

HARNESSING PERIPHERAL SURFACE INFORMATION ENTROPY TO
CHARACTERIZE BIOMOLECULAR RECOGNITION AND OSMOTIC STRESS
RESPONSE

by

Tyler J. Grear

A dissertation submitted to the faculty of
The University of North Carolina at Charlotte
in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in
Bioinformatics and Computational Biology
Concentration: Molecular Biophysics

Charlotte

2026

Approved by:

Dr. Donald J. Jacobs

Dr. Jessica Roseberry

Dr. Jun-tao Guo

Dr. Cynthia Gibas

Dr. Kausik Chakrabarti

ABSTRACT

TYLER J. GREAR. Harnessing Peripheral Surface Information Entropy to Characterize Biomolecular Recognition and Osmotic Stress Response. (Under the direction of DR. DONALD J. JACOBS)

Predicting favorable protein-peptide binding events remains a central challenge in computational biology and biophysics, with uncertainty surrounding how nonlocal effects shape the global energy landscape. Here, peripheral surface information (PSI) entropy is introduced, a quantitative measure of statistical variability in apolar and charged non-interacting surface (NIS) proportions of conformational ensembles. Using HADDOCK3 molecular docking and explicit-solvent molecular dynamics simulations (MDS), it was demonstrated that favorable binding partners exhibit low-entropy macrostates indicative of preferential surface configurations. Across dozens of peptides and multiple receptor systems, dominant modes persisted under varied docking parameters and initial conditions. An experimental meta-ensemble of 36 WW domains confirmed dominant NIS modes independent of *in silico* methodology. Elevated p53 concentration under osmotic stress has been established at the systems biology level, this dissertation provides a molecular-scale explanation using MDS under varying osmotic pressures. Experimental NMR chemical shifts for the MDM2 disordered lid were compared with ensemble-based predictions, yielding a correlation coefficient of 0.998. With increasing osmotic pressure, lid-base contacts decreased and conformational populations redistributed, supporting a mechanism where increased lid rigidity governs the stress response. PSI entropy was then applied to apo MDM2 simulation trajectories, establishing a novel cancer-therapeutic strategy centered on peripheral-state targeting and destabilization.

ACKNOWLEDGEMENTS

I offer my sincere gratitude to Professor Donald J. Jacobs for his constant guidance, support, and mentorship throughout my academic career. I am likewise grateful to my thesis committee members for their time, expertise, and thoughtful consideration of this work. Their insight and encouragement have contributed meaningfully to the development of this dissertation.

This work was supported by the National Institute of General Medical Sciences of the National Institutes of Health (NIH) under Award Number R15GM146200. The content is solely the responsibility of the author and does not necessarily represent the official views of the National Institutes of Health.

I further acknowledge the University of North Carolina at Charlotte University Research Computing (URC) group for facilitating the use of high-performance computing (HPC) clusters and systems that were essential to this research.

DEDICATION

This dissertation is dedicated to Sergeant John Alden Lyons, whose courage, service, and sacrifice continue to guide my life and work. Per me tua virtus vivit (through me, your virtue lives on).

"The nation that will insist upon drawing a broad line of demarcation between the fighting man and the thinking man is liable to find its fighting done by fools and its thinking by cowards."

Sir William Francis Butler
Colonel, British Army

TABLE OF CONTENTS

LIST OF TABLES	ix
LIST OF FIGURES	xi
LIST OF ABBREVIATIONS	xvii
CHAPTER 1: INTRODUCTION	1
1.1. The Problem of Binding Affinity Predictions	1
1.1.1. Sampling in High-Dimensional State Spaces	2
1.1.2. A Need for Context-Specific Descriptors	3
1.1.3. Accounting for Global vs Local Thermodynamics	5
1.2. Non-Interacting Surfaces	6
1.3. MDM2 Under Osmotic Stress	7
1.4. Thesis Statement	9
1.5. Contributions of this Dissertation	9
1.5.1. Peripheral surface information entropy as an ensemble descriptor	9
1.5.2. $\mathcal{N}$ -space as a representation of peripheral-state populations	10
1.5.3. A peripheral destabilization strategy for WW domains	10
1.5.4. Peripheral design targets in apo MDM2	10
CHAPTER 2: SYSTEMS OF STUDY	11
2.1. WW Domain	11
2.2. PDZ Domain	13
2.3. MDM2 Protein	15

CHAPTER 3: THEORETICAL FRAMEWORK	17
3.1. Statistical Mechanics of Conformational Ensembles	17
3.2. Peripheral Surface Information Entropy, S_Ψ	18
3.3. Normalization of S_Ψ	20
3.4. Collective Variables, Order Parameters, and Reaction Coordinates	22
3.5. Potential of Mean Force and Free Energy Landscapes	23
3.6. Relevant NMR Chemical Shift Theory	24
CHAPTER 4: COMPUTATIONAL METHODS	27
4.1. Solvent Accessible Surface Area and Relative Solvent Accessibility	27
4.2. Ensemble Generation	27
4.2.1. Molecular Docking	27
4.2.2. Molecular Dynamics Simulation	28
4.3. S_Ψ Computation	29
4.4. Chemical Shift Prediction	31
CHAPTER 5: APPLICATION OF S_Ψ	33
5.1. Emergence of $\mathcal{N}$ -States Under Diverse Conditions and Conformations	33
5.2. Robustness of PSI Entropy Across Varied Systems	36
5.3. Experimental Validation of Dominant $\mathcal{N}$ -States	39
CHAPTER 6: A MOLECULAR-SCALE MECHANISM FOR P53 ABUNDANCE UNDER OSMOTIC STRESS	43
6.1. A Comparison of Computational and Experimental Chemical Shifts	43

	viii
6.2. Lid Rigidity and Lid-Base Dissociation	46
6.3. Osmotic Stress Induces State Population Shifts	49
CHAPTER 7: DISCUSSION AND FUTURE DIRECTIONS	52
REFERENCES	55
APPENDIX A: SUPPLEMENTAL TABLES	64
APPENDIX B: SUPPLEMENTAL FIGURES	71
APPENDIX C: SUPPLEMENTAL CALCULATIONS	79

LIST OF TABLES

TABLE 5.1: $\mathcal{N}$ -state properties between Step 2 (semi-flexible refinement) and Step 3 (MDS in explicit solvent) extracted from ensembles of protein-peptide complexes. Sixteen proline-rich peptides with motifs PPxY, LPxY, and PPGW were complexed with the WW domain (PDB ID 2LTW). The terms $\Delta\mathcal{N}$, ΔS_Ψ , and ΔK indicate the Step 3 minus Step 2 changes in the number of $\mathcal{N}$ -states, PSI entropy (S_Ψ), and the normalization constant $K = Q/M$ respectively.	35
TABLE 5.2: $\mathcal{N}$ -state properties between Step 2 (semi-flexible refinement) and Step 3 (MDS in explicit solvent) extracted from ensembles of protein-peptide complexes. Sixteen random sequences were predicted using AF3 then equilibrated with MDS in explicit TIP3P solvent (RD01 to RD16). The terms $\Delta\mathcal{N}$, ΔS_Ψ , and ΔK indicate the Step 3 minus Step 2 changes in the number of $\mathcal{N}$ -states, PSI entropy (S_Ψ), and the normalization constant $K = Q/M$ respectively.	35
TABLE 6.1: Contact summary for the N-terminal lid-base (residues 19 to 24), aggregated by osmotic condition. Values are computed over all frames pooled within each condition. The mean is the per-frame average over all runs. For total contacts and favorable and unfavorable sums, the per-frame standard deviation, reported as $\mu \pm \sigma$. Reduced contacts indicates diminished lid-base coupling to the MDM2 protein core.	49
TABLE A.1: Pairwise chemical weights for residue-residue contact classes.	64
TABLE A.2: NACCESS residue maximum solvent-accessible surface area values (ASA_{max}) for SASA to RSA calculations.	65
TABLE A.3: List of {PY} (PPxY, LPxY, and PPGW) protein sequences.	66
TABLE A.4: List of randomly generated protein sequences in the {RD} set.	66

TABLE A.5: $\mathcal{N}$ -state properties between Step 2 (semi-flexible refinement) and Step 3 (MDS in explicit solvent) extracted from conformational ensembles (docking solutions). Sixteen proline-rich peptides with motifs PPxY, LPxY, and PPGW were used to generate ensembles of protein-peptide complexes. The second column from the left contains the variables $\mathcal{N}_a$ and $\mathcal{N}_c$ being the NIS apolar and charged surface percentages respectively. P_{max} represents the maximum probability state (mode) in $\mathcal{N}$ -space, and S_Ψ is the peripheral surface information (PSI) entropy for a given ensemble. The terms $\Delta\mathcal{N}$ and ΔS_Ψ indicate the changes from Step 2 to Step 3 in the number of $\mathcal{N}$ -states and PSI entropy respectively.	67
TABLE A.6: $\mathcal{N}$ -state properties between Step 2 (semi-flexible refinement) and Step 3 (MDS in explicit solvent) extracted from conformational ensembles (docking solutions). Sixteen random sequences (decoys) were predicted using AF3 then equilibrated with MDS in explicit TIP3P solvent (RD01 to RD16). The second column contains the variables $\mathcal{N}_a$ and $\mathcal{N}_c$ which are the NIS apolar and charged surface percentages respectively. P_{max} represents the maximum probability state (mode) in $\mathcal{N}$ -space, and S_Ψ is the peripheral surface information (PSI) entropy for a given conformational ensemble. The terms $\Delta\mathcal{N}$ and ΔS_Ψ indicate the changes from Step 2 to Step 3 in the number of $\mathcal{N}$ -states and PSI entropy respectively.	68
TABLE A.7: Summary of the cross-fertilization procedure pairings for the MDM2 (4HFZ) and PDZ (1ZUB) receptors. Each phase involved 10 independent docking runs for both the proper (cognate) and improper (non-cognate) peptide candidates.	69
TABLE A.8: Experimental meta-ensemble systems used in Chapter 5.3 (experimental WW domains). Year denotes the initial PDB release year, the number of conformers in the NMR ensembles, and resolution is reported only for X-ray diffraction entries.	70

LIST OF FIGURES

- FIGURE 2.1: The WW domain. (a) The binding groove of the WW domain exists within the blue region bound by two highly conserved tryptophans (Trp17 and Trp39). Taking on the activated conformational state, the Tyr28 and Trp39 amino acids exhibit upward orientations relative to the beta sheet, likely due to π effects which are known contributors of favorable binding enthalpy. 12
- FIGURE 2.2: The crystallized 1BE9 PSD-95 PDZ domain characterized by X-ray diffraction. Target A (blue) contains the GLGF repeat within the carboxylate binding loop between the β A and β B strands. Target B (green) contains the α B helix which is vital for canonical binding interactions. 14
- FIGURE 3.1: The classical double pendulum. The Lagrangian generalized coordinates θ_1 and θ_2 capture the full configuration space of the system. (b) The double pendulum trajectories during a simulation of 120 seconds. (c) Centered at (0,0) is the suspension point p_1 . (d) The red end, p_2 , of the first pendulum shows a structured trajectory over time where any point is equidistant to p_1 . (e) When monitoring the blue end of the second pendulum a standard example of chaotic dynamics is revealed. Given enough time the blue trajectories would fill in this region completely having sampled all possible configurations. How long will it take to visit all possible configurations? This is just the sampling problem from another vantage. 23
- FIGURE 4.1: The $\mathcal{N}$ -space hexbin plot. Each hexagon represents a unit cell in NIS state space. After RSA thresholding and explicit interface exclusion, the exemplar ensemble contained $\Omega = 227$ microstates that partitioned into $N = 67$ distinct macrostates, with color indicating the multiplicity, $g(\mathcal{N}_a, \mathcal{N}_c)$. 30

FIGURE 5.1: An overview of $\mathcal{N}$ -state properties while varying: the Smad7 peptide conformations, number of rigid-body models, and number of models refined by semi-flexible MDS. (top row) Conformations are shown with N-termini on bottom and C-termini on top where backbone geometries are depicted in silver. (SMDA) Was taken from the NMR ensemble 2LTW without modification. (SMDB) This conformation was predicted with AF2. (SMDC) The prior AF2 conformation was equilibrated with MDS where a partial helix emerges. (SMDD) The final conformation was generated using AF3. (a) The number of macrostates, N , vs the number of microstates (Ω) where the number of rigid-body (RB) models was 25000. As the number of microstates increases, the number of emergent macrostates follow a logarithmic trend, reaching a maximum of $N = 157$ for SMDB. (b) Kernel density estimations of $\mathcal{N}_a$ proportions for the four Smad7 peptides. A dominant $\mathcal{N}_a$ mode emerges for all 4 peptide candidates, attaining a higher amplitude for the SMDA and SMDD partners.

34

FIGURE 5.2: An overview for the cross-fertilization procedure. (a) Under matched protocols, proper ensembles exhibit lower S_Ψ than improper ensembles for both receptors. Note, the separation $S_\Psi^{\text{phase}}(\text{proper}) < S_\Psi^{\text{phase}}(\text{improper})$ persisted even when considering uncertainty (shown by error bars). (b) The enthalpy-weighted contact masses with an upweighting of spatially-dense favorable contacts for the proper ELKS1b peptide engagement. This exemplifies Regime 1 with low S_Ψ and low Q/M .

38

FIGURE 5.3: The WW domain side (a) and front (b) views. (a) The binding groove of the WW domain exists within the blue region bound by two highly conserved tryptophans (Trp17 and Trp39). Taking on the activated conformational state, the Tyr28 and Trp39 amino acids exhibit upward orientations relative to the beta sheet, likely due to π effects which are known contributors of favorable binding enthalpy. (b) An alignment and superposition of all 36 PDB entries (34 NMR, 2 X-ray diffraction). The varying WW receptors (gray) are complexed with their corresponding peptide, ligand, or small molecule (multicolored ribbons).

41

FIGURE 5.4: Bivariate distributions derived from a meta-ensemble of varying experimental WW domains that were complexed with small molecules, ligands, and peptides. (a) The continuous $\mathcal{N}$ -space containing all 657 curated experimental protein-partner complexes. It is readily apparent that a dominant mode of the bivariate distribution exists, shown by red/dark-red regions. (b) The discrete $\mathcal{N}$ -space (hexbin plot) explicitly shows 3 general regions of high multiplicity across all microstates. Trifurcating the meta-ensemble based on known classification, the corresponding bivariate distributions are shown in (c-e). The dominant class contributions to the meta-ensemble are indicated by (I, II, and III) in (a).

42

FIGURE 6.1: Isotonic ensemble validation of N-terminal lid-base chemical shifts. Predicted shifts are plotted against Biological Magnetic Resonance Bank (BMRB) 6612 experimental shifts for residues 19 to 24. Each point is one matched (residue, atom) pair and colors indicate atom type (^{15}N , H^{N} , H_{α} , C_{α} , C_{β}). The dashed line is the identity ($y = x$), $r = 0.998$. Points lie close to that line, so the isotonic ensemble reproduces the experimental shift pattern for the matched nuclei.

44

FIGURE 6.2: Bland and Altman plot for isotonic predictions. Difference is $\delta_{\text{pred}} - \delta_{\text{exp}}$ for each matched residue and atom. Horizontal lines indicate zero, mean difference (bias), and limits of agreement (LoA) ($\mu \pm 1.96\sigma$). These results indicate a modest overall bias and generally good agreement between predicted and experimental chemical shifts.

45

FIGURE 6.3: One-dimensional free-energy basins along CV1 (lid extension; top row) and CV2 (lid-core angle; bottom row) for isotonic, hyperosmotic, and supraosmotic conditions. Histograms pool the same decimated frames (stride 600, approximately 12 ns between samples at a 20 ps sampling frequency). Potentials of mean force are shown with a fixed 0-5 $k_{\text{B}}T$ range, and vertical dotted lines mark the abscissa of the global minimum of each discrete PMF. The shaded $F < 2 k_{\text{B}}T$ region and horizontal dotted line retain their usual interpretation as a thermally accessible basin. With increasing osmolarity, CV1 shifts toward larger extension values, while CV2 exhibits stress-induced redistribution of hinge-angle geometry.

47

FIGURE 6.4: Condition-wise summary of the lid-extension coordinate (CV1; residue 1 to residue 23 $C_\alpha - C_\alpha$ distance). Bar height gives the condition level mean over trajectories (μ), error bars report the mean per-trajectory standard deviation (σ). From isotonic to hyperosmotic to supraosmotic conditions, μ shifts upward while σ decreases, indicating lid dynamics that are more extended and less flexible (more rigid) under osmotic stress. 48

FIGURE 6.5: Free-energy landscapes in the (CV1,CV2) plane. Free energy in kcal mol⁻¹ is relative to the minimum. Each column is one osmotic condition (isotonic, hyperosmotic, supraosmotic). Free energy increases away from the dominant basins, shifts in state populations show osmotic redistribution of the lid-base ensemble. 50

FIGURE 6.6: Stacked global $\mathcal{N}$ -space panels for osmotic-stress-driven population redistribution. The upper row shows condition-dependent occupancy multiplicity $g(N_a, N_c)$ for isotonic, hyperosmotic, and supraosmotic ensembles; the annotation above that row summarizes the scale of pooled sampling ($\mathcal{N}$ -space occupancy, 1.2×10^6 frames per osmotic condition). The lower row shows the corresponding change relative to the isotonic baseline (Δg), highlighting regions of decreased and increased occupancy and newly populated $\mathcal{N}$ -states. Data are derived from 96 trajectories of 250 ns each per condition, i.e. 24 μ s of aggregate sampling per condition and all twenty-four systems. The redistribution indicates that osmotic stress reorganizes equilibrium peripheral macrostates and reweights state populations relative to isotonic conditions. 51

FIGURE B.1: Dual energy-distribution plots for the WW domain-GESPPPPYSRYPMD conformational ensemble. The x-axes represent a 227 dimensional state space where each conformational state is ranked by HADDOCK score. The y-axes correspond to the energies: (a) van der Waals; (b) electrostatics; (c) desolvation; and (d) restraint violations. A Gaussian distribution indicates consistency of that energy over the ensemble of complexes as with (c). 71

FIGURE B.2: Complete $\mathcal{N}_a$ and $\mathcal{N}_c$ distributions for the Smad7 parameter sweep at RB = 10k (corresponding to the distributions summarized in Results 1, Figure 3b). Kernel density estimates are shown for $\mathcal{N}_a$ (top four panels) and $\mathcal{N}_c$ (bottom four panels) across the four Smad7 conformations (SMDA-SMDD; legend colors). Columns correspond to the maximum retained ensemble size $n_{\text{ret}} \in \{200, 1000, 5000, 10000\}$. 72

FIGURE B.3: Complete $\mathcal{N}_a$ and $\mathcal{N}_c$ distributions for the Smad7 parameter sweep at RB = 25k (corresponding to the distributions summarized in Results 1, Figure 3b). Kernel density estimates are shown for $\mathcal{N}_a$ (top four panels) and $\mathcal{N}_c$ (bottom four panels) across the four Smad7 conformations (SMDA-SMDD; legend colors). Columns correspond to the maximum retained ensemble size $n_{\text{ret}} \in \{200, 1000, 5000, 10000\}$.

73

FIGURE B.4: Contact map summaries for receptor 4HFZ under the cross-fertilization protocol. (top) Ensemble-averaged contact map probabilities, $\rho(i, j)$. (middle) Contact-class weighted probability map, $\tilde{m}_{i,j}$, reflecting the enthalpic contribution of individual contacts. (bottom) The ensemble-level net contact-class modulation, $\sum_{\Omega}(\gamma(c_i, c_j) - 1.0) \cdot m_{i,j}$, where positive (red) and negative (blue) values denote regions of favorable and unfavorable biophysical modulation respectively.

74

FIGURE B.5: Contact map summaries for receptor 1ZUB under the cross-fertilization protocol. (top) Ensemble-averaged contact map probabilities, $\rho(i, j)$. (middle) Contact-class weighted probability map, $\tilde{m}_{i,j}$, reflecting the enthalpic contribution of individual contacts. (bottom) The ensemble-level net contact-class modulation, $\sum_{\Omega}(\gamma(c_i, c_j) - 1.0) \cdot m_{i,j}$, where positive (red) and negative (blue) values denote regions of favorable and unfavorable biophysical modulation respectively. For 1ZUB, the active residues are discontinuous in sequence shown by a white horizontal space between residues 638 and 683.

75

FIGURE B.6: Phase-pooled bivariate $\mathcal{N}$ -space density for receptor 4HFZ under the cross-fertilization protocol (proper vs improper). Microstates are pooled across replicas within each phase.

76

FIGURE B.7: Phase-pooled bivariate $\mathcal{N}$ -space density for receptor 1ZUB under the cross-fertilization protocol (proper vs improper). Microstates are pooled across replicas within each phase.

77

FIGURE B.8: Phase-pooled bivariate $\mathcal{N}$ -space density for the three experimental meta-ensemble classes: (top) Class 1, (middle) Class 2, and (bottom) Class 3. Densities are computed using the RSA-thresholded count metric with explicit interface exclusion.

78

FIGURE C.1: Autocorrelation functions for CV1 (lid extension, red) and CV2 (lid-core angle, black) computed from the isotonic simulation ensemble. The dashed horizontal line indicates the $1/e$ threshold used to estimate characteristic decorrelation time. CV1 decays more rapidly, whereas CV2 remains correlated over substantially longer lag times, consistent with slower collective relaxation of the lid-core angular DoF.

LIST OF ABBREVIATIONS

ACF	Autocorrelation function.
AF2	AlphaFold2.
AF3	AlphaFold3.
AIR	Ambiguous interaction restraint.
BMRB	Biological Magnetic Resonance data Bank.
BSA	Buried surface area.
CV	Collective variable.
DoF	Degrees of freedom.
ELKS1b	Protein rich in glutamate, leucine, lysine, and serine 1b.
FES	Free-energy surface.
HADDOCK3	High Ambiguity Driven biomolecular DOCKing 3.
iRMSD	Interface root-mean-square deviation.
MDM2	Murine double minute 2.
MDS	Molecular dynamics simulation(s).
NIS	Non-interacting surface.
NLS	Nuclear localization sequence.
NMR	Nuclear magnetic resonance.
NPT	Isothermal-isobaric ensemble.
NVE	Microcanonical ensemble.
NVT	Canonical ensemble.
OPLS	Optimized potentials for liquid simulations.
PDB	Protein Data Bank.
PDZ	Postsynaptic density-95/discs-large/zonula occludens-1.
PME	Particle-mesh Ewald.
PMF	Potential of mean force.
PSI	Peripheral surface information.
PTPN14	Protein tyrosine phosphatase non-receptor type 14.
RMSE	Root-mean-square error.
RSA	Relative solvent accessibility.
SASA	Solvent-accessible surface area.
TIP3P	Three-point transferable intermolecular potential.
WW	Protein domain containing two conserved tryptophan residues.
YAP	Yes-associated protein.
YAP65	65 kDa Yes-associated protein.

CHAPTER 1: INTRODUCTION

1.1 The Problem of Binding Affinity Predictions

Predicting binding affinity from molecular structure remains challenging because experimentally measured affinity is an equilibrium property of ensembles rather than a property of a single representative structure. The relevant ensembles occupy extremely high-dimensional state spaces shaped by conformational variability, intermolecular recognition, and solvent-coupled reorganization; subsequently, the primary bottleneck is sampling. The curse of dimensionality implies that only a negligible fraction of configuration space is accessible to any finite computational protocol, so models must extract reproducible signals from incomplete conformational coverage.

A second limitation is discriminatory power. Many biomolecular interactions are stabilized by numerous weak contributions distributed across the system rather than by a small number of dominant contacts, where enthalpy-entropy compensation and related free energy window effects can map mechanistically distinct binding processes onto similar values of ΔG . This reduces the interpretability and predictive value of ΔG as a standalone descriptor and motivates context-specific descriptors that characterize how binding is stabilized, not only the net favorability.

Finally, an interface-focused description is often incomplete because binding is conditioned by global thermodynamic factors. Solvent reorganization, peripheral surface properties, and nonlocal coupling can modulate which bound states are populated and how stability is achieved even when interfacial contacts appear similar. This dissertation therefore emphasizes an extended-interface perspective that integrates local interfacial energetics, global protein-solvent organization, and develops ensemble-level thermoinformatic measures to quantify binding constraints.

1.1.1 Sampling in High-Dimensional State Spaces

The curse of dimensionality is not an abstract concern in biomolecular modeling. It is a concrete barrier that arises the moment one treats binding as an ensemble problem. Even when ignoring explicit solvent, the conformational degrees of freedom (DoF) for flexible proteins and peptides produce a state space whose effective volume grows exponentially with system dimension [1, 2]. With explicit solvent, the number of microstates explodes further, and the equilibrium ensemble becomes a small, structured subset of a vast configuration space. The practical consequence is that no finite simulation campaign can claim exhaustive exploration, and no single protocol can guarantee that the dominant thermodynamic basins have been identified [3, 4].

This sampling limitation creates two distinct risks. The first is omission, where relevant states may never be visited, low probability events may be important concerning functional dynamics or transition pathways. The second is bias, in which the states that are visited may reflect initial conditions, restraints, force-field idiosyncrasies, or protocol choices more strongly than true equilibrium structure. Docking can generate many candidate poses, often under simplified treatments of receptor flexibility and approximate scoring functions [5]. Molecular dynamics simulation (MDS) can incorporate richer physical detail, yet it frequently remains confined to local free-energy basins over accessible simulation timescales, precisely why enhanced-sampling methods have become so important in biomolecular simulation [6, 7].

An intuitive geometric analogy is to imagine standing in an empty room and tossing a handful of pebbles onto the floor, where they spread uniformly over the surface. In the approximately two-dimensional plane occupied by the pebbles, the coverage may appear extensive. Once the full three-dimensional room is considered, those same points sample only a narrow slice of the available volume. The same logic extends to high-dimensional spaces, coverage that appears substantial on a low-dimensional manifold may only correspond to a vanishingly small subset of the full state space.

Both docking and simulation operate under partial information, and the central question becomes how to extract reliable conclusions from incomplete ensembles.

1.1.2 A Need for Context-Specific Descriptors

Affinity prediction faces a second challenge beyond sampling. Binding is commonly summarized by the standard thermodynamic relation

$$\Delta G = \Delta H - T\Delta S, \quad (1.1)$$

where ΔG is the Gibbs free energy, ΔH is the change in enthalpy, given $\Delta H = U + PV$, and $T\Delta S$ is the absolute temperature-scaled change in entropy. This observable compresses distinct physical contributions into a single scalar. Experimental affinity is reported primarily as the dissociation constant, K_d , or ΔG , while decomposition into enthalpic and entropic terms requires additional thermodynamic analysis [8]. Even when ΔH and ΔS are available, multiple mechanistically distinct combinations can yield similar ΔG values. This degeneracy is especially important in biomolecular recognition such as antibody-antigen complexation, where binding often reflects many weak and distributed contributions including hydrophobic contacts, electrostatics, hydrogen bonding, and solvent reorganization. Different mechanisms can converge to similar affinities, masked by enthalpy-entropy compensation [9].

Equation 1.1 is deceptively simple. Although $\Delta G = \Delta H - T\Delta S$ provides the correct thermodynamic decomposition, it does not by itself resolve the many distinct physical contributions embedded within the entropic term, including conformational, solvent, and rotational-translational effects [10]. This compression is one reason that binding-affinity calculations often lack generalizability and transferability across systems unseen [11]. As an initial step toward more transferable free-energy determination, this dissertation introduces an entropic component that encodes both the peripheral surface composition of the protein and the structural organization of the

first hydration shell, grounded by the known importance of hydration-shell water in biomolecular recognition [12, 13].

There is something subtle lurking behind Equation 1.1. The entropy term, S , is not a single mechanistic quantity that can be identified with any one isolated source of disorder. In the context of biomolecular binding, it denotes the total thermodynamic entropy of the process, and therefore subsumes multiple distinct contributions arising from molecular motion, internal rearrangement, and environmental reorganization [14]. A more explicit decomposition may be written as

$$\Delta S_{\text{total}} = \Delta S_{\text{trans}} + \Delta S_{\text{rot}} + \Delta S_{\text{vib}} + \Delta S_{\text{conf}} + \Delta S_{\text{solv}} + \Delta S_{\text{ion}} + \cdots, \quad (1.2)$$

where the terms respectively denote translational, rotational, vibrational, conformational, solvent, and ionic contributions to the total entropy change upon complex formation. Under this view, the persistent lack of transferability and generalizability in computational free energy determination does not arise simply due to numerical insufficiency. It reflects a more basic informational deficit, I lean towards the view that the set of entropic quantities typically retained is not robust enough to represent the full state dependence ingrained in different binding processes of biomolecules.

This limitation constitutes a fundamental barrier in free energy research. Even for the seemingly narrower case of conformational entropy, exact evaluation is formally tied to the full distribution of accessible microstates. In statistical mechanical form,

$$S_{\text{conf}} = -k_{\text{B}} \sum_i p_i \ln p_i, \quad (1.3)$$

which requires knowledge of the probabilities p_i over the complete conformational ensemble. For realistic biomolecular systems, this ensemble is neither finitely enumerable or analytically tractable. Consequently, exact computation of even one com-

ponent of ΔS_{total} is already obstructed by the need for information that is inaccessible.

If the full entropic content of binding cannot be directly resolved, then the problem shifts from exhaustive accounting to principled representation. From this view, a global entropy may be interpreted as an entropic proxy designed to capture otherwise unrepresented constraints encoded in the organization of the system. The broader objective is not to enumerate every entropy explicitly, it is to curate a set of entropic proxies that are sufficient for the research question being posed. Progress in binding free-energy prediction depends on whether these proxies can preserve the latent information that conventional decompositions leave behind or never extract at all.

1.1.3 Accounting for Global vs Local Thermodynamics

A common simplifying assumption in affinity prediction is that binding is determined primarily by local interactions at the interface. This assumption is computationally attractive because interfacial contacts are straightforward to enumerate and convert into approximate energetic terms, and contact-based models are able to recover substantial predictive signal from interface structure alone [15]. The surrounding environment provides thermodynamic context, and that context can reshape the global energy landscape even when local contacts appear dominant. This motivates an extended-interface viewpoint in which nonlocal interactions and peripheral effects can contribute meaningfully to binding strength and specificity [16, 17].

Several lines of evidence support this broader view. Solvent reorganization can contribute materially to binding thermodynamics through hydration, electrostatics, and water-mediated enthalpy-entropy balance, while binding heat-capacity changes provide an experimental signature of restructuring that extends beyond a purely local contact picture [18]. Allosteric coupling provides a second route by which distal regions can influence binding, since perturbations far from the interface can redistribute conformational populations and alter binding-relevant states [19]. A model restricted to local interface energetics is likely to remain systematically incomplete, especially

in systems dominated by weak interactions and solvent-coupled effects.

This dissertation treats binding as a joint constraint with both local and global components. Local interfacial energetics remain necessary, yet they are embedded in a larger thermodynamic setting determined by protein-solvent interactions and peripheral surface composition. The practical question is how to represent that global contribution in a form that is measurable, computable, and predictive. Ensemble-level metrics that quantify organization, variability, or concentration of peripheral states provide a natural route, because they express nonlocal effects as statistical constraints rather than ad hoc corrections to an interface score.

The resulting design philosophy is to optimize interfacial contacts together with the peripheral and solvent-coupled organization that stabilizes binding. In this dissertation, the peripheral surface information (PSI) entropy is introduced as one realization of that viewpoint. It combines an enthalpy-aware contact-statistical component motivated by the known predictive value of interfacial contact networks [15], with a global entropy supported by prior evidence that the non-interacting surface (NIS), its solvation, and its dynamics contribute to favorable binding conditions [16, 17].

1.2 Non-Interacting Surfaces

Ideas that the non-interacting surface (NIS) during protein-protein association can contribute, in a non-negligible manner, to binding affinity calculations have progressed over the last few decades. They were largely debated within the context of allosteric effects and entropy-driven solvent reorganization [20, 21]. In the mid-2000s, researchers began to discuss the possibility of an extended interface [22], asking whether nonlocal interactions can meaningfully affect binding strength and/or specificity. Volumetric and calorimetric measurements, together with nuclear magnetic resonance (NMR) spectroscopy have enabled quantitative monitoring of these nonlocal influences on binding mechanisms. Additionally, changes in heat capacity upon formation of a complex can provide information on nonlocal effects [22, 23]. In systems governed

by complex weak interactions, enthalpy-entropy compensation can limit the discriminative value of ΔG -based descriptors [24]. Free-energy functions that are overly concentrated on the binding interface often miss the global thermodynamic context provided by the environment. In this dissertation, the heterogeneity of free-energy contributions is related to information entropy, establishing that conformations corresponding to favorable energetics can be characterized by unique statistical signatures that are encoded on the peripheral surface.

Since the mid-2010s, researchers have called for quantitative descriptions of NIS characteristics in order to build a global model of binding affinity and identify an Archimedean point for free-energy determination [25, 26, 27]. Furthermore, new entropic considerations are expected for flexible complexes, where the correlation of buried surface area (BSA) and binding affinity does not hold [26]. This demands new global metrics derived from entire conformational ensembles rather than representative binding poses or energy profiles. It has been shown that the percentages of apolar and charged NIS (denoted $\mathcal{N}_a$ and $\mathcal{N}_c$ respectively) exhibit significant correlations with binding affinity [25, 28, 15, 29], and NIS effects have been verified through alanine scanning [25, 28]. It has also been observed that proportions of $\mathcal{N}_a$ and $\mathcal{N}_c$ are conserved over orthologous complexes indicating an evolutionary selection pressure [28, 30]. Assuming evolution has performed the functional optimization *a priori*, an effective bioengineering protocol should harness the fingerprints that emerge from this selective pressure. A new information-theoretic entropy is proposed that encodes NIS properties over ensembles of molecular complexes consistent with favorable global binding conditions. This emergent entropy is verified to exist independent of *in silico* method, extracted from ensembles of experimentally resolved molecular complexes.

1.3 MDM2 Under Osmotic Stress

Osmotic pressure, Π , arises from differences in solute concentration across a semipermeable boundary. In cellular settings, Π provides a measure of the thermodynamic

tendency of water to displace in response to those concentration differences [31], not to be confused with mechanical pressure often denoted P . Hyperosmotic conditions increase the effective concentration of solutes in the surrounding environment, altering hydration, macromolecular crowding, and the physical constraints experienced by proteins and other cellular components [32]. The murine double minute 2 (MDM2) oncoprotein and the tumor suppressing protein p53, are also relevant in the context of osmotic stress, where prior work has shown that hyperosmotic conditions perturb this regulatory axis at the cellular and systems level. MDM2 is the principal negative regulator of p53, acting through direct binding to p53 and ubiquitin-dependent control of its stability [33, 34]. The p53 protein functions as a central stress-responsive tumor suppressor that coordinates cell-cycle arrest, DNA damage responses, and apoptosis [35]. Across multiple experimental settings, hyperosmotic stress has been observed to stabilize or activate p53, promote p53-dependent checkpoint signaling, and alter the regulatory state of MDM2. In renal inner medullary cells, hypertonic stress induces p53 activation in a manner that is protective against apoptosis, and subsequent work showed that this protection is associated with restriction of DNA replication and delayed cell-cycle progression under hypertonic conditions [36, 37]. In fibroblast systems, osmotic shock was shown to transiently stabilize p53, promote p38-dependent phosphorylation of p53, induce p21, and drive G_1 arrest, with this response linked mechanistically to altered MDM2-mediated control of p53 stability [38].

More recent work in NIH3T3 cells further demonstrated that increasing hyperosmotic stress produces a pronounced p53 response over a broad osmolarity range, while also modulating phosphorylated MDM2 and phosphorylated p38, reinforcing the view that osmotic challenge is transmitted through the MDM2-p53 regulatory network rather than through p53 alone [39]. At a broader systems level, transcriptomic, proteomic, and metabolomic analyses of hypertonic saline responses in airway epithelial cells likewise identified activation of the p53-p21 signaling axis as a major

outcome of modest hyperosmolar treatment [40]. These studies establish that osmotic perturbation is a well-known upstream regulator of the MDM2-p53 system.

What remains much less developed is a molecular-scale description of how this phenomenon is encoded in the conformational ensemble of MDM2 itself. Existing work on MDM2 structural dynamics has largely focused on the intrinsic disorder of the N-terminal lid, pseudosubstrate-like self-occlusion of the p53-binding cleft, phosphorylation-dependent regulation, and ligand-induced plasticity of the binding groove [41, 42, 43]. These studies are highly informative; however, they do not explain how an osmotic perturbation to the surrounding environment redistributes the conformational states of apo MDM2 or reshapes the accessibility of the p53-binding site. Thus, although the biological consequence of osmotic stress on the p53 pathway has been recognized for years, the mechanistic bridge from osmotic stress to MDM2 conformational response has remained unresolved.

1.4 Thesis Statement

This dissertation advances the thesis that peripheral surface organization encodes thermodynamic information relevant to biomolecular recognition and environmental response that is not recoverable from interface-centered descriptions alone. To test this thesis, peripheral surface information (PSI) entropy is introduced as an information-theoretic measure of ensemble variability in apolar and charged non-interacting surface states, and $\mathcal{N}$ -space is developed as a reduced representation of peripheral-state populations across docking and molecular simulation ensembles.

1.5 Contributions of this Dissertation

1.5.1 Peripheral surface information entropy as an ensemble descriptor

This dissertation introduces peripheral surface information (PSI) entropy, an information-theoretic descriptor of peripheral surface organization in biomolecular ensembles. By quantifying variability in apolar and charged non-interacting surface states across

docking and MDS ensembles, PSI entropy extends affinity analysis beyond interfacial contacts and captures thermodynamically relevant peripheral heterogeneity.

1.5.2 $\mathcal{N}$ -space as a representation of peripheral-state populations

$\mathcal{N}$ -space is introduced as a reduced representation of peripheral-state populations in conformational ensembles. Monitoring shifts in the statistical distribution of apolar and charged peripheral surface compositions, $\mathcal{N}$ -space resolves ensemble-level population changes, including stress-induced responses in apo MDM2.

1.5.3 A peripheral destabilization strategy for WW domains

A peripheral destabilization strategy for WW domains is introduced that is based on perturbation of non-binding surface features rather than direct modification of the canonical peptide-recognition groove. In this framework, WW-domain design is recast in terms of distal thermodynamic control, with the goal of modulating PPxY-mediated recognition and its context-dependent effects on YAP signaling.

1.5.4 Peripheral design targets in apo MDM2

This work identifies peripheral design targets in apo MDM2 outside the canonical p53-binding groove. These regions are proposed as distal thermodynamic controls through which lid organization, binding-site accessibility, and ensemble populations may be modulated, establishing a foundation for allosteric design strategies.

CHAPTER 2: SYSTEMS OF STUDY

2.1 WW Domain

The first system considered is the WW-domain complex deposited as PDB ID 2LTW, determined by solution NMR. This structure corresponds to the first WW domain of human Yes-associated protein (YAP WW1) in complex with a 14-residue Smad7-derived peptide, GESPPPPYSRYPMD [44]. In the context of the full-length protein, YAP was originally identified as a 65 kDa phosphoprotein (YAP65) and is widely recognized as a transcriptional co-regulator that most often acts in a pro-oncogenic manner upon nuclear translocation, although tumor-suppressive roles have also been reported in specific biological contexts, including breast cancer [45, 46, 47]. The WW domain is a compact globular module of 36 residues containing the canonical three-stranded antiparallel β -sheet characterized by two conserved tryptophan amino acids flanking each end. This fold is widely used for investigating the thermodynamics, folding kinetics, and peptide-recognition properties of small β -sheet proteins [48].

The Class I WW domain recognizes proline-rich peptide motifs of the general form PPxY or LPxY [49]. In 2LTW, the WW1 complexed with a Smad7 peptide is an example of canonical WW-mediated peptide recognition as shown in Figure 2.1. The interface is formed by residues that are discontinuous in sequence but contiguous in three-dimensional space, notably Tyr19, Leu21, His23, Gln26, and Trp30, which together define the established XP and Tyr binding grooves [49, 44]. This enables the domain to engage both the proline-rich segment of the ligand and the critical tyrosine side chain that specifies Class I recognition. PDB entry 2LTW serves as a particularly informative model system because it captures the structural logic of small β -sheet recognition and is suitable for studying how local surface chemistry and

topology encode molecular recognition in peptide-binding domains.

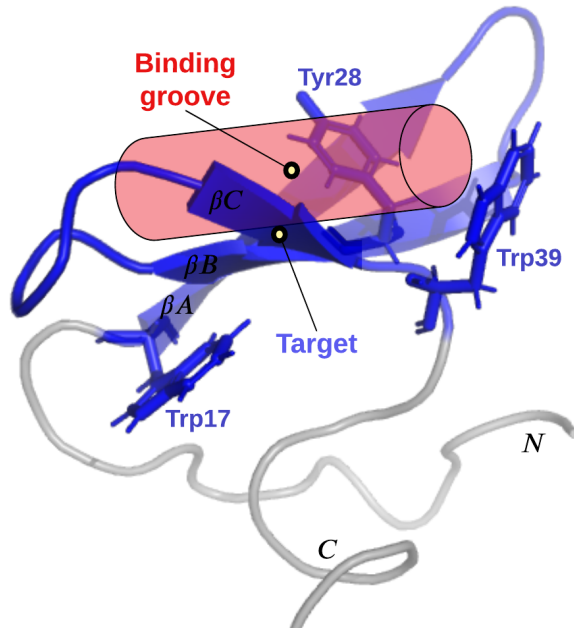


Figure 2.1: The WW domain. (a) The binding groove of the WW domain exists within the blue region bound by two highly conserved tryptophans (Trp17 and Trp39). Taking on the activated conformational state, the Tyr28 and Trp39 amino acids exhibit upward orientations relative to the beta sheet, likely due to π effects which are known contributors of favorable binding enthalpy.

Destabilizing the WW domains of YAP is expected to shift signaling toward tumor suppression by reducing two cancer-relevant functions of the protein. First, WW-domain mutation diminishes YAP-driven cell proliferation and oncogenic transformation, indicating that intact WW domains are required for a substantial portion of the pro-growth transcriptional program [50]. Second, the folded WW domains of YAP1 also function as a WW-nuclear localization signal (NLS) that promotes nuclear import where interference with this module attenuates YAP1 nuclear translocation, impairs stemness gene activation, and reduces sphere formation in malignant breast cancer cells [51]. This strategy remains context-dependent because WW domains also mediate binding to negative regulators such as protein tyrosine phosphatase, non-receptor type 14 (PTPN14), which restrains YAP by promoting its inactive cytoplasmic state

[52]. Destabilization of WW-domains is best viewed as a plausible anti-oncogenic strategy that can suppress YAP nuclear function and growth-promoting signaling, while requiring careful evaluation of partner-specific effects.

2.2 PDZ Domain

PDZ (postsynaptic density-95/discs-large/zonula occludens-1) domains are a large class of compact globular protein-protein interaction modules distributed broadly across eukaryotes, from yeast to humans [53, 54]. PDZ-containing proteins have attracted sustained interest due to their central roles in organizing signaling assemblies; subsequently, they represent viable targets for molecular inhibition in therapeutic design [55, 53, 56, 57]. The structural basis of PDZ-mediated binding specificity is well studied and is commonly described using a simplified motif-based classification framework [58, 59]. Class I PDZ domains recognize motifs of the form X-S/T-X- Φ_{COO^-} , class II domains recognize X- Φ -X- Φ_{COO^-} , and class III domains recognize X-D/E-X- Φ_{COO^-} , where X denotes any residue, A/B is A or B, and Φ denotes a hydrophobic residue, typically L, V, I, or F [53, 60]. Although broader and more nuanced classification schemes have been proposed [59], this simplified framework is sufficient for general context. PDZ domains often display substantial degeneracy and contextual plasticity in ligand recognition, including interactions with noncanonical internal motifs, lipids, and even other PDZ domains, giving rise to the broader and promiscuous PDZ interactome [61, 60, 62].

The second system of study is the RIM1 α PDZ domain in complex with an ELKS1b C-terminal peptide (CDQDEEEGIWA), deposited as PDB ID 1ZUB and determined by solution NMR [63]. 1ZUB captures a presynaptic active-zone scaffold complex in which the PDZ domain of RIM1 α recognizes the C terminus of ELKS1b [64, 63]. The deposited construct contains the *Rattus norvegicus* RIM1 α PDZ-domain fragment together with an ELKS1b peptide derived from the extreme C terminus of the binding partner. Structurally, the receptor retains the canonical PDZ fold, comprising a

compact globular architecture built from a six-stranded β -sheet surrounded by two α helices. As with other PDZ domains, the ligand binds along the groove formed between the carboxylate-binding loop and the α B helix, extending the β -sheet in an antiparallel manner as shown in Figure 2.2.

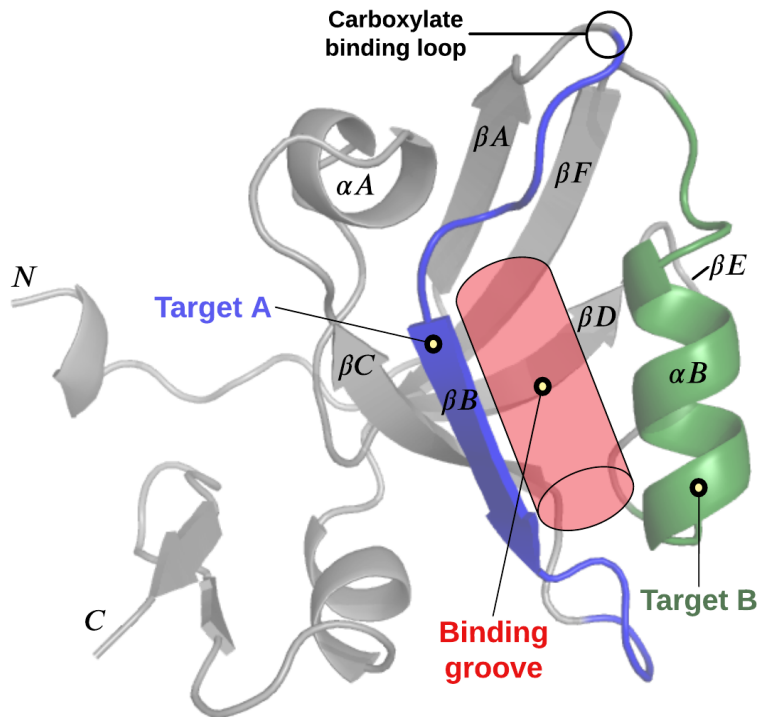


Figure 2.2: The crystallized 1BE9 PSD-95 PDZ domain characterized by X-ray diffraction. Target A (blue) contains the GLGF repeat within the carboxylate binding loop between the βA and βB strands. Target B (green) contains the αB helix which is vital for canonical binding interactions.

This system is especially informative because its specificity depends on the combined contribution of the four C-terminal ELKS residues, consistent with prior biochemical mapping of the ERC/ELKS C terminus [64, 63]. For the purposes of this study, 1ZUB provides a compact and mechanistically rich model of peptide-mediated recognition in which the binding surface is well resolved, the complex is experimentally characterized in solution, and the molecular determinants of specificity are stringent. These features make the RIM1 α -ELKS1b complex a useful complement to more classical PDZ model systems, particularly for examining how surface topology and local

chemistry constrain selective binding in small modular interaction domains.

2.3 MDM2 Protein

The third system of study is murine double minute 2 (MDM2), a cancer-related oncoprotein and central negative regulator of the tumor suppressor p53 [33]. Through direct recognition of the N-terminal transactivation region of p53, MDM2 suppresses p53 activity and promotes its ubiquitin-dependent downregulation, making the p53-MDM2 axis one of the most extensively studied protein-protein interactions in cancer biology [65, 66, 34, 67]. For the aims of this dissertation, MDM2 introduces a complementary regime of peptide-mediated recognition in which binding is concentrated into a hydrophobic pocket, yet remains sensitive to conformational heterogeneity and the broader thermodynamic state of the receptor surface.

Two experimentally resolved MDM2 structures are considered in this work in order to distinguish between the open, ligand-compatible receptor state and the autoinhibitory lid-bearing state. The first is PDB ID 4HFZ, an X-ray structure resolved at 2.69 Å in which the p53-binding domain of MDM2 is observed in complex with an N-terminal p53-derived peptide. In this bound state, the peptide adopts the canonical α -helical recognition geometry and inserts its hydrophobic side chains into the well-known MDM2 groove. Because this structure presents an open and well-defined binding cleft, it serves as an appropriate model for analyses centered on the classical p53-recognition pocket. More broadly, this lidless or effectively core-domain representation is the form most commonly employed in docking protocols, virtual screening studies, and structure-based inhibitor design targeting the p53-MDM2 interface. As a result, 4HFZ is the form used in Chapter 5, where MDM2 is treated as a conventional deep-pocket peptide receptor for docking.

To complement this open-pocket representation, the second MDM2 structure considered here is PDB ID 1Z1M, an apo solution NMR structure that retains the intrinsically disordered N-terminal lid. Unlike peptide-bound crystal structures of the

open cleft, 1Z1M captures the receptor in a state in which the N-terminal lid can associate with and partially occlude the p53-binding groove. This lid-bearing state is especially important because it is a dynamically regulated surface capable of transient self-inhibition. Prior structural and dynamical studies have shown that the lid can compete with ligands for access to the cleft, exchange between docked and undocked conformations, and act as a regulatory element that modulates binding competence. 1Z1M is therefore studied specifically to investigate the dynamics of the disordered lid and its role in self-inhibition, whereas 4HFZ is used as the corresponding lidless, ligand-compatible receptor state that underlies much of the existing docking literature. These two MDM2 structures provide a holistic view of the same cancer-relevant receptor system. The 4HFZ structure emphasizes the canonical peptide-bound geometry of the open hydrophobic pocket, while 1Z1M emphasizes the conformational plasticity of the apo receptor and the regulatory influence of the N-terminal lid.

CHAPTER 3: THEORETICAL FRAMEWORK

3.1 Statistical Mechanics of Conformational Ensembles

A single experimentally resolved or energy-minimized structure is often useful as a reference; however, it is not a complete physical description of a biomolecule. Under physiological conditions, biomolecular systems undergo continual thermally driven fluctuations and therefore occupy a distribution of accessible conformations rather than one fixed geometry [68, 69]. From the perspective of molecular biophysics, any observed structure is best understood as one member of a broader conformational space, and the biologically relevant object of description is the ensemble of accessible states. This statistical representation may be expressed by assigning a probability p_i to each accessible conformation x_i , such that $\sum_i p_i = 1$.

In statistical mechanics, each conformation, x_i , may be treated as a microstate with an associated energy, where its contribution depends on both accessibility and the likelihood of observing the i -th microstate. Subsequently, a favorable binding geometry or low interaction energy does not guarantee substantial macroscopic association when the corresponding bound state is only sparsely populated. Experimentally and computationally relevant quantities are best described by population-weighted averages, such that an observable A is given by $\langle A \rangle = \sum_i p_i A(x_i)$.

Many biological processes are governed by shifts in ensemble state populations instead of a single uniquely optimal structure. Even relatively low-population conformations may remain mechanistically important when they participate in binding or transition pathways [70]. It follows that a single conformation may fail to capture the entropic and dynamical contributions that shape biological function. Because many experimental and simulation-derived observables are inherently ensemble quantities,

computational modeling requires adequate sampling of conformational space [4]. This statistical-mechanical viewpoint motivates the ensemble-based analyses throughout this dissertation, where the objective is to characterize distribution-level properties.

3.2 Peripheral Surface Information Entropy, S_Ψ

The contribution of the non-interacting surface (NIS) to biomolecular recognition suggests that descriptors restricted to the binding interface may omit global thermodynamic information [16, 17]. This limitation becomes especially relevant in flexible systems, where single representative poses do not adequately capture ensemble-level phenomena [70]. The peripheral surface information (PSI) entropy, S_Ψ , is introduced here as a statistical measure of the variability in peripheral NIS states.

Consider a conformational ensemble containing Ω microstates, denoted $\{x_i\}_{i=1}^\Omega$. Solvent-accessible surface area (SASA) and relative solvent accessibility (RSA) calculations were performed [71], for which technical descriptions are provided in Section 4.1 [72]. Each microstate is then assigned a peripheral macrostate according to the remaining NIS composition. The macrostate label is defined as the exact integer tuple (n_A, n_C, n_P) , where n_A , n_C , and n_P denote the number of apolar, charged, and polar components of the peripheral surface respectively. These macrostates map onto points in NIS state space through their corresponding fractional compositions. The occupied states may be visualized in the projected $(\mathcal{N}_a, \mathcal{N}_c)$ plane referred to as $\mathcal{N}$ -space, where $\mathcal{N}_a = n_A/n_{\text{tot}}$ and $\mathcal{N}_c = n_C/n_{\text{tot}}$, with $n_{\text{tot}} = n_A + n_C + n_P$, while the full (n_A, n_C, n_P) tuple remains the distinguishing macrostate definition.

Let $\mathcal{N}_j$ denote the j -th occupied macrostate in $\mathcal{N}$ -space, and let N be the number of distinct occupied macrostates. The number of microstates that exist in a given macrostate is quantified by the multiplicity function, $g(\mathcal{N}_j)$. In the projected plane, this structured occupancy may be viewed equivalently as $g(\mathcal{N}_a, \mathcal{N}_c)$, and the corresponding empirical probability is $p_i = g(\mathcal{N}_j)/\Omega$. A nonuniform occupancy in $\mathcal{N}$ -space indicates that the ensemble does not populate all peripheral states equally and instead

concentrates probability within a preferential subset of accessible NIS patterns. This structured occupancy motivates an entropy-based description of NIS state variability.

Introduced by Claude Shannon in 1948, information entropy quantifies the amount of information contained in a message or the unpredictability of a system [73]. For a finite countable set of states, Shannon entropy is written as $S = -K \sum_i p_i \log_2(p_i)$, where K is a positive scaling factor. Choosing $K = 1$ gives the entropy in binary digits (bits), given a log base of 2. The macrostate entropy is introduced first in its intermediate form, while the physically motivated specification of K is given in Section 3.3. The mapping of ensemble occupancy to macrostate-level structure in $\mathcal{N}$ -space allows a peripheral surface information (PSI) entropy to be expressed as

$$S'_\Psi = -K \sum_{j=1}^N \left(\frac{g(\mathcal{N}_j)}{\Omega} \right) \log_2 \left(\frac{g(\mathcal{N}_j)}{\Omega} \right). \quad (3.1)$$

The prime notation indicates the intermediate expression prior to contact-based normalization. In this form, S'_Ψ serves as a descriptor of the ensemble's statistical state rather than a direct component of the classical Gibbs free energy. Its role is to quantify the variability of peripheral NIS organization across the ensemble after exclusion of the interface atoms. Lower values of S'_Ψ indicate that the ensemble is concentrated within a smaller set of recurring peripheral NIS macrostates, whereas higher values indicate broader occupancy and greater peripheral-state heterogeneity. The guiding hypothesis, as introduced in Section 1.4, is that energetically favorable recognition is accompanied by reduced variability of peripheral NIS macrostates and therefore a lower measured PSI entropy. This is treated as a working hypothesis rather than an *a priori* guarantee, and the relevant question is whether S_Ψ separates favorable from unfavorable binding under controlled comparisons.

3.3 Normalization of S_Ψ

Intermolecular contact statistics over complete ensembles were utilized to construct the normalization constant, K , in Equation 3.1. Docking runs yield an inter-chain contact table listing residue-residue pairs (i, j) with an associated probability-like value treated here as an unnormalized "mass" of contact, $m_{i,j}$, defined from a C_α - C_α contact probability. To incorporate the biophysical affinity of different contact types, a contact class label was assigned to each amino acid (e.g., apolar, polar, charged-positive, charged-negative). For a contact between class (c_i) and (c_j) , a dimensionless scaling factor $\gamma(c_i, c_j)$ was applied yielding the contact class-weighted mass, $\tilde{m}_{i,j} = \gamma(c_i, c_j) m_{i,j}$. This mapping generally upweights contacts known to be favorable (e.g., apolar-apolar, polar-polar, and opposite charges) and downweights unfavorable combinations (e.g., apolar-polar, apolar-charged, and like charges) with representative values $\gamma(c_i, c_j) \in [0.61, 1.65]$, heuristically established from atomic contact statistics [74]. The enthalpy-aware contact-classes, weights, and brief descriptions are exhaustively provided in Appendix A, Table A.1.

Given an ensemble with nonzero contact masses, duplicate (i, j) entries were summed over all members of the run to obtain a single class-weighted mass per distinct contact pair. Aggregating over the resulting set of distinct pairs defines an ensemble-level weighted total contact mass, $M = \sum_{(i,j)} \tilde{m}_{i,j}$. Taking the cardinality of the set of distinct (i, j) pairs as Q , the following reciprocal relationship is defined as

$$\frac{1}{K} = \frac{M}{Q}, \quad (3.2)$$

M grows with consistent complementary interactions that concentrate on favorable contact modes shaped by hydrophobic packing and electrostatics. M/Q is sensitive to recurrent binding motifs, while a large Q with large M signals energetic concentrations that are diffuse along the receptor surface. Inserting Equation 3.2 into Equation 3.1

provides the explicit form

$$S_{\Psi} = -\left(\frac{Q}{M}\right) \sum_{j=1}^N \left(\frac{g(\mathcal{N}_j)}{\Omega}\right) \log_2 \left(\frac{g(\mathcal{N}_j)}{\Omega}\right). \quad (3.3)$$

Each macrostate, $\mathcal{N}_j$, contributes $(-K p_j \log_2 p_j)$ to S_{Ψ} . Subsequently, the relevant contributions are those from the $\mathcal{N}$ -states rather than conventionally with microstate contributions. This was due to microstates only considered through their multiplicities, $g(\mathcal{N}_j)$, and being otherwise indistinguishable, a strategic coarse graining of information. The unnormalized PSI entropy, S'_{Ψ} , monitors how many distinct peripheral NIS patterns are explored by favorable poses in $\mathcal{N}$ -space. Incorporating the normalization factor $K = Q/M$, where M is the class-weighted total contact mass and Q is the number of distinct contact pairs, results in an entropy per unit favorable contact enthalpy. The following regimes are described to provide an intuitive understanding for S_{Ψ} as a measure of the global thermodynamic landscape of binding events.

Regime I: low S_{Ψ} , low Q/M . The interface is dominated by a focused set of high-affinity contacts that recur across the ensemble, yielding small Q/M , and the emergent NIS patterns concentrate into a subset of macrostates such that S_{Ψ} is small.

Regime II: high S_{Ψ} , high Q/M . Here, the class-weighted contact mass is spread over many weak or sporadic residue-residue pairs such that Q/M is large; additionally, the ensemble populates a broad family of distinct $\mathcal{N}$ -states giving a large S_{Ψ} . This regime represents high degeneracy and polyspecific recognition, consistent with transient or hub-like interactions.

Regime III: intermediate S_{Ψ} , low Q/M . In this intermediate S_{Ψ} regime, the interface is built from strong recurrent contacts so Q/M remains small, yet S_{Ψ} is only moderately reduced, indicating the peripheral NIS patterns retain appreciable variability across binding poses. This corresponds to a landscape that is well-focused at

the interface with strong contacts yet is globally adaptable at the periphery, allowing multiple $\mathcal{N}$ -states to remain thermodynamically probable.

3.4 Collective Variables, Order Parameters, and Reaction Coordinates

The terms collective variable (CV), order parameter, and reaction coordinate are often used interchangeably in computational biology and biophysics, although they are not identical. Each represents a reduced description of a much higher-dimensional system, differing primarily in what information is retained and for what purpose. A useful analogy is the classical double pendulum shown in Figure 3.1a. Although the system can be written in Cartesian coordinates, its motion is more naturally described by two angular variables, θ_1 and θ_2 , that respect the geometric constraints. These angles are Lagrangian generalized coordinates, they replace a redundant description with a compact set of variables that still completely determines the admissible states of the constrained system [75]. The geometric constraints of the double pendulum are most informative when treated as an inverse problem, where the observed trajectories provide evidence of the physical conditions that confine motion to a reduced manifold, as shown in Figure 3.1b. By the same logic, trajectories in $\mathcal{N}$ -space provide information for inferring the biophysical constraints that govern complexation.

The analogy to biomolecular CVs is helpful, although it is not exact. Generalized coordinates are complete within the constrained configuration space, whereas biomolecular CVs are usually incomplete, many-to-one projections of the microscopic state. A distance, dihedral angle, contact number, or RMSD may isolate a physically meaningful mode of change, yet none of these generally specifies the full molecular configuration. Thus, CVs are best understood as deliberately chosen reduced descriptors rather than full replacements for the underlying coordinates. Within this framework, an order parameter is a measurable quantity used to distinguish states or quantify structural organization, often without implying that it fully captures the mechanism of transition. A reaction coordinate is more specialized, intended to de-

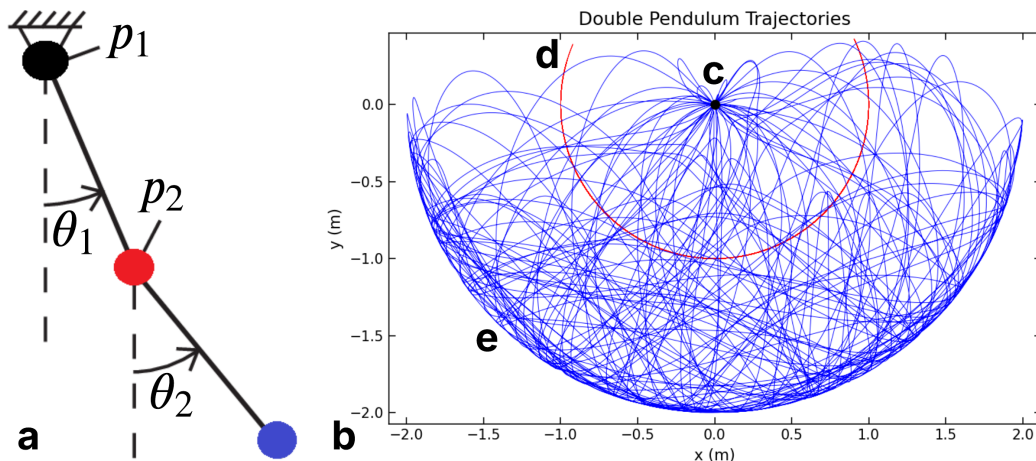


Figure 3.1: The classical double pendulum. The Lagrangian generalized coordinates θ_1 and θ_2 capture the full configuration space of the system. (b) The double pendulum trajectories during a simulation of 120 seconds. (c) Centered at (0,0) is the suspension point p_1 . (d) The red end, p_2 , of the first pendulum shows a structured trajectory over time where any point is equidistant to p_1 . (e) When monitoring the blue end of the second pendulum a standard example of chaotic dynamics is revealed. Given enough time the blue trajectories would fill in this region completely having sampled all possible configurations. How long will it take to visit all possible configurations? This is just the sampling problem from another vantage.

scribe progress along a transition pathway and to align with the slow dynamical mode governing barrier crossing [76]. A CV may serve as an order parameter, and an order parameter may sometimes approximate a reaction coordinate, though these roles are not equivalent. The central distinction is informational, generalized coordinates are complete on a constrained manifold, whereas CVs, order parameters, and reaction coordinates are projections that make otherwise intractable high-dimensional biomolecular processes interpretable, at the cost of losing information [70, 76].

3.5 Potential of Mean Force and Free Energy Landscapes

The potential of mean force (PMF) provides a reduced thermodynamic description of a molecular system by projecting the equilibrium ensemble onto one or more CVs. For a pair of CVs, $\boldsymbol{\xi} = (\xi_1, \xi_2)$, the corresponding two-dimensional free energy surface (FES) is obtained from the equilibrium probability density $P(\xi_1, \xi_2)$ through the Boltzmann inversion formula

$$F(\xi_1, \xi_2) = -k_B T \ln P(\xi_1, \xi_2) + C, \quad (3.4)$$

where k_B is the Boltzmann constant, T is the absolute temperature, and C is an arbitrary additive constant [77, 78]. All degrees of freedom (DoF) orthogonal to ξ are integrated out in this construction; subsequently, the PMF represents an effective FES rather than the full Gibbs free energy of association [79, 78]. In the reduced representation, minima identify thermodynamically favored macrostates, while elevated regions mark unfavorable regions of the (ξ_1, ξ_2) plane.

A two-dimensional PMF is especially useful when the process of interest is governed by coupled structural modes. In biomolecular systems, conformational rearrangements, partial ordering events, and binding-associated deformations are often not captured adequately by a single coordinate. This is because a one-dimensional projection may merge distinct substates or obscure the structure of transition regions. When two CVs are retained, the resulting free energy landscape may resolve distinct basins, anisotropic barrier regions, and preferred transition channels in a manner that is more faithful to the underlying ensemble [80, 77]. A low-dimensional PMF should be regarded as a thermodynamic map over the chosen observables rather than as a complete representation of the full configurational surface [77, 81]. The 1D PMF and 2D FES are constructed later in Chapter 6, extracted from apo MDM2 ensembles.

3.6 Relevant NMR Chemical Shift Theory

The NMR chemical shift is conventionally defined on the parts-per-million scale as

$$\delta = \frac{\nu - \nu_{\text{ref}}}{\nu_0} \times 10^6, \quad (3.5)$$

where ν is the observed resonance frequency of the nucleus, ν_{ref} is the resonance frequency of the reference compound, and ν_0 is the spectrometer frequency for that nucleus [82]. Because δ is referenced and normalized by ν_0 , the same nucleus in the

same local environment should yield the same chemical shift even when the absolute resonance frequency differs across instruments.

The chemical shift is determined by the local electronic environment of the nucleus. In an external magnetic field B_0 , the local field may be written as $B_{\text{loc}} = B_0(1 - \sigma)$, where σ is the shielding constant [83]. Variations in shielding arise from local structure and environment, including bonding geometry, hydrogen bonding, ring-current effects, electrostatics, and conformational state. In solution, the observed value reflects the isotropic average of the shielding tensor, $\sigma_{\text{iso}} = (\sigma_{xx} + \sigma_{yy} + \sigma_{zz})/3$ [83, 84]. Chemical shifts therefore serve as local structural observables that are sensitive to changes in residue-level geometry and local environment.

In this dissertation, experimentally assigned chemical shifts obtained from the biological magnetic resonance data bank (BMRB) are treated as reference observables on the standard ppm scale [85]. Computationally predicted shifts are interpreted on the same scale and are compared directly to experiment after atom identity and referencing are matched. The corresponding prediction residual is written as

$$\Delta\delta = \delta_{\text{pred}} - \delta_{\text{exp}}, \quad (3.6)$$

where δ_{pred} and δ_{exp} are the computationally predicted and experimental chemical shifts respectively. This residual quantifies the agreement between a structural model and the experimental local magnetic environment.

Chemical shifts are ensemble-level observables and should therefore be interpreted within the statistical mechanical framework introduced earlier in this chapter. If the system occupies microstates x_i with probabilities p_i , then the relevant predicted observable is the ensemble average

$$\langle\delta\rangle = \sum_i p_i \delta(x_i). \quad (3.7)$$

This relation is important in flexible systems, where the experimentally observed shift may correspond to a population-weighted average over multiple conformations rather than to any single structure [86, 87]. The theoretical role of chemical shifts in this dissertation is therefore to provide experimentally grounded, local observables that can be compared directly with structure-derived predictions from simulation. The specific SPARTA-based computational procedure that was used to generate the predicted chemical shifts is introduced later in Chapter 4.

CHAPTER 4: COMPUTATIONAL METHODS

4.1 Solvent Accessible Surface Area and Relative Solvent Accessibility

For each complex, per-residue solvent-accessible surface area (SASA) was computed from the atomic coordinates using FreeSASA [71]. In this geometric construction, SASA is the area traced by the center of a spherical solvent probe rolled over the atomic van der Waals surface, and FreeSASA implements both the Lee and Richards and the Shrake and Rupley approximations for this calculation. These SASA values were then converted to relative solvent accessibility (RSA) by normalization with the NACCESS residue maximum accessible surface area values [72],

$$\text{RSA}_i = \frac{\text{SASA}_i}{\text{ASA}_{\text{max},i}}. \quad (4.1)$$

See Appendix A, Table A.2 for the full set of ASA_{max} values. Residues with $\text{RSA}_i \geq 0.05$ were assigned to three chemical classes: apolar (A), charged (C), and polar (P). Interface residues were identified geometrically by a 5.0 Å heavy-atom distance cutoff between receptor Chain A and partner Chain B, and all such interacting residues were excluded for the RSA/SASA calculations.

4.2 Ensemble Generation

4.2.1 Molecular Docking

The docking software used to generate results was HADDOCK 3.0 (HADDOCK3) [88]. This method was selected due to its modular and open-source framework allowing for customization of parameters/energy functions, providing maximal control over conformational ensemble generation. The standard HADDOCK3 pipeline consists of 3 steps: 1) rigid-body docking; 2) semi-flexible refinement in torsion angle space; and

3) molecular dynamics simulation (MDS) in explicit solvent. Parameters were set to the defaults unless explicitly stated in this dissertation. Correspondent with the steps above, there are 3 energy-based HADDOCK scoring (HS) functions given by

$$HS = \begin{cases} 0.01 E_{vdw} + E_{elec} + E_{desolv} + 0.01 E_{AIR} - 0.01 BSA, & \text{Step 1} \\ E_{vdw} + E_{elec} + E_{desolv} + 0.1 E_{AIR} - 0.01 BSA, & \text{Step 2} \\ E_{vdw} + 0.2 E_{elec} + E_{desolv} + 0.1 E_{AIR}, & \text{Step 3} \end{cases} \quad (4.2)$$

where E_{vdw} and E_{elec} represent non-bonded van der Waals and Coulomb intermolecular energies respectively. E_{desolv} is an empirical desolvation term, BSA is the buried surface area upon complexation in $\text{\AA}^2$, and E_{AIR} is the restraint violation energy which expresses an agreement between experimental and back-calculated data. All energies were computed using the optimized potentials for liquid simulations (OPLS) force field [89], reported in units of kJ/mol.

4.2.2 Molecular Dynamics Simulation

All molecular dynamics simulations were performed using open-source molecular mechanics (OpenMM), a high performance molecular simulation toolkit designed for both routine biomolecular simulation and method development [90]. OpenMM provides a layered software architecture in which a high-level application interface is coupled to a lower-level programmable library, allowing the same framework to be used either as a stand-alone engine for conventional simulations or as a flexible platform for constructing custom molecular modeling workflows. This design enables direct specification of molecular topology, force-field parameters, boundary conditions, integrators, thermostats, barostats, and constraints in a unified environment.

A major strength of OpenMM is that it combines broad force-field interoperability with hardware-accelerated numerical performance. The software supports common molecular simulation inputs and force-field ecosystems used for proteins, nucleic acids,

lipids, solvents, and small molecules, while also exposing programmable force objects and custom energy terms that permit the implementation of user-defined Hamiltonians and enhanced-sampling protocols [91]. OpenMM further provides multiple computational platforms, including CPU and GPU backends, allowing simulations to be executed efficiently on modern heterogeneous hardware. This makes the package particularly well suited for large-scale biomolecular simulations in which both numerical throughput and methodological flexibility are important.

Within a simulation workflow, OpenMM propagates the atomic coordinates by numerically integrating Newtonian equations of motion under the potential energy function defined by the selected molecular mechanics model. Accordingly, the force on atom i is obtained from the gradient of the potential energy,

$$\mathbf{F}_i = -\nabla_i U(\mathbf{r}_1, \mathbf{r}_2, \dots, \mathbf{r}_N), \quad (4.3)$$

and the time evolution of the system is generated by an integration algorithm such as Langevin or Verlet dynamics [92, 93]. OpenMM acts as both a program for trajectory generation, and general computational framework for evaluating energies, forces, and thermodynamic observables from atomistic models. Its combination of efficiency, extensibility, and close integration with the Python ecosystem has made it a widely used platform for molecular biophysics and computational chemistry [91, 90].

4.3 S_Ψ Computation

For the purpose of establishing the methods, the exemplar model system was the WW domain (PDB ID 2LTW, experimental method NMR), a small highly flexible 36-residue 4.15 kDa globular protein characterized by two conserved tryptophan amino acids flanking each end. WW domains serve as a model system for understanding the thermodynamics and kinetics of the β -sheet fold [48]. The protein receptor was experimentally complexed with a 14-residue peptide, GESPPPPYSRYPMD. The first of

30 NMR conformers was selected, the protein and peptide were structurally separated then prepared and protonated using PDB2PQR [94]. Next, independent structures were equilibrated in explicit TIP3P solvent using OpenMM [90] to decorrelate the structures from the experimentally bound conformations. The active residues were defined as (Tyr19, Leu21, His23, Gln26, and Trp30), that are discontinuous in sequence while being spatially contiguous, forming the known XP and Tyr grooves of the WW domain [49]. There were 3000 rigid-body conformations generated where the top-ranked 400 (according to Equation 4.2) were retained for the next stage. The second step of the docking protocol introduced flexibility to the interacting partners through a three-step MDS-based refinement in order to optimize interface packing. After the completion of Step 2, the top-ranked 400 models were further refined to remove all complexes having an interface root mean square deviation (iRMSD) greater than 10 Å, resulting in 227 complexes in the exemplar ensemble. Dual energy-distribution plots are provided in Appendix B, Figure B.1 for the generated ensemble described.

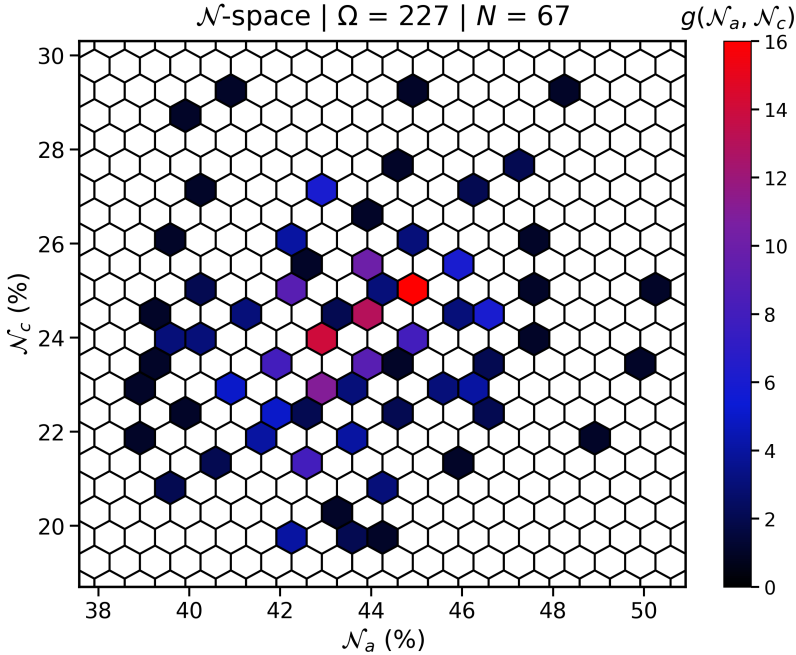


Figure 4.1: The $\mathcal{N}$ -space hexbin plot. Each hexagon represents a unit cell in NIS state space. After RSA thresholding and explicit interface exclusion, the exemplar ensemble contained $\Omega = 227$ microstates that partitioned into $N = 67$ distinct macrostates, with color indicating the multiplicity, $g(\mathcal{N}_a, \mathcal{N}_c)$.

The notable observation at this time was the self-organizing behavior of the NIS apolar and charged proportions ($\mathcal{N}_a$ and $\mathcal{N}_c$) as shown in Figure 4.1. The $(\mathcal{N}_a, \mathcal{N}_c)$

points span a two-dimensional space which contains 227 microstates ($\Omega = 227$) that assembled into $N = 67$ macrostates. The number of microstates that exist in a given macrostate is quantified by the multiplicity function, $g(\mathcal{N}_a, \mathcal{N}_c)$. With $\mathcal{N}$ -space defined, S_Ψ is then calculated using Equation 3.3.

4.4 Chemical Shift Prediction

To compare structural models against experiment, backbone chemical shifts were back-calculated from atomic coordinates using the synth-nmr python package [95]. This is a SPARTA-style method that combines residue-specific random-coil reference shifts, secondary-structure-dependent offsets inferred from backbone dihedral angles, aromatic ring-current corrections for proton nuclei, and an additive stochastic term [96]. For nucleus type a at residue i , the predicted chemical shift is

$$\hat{\delta}_i^{(a)} = \delta_{\text{rc}}^{(a)}(r_i) + \Delta_{\text{ss}}^{(a)}(s_i) + \Delta_{\text{ring},i}^{(a)} + \epsilon_i^{(a)}, \quad (4.4)$$

where $\delta_{\text{rc}}^{(a)}(r_i)$ is the residue-specific random-coil reference value for residue r_i , $\Delta_{\text{ss}}^{(a)}(s_i)$ is the secondary structure-dependent correction associated with local conformational class s_i , $\Delta_{\text{ring},i}^{(a)}$ is the aromatic ring-current contribution, and $\epsilon_i^{(a)}$ is an additive Gaussian noise term. The calculation begins from a residue-dependent random-coil baseline, then applies structure-dependent corrections that account for local backbone geometry and tertiary-structure effects arising from nearby aromatic side chains.

This formulation is especially useful because it places the predicted observable in direct correspondence with experiment. Whereas NMR measurement yields the experimental chemical shift, $\delta_{i,\text{exp}}^{(a)}$, the synth-nmr approach provides the corresponding structure-conditioned estimate, $\hat{\delta}_i^{(a)}$. Agreement between model and experiment may be assessed at the residue level through the prediction residual

$$r_i^{(a)} = \hat{\delta}_i^{(a)} - \delta_{i,\text{exp}}^{(a)}, \quad (4.5)$$

or more globally, the root-mean-square error

$$\text{RMSE}^{(a)} = \left[\frac{1}{N_a} \sum_{i=1}^{N_a} \left(\hat{\delta}_i^{(a)} - \delta_{i,\text{exp}}^{(a)} \right)^2 \right]^{1/2}, \quad (4.6)$$

where N_a is the number of observations available for nucleus type a .

The synth-nmr method functions as a forward model that maps candidate protein structures to expected NMR chemical shifts, enabling quantitative comparison of computed conformations or structural ensembles against experiment. In particular, the predicted backbone and $C\beta$ shifts provide a residue-resolved probe of local conformation and environment because these nuclei are sensitive to backbone geometry, secondary structure, and the nearby chemical environment [96]. In the present implementation, the empirical fallback should be interpreted as a reduced SPARTA-style approximation rather than a full reproduction of the original SPARTA model.

CHAPTER 5: APPLICATION OF S_Ψ

5.1 Emergence of $\mathcal{N}$ -States Under Diverse Conditions and Conformations

Four conformations of the Smad7 peptide (GESPPPPYSRYPMD) were generated as shown along the top row of Figure 5.1. The first conformation (SMDA) was taken from the NMR ensemble (PDB ID 2LTW; conformer 1/30). The second conformation (SMDB) was predicted using AlphaFold2 (AF2). The third conformation (SMDC) resulted from performing an equilibration for 500 ps in explicit solvent on SMDB. The fourth peptide conformation (SMDD) was predicted with AlphaFold3 (AF3). This subset acts as an initial probe into whether the emergent $\mathcal{N}$ -states are sensitive to peptide conformation. For all 4 conformations, the same protocols to dock Smad7 on the WW receptor were executed with two parameter sweeps of: *i*) the number of rigid body models to generate in Step 1 with increments of (10000 and 25000); and *ii*) the number of rigid models retained (n_{ret}) for Step 2 semi-flexible refinement were varied in maximal increments of (200, 1000, 5000, 10000).

The primary finding was that the organization into a subset of macrostates continues under the parameter sweeps. When the number of rigid-body models in Step 1 was 10000, the mean number of $\mathcal{N}$ -states across all Smad7 peptide conformations was: 24.25 ± 5.72 when $n_{ret} = 200$, 65.50 ± 12.60 if $n_{ret} = 1000$, 106.75 ± 16.68 at $n_{ret} = 5000$, and 145.00 ± 19.17 given $n_{ret} = 10000$. This trend persisted when the number of rigid-body models was 25000. The smaller mean number of $\mathcal{N}$ -states shows that the self-organization into subsets of macrostates occurs across the varied conformations, parameters, and conditions as exemplified in Figure 5.1a. Figure 5.1b shows the emergence of a dominant mode for each Smad7 ensemble. For a complete set of $\mathcal{N}_a$ and $\mathcal{N}_c$ distributions see Appendix B, Figures B.2 and B.3.

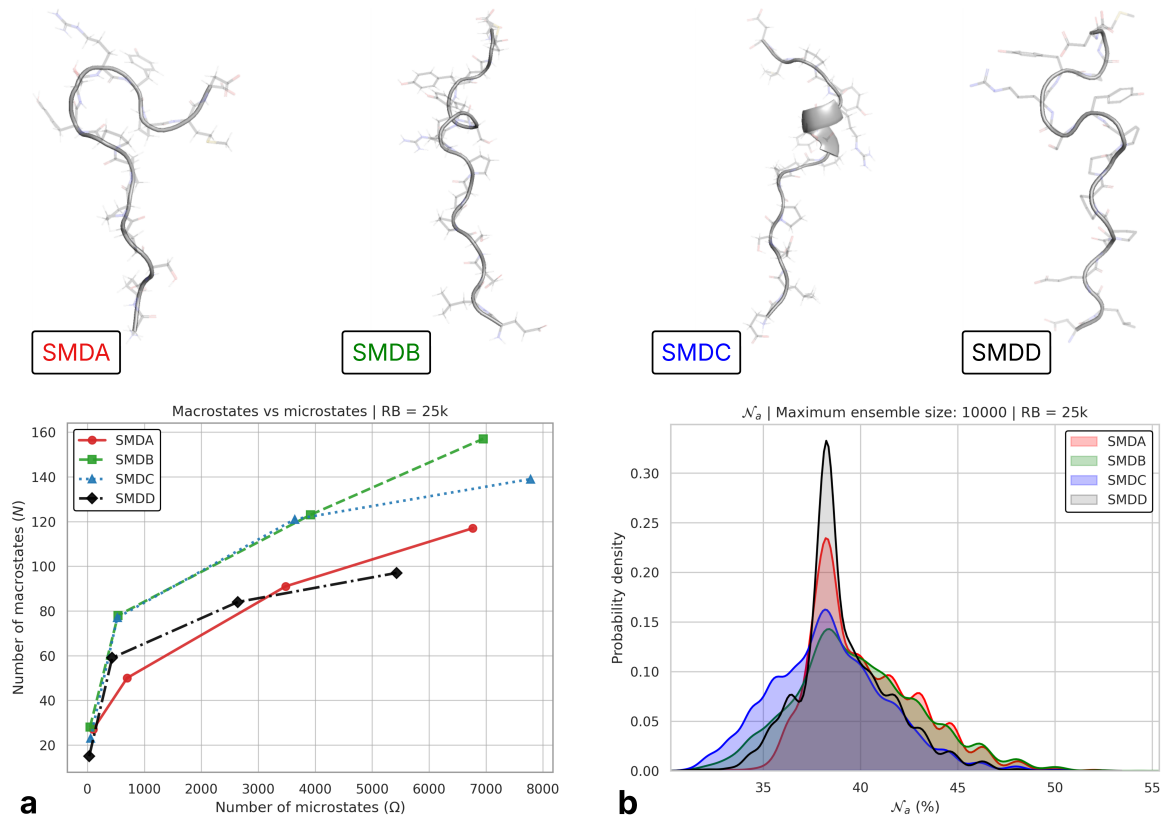


Figure 5.1: An overview of $\mathcal{N}$ -state properties while varying: the Smad7 peptide conformations, number of rigid-body models, and number of models refined by semi-flexible MDS. (top row) Conformations are shown with N-termini on bottom and C-termini on top where backbone geometries are depicted in silver. (SMDA) Was taken from the NMR ensemble 2LTW without modification. (SMDB) This conformation was predicted with AF2. (SMDC) The prior AF2 conformation was equilibrated with MDS where a partial helix emerges. (SMDD) The final conformation was generated using AF3. (a) The number of macrostates, N , vs the number of microstates (Ω) where the number of rigid-body (RB) models was 25000. As the number of microstates increases, the number of emergent macrostates follow a logarithmic trend, reaching a maximum of $N = 157$ for SMDB. (b) Kernel density estimations of $\mathcal{N}_a$ proportions for the four Smad7 peptides. A dominant $\mathcal{N}_a$ mode emerges for all 4 peptide candidates, attaining a higher amplitude for the SMDA and SMDD partners.

Next, the distributions of $\mathcal{N}$ -states for a diverse set of WW domain-peptide complexes were evaluated to determine if common macrostates emerge. Step 3 from the standard HADDOCK pipeline is now included, executing MDS with explicit TIP3P solvent. The general expectation from the inclusion of Step 3 is that as the system relaxes there is an expected general increase in S_Ψ across all systems due to ther-

mal fluctuations. This is expected to occur along with an increase in the number of $\mathcal{N}$ -states due to an increased sampling of binding conformations, reported in Tables 5.1 and 5.2. Two sets of peptides were created, each used to generate ensembles of complexes with the WW domain. The first set, $\{PY\}$, contains 16 proline-rich peptides with motifs PPxY (9), LPxY (3), and PPGW (4). The second set, $\{RD\}$, is comprised of 16 randomly generated sequences where peptide lengths varied between 10 and 15. Each structure of $\{RD\}$ was predicted using AF3 then both sets were decorrelated from initial states by equilibrating with short MDS prior to docking protocols. Full sequences for the $\{PY\}$ and $\{RD\}$ sets are provided in Appendix A, Tables A.3 and A.4 respectively. The number of Step 1 rigid-body models was 10000, and the number of retained models (n_{ret}) between steps was held fixed at $n_{ret} = 200$.

Table 5.1: $\mathcal{N}$ -state properties between Step 2 (semi-flexible refinement) and Step 3 (MDS in explicit solvent) extracted from ensembles of protein-peptide complexes. Sixteen proline-rich peptides with motifs PPxY, LPxY, and PPGW were complexed with the WW domain (PDB ID 2LTW). The terms $\Delta\mathcal{N}$, ΔS_Ψ , and ΔK indicate the Step 3 minus Step 2 changes in the number of $\mathcal{N}$ -states, PSI entropy (S_Ψ), and the normalization constant $K = Q/M$ respectively.

	PPxY									LPxY			PPGW			
	2DJY	2KXQ	2LB1	2LB2	2LTV	2LTY	2LAW	SMDA	SMDC	2EZ5	7LP5a	2MPT	1ZCN	2ZQS	6JIZ	7BQF
$\Delta\mathcal{N}$	8.00	4.00	3.00	6.00	4.00	14.00	12.00	6.00	8.00	2.00	-13.00	8.00	-3.00	5.00	-7.00	-6.00
ΔS_Ψ	0.22	-0.84	0.08	-0.29	0.22	0.20	0.06	0.16	0.22	-0.97	0.02	-0.48	0.51	-0.55	-0.30	-1.63
ΔK	0.02	-0.16	0.01	-0.06	0.01	-0.01	0.00	0.03	0.02	-0.15	0.02	-0.10	0.08	-0.15	-0.04	-0.33

Table 5.2: $\mathcal{N}$ -state properties between Step 2 (semi-flexible refinement) and Step 3 (MDS in explicit solvent) extracted from ensembles of protein-peptide complexes. Sixteen random sequences were predicted using AF3 then equilibrated with MDS in explicit TIP3P solvent (RD01 to RD16). The terms $\Delta\mathcal{N}$, ΔS_Ψ , and ΔK indicate the Step 3 minus Step 2 changes in the number of $\mathcal{N}$ -states, PSI entropy (S_Ψ), and the normalization constant $K = Q/M$ respectively.

	RD01	RD02	RD03	RD04	RD05	RD06	RD07	RD08	RD09	RD10	RD11	RD12	RD13	RD14	RD15	RD16
$\Delta\mathcal{N}$	15.00	8.00	6.00	9.00	9.00	13.00	-6.00	1.00	4.00	1.00	14.00	2.00	-1.00	14.00	4.00	7.00
ΔS_Ψ	-1.38	0.23	-1.13	-0.85	0.21	-0.96	0.23	0.10	-0.18	-0.15	-0.45	-1.65	-0.62	1.11	-0.01	0.34
ΔK	-0.44	0.01	-0.24	-0.17	0.02	-0.32	0.04	0.02	-0.06	-0.03	-0.22	-0.29	-0.09	0.11	-0.02	0.04

Table 5.1 and Table 5.2 summarize how the number of $\mathcal{N}$ -states, the normalized PSI entropy, and the normalization constant $K = Q/M$ evolve across systems. The

number of $\mathcal{N}$ -states more frequently increased than decreased from Step 2 to Step 3, consistent with added sampling during solvent relaxation. However, this trend was not universal as the condition $\Delta\mathcal{N} \geq 0$ held for 26/32 ensembles, with decreases observed for 1ZCN, 6JIZ, 7BQF, 7LP5a, RD07, and RD13. The S_Ψ levels at Step 2 and Step 3 separate the two peptide sets. At Step 2, $S_\Psi = 2.95 \pm 1.09$ for $\{PY\}$ compared to 3.94 ± 1.24 for $\{RD\}$. At Step 3, $S_\Psi = 2.74 \pm 0.98$ for $\{PY\}$ compared to 3.62 ± 1.10 for $\{RD\}$. In contrast, the unnormalized Shannon term (S'_Ψ with $K = 1$) does not separate $\{PY\}$ vs $\{RD\}$. This indicates that the PY-RD separation is introduced by the $K = Q/M$ scaling rather than by macrostate diversity alone. For full tables and metrics, see Appendix A, Tables A.5 and A.6.

5.2 Robustness of PSI Entropy Across Varied Systems

Two new receptors were included to monitor the $\mathcal{N}$ -space trends already established with the WW domain. The first receptor was the murine double minute 2 (MDM2) protein, a vital regulator of the tumor suppressing protein p53. The MDM2 structure (PDB ID 4HFZ, lidless MDM2) was resolved by X-ray diffraction (resolution 2.69 Å) and originally complexed with an N-terminal p53 segment ETFSDLWKLLP. The second receptor was the postsynaptic density-95/discs large/zonula occludens-1 (PDZ) protein (PDB ID 1ZUB) attributed with the organization of neurotransmitters and signaling pathways at neuronal synapses. The structure of 1ZUB was resolved by NMR where the first of 20 conformers was selected. The 1ZUB protein was complexed with an ELKS1b C-terminal peptide, CDQDEEEGIWA. The receptor and peptide of 4HFZ/1ZUB were separated then decorrelated from their initial conformations by: 1) equilibrating the receptors in explicit solvent using OpenMM; and 2) peptide conformations were diversified by performing a 20 ns simulation on each in TIP3P solvent then sampling every 2 ns resulting in 10 conformations for each peptide candidate.

The active sites provided to HADDOCK3 were expanded compared to the exemplar ensemble in Chapter 4.3 with the WW domain (5 active residues). For MDM2 the

active site was set to 54-99 (45 residues), containing the known binding groove that is lined with apolar amino acids. When peptides were docked to the PDZ protein, active sites were defined as 623-639 and 681-695 (a total of 32 residues) covering the binding groove formed by a carboxylate binding loop, βB beta strand, and αB helix. A cross-fertilization procedure was performed to test the working assumption established in the Methods following Equation 3.1 (see Appendix A, Table A.7 for a summary of pairings). This can be succinctly stated as: given a biologically cognate (proper) and non-cognate (improper) peptide candidate docked to the same protein receptor, a favorable candidate should measure a lower S_Ψ compared to candidates that do not bind well. The procedure involved 2 phases where each phase utilized a different protein receptor, a total of 40 HADDOCK3 runs (10 proper and 10 improper runs per phase). The ambiguous interaction restraints (AIRs) were removed from Step 3 during MDS. This acts as a molecular stress test for favorable interactions; consequently, a decrease in Ω of final ensembles was expected.

Proper and improper ensembles are summarized separately at the phase level. The unnormalized Shannon term (S'_Ψ with $K = 1$) is computed per run from the macrostate probability distribution $\{p_i\}$, then averaged across 10 replicas. The scaling coefficient, K , is computed from contact statistics by pooling the nonzero contact pairs (edges) and their contact weights over all replicas in the phase. The reported phase-level PSI entropy is then $S_\Psi^{\text{phase}} = K_{\text{phase}} \langle S'_\Psi \rangle_{\text{runs}}$, ensuring that the normalization reflects the collective binding-site contact structure of the phase rather than an average of per-run normalizers. Pooling is appropriate because both Q and M depend on the support and weights of the contact network, estimated more stably at the phase level than from any single replica. Averaging per-run ratios would conflate run-to-run sampling variability in contact support with the phase-level normalization.

Across 10 replicas per phase, ensemble sizes and macrostate counts varied by receptor and condition. Despite replica-to-replica measurement uncertainty, the order-

ing $S_{\Psi}^{\text{phase}}(\text{proper}) < S_{\Psi}^{\text{phase}}(\text{improper})$ was preserved for both receptors. For 4HFZ, the proper ensembles had $\Omega = 47.6 \pm 32.0$ and $N = 42.2 \pm 26.4$, while the improper ensembles had $\Omega = 50.1 \pm 19.5$ and $N = 46.1 \pm 16.4$. For 1ZUB, the proper ensembles had $\Omega = 63.1 \pm 17.7$ and $N = 58.3 \pm 15.3$, while the improper ensembles had $\Omega = 24.7 \pm 8.8$ and $N = 23.2 \pm 7.2$. Despite this variability, the phase-level S_{Ψ} values followed the working assumption under matched protocols, as shown in Figure 5.2a. For the 4HFZ receptor phase, $S_{\Psi}^{\text{phase}} = 0.5086$ (proper) versus $S_{\Psi}^{\text{phase}} = 0.6489$ (improper), corresponding to $\Delta\% = 21.6\%$. For the 1ZUB receptor phase, $S_{\Psi}^{\text{phase}} = 0.3986$ (proper) versus $S_{\Psi}^{\text{phase}} = 0.6969$ (improper), with $\Delta\% = 42.8\%$. Here, $\Delta\% = (1 - S_{\Psi}^{\text{proper}}/S_{\Psi}^{\text{improper}}) \times 100$.

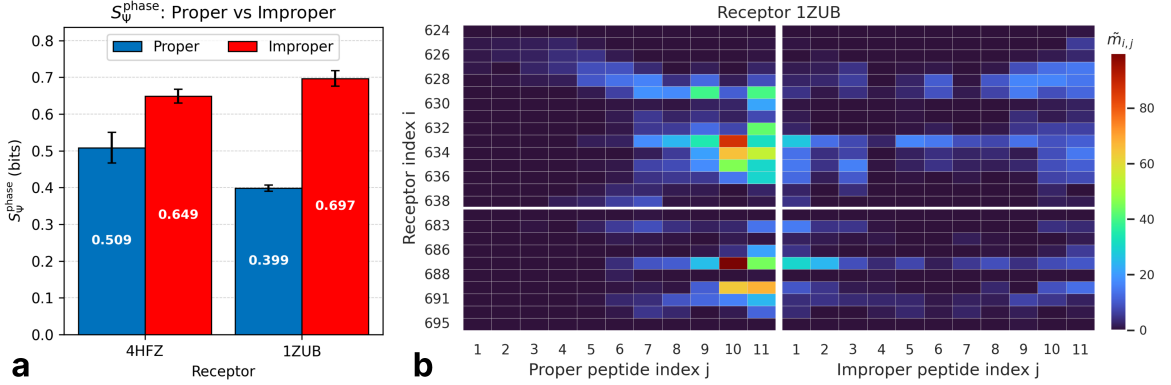


Figure 5.2: An overview for the cross-fertilization procedure. (a) Under matched protocols, proper ensembles exhibit lower S_{Ψ} than improper ensembles for both receptors. Note, the separation $S_{\Psi}^{\text{phase}}(\text{proper}) < S_{\Psi}^{\text{phase}}(\text{improper})$ persisted even when considering uncertainty (shown by error bars). (b) The enthalpy-weighted contact masses with an upweighting of spatially-dense favorable contacts for the proper ELKS1b peptide engagement. This exemplifies Regime 1 with low S_{Ψ} and low Q/M .

The enthalpy-weighted contact maps shown in Figure 5.2b refine this view by spatially localizing the interactions of proper ensembles. Appendix B, Figures B.4 and B.5 display cross-fertilization contact map probabilities, $\rho(i, j)$, contact-class weighted probability maps, $\tilde{m}_{i,j}$, and the ensemble-level net contact-class modulation, $\sum_{\Omega}(\gamma(c_i, c_j) - 1.0) \cdot m_{i,j}$, for 4HFZ and 1ZUB receptor phases. For the ELKS1b C-terminal peptide engagement, the contact class-weighted mass, $\tilde{m}_{i,j}$, collapses to a

compact patch of favorable contacts in the canonical groove, explicitly upweighted by $\gamma(c_i, c_j)$. This spatial focusing exemplifies Regime 1 (low S_Ψ and low Q/M), where a small subset of contacts carries a disproportionate share of the total engagement. The contact masses that dominate $\tilde{m}_{i,j}$ also dominate the contributions to S_Ψ ; subsequently, Figure 5.2 links reduced PSI entropy directly to a concentrated pattern of biophysically favorable contacts on the active site. The improper ensemble in Figure 5.2b falls into Regime 2 with elevated Q/M and larger S_Ψ per unit contact density. For detailed cross-fertilization calculations of phase-level statistics and associated error see Appendix C, Calculations C.1 and C.2 respectively.

Pooling all replicas within each phase, the phase-pooled $\mathcal{N}$ -space densities provide a distribution-level view of peripheral organization under matched protocols (see Appendix B, Figures B.6-B.7). Because Ω differs across conditions, both the absolute effective macrostate count ($N_{\text{eff}} = 2^S$) and its size-normalized fraction (N_{eff}/Ω) are used to distinguish changes in probability mass. For 4HFZ, the proper ensemble is more concentrated in $\mathcal{N}$ -space, with a smaller effective number of macrostates ($N_{\text{eff}} = 220.5$ for proper versus 258.2 with improper). For 1ZUB, the dominant $\mathcal{N}$ -space mode is preserved while the normalized effective state fraction decreases ($N_{\text{eff}}/\Omega = 0.481$ for proper versus 0.629 with improper). These pooled distributions show that reduced S_Ψ^{phase} coincides with a tightening of peripheral state occupancy in a receptor-dependent manner. For the corresponding definitions of N_{eff} and N_{eff}/Ω under phase pooling see Appendix C, Calculations C.3.

5.3 Experimental Validation of Dominant $\mathcal{N}$ -States

To investigate whether the emergence of dominant $\mathcal{N}$ -space modes is a general phenomenon independent of *in silico* docking, ensembles of experimentally resolved structures were analyzed. The goal is to identify conserved surface configurations independent of simulation protocols. Such convergence of NIS properties with specific interfaces would suggest an evolutionary selection pressure. This link between

interfacial properties and selection pressure has been discussed previously [28, 30], though not through the lens of NIS composition. For demonstration purposes, the WW domain is again used as a model system. The PDB RESTful API was utilized in Python to extract all structures with Pfam Protein Family (PF00397) annotations related to the WW domain. This query resulted in a total of 206 entries resolved from 4 experimental techniques: Solution NMR (107), X-ray Diffraction (96), Electron Microscopy (2), and Solid-State NMR (1). The 206 entries were spread over the following organisms: *Homo sapiens* (166), *Mus musculus* (25), *Rattus norvegicus* (4), *Drosophila melanogaster* (3), *Saccharomyces cerevisiae* (3), and others (5). Further refinement to all monomer WW domains complexed with small molecules, ligands, and peptides (including engineered molecules) resulted in 36 entries, 34 from NMR and 2 from X-ray experiments (see Appendix A, Table A.8) that were then pooled into a meta-ensemble of $\Omega = 657$ microstates (complexes). An alignment (reference structure 2LTW, Figure 5.3a) followed by a superposition is presented in Figure 5.3b.

The multi-colored ribbons shown in Figure 5.3b exemplify a variety of partners in terms of both binding poses and amino acid composition. This conformational messiness was intentionally curated as the expected emergence of a dominant $\mathcal{N}$ -state should occur along with statistical outliers and intrinsic conformational variance. Using the same process as previous sections, the corresponding $\mathcal{N}$ -space was constructed from the experimental meta-ensemble. A kernel density estimation was performed for the variables $\mathcal{N}_a$ and $\mathcal{N}_c$, spanning a two-dimensional continuous $\mathcal{N}$ -space. The bivariate distribution of all 657 microstates is presented in Figure 5.4a where a dominant mode in $\mathcal{N}$ -space is confirmed by a dense area of probability. The discrete $\mathcal{N}$ -space is shown in Figure 5.4b where 3 regions of higher multiplicities are revealed. There are three well-known classes of WW domains based on peptide recognition [49]; accordingly, the meta-ensemble was trifurcated and the bivariate distributions were calculated for each class (Figures 5.4c-5.4e), see Appendix B, Figure

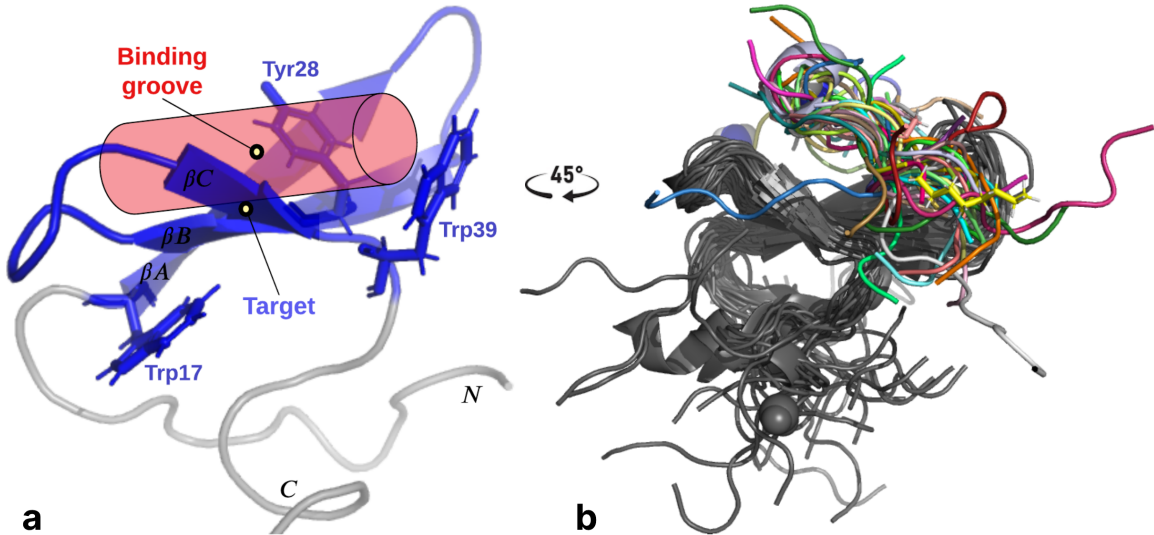


Figure 5.3: The WW domain side (a) and front (b) views. (a) The binding groove of the WW domain exists within the blue region bound by two highly conserved tryptophans (Trp17 and Trp39). Taking on the activated conformational state, the Tyr28 and Trp39 amino acids exhibit upward orientations relative to the beta sheet, likely due to π effects which are known contributors of favorable binding enthalpy. (b) An alignment and superposition of all 36 PDB entries (34 NMR, 2 X-ray diffraction). The varying WW receptors (gray) are complexed with their corresponding peptide, ligand, or small molecule (multicolored ribbons).

B.8 for full plots with axis labels. A dominant mode was observed for both the WW family as a whole and each individual class. These insights preliminarily support the hypothesis that as globular protein-peptide systems tend towards specificity over evolutionary timescales, the NIS properties converge towards a dominant mode in $\mathcal{N}$ -space, phenomena that are driven by evolutionary selection pressure.

The effective macrostate fraction, $N_{\text{eff}}/\Omega = 0.246$ ($N_{\text{eff}} = 161.8$), is consistent with state occupancy being concentrated rather than spreading uniformly across the experimental space. Notably, $N = 234$ distinct $\mathcal{N}$ -states (macrostates) were observed where 103 were singletons ($g_i = 1$), while the maximum macrostate occupancy reached $g_{\text{max}} = 20$. This heavy-tailed occupancy profile indicates that most states are sparsely sampled whereas a small subset repeatedly concentrates probability mass into dominant modes. Analysis of the $g_{\text{max}} = 20$ macrostate indicates that its membership is not restricted to a single experimental ensemble, it is partitioned across

multiple PDB entries (e.g., 2EZ5 and 2M3O); furthermore, 131 macrostates ($\sim 56\%$) contain microstates from more than one NMR ensemble, confirming that $\mathcal{N}$ -space configurations are robust to experimental determinations over 15 years. These results establish a consistent dominant-mode picture in $\mathcal{N}$ -space from energy-directed *in silico* ensembles to experimentally resolved structures. WW-domain destabilization is framed here as a peripheral design strategy based on surface modulation rather than direct alteration of the binding groove. By using small molecules or ligands to perturb NIS properties and drive NIS proportions away from dominant-mode regions in $\mathcal{N}$ -space, ensemble populations are expected to redistribute, and tumor-promoting YAP-mediated nuclear signaling is expected to attenuate [50, 51].

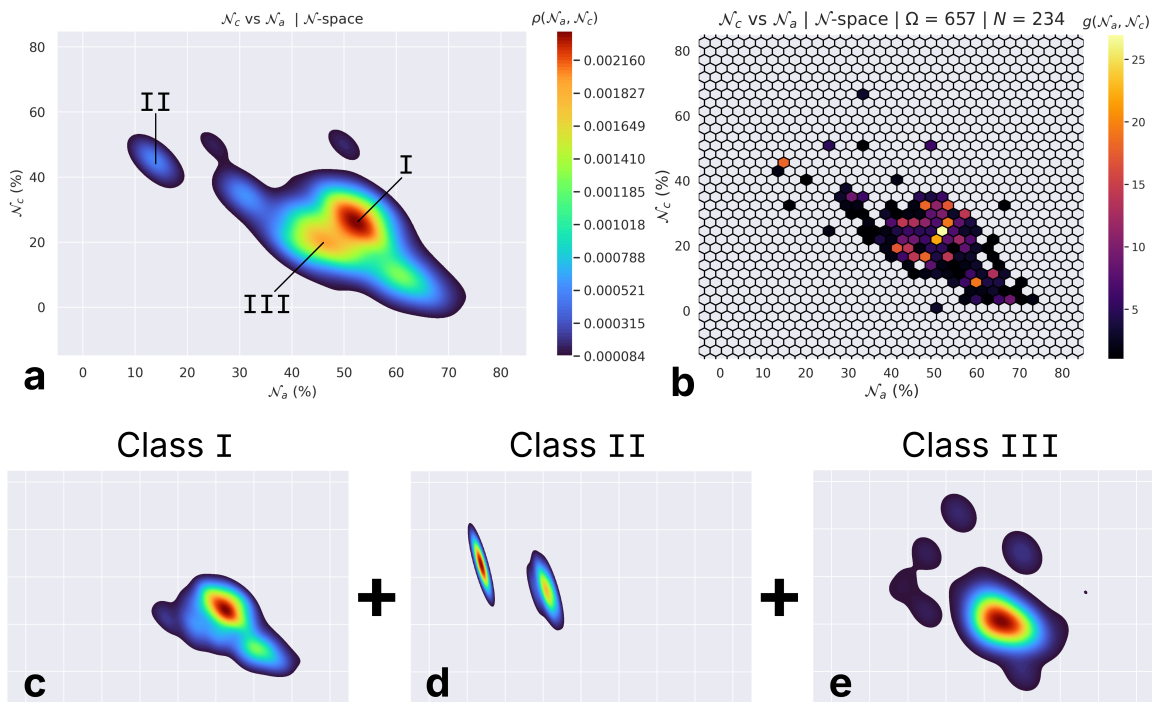


Figure 5.4: Bivariate distributions derived from a meta-ensemble of varying experimental WW domains that were complexed with small molecules, ligands, and peptides. (a) The continuous $\mathcal{N}$ -space containing all 657 curated experimental protein-partner complexes. It is readily apparent that a dominant mode of the bivariate distribution exists, shown by red/dark-red regions. (b) The discrete $\mathcal{N}$ -space (hexbin plot) explicitly shows 3 general regions of high multiplicity across all microstates. Trifurcating the meta-ensemble based on known classification, the corresponding bivariate distributions are shown in (c-e). The dominant class contributions to the meta-ensemble are indicated by (I, II, and III) in (a).

CHAPTER 6: A MOLECULAR-SCALE MECHANISM FOR P53 ABUNDANCE UNDER OSMOTIC STRESS

6.1 A Comparison of Computational and Experimental Chemical Shifts

The first objective is to validate simulated dynamics by contrasting computational against experimental chemical shifts. All structures were prepared using PDBFixer [97] and protonated with PDB2PQR [98]. Conventional MDS was performed for the human MDM2 N-terminal p53-binding domain (PDB 1Z1M, lid residues 1-24 and protein core residues 25-118) in OpenMM (CUDA, mixed precision) [91]. The simulation configuration included the Amber ff14SB force field [99], explicit TIP3P-FB waters [100], particle-mesh Ewald electrostatics [101], a 1.0 nm nonbonded cutoff, and hydrogen bonds constrained. Each trajectory used the protocol: *i*) energy minimization to convergence; *ii*) a 50 ps canonical (NVT) warm-up at 1 fs step size; *iii*) 1 ns NVT equilibration; *iv*) 2 ns isothermal-isobaric (NPT) equilibration; *v*) a 1 ns microcanonical (NVE) diagnostic segment to monitor energy drift and numerical stability [102]; and *vi*) a 250 ns NPT production run at 2 fs time steps, using a Langevin-middle integrator at 300 K [103], and Monte Carlo barostat that maintained isotropic mechanical pressure (1 atm). Isotonic (physiological conditions ≈ 300 mOsm L⁻¹), hyperosmotic (≈ 388 mOsm L⁻¹), and supraosmotic (≈ 690 mOsm L⁻¹) systems were built with 0.15 M NaCl and sorbitol. Raising total solute concentration in this way lowers solvent activity and mimics the osmotic pressure increase expected when macromolecules crowd. Sorbitol is uncharged, so osmolarity was tuned without a charged cosolute that would add long-ranged electrostatic effects beyond those already present from protein and ions. All structures were cleaned using PDBFixer and protonated using PDB2PQR. For each osmotic condition, 24 NMR conformers

acted as initial structures each having 4 replicas (96 trajectories per condition, 288 total). Total simulation time summed to $\sim 24 \mu\text{s}$ per condition and $\sim 72 \mu\text{s}$ overall.

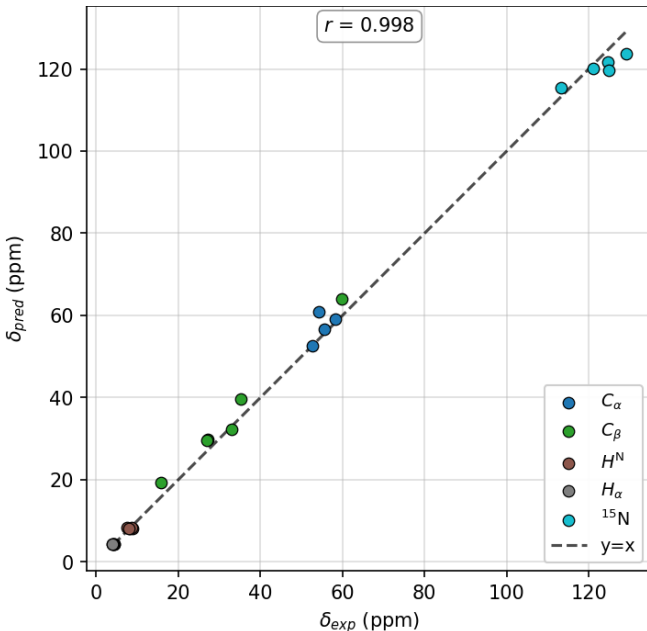
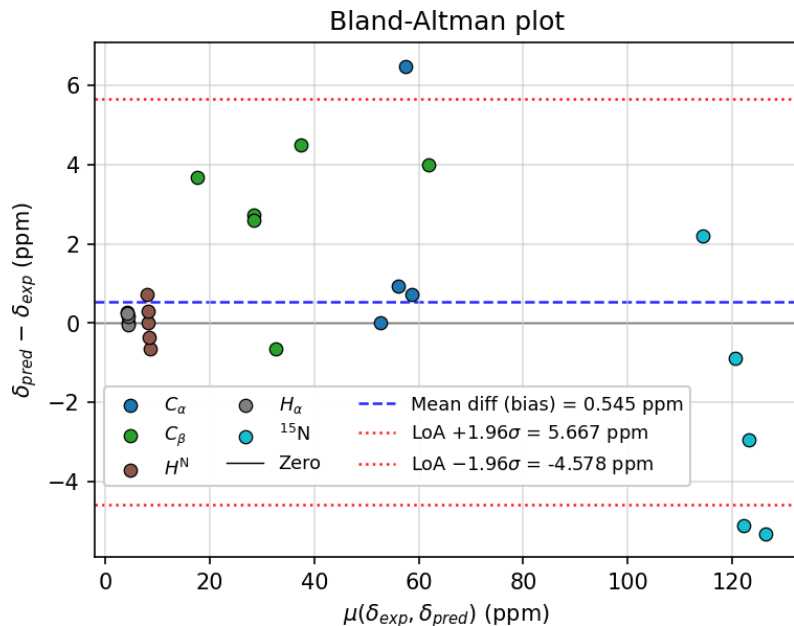


Figure 6.1: Isotonic ensemble validation of N-terminal lid-base chemical shifts. Predicted shifts are plotted against Biological Magnetic Resonance Bank (BMRB) 6612 experimental shifts for residues 19 to 24. Each point is one matched (residue, atom) pair and colors indicate atom type (^{15}N , H^N , H_α , C_α , C_β). The dashed line is the identity ($y = x$), $r = 0.998$. Points lie close to that line, so the isotonic ensemble reproduces the experimental shift pattern for the matched nuclei.

The normalized autocorrelation function (ACF) measures how strongly a CV linearly correlates with itself after a fixed time lag along an equilibrium trajectory, and provides a standard route to a correlation time for that observable [104]. The decorrelation stride for chemical-shift averaging was taken from the normalized autocorrelation of the lid extension CV ($C_\alpha 1$ to $C_\alpha 23$), the first $1/e$ crossing at $\tau \approx 600$ frames (~ 12 ns) set synth-nmr predictions to stride 600 for residues 19 to 24 [105], for a detailed ACF construction see Appendix C, Calculations C.4. Isotonic means per (residue, atom) were compared to Biological Magnetic Resonance Bank (BMRB) entry 6612 for the same PDB 1Z1M NMR ensemble as in the simulations. Fifty-six experimental and 34 predicted nuclei were identified. 25 intersecting pairs, dominated by backbone assignments, supplied scatter and Bland and Altman points with matched δ_{exp} and δ_{pred} [106] as shown in Figure 6.2. The 31 BMRB-only nuclei were mostly flexible side chains (for example H_β , H_γ , and C_γ), nuclei without a counterpart [106]. Predicted and experimental isotonic shifts agree closely ($r = 0.998$), showing

ensemble-mean predictions track the NMR shifts for this residue set [106]. Agreement largely holds across atom types present in the matched set. Note that atomic notation places mass number at the upper left, as in ^{15}N , atom-specific position at the lower right, as in H_α or C_β , formal charge at the upper right, as in Na^+ , and atomic number at the lower left only in specialized nuclide notation, as in $^{15}_7\text{N}$.

Figure 6.2: Bland and Altman plot for isotonic predictions. Difference is $\delta_{\text{pred}} - \delta_{\text{exp}}$ for each matched residue and atom. Horizontal lines indicate zero, mean difference (bias), and limits of agreement (LoA) ($\mu \pm 1.96\sigma$). These results indicate a modest overall bias and generally good agreement between predicted and experimental chemical shifts.



Agreement was further assessed with a Bland-Altman plot [106], which showed generally small residuals and limits of agreement (LoA) that contained 22 of the 25 matched nuclei. The three points outside those limits were confined to amide ^{15}N of residues 21 and 23 and C_α of residue 19, consistent with nucleus type-dependent limitations of the empirical predictor rather than a broad failure of the isotonic ensemble [107, 106]. The near-unity correlation ($r = 0.998$) and the small residual structure observed in the Bland-Altman analysis support use of the isotonic run as a calibrated baseline for subsequent comparisons. Across osmotic conditions, the ensemble-mean predicted shifts at the lid base changed marginally.

It has been demonstrated that the isotonic ensemble reproduces the local lid-base (residues 19-24) chemical-shift signal with strong agreement to experiment, small residuals, and near-unity correlation. The simulation trajectories provide an experi-

mentally anchored isotonic baseline for the insights that follow. Subsequent osmotic stress responses can be interpreted against a reference state that is locally calibrated by NMR rather than by an arbitrary starting geometry.

6.2 Lid Rigidity and Lid-Base Dissociation

The length of lid-extension (CV1) and the lid-core angle (CV2), were defined using the same collective-variable scheme previously applied to MDM2 simulations by Bueren-Calabuig and Michel in [105], expressed as

$$\text{CV1} = \|\mathbf{r}_{1,\text{C}_\alpha} - \mathbf{r}_{23,\text{C}_\alpha}\|, \quad (6.1)$$

$$\text{CV2} = \angle(\mathbf{r}_{11,\text{C}_\alpha}, \mathbf{r}_{50,\text{C}_\alpha}, \mathbf{r}_{62,\text{C}_\alpha}), \quad (6.2)$$

where $\mathbf{r}_{x,\text{C}_\alpha}$ denotes the position vector for the C_α atom of residue x , and $\angle(a, b, c)$ denotes the angle subtended by the three points a , b , and c , with b as the vertex. The osmotic-pressure effect on lid motion was characterized by one-dimensional potentials of mean force (PMFs) along CV1 and CV2. Both the one-dimensional PMFs and the joint two-dimensional energy landscape in CV space were constructed from the same decimated trajectories adopted for chemical-shift prediction, with maximally independent samples. As in most physical systems, the ensemble tends toward minimum-energy states; subsequently, the energy basin minimum provides an informative measure of the dominant CV state under a given osmotic condition [104].

Reconstruction of the energy basins revealed condition-dependent shifts in dominant lid conformations. For CV1, the dominant basin shifts toward larger lid-extension values with increasing osmotic pressure, the PMF minima occur at CV1 = 19.74 Å, 25.62 Å, and 28.97 Å under isotonic, hyperosmotic, and supraosmotic conditions respectively, this progression is shown along the top row of Figure 6.3. The progression along CV2 (see Figure 6.3 bottom row) shows an unexpected trend of state populations shifting towards a more open lid-core hinge angle. The pro-

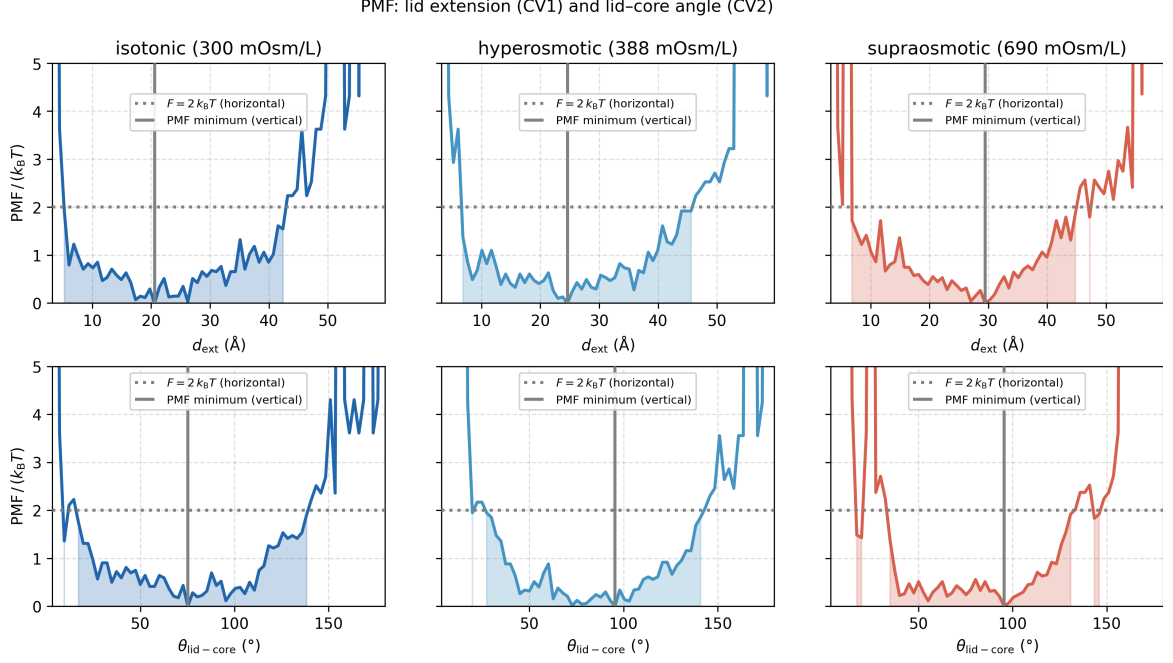


Figure 6.3: One-dimensional free-energy basins along CV1 (lid extension; top row) and CV2 (lid-core angle; bottom row) for isotonic, hyperosmotic, and supraosmotic conditions. Histograms pool the same decimated frames (stride 600, approximately 12 ns between samples at a 20 ps sampling frequency). Potentials of mean force are shown with a fixed $0-5 k_B T$ range, and vertical dotted lines mark the abscissa of the global minimum of each discrete PMF. The shaded $F < 2 k_B T$ region and horizontal dotted line retain their usual interpretation as a thermally accessible basin. With increasing osmolarity, CV1 shifts toward larger extension values, while CV2 exhibits stress-induced redistribution of hinge-angle geometry.

gression along CV2 was $= 67.60^\circ$, 80.91° , and 96.90° for isotonic, hyperosmotic, and supraosmotic ensemble minimum states respectively. The progression along CV2 is consistent with hinge geometry opening under stress and then plateauing in the sampled coordinate. More broadly, condition-dependent deformation of the CV2 energy basin indicates that osmotic stress reshapes the hinge geometry that governs the conformational compatibility at the lid base [104, 105].

The stress response is further refined when monitoring the number of atomic contacts at the lid-core hinge as a function of osmotic pressure. The statistical fluctuations represented by σ as the mean per-trajectory standard deviation, decreases concurrently with $\sigma = 5.46 \text{ \AA} \rightarrow 4.72 \text{ \AA} \rightarrow 4.03 \text{ \AA}$. The monotonic decrease in σ im-

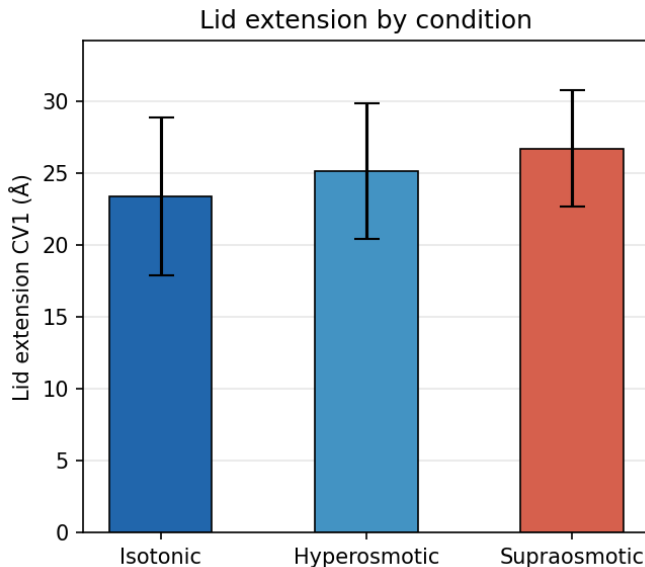


Figure 6.4: Condition-wise summary of the lid-extension coordinate (CV1; residue 1 to residue 23 $C_\alpha - C_\alpha$ distance). Bar height gives the condition level mean over trajectories (μ), error bars report the mean per-trajectory standard deviation (σ). From isotonic to hyperosmotic to supraosmotic conditions, μ shifts upward while σ decreases, indicating lid dynamics that are more extended and less flexible (more rigid) under osmotic stress.

plies reduced flexibility along the lid extension coordinate. Direct connection of this trend to lower configurational entropy requires further analysis, which is carried out in the next section through CV-space and $\mathcal{N}$ -space occupancy [104].

Total lid-base contact counts decrease monotonically with osmotic pressure, implying a partial dissociation of the lid-core hinge with mean contact totals of 1003.79 for isotonic, 924.70 for hyperosmotic, and 879.75 for supraosmotic conditions. The decline in lid-base hinge contacts is summarized in Table 6.1, consistent with osmotic pressure dissociating or loosening lid-base coupling. An interesting finding was that the proportions of favorable and unfavorable contacts remain surprisingly similar. The mean favorable fraction, $n_{\text{fav}}/(n_{\text{fav}} + n_{\text{unfav}})$, stays within about half a percentage point of exact parity for isotonic, hyperosmotic, and supraosmotic ensembles ($\sim 50.7\%$, $\sim 50.9\%$, and $\sim 50.4\%$, respectively). The favorable lid-base contacts are only slightly in excess of unfavorable ones even as total contact density falls. That balance is nonetheless compatible with large shifts in the CV projections, because equilibrium weights amplify small local biases when they enter exponential combinations of probabilities and free-energy profiles [104, 108, 109].

Importantly, the same stressed-regime trends in lid extension, hinge geometry, lid-

Table 6.1: Contact summary for the N-terminal lid-base (residues 19 to 24), aggregated by osmotic condition. Values are computed over all frames pooled within each condition. The mean is the per-frame average over all runs. For total contacts and favorable and unfavorable sums, the per-frame standard deviation, reported as $\mu \pm \sigma$. Reduced contacts indicates diminished lid-base coupling to the MDM2 protein core.

Condition	n_{runs}	$n_{\text{total}} (\mu \pm \sigma)$	$n_{\text{fav}} (\mu \pm \sigma)$	$n_{\text{unfav}} (\mu \pm \sigma)$
isotonic	96	1003.79 ± 272.67	509.30 ± 187.74	494.49 ± 148.54
hyperosmotic	96	924.70 ± 269.38	471.46 ± 189.08	453.25 ± 134.16
supraosmotic	96	879.75 ± 250.92	443.84 ± 171.77	435.91 ± 138.38

base contacts, and PMFs are observed while the isotonic reference remains consistent with experimental chemical-shift benchmarks, which supports treating these computational observables as physically meaningful [106, 107]. However, there were the unexpected predictions in the 1Z1M lid hinge (19-24), mean $|\delta_{\text{pred}} - \delta_{\text{exp}}|$ differences with isotonic were only 0.052 ppm and 0.049 ppm under hyperosmotic and supraosmotic conditions respectively, despite clear lid-base decoupling.

These observed trends portray osmotic stress as driving a more extended, less fluctuating lid with weaker hinge coupling that raises the effective barrier for large lid displacements thereby hindering the p53 induced-fit engagement with the groove [105]. The next section investigates CV-space and $\mathcal{N}$ -space occupancies to assess whether thermally accessible microstates in (CV1,CV2) and peripheral configurational entropy shrink in tandem, linking rigidification to fewer accessible states.

6.3 Osmotic Stress Induces State Population Shifts

The isotonic ensemble matches experimental lid-base chemical shifts at high correlation; consequently, the structural observables under osmotic stress are tied to a locally validated reference [106, 107]. Relative to that baseline, hyperosmotic and supraosmotic conditions show larger mean lid extension, a higher hinge angle at the one-dimensional PMF minimum, lower mean per-trajectory fluctuation along CV1, and lower lid-base contact totals with favorable and unfavorable classes declining in parallel (Table 6.1 and Figure 6.3). This is consistent with a hinge-lid ensemble that

is more extended and less fluctuating while partially decoupled from the core. The two-dimensional free-energy surfaces in the (CV1,CV2) plane (Figure 6.5) combine CVs into one equilibrium view where each condition contains a minimum-energy state.

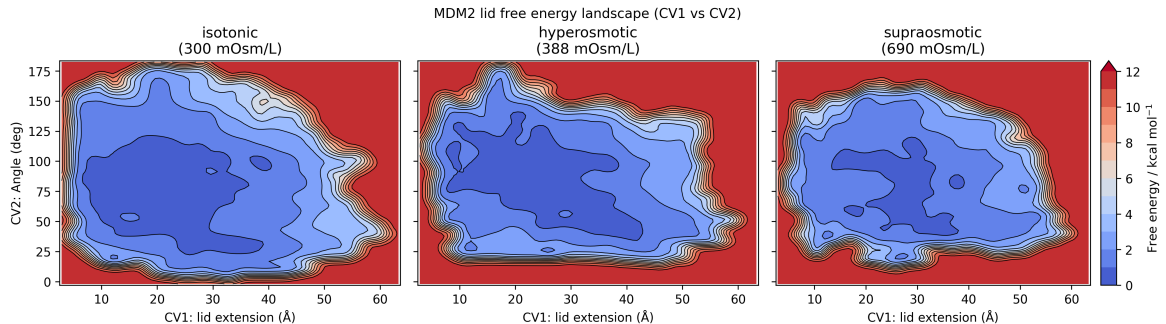


Figure 6.5: Free-energy landscapes in the (CV1,CV2) plane. Free energy in kcal mol^{-1} is relative to the minimum. Each column is one osmotic condition (isotonic, hyperosmotic, supraosmotic). Free energy increases away from the dominant basins, shifts in state populations show osmotic redistribution of the lid-base ensemble.

Peripheral organization appears in $\mathcal{N}$ -space as macrostate multiplicity $g(\mathcal{N}_a, \mathcal{N}_c)$, and isotonic-referenced Δg in Figure 6.6. Here, the analysis is restricted to intramolecular ensemble redistribution within a single MDM2 system, so the Q/M normalization was not applied and entropy is reported as the unnormalized Shannon quantity S'_Ψ . PSI entropy, falls monotonically from isotonic ($S'_\Psi=6.39$ bits) to hyperosmotic ($S'_\Psi=5.72$ bits) to supraosmotic ($S'_\Psi=5.44$ bits), concentrating state populations into fewer effective peripheral macrostates, a stress-induced tightening of the periphery. The 2D free-energy surface (FES) and $\mathcal{N}$ -space views are consistent with lid stiffening and reduced peripheral entropy. The bottom row of Figure 6.6 (Δg vs. isotonic) highlights (N_a, N_c) states that gain multiplicity under stress response alongside the CV1/CV2, PMFs, contacts, and FES. Hex cells with the strongest stress-induced enrichment are natural targets for biased sampling or design (ligands, staples, cross-links) targeting the stressed periphery to modulate the $(\mathcal{N}_a, \mathcal{N}_c)$ profile under isotonic conditions, possibly destabilizing the target.

Figure 6.5 supplies a joint (CV1,CV2) free-energy measurement on the trajectories

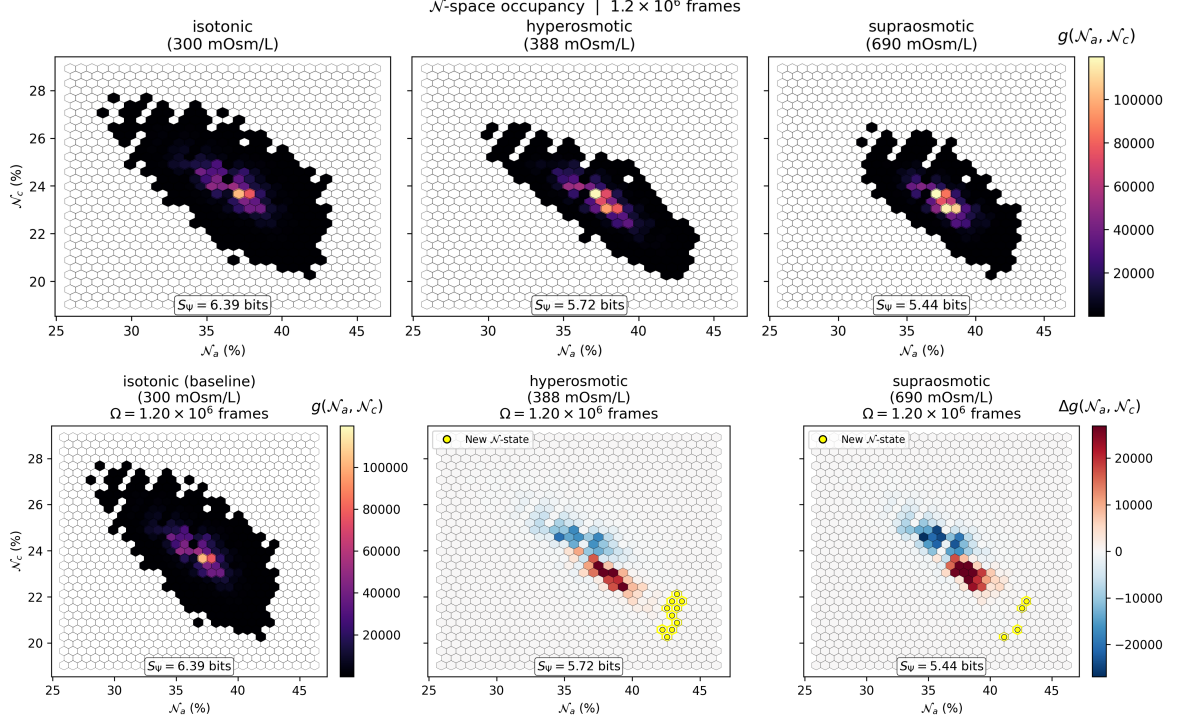


Figure 6.6: Stacked global $\mathcal{N}$ -space panels for osmotic-stress-driven population redistribution. The upper row shows condition-dependent occupancy multiplicity $g(N_a, N_c)$ for isotonic, hyperosmotic, and supraosmotic ensembles; the annotation above that row summarizes the scale of pooled sampling ($\mathcal{N}$ -space occupancy, 1.2×10^6 frames per osmotic condition). The lower row shows the corresponding change relative to the isotonic baseline (Δg), highlighting regions of decreased and increased occupancy and newly populated $\mathcal{N}$ -states. Data are derived from 96 trajectories of 250 ns each per condition, i.e. $24 \mu\text{s}$ of aggregate sampling per condition and all twenty-four systems. The redistribution indicates that osmotic stress reorganizes equilibrium peripheral macrostates and reweights state populations relative to isotonic conditions.

experimentally validated with the BRMB chemical shifts, while Figure 6.6 shows how osmotic stress reallocates peripheral state populations, decreasing S_Ψ quantifies the loss of peripheral diversity, and the Δg plots show which $(\mathcal{N}_a, \mathcal{N}_c)$ macrostates gain or lose multiplicity relative to isotonic sampling. Hex cells with the strongest stress-induced enrichment identify concrete $(\mathcal{N}_a, \mathcal{N}_c)$ objectives for interventions (ligands, staples, cross-links) aimed at reproducing or counteracting the stressed periphery.

CHAPTER 7: DISCUSSION AND FUTURE DIRECTIONS

The isotonic ensemble was anchored to experiment by the strong agreement between predicted and measured chemical shifts at the N-terminal lid base, with $r = 0.998$ for residues 19 to 24 relative to BMRB entry 6612 [110]. Against that validated reference, osmotic stress was accompanied by a coordinated redistribution of the lid ensemble, where mean CV1 increased from ~ 23.4 Å to ~ 25.1 Å to ~ 26.7 Å, the corresponding standard error decreased from ~ 5.5 Å to ~ 4.7 Å to ~ 4.0 Å, the one-dimensional PMF minima shifted toward larger CV1 and CV2 values, and total lid-base contacts declined from isotonic to hyperosmotic to supraosmotic conditions (see Table 6.1, Figure 6.4, and Figure 6.3). These observables collectively support the interpretation that osmotic stress biases MDM2 toward a lid ensemble that is more extended, less fluctuating, less tightly coupled to the core, and importantly the lid rigidity increases with osmotic pressure, ultimately reducing opportunities of induced fit.

The mechanistic significance of that redistribution is strengthened by the free-energy and state-population analyses. The joint CV surface indicates that the stressed ensembles remain organized around a dominant low-free-energy region rather than dispersing across unrelated lid geometries and the accompanying $\mathcal{N}$ -space analysis shows that peripheral organization is concentrated into fewer distinct macrostates. PSI entropy decreases monotonically from 6.39 bits to 5.72 bits to 5.44 bits from isotonic to hyperosmotic to supraosmotic conditions, and the corresponding entropy on the $(\mathcal{N}_a, \mathcal{N}_c)$ plane decreases from 4.11 bits to 3.70 bits to 3.46 bits. Support for the proposed physical mechanism is therefore obtained at two levels. Locally, through PMFs and CVs that indicate a narrower and more persistent lid-extension distribution; and globally, through state-population shifts that reveal stress-dependent tightening of

the periphery. Within that framework, increased lid rigidity is inferred to render displacement of the lid volume less favorable, even though lid-base contacts are reduced.

Across the dissertation, PSI entropy serves as a common statistical language for peripheral organization. In the broader protein-peptide setting, preferential occupation of restricted $\mathcal{N}$ -state regions identifies constrained NIS organization. In the MDM2 osmotic-stress system, the same formalism resolves how an environmental perturbation redistributes lid-associated peripheral states without requiring convergence to a single structural endpoint. Structural observables CV1, CV2, PMFs, contacts, free-energy basins, Δg , and S_Ψ become mutually interpretable descriptors of one ensemble shift. High Δg hexbins may therefore be regarded as candidate control points for modulating peripheral organization, groove accessibility, and recognition.

The peripheral, NIS-driven WW-domain destabilization strategy proposed here appears, to the best of my knowledge, distinct from prior WW-domain design efforts in its exact form. Existing work has established that WW-domain stability and function can be altered by mutation, and the closest conceptual precedent is the Pin1 WW-domain study in which a single peripheral surface substitution, Q33E, reduced thermostability through propagated electrostatic and hydrophobic reorganization and was associated with decreased ligand binding despite lying outside the direct binding site [111]. By contrast, a validated strategy in which distal surface perturbation or small molecules are used to globally destabilize YAP WW domains as a therapeutic design principle was not identified in the literature surveyed here, although mutation of the YAP WW domains has been shown to diminish oncogenic activity [50]. This proposed design strategy should be regarded as context dependent, since YAP/TAZ WW domains mediate interactions with both positive and negative regulators.

Two future directions arise naturally from this dissertation. First, a targeted NMR validation should initially be developed under isotonic, hyperosmotic, and supraosmotic conditions to experimentally test whether residues 19 to 24 and groove-proximal

sites exhibit the predicted reduction in conformational heterogeneity and redistribution in exchange behavior. Condition-resolved chemical shifts or intensity measurements would be particularly informative because both kinetics and proximity could then be assessed in the same lid-base region identified by simulation. An "anti-directed" design strategy may be formulated in $\mathcal{N}$ -space where ligands, staples, cross-links, or mutations are selected to drive peripheral occupancy away from the dominant modes, destabilizing specific recognition or perturbing allosteric communication.

REFERENCES

- [1] Levinthal C. How to Fold Graciously. *Mossbauer Spectroscopy in Biological Systems*. 1969:22–24.
- [2] Dill KA, MacCallum JL. The Protein-Folding Problem, 50 Years On. *Science*. 2012;338(6110):1042–1046.
- [3] Hollingsworth SA, Dror RO. Molecular Dynamics Simulation for All. *Neuron*. 2018;99(6):1129–1143.
- [4] Bernardi RC, Melo MCR, Schulten K. Enhanced Sampling Techniques in Molecular Dynamics Simulations of Biological Systems. *Biochimica et Biophysica Acta (BBA) - General Subjects*. 2015;1850(5):872–877.
- [5] Meng XY, Zhang HX, Mezei M, Cui M. Molecular Docking: A Powerful Approach for Structure-Based Drug Discovery. *Current Computer-Aided Drug Design*. 2011;7(2):146–157.
- [6] Torrie GM, Valleau JP. Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling. *Journal of Computational Physics*. 1977;23(2):187–199.
- [7] Hénin J, Lelièvre T, Shirts MR, Valsson O, Delemotte L. Enhanced Sampling Methods for Molecular Dynamics Simulations. *Living Journal of Computational Molecular Science*. 2022;4(1):1583.
- [8] Pierce MM, Raman CS, Nall BA. Isothermal Titration Calorimetry of Protein-Protein Interactions. *Methods*. 1999;19(2):213–221.
- [9] Chodera JD, Mobley DL. Entropy-Enthalpy Compensation: Role and Ramifications in Biomolecular Ligand Recognition and Design. *Annual Review of Biophysics*. 2013;42:121–142.
- [10] Caro JA, Harpole KW, Kasinath V, Lim J, Granja J, Valentine KG, et al. Entropy in molecular recognition by proteins. *Proceedings of the National Academy of Sciences of the United States of America*. 2017;114(25):6563–6568.
- [11] Wallach I, Heifets A. Most Ligand-Based Classification Benchmarks Reward Memorization Rather than Generalization. *Journal of Chemical Information and Modeling*. 2018;58(5):916–932.
- [12] Maurer M, Oostenbrink C. Water in protein hydration and ligand recognition. *Journal of Molecular Recognition*. 2019;32(12):e2810.
- [13] Konstantinovskiy D, Perets EA, Santiago T, Velarde L, Hammes-Schiffer S, Yan ECY. Detecting the First Hydration Shell Structure around Biomolecules at Interfaces. *ACS Central Science*. 2022;8(10):1404–1414.

- [14] Gilson MK, Given JA, Bush BL, McCammon JA. The statistical-thermodynamic basis for computation of binding affinities: a critical review. *Biophysical Journal*. 1997 mar;72(3):1047–1069.
- [15] Vangone A, Bonvin AM. Contacts-based prediction of binding affinity in protein–protein complexes. *elife*. 2015;4:e07454.
- [16] Kastritis PL, Rodrigues JPGLM, Folkers GE, Boelens R, Bonvin AMJJ. Proteins feel more than they see: fine-tuning of binding affinity by properties of the non-interacting surface. *Journal of Molecular Biology*. 2014;426(14):2632–2652.
- [17] Visscher KM, Kastritis PL, Bonvin AMJJ. Non-interacting surface solvation and dynamics in protein–protein interactions. *Proteins: Structure, Function, and Bioinformatics*. 2015;83(3):445–458.
- [18] Cooper A. Heat capacity effects in protein folding and ligand binding: a re-evaluation of the role of water in biomolecular thermodynamics. *Biophysical Chemistry*. 2005;115(2-3):89–97.
- [19] Motlagh HN, Wrabl JO, Li J, Hilser VJ. The ensemble nature of allostery. *Nature*. 2014;508(7496):331–339.
- [20] Kauzmann W. Some factors in the interpretation of protein denaturation. In: *Advances in protein chemistry*. vol. 14. Elsevier; 1959. p. 1–63.
- [21] Bacon F. On the nature of allosteric transitions: a plausible model. *J Mol Biol*. 1965;12(1):88–118.
- [22] Ladbury JE, Williams MA. The extended interface: measuring non-local effects in biomolecular interactions. *Current Opinion in Structural Biology*. 2004;14(5):562–569.
- [23] Jacobs D, Dallakyan S. Elucidating protein thermodynamics from the three-dimensional structure of the native state using network rigidity. *Biophysical journal*. 2005;88 2:903–15.
- [24] Cooper A, Johnson CM, Lakey JH, Nöllmann M. Heat does not come in different colours: entropy–enthalpy compensation, free energy windows, quantum confinement, pressure perturbation calorimetry, solvation and the multiple causes of heat capacity effects in biomolecular interactions. *Biophysical chemistry*. 2001;93(2-3):215–230.
- [25] Kastritis PL, Bonvin AM. Molecular origins of binding affinity: seeking the Archimedean point. *Current Opinion in Structural Biology*. 2013;23(6):868–877.
- [26] Kastritis PL, Bonvin AM. On the binding affinity of macromolecular interactions: daring to ask why proteins interact. *Journal of The Royal Society Interface*. 2013;10(79):20120835.

- [27] Visscher KM, Kastritis PL, Bonvin AM. Non-interacting surface solvation and dynamics in protein–protein interactions. *Proteins: Structure, Function, and Bioinformatics*. 2015;83(3):445–458.
- [28] Kastritis PL, Rodrigues JP, Folkers GE, Boelens R, Bonvin AM. Proteins feel more than they see: fine-tuning of binding affinity by properties of the non-interacting surface. *Journal of molecular biology*. 2014;426(14):2632–2652.
- [29] Vangone A, Schaarschmidt J, Koukos P, Geng C, Citro N, Trellet ME, et al. Large-scale prediction of binding affinity in protein–small ligand complexes: The PRODIGY-LIG web server. *Bioinformatics*. 2019;35(9):1585–1587.
- [30] Schweke H, Mucchielli MH, Sacquin-Mora S, Bei W, Lopes A. Protein interaction energy landscapes are shaped by functional and also non-functional partners. *Journal of molecular biology*. 2020;432(4):1183–1198.
- [31] Finan JD, Guilak F. The effects of osmotic stress on the structure and function of the cell nucleus. *Journal of Cellular Biochemistry*. 2010;109(3):460–467.
- [32] Mourão MA, Hakim JB, Schnell S. Connecting the dots: the effects of macromolecular crowding on cell physiology. *Biophysical Journal*. 2014;107(12):2761–2766.
- [33] Momand J, Zambetti GP, Olson DC, George D, Levine AJ. The mdm-2 oncogene product forms a complex with the p53 protein and inhibits p53-mediated transactivation. *Cell*. 1992;69(7):1237–1245.
- [34] Honda R, Tanaka H, Yasuda H. Oncoprotein MDM2 is a ubiquitin ligase E3 for tumor suppressor p53. *FEBS Letters*. 1997;420(1):25–27.
- [35] Levine AJ. p53, the cellular gatekeeper for growth and division. *Cell*. 1997;88(3):323–331.
- [36] Dmitrieva N, Kultz D, Michea L, Ferraris J, Burg M. Protection of renal inner medullary epithelial cells from apoptosis by hypertonic stress-induced p53 activation. *Journal of Biological Chemistry*. 2000;275(24):18243–18247.
- [37] Dmitrieva N, Michea L, Burg M. p53 protects renal inner medullary cells from hypertonic stress by restricting DNA replication. *American Journal of Physiology - Renal Physiology*. 2001;281(3):F522–F530.
- [38] Kishi H, Nakagawa K, Matsumoto M, Suga M, Ando M, Taya Y, et al. Osmotic shock induces G1 arrest through p53 phosphorylation at Ser33 by activated p38MAPK without phosphorylation at Ser15 and Ser20. *Journal of Biological Chemistry*. 2001;276(42):39115–39122.
- [39] Lambert IH, Enghoff MS, Brandi ML, Hoffmann EK. Regulation of p53 in NIH3T3 mouse fibroblasts following hyperosmotic stress. *Physiological Reports*. 2015;3(6):e12412.

- [40] Gamboni F, Anderson C, Mitra S, Reisz JA, Nemkov T, Dzieciatkowska M, et al. Hypertonic saline primes activation of the p53-p21 signaling axis in human small airway epithelial cells that prevents inflammation induced by pro-inflammatory cytokines. *Journal of Proteome Research*. 2016;15(10):3813–3826.
- [41] Verkhivker GM. Simulating molecular mechanisms of the MDM2-mediated regulatory interactions: a conformational selection model of the MDM2 lid dynamics. *PLoS ONE*. 2012;7(7):e40897.
- [42] Mukherjee S, Pantelopulos GA, Voelz VA. Markov models of the apo-MDM2 lid region reveal diffuse yet two-state binding dynamics and receptor poses for computational docking. *Scientific Reports*. 2016;6:31631.
- [43] Bueren-Calabuig JA, Michel J. Elucidation of ligand-dependent modulation of disorder-order transitions in the oncoprotein MDM2. *PLoS Computational Biology*. 2015;11(6):e1004282.
- [44] Aragón E, Goerner N, Xi Q, Gomes T, Gao S, Massagué J, et al. Structural Basis for the Versatile Interactions of Smad7 with Regulator WW Domains in TGF- β Pathways. *Structure*. 2012;20(10):1726–1736.
- [45] Sudol M. Yes-Associated Protein (YAP65) Is a Proline-Rich Phosphoprotein That Binds to the SH3 Domain of the Yes Proto-Oncogene Product. *Oncogene*. 1994;9(8):2145–2152.
- [46] Yuan M, Tomlinson V, Lara R, Holliday D, Chelala C, Harada T, et al. Yes-Associated Protein (YAP) Functions as a Tumor Suppressor in Breast. *Cell Death and Differentiation*. 2008;15(11):1752–1759.
- [47] Raj N, Bam R. Reciprocal Crosstalk Between YAP1/Hippo Pathway and the p53 Family Proteins: Mechanisms and Outcomes in Cancer. *Frontiers in Cell and Developmental Biology*. 2019;7:159.
- [48] Pires JR, Taha-Nejad F, Toepert F, Ast T, Hoffmüller U, Schneider-Mergener J, et al. Solution structures of the YAP65 WW domain and the variant L30 K in complex with the peptides GTPPPPYTVG, N-(n-octyl)-GPPPY and PLPPY and the application of peptide libraries reveal a minimal binding epitope. *Journal of molecular biology*. 2001;314(5):1147–1156.
- [49] Lu H, Li H, Banu Bte Sm Rashid S, Leow WK, Liou YC. Knowledge-guided docking of WW domain proteins and flexible ligands. In: *Pattern Recognition in Bioinformatics: 4th IAPR International Conference, PRIB 2009, Sheffield, UK, September 7-9, 2009. Proceedings 4*. Springer; 2009. p. 175–186.
- [50] Zhao B, Kim J, Ye X, Lai ZC, Guan KL. Both TEAD-binding and WW domains are required for the growth stimulation and oncogenic transformation activity of yes-associated protein. *Cancer Research*. 2009;69(3):1089–1098.

- [51] Yang Y, Wu M, Pan Y, Hua Y, He X, Li X, et al. WW domains form a folded type of nuclear localization signal to guide YAP1 nuclear import. *Journal of Cell Biology*. 2024;223(6):e202308013.
- [52] Liu X, Yang N, Figel SA, Wilson KE, Morrison CD, Gelman IH, et al. PTPN14 interacts with and negatively regulates the oncogenic function of YAP. *Oncogene*. 2013;32(10):1266–1273.
- [53] Liu X, Fuentes EJ. Emerging Themes in PDZ Domain Signaling: Structure, Function, and Inhibition. *International Review of Cell and Molecular Biology*. 2019;343:129–218.
- [54] Lee HJ, Zheng JJ. PDZ Domains and Their Binding Partners: Structure, Specificity, and Modification. *Cell Communication and Signaling*. 2010;8:8.
- [55] Dev KK. Making Protein Interactions Druggable: Targeting PDZ Domains. *Nature Reviews Drug Discovery*. 2004;3(12):1047–1056.
- [56] Christensen NR, Čalyševa J, Fernandes EFA, Lüchow S, Clemmensen LS, Haugaard-Kedström LM, et al. PDZ Domains as Drug Targets. *Advanced Therapeutics*. 2019;2(7):1800143.
- [57] Nardella C, Visconti L, Malagrino F, Pagano L, Bufano M, Nalli M, et al. Targeting PDZ Domains as Potential Treatment for Viral Infections, Neurodegeneration and Cancer. *Biology Direct*. 2021;16(1):15.
- [58] Kim E, Sheng M. PDZ Domain Proteins of Synapses. *Nature Reviews Neuroscience*. 2004;5(10):771–781.
- [59] Tonikian R, Zhang Y, Sazinsky SL, Currell B, Yeh JH, et al. A Specificity Map for the PDZ Domain Family. *PLoS Biology*. 2008;6(9):e239.
- [60] Amacher JF, Brooks L, Hampton TH, Madden DR. Specificity in PDZ-Peptide Interaction Networks: Computational Analysis and Review. *Journal of Structural Biology X*. 2020;4:100022.
- [61] Kalyoncu S, Keskin Ö, Gursoy A. Interaction Prediction and Classification of PDZ Domains. *BMC Bioinformatics*. 2010;11:357.
- [62] Nourry C, Grant SGN, Borg JP. PDZ Domain Proteins: Plug and Play! *Science's STKE*. 2003;(179):re7.
- [63] Lu J, Li H, Wang Y, Südhof TC, Rizo J. Solution Structure of the RIM1 α PDZ Domain in Complex with an ELKS1b C-Terminal Peptide. *Journal of Molecular Biology*. 2005;352(2):455–466.
- [64] Wang Y, Liu X, Biederer T, Südhof TC. A Family of RIM-Binding Proteins Regulated by Alternative Splicing: Implications for the Genesis of Synaptic Active Zones. *Proceedings of the National Academy of Sciences of the United States of America*. 2002;99(22):14464–14469.

- [65] Kussie PH, Gorina S, Marechal V, Elenbaas B, Moreau J, Levine AJ, et al. Structure of the MDM2 oncoprotein bound to the p53 tumor suppressor trans-activation domain. *Science*. 1996;274(5289):948–953.
- [66] Haupt Y, Maya R, Kazaz A, Oren M. Mdm2 promotes the rapid degradation of p53. *Nature*. 1997;387(6630):296–299.
- [67] Alarcon-Vargas D, Ronai Z. p53–Mdm2—the affair that never ends. *Carcinogenesis*. 2002;23(4):541–547.
- [68] Frauenfelder H, Sligar SG, Wolynes PG. The Energy Landscapes and Motions of Proteins. *Science*. 1991;254(5038):1598–1603.
- [69] Henzler-Wildman K, Kern D. Dynamic Personalities of Proteins. *Nature*. 2007;450(7172):964–972.
- [70] Boehr DD, Nussinov R, Wright PE. The Role of Dynamic Conformational Ensembles in Biomolecular Recognition. *Nature Chemical Biology*. 2009;5(11):789–796.
- [71] Mitternacht S. FreeSASA: An open source C library for solvent accessible surface area calculations. *F1000Research*. 2016;5:189.
- [72] Hubbard SJ, Thornton JM. NACCESS: A computer program for calculating accessibilities; 1993. Department of Biochemistry and Molecular Biology, University College London.
- [73] Shannon CE. A Mathematical Theory of Communication. *Bell System Technical Journal*. 1948;27(3–4):379–423, 623–656.
- [74] Miyazawa S, Jernigan RL. Residue–residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. *Journal of molecular biology*. 1996;256(3):623–644.
- [75] Goldstein H, Poole CP, Safko JL. *Classical Mechanics*. 3rd ed. San Francisco: Addison-Wesley; 2002.
- [76] Peters B. Reaction Coordinates and Mechanistic Hypothesis Tests. *Annual Review of Physical Chemistry*. 2016;67:669–690.
- [77] Roux B. The Calculation of the Potential of Mean Force Using Computer Simulations. *Computer Physics Communications*. 1995;91:275–282.
- [78] Chipot C. Free Energy Methods for the Description of Molecular Processes. *Annual Review of Biophysics*. 2023;52:113–138.
- [79] Kumar S, Rosenberg JM, Bouzida D, Swendsen RH, Kollman PA. The Weighted Histogram Analysis Method for Free-Energy Calculations on Biomolecules. I. The Method. *Journal of Computational Chemistry*. 1992;13(8):1011–1021.

- [80] Dill KA, Ozkan SB, Shell MS, Weikl TR. The Protein Folding Problem. *Annual Review of Biophysics*. 2008;37:289–316.
- [81] Kabelka I, Brožek R, Vácha R. Selecting Collective Variables and Free-Energy Methods for Peptide Translocation across Membranes. *Journal of Chemical Information and Modeling*. 2021;61(2):819–830.
- [82] Harris RK, Becker ED, Cabral de Menezes SM, Goodfellow R, Granger P. NMR Nomenclature. Nuclear Spin Properties and Conventions for Chemical Shifts (IUPAC Recommendations 2001). *Pure and Applied Chemistry*. 2001;73(11):1795–1818.
- [83] Case DA. Chemical shifts in biomolecules. *Current Opinion in Structural Biology*. 2013;23(2):172–176.
- [84] Saitô H, Ando I, Ramamoorthy A. Chemical Shift Tensor – the Heart of NMR: Insights into Biological Aspects of Proteins. *Progress in Nuclear Magnetic Resonance Spectroscopy*. 2010;57(2):181–228.
- [85] Ulrich EL, Akutsu H, Doreleijers JF, Harano Y, Ioannidis YE, Lin J, et al. BioMagResBank. *Nucleic Acids Research*. 2008;36(Database issue):D402–D408.
- [86] Torchia DA. NMR Studies of Dynamic Biomolecular Conformational Ensembles. *Progress in Nuclear Magnetic Resonance Spectroscopy*. 2015;84–85:14–32.
- [87] Ramelot TA, Tejero R, Montelione GT. Representing Structures of the Multiple Conformational States of Proteins. *Current Opinion in Structural Biology*. 2023;83:102703.
- [88] Giulini M, Reys V, Teixeira JM, Jiménez-García B, V Honorato R, Kravchenko A, et al. HADDOCK3: A modular and versatile platform for integrative modeling of biomolecular complexes. *Journal of Chemical Information and Modeling*. 2025.
- [89] Jorgensen WL, Tirado-Rives J. The OPLS [optimized potentials for liquid simulations] potential functions for proteins, energy minimizations for crystals of cyclic peptides and crambin. *Journal of the American Chemical Society*. 1988;110(6):1657–1666.
- [90] Eastman P, Galvelis R, Peláez RP, Abreu CR, Farr SE, Gallicchio E, et al. OpenMM 8: molecular dynamics simulation with machine learning potentials. *The Journal of Physical Chemistry B*. 2023;128(1):109–116.
- [91] Eastman P, Swails J, Chodera JD, McGibbon RT, Zhao Y, Beauchamp KA, et al. OpenMM 7: Rapid development of high performance algorithms for molecular dynamics. *PLoS computational biology*. 2017;13(7):e1005659.
- [92] Langevin P. Sur la théorie du mouvement brownien. *Comptes Rendus de l’Académie des Sciences (Paris)*. 1908;146:530–533.

- [93] Verlet L. Computer “Experiments” on Classical Fluids. I. Thermodynamical Properties of Lennard-Jones Molecules. *Physical Review*. 1967;159(1):98–103.
- [94] Jurrus E, Engel D, Star K, Monson K, Brandi J, Felberg LE, et al. Improvements to the APBS biomolecular solvation software suite. *Protein science*. 2018;27(1):112–128.
- [95] Elkins G. synth-nmr: NMR spectroscopy calculations for protein structures; 2026. Available from: <https://github.com/elkins/synth-nmr>.
- [96] Shen Y, Bax A. SPARTA+: a modest improvement in empirical NMR chemical shift prediction by means of an artificial neural network. *Journal of Biomolecular NMR*. 2010;48(1):13–22.
- [97] Eastman P, Friedrichs MS, Chodera JD, Radmer RJ, Bruns CM, Ku JP, et al. OpenMM 4: A Reusable, Extensible, Hardware Independent Library for High Performance Molecular Simulation. *Journal of Chemical Theory and Computation*. 2013;9(1):461–469.
- [98] Dolinsky TJ, Czodrowski P, Li H, Nielsen JE, Jensen JH, Klebe G, et al. PDB2PQR: An Automated Pipeline for the Setup of Poisson–Boltzmann Electrostatics Calculations. *Nucleic Acids Research*. 2004;32(suppl_2):W665–W667.
- [99] Maier JA, Martinez C, Kasavajhala K, Wickstrom L, Hauser KE, Simmerling C. ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *Journal of Chemical Theory and Computation*. 2015;11(8):3696–3713.
- [100] Wang LP, Martinez TJ, Pande VS. Building Force Fields: An Automatic, Systematic, and Reproducible Approach. *The Journal of Physical Chemistry Letters*. 2014;5(11):1885–1891.
- [101] Darden T, York D, Pedersen L. Particle Mesh Ewald: An N-log(N) Method for Ewald Sums in Large Systems. *The Journal of Chemical Physics*. 1993;98(12):10089–10092.
- [102] Braun E, Gilmer J, Mayes HB, Mobley DL, Monroe JI, Prasad S, et al. Best Practices for Foundations in Molecular Simulations [Article v1.0]. *Living Journal of Computational Molecular Science*. 2019;1(1):5957.
- [103] Zhang Z, Liu X, Yan K, Tuckerman ME, Liu J. Unified Efficient Thermostat Scheme for the Canonical Ensemble with Holonomic or Isokinetic Constraints via Molecular Dynamics. *The Journal of Physical Chemistry A*. 2019;123(28):6056–6079.
- [104] Allen MP, Tildesley DJ. *Computer Simulation of Liquids*. Oxford: Oxford University Press; 1987.

- [105] Bueren-Calabuig JA, Michel J. Elucidation of ligand-dependent modulation of disorder-order transitions in the oncoprotein MDM2. *PLOS Computational Biology*. 2015;11(6):e1004282.
- [106] Montelione GT, Nilges M, Bax A, Guntert P, Herrmann T, Richardson JS, et al. Recommendations of the wwPDB NMR Validation Task Force. *Structure*. 2013;21(9):1563–1570.
- [107] Koes DR, Vries JK. Error assessment in molecular dynamics trajectories using computed NMR chemical shifts. *Computational and Theoretical Chemistry*. 2017;1099:152–166.
- [108] McQuarrie DA. *Statistical Mechanics*. New York: Harper & Row; 1976.
- [109] Dill KA, Bromberg S. *Molecular Driving Forces: Statistical Thermodynamics in Biology, Chemistry, Physics, and Nanoscience*. 2nd ed. New York: Garland Science; 2011.
- [110] Uhrinova S, Uhrin D, Powers H, Watt K, Zheleva D, Fischer P, et al. Structure of Free MDM2 N-Terminal Domain Reveals Conformational Adjustments That Accompany p53-Binding. *Journal of Molecular Biology*. 2005;350(3):587–598.
- [111] Zhang M, Case DA, Peng JW. Propagated Perturbations from a Peripheral Mutation Show Interactions Supporting WW Domain Thermostability. *Structure*. 2018;26(11):1474–1485.e5.

APPENDIX A: SUPPLEMENTAL TABLES

Table A.1: Pairwise chemical weights for residue-residue contact classes.

Contact Class	$\gamma(c_i, c_j)$	Note
positive-negative	1.65	Salt-bridge attraction, strongly favorable
negative-positive	1.65	Salt-bridge attraction, strongly favorable
apolar-apolar	1.35	Hydrophobic contacts, drives concentration, moderately favorable
polar-polar	1.22	H-bond/dipole interactions, moderately favorable
polar-positive	1.22	Polar assists charge via H-bond/electrostatics, moderately favorable
polar-negative	1.22	Polar assists charge via H-bond/electrostatics, moderately favorable
apolar-polar	0.90	Induces desolvation costs, mildly unfavorable
apolar-positive	0.82	Hydrophobe-charge mismatch, moderately unfavorable
apolar-negative	0.82	Hydrophobe-charge mismatch, moderately unfavorable
positive-positive	0.61	Like-charge repulsion, strongly unfavorable
negative-negative	0.61	Like-charge repulsion, strongly unfavorable

Table A.2: NACCESS residue maximum solvent-accessible surface area values (ASA_{max}) for SASA to RSA calculations.

Residue	ASA_{max}
Alanine (ALA)	$107.95 \text{ \AA}^2$
Cysteine (CYS)	$134.28 \text{ \AA}^2$
Aspartic acid (ASP)	$140.39 \text{ \AA}^2$
Glutamic acid (GLU)	$172.25 \text{ \AA}^2$
Phenylalanine (PHE)	$199.48 \text{ \AA}^2$
Glycine (GLY)	$80.10 \text{ \AA}^2$
Histidine (HIS)	$182.88 \text{ \AA}^2$
Isoleucine (ILE)	$175.12 \text{ \AA}^2$
Lysine (LYS)	$200.81 \text{ \AA}^2$
Leucine (LEU)	$178.63 \text{ \AA}^2$
Methionine (MET)	$194.15 \text{ \AA}^2$
Asparagine (ASN)	$143.94 \text{ \AA}^2$
Proline (PRO)	$136.13 \text{ \AA}^2$
Glutamine (GLN)	$178.50 \text{ \AA}^2$
Arginine (ARG)	$238.76 \text{ \AA}^2$
Serine (SER)	$116.50 \text{ \AA}^2$
Threonine (THR)	$139.27 \text{ \AA}^2$
Valine (VAL)	$151.44 \text{ \AA}^2$
Tryptophan (TRP)	$249.36 \text{ \AA}^2$
Tyrosine (TYR)	$212.76 \text{ \AA}^2$

Table A.3: List of $\{PY\}$ (PPxY, LPxY, and PPGW) protein sequences.

PDB ID	Sequence	Length
1ZCN	LPPGWEKRMS	10
2LB1	PPAYLPPEDP	10
2LB2	PPPGYLSGDG	10
2LTV	PPPYSRYPMD	10
2ZQS	KLPPGWEKRM	10
2EZ5	TGLPSYDEALH	11
2KXQ	SPPPPYSRYPM	11
6J1Z	PPGWEKRTDPR	11
7BQF	LPLPPGWSVDW	11
2LAW	TPPPAYLPPEDP	12
2LTY	ESPPPPYSRYPM	13
2MPT	RLDLPPYETFEDL	13
7LP5a	QNNEELPTYEEAKV	14
SMDA	GESPPPPYSRYPM	14
SMDC	GESPPPPYSRYPM	14
2DJY	ELESPPPPYSRYPM	15

Table A.4: List of randomly generated protein sequences in the $\{RD\}$ set.

Name	Sequence	Length
RD01	YSEEITHQQM	10
RD02	RHAKHWMSRG	10
RD03	FWISKKCTKW	10
RD04	HATCCYDVKDF	11
RD05	FWEVFMGRLL	11
RD06	MQQMGEFKKNW	11
RD07	EDAEIYEPYFNW	12
RD08	CFPYYGYTTIMH	12
RD09	MMIRYLHNVKSDW	13
RD10	DFNTHYGVVSIET	13
RD11	SIVFMYTLMFMMIN	14
RD12	PIHLDARAPRYWQN	14
RD13	KNQFLMIWHAYYIL	14
RD14	VMWECAYSNTHWSGI	15
RD15	YDVLQQVSFMTNYWI	15
RD16	SYYSFWDGRIHQTGA	15

Table A.5: $\mathcal{N}$ -state properties between Step 2 (semi-flexible refinement) and Step 3 (MDS in explicit solvent) extracted from conformational ensembles (docking solutions). Sixteen proline-rich peptides with motifs PPxY, LPxY, and PPGW were used to generate ensembles of protein-peptide complexes. The second column from the left contains the variables $\mathcal{N}_a$ and $\mathcal{N}_c$ being the NIS apolar and charged surface percentages respectively. P_{max} represents the maximum probability state (mode) in $\mathcal{N}$ -space, and S_Ψ is the peripheral surface information (PSI) entropy for a given ensemble. The terms $\Delta\mathcal{N}$ and ΔS_Ψ indicate the changes from Step 2 to Step 3 in the number of $\mathcal{N}$ -states and PSI entropy respectively.

		PPxY									LPxY			PPGW			
		2DJY	2KXQ	2LB1	2LB2	2LTV	2LTY	2LAW	SMDA	SMDC	2EZ5	7LP5a	2MPT	1ZCN	2ZQS	6JIZ	7BQF
Step 2	$\mathcal{N}_a$ (%)	40.00	41.30	44.44	42.22	40.00	39.58	44.68	40.82	40.00	39.13	34.69	39.58	40.00	40.00	36.96	43.48
	$\mathcal{N}_c$ (%)	24.00	19.57	22.22	22.22	22.22	22.92	21.28	22.45	24.00	21.74	26.53	27.08	24.44	26.67	28.26	19.57
	P_{max}	0.80	0.56	0.88	0.86	0.91	0.70	0.80	0.79	0.80	0.62	0.76	0.87	0.43	0.93	0.43	0.88
	$\mathcal{N}$ -states	4.00	7.00	4.00	6.00	3.00	5.00	6.00	4.00	4.00	9.00	6.00	3.00	14.00	4.00	6.00	3.00
	S_Ψ	-61.23	-43.94	-76.27	-50.27	-242.69	-29.28	-69.30	-85.27	-61.23	-22.70	-37.29	-123.54	-7.32	-73.41	-44.42	-76.43
Step 3	$\mathcal{N}_a$ (%)	40.00	41.30	44.44	42.22	40.00	39.58	44.68	40.82	40.00	39.13	34.69	39.58	40.91	40.00	36.96	43.48
	$\mathcal{N}_c$ (%)	24.00	19.57	22.22	22.22	22.22	22.92	21.28	22.45	24.00	21.74	26.53	27.08	25.00	26.67	28.26	19.57
	P_{max}	0.68	0.50	0.69	0.74	0.79	0.54	0.60	0.69	0.68	0.46	0.53	0.64	0.32	0.69	0.29	0.88
	$\mathcal{N}$ -states	8.00	9.00	7.00	8.00	6.00	10.00	7.00	7.00	8.00	12.00	8.00	4.00	12.00	7.00	10.00	4.00
	S_Ψ	-20.10	-24.65	-27.74	-27.64	-83.38	-7.62	-44.88	-41.78	-20.10	-9.10	-13.23	-68.95	-9.12	-22.57	-13.72	-30.79
	$\Delta\mathcal{N}$	4.00	2.00	3.00	2.00	3.00	5.00	1.00	3.00	4.00	3.00	2.00	1.00	-2.00	3.00	4.00	1.00
	ΔS_Ψ	41.13	19.29	48.53	22.63	159.31	21.66	24.42	43.49	41.13	13.60	24.05	54.59	-1.79	50.84	30.70	45.63

Table A.6: $\mathcal{N}$ -state properties between Step 2 (semi-flexible refinement) and Step 3 (MDS in explicit solvent) extracted from conformational ensembles (docking solutions). Sixteen random sequences (decoys) were predicted using AF3 then equilibrated with MDS in explicit TIP3P solvent (RD01 to RD16). The second column contains the variables $\mathcal{N}_a$ and $\mathcal{N}_c$ which are the NIS apolar and charged surface percentages respectively. P_{max} represents the maximum probability state (mode) in $\mathcal{N}$ -space, and S_Ψ is the peripheral surface information (PSI) entropy for a given conformational ensemble. The terms $\Delta\mathcal{N}$ and ΔS_Ψ indicate the changes from Step 2 to Step 3 in the number of $\mathcal{N}$ -states and PSI entropy respectively.

		RD01	RD02	RD03	RD04	RD05	RD06	RD07	RD08	RD09	RD10	RD11	RD12	RD13	RD14	RD15	RD16
Step 2	$\mathcal{N}_a$ (%)	33.33	35.56	33.33	34.78	43.48	36.96	36.17	38.30	37.50	37.50	46.94	38.78	40.82	36.00	38.00	36.00
	$\mathcal{N}_c$ (%)	22.22	24.44	24.44	23.91	23.91	23.91	25.53	17.02	22.92	20.83	16.33	22.45	18.37	18.00	18.00	20.00
	P_{max}	0.92	0.63	0.89	0.69	0.83	0.95	0.71	0.58	0.67	0.67	0.82	0.76	0.70	0.94	0.87	0.90
	$\mathcal{N}$ -states	4.00	7.00	4.00	7.00	5.00	2.00	5.00	6.00	6.00	6.00	3.00	5.00	7.00	2.00	5.00	5.00
	S_Ψ	-142.64	-33.60	-117.95	-23.97	-70.79	-151.29	-79.61	-34.26	-19.42	-44.17	-24.71	-35.29	-9.68	-140.52	-27.13	-81.42
Step 3	$\mathcal{N}_a$ (%)	33.33	35.56	33.33	34.78	43.48	36.96	36.17	38.30	37.50	37.50	46.94	38.78	40.82	36.00	38.00	36.00
	$\mathcal{N}_c$ (%)	22.22	24.44	24.44	23.91	23.91	23.91	25.53	17.02	22.92	20.83	16.33	22.45	18.37	18.00	18.00	20.00
	P_{max}	0.73	0.52	0.71	0.46	0.76	0.83	0.57	0.38	0.56	0.48	0.49	0.66	0.67	0.70	0.70	0.72
	$\mathcal{N}$ -states	6.00	9.00	6.00	11.00	4.00	6.00	8.00	8.00	9.00	8.00	8.00	5.00	10.00	6.00	7.00	7.00
	S_Ψ	-67.30	-18.22	-60.30	-7.79	-87.38	-32.19	-30.90	-15.71	-5.19	-18.71	-3.57	-34.00	-5.21	-26.59	-11.50	-40.12
	$\Delta\mathcal{N}$	2.00	2.00	2.00	4.00	-1.00	4.00	3.00	2.00	3.00	2.00	2.00	0.00	3.00	4.00	2.00	2.00
	ΔS_Ψ	75.34	15.38	57.65	16.17	-16.59	119.10	48.71	18.56	14.23	25.46	21.14	1.30	4.47	113.93	15.63	41.31

Table A.7: Summary of the cross-fertilization procedure pairings for the MDM2 (4HFZ) and PDZ (1ZUB) receptors. Each phase involved 10 independent docking runs for both the proper (cognate) and improper (non-cognate) peptide candidates.

Phase	Receptor (PDB ID)	Proper Peptide (Cognate)	Improper Peptide (Non-cognate)
1	MDM2 (4HFZ)	p53 segment (ETFSDLWKLLP)	ELKS1b segment (CDQDEEEGIWA)
2	PDZ (1ZUB)	ELKS1b segment (CDQDEEEGIWA)	p53 segment (ETFSDLWKLLP)

Table A.8: Experimental meta-ensemble systems used in Chapter 5.3 (experimental WW domains). Year denotes the initial PDB release year, the number of conformers in the NMR ensembles, and resolution is reported only for X-ray diffraction entries.

PDB ID	# conformers	Experimental method	Resolution ($\text{\AA}$)	Year
1I5H	15	Solution NMR	-	2001
1I8G	10	Solution NMR	-	2001
1I8H	10	Solution NMR	-	2001
1JMQ	20	Solution NMR	-	2001
1K5R	10	Solution NMR	-	2001
1K9Q	20	Solution NMR	-	2001
1K9R	20	Solution NMR	-	2001
2DJY	30	Solution NMR	-	2006
2DYF	20	Solution NMR	-	2006
2EZ5	30	Solution NMR	-	2006
2JMF	20	Solution NMR	-	2007
2JO9	10	Solution NMR	-	2007
2JUP	10	Solution NMR	-	2007
2KQ0	20	Solution NMR	-	2010
2LAJ	15	Solution NMR	-	2011
2LAW	20	Solution NMR	-	2011
2LAX	20	Solution NMR	-	2011
2LAY	20	Solution NMR	-	2011
2LAZ	24	Solution NMR	-	2011
2LB0	21	Solution NMR	-	2011
2LB1	20	Solution NMR	-	2011
2LB2	25	Solution NMR	-	2011
2LB3	20	Solution NMR	-	2011
2LTV	40	Solution NMR	-	2012
2LTW	30	Solution NMR	-	2012
2LTX	25	Solution NMR	-	2012
2LTY	30	Solution NMR	-	2012
2LTZ	20	Solution NMR	-	2012
2M3O	15	Solution NMR	-	2013
2MPT	20	Solution NMR	-	2014
2N1O	19	Solution NMR	-	2015
2N8T	10	Solution NMR	-	2016
2RLY	8	Solution NMR	-	2007
2RM0	8	Solution NMR	-	2007
2HO2	1	X-ray diffraction	1.33	2007
2OEI	1	X-ray diffraction	1.35	2007

APPENDIX B: SUPPLEMENTAL FIGURES

Figure B.1: WW-domain ensemble energy distributions

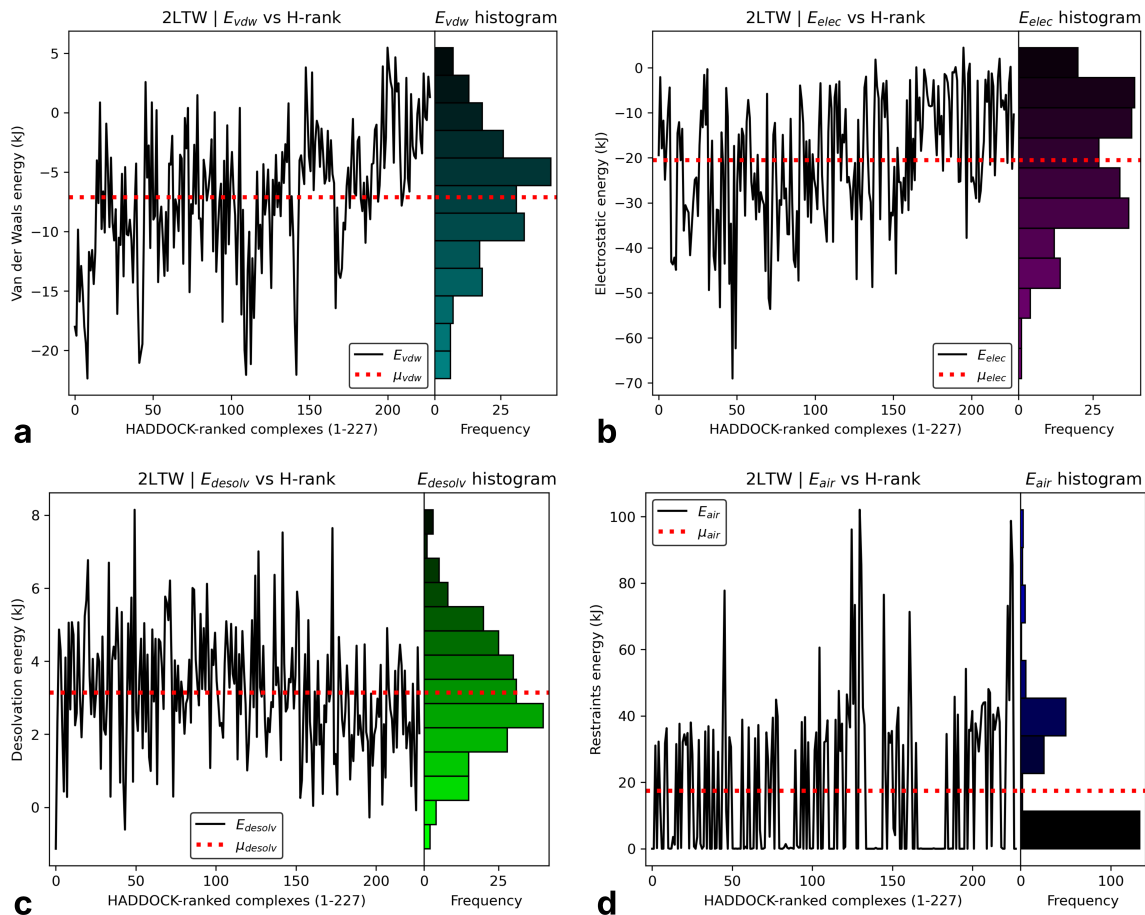


Figure B.1: Dual energy-distribution plots for the WW domain-GESPPPPYSRYPMD conformational ensemble. The x-axes represent a 227 dimensional state space where each conformational state is ranked by HADDOCK score. The y-axes correspond to the energies: (a) van der Waals; (b) electrostatics; (c) desolvation; and (d) restraint violations. A Gaussian distribution indicates consistency of that energy over the ensemble of complexes as with (c).

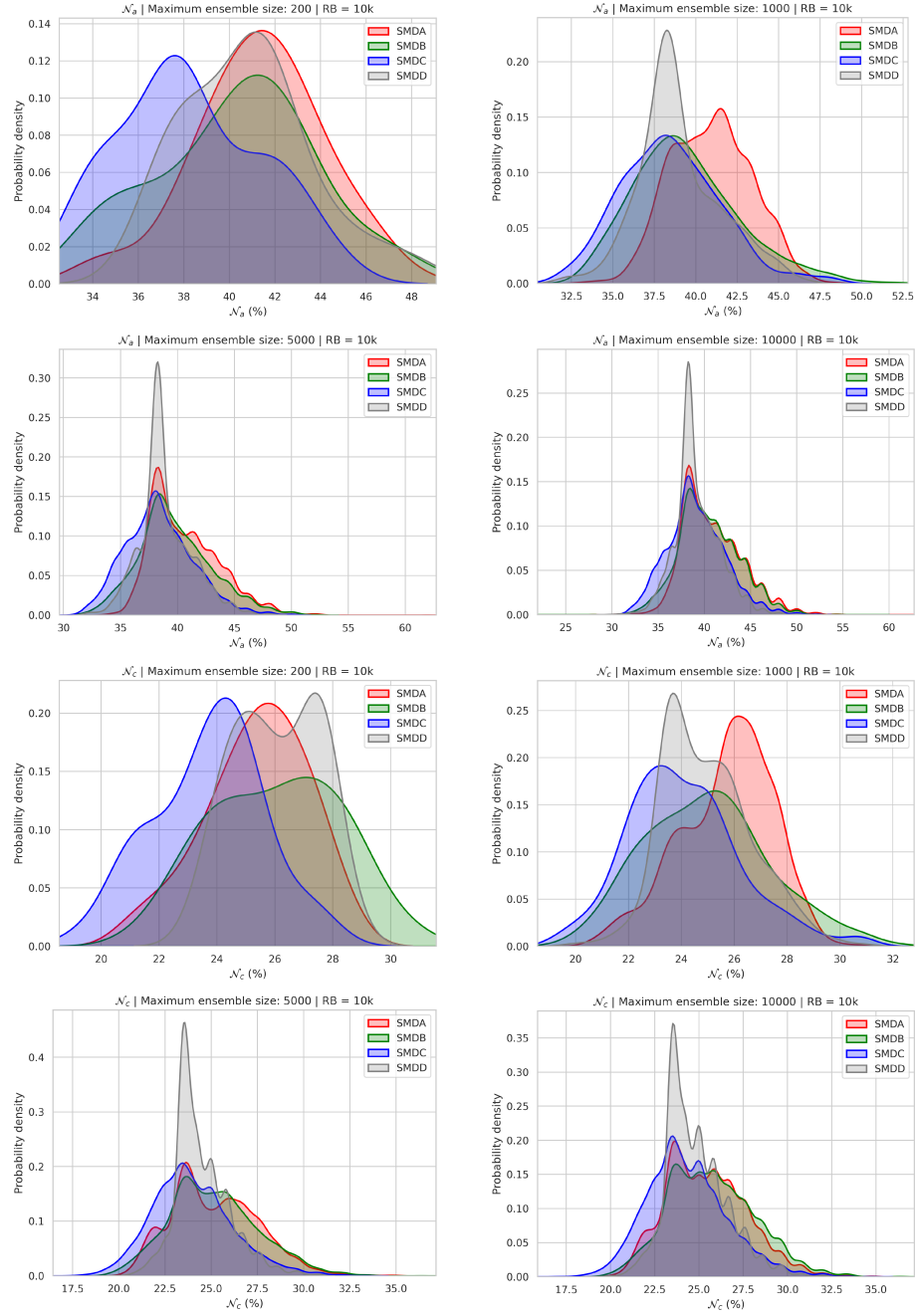
Figure B.2: Smad7 $\mathcal{N}_a$ and $\mathcal{N}_c$ distributions at RB = 10k

Figure B.2: Complete $\mathcal{N}_a$ and $\mathcal{N}_c$ distributions for the Smad7 parameter sweep at RB = 10k (corresponding to the distributions summarized in Results 1, Figure 3b). Kernel density estimates are shown for $\mathcal{N}_a$ (top four panels) and $\mathcal{N}_c$ (bottom four panels) across the four Smad7 conformations (SMDA-SMDD; legend colors). Columns correspond to the maximum retained ensemble size $n_{\text{ret}} \in \{200, 1000, 5000, 10000\}$.

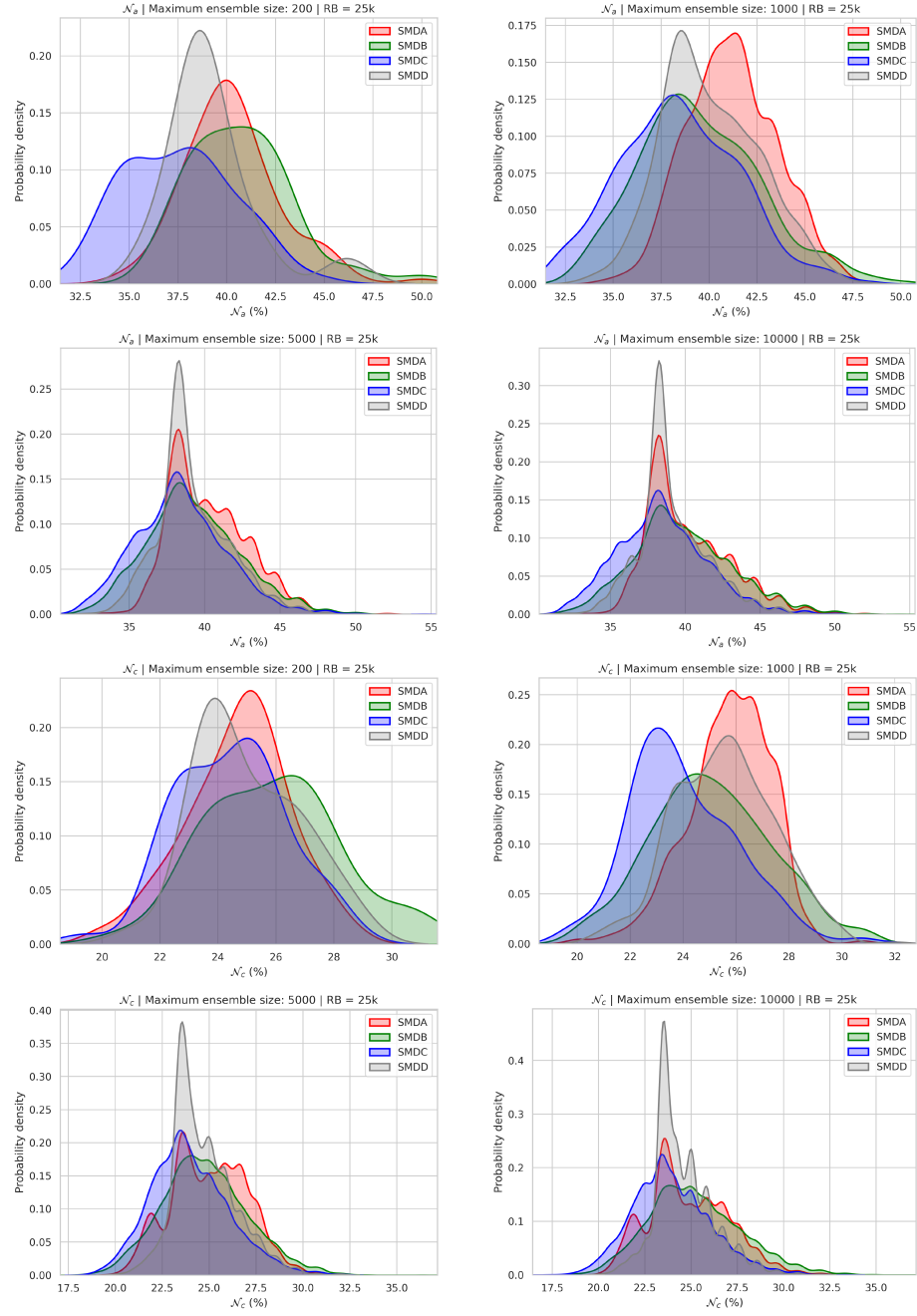
Figure B.3: Smad7 $\mathcal{N}_a$ and $\mathcal{N}_c$ distributions at RB = 25k

Figure B.3: Complete $\mathcal{N}_a$ and $\mathcal{N}_c$ distributions for the Smad7 parameter sweep at RB = 25k (corresponding to the distributions summarized in Results 1, Figure 3b). Kernel density estimates are shown for $\mathcal{N}_a$ (top four panels) and $\mathcal{N}_c$ (bottom four panels) across the four Smad7 conformations (SMDA-SMDD; legend colors). Columns correspond to the maximum retained ensemble size $n_{\text{ret}} \in \{200, 1000, 5000, 10000\}$.

Figure B.4: Cross-fertilization contact maps for 4HFZ

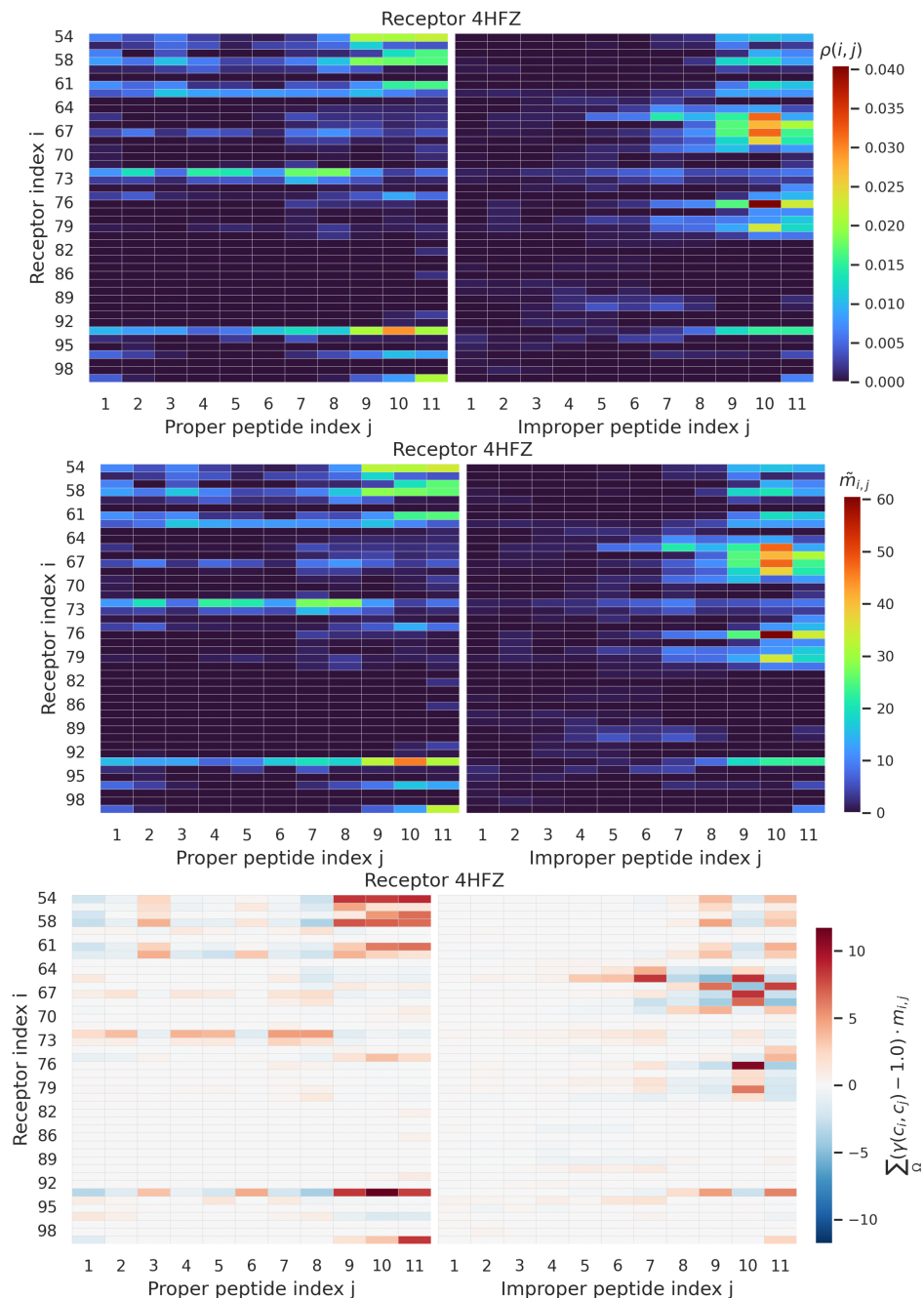


Figure B.4: Contact map summaries for receptor 4HFZ under the cross-fertilization protocol. (top) Ensemble-averaged contact map probabilities, $\rho(i, j)$. (middle) Contact-class weighted probability map, $\tilde{m}_{i,j}$, reflecting the enthalpic contribution of individual contacts. (bottom) The ensemble-level net contact-class modulation, $\sum_{\Omega} (\gamma(c_i, c_j) - 1.0) \cdot m_{i,j}$, where positive (red) and negative (blue) values denote regions of favorable and unfavorable biophysical modulation respectively.

Figure B.5: Cross-fertilization contact maps for 1ZUB

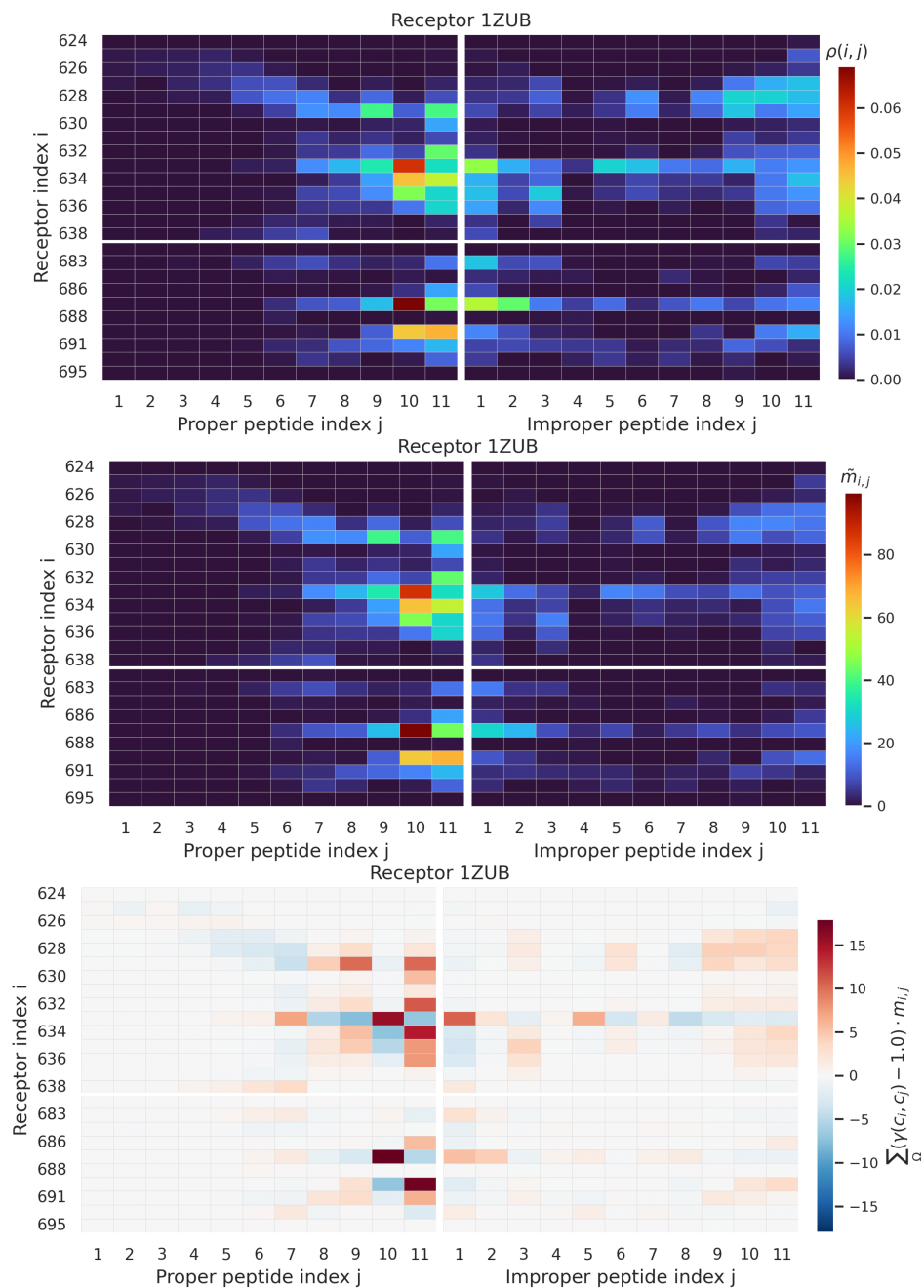


Figure B.5: Contact map summaries for receptor 1ZUB under the cross-fertilization protocol. (top) Ensemble-averaged contact map probabilities, $\rho(i, j)$. (middle) Contact-class weighted probability map, $\tilde{m}_{i,j}$, reflecting the enthalpic contribution of individual contacts. (bottom) The ensemble-level net contact-class modulation, $\sum_{\Omega}(\gamma(c_i, c_j) - 1.0) \cdot m_{i,j}$, where positive (red) and negative (blue) values denote regions of favorable and unfavorable biophysical modulation respectively. For 1ZUB, the active residues are discontinuous in sequence shown by a white horizontal space between residues 638 and 683.

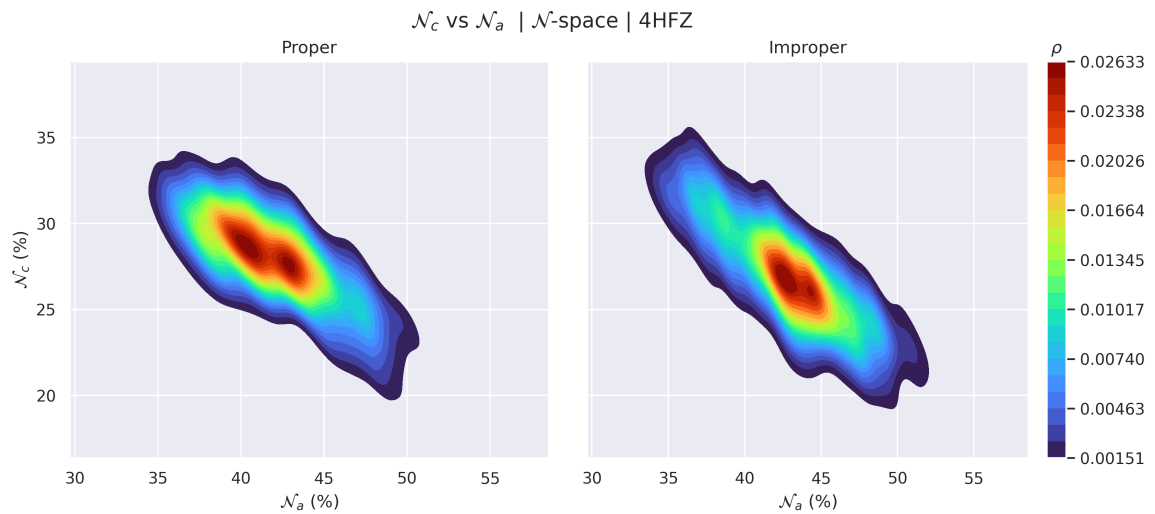
Figure B.6: Phase-pooled $\mathcal{N}$ N-space density for 4HFZ

Figure B.6: Phase-pooled bivariate $\mathcal{N}$ -space density for receptor 4HFZ under the cross-fertilization protocol (proper vs improper). Microstates are pooled across replicas within each phase.

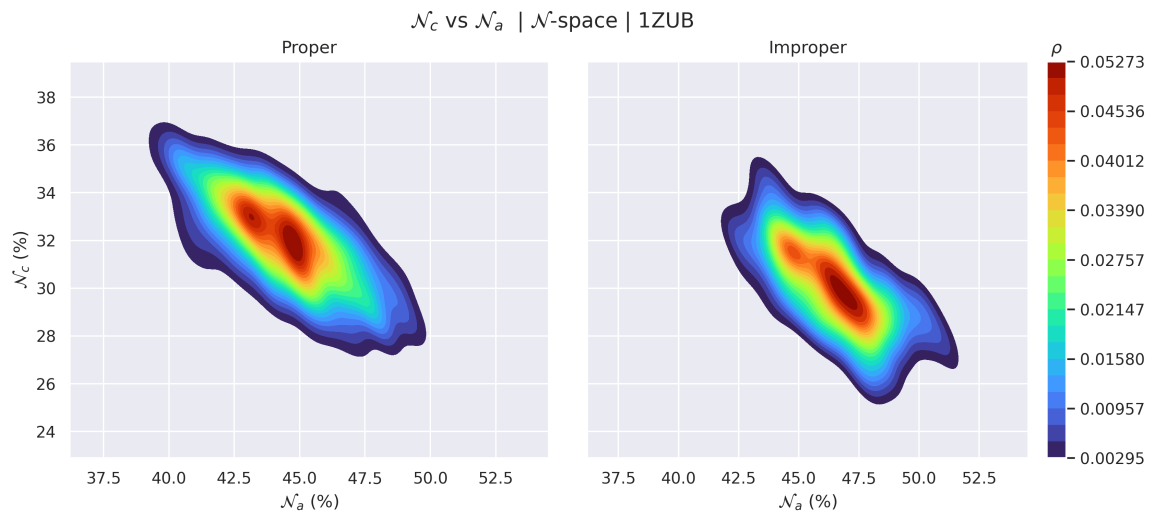
Figure B.7: Phase-pooled $\mathcal{N}$ -space density for 1ZUB

Figure B.7: Phase-pooled bivariate $\mathcal{N}$ -space density for receptor 1ZUB under the cross-fertilization protocol (proper vs improper). Microstates are pooled across replicas within each phase.

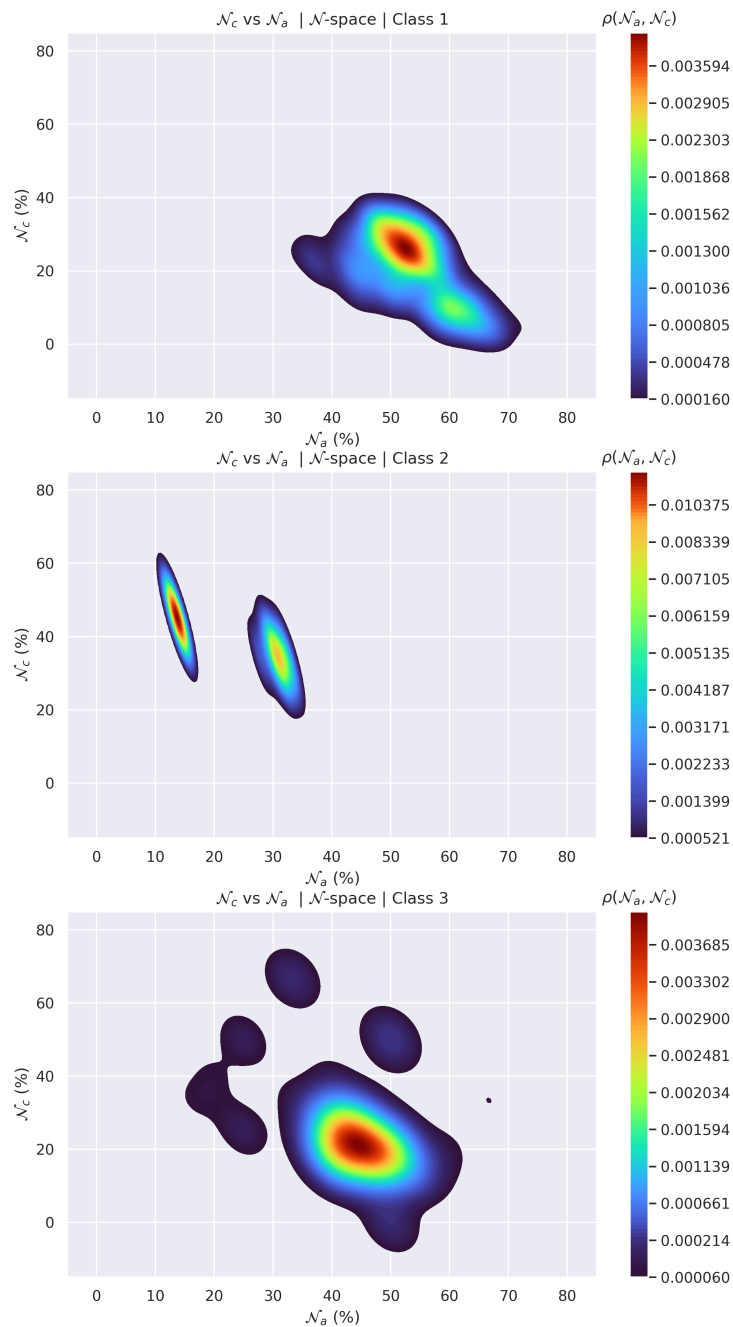
Figure B.8: Experimental WW-domain meta-ensemble $\mathcal{N}$ -space densities

Figure B.8: Phase-pooled bivariate $\mathcal{N}$ -space density for the three experimental meta-ensemble classes: (top) Class 1, (middle) Class 2, and (bottom) Class 3. Densities are computed using the RSA-thresholded count metric with explicit interface exclusion.

APPENDIX C: SUPPLEMENTAL CALCULATIONS

Calculation C.1: Phase-level ensemble means and proper-improper differences

For each receptor, two phases are considered: a proper phase $\mathcal{P}$ and an improper phase $\mathcal{I}$. Within a given phase $X \in \{\mathcal{P}, \mathcal{I}\}$, n_X independent ensembles are indexed by $e = 1, \dots, n_X$. Each ensemble e yields: (i) an unnormalized Shannon term $S'_{\Psi,e}$ (bits) computed from the macrostate probability distribution; and (ii) contact statistics that determine a phase-level normalization factor. For each phase X , the phase-pooled normalization is

$$\left(\frac{Q}{M}\right)^{(X)} = \frac{Q^{(X)}}{M^{(X)}},$$

where $Q^{(X)}$ and $M^{(X)}$ are obtained by pooling the nonzero contact pairs and their contact weights over all replicas within the phase. The reported phase-level PSI entropy is then

$$S_{\Psi}^{\text{phase},(X)} = \left(\frac{Q}{M}\right)^{(X)} \bar{S}'_{\Psi,(X)}, \quad \bar{S}'_{\Psi,(X)} = \frac{1}{n_X} \sum_{e \in X} S'_{\Psi,e}.$$

Phase differences are reported either as absolute differences,

$$\Delta S_{\Psi}^{\text{phase}} = S_{\Psi}^{\text{phase},(\mathcal{P})} - S_{\Psi}^{\text{phase},(\mathcal{I})}, \quad \Delta \left(\frac{Q}{M}\right) = \left(\frac{Q}{M}\right)^{(\mathcal{P})} - \left(\frac{Q}{M}\right)^{(\mathcal{I})},$$

or as percent differences

$$\Delta\% \left(S_{\Psi}^{\text{phase}}\right) = \left(1 - \frac{S_{\Psi}^{\text{phase},(\mathcal{P})}}{S_{\Psi}^{\text{phase},(\mathcal{I})}}\right) \times 100\%, \quad \Delta\% \left(\frac{Q}{M}\right) = \left(1 - \frac{(Q/M)^{(\mathcal{P})}}{(Q/M)^{(\mathcal{I})}}\right) \times 100\%.$$

Calculation C.2: Standard errors for S_{Ψ}^{phase} and Q/M

Ensembles are treated within a phase as independent replicas, so error bars represent the standard error (SE) of the phase mean, estimated from the across-ensemble

variability. For S'_Ψ , the sample variance in phase X is

$$\sigma_{S'_\Psi}^{2,(X)} = \frac{1}{n_X - 1} \sum_{e \in X} (S'_{\Psi,e} - \bar{S}'_{\Psi,(X)})^2,$$

and the corresponding SE of the mean is

$$\text{SE}(\bar{S}'_{\Psi,(X)}) = \frac{\sigma_{S'_\Psi}^{(X)}}{\sqrt{n_X}}.$$

For the ensemble unit-measure ratio Q/M , the sample variance is

$$\sigma_{Q/M}^{2,(X)} = \frac{1}{n_X - 1} \sum_{e \in X} \left(\left(\frac{Q}{M} \right)_e - \overline{\left(\frac{Q}{M} \right)}^{(X)} \right)^2,$$

with corresponding standard error

$$\text{SE}\left(\overline{\left(\frac{Q}{M} \right)}^{(X)}\right) = \frac{\sigma_{Q/M}^{(X)}}{\sqrt{n_X}}.$$

Error in the reported values and figures correspond to these phase-specific SE calculations for $\bar{S}'_{\Psi,(X)}$ and $\overline{(Q/M)}^{(X)}$ in the proper and improper phases.

Calculation C.3: Effective number of macrostates and normalized effective state fraction

Pooling all microstates across replicas within a phase yields a phase-pooled macrostate probability distribution

$$p_i = \frac{g_i}{\Omega}, \quad \sum_i p_i = 1,$$

where g_i is the pooled macrostate multiplicity and Ω is the total number of pooled microstates. The unnormalized Shannon entropy (bits) is

$$S'_\Psi = - \sum_i p_i \log_2 p_i,$$

and the corresponding Shannon effective number of macrostates (for $\log_x$ with base 2) is defined as

$$N_{\text{eff}} = 2^{S'_{\Psi}}.$$

For convenience, the normalized effective state fraction is reported as N_{eff}/Ω , which measures the effective support of the pooled macrostate distribution relative to the total microstate count.

For the two receptors, the phase-pooled values are:

	$\Omega_{\mathcal{P}}$	$N_{\text{eff}}^{(\mathcal{P})}$	$\Omega_{\mathcal{I}}$	$N_{\text{eff}}^{(\mathcal{I})}$
4HFZ	476	220.5	501	258.2
1ZUB	631	303.3	247	155.3

with corresponding normalized effective state fractions N_{eff}/Ω of 0.463 (4HFZ proper), 0.515 (4HFZ improper), 0.481 (1ZUB proper), and 0.629 (1ZUB improper).

Calculation C.4: Autocorrelation function

For a scalar observable x_t sampled along an equilibrium trajectory at constant frame interval Δt , the normalized autocorrelation function (ACF) at a lag of ℓ frames is defined as

$$C(\ell) = \frac{\langle (x_t - \bar{x})(x_{t+\ell} - \bar{x}) \rangle}{\langle (x_t - \bar{x})^2 \rangle}, \quad C(0) = 1,$$

where $\bar{x} = \langle x_t \rangle$ is the trajectory mean and $\langle \cdot \rangle$ denotes the time average. $C(\ell)$ measures the extent to which the signal retains memory of its earlier value after a lag of $\ell \Delta t$. A large positive $C(\ell)$ indicates that successive frames remain strongly correlated, whereas decay of $C(\ell)$ toward zero indicates progressive decorrelation.

A practical decorrelation time may be defined as the first lag at which the ACF drops below $1/e$, where $1/e \approx 0.368$. This criterion provides an operational timescale beyond which residual memory is modest, and it is convenient for choosing a decimation stride for histogramming, PMF reconstruction, and ensemble averaging. For

the MDM2 lid collective variables, frames were saved every $\Delta t = 20$ ps. In the pooled CV1 ACF, the first crossing below $1/e$ occurred at $\ell_{1/e}^{(\text{CV1})} \approx 594$ frames. The corresponding decorrelation time is

$$\tau_{\text{CV1}} \approx 594 \times 20 \text{ ps} = 11,880 \text{ ps} \approx 11.9 \text{ ns} \approx 12 \text{ ns}.$$

Using this value, the stride in trajectory frames is

$$\frac{\tau_{\text{CV1}}}{\Delta t} = \frac{12,000 \text{ ps}}{20 \text{ ps/frame}} = 600 \text{ frames}.$$

Accordingly, a stride of 600 frames was adopted for the decimated analyses in the main text, corresponding to one retained sample approximately every 12 ns. CV2 reports a slower collective mode than CV1, as shown in Figure C.1. The 12 ns stride was defined from the CV1 decorrelation analysis and then used consistently for the decimated chemical-shift, PMF, and free-energy calculations reported.

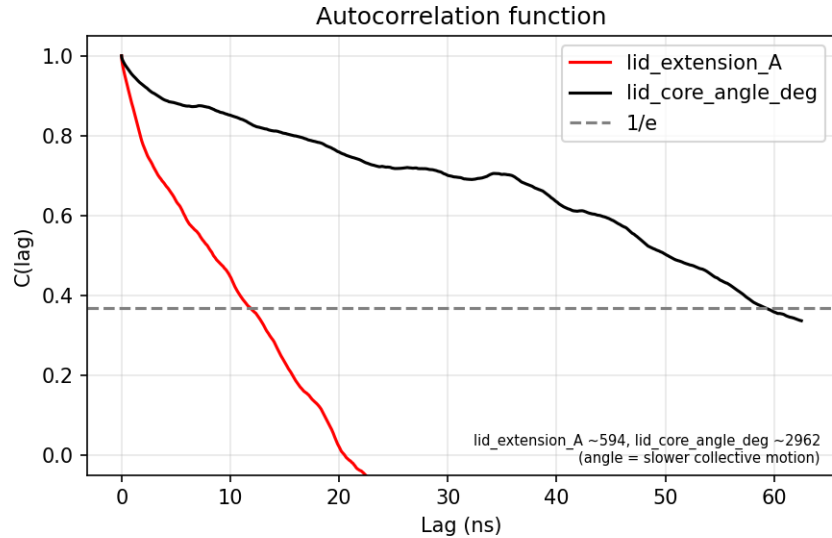


Figure C.1: Autocorrelation functions for CV1 (lid extension, red) and CV2 (lid-core angle, black) computed from the isotonic simulation ensemble. The dashed horizontal line indicates the $1/e$ threshold used to estimate characteristic decorrelation time. CV1 decays more rapidly, whereas CV2 remains correlated over substantially longer lag times, consistent with slower collective relaxation of the lid-core angular DoF.