, D.I., *Advanced ensemble modelling of flexible macromolecules using X-ray solution scattering*. IUCrJ, 2015. **2**(2): p. 207-217.
102. Rader, A.J., *Coarse-grained models: getting more with less*. Curr Opin Pharmacol, 2010. **10**(6): p. 753-9.
103. Noid, W.G., *Perspective: Coarse-grained models for biomolecular systems*. J Chem Phys, 2013. **139**(9): p. 090901.
104. Hornak, V., Abel, R., Okur, A., Strockbine, B., et al., *Comparison of multiple Amber force fields and development of improved protein backbone parameters*. Proteins, 2006. **65**(3): p. 712-25.
105. Konarev, P.V. and Svergun, D.I., *A posteriori determination of the useful data range for small-angle scattering experiments on dilute monodisperse systems*. IUCrJ, 2015. **2**(Pt 3): p. 352-60.
106. Metskas, L.A. and Rhoades, E., *Conformation and Dynamics of the Troponin I C-Terminal Domain: Combining Single-Molecule and Computational Approaches for a Disordered Protein Region*. J Am Chem Soc, 2015. **137**(37): p. 11962-9.
107. Dunker, A.K., Brown, C.J., Lawson, J.D., Iakoucheva, L.M., et al., *Intrinsic disorder and protein function*. Biochemistry, 2002. **41**(21): p. 6573-82.

108. Solomon, I., *Relaxation processes in a system of two spins*. Phys Rev, 1955. **99**(2): p. 559.
109. Bloembergen, N. and Morgan, L.O., *Proton Relaxation Times in Paramagnetic Solutions Effects of Electron Spin Relaxation*. J Chem Phys, 1961. **34**(3): p. 842-&.
110. Tang, C., Ghirlando, R., and Clore, G.M., *Visualization of transient ultra-weak protein self-association in solution using paramagnetic relaxation enhancement*. J Am Chem Soc, 2008. **130**(12): p. 4048-4056.
111. Suh, J.Y., Tang, C., and Clore, G.M., *Role of electrostatic interactions in transient encounter complexes in protein-protein association investigated by paramagnetic relaxation enhancement*. J Am Chem Soc, 2007. **129**(43): p. 12954-5.
112. Tang, C., Schwieters, C.D., and Clore, G.M., *Open-to-closed transition in apo maltose-binding protein observed by paramagnetic NMR*. Nature, 2007. **449**(7165): p. 1078-82.
113. Volkov, A.N., Worrall, J.A.R., Holtzmann, E., and Ubbink, M., *Solution structure and dynamics of the complex between cytochrome c and cytochrome c peroxidase determined by paramagnetic NMR*. Proc Natl Acad Sci USA, 2006. **103**(50): p. 18945-18950.
114. Tang, C., Iwahara, J., and Clore, G.M., *Visualization of transient encounter complexes in protein-protein association*. Nature, 2006. **444**(7117): p. 383-6.
115. Iwahara, J. and Clore, G.M., *Detecting transient intermediates in macromolecular binding by paramagnetic NMR*. Nature, 2006. **440**(7088): p. 1227-30.

116. Iwahara, J., Zweckstetter, M., and Clore, G.M., *NMR structural and kinetic characterization of a homeodomain diffusing and hopping on nonspecific DNA*. Proc Natl Acad Sci U S A, 2006. **103**(41): p. 15062-7.
117. Iwahara, J., Schwieters, C.D., and Clore, G.M., *Ensemble approach for NMR structure refinement against ¹H paramagnetic relaxation enhancement data arising from a flexible paramagnetic group attached to a macromolecule*. J Am Chem Soc, 2004. **126**(18): p. 5879-5896.
118. Anderson, J.T., Paddy, M.R., and Swanson, M.S., *Pub1 Is a Major Nuclear and Cytoplasmic Polyadenylated Rna-Binding Protein in Saccharomyces-Cerevisiae*. Mol Cell Biol, 1993. **13**(10): p. 6102-6113.
119. Matunis, M.J., Matunis, E.L., and Dreyfuss, G., *PUB1: a major yeast poly(A)+ RNA-binding protein*. Mol Cell Biol, 1993. **13**(10): p. 6114-23.
120. Li, H., Shi, H., Wang, H., Zhu, Z., et al., *Crystal structure of the two N-terminal RRM domains of Pub1 and the poly(U)-binding properties of Pub1*. J Struct Biol, 2010. **171**(3): p. 291-7.
121. Liu, W., Zhang, J., Fan, J.S., Tria, G., et al., *A New Method for Determining Structure Ensemble: Application to a RNA Binding Di-Domain Protein*. Biophys J, 2016. **110**(9): p. 1943-56.
122. Neri, M., Anselmi, C., Carnevale, V., Vargiu, A.V., et al., *Molecular dynamics simulations of outer-membrane protease T from E-coli based on a hybrid coarse-grained/atomistic potential*. J Phys: Condens Matter, 2006. **18**(14): p. S347-S355.
123. Yang, T.Y., Wu, J.C., Yan, C.L., Wang, Y.F., et al., *Virtual screening using molecular simulations*. Proteins, 2011. **79**(6): p. 1940-1951.
124. Wang, J., Wolf, R.M., Caldwell, J.W., Kollman, P.A., et al., *Development and testing of a general amber force field*. J Comput Chem, 2004. **25**(9): p. 1157-74.

125. Fogolari, F., Zuccato, P., Esposito, G., and Viglino, P., *Biomolecular electrostatics with the linearized Poisson-Boltzmann equation*. Biophys J, 1999. **76**(1 Pt 1): p. 1-16.
126. Fogolari, F., Brigo, A., and Molinari, H., *Protocol for MM/PBSA molecular dynamics simulations of proteins*. Biophys J, 2003. **85**(1): p. 159-66.
127. Meyer, T. and Knapp, E.W., *pKa values in proteins determined by electrostatics applied to molecular dynamics trajectories*. J Chem Theory Comput, 2015. **11**(6): p. 2827-40.
128. Kumari, R., Kumar, R., Open Source Drug Discovery, C., and Lynn, A., *g_mmpbsa--a GROMACS tool for high-throughput MM-PBSA calculations*. J Chem Inf Model, 2014. **54**(7): p. 1951-62.
129. Ponti, A., *Simulation of magnetic resonance static powder lineshapes: A quantitative assessment of spherical codes*. J Magn Reson, 1999. **138**(2): p. 288-97.
130. Pronk, S., Pall, S., Schulz, R., Larsson, P., et al., *GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit*. Bioinformatics, 2013. **29**(7): p. 845-54.
131. Kollman, P.A., Massova, I., Reyes, C., Kuhn, B., et al., *Calculating structures and free energies of complex molecules: combining molecular mechanics and continuum models*. Acc Chem Res, 2000. **33**(12): p. 889-97.
132. Miller III, B.R., McGee Jr, T.D., Swails, J.M., Homeyer, N., et al., *MMPBSA.py: an efficient program for end-state free energy calculations*. J Chem Theory Comput, 2012. **8**(9): p. 3314-3321.
133. Honig, B. and Nicholls, A., *Classical electrostatics in biology and chemistry*. Science, 1995. **268**(5214): p. 1144-9.

134. Warwicker, J. and Watson, H.C., *Calculation of the Electric-Potential in the Active-Site Cleft Due to Alpha-Helix Dipoles*. J Mol Biol, 1982. **157**(4): p. 671-679.
135. Tan, C., Tan, Y.H., and Luo, R., *Implicit nonpolar solvent models*. J Phys Chem B, 2007. **111**(42): p. 12263-74.
136. Frisch, M., Trucks, G., Schlegel, H.B., Scuseria, G., et al., *Gaussian 09, revision D. 01*. 2009, Gaussian, Inc., Wallingford CT.
137. Czerminski, R. and Elber, R., *Computational studies of ligand diffusion in globins: I. Leghemoglobin*. Proteins, 1991. **10**(1): p. 70-80.
138. Levy, Y., Cho, S.S., Onuchic, J.N., and Wolynes, P.G., *A survey of flexible protein binding mechanisms and their transition states using native topology based energy landscapes*. J Mol Biol, 2005. **346**(4): p. 1121-1145.
139. Allain, F.H., Bouvet, P., Dieckmann, T., and Feigon, J., *Molecular basis of sequence-specific recognition of pre-ribosomal RNA by nucleolin*. EMBO J, 2000. **19**(24): p. 6870-81.
140. Allain, F.H., Gilbert, D.E., Bouvet, P., and Feigon, J., *Solution structure of the two N-terminal RNA-binding domains of nucleolin and NMR study of the interaction with its RNA target*. J Mol Biol, 2000. **303**(2): p. 227-41.
141. Crowder, S.M., Kanaar, R., Rio, D.C., and Alber, T., *Absence of interdomain contacts in the crystal structure of the RNA recognition motifs of Sex-lethal*. Proc Natl Acad Sci U S A, 1999. **96**(9): p. 4892-7.
142. Zhu, G., Saw, W.G., Nalaparaju, A., Gruber, G., et al., *Coarse-Grained Molecular Modeling of the Solution Structure Ensemble of Dengue Virus Nonstructural Protein 5 with Small-Angle X-ray Scattering Intensity*. J Phys Chem B, 2017. **121**(10): p. 2252-2264.

143. Francis, D.M., Różycki, B., Koveal, D., Hummer, G., et al., *Structural basis of p38 α regulation by hematopoietic tyrosine phosphatase*. Nat Chem Biol, 2011. **7**(12): p. 916-924.
144. Frank, A.T., Stelzer, A.C., Al-Hashimi, H.M., and Andricioaei, I., *Constructing RNA dynamical ensembles by combining MD and motionally decoupled NMR RDCs: new insights into RNA dynamics and adaptive ligand recognition*. Nucleic Acids Res, 2009. **37**(11): p. 3670-3679.
145. Schwieters, C.D., Kuszewski, J.J., and Clore, G.M., *Using Xplor-NIH for NMR molecular structure determination*. Prog Nucl Magn Reson Spectrosc, 2006. **48**(1): p. 47-62.
146. Schwieters, C.D., Kuszewski, J.J., Tjandra, N., and Clore, G.M., *The Xplor-NIH NMR molecular structure determination package*. J Magn Reson, 2003. **160**(1): p. 65-73.
147. Gokhale, R.S. and Khosla, C., *Role of linkers in communication between protein modules*. Curr Opin Chem Biol, 2000. **4**(1): p. 22-27.
148. George, R.A. and Heringa, J., *An analysis of protein domain linkers: their classification and role in protein folding*. Protein Eng, 2002. **15**(11): p. 871-879.
149. Gorantla, N.V., Shkumatov, A.V., and Chinnathambi, S., *Conformational Dynamics of Intracellular Tau Protein Revealed by CD and SAXS*. Methods Mol Biol, 2017. **1523**: p. 3-20.
150. Deshmukh, L., Schwieters, C.D., Grishaev, A., and Clore, G.M., *Quantitative Characterization of Configurational Space Sampled by HIV-1 Nucleocapsid Using Solution NMR, X-ray Scattering and Protein Engineering*. ChemPhysChem, 2016. **17**(11): p. 1548-1552.

151. Mertens, H.D.T. and Svergun, D.I., *Combining NMR and small angle X-ray scattering for the study of biomolecular structure and dynamics*. Arch Biochem Biophys, 2017.
152. Sterckx, Y.G., Volkov, A.N., Vranken, W.F., Kragelj, J., et al., *Small-angle X-ray scattering- and nuclear magnetic resonance-derived conformational ensemble of the highly flexible antitoxin PaaA2*. Structure, 2014. **22**(6): p. 854-65.
153. Cornilescu, G., Didychuk, A.L., Rodgers, M.L., Michael, L.A., et al., *Structural Analysis of Multi-Helical RNAs by NMR-SAXS/WAXS: Application to the U4/U6 di-snRNA*. J Mol Biol, 2016. **428**(5 Pt A): p. 777-89.
154. Aznauryan, M., Delgado, L., Soranno, A., Nettels, D., et al., *Comprehensive structural and dynamical view of an unfolded protein from the combination of single-molecule FRET, NMR, and SAXS*. Proc Natl Acad Sci U S A, 2016. **113**(37): p. E5389-98.
155. Grishaev, A., Wu, J., Trewhella, J., and Bax, A., *Refinement of multidomain protein structures by combination of solution small-angle X-ray scattering and NMR data*. J Am Chem Soc, 2005. **127**(47): p. 16621-8.
156. Allison, J.R., *Using simulation to interpret experimental data in terms of protein conformational ensembles*. Curr Opin Struct Biol, 2017. **43**: p. 79-87.
157. Schneidman-Duhovny, D., Pellarin, R., and Sali, A., *Uncertainty in integrative structural modeling*. Curr Opin Struct Biol, 2014. **28**: p. 96-104.
158. Henriques, J., Cragnell, C., and Skepo, M., *Molecular Dynamics Simulations of Intrinsically Disordered Proteins: Force Field Evaluation and Comparison with Experiment*. J Chem Theory Comput, 2015. **11**(7): p. 3420-31.

159. Rauscher, S., Gapsys, V., Gajda, M.J., Zweckstetter, M., et al., *Structural Ensembles of Intrinsically Disordered Proteins Depend Strongly on Force Field: A Comparison to Experiment*. J Chem Theory Comput, 2015. **11**(11): p. 5513-24.
160. Palazzesi, F., Prakash, M.K., Bonomi, M., and Barducci, A., *Accuracy of current all-atom force-fields in modeling protein disordered states*. J Chem Theory Comput, 2015. **11**(1): p. 2-7.
161. Bonomi, M., Heller, G.T., Camilloni, C., and Vendruscolo, M., *Principles of protein structural ensemble determination*. Curr Opin Struct Biol, 2017. **42**: p. 106-116.

Publications

Zhu, G., Saw, W.G., Nalaparaju, A., Grüber, G., and Lu, L., *Coarse-Grained Molecular Modeling of Solution Structure Ensemble of Dengue Virus Non-Structural Protein 5 with Small-Angle X-ray Scattering Intensity*. J Phys Chem B, 2017, **121**(10): p. 2252-2264.

Zhu, G., Liu, W., Bao, C., Tong, D., Ji, H., Shen, Z., Yang, D., and Lu, L., *Investigating the Effect of Energy-Based Pool Structure Selection in the Structure Ensemble Modeling with Experimental Distance Constraints*. Proteins. (submitted)

Yang, J., **Zhu, G.**, Tong, D., Shen, Z., and Lu, L., *B-Spline Tight Frame Based Force Matching Method*. J. Comput. Phys. (submitted)

Zhu, G., Pan, A., Grüber, G., and Lu, L., *Replica-Exchange Molecular Dynamics Simulations Identify the Conformational States of ZIKA NS3*. (manuscript in preparation)

Liu, W., **Zhu, G.**, Liu, X., Lu, L., and Yang, D., *MPGA: Matching Pursuit Genetic Algorithm for Structure Ensemble Determination of Large Intrinsically Disordered Proteins*. (manuscript in preparation)