MOLECULAR LEVEL UNDERSTANDING OF THE FUNCTIONALITY OF PDZ3 VARIANTS VIA ADVANCED ALL-ATOM SIMULATIONS AND DYNAMIC RESIDUE NETWORK ANALYSES

by

TANDAÇ FÜRKAN GÜÇLÜ

Submitted to the Graduate School of Engineering and Natural Sciences

in partial fulfillment of

the requirements for the degree of

Doctor of Philosophy

Sabancı University

May 2021

MOLECULAR LEVEL UNDERSTANDING OF THE FUNCTIONALITY OF PDZ3 VARIANTS VIA ADVANCED ALL-ATOM SIMULATIONS AND DYNAMIC RESIDUE NETWORK ANALYSES

APPROVED BY:

DATE OF APPROVAL: 17/05/2021

ABSTRACT

MOLECULAR LEVEL UNDERSTANDING OF THE FUNCTIONALITY OF PDZ3 VARIANTS VIA ADVANCED ALL-ATOM SIMULATIONS AND DYNAMIC RESIDUE NETWORK ANALYSES

Tandaç Fürkan Güçlü

Molecular Biology, Genetics and Bioengineering, Ph.D. Thesis, 2021

Thesis Supervisor: Canan Atılgan

Thesis Co-supervisor: Ali Rana Atılgan

Keywords: PDZ3 domain, Molecular Dynamics, Free Energy Perturbation, Graph Theory Girvan-Newman Algorithm.

The third PDZ domain of PSD-95 (PDZ3) constitutes a common model to study single domain allostery without significant structural changes. In PDZ3, H372 directly connected to the binding site and G330 holding an off-binding-site position, were designated to assess the effect of mutations on binding selectivity. It has been observed that the H372A and G330T-H372A mutations change ligand preferences from class I (T/S amino acid at position -2 of the ligand) to class II (hydrophobic amino acid at the same position). Alternatively, the G330T single mutation leads to the recognition of both ligand classes. We have performed a series of molecular dynamics (MD) simulations for previously mentioned PDZ3 variants in the absence and presence of both types of ligands. With the combination of free energy difference calculations and a detailed analysis of MD trajectories, binding behavior of PDZ3 mutants, as well as their effects on ligand selection and binding affinities are explained. To scrutinize the residue-by-residue interaction we employ graph theory, and we assess dynamical community composition by using Girvan-Newman algorithm. We find that the highly charged and distal N-terminus share the same community with the ligand in the functional complexes. N- and C-termini of PDZ3 share communities, and α3 acts as a hub for the whole protein by sustaining the communication with all structural segments. Thus, ligand binding fate in PDZ3 is traced to the population of community compositions extracted from dynamics despite the lack of significant conformational changes.

ÖZET

ATOMİK ÇÖZÜNÜRLÜKTE BENZETİMLER VE DİNAMİK AMİNO ASİT ÇİZGE ANALİZLERİ KULLANILARAK PDZ3 VARYANTLARININ FONKSİYONUN MOLEKÜLER SEVİYEDE ARAŞTIRILMASI

Tandaç Fürkan Güçlü

Moleküler Biyoloji, Genetik ve Biyomühendislik, Doktora Tezi, 2021

Tez Danışmanı: Canan Atılgan

Tez İkinci Danışmanı: Ali Rana Atılgan

Anahtar kelimeler: PDZ3 proteini, Moleküler Dinamik, Serbest Energy Sarsımı, Çizge Kuramı, Girvan-Newman Algoritması

PSD-95 yapısının üçüncü PDZ proteini (PDZ3), önemli yapısal değişiklik göstermeden sergilediği alosterik özelliğinden ötürü sık kullanılan bir model proteindir. PDZ3 proteininde H372 ligand bağlanma bölgesiyle direkt ilişkili, G330 ise bu bölgeden uzakta yer alır; her iki amino-asit bölgesi de mutasyona uğradığında, ligand seçimini etkiler. Literatüre göre, H372A ve G330T-H372A mutasyonları sınıf I'den (ligandın -2 pozisyonundaki T/S rezidülerine karşı gelen) sınıf II'ye (aynı pozisyonda hidrofobik rezidü) dönüşen ligand bağlanmasına sebep olurken, G330T tekli mutastonu her iki liganda da bağlanmayla sonuçlanır. Bu olguyu araştımak için, ligandsız ve iki liganda bağlı her yapı için moleküler dinamik (MD) benzetimleri yürüttük. MD yörüngelerinin detaylı analizleri ve serbest enerji farkı hesaplamalarıyla birlikte PDZ3 mutant yapılarının ligand bağlanma davranışı ve seçimi açıklandı. Rezidüler arası etkileşimi araştırmak amacıyla, çizge kuramını kullandık ve çizgelerdeki komünite yapılarını ve içeriğini Girvan-Newman algoritmasıyla inceledik. Yüklü yapısıyla N-ucunun, fonksiyonel PDZ3 komplekslerinde, ligand ile aynı komünitede daha çok zaman geçirdiğini tespit ettik. N ve C uçları yüksek oranda aynı komünitede bulunmakta ve α_3 de bir merkez gibi davranarak tüm iç iletişimi sağlamaktadır. Sonuç olarak, PDZ3 proteninin ligand seçimi, büyük yapısal değişiklikler göstermemesine rağmen, komünite yapılarıyla açıklandı.

Viva fui in sylvis, sum dura occisa securi,

dum vixi, tacui, mortua dulce cano.

ACKNOWLEDGEMENTS

I would like to thank my advisors Canan Atilgan and Ali Rana Atilgan for their support. Especially, the humane and collective research environment they created was incomparable.

Deniz Sezer was my nemesis in PhD. He always pushed me to my limits in terms of knowledge. With his nice remarks and critics, I improved myself significantly. Afterwards, we had a lot of conversations with him about many subjects.

Further, Zehra Sayers, Mert Gur, Ogun Adebali were in my jury; I thank them for their comments, and even their nice e-mails made me happy. Especially Ozge Sensoy who had been in my progress committee for years. I really appreciate her comments and questions through my PhD. Onur Varol was also present as an 'honorary jury member', I thank him for coming and his questions/remarks. Smiling faces of the committee members were really encouraging.

Ozlem Tastan Bishop invited me to South Africa and gave me a research opportunity. It has been my only foreign country experience, and I thank her for this great chance.

I want to mention my friends outside the office, and they are Tugdem, Kadriye, İpek, Senem, Tugce, Omid, Canhan, Deniz and Leila. I had a great time with them.

Esra and Berika were my friends from another dimension, away from the engineering, numbers and codes. They were a hidden refuel to my soul, and our conversations about literature, politics and art were invaluable. Also, I never ever should forget the food and having me as a guest at their dorms. Our Starbucks (actually, the name was different) times with Berika were always soothing and happy.

Ronay, Ozlem, Oznur, Sevde, Liyne and Irem were from the faculty. We share talks, coffee and meals together; they were a huge support in those 'short' breaks. Ceren's smile and gifts were always nice; she (with Oguzhan) had a surprise party for my thirtieth age.

At first there was Sofia, then I met with Kaan, Haleh and Goksin. Sofia taught me everything about this foreign place called 'Sabanci University', and I still do not forget fruit juice at night to empower me to finish my work. She meticulously helped me with science/music and

even with my all in-campus room transfers, never complained. We played guitar and had some ‘mini’ concerts with Kaan, that was fun. We always talk about various subjects with Goksin, and she was the bridge connecting me to people. She always supported me, particularly after the qualification, her kind and warm speech encouraged me a lot. She is one of the most ‘humane’ elements during my PhD. Ömer helped me a lot with the codes, when I was trying to settle down and have a working environment. We had evening coffee with him, and that was valuable. We spent quality time with Aygul, talking about books, music and movies; when I got interested in ‘intellectual’ things, she stimulated me with her discussion and knowledge. Caroline came to our lab, and we had a great time together. I ‘taught’ her the little knowledge about Istanbul I know. Then, when I was in South Africa, she opened her house to me and introduced me to different people, rendering my social knowledge wider. I learnt a different set of things from her, and it was unforgettable. Ebru usually criticized my ‘pessimist scientist’ side, where I bragged about today’s scientific view and PhD. She always inspired me to tolerate those thoughts, so that I can finish my PhD. I had most of my scientific talks with Ebru, which I appreciate. Kurt showed me different views about many things; our conversations about behavior and language changes between the cultures were very informative. He never got offended with my sarcastic behavior and always talked with me nicely. Tamay was the hard-working ant of our lab; she learnt and elevated herself very fast, which inspired me significantly. Işık was the person that understood my quirky behavior and language swiftly; she taught me ‘new’ things by fighting my stubborn aspect of the new. Her approach to me always enlightened my dark mood, that I expertly hid under a smile. Erhan started as an intern, ended up as a colleague; he is gentle and thoughtful, impartial from the people he faced, therefore, he tolerated and befriended me beautifully.

Nazlı started the PDZ3 project and ran most of the MD simulations. She started the FEP pipeline leading me to automate the process. We worked together to understand the function of PDZ3 and continued working, after she was in Canada. As a smart colleague and an excellent friend, she always supported me during my PhD. I learnt a lot from her, especially, her presentation style led me to improve my presentation skills. Her positive effect on the environment was important for the collectivity, and her selfless acts of helping the people around made a significant impact.

Oguzhan continued a work about DPD and with his questions, he added me to the process, leading me to learn those methods. His struggle to improve everything around him and constant helping to people made a difference in the office. He also connected many people enabling an interacting environment. I learnt a lot from his ‘new gadgets’; his appetite to do new things was inspiring. Unlike the current trend to use science as a stepping stone to go abroad, he was thrilled to do science.

My students from the lectures and undergrad projects were always nice; in particular, I want to mention Ogulcan, Umit, Gokce and Zeynep. I hope I contributed a bit to their knowledge.

My family always left their star burning. They supported me indefinitely.

In this garden of academia, I witnessed a vast amount of selfish and pragmatist people living without a glimpse of philosophical depth; measuring quality by only success, which will end up as a few lines in a cv. On the other hand, under the name of freedom and individuality, there are a significant number of people who turn their back to everything. Thankfully, my advisors were a soothing and directing narration through my PhD, and they enveloped me with humane and collective people. In this good environment, I tried to connect with many people to set an example of this ideal. Also, by exchanging thoughts, I hope to reflect this behavioral pattern, which will be disseminated, evolved and corrected through the flow of time. So that, those fragments of ideas might make a difference in the future. I believe in this process; since, this is the sole method that I know to rebel against this cruel world.

This work was funded by the TÜBİTAK 117F389 project.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVATIONS

°C	Celsius degrees
Å	Ångström
BC	Betweenness Centrality
C	Clustering coefficient
Cl	Chlorine
CRIPT	CXXC Repeat Containing Interactor of PDZ3 Domain
DM	Double Mutant
FEP	Free Energy Perturbation
fs	femtosecond
GK	Guanylate Kinase-Like Domain
K	Potassium or Kelvin
k	Degree of a node
kcal	kilocalorie
L	Shortest-path length of a node
MAGUK	Membrane-Associated Guanylate Kinases
MD	Molecular Dynamics
MM-GBSA	Molecular Mechanics-Generalized Born Surface Area
MSA	Multiple Sequence Alignment
MuMi	Mutation and Minimization
ns	nanosecond
PDB	Protein Data Bank

PDZ	Post synaptic density protein, Drosophila disc large tumor suppressor and Zonula occludens-1 protein
PSD-95	Postsynaptic density protein 95
RDF	Radial Distribution Function
RGB	Red Green Blue
RMSD	Root Mean Square Deviation
RMSF	Root Mean Square Fluctuation
RN	Residue network
SASA	Solvent Accessible Surface Area
SCA	Statistical Coupling Analysis
SH3	SRC Homology 3 Domain
WT	Wild-Type

1. INTRODUCTION

1.1 The third PDZ domain of PSD-95 Complex

PDZ domains are abundant in various complexes, and the well-studied PSD-95 is from the MAGUK family with a PDZ-SH3-GK pattern. The third PDZ (PDZ3) domain of this complex is important for the formation of this supramodule and its binding to C-terminal ligands.[1-6] Most of the PDZ domains have two α-helices and six β-strands. However, PDZ3 is unique with an extra α-helix (α_3) located at its C-terminus. α_3 has been shown to be important in the folding of PDZ3 as well as its interactions with SH3 and ligands.[1, 5, 7-11]

The structure of wild-type (WT) PDZ3 has been solved with its cognate ligand CRIPT;[12, 13] it has been demonstrated that PDZ3 binds specific motifs of ligands.[14] The ligands are classified by amino acid type at the second position from the C-terminus (-2 position), with those having Thr/Ser defined as Class-I; this class contains CRIPT.[13] Class-II has hydrophobic, and Class-III has Asp/Glu amino acids at this position.[14]

In this thesis, we focus on the WT, G330T, H372A and G330T-H372A (double mutant, DM) variants in complex with the ligand CRIPT (L_1) and its T-2F form (L_2).[15-17] These theoretical variants of PDZ3 have been investigated by using mutational scans, their functionality have been assessed by experimental binding constants, and their crystal structures have been deposited.[16-19] In an elegant work on PDZ, Raman et al. have shown that the adaptive evolutionary pathway for switching ligand binding from Class I to Class II utilizes a class bridging mutation such as G330T while the H372A mutation may only be gained at a second step.[17] Hence, the binding experiments display that the WT forms a functional complex only with L_1 (Figure 1.1a). On the other hand, the G330T single mutant prefers binding to both L_1 and L_2 (Figure 1.1b); however, H372A single mutant and DM proteins favor only L_2-binding (Figure 1.1c, d).[16] Thus, of the eight possible complexes, WT_{L1}, $G330T_{L1}$, $G330T_{L2}$, $H372A_{L2}$ and DM_{L2} are functional, while WT_{L2}, $H372A_{L1}$ and DM_{L1} forms are unfavorable.

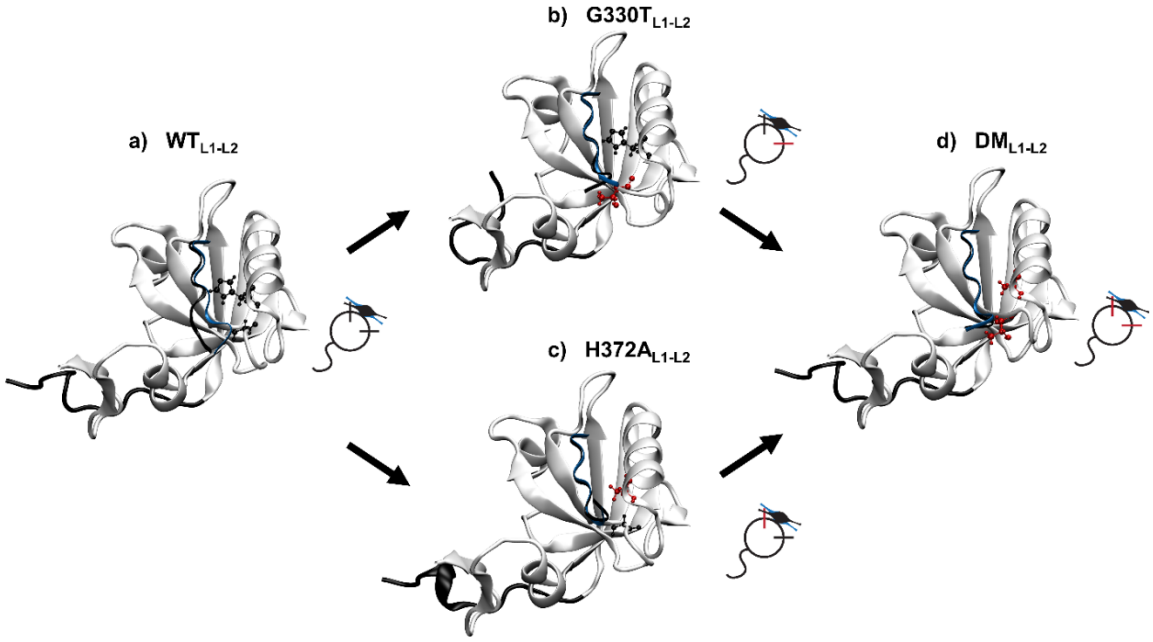

Figure 1.1 PDZ domain complexes for WT, G330T, H372A and DM cases are visualized with the mutations and the ligands. On the protein structure, G330 and H372 sites are shown in black for the wild-type form, and the mutated residues are shown in red. CRIPT (L_1) and N-terminus region (residues between 299 and 310) are shown as black ribbons, while T-2F (L_2) is illustrated in blue. The shorthand cartoons are used to differentiate each case: Circle represents the body of the protein; ticks, whiskers and violin shapes correspond to the mutation sites, the N-terminus regions and the ligands, respectively. The color code for each component in cartoons is the same as the protein structure. Residue 330 is not directly located at the binding site, whereas 372 directly interacts with the ligand. **a** WT structure binds strongly to the L_1. **b** G330T mutant is the class bridging mutant; thus, it binds both L_1 and L_2 with similar free energy difference. **c, d** H372A and DM structures prefer to bind L_2 with a lower energy value; therefore, H372A mutation is defined as the class switching mutant.

1.2 Synopsis of the Thesis

Mutation-Minimization (MuMi)[20] method is applied on a protein ensemble consisting of forty non-homologous PDB structures (see Materials and Methods). Alanine scanning is used to investigate the mutational perturbations for each protein, and emergent structures are analyzed as residue networks (RNs) by assessing degree (k), clustering coefficient (C), average shortest path length ($<L>$) and betweenness centrality (BC) (Figure 1.2a). After the

insight of the ensemble view of static RNs, we focus on PDZ3 to examine the molecular basis of biological function by employing more detailed methods and dynamic RNs.

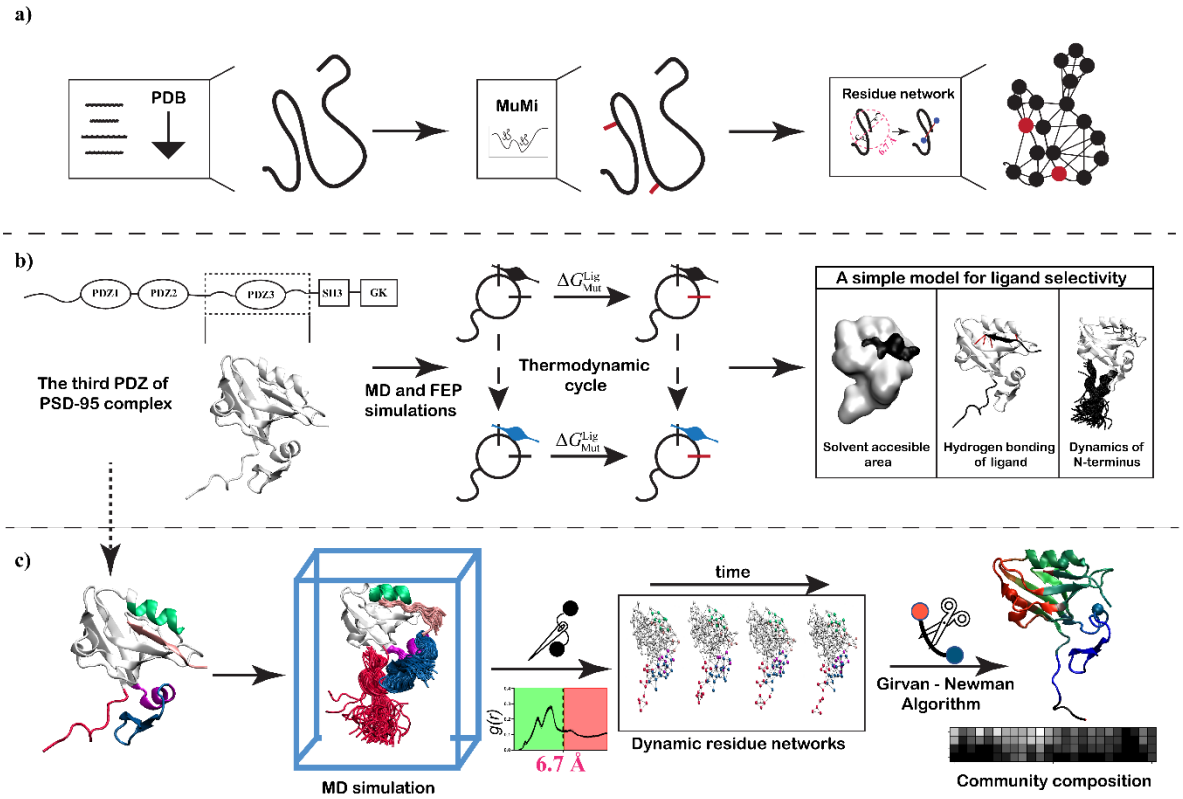

Figure 1.2 Scope of the thesis. **a** MuMi scheme is visualized: 40 PDB structures are downloaded, mutated, and minimized; then each is coarse-grained to a RN. **b** The detailed investigation of PDZ3: MD and FEP simulations are performed, and binding/mutation is assessed by using thermodynamic integration. A simple model is constructed to explain the binding energies by using solvent accessible area, hydrogen bonds between protein-ligand and solvent interaction of charged residues in N-terminus as the main contributing factors. **c** Community composition study of PDZ3; following the MD simulations, the graphs are constructed for one ns apart snapshots. Then, communities are detected by the Girvan-Newman algorithm, and their composition is investigated and visualized by RGB color-codes.

Hence, to investigate in detail, we perform molecular dynamics (MD) simulations for the apo and L_1/L_2-bound form of the WT, G330T, H372A and DM PDZ3 structures (Figure 1.1). By

using the conformations obtained from the MD trajectories, we conduct free energy perturbation (FEP) simulations[21, 22] to investigate the energy cost of the mutations. Integrating the binding and mutation energies into thermodynamic cycles allows us to relate the computational results to the experimental binding energies from a previous study.[16] Further, our detailed analyses of the MD trajectories reveal that the charged N-terminus region of PDZ3 has a significant impact on the ligand specificity in addition to the direct interactions occurring at the binding site. To test this hypothesis, we replicate the simulations on the N-terminus truncated complexes. We show that the free energy differences leading to the class bridging/switching behavior are nullified in the absence of the N-terminus. Thus, we demonstrate the electrostatic contributions due to the dynamics of the N-terminus region is essential for the formation of the functional PDZ complexes (Figure 1.2b, see ref[23]).

To understand the molecular basis of binding, we investigate the previously studied structural segments, which are the C terminus, α_2 helix and the aforementioned α_3 helix (Figure 1.3),[5, 9-11, 23-25] along with the N-terminus. The dynamics of the highly charged N-terminus region is shown to be important in the binding mechanism, especially due to its electrostatic contributions to the total free energy.[23] The effect of the N-terminus might be partnered with that of the C-terminus.[24] The α_2 helix lines up the ligand, and its effect has been investigated by deep mutational scanning.[25] Mutants of H372, residing on the α_2 helix, have been shown to cause significant binding constant shifts when coupled with other point mutations, and this effect is conserved in other PDZ domains.[16, 25] In PDZ3, H372 is in direct contact with the T-2 residue of L_1 implying an essential role for ligand binding.[12, 16, 25, 26]

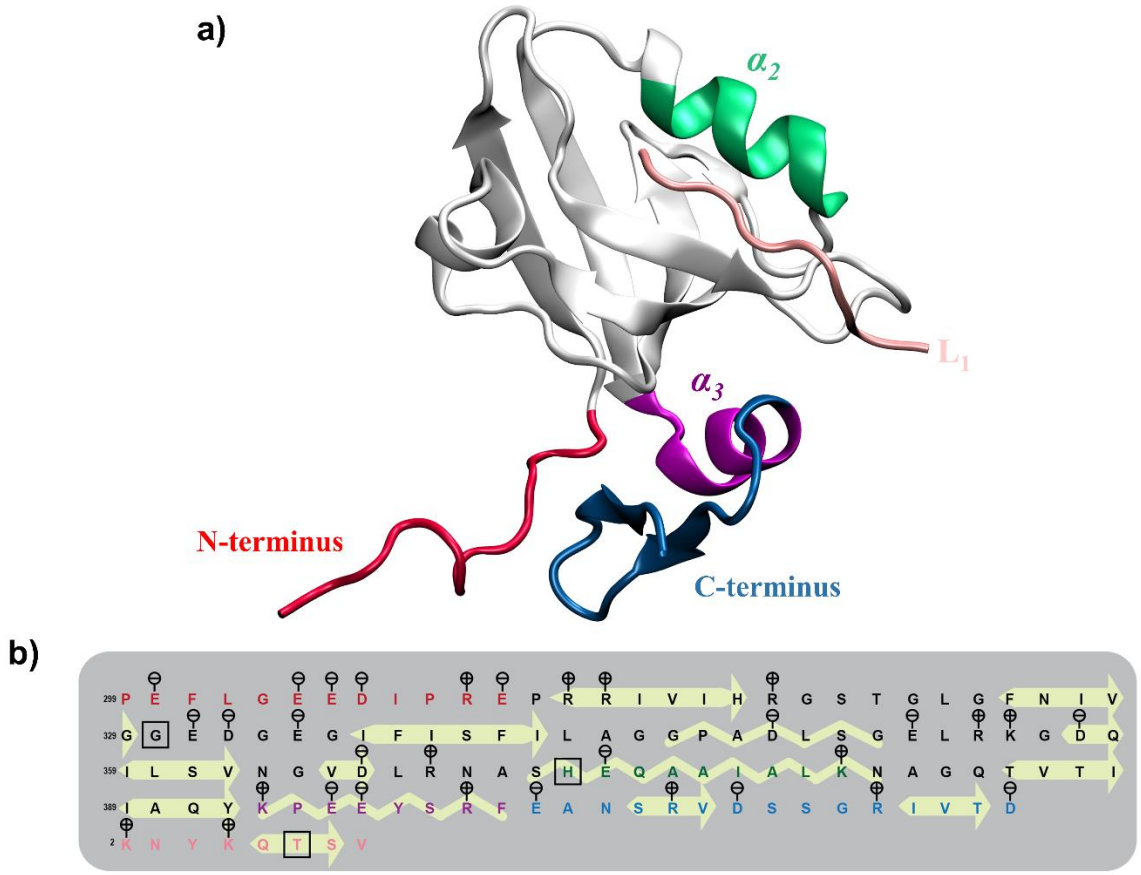

Figure 1.3 **a** Structure and sequence of WT_{L1} complex (PDB code 5HEB). N-terminus (residues 299-310), α_2 (residues 372-380), α_3 (residues 393-400), C-terminus (residues 401-415) and ligand (residues 2-9) are illustrated in red, green, purple, blue and magenta, respectively. **b** PDZ sequence and details of the structural elements. Charges of residues are marked in circles above amino acid symbols. Yellow shapes indicate secondary structures, arrow for β-strands and zigzag for α-helices. Squared-residues display the mutation positions, which are G330, H372 in the protein and T-2 in the ligand.

To further our understanding on the role of structural segments on ligand specificity, we employ graph theory to investigate the communication between these functionally important structural segments.[20, 27-29] Coarse-graining the protein structure and projecting it to a RN reduces that structure to nodes (vertices) and edges (links).[30-32] In an RN, centrality of nodes reveals the residues that manipulate information flow, and identifying residues with high centrality provides a profile of biological function and evolutionarily conservation.[20, 33-36] On the other hand, focusing on edge centrality in order to detect 'communities' has been put

forth as an approach to distinguish modular units with interdependent functions for any network,[37-39] and these ideas have been applied to residue networks to shed light on communication patterns between structural segments.[37-39] Here, we analyze MD simulations to gather a range of conformations sampled by PDZ3 with a novel approach. We first apply network analysis on extracted MD snapshots;[40] then, we relate the dynamical changes of community members to their structural and functional origins. Community analysis reveals hidden allostery in protein structures by assessing the communication scenarios between the structural segments.[38, 39] We propose that this method may be used to cluster the conformational dynamics of protein structures, and to infer information flow underlying functional mechanisms (Figure 1.2c, see ref[41]).

2. MATERIALS AND METHODS

2.1 MD and FEP Simulations

2.1.1 PDB Files and MD Simulations

Protein structures are downloaded from PDB (Table 2.1, 2.2).[42] For PDZ3 complexes, sequences of protein and ligand are arranged to be between 299-415 and 2-9 using SWISS-MODEL[43] server. MD simulations are performed by using NAMD[44, 45] software, and CHARMM36[46] force-field is used for topologies and parameters. VMD[47] is utilized for preprocessing of structures, such as structure specific topology file (PSF file) generation, solvation (constructing water-box), ionization and visualization of MD trajectories. By using the solvent plug-in VMD 1.9.3, protein structures are solvated in a rectangular water box with a minimum distance of 10 Å between the protein and the nearest edge of the water box.

Charge of the system is neutralized, and the ionic strength is tuned to 150 mM by adding a sufficient number of potassium chloride ions (KCl). The particle mesh Ewald method[48] is utilized for long-range electrostatics with a cut-off distance of 12 Å. Temperature control is maintained by Langevin dynamics. The system is simulated in the NPT ensemble achieving constant 1 atm and 310 K. The time step is 2 fs; hence, 10,000 steps of minimization and 100,000,000 steps of equilibration are conducted for PDZ3 systems. PDZ3 complexes of WT, G330T, H372A and DM for apo, L_1 and L_2 forms (Table 2.2) are simulated for 200 ns, the runs are duplicated to enhance sampling. Nevertheless, the N-termini-truncated versions of the ligand-bound forms are run for 50 ns each.

2.1.2 Trajectory Analyses

The equilibrated last 120 ns of each MD simulations is divided into three equal 40 ns chunks, and a total of six chunks for each PDZ3 complex are utilized for RMSF (root-mean-square fluctuation), cross-correlation, hydrogen-bond occupancy and SASA (solvent-accessible-surface area) calculations. The first frame of each trajectory is used as a reference for RMSD (root-mean-square deviation) and RMSF calculations. $N \times N$ cross-correlation matrix is calculated by taking the trace of each 3×3 element of the covariance matrix to observe the correlated motions between residues.[49] Heatmaps of cross-correlation matrices are normalized between -1 and 1 for visualization purposes. Radial distribution function ($g(r)$), hydrogen bond and SASA analyses are performed in VMD with the default settings.[47] Additionally, manipulation of trajectory files and basic analyses are done by using ProDy[50] package of python programming language.

2.1.3 FEP Simulations

Free energy changes of the system space between reference (A) and target (B) states is sampled in forward/backward directions through the coupling parameter, λ (0→1).[51] The free energy difference is calculated by,[52]

$$\Delta F(\mathbf{A} \rightarrow \mathbf{B}) = -k_B T \langle exp\left(-\frac{E_{\mathbf{B}}-E_{\mathbf{A}}}{k_B T}\right)\rangle_{\mathbf{A}} \quad (2.1)$$

where the energy difference between states A and B is denoted with ΔF, and k_B and T are the Boltzmann constant and temperature, and the energy of state A is $E_{\mathbf{A}}$. The angular brackets indicate ensemble average over a trajectory for state A. While each 200 ps long windows consist of 50 ps equilibration and 150 ps data generation, through 32 window, λ varies between WT (λ = 0) and mutated (λ = 1) states to maintain overlaps of probability distributions.[53] The average energy change is calculated by using Bennet acceptance ratio algorithm[54] to minimize the error.

To produce a large variety of energy data, the starting structures for FEP simulations are selected from the 50, 100, 150 and 200 ns time points of the two duplicate MD simulations. Thus, the FEP calculation are conducted for 8 times for each PDZ3 complex. For N-terminus removed PDZ3 complexes, only the last snapshots of the 50 ns-long MD simulation belonging to the L_1/L_2 bound forms of WT^{Δ} complexes are used for the single mutations ($WT^{\Delta} \rightarrow G330T^{\Delta}$ and $WT^{\Delta} \rightarrow H372A^{\Delta}$), and these FEP simulations are replicated for four times. Exponential averaging is used to obtain the reported ΔG values, and error values are calculated by taking the square root of squared sums. Lastly, $\Delta G = -k_B T \ln(K_d)$ equation is employed to calculate binding energies by using the dissociation constant (K_d) values obtained from the binding affinity experiments.[16, 17]

2.1.4 MuMi Scheme

WT PDB structures (Table 2.1) are minimized for 50,000 steps, and after the insertion of mutation, in-silico mutated structures are minimized for another 50,000 steps to equilibrate the protein structures. Each residue of the WT structure is mutated to alanine, then minimized with the same process.

Table 2.1 PDB ID of 40 proteins.

PDB ID	Residues	PDB ID	Residues
5rxn	54	1lis	131
1dtx	58	1kuh	132
1tfs	60	1irl	133
1tgx	60	1jac	133
1pi2	61	2tbd	134
1cse	63	1cof	135
1ptx	64	1pms	135
2bbi	71	1gen	200
1hrz	73	1iae	200
1hcp	75	1nox	200
1iml	76	2gsq	202
1cdq	77	1cfb	205
1kve	77	1dyr	205
1vcc	77	1thv	207
1cyu	98	2abk	211
1be9	115	1ctm	250
1slt	129	1mml	251
1sei	130	1vin	252
1hmt	131	1plq	258
1htp	131	1lxa	262

Table 2.2 PDZ complexes and their PDB IDs.

Abbreviation	PDB ID	Structure Details
WT_0	5HDY	PDZ3 of PSD-95
$G330T_0$	5HET	PDZ3 of PSD-95 (G330T mutant)
$H372A_0$	5HF4	PDZ3 of PSD-95 (H372A mutant)
DM_0	5HFD	PDZ3 of PSD-95 (G330T, H372A double mutant)
WT_{L1}	5HEB	PDZ3 of PSD-95 in complex with the peptide derived from CRIPT
$G330T_{L1}$	5HEY	PDZ3 of PSD-95 (G330T mutant) in complex with the peptide derived from CRIPT
$H372A_{L1}$	5HFB	PDZ3 of PSD-95 (H372A mutant) in complex with the peptide derived from CRIPT
DM_{L1}	5HFE	PDZ3 of PSD-95 (G330T, H372A double mutant) in complex with the peptide derived from CRIPT
WT_{L2}	5HED	PDZ3 of PSD-95 in complex with the mutant peptide derived from CRIPT (T-2F)
$G330T_{L2}$	5HF1	PDZ3 of PSD-95 (G330T mutant) in complex with the mutant peptide derived from CRIPT (T-2F)
$H372A_{L2}$	5HFC	PDZ3 of PSD-95 (H372A mutant) in complex with the mutant peptide derived from CRIPT (T-2F)
DM_{L2}	5HFF	PDZ3 of PSD-95 (G330T, H372A double mutant) in complex with the mutant peptide derived from CRIPT (T-2F)

2.2 Construction of Networks from Proteins and Measures on a Toy Graph

2.2.1 Construction of Residue Networks (RNs)

C_β of each residue (C_α for glycine) is taken as a node to preserve side-chain sensitivity in calculations to construct a graph of the protein structure. Nodes within a 6.7 Å distance are taken as interacting, and an edge is assigned between them. The cut-off distance of 6.7 Å is chosen for linking the first coordination shell of C_β atoms in RDF which belongs to adjoint residues and other residues that locate close to the central residue.[20, 29, 55] RNs are unweighted, undirected, and they do not have self-loops or parallel edges.

2.2.2 Graph Measures on a Toy Network

A graph consists of nodes (vertices) and links (vertices), and there is a variety of measures that focuses on nodes/edges that are informative about features of networks. Degree (number of neighbors, *k*) is local measure which indicates the number of links, thus nodes that are connected to a certain node. In G_0 (Figure 2.1), Node 7 and 10 have the highest number of neighbors, while Node 2 and 7 have the lowest degree.

Clustering coefficient (*C*) indicates number of triangles that goes through a node, which is interpreted as the fraction of neighbors becoming a neighbor of each other. Clustering coefficient of a certain node is calculated by,

$$C = \frac{2T(u)}{\mathrm{k}\,(u)(\mathrm{k}(u)-1)} \tag{2.2}$$

where T is triangles through the node u, and $\mathrm{k}(u)$ is the number of neighbors of u.

Although, C is informative about local structure, it is impartial to the number of adjacent nodes. For example, distal nodes of G_0, such as nodes 2, 5, 8 and 9 have the highest C, even though they have lower connection overall, considering k values (Figure 2.1b).

Average shortest path length ($<L>$) of nodes displays the spatial accessibility in units of edges in a network. The $<L>$ of a node is calculated by,

$$< L_i >= \frac{1}{n-1}\sum_i^n L_{ij} \tag{2.3}$$

where i is an arbitrary node which is targeted to calculate L, j is any other node, and n is the number of nodes in a graph. Nodes 2 and 5 have lowest number of neighbors and high $<L>$ values, which indicates their low connectivity in G_0 graph.

Betweenness centrality (BC) is a general term to calculate a node's impact on information flow, by assessing communication of all unique node pairs that traverse through a certain node; here, we focus on shortest path BC,[56] current flow/random walk BC[57] and communicability BC.[58, 59] To distinguish current flow and communicability BC measures from shortest path BC, we refer them by using their path/walk calculation methods; on the other hand, the term 'BC' always corresponds to shortest path BC, in this thesis.

Assuming that information exchange happens through the shortest paths on a network; the BC of an arbitrary element (node or edge) is calculated by,[56, 60]

$$BC(X) = \sum \left(\frac{P(a,b|X)}{P(a,b)} \right) \quad (2.4)$$

where $P(a, b|X)$ is the number of the shortest paths travelled through this element, and $P(a, b)$ is the number of all shortest path between the nodes a and b. BC is normalized by $(n-1)(n-2)/2$ for nodes and $n(n-1)/2$ for edges, where n is the number of nodes in the network. The range of BC is [0, 1]; if all the shortest paths traverse through the certain element (X), it is 1. Hence, BC gives both local and global information about a graph. In G_0, Node 1, 3, 4, 6, 7, 10 demonstrate high BC values and are located at connection between highly connected node groups, independent from the number of neighbors. Further, information flow may be assessed as number of walks[58, 59] and current flow[57]. The centrality measures based on these computations are displayed in Figure 2.1b, and the results indicate that three types of centrality measures show similar pattern, thus shortest path BC is chosen for further analyses.

Nonetheless, edge BC (Equation 2.4) is used to detect close-knit node groups which are defined as 'communities' in Girvan-Newman algorithm.[37] Girvan-Newman algorithm calculates edge BC and removes the edge with the highest BC in a loop, hence leading two and more connected components to emerge in a graph. These connected components correspond to 'communities' and denoted by Ω. In Figure 2.1a, c, the detection of communities is illustrated for G_0; where Ω is 1 for first edge removal. After the removal of 3 highest BC edges, three communities separate (Ω = 3) in G_3. The community separation of G_0 also exemplifies that the edges with high BC have propensity to connect nodes with high node BC. Nodes 1, 3, 4, 6, 7 and 10 are connected with the highest BC edges, as a result, they are removed as per Girvan-Newman algorithm. The computation of the graphs are done by using NetworkX[61] package of python programming language, in this thesis.

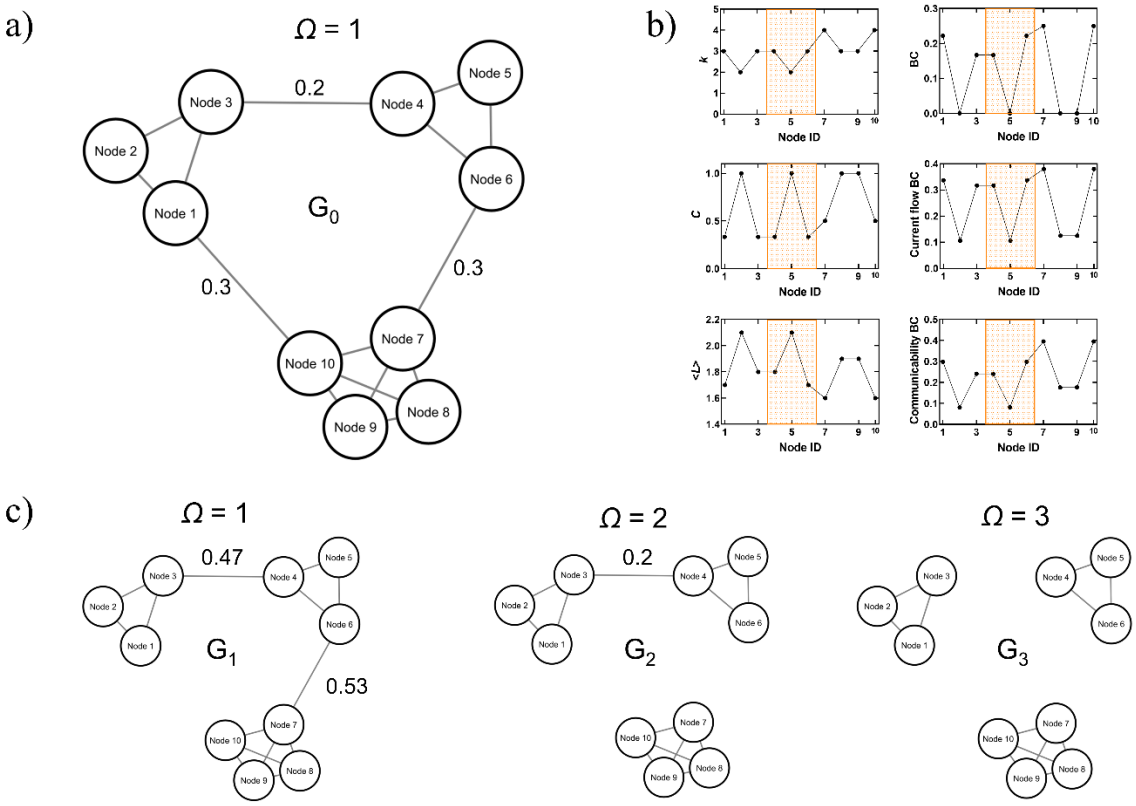

Figure 2.1 Toy network that consists of 10 nodes and 15 edges. **a**, **b** G_0 graph and *k*, *C*, <*L*> and centrality measures are illustrated. Edge BC values are written on top of the concerning edges. **c** Community detection of G_0 by using Girvan-Newman and changing edge BC values are displayed for each edge cut.

3. RESULTS

3.1 MuMi Results

MuMi[20] scheme is employed to investigate structural impact of mutational perturbations by assessing the changes in graph measures after insertion of mutation (Figure 1.2a). 40 PDB

structures (Table 2.1) are fetched, and each residue is mutated to alanine. To investigate the overall change upon mutations we calculate,

$$\Delta A = A_{WT} - A_{Mut} \tag{3.1}$$

$$Normalized\ variance\ of\ A = \frac{\langle \Delta A^2 \rangle^{1/2}}{A_{WT}} \tag{3.2}$$

where *A* is a graph parameter, and the change is averaged over all possible single alanine mutations of protein structures. The effect of alanine mutations is quantified as a function of conservation whereby the conservation scores are computed by the ConSurf server.[62]

k, *C*, BC and <*L*> are calculated for protein structures. *k* infers the local connectivity of the protein, and mean *k* for the ensemble is ~6. Residues with high *k* tend to have high conservation scores indicating that residues with high connectivities oppose changes during evolutionary processes. Interestingly, the number of neighbors of an amino acid do not alter significantly upon mutations (Figure 3.1a). Further, mean values of *C* and <*L*> for the ensemble are 0.4 and 5, respectively. With the elevating conservation values, *C* and <*L*> have propensity to get lower. The change of *C* is lower in evolutionarily conserved residues, while results are insignificant for <*L*>, after the alanine scanning (Figure 3.1b, c). BC, with an ensemble mean of 0.03, demonstrates significant results; hence, conserved residues have a higher BC, and BC is sensitive and indicative for mutational perturbations (Figure 3.1d). Nevertheless, BC is employed in various studies to understand protein function.[20, 34, 35, 40]

Additionally, average displacement of C_β (C_α for glycine) after the alanine scanning is computed (Figure 3.1e). Based on this computation, we show that conserved residues tend to deviate less upon mutational perturbations.

To understand the BC results further, WT BC and the change of BC of each protein structure is compared with conservation scores and DEPTH[63] values. DEPTH calculates distance of a buried residue to a closest solvent accessible surface. In Figure 3.1f, Pearson *R* values between BC/change of BC and DEPTH (~0.6-0.5) are higher than correlation of BC and conservation scores (~0.3). Thus, central residues are partial towards being in core and buried; however, BC is not an indicative of evolutionary conservation. Additionally, after the alanine scan, BC does not display distinguishable results compared to the WT calculations.

Overall, this general approach is not efficient to investigate structure and function of proteins. The reasons are several; first, minimization is not a suitable process for emergence of variant behaviors. Second, averaging of mutational responses lead cancellation of slight changes, which may be important for protein function. Lastly, without a specific biological hypothesis on each protein, batch analysis of a pool of structures is not comprehensible and illuminating. Therefore, to tackle these problems, we focus on mutation and/or binding of PDZ3 proteins and assess their functionality by calculating relative binding energies. Then, we revisit the graph theoretical computations through the lens of molecular basis of functionality and temporality.

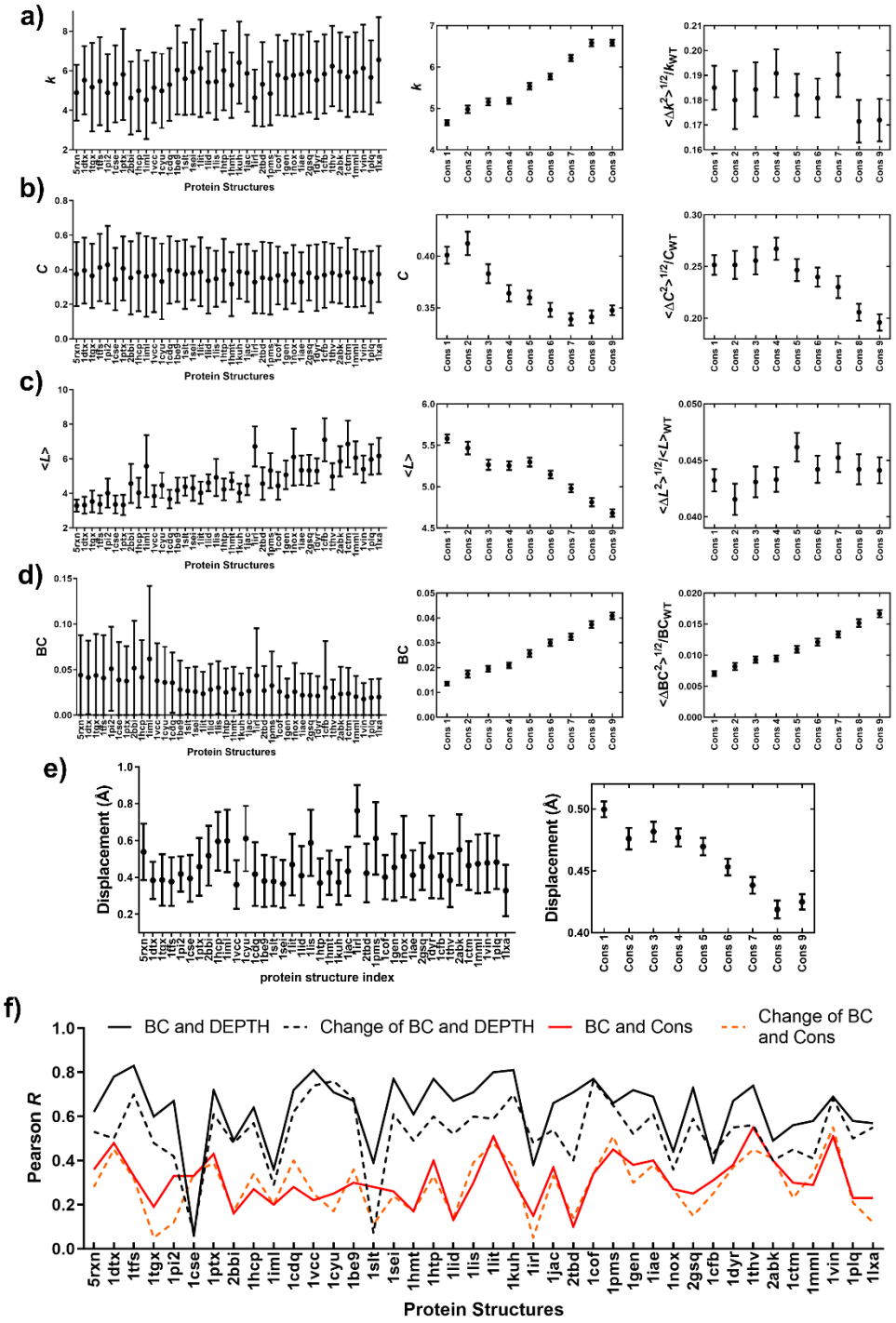

Figure 3.1 **a-e** *k*, *C*, <*L*>, BC and Displacement values are computed and visualized as per protein structure, classified by conservation value and their changes based on conservation value. **f** Correlation between BC/change of BC and DEPTH/conservation scores for each protein structure are displayed. PDB IDs on x-axis are sorted by increasing residue number.

3.2 Investigation of PDZ3 Function: Relative Binding Energies and All-Atom MD Simulations

After the assessment of the protein ensemble by using MuMi, now we focus on PDZ3, where we study its biological functionality by employing mutation and binding energies in thermodynamic cycles. Apo and L_1/L_2 bound forms of WT, G330T, H372A and DM complexes are taken for MD and FEP simulations as it was explained in Methods section. After the investigation of relative binding energies, MD simulations are analyzed to understand the molecular basis of these occurring energies. Then, a simple model is constructed to approximate the experimental binding energies (Figure 1.2b).[23]

3.2.1 Construction of Thermodynamic Cycles and MM/GBSA calculation

FEP calculations are compared to the experimental findings as schematically displayed in Figure 3.2a. K_d values obtained from binding affinity experiments[16, 17] are employed to calculate standard binding free energies through $\Delta G = -k_B T \ln(K_d)$.

By utilizing the ΔG results from the experimental data from Cycle *A* in Figure 3.2a,

$$\Delta G_{M-L1}^{Bind} - \Delta G_{W-L1}^{Bind} = \Delta\Delta G_A \tag{3.3}$$

with the FEP results,

$$\Delta G_{L1}^{Mut} - \Delta G_{0}^{Mut} = \Delta\Delta G_A \tag{3.4}$$

Equations 3.3 and 3.4 are arranged as,

$$\Delta G_{M-L1}^{Bind} - \Delta G_{W-L1}^{Bind} = \Delta G_{L1}^{Mut} - \Delta G_{0}^{Mut} \quad (3.5)$$

All variables are illustrated in Figure 3.2a, and the same calculations are conducted for Cycle *B*. $\Delta\Delta\Delta G$ is utilized for the validation of the FEP results.

$$\Delta\Delta G_B - \Delta\Delta G_A = \Delta\Delta\Delta G \quad (3.6)$$

Thermodynamic cycles are constructed to compute various relative free energy changes of mutations or ligand binding (Figure 3.2a). For each cycle connected by solid arrows, the difference between two vertical (binding) free energy changes (Figure 3.2a, red) signifies the more stable bound form; hence, for a negative $\Delta\Delta G_A$ (Equation 3.3) the mutated complex with L_1 is more favorable than the WT complex. Since each of the cycles completed by the solid arrows should sum to 0, the $\Delta\Delta G$ calculated from experimental binding constants may also be obtained by computational means which is the difference between the two horizontal free energy changes (Equations 3.4, 3.5). Nonetheless, $\Delta\Delta\Delta G$ is calculated to investigate the favorable ligand, after the mutation of the protein (Equation 3.6). A negative $\Delta\Delta\Delta G$ indicates that after the mutation, the protein tends to form a favorable complex with L_1 (Figure 3.2a).

Further, MM-GBSA takes into account the sum of (i) the protein self-energy, (ii) a non-electrostatic solvation energy proportional to the SASA of the whole protein, and (iii) the electrostatic component of solvation expressed by the generalized Born model. In particular, the first term is exact, in that it computes all the bonded and non-bonded interactions between the atoms of the protein. MM/GBSA calculations were performed by the MolAICal tool.[64]

MM/GBSA calculated energies are much lower than the experimental energies; as such, they are scores rather than estimates of binding energies.[65]

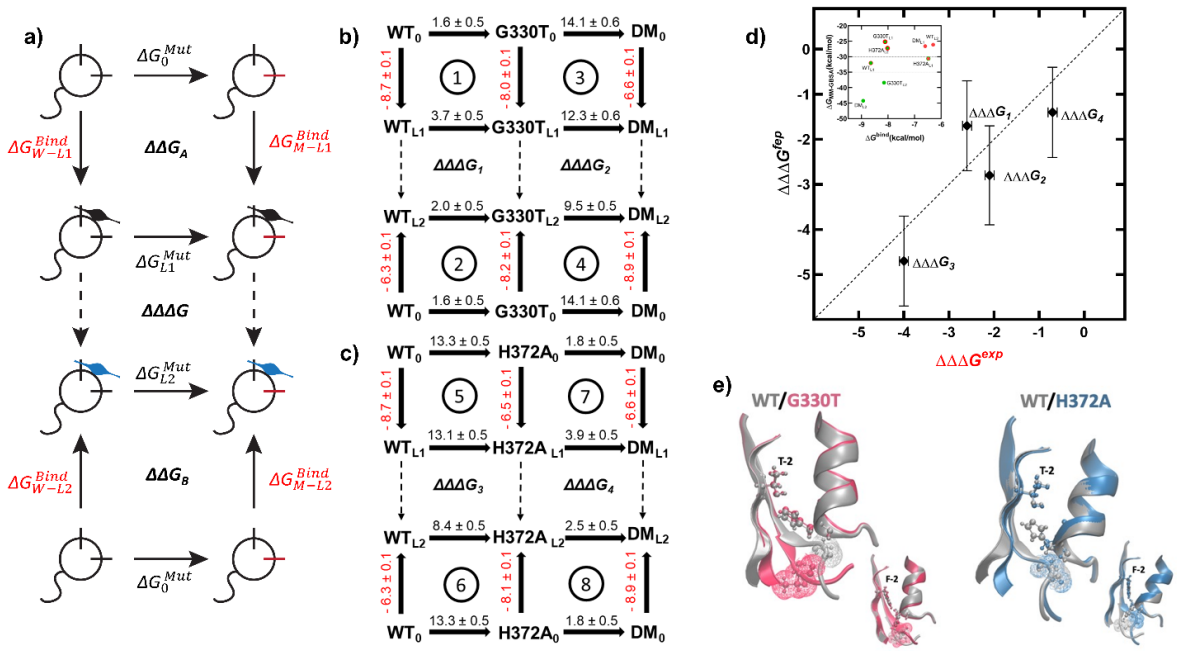

Figure 3.2 **a** Cartoon illustration for mutation and ligand binding thermodynamic cycles of the PDZ domain. Violin shapes in black and blue represent L_1 and L_2, respectively. ΔG^{Mut} values (in black) show the FEP simulation derived free energy differences between the WT and mutated PDZ3. ΔG^{Bind} values (in red) are the standard binding free energy differences calculated from experimentally reported K_d. Dashed arrows indicate the ligand switching process (not directly calculated). **b, c** Mutational cost (black) and ligand binding free energies (in red) for G330T, H372A and DM; all values in kcal/mol. $\Delta\Delta\Delta G$ may be calculated by using either ΔG^{Mut} or ΔG^{Bind} values. **d** Plot of FEP calculated vs. experimental $\Delta\Delta\Delta G$ (Pearson R = 0.84, p-value < 0.01) attest to the precision of FEP calculations; MM-GBSA vs. ΔG^{Bind} is displayed in the inset whereby the approximate approach falls short of distinguishing the binding propensities of the complexes (color of symbol is green for experimentally functional ligand, red otherwise; outline color of symbol indicates how it would be classified by MM-GBSA; region between dashed lines is for uncertain classification.) **e** Close-up views of side chain positioning and residue 330 solvent accessibility affected by the mutations (WT gray, G330T pink, H372A blue). L_1 bound forms are displayed large, L_2 bound forms are small. G330T mutation (left) does not affect interactions between H372 and position -2 in either ligand but leads to a conformation shift in the loop carrying position 330, increasing its solvent accessibility. H372A mutation (right) in L_1 bound form leads to loss of polar-polar interactions in the binding pocket; in L_2 bound form, the kink in the loop containing G330 due to rotation of H372 side chain to accommodate the large F-2 is relieved upon mutation. Solvent accessibility of G330 remains the same in both cases.

3.2.2 Mutation Pathways and Relative Binding Energies of PDZ3

The binding constants from a previous study[17] are employed to calculate the standard binding free energies of WT, G330T, H372A and DM PDZ3 to L_1 and L_2 (shown in red in Figure 3.2 b-c). PDZ domains were shown to operate at the 1-15 μM dissociation constant range,[19] by following previous work, we have used -7.0 kcal/mol as a standard binding free energy threshold (~10 μM at physiological temperatures) to classify the PDZ complexes. Therefore, WT_{L1}, $G330T_{L1}$, $G330T_{L2}$, $H372A_{L2}$ and DM_{L2} are the functional PDZ complexes based on their binding free energies, which are -8.7, -8.0, -8.2, -8.1 and -8.9 kcal/mol, respectively. The binding free energies of non-functional complexes are -6.3, -6.5 and -6.6 kcal/mol for WT_{L2}, $H372A_{L1}$ and DM_{L1}, respectively. The FEP calculated free energy costs of the mutations are also displayed in the same figures in black (Figure 3.2).

Additionally, the tautomeric states of key residues, such as H372, may change upon ligand binding and that they may have significant populations in multiple states in bound or apo forms.[66, 67] Since each cycle should sum to zero, we find from the deviations in cycles ①, ②, ⑦ and ⑧ that the apo form G330T mutation has an additional cost of ~1.8 kcal/mol, and from those in cycles ③, ④, ⑤ and ⑥ that the H372A mutation in the apo form has an additional ~3.2 kcal/mol cost due to such population shifts.

Although each leg of the cycles in Figure 3.2b-c may be obtained computationally, such calculations are subject to various errors inherent in the employed methodology. In Figure 3.2a, $\Delta\Delta\Delta G$, the difference between the dashed arrows, is equal to $\Delta\Delta G_B - \Delta\Delta G_A$ (horizontal, mutation energies) as well as that between the vertical binding free energies (Equation 3.6). Thus, we employ $\Delta\Delta\Delta G$ to validate our simulations against the experimental binding energies (Figure 3.2d), and we find that within the 1 kcal/mol accuracy limit provided by FEP calculations,[68] the energy differences obtained by using conformations sampled throughout the MD trajectories are consistent with the experimental binding energies. Note

in particular that this approach eliminates the use of mutational cost calculations in the apo forms, in particular the above-mentioned energy cost due to shifts in dynamic tautomeric populations. We also note that the MM-GBSA method widely used for efficient scoring of ligand binding does not have the precision necessary to distinguish the binding selectivity of PDZ3 to ligands (inset to Figure 3.2d).

Since we do not calculate binding free energies, we are not positioned to directly comment on which mutants will be functional. However, we are now able to discuss the mechanisms by which these mutations operate. First off, the general effect of ligand type on free energetic cost of the mutations is clear from the calculations where changes in the L_2 bound form is always less costly than those in the L_1 bound form, hence the negative $\Delta\Delta\Delta G$ values displayed in Figure 3.2d.

As discussed in a previous study on the mutational pathways of PDZ,[17] G330T is the most abundant variant owing to its class bridging behavior (Figure 3.2e, left). This phenomenon is explained by the low free energy cost of this mutation, irrespective of the presence of the ligand in the binding pocket and of the prior H372A mutation (in the range of up to 4 kcal/mol; Figure 3.2, cycles ①, ②, ⑦ and ⑧; numbers in black). Considering that the G330T mutation requires relatively low number of atom additions and that residue 330 is located on a flexible loop (Figure 3.2e, left) where the newly created polar side chain may easily be repositioned to get in contact with water, the low cost is plausible. Nevertheless, the L_2 bound form is able to accommodate this mutation with ~2 kcal/mol less cost than either the ligand free or the L_1 bound forms, leading to the ligand bridging behavior (compare cycles ① and ②). Similarly, the H372A mutant is also able to contain the additional G330T mutant with ~1.5 kcal/mol less cost in the presence of L_2 (compare cycles ⑦ and ⑧), retaining the ligand switching behavior brought on by this first mutation.

In contrast to G330T, the H372A mutation is rather costly (on the order of 10 kcal/mol) under all circumstances due to the large change of the side chain volume, as well as the shift this

position makes from polar to hydrophobic (Figure 3.2e, right). Moreover, this residue is in the binding pocket in direct contact with the ligand in the liganded cases, which makes the free energetic cost highly dependent on the rearrangements of local contacts. Especially in L_1-bound cases, the bond between H372 and the threonine residue at the -2 position of L_1 is pertinent to the formation of the WT_{L1} complex.[12] Note that H372 was also shown via deep mutagenesis to be the most sensitive residue to mutations.[16] Nevertheless, the cost of the H372A mutation is significantly lower when L_2 is bound to the protein, by 4.7 kcal/mol (compare cycles ⑤ and ⑥), helping bring the binding free energy of this mutant in the functional range, and thus leading to ligand switching. When H372A mutation follows G330T, the cost is again lower in the presence of L_2, but more importantly, it is enough to compensate for the effect of the mutation in the apo form and to retain a physiologically significant degree of specificity.

3.2.3 N-Terminus Fluctuation Patterns are Distinct for Each PDZ3 Complex

The ligand-bound trajectories we have generated reliably represent the equilibrium properties of the bound complexes, as validated in the previous subsection since the FEP calculations are based on the conformations generated in these simulations. We now focus on the dynamics of the complexes to delineate the allosteric behavior observed in these systems. We thus compute the RMSD, RMSF and the cross-correlations of the trajectories.

RMSD for each MD trajectory is shown in Figure 3.4. We find an unusual property for these trajectories whereby there are stretches of times having plateaus with small fluctuations, separated by regions of relatively large fluctuations. RMSF results show that the most mobile site of PDZ3 is the N-terminus region (Figure 3.3). Hence, we determine that the overall mobility of the protein is dominated by the fluctuation of the N-terminal region.

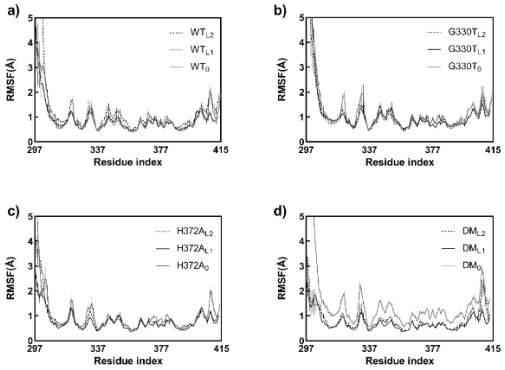

Figure 3.3 The RMSF results of WT, G330T, H372A and DM for apo and L_1/L_2 bound cases. 40 ns-long chunks from duplicate trajectories are averaged; error bars indicate the standard error of these eight chunks. **a, b** WT and G330T show similar patterns and a highly mobile N-terminus region in all cases. **c, d** H372A and DM display similar regimes with a mobile whisker and peak around residue 408. Interestingly, DM_0 has a significantly high mobility compared to the ligand-bound complexes.

All apo complexes have highly mobile and disorganized N-termini, while those of the ligand-bound complexes have clusters around preferred conformational states (Figure 3.4). The ligand-bound forms of WT and G330T have similar cluster shapes for the N-terminus residues (Figure 3.4a-b). $H372A_{L1}$ and DM_{L1} exhibit nearly identical conformational preferences of the N-terminus (Figure 3.4c-d). The N-terminus of $H372A_{L2}$ adopts an extended conformation, while that of DM_{L2} displays a collapsed form (Figure 3.4c-d).

To scrutinize the dynamical behavior of the PDZ structures further, the cross-correlation matrices of the protein complexes are calculated (Figure 3.5). The correlations of residue displacements are very similar, with the inner product between WT_0 and each of the other systems varying in the range 0.80-0.97 when the 12 residue-long N-terminus spanning residues 299-310 is omitted in the calculations. However, there are no other dominant motions that directly manifest themselves in the cross correlations between the regions of the protein body that are unique to the variant studied.

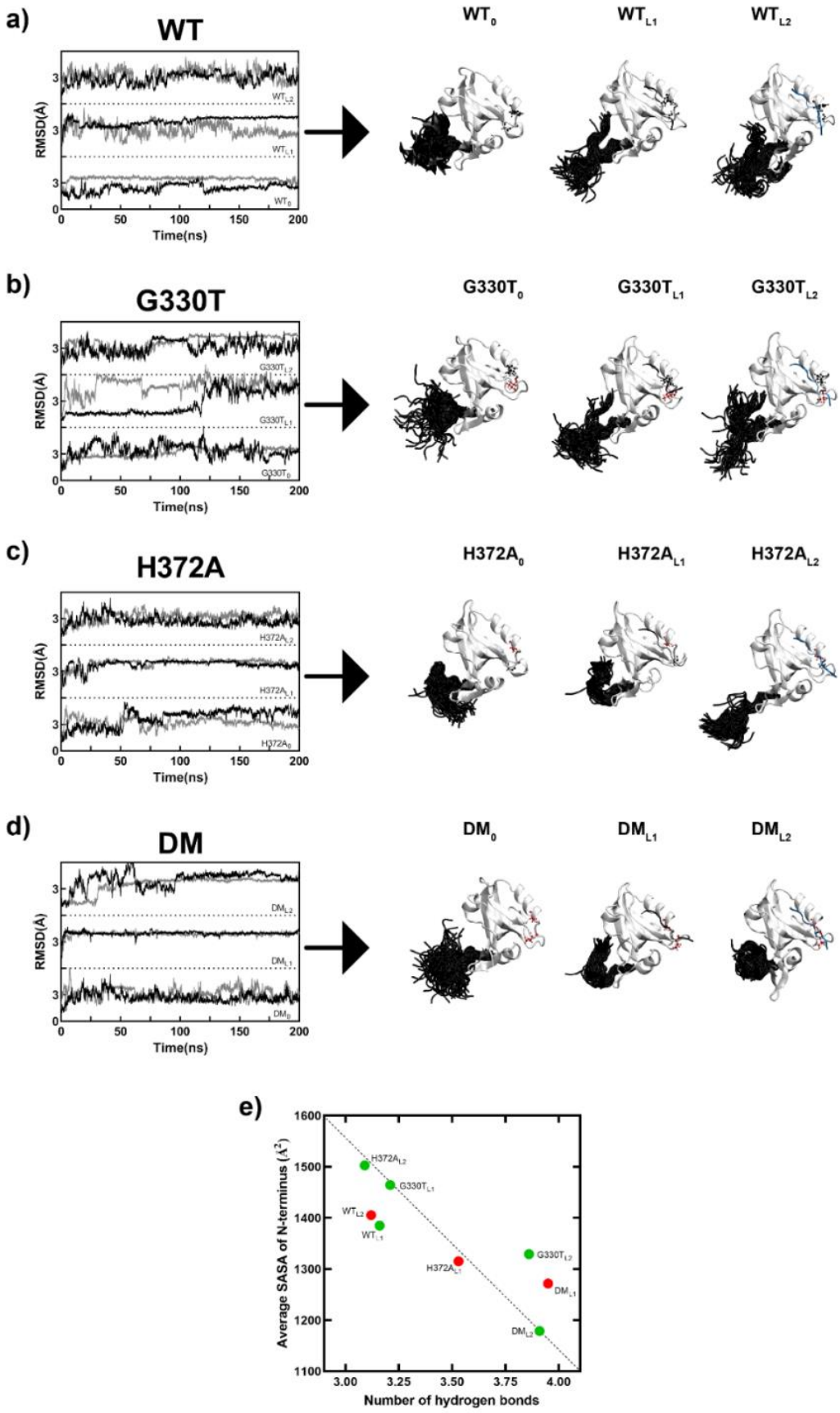
a)
WT
WT0
WTL1
WTL2
RMSD(Å)
Time(ns)
b)
G330T
G330T0
G330TL1
G330TL2
c)
H372A
H372A0
H372AL1
H372AL2
d)
DM
DM0
DML1
DML2
e)
Average SASA of N-terminus (Å2)
Number of hydrogen bonds

Figure 3.4 **(a-d)** RMSD and N-terminus conformers observed in the MD simulations for apo and L_1/L_2 bound cases. In RMSD plots, duplicate simulations are displayed separately by black and gray curves. In protein structure visualizations, evenly separated 500 snapshots of only the N terminus region (shown in black) from duplicate trajectories for each complex are illustrated; the rest of the protein (in white) is displayed by its average structure with side chains of residues 330 and 372 displayed in ball and stick (red if mutated, black if native). L_1 and L_2 are shown in black and blue, respectively. **(e)** Regression between $N_{\text{H-bonds}}$ (Table 3.1) and SASA of N-terminus residues (Figure 3.6); Pearson R = -0.86, p-value < 0.01). The functional complexes are shown in green, while unfavorable ones are displayed in red.

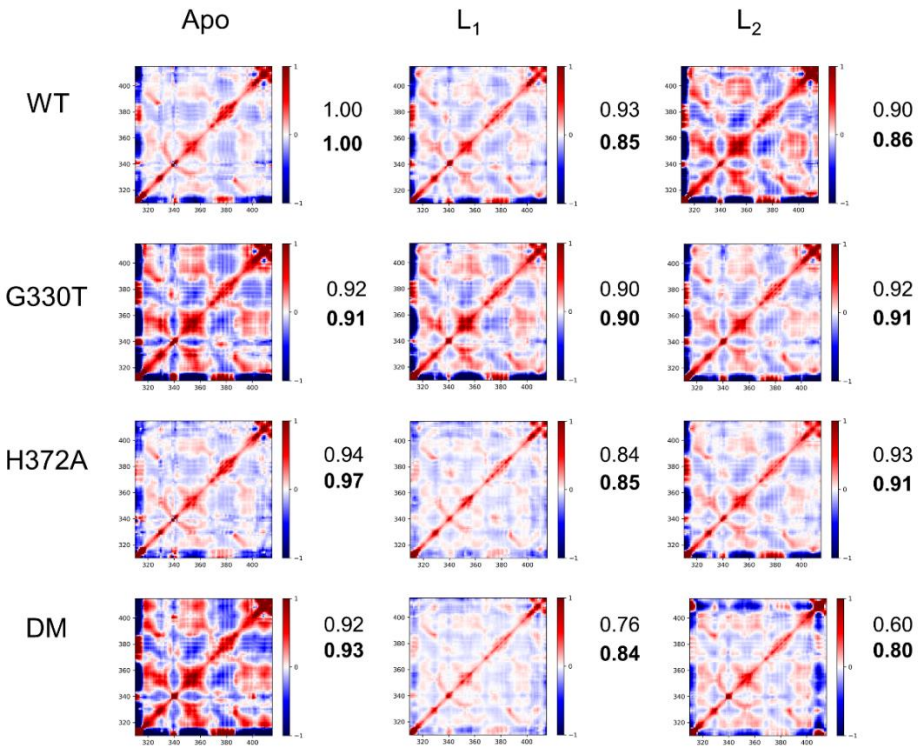

Figure 3.5 Cross-correlation maps of WT, G330T, H372A and DM for apo and L_1/L_2 bound cases. 40 ns-long chunks from duplicate trajectories are averaged for each complex. Graphs are built for the whole protein. Numbers display the similarity to WT_0, 1 being for identical correlation maps; bold values are computed omitting the N-terminus in the correlation map calculations. DM_{L1} and DM_{L2} display the largest departure in fluctuation patterns compared to the apo WT complex.

3.2.4 N-terminus is an allosteric partner essential in determining binding ligand.

The N-terminus has the NH_2-P**E**FLG**EED**IP**RE** sequence; half of its 12 residues are charged, and the rest are hydrophobic (Figure 1.3b). It also renders the -4 net charge of the protein. This sequence possibly contributes to the long-range control over binding affinities. To quantify the various degrees of flexibility observed in the conformations discussed in the previous section (Figure 3.3-3.5), we plot the SASA distributions of the N-terminus residues (Figure 3.6). The disordered flexibility of the region in the apo forms is characterized by their broad distributions (with variance ~60 $Å^2$), while SASA also delineates the two-state nature of the conformations of the WT_{L1}, $G330T_{L2}$ and even distinguishes minor conformations such as that observed for $G330T_{L1}$; DM_{L2} is particularly characterized by a peakish single conformation. SASA distributions of only the charged N-terminus residues, on the other hand, display starkly different features (Figure 3.7).

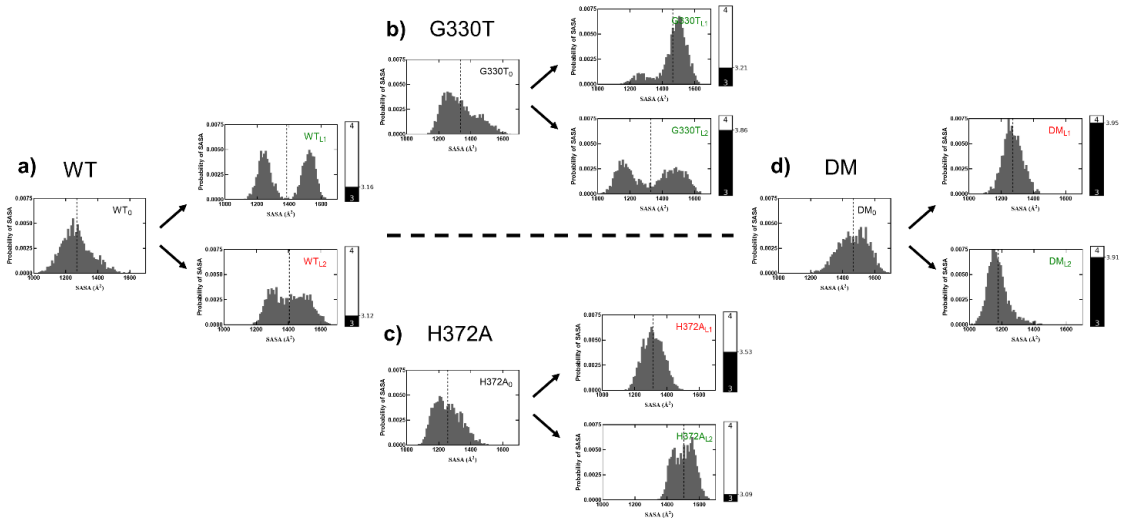

Figure 3.6 Probability distribution of SASA for the full-length N-terminus; vertical dashed lines indicate average SASA; side bars display the average hydrogen bond count between the PDZ domain and the ligand in the MD simulations in ligand-bound cases. Complexes with favorable binding to the ligands are labelled in green, and complexes with unfavorable binding are in red.

Table 3.1 Average number of hydrogen bonds ($N_{\text{H-bonds}}$) between the protein and the ligand for the full and N-terminus truncated protein structures

	Full length protein	Truncated (Δ) protein
WT_{L1}	3.2 ± 1.4	3.4 ± 1.6
WT_{L2}	3.1 ± 1.6	3.4 ± 1.7
$G330T_{L1}$	3.2 ± 1.6	3.7 ± 1.7
$G330T_{L2}$	3.9 ± 1.7	4.3 ± 1.7
$H372A_{L1}$	3.5 ± 1.7	3.0 ± 1.4
$H372A_{L2}$	3.1 ± 1.5	3.5 ± 1.7
DM_{L1}	4.0 ± 1.7	4.4 ± 1.9
DM_{L2}	3.9 ± 1.7	4.3 ± 2.0

We emphasize that it is not possible to determine the binding fate of the ligands by focusing on this property of the N-terminus alone. In fact, investigating the main interactions at the binding site is in order. We find that the average number of hydrogen bonds established between PDZ3 and the ligands (varying between 3 and 4) also differs between the various complexes (displayed in Figure 3.7 sidebars and listed in Table 3.1). Although, the difference is small, hydrogen bonds established between protein and ligand have a significant impact on ligand binding.[69] Moreover, there is a meaningful negative correlation between the average SASA of the N-terminus and the number of hydrogen bonds (Pearson R = -0.86; Figure 3.4e) implying allosteric communication between binding site and the charged tail. The inverse relationship between the overall SASA of the N-terminus and the number of hydrogen bonds at the binding site has mechanical origins. The fraction of time the N-terminus having a net charge of +4 is solvated or shielded depends on the differential modulation of electrostatic interactions in the variant studied. However, an increased shielding from the solvent means it interacts with the main body of PDZ3, particularly the nearby C-terminus, as depicted in the cartoons of Figure 3.4. When stabilized by the N-terminus, these regions mechanically support the binding site from below, sandwiching it with the well-known α_2 helix,[25] thus increasing the lifetime of hydrogen bonds at the binding site.

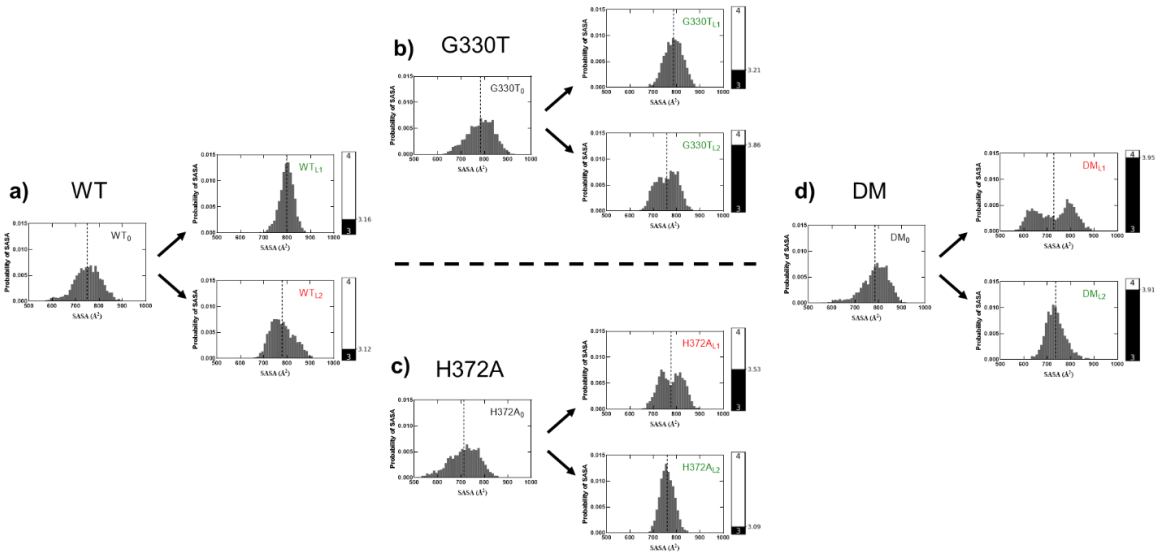

Figure 3.7 Probability distribution of SASA for the N-terminus charged residues; side bars display the average hydrogen bond count between the PDZ domain and the ligand in the MD simulations in ligand-bound cases. The complexes with favorable binding to the ligands are labelled in green, while the complexes with unfavorable binding are in red. **a** Regarding the WT complexes, solvent accessibility of the charged residues has a dominant effect on ligand-binding. WT_0 and WT_{L2} have similar broad distributions while WT_{L1} is peakish leading to favorable binding. Average hydrogen bond counts are indifferent in both cases. **b** SASA profiles of G330T complex do not differ drastically after binding. The L_1 bound complex has a slightly sharper peak and a lower hydrogen bond count. However, the L_2 bound form has a wider SASA distribution and a higher hydrogen bond count. Additional hydrogen bonds and sharper SASA distributions compensate one another to induce the favorable binding in the alternate cases. **c, d** In H372A and DM cases, the favorable binding to L_2 is mainly due to the sharp SASA distribution of the charged residues on the N-terminus.

To explain the observed binding preferences of PDZ, we have identified two main contributions: (i) The direct effect at the binding site quantified by the average number of hydrogen bonds between the protein and the ligand; and (ii) the allosteric effect due to conformational multiplicity of the N-terminus and its resulting dynamic interactions. The latter is related to the electrostatic free energy change of the system through the generalized Born (GB) model,[70, 71]

$$\Delta G_{elec}^{solv} = \frac{1}{2}\sum_{i=1}^{M}\sum_{j=1}^{M}\frac{q_i q_j}{4\pi\varepsilon_0 f_{GB}(r_{ij})}\left(\frac{1}{\epsilon_w}-\frac{1}{\epsilon_m}\right) = \frac{q_i q_j}{8\pi\varepsilon_0}\left(\frac{1}{\epsilon_w}-\frac{1}{\epsilon_m}\right)\sum_{i=1}^{M}\sum_{j=1}^{M}\frac{1}{f_{GB}(r_{ij})} \quad (3.7)$$

While not defined uniquely, several simple but effective forms were used for the function $f_{GB}(r_{ij})$ including inverse relation to charge-charge distances r_{ij} and screening effects. Here, we assume $\sum_{i=1}^{M}\sum_{j=1}^{M} f_{GB}(r_{ij})^{-1} \propto \sigma^{-\frac{1}{2}}$, where σ is the variance of the SASA of the charged residues. Moreover, the two contributions also determine the overall conformations adopted by the protein.

Thus, our simplified model to predict binding free energies from MD trajectories is given by the sum of three effects:

$$\Delta G_L - \Delta G_0 = \Delta\Delta G_A = \Delta\Delta G^{\text{solvation}} + \Delta\Delta G^{\text{H-bonds}} + \Delta\Delta G^{\text{electrostatics}}$$
$$= \alpha\,(S_L - S_0) + \beta N_{\text{H-bonds}} + \frac{q_i q_j}{8\pi\varepsilon_0}\left(\frac{1}{\epsilon_w} - \frac{1}{\epsilon_m}\right)\gamma\left(\frac{1}{\sqrt{\sigma_L}} - \frac{1}{\sqrt{\sigma_0}}\right) \qquad (3.8)$$

The first is the usual term accounting for the change in the solvation free energy of the protein,[72-76] proportional to the difference in the average SASA of the whole protein in the ligand-bound and ligand-free forms, S_L and S_0, respectively (

Table 3.2). The second term accounts for the interaction energy between the ligand and the binding site, dominated by the number of hydrogen bonds formed at the binding cavity for each ligand-bound trajectory, $N_{H-bonds}$. The last term approximates the electrostatic free energy change, with ϵ_w and ϵ_m being the dielectric constants of water and the buried medium, respectively, and σ_L and σ_0 representing the variance in the SASA of the charged residues of the N-terminus for the liganded and ligand free forms, respectively (

Table 3.2). Equation 3.8 regresses the experimental free energy of binding data with α = 2 cal/mol/Å^2, β = -1.6 kcal/mol, and γ = 0.75 for ϵ_w = 80 and ϵ_m = 2 (Pearson R = 0.94; Figure 3.8). These pre-factors are physiologically relevant: Previously reported data for α is in the

range of a few to tens of cal/mol/Å^2,[73, 77-79] and for β, the rupture of each hydrogen bond may cost the total energy of a protein ~1.6 kcal/mol in aqueous environment.[80-82]

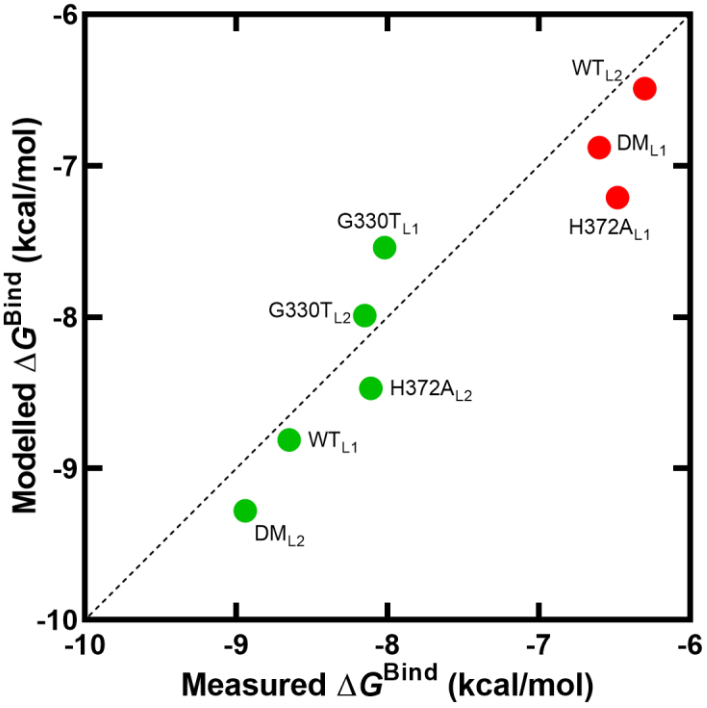

Figure 3.8 Regression between modelled (Equation 2.8) and measured binding free energies (Table 3.1); Pearson R = 0.94, p-value < 0.001). The functional complexes are shown in green, while unfavorable ones are displayed in red.

We emphasize that while being similar to MM-GBSA that is a computationally efficient approach to score ligand binding, our approximate form in equation 2 attempts to model the features that best distinguish binding fates in PDZ3. We thus find that the minimal model that optimizes the experimentally measured energies is comprised of (i) the hydrogen bonds at the binding site which is only a very small proportion of all the terms contained in the protein self-energy of MM-GBSA, (ii) the non-electrostatic solvation energy, modeled the same way as in MM-GBSA, and (iii) the electrostatic solvation energy due to the charged N-terminal residue fluctuations which is different from the GB model which takes into account the sum of the inverse distances between all charges amino acids in the protein-ligand system.

The contribution of each term to the binding free energy is listed in Table 3.3. Note that the largest contribution is due to the hydrogen bonds established in the binding pocket in each case. Nevertheless, the overall conformational change of the protein as well as the specific conformations adopted by the N-terminus charged residues determine the binding fate.

Table 3.2 Average SASA of the whole protein (S) and SASA variances of N-terminus region for charged residues (σ) used in Equation 2.8.

	S	σ
WT_0	7128	60
WT_{L1}	6746	32
WT_{L2}	6736	51
$G330T_0$	7171	58
$G330T_{L1}$	6875	40
$G330T_{L2}$	6818	47
$H372A_0$	7105	66
$H372A_{L1}$	6734	50
$H372A_{L2}$	6937	31
DM_0	7300	62
DM_{L1}	6533	81
DM_{L2}	6506	44

Thus, the L_1 preference of the WT protein is reinforced by the fixed conformations adopted by the N-terminus charged residues, despite similar hydrogen bonding interactions it has with both ligands (Table 3.3 and Figure 3.7). The G330T single mutation is a special case which prefers to bind both L_1 and L_2. However, the preference to L_1 is reinforced by the electrostatic contributions of the complex, while that to L_2 is due to the increased number of hydrogen bonds established at the binding cavity. Accordingly, the ligand-bridging behavior of the G330T mutation is conducted by the compensation between the conformational dynamics of the N-terminus region and the protein-ligand hydrogen bonding. The H372A single mutation prefers to bind L_2 due to significant contributions from the N-terminus electrostatics, a factor that is nearly absent for the L_1 bound complex. In the presence of both mutations, we find that there is significant increase in binding pocket interactions as well as the overall solvation

free energy of both DM_{L1} and DM_{L2} complexes. However, the N-terminus charged residue fluctuations in $DM_{L1,}$ which uniquely exceed those in the unbound form (with a positive $\Delta\Delta G^{\text{electrostatics}}$) offsets this advantage for L_1.

Table 3.3 Modeled vs. measured binding free energies and individual contributions from equation 2.8 (kcal/mol).*

Complex	$\Delta\Delta G^{\text{solvation}}$	$\Delta\Delta G^{\text{H-bonds}}$	$\Delta\Delta G^{\text{electrostatics}}$	Modelled ΔG^{Bind}	Measured ΔG^{Bind}
WT_{L1}	-0.76	-5.05	**-3.00**	-8.81	-8.65
WT_{L2}	-0.78	-4.99	-0.71	-6.49	-6.30
$G330T_{L1}$	-0.59	-5.13	**-1.82**	-7.54	-8.02
$G330T_{L2}$	-0.71	**-6.18**	-1.10	-7.99	-8.15
$H372A_{L1}$	-0.74	-5.65	-0.82	-7.21	-6.48
$H372A_{L2}$	-0.34	-4.94	**-3.19**	-8.47	-8.11
DM_{L1}	**-1.53**	**-6.31**	**0.97**	-6.88	-6.60
DM_{L2}	**-1.59**	**-6.25**	-1.44	-9.28	-8.94

* functional complexes are displayed in green; interactions dominating the binding fate are shown in bold.

3.2.5 Removal of the charged N-terminus exposes its key role in ligand binding.

We design a knock-out computer experiment by removing residues 299-310 constituting the N-terminus to test its role on binding affinities; in what follows, these systems are referred by the superscript Δ. We run 50 ns-long MD simulations for the N-terminus deleted forms of all eight ligand bound systems. Additionally, we perform FEP calculations for the single mutations ($G330T^{\Delta}$ and $H372A^{\Delta}$).

In Figure 3.9 a, we display the free energetic cost of removal the N-terminus from the two single mutations. In the full-length PDZ, the N-terminus appears to have the largest favorable

impact on WT_{L1}, $G330T_{L1}$ and $H372A_{L2}$ (Table 3.3). Thus, we expect cycle ② to be the least affected by its absence as is the case, with the G→T transition in the presence of L_2 costing 2.0 ± 0.5 to 1.9 ± 0.3 kcal/mol, for G330T and $G330T^{\Delta}$, respectively. In cycles ①, ③ and ④ at least one of the constituents of the full-length proteins is highly dependent on N-terminus dynamics, and it is therefore not straightforward to judge which will cost the most. We do find through FEP calculations, however, that cycle ① which has both WT_{L1}, $G330T_{L1}$ involved is the most affected, with the cost of the G →T transition in the presence of L_1 costing 3.3 kcal/mol more when the charged N-terminus is removed. This is followed by cycle ④ where the binding pocket H→A transition in the presence of L_2 has an additional cost of 2.6 kcal/mol.

The loss of the N-terminus translates into modified binding pocket interactions where we may trace the origin of the differing free energy differences. Removal of N-terminus increases $N_{H-bonds}$ in all cases by an average of 0.2-0.5, except in $H372A^{\Delta}{}_{L1}$ where it decreases by 0.5 (Table 3.1). For the G330T transition, L_2 binding is more favorable, mainly because the deletion of the N-terminus leads to a net gain of nearly one full hydrogen bond (from 3.4 to 4.3) while it is a mere 0.3 gain (from 3.4 to 3.7) in its L_1 bound form, also corroborated by $\Delta\Delta\Delta G_1^{\Delta}$ = -5.0 ± 0.4 kcal/mol (Figure 3.9 b). For the H372A transition, L_2 binding is again more favorable, but with a lower propensity ($\Delta\Delta\Delta G_2^{\Delta}$ = -2.7 ± 0.4 kcal/mol), as the L_1 and L_2 binding cost -0.4, and +0.1 hydrogen bonds, respectively. Finally, we expect the DM^{Δ} systems to favor binding both L_1 and L_2 due to the gain of a full hydrogen bond in the binding pocket in the absence of the N-terminus. We thus expect that if the total binding free energy is in the functional range, then the ligand bridging behavior of the $G330T^{\Delta}$ mutation will change to ligand switching while the reverse is expected to happen for DM^{Δ}.

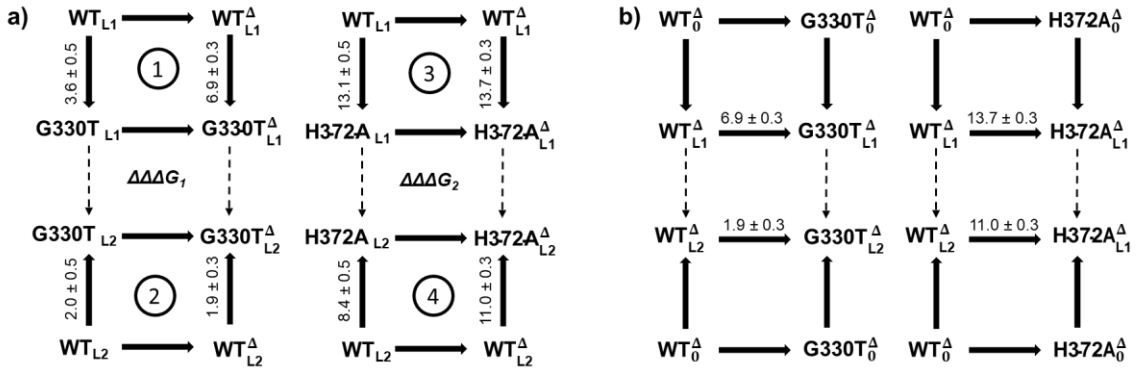

Figure 3.9 **a** Thermodynamic cycle depicting the role of N-terminus removal on the single mutations. Vertical arrows indicate the mutation process, and the horizontal arrows display the deletion operation on complexes denoted by the superscript Δ. The difference between the horizontal changes is equivalent to that of the vertical ones in each cycle. $\Delta\Delta G$ values of the cycles ① and ④ indicate that the $G330T_{L1}$ and $H372A_{L2}$ forms are highly unfavorable after the operation, while those for ② and ③ have low or no cost. **b** Thermodynamic cycles depicting mutation and ligand binding of the truncated complexes; ΔΔΔG represent the grand difference between the binding affinities of the WT versus mutant towards either ligand as ΔG of the apo complexes cancel out. $\Delta\Delta\Delta G_1^{\Delta}$ and $\Delta\Delta\Delta G_2^{\Delta}$ are -5.0 ± 0.4 and -2.7 ± 0.4 kcal/mol, respectively.

3.3 Intramolecular Residue Interaction of PDZ3 explained by Dynamic Community Composition

After investigating the functionality of PDZ3 by FEP simulations and thermodynamic integration, we scrutinize the residue interaction to explain the biophysics of binding selectivity of PDZ3, by employing graph theoretical approach. Hence, RNs are constructed through MD trajectories of PDZ3 complexes; then, by assessing BC of nodes and edges, we detect close-knit residue group (communities) and survey the structural origins of the community members (Figure 1.2c).[41]

3.3.1 Construction of Dynamic RNs

To construct a graph of the protein structure, C_β of each residue (C_α for glycine) is taken as a node to preserve side-chain sensitivity in calculations. Nodes within a 6.7 Å distance are taken as interacting, and an edge is assigned between them. The cut-off distance of 6.7 Å is chosen for linking the first coordination shell of C_β atoms in radial distribution function (RDF) which belongs to adjoint residues and other residues that locate close to the central residue;[20, 29] for a detailed account of the choice of cutoffs in residue networks, see ref[55]. This cut-off is validated by the RDF analysis of our MD simulations, which shows that the dynamics and sampled system do not affect the overall coordination shell location (Figure 3.10). The first 80 ns portion of all MD trajectories are discarded for equilibration, and the last 120 ns portions are used in all analyses; therefore, snapshots of 240 ns long MD simulation for WT, G330T, H372A and DM in complex with L_1 and L_2 are utilized for further analyses. Additionally, MD simulations for truncated (Δ) complexes are performed for 50 ns, and the whole trajectory is utilized for calculations in these systems with minimal amounts of fluctuations.

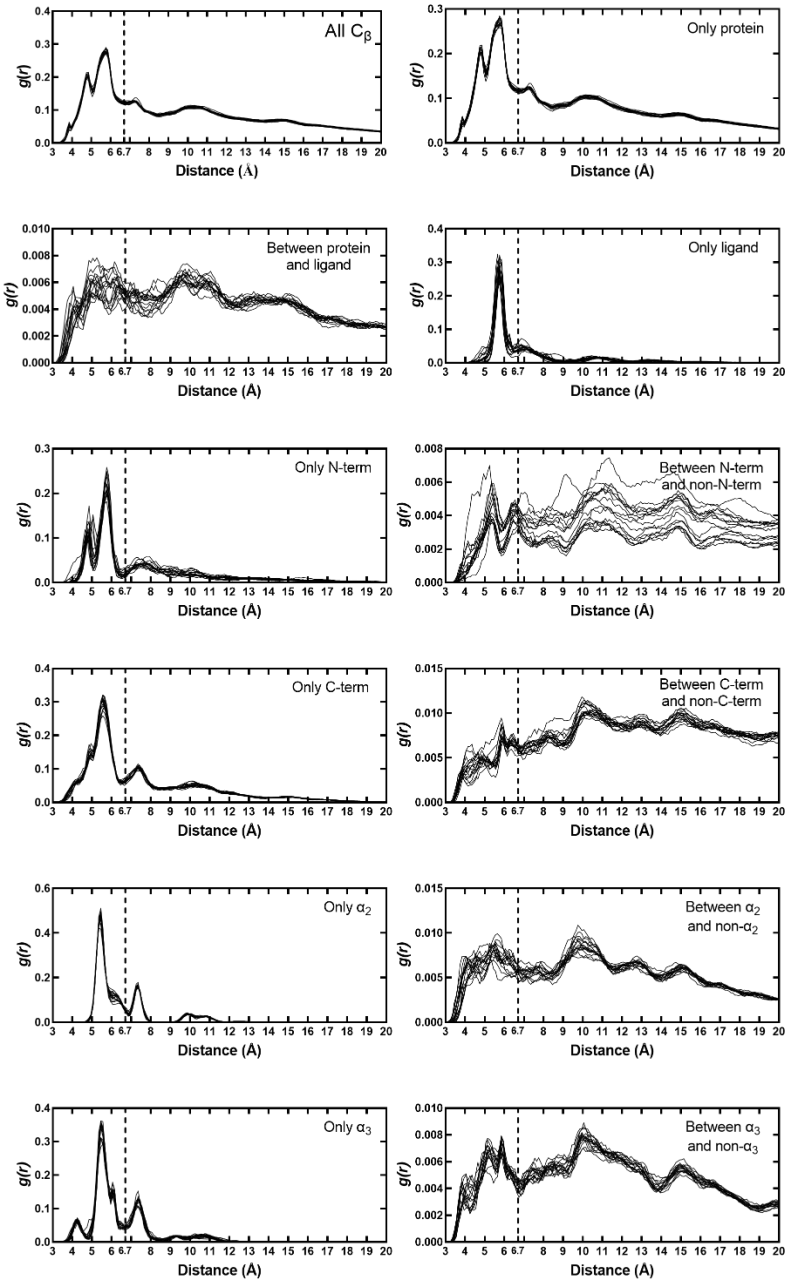

Figure 3.10 Radial distribution function, $g(r)$, between residue centers, results from the 16 simulations belonging to all full length PDZ3 complexes are superimposed. The calculations are done for different structural segments, such as only ligand, α_2, α_3 and N/C termini. Additionally, $g(r)$ between each structural segment and every other residue that does not belong to that segment is computed. The cut-off distance of 6.7 Å signified by dashed line marks the completion location of the first coordination shell of non-bonded Cβ atoms; this cut-off is utilized for all network analyses in this study. Note that selection of Cβ atoms as coarse-graining centers is essential to capture dynamical communication between residue side chains.

3.3.2 Node BC, Detection of Communities and Structural Origins of Members

Node BC gives both local and global information about a graph, and it has been effectively used for structural and functional assessment of proteins, and slight differences of node BC lead significant changes in a network.[20, 34-36, 38, 39]

Further, Girvan-Newman algorithm is employed to detect communities.[37, 83] A community is a group of nodes that are connected to each other without having an edge to the nodes out of this group. Hence, members of same community are in cooperation, which, in the context of proteins, translates to function. The algorithm searches for communities by breaking the most central (highest BC) edge in each iteration, and the occurring communities (Ω) are utilized for further analyses. Along with the total edge count, number of removed edges until community separation is achieved proves informative about the state of the whole system. To scrutinize the structural origins of community members, we devise a simple algorithm. For an RN belonging to an MD snapshot, each community in Ω is checked whether at least one member from desired two different structural segments is located in the same community; if they are, the score for that Ω is 1, otherwise it is 0. The sum of scores is normalized by the total number of MD snapshots for average results.

Accordingly, 240 snapshots from MD simulations (extracted from the equilibrated portions of the trajectories at 1 ns intervals) of each complex are utilized for the average results, and 24 snapshots (extracted at 10 ns intervals) are used for the detailed investigation of the conformations.

3.3.3 Visualization of Community Dynamics on Three-Dimensional Protein Structure

To visualize the dynamical shifts in shared communities, we apply the following procedure: Our aim is to color each residue according to its persistence in a given community. Thus, we first select three residues that predominantly remain in separate communities and are in rigid structural elements. Here we have selected I316 in β_1, A375 in α_2 and F400 in β_3. These residues are attributed the colors red (R), green (G) and blue (B), respectively. After performing community detection at each snapshot for a given Ω, we assign an attribute of R, G, B or null to each residue at every time point t in vector $\mathbf{c}_i(t)$ such that if the residue is in the same community with I316, $\mathbf{c}_i(t)$ = [1 0 0], if it is with A375, $\mathbf{c}_i(t)$ = [0 1 0], if it is with F400, $\mathbf{c}_i(t)$ = [0 0 1]; $\mathbf{c}_i(t)$ = [0 0 0] otherwise. The color of the protein is accumulated in the color matrix $\mathbf{C}$ of dimensions $n \times 3$ with each row holding the RGB color code of the residue, $\mathbf{C}_i = \sum_t \mathbf{c}_i(t) \times 255/T$, where T is the total number of time points so that the normalization by $255/T$ allows the use of decimal RGB code. The protein three-dimensional structure is then colored according to these values. As a result, residues always co-inhabiting a community with I316, A375 or F400 will have a pure R, G or B color, but those switching between regions will have blended colors, e.g., a residue spending half of its time in the same community with I316 and the other with A375 will appear yellow. Finally, those residues that are never clustered with any of the three will be colored black. In this study, this visualization has been applied to community sizes of $\Omega = 4$.

3.3.4 Betweenness Centrality Unveils Hinge Residues Affecting Function of the Complex

We first investigate if the BC of residues calculated as an average over the snapshots obtained through MD provide information additional to their mean-squared fluctuations (MSF). Selected case of H372A_{L2} is displayed in Figure 3.11a, and data for all cases are provided in

Figure 3.12. It is well known that high MSF residues are in mobile regions, mostly on solvent exposed loops, while residues in secondary structural elements have low MSF. Meanwhile, BC works at the resolution of single residues and reveals that even in flexible loops there are residues with high BC, undertaking hinge roles in PDZ binding mechanics (Figure 3.11a). We note that for the same mutant, binding different ligands may significantly shift the centrality of resides, best displayed by the ΔBC curves as exemplified for H372A in Figure 3.11b wherein the N-terminus residues G303 and E304 as well as the turn residue S409 have much increased centrality in the functional $H372A_{L2}$ compared to the low binding affinity complex $H372A_{L1}$ (Figure 3.11b).

In fact, these residues arise frequently amongst those with the largest ΔBC in all variants, whose locations and interactions are displayed in Figure 3.11c indicating the high centrality of N and C termini. We find G303/E304/I307 in the N-terminus, Y392 at the beginning of the α_3 helix, and S409/G410 in the C-terminus to significantly shift their centrality depending on the variant. Moreover, in truncation simulation series, $PDZ3^{\Delta}$, whereby the N-terminus has been deleted, the BC of C-terminus region drops in all cases, indicating that N and C termini interaction is substantial for the formation of the PDZ complexes. In these truncation variants, Y392, Y397 and S409 commonly lose centrality. Note that Y397 resides in the middle of the α_3 helix and does not directly interact with the N-terminus in full length proteins providing an example of how the lost interactions lead to a domino effect that reflect into the bulk of the protein in the communication. Finally, note how for all complexes, BC of Y392 shifts, emphasizing the function of this residue on information flow. Y392 is located at the beginning of α_3, and it is central in all complexes; along with the high centrality of N/C termini, this residue appears to hold a mediator function for PDZ3 without conferring ligand specificity.

Additionally, we compute aforementioned current flow and communicability BC analyses[57, 58] for every PDZ3 complex with a dynamical approach; however, we find that three BC measures have a similarity of higher than ~0.9 Pearson *R*. Hence, we employed only 'shortest path' BC to investigate PDZ3.

While residues with high BC changes in the different complexes may indicate regions to target to alter the function of a protein, they alone cannot pinpoint how function is dynamically orchestrated. For this purpose, we turn to analyze how communities of residues are structured.

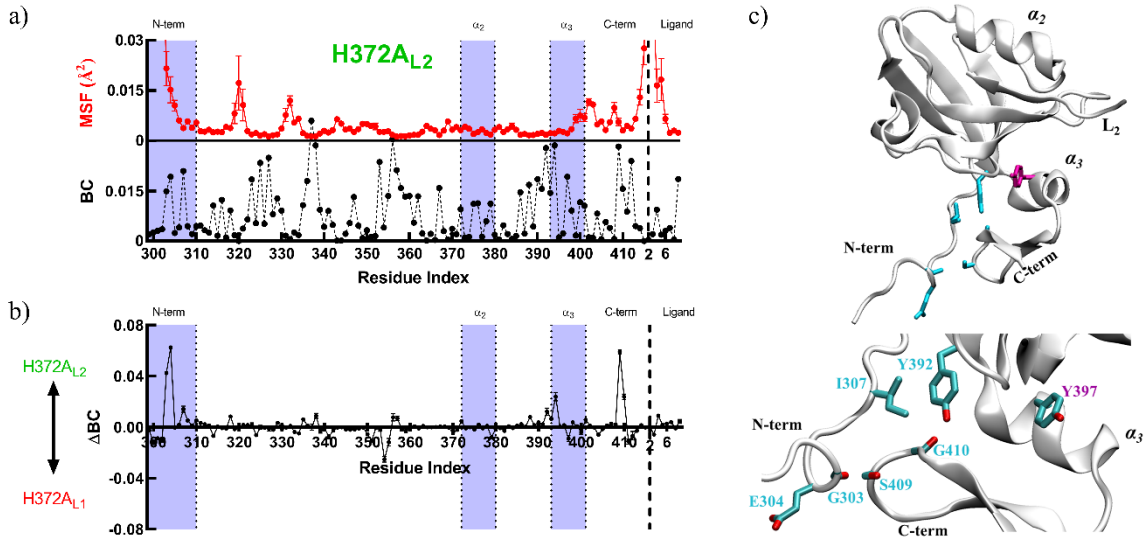

Figure 3.11 **a** Sample MSF and BC for H372A$_{L2}$. Structural segments are labeled at the top of the graph, and segments of interest are highlighted in purple. MSF results are averages over six chunks of 40 ns each obtained from the equilibrated duplicate MD trajectories. **b** ΔBC between L_1 and L_2 bound forms of H372A. Along with α_3, N and C termini are more central in H372A$_{L2}$. **c** Residues with $|\Delta BC| > 0.04$ emerging in any of the studies complexes are mapped on the three-dimensional PDZ structure; residues with large changes in full length proteins are shown in cyan; Y397 (magenta) is highlighted only in truncation mutants along with Y392 and S409 that appear in both types of systems. All these residues reside on the N-terminus, C-terminus or the α_3 helix; their locations and interactions shown in detail below.

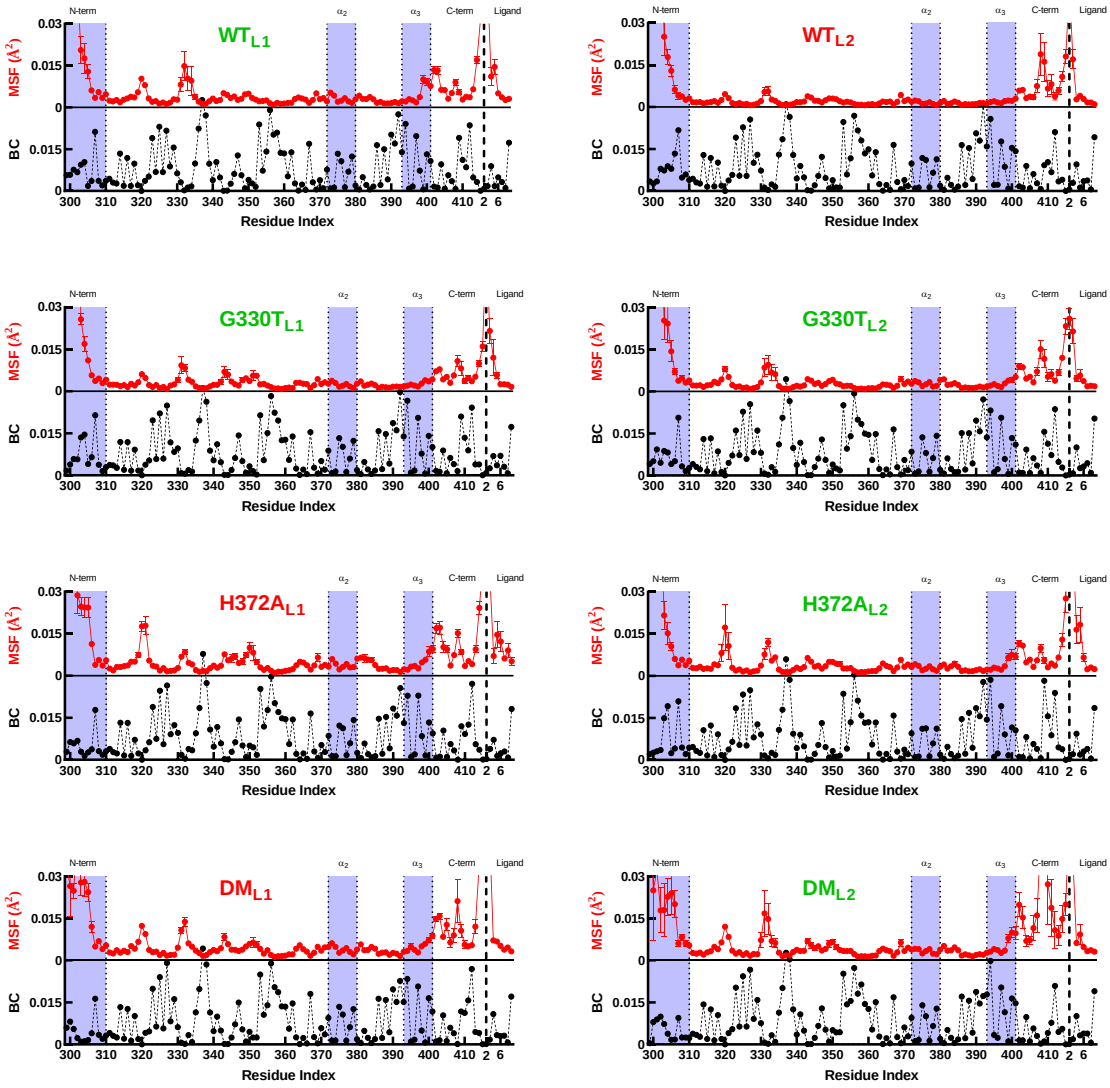

Figure 3.12 BC and MSF of each PDZ3 complex. Structural segments are labeled at the top of the graph, and segments of interest are highlighted. Favorable and unfavorable complexes are labelled by green and red, respectively. MSF results are computed by using C_β atoms of 40-ns long chunks of MD trajectories in the 120-200 ns time window; thus, six chunks are averaged, and error bars are displayed.

3.3.5 Number of Broken Edges and Size of the Communities Illustrate Diverse Organizations of PDZ3

Rather than individual residues, we now focus on the edges of the PDZ3-RNs, i.e., the interactions between residue pairs. In particular, an analysis of groups of residues working together is deciphered by studying the composition of communities formed by structural segments during the course of the time in the trajectories. To detect the communities, the edges are removed one-by-one hierarchically, starting from the most central, until a group of residues separate out into a disconnected community. The total number of edges in the range of 360-380 show that RNs of PDZ3 are very sparse (Table 3.4), compared to the maximum number of possible edges, which is $n(n-1)/2 = 7750$ where $n = 125$ is the number of nodes. Considering the sparse character of the networks and the prior assessment of the communities, a community window of Ω = 3-6 is selected for detailed study. At Ω = 2, the flexible N-terminus and protein body are grouped separately as a trivial result; for Ω > 6 single residue communities start to dominate.

Table 3.4 Averages for total number of edges, and number of broken edges to achieve a community of size Ω.

System label	Total number of edges	Average number of broken edges			
		Ω = 3	**Ω = 4**	**Ω = 5**	**Ω = 6**
WT_{L1}	365 ± 1	31 ± 1	51 ± 1	63 ± 1	72 ± 1
WT_{L2}	364 ± 1	29 ± 1	50 ± 1	63 ± 1	73 ± 1
$G330T_{L1}$	363 ± 1	26 ± 1	47 ± 1	60 ± 1	70 ± 1
$G330T_{L2}$	364 ± 1	30 ± 1	49 ± 1	62 ± 1	71 ± 1
$H372A_{L1}$	366 ± 1	36 ± 1	54 ± 1	64 ± 1	71 ± 1
$H372A_{L2}$	367 ± 1	27 ± 1	48 ± 1	60 ± 1	71 ± 1
DM_{L1}	376 ± 1	38 ± 1	57 ± 1	68 ± 1	77 ± 1
DM_{L2}	374 ± 1	40 ± 1	56 ± 1	67 ± 1	75 ± 1

To provide an insight on how the community sharing and BC data lead to complementary information, in Figure 3.13 we display as heatmaps the BC value at each instant of the trajectories, accompanied below them with the fraction of time the N-terminus and the ligand share a community for the range of Ω = 3-6. The ligand has low BC in all cases (black stripes in the topmost part of the figures), and several residues of the N-terminus has high BC (light colored instants in the lowest part of the figures). Nevertheless, these two regions frequently share a community for Ω = 3, but their communities may further separate out depending on the system studied for larger Ω values (checkerboxes in Figure 3.13).

The total number of edges averaged over the trajectory snapshots as well as the number of edges needed to be broken to reach a given community size are listed in Table 3.4. They are ~365 in WT and the single mutants, while DM forms have ~375 edges. This is a result of the close-knit character of DM complexes, having more hydrogen bonds at the binding pocket and low overall solvent accessibility, as quantified in detail in previous section.[23] Although the averages are similar, the number of broken edges fluctuate over the course of the trajectories (Figure 3.14). These changes indicate that the configurations for information flow are modified throughout the MD trajectories. When we focus on the size of the communities, we find that at Ω = 3 there is one large community having 60-100 residues accompanied by two smaller ones (compare the green curves to red and blue curves in Figure 3.15 for Ω = 3). As more edges are removed, the variations at different time points smooth out and the size of the communities gets more similar, with the largest community having 20-60 members and the smallest one having a few members (see Figure 3.15 for Ω = 6).

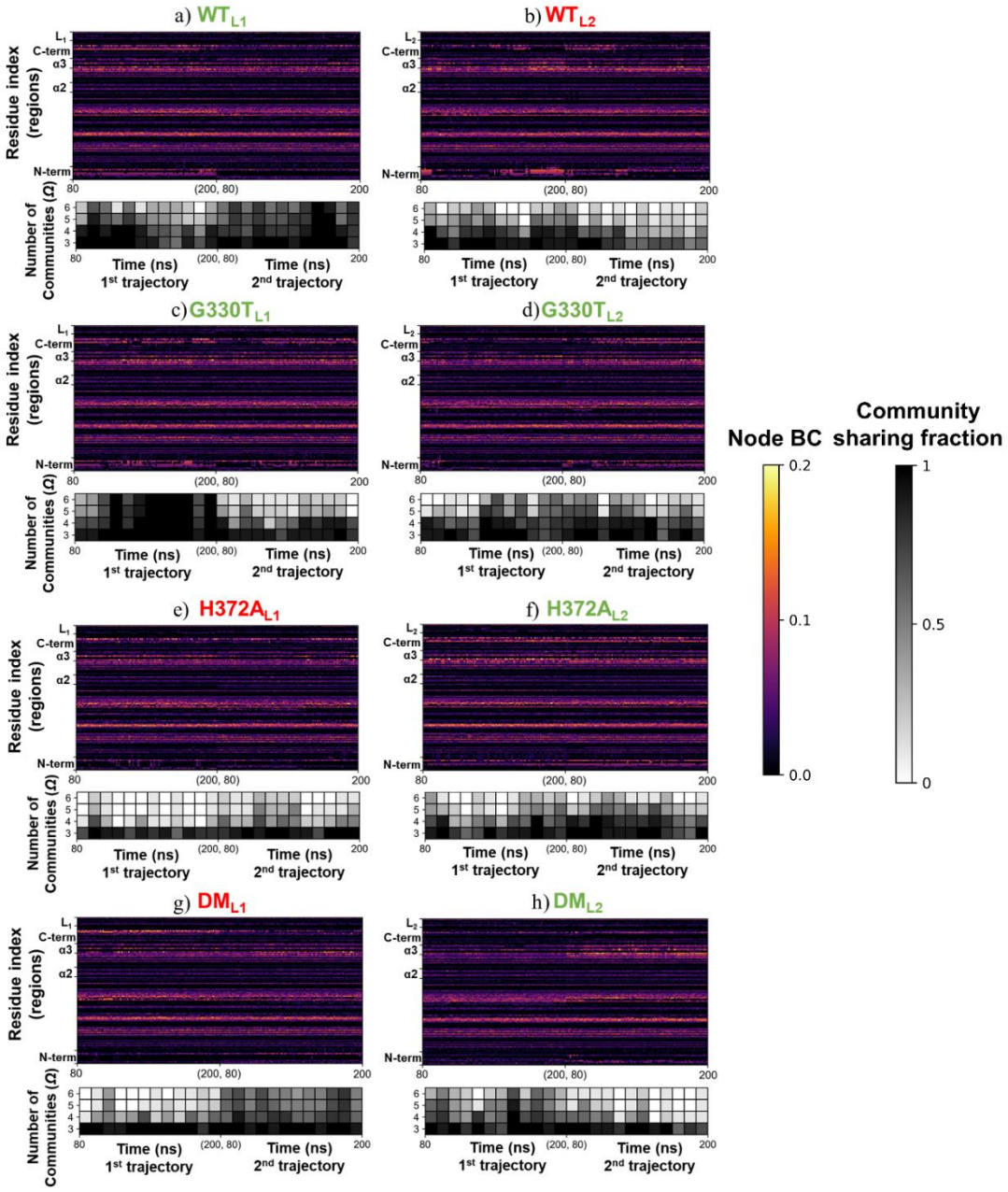

Figure 3.13 Heatmap of Node BC (colored; labelled by structural segments instead of residue indices) paired to the fraction of community sharing of N-terminus and ligand for Ω = 3-6 (grayscale cells, each cell average of 10 snapshots) throughout the MD trajectories. The results for two replica trajectories are concatenated, with time points covering 80-200 ns for each as shown in x-axis labels.

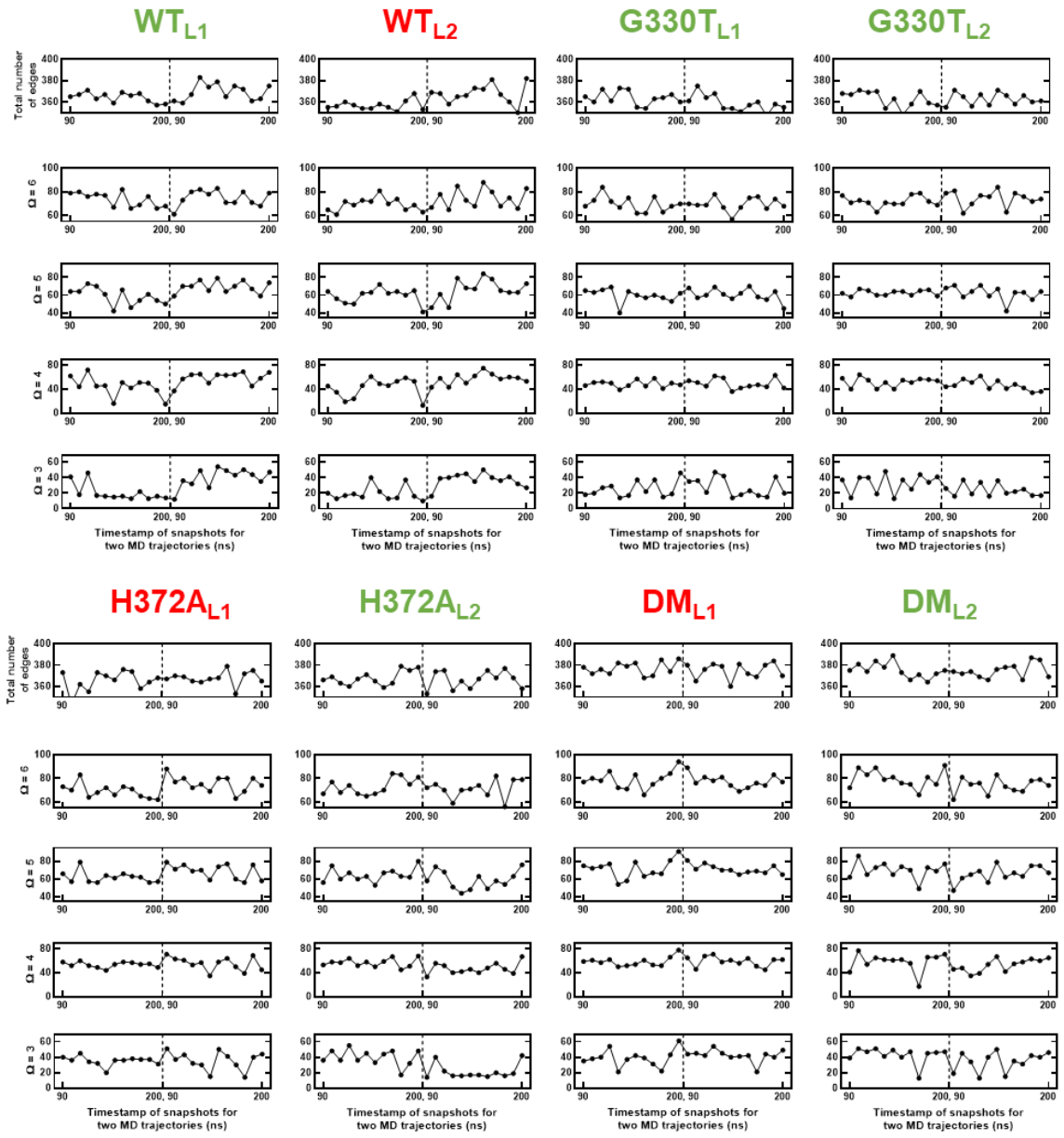

Figure 3.14 Time evolution of the number of broken edges for emergence of Ω =3-6 communities.

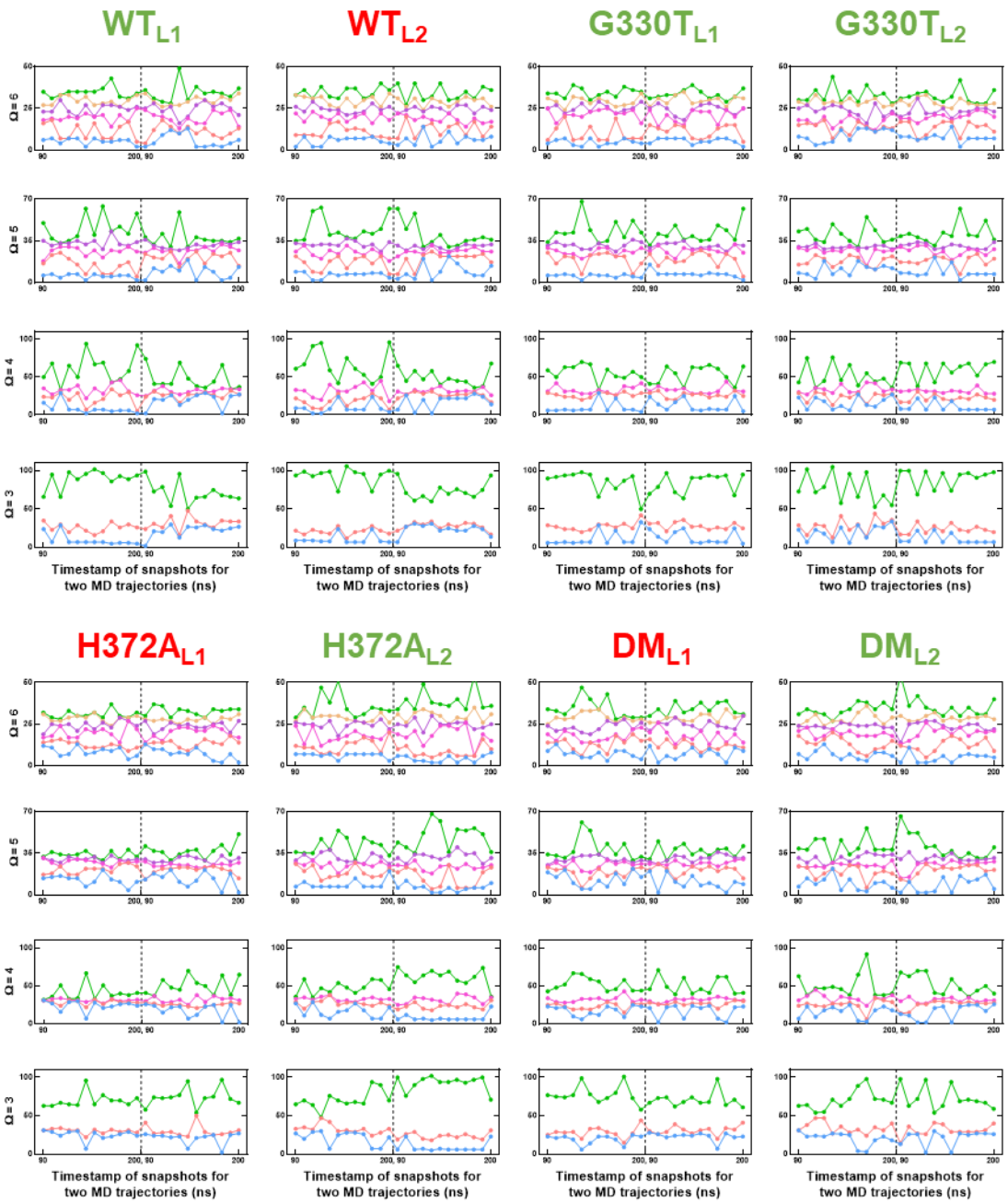

Figure 3.15 Time evolution of number of community members for Ω =3-6 communities. Smallest community is shown in blue, while the largest in green.

The composition of the communities is investigated by assessing the structural origins of its members. The fraction of being a member of the same community is calculated for all available pairs between five structural segments that have been determined in the BC analysis to be essential for PDZ3; namely the N/C termini, α_2/α_3 helices and ligand (refer to Figure 1.3 for residue indices). Note that two segments are classified to be members of the same community if at least one residue is shared; therefore, the fractions do not sum to 1. The results for N-terminus and ligand and community sharing for the range of Ω = 3-6 are listed in Table 3.5 and that for all pairs of structural segments is in Table 3.6.

Table 3.5 Fraction of instances that N-terminus and ligand co-inhabit a community.*

System label	$\Omega = 3$	$\Omega = 4$	$\Omega = 5$	$\Omega = 6$
WT_{L1}	**0.93 ± 0.02**	**0.84 ± 0.02**	0.66 ± 0.03	0.56 ± 0.03
WT_{L2}	**0.84 ± 0.02**	0.63 ± 0.03	0.32 ± 0.03	0.14 ± 0.02
$G330T_{L1}$	**0.90 ± 0.02**	**0.80 ± 0.03**	0.44 ± 0.03	0.26 ± 0.03
$G330T_{L2}$	**0.93 ± 0.02**	**0.79 ± 0.03**	0.55 ± 0.03	0.48 ± 0.03
$H372A_{L1}$	**0.88 ± 0.02**	0.31 ± 0.03	0.14 ± 0.02	0.09 ± 0.02
$H372A_{L2}$	**0.88 ± 0.02**	**0.70 ± 0.03**	0.35 ± 0.03	0.15 ± 0.02
DM_{L1}	**0.97 ± 0.01**	0.56 ± 0.03	0.36 ± 0.03	0.33 ± 0.03
DM_{L2}	**0.90 ± 0.02**	0.50 ± 0.03	0.32 ± 0.03	0.21 ± 0.03

*values greater than 0.7 shown in bold; those less than 0.5 are colored blue.

Table 3.6 Fraction of instances pairs of structural segments co-inhabit a community.*

System label	N-term and C-term				N-term and α_3				N-term and α_2			
	$\Omega = 3$	$\Omega = 4$	$\Omega = 5$	$\Omega = 6$	$\Omega = 3$	$\Omega = 4$	$\Omega = 5$	$\Omega = 6$	$\Omega = 3$	$\Omega = 4$	$\Omega = 5$	$\Omega = 6$
WT_{L1}	**1.00 ± 0.01**	**1.00 ± 0.01**	**0.98 ± 0.01**	**0.97 ± 0.01**	**0.99 ± 0.01**	**0.96 ± 0.01**	**0.88 ± 0.02**	**0.78 ± 0.03**	**0.85 ± 0.02**	0.60 ± 0.03	0.33 ± 0.03	0.16 ± 0.02
WT_{L2}	**0.80 ± 0.03**	**0.79 ± 0.03**	**0.75 ± 0.03**	**0.70 ± 0.03**	**0.89 ± 0.02**	**0.84 ± 0.02**	**0.73 ± 0.03**	0.67 ± 0.03	**0.81 ± 0.03**	0.61 ± 0.03	0.33 ± 0.03	0.12 ± 0.02
$G330T_{L1}$	**0.97 ± 0.01**	**0.97 ± 0.01**	**0.96 ± 0.01**	**0.95 ± 0.01**	**0.95 ± 0.01**	**0.90 ± 0.02**	**0.85 ± 0.02**	**0.83 ± 0.02**	**0.87 ± 0.02**	0.68 ± 0.03	0.31 ± 0.03	0.11 ± 0.02
$G330T_{L2}$	**0.92 ± 0.02**	**0.91 ± 0.02**	**0.88 ± 0.02**	**0.85 ± 0.02**	**0.97 ± 0.01**	**0.93 ± 0.02**	**0.86 ± 0.02**	**0.79 ± 0.03**	**0.86 ± 0.02**	0.62 ± 0.03	0.35 ± 0.03	0.17 ± 0.02
$H372A_{L1}$	**1.00 ± 0.01**	**0.99 ± 0.01**	**0.99 ± 0.01**	**0.98 ± 0.01**	**0.98 ± 0.01**	**0.96 ± 0.01**	**0.91 ± 0.02**	**0.83 ± 0.02**	**0.85 ± 0.02**	0.43 ± 0.03	0.28 ± 0.03	0.15 ± 0.02
$H372A_{L2}$	**1.00 ± 0.01**	**1.00 ± 0.01**	**1.00 ± 0.01**	**1.00 ± 0.01**	**0.96 ± 0.01**	**0.95 ± 0.01**	**0.94 ± 0.02**	**0.90 ± 0.02**	**0.86 ± 0.02**	0.68 ± 0.03	0.38 ± 0.03	0.17 ± 0.02
DM_{L1}	**1.00 ± 0.01**	**1.00 ± 0.01**	**0.99 ± 0.01**	**0.99 ± 0.01**	**0.99 ± 0.01**	**0.94 ± 0.02**	**0.82 ± 0.02**	**0.72 ± 0.03**	**0.92 ± 0.02**	0.46 ± 0.03	0.19 ± 0.03	0.10 ± 0.02
DM_{L2}	**0.99 ± 0.01**	**0.98 ± 0.01**	**0.97 ± 0.01**	**0.96 ± 0.01**	**0.99 ± 0.01**	**0.98 ± 0.01**	**0.95 ± 0.01**	**0.94 ± 0.02**	**0.86 ± 0.02**	0.41 ± 0.03	0.21 ± 0.03	0.13 ± 0.02
System label	C-term and ligand				C-term and α_3				C-term and α_2			
	$\Omega = 3$	$\Omega = 4$	$\Omega = 5$	$\Omega = 6$	$\Omega = 3$	$\Omega = 4$	$\Omega = 5$	$\Omega = 6$	$\Omega = 3$	$\Omega = 4$	$\Omega = 5$	$\Omega = 6$
WT_{L1}	0.38 ± 0.03	0.35 ± 0.03	0.32 ± 0.03	0.31 ± 0.03	**0.95 ± 0.01**	**0.95 ± 0.01**	**0.95 ± 0.01**	**0.95 ± 0.01**	0.17 ± 0.02	0.07 ± 0.02	0.03 ± 0.01	0.01 ± 0.01
WT_{L2}	0.35 ± 0.03	0.31 ± 0.03	0.28 ± 0.03	0.26 ± 0.03	**0.96 ± 0.01**	**0.95 ± 0.01**	**0.95 ± 0.01**	**0.95 ± 0.01**	0.28 ± 0.03	0.17 ± 0.02	0.12 ± 0.02	0.11 ± 0.02
$G330T_{L1}$	0.42 ± 0.03	0.39 ± 0.03	0.37 ± 0.03	0.35 ± 0.03	**0.97 ± 0.01**	**0.96 ± 0.01**	**0.96 ± 0.01**	**0.95 ± 0.01**	0.18 ± 0.03	0.07 ± 0.02	0.03 ± 0.01	0.01 ± 0.01
$G330T_{L2}$	0.31 ± 0.03	0.24 ± 0.03	0.23 ± 0.03	0.21 ± 0.03	**0.96 ± 0.01**	**0.96 ± 0.01**	**0.96 ± 0.01**	**0.96 ± 0.01**	0.26 ± 0.03	0.10 ± 0.02	0.06 ± 0.02	0.04 ± 0.01
$H372A_{L1}$	0.26 ± 0.03	0.15 ± 0.02	0.11 ± 0.02	0.09 ± 0.02	**0.92 ± 0.02**	**0.92 ± 0.02**	**0.91 ± 0.02**	**0.91 ± 0.02**	0.22 ± 0.03	0.10 ± 0.02	0.04 ± 0.01	0.01 ± 0.01
$H372A_{L2}$	0.27 ± 0.03	0.22 ± 0.03	0.19 ± 0.03	0.17 ± 0.02	**0.98 ± 0.01**	**0.98 ± 0.01**	**0.98 ± 0.01**	**0.98 ± 0.01**	0.15 ± 0.02	0.08 ± 0.02	0.04 ± 0.01	0.02 ± 0.01
DM_{L1}	0.48 ± 0.03	0.25 ± 0.03	0.17 ± 0.02	0.15 ± 0.02	**0.91 ± 0.02**	**0.91 ± 0.02**	**0.91 ± 0.02**	**0.91 ± 0.02**	0.46 ± 0.03	0.20 ± 0.03	0.10 ± 0.02	0.08 ± 0.02
DM_{L2}	0.23 ± 0.03	0.17 ± 0.02	0.16 ± 0.02	0.14 ± 0.02	**0.98 ± 0.01**	**0.98 ± 0.01**	**0.98 ± 0.01**	**0.98 ± 0.01**	0.19 ± 0.03	0.10 ± 0.02	0.07 ± 0.02	0.05 ± 0.01
System label	α_3 and ligand				α_3 and α_2				α_2 and ligand			
	$\Omega = 3$	$\Omega = 4$	$\Omega = 5$	$\Omega = 6$	$\Omega = 3$	$\Omega = 4$	$\Omega = 5$	$\Omega = 6$	$\Omega = 3$	$\Omega = 4$	$\Omega = 5$	$\Omega = 6$
WT_{L1}	**0.84 ± 0.02**	**0.80 ± 0.03**	**0.73 ± 0.03**	0.69 ± 0.03	**0.74 ± 0.03**	0.58 ± 0.03	0.41 ± 0.03	0.28 ± 0.03	**0.99 ± 0.01**	**0.92 ± 0.02**	**0.86 ± 0.02**	**0.80 ± 0.03**
WT_{L2}	**0.97 ± 0.01**	**0.94 ± 0.02**	**0.88 ± 0.02**	**0.85 ± 0.02**	**0.96 ± 0.01**	**0.83 ± 0.02**	**0.76 ± 0.03**	0.64 ± 0.03	**1.00 ± 0.01**	**0.95 ± 0.01**	**0.94 ± 0.02**	**0.89 ± 0.02**
$G330T_{L1}$	**0.86 ± 0.02**	**0.80 ± 0.03**	0.68 ± 0.03	0.63 ± 0.03	**0.77 ± 0.03**	0.55 ± 0.03	0.35 ± 0.03	0.23 ± 0.03	**0.99 ± 0.01**	**0.93 ± 0.02**	**0.85 ± 0.02**	**0.78 ± 0.03**
$G330T_{L2}$	**0.80 ± 0.03**	**0.72 ± 0.03**	0.66 ± 0.03	0.63 ± 0.03	**0.75 ± 0.03**	0.54 ± 0.03	0.44 ± 0.03	0.34 ± 0.03	**0.98 ± 0.01**	**0.91 ± 0.02**	**0.86 ± 0.02**	**0.82 ± 0.03**
$H372A_{L1}$	**0.89 ± 0.02**	0.62 ± 0.03	0.57 ± 0.03	0.56 ± 0.03	**0.87 ± 0.02**	0.65 ± 0.03	0.58 ± 0.03	0.51 ± 0.03	**1.00 ± 0.01**	**1.00 ± 0.01**	**1.00 ± 0.01**	**1.00 ± 0.01**
$H372A_{L2}$	**0.91 ± 0.02**	**0.80 ± 0.03**	0.68 ± 0.03	0.62 ± 0.03	**0.88 ± 0.02**	**0.75 ± 0.03**	0.64 ± 0.03	0.54 ± 0.03	**1.00 ± 0.01**	**1.00 ± 0.01**	**0.98 ± 0.01**	**0.96 ± 0.01**
DM_{L1}	**0.97 ± 0.01**	**0.73 ± 0.03**	0.68 ± 0.03	0.66 ± 0.03	**0.95 ± 0.01**	0.67 ± 0.03	0.58 ± 0.03	0.54 ± 0.03	**1.00 ± 0.01**	**1.00 ± 0.01**	**0.99 ± 0.01**	**0.98 ± 0.01**
DM_{L2}	**0.90 ± 0.02**	**0.77 ± 0.03**	**0.70 ± 0.03**	0.66 ± 0.03	**0.85 ± 0.02**	0.67 ± 0.03	0.57 ± 0.03	0.50 ± 0.03	**1.00 ± 0.01**	**0.99 ± 0.01**	**0.98 ± 0.01**	**0.97 ± 0.01**

*values greater than 0.7 shown in bold; those less than 0.5 are colored blue.

First and foremost, N-terminus – C-terminus – α_3 are located in the same community even for the largest value considered of $\Omega = 6$ with fraction of time they spend together in pairs exceeding values 0.7 throughout and nearly equal to 1 in most cases (Table 3.6). In addition to the previously mentioned node BC results, the high fraction of N/C termini co-inhabiting the same community along with α_3 emphasizes the interaction between the two segments is critical for all complexes. The two termini exert a clamping effect that facilitates the overall dynamics.[24] Considered with the high node BC (Figure 3.11, 3.12), α_3 acts as a hub between structural segments, jointly with its neighboring Y392. Similarly, α_2 that hosts H372 and the ligand reside in the same community throughout the trajectories (Table 3.6), an expected result since α_2 lines the ligand (Figure 1.3). A series of manuscripts discuss the allosteric communication between the ligand binding site and the α_3 helix.[5, 7-11] Our analyses show that the two regions share a community with a fraction higher than ~0.6 in all PDZ complexes even at $\Omega = 6$. However, community sharing between α_3 and α_2 is relatively low, particularly for WT_{L1} and G330T single mutation systems.

The fraction of community sharing between N-terminus and ligand is a measure of the extent to which the distal mobile region affects ligand binding (Table 3.5). We find that this couple tends to be in the same community in favorable complexes up to $\Omega = 5$, except for DM_{L2}. Also, the analyses between N/C termini, α_2 and α_3 do not indicate this specificity, which means that N-terminus communicates directly with the ligand. The effect is due to the long-range electrostatic interactions between the N-terminus having -4 net charge and the ligand with +2 charge. It is filtered through the protein core, and its range is manipulated by point mutations at positions 330 and 372. As a result, the flexible N-terminus interacts with the ligand and affects binding specificity, significantly for WT and single mutations. We have shown previously[23] that the strong ligand binding preference of DM is dominated by the tight structure attained by this variant upon ligand binding that affects the overall solvation free energy change due to the reduction in solvent accessible surface area and additional hydrogen bonds counts at the binding site. The preference of L_2 binding over L_1 in this case is explained by the enhanced BC of the α_3 hub that manifests itself as the tendency of the N-terminus, C-

terminus and α_3 to coinhabit the same community even at $\Omega = 6$ (fraction of finding all three together is 0.88 for DM_{L2} vs. only 0.47 for DM_{L1}.)

The analyses of the remaining segments do not provide information on binding specificity of PDZ3. For example, the C-terminus does not communicate with structural segments other than the N-terminus. Although, PDZ3-specific α_3 has a higher fraction of communication with ligand and α_2, the values do not differentiate functional complexes implying that α_3 does not directly affect ligand selectivity. We therefore find that the N-terminus is the main region affected by the single residue changes that lead to ligand selectivity in PDZ3.

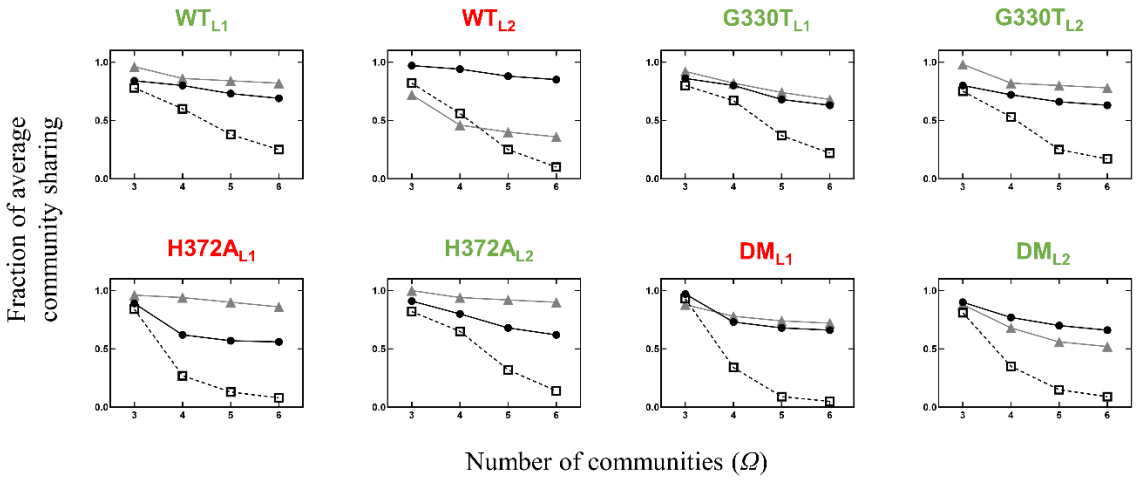

Figure 3.16 Average community co-occupancy fraction for, (☐) N-terminus, α_3 and ligand triplet; (●) α_3 and ligand pair in full length PDZ3 systems; (Δ) α_3 and ligand in $PDZ3^{\Delta}$ systems.

In fact, removal of the N-terminus significantly alters the community structure of PDZ3. In Table 3.7 we display the community sharing fractions for all pairs of the remaining regions in $PDZ3^{\Delta}$ simulation series where the values that differ from the full length PDZ3 variant (Table 3.6) by more than the error margins are colored in red. We find that the already low community sharing between C-terminus and α_2/ligand in the full length PDZ3 is nearly completely lost even for $\Omega = 3$ in all variants except $H372A_{L2}$ where there is slight increase in communication between these regions. Moreover, focusing on the central role attributed

to α_3 to ligand binding (Figure 3.16), their community structure is also significantly altered, especially for WT_{L2} whose communication with the ligand is substantially disrupted for WT_{L2}, but reinforced for $H372A_{L1}$, and $H372A_{L2}$. The N-terminus is almost always grouped together with α_3 and ligand for $\Omega = 3$ but is the first region to separate out for $\Omega > 3$ (Figure 3.16, empty squares). If this region were not to affect the community dynamics of α_3 and ligand, one would expect α_3 and ligand to have the same occupancy fractions in PDZ3 and $PDZ3^{\Delta}$ simulations (Figure 3.16 circles and triangles). Indeed, we have discussed how the binding affinity of the DM variants is mainly governed by the tight overall structure, unlike the rest of the systems. Therefore, as expected, the absence of the N-terminus has no effect on the community structure of α_3 and ligand in the DM systems.

Table 3.7 Fraction of instances pairs of structural segments co-inhabit a community in truncated (Δ) PDZ3.*

System label	C-term and ligand				C-term and α_3				C-term and α_2			
	$\Omega = 3$	$\Omega = 4$	$\Omega = 5$	$\Omega = 6$	$\Omega = 3$	$\Omega = 4$	$\Omega = 5$	$\Omega = 6$	$\Omega = 3$	$\Omega = 4$	$\Omega = 5$	$\Omega = 6$
WT^{Δ}_{L1}	0.20 ± 0.06	0.18 ± 0.05	0.18 ± 0.05	0.16 ± 0.05	**0.98 ± 0.02**	**0.98 ± 0.02**	**0.98 ± 0.02**	**0.98 ± 0.02**	0.08 ± 0.03	0.06 ± 0.03	0.04 ± 0.03	0.02 ± 0.02
WT^{Δ}_{L2}	0.02 ± 0.02	0.02 ± 0.02	0.02 ± 0.02	0.02 ± 0.02	**1.00 ± 0.01**	**1.00 ± 0.01**	**1.00 ± 0.01**	**1.00 ± 0.01**	0	0	0	0
$G330T^{\Delta}_{L1}$	0.22 ± 0.06	0.22 ± 0.06	0.20 ± 0.06	0.18 ± 0.05	**1.00 ± 0.01**	**1.00 ± 0.01**	**1.00 ± 0.01**	**1.00 ± 0.01**	0.16 ± 0.05	0.06 ± 0.03	0.06 ± 0.03	0.04 ± 0.03
$G330T^{\Delta}_{L2}$	0.26 ± 0.06	0.22 ± 0.06	0.22 ± 0.06	0.16 ± 0.05	**1.00 ± 0.01**	**1.00 ± 0.01**	**1.00 ± 0.01**	**1.00 ± 0.01**	0.02 ± 0.02	0	0	0
$H372A^{\Delta}_{L1}$	0.16 ± 0.05	0.16 ± 0.05	0.10 ± 0.04	0.06 ± 0.03	**0.98 ± 0.02**	**0.98 ± 0.02**	**0.98 ± 0.02**	**0.98 ± 0.02**	0.06 ± 0.03	0.04 ± 0.03	0.02 ± 0.02	0.02 ± 0.02
$H372A^{\Delta}_{L2}$	0.54 ± 0.07	0.52 ± 0.07	0.38 ± 0.07	0.36 ± 0.07	**0.98 ± 0.02**	**0.98 ± 0.02**	**0.98 ± 0.02**	**0.98 ± 0.02**	0.32 ± 0.07	0.26 ± 0.06	0.22 ± 0.06	0.18 ± 0.05
DM^{Δ}_{L1}	0.16 ± 0.05	0.16 ± 0.05	0.14 ± 0.05	0.10 ± 0.04	**0.96 ± 0.03**	**0.96 ± 0.03**	**0.96 ± 0.03**	**0.96 ± 0.03**	0.10 ± 0.04	0.08 ± 0.04	0.06 ± 0.03	0.02 ± 0.02
DM^{Δ}_{L2}	0.16 ± 0.05	0.14 ± 0.05	0.08 ± 0.04	0.06 ± 0.03	**0.96 ± 0.03**	**0.96 ± 0.03**	**0.96 ± 0.03**	**0.96 ± 0.03**	0.14 ± 0.05	0.10 ± 0.02	0.02 ± 0.02	0.02 ± 0.02
System label	α_3 and ligand				α_3 and α_2				α_2 and ligand			
	$\Omega = 3$	$\Omega = 4$	$\Omega = 5$	$\Omega = 6$	$\Omega = 3$	$\Omega = 4$	$\Omega = 5$	$\Omega = 6$	$\Omega = 3$	$\Omega = 4$	$\Omega = 5$	$\Omega = 6$
WT^{Δ}_{L1}	**0.96 ± 0.03**	**0.86 ± 0.05**	**0.84 ± 0.05**	**0.82 ± 0.05**	**0.82 ± 0.05**	0.62 ± 0.07	0.44 ± 0.07	0.24 ± 0.06	**0.92 ± 0.04**	**0.76 ± 0.06**	0.68 ± 0.07	0.60 ± 0.07
WT^{Δ}_{L2}	**0.72 ± 0.06**	0.46 ± 0.07	0.40 ± 0.07	0.36 ± 0.07	0.66 ± 0.07	0.44 ± 0.07	0.30 ± 0.07	0.22 ± 0.06	**0.98 ± 0.02**	**0.94 ± 0.03**	**0.90 ± 0.04**	**0.86 ± 0.05**
$G330T^{\Delta}_{L1}$	**0.92 ± 0.04**	**0.82 ± 0.05**	0.74 ± 0.06	0.68 ± 0.07	**0.86 ± 0.05**	0.46 ± 0.07	0.18 ± 0.05	0.12 ± 0.05	**1.00 ± 0.01**	**0.80 ± 0.06**	**0.70 ± 0.07**	0.66 ± 0.07
$G330T^{\Delta}_{L2}$	**0.98 ± 0.02**	**0.82 ± 0.05**	**0.80 ± 0.06**	0.78 ± 0.06	**0.90 ± 0.04**	**0.72 ± 0.06**	0.52 ± 0.07	0.38 ± 0.07	**0.98 ± 0.02**	**0.94 ± 0.03**	**0.92 ± 0.04**	**0.92 ± 0.04**
$H372A^{\Delta}_{L1}$	**0.96 ± 0.03**	**0.94 ± 0.03**	**0.90 ± 0.04**	**0.86 ± 0.05**	**0.90 ± 0.04**	**0.80 ± 0.06**	0.64 ± 0.07	0.56 ± 0.07	**1.00 ± 0.01**	**0.92 ± 0.04**	**0.90 ± 0.04**	**0.88 ± 0.05**
$H372A^{\Delta}_{L2}$	**1.00 ± 0.01**	**0.94 ± 0.03**	**0.92 ± 0.04**	**0.90 ± 0.04**	**0.96 ± 0.03**	**0.86 ± 0.05**	**0.78 ± 0.06**	0.66 ± 0.07	**1.00 ± 0.01**	**0.94 ± 0.03**	**0.90 ± 0.04**	**0.86 ± 0.05**
DM^{Δ}_{L1}	**0.88 ± 0.05**	**0.78 ± 0.06**	**0.74 ± 0.06**	**0.72 ± 0.06**	**0.84 ± 0.05**	**0.80 ± 0.06**	0.64 ± 0.07	0.50 ± 0.07	**1.00 ± 0.01**	**0.98 ± 0.02**	**0.96 ± 0.03**	**0.96 ± 0.03**
DM^{Δ}_{L2}	**0.88 ± 0.05**	0.68 ± 0.07	0.56 ± 0.07	0.52 ± 0.07	**0.84 ± 0.05**	0.68 ± 0.07	0.46 ± 0.07	0.38 ± 0.07	**1.00 ± 0.01**	**1.00 ± 0.01**	**1.00 ± 0.01**	**0.98 ± 0.02**

*values greater than 0.7 shown in bold; those differing from the full-length variant value (Table 3.6) by more than the sum of the error bars colored red.

3.3.6 Major Communities, Dedicated Membership and Ubiquitous Residues

Our discussion so far has utilized community sharing of structural units whereby if at least one residue from each element appears in the same community, they are listed in the co-occupancy fractions. To better explain the dynamical shifts in shared communities, we have superposed the separation into communities as a visualization on the protein structures as explained under methods. In Figure 3.17 we observe that all the complexes have split into three main regions. One major community organizes around the binding site, including most of the ligand and the α_2 helix; this community is predominantly green. A second one is dominated by the α_3 helix colored blue. A third community, colored red, includes the β_1 strand and the surrounding loops as well as the residue at position 0 of the ligand (residue 9 in Figure 1.3). There are stark differences in the behavior of some other regions, however.

For example, in the functional complexes WT_{L1} and $G330T_{L1}$, the regions are well separated from each other, consisting of "pure" RGB colors except for the ligand and some parts in the central β sheet. In addition, the first few residues of the N-terminus are always separated out from the rest of the protein, indicated by their black color. In these complexes most of the N-terminus communicates with the α_3 helix throughout the trajectory, together forming the blue region which dominates the dynamics in this variant. The shades residues take indicate they share communities with the red, green, and blue regions proportionately. In particular, except at position 0 which is red, the ligand has the shade of teal with green:blue ratio of 2, i.e., it co-inhabits the blue region a third of the time. Note that the ligand forming an ingroup with the blue region is a must for favorable complexes; e.g., in the unfavorable $H372A_{L1}$, the ligand is pure green hence lacking a dynamical unification with the whole protein, although all other features of this complex is similar to WT_{L1} and $G330T_{L1}$.

In $G330T_{L2}$ and $H372A_{L2}$, though node BC is stable throughout the MD trajectory, community compositions show that the underlying interactions change over time. In particular, the groups containing the α_2/α_3 helices are blended, the teal color of the former

signifying 1:1 green:blue ratio, including the ligand. In this case, the whole N-terminus groups with the C-terminus and the α_3 helix. However, the α_3 helix itself is not a separate group but blends with the ligand/α_2. This high grouping of the core with the N/C termini and the α_3 helix is proposed to reinforce the high binding affinity of these variants.

The DM_{L2} complex, which is the one with the highest binding affinity to its ligand,[17] displays a further property of community sharing. While the average number of hydrogen bonds between the ligand and PDZ3 increased from ~3.2 to 3.9 in the double mutants[23] due to the decreased overall size of these variants, the central (green) region still communicates a great deal with the blue part. Unique to this variant is the behavior of the N-terminus which shares its time partnering with the green and red regions reinforcing the binding by interacting with the β_1 strand with some of its (orange colored) residues in addition to its α_3 interactions.

Finally, it is clear from the coloring of the unfavorable WT_{L2} and DM_{L1} complexes, when the majority of the N-terminus does not co-inhabit communities with the body of the protein, favorable binding does not take place.

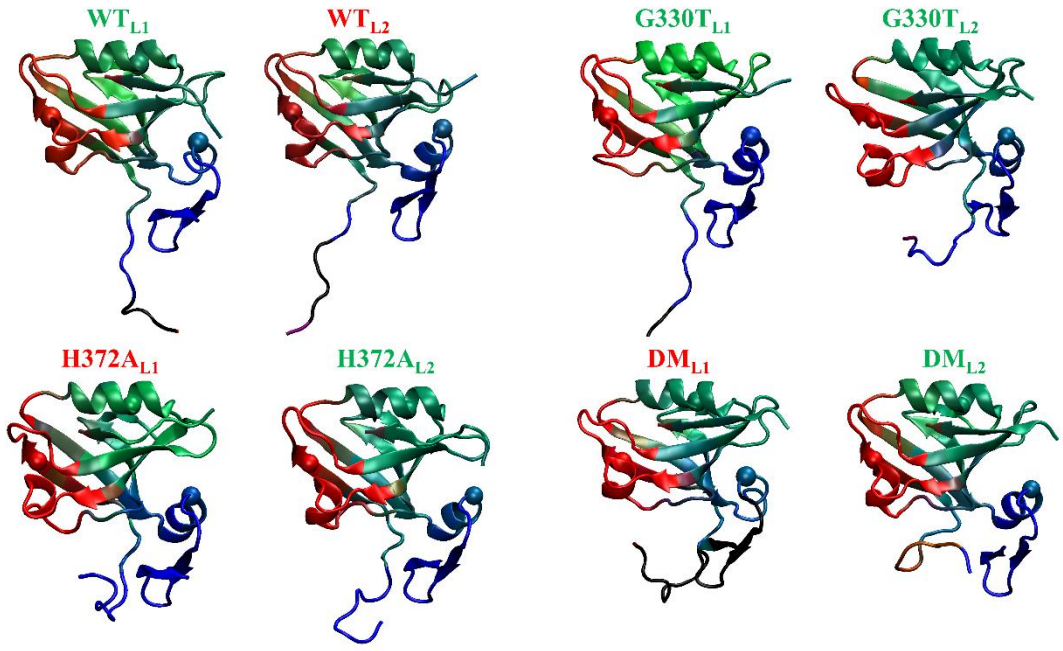

Figure 3.17 Visuals of dynamical community composition for selected variants, superimposed on the average structure for each variant. Community composition for each residue at $\Omega = 4$ is accumulated throughout the trajectories. Residues predominantly sharing the same community with I316, A375 and F400 (as ball representation) are in red, green and blue (RGB), respectively. Dynamically changing community neighbors relative to these reference residues are displayed as a mixture of RGB colors. The red community is separated out from the rest in all variants. The blue community carrying the α_3 helix dominates the green community carrying α_2 and the ligand since even A375 residing at the center of the latter has blue components in H372A and DM variants (hence the tinted green color). Black residues never share a community with the main selections, always separating out into a separate community.

3.3.7 Evolution of PDZ3 Is Investigated by Node BC and Conservation Scores

Node BC and community composition analyses are applied on 240 ns long MD trajectory of WT_{L1}. First, conservation score of each residue is calculated by using Consurf.[62] Inside the Consurf pipeline, multiple sequence alignments (MSA) of 150 sequence homologs of PDZ3 are recorded for further analyses. Consurf scores show that N-terminus, α_3 and C-terminus have lower conservation, compared to the protein core [313-392] (Figure 3.18a). Positions

containing a large number of gaps in PDZ3 MSA leading to an insufficient sample size are signified as low confidence values in Consurf pipeline. Hence, the confidence interval (0.75) indicates that MSA has inadequate data for these terminal structural segments.

Further, by using the recorded MSA, statistical coupling analysis (SCA) is conducted by utilizing the pipeline in a previous study.[84] First, to eliminate the sequences with gaps for more precision, pipeline reduces the MSA to 126 sequences, then the SCA calculation converges on the 80 residue-long protein region [313-392] for further calculation of amino acid positions (D_i) (Figure 3.18a). This reduction of MSA sequences and amino-acid positions to fine-tune the conservation calculation stems from a higher confidence bound (0.95) belonging to the SCA scheme. These findings are also in agreement with the Mclaughlin's study,[16] where sectors are defined only for the 83 residue long core. Overall, the results display that the evolutionary information is insufficient in terms of MSA for PDZ3. Hence, the Pearson's R for Consurf and D_i (SCA) scores for the protein core is 0.88 indicating the high similarity between the two measures; on the other hand, correlation of each of these scores for node BC is the same with a value of 0.57 meaning that node BC does not approximate the evolutionary conservation of residues. This is an expected outcome since the evolution of a sequence cannot be constrained to only binding.[85]

Residue by residue investigation of the average community composition ($\Omega = 4$) is compared with the SCA weighted correlation matrix for the protein core (Figure 3.18b, c). In Figure 3.18b, there are 7-8 observable communities with different sizes indicating that, even in the immobile protein core, a modular structure occurs. On the other hand, SCA which indicates coevolving regions shows the effect of residues between 320-340 (Figure 3.18c). Although, there is no similarity in terms of modularity between community sharing and coevolving residues (Figure 3.18b, c); the results in SCA weighted matrix (Figure 3.18c) are in agreement with Lockless' findings.[15]

An insufficient amount of evolutionary information precludes inferring conclusions for the regions of interest for us in the current study. One strength of using network-based measures is that they work on a single structure and using them on a series of plausible conformations obtained from sufficiently long MD simulations adds to the wealth of information one can obtain from coarse grained designations of proteins.

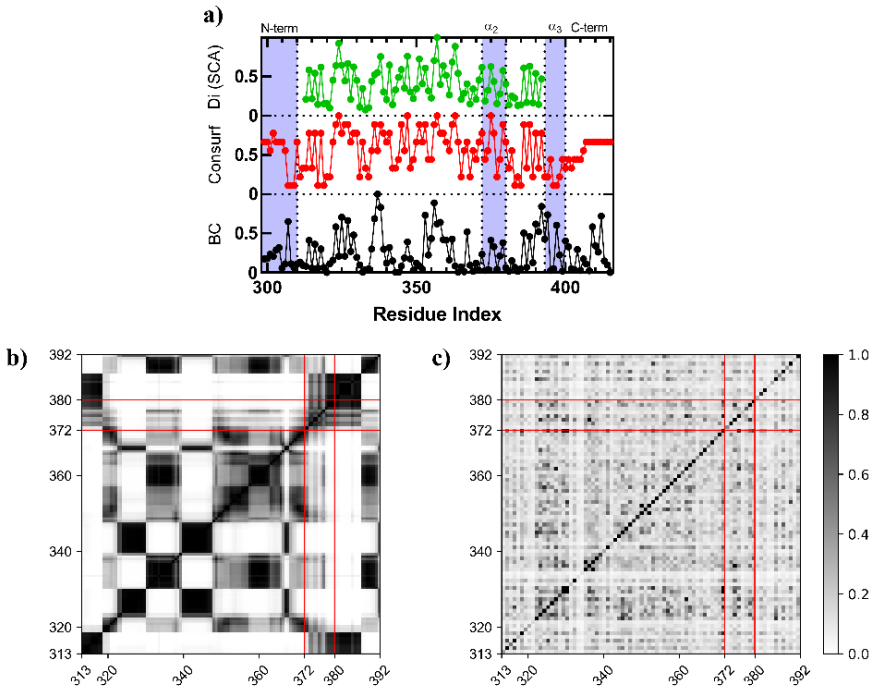

Figure 3.18 **a** Node BC (in black), Consurf (in red) and D_i (SCA, in green) results are calculated for WT_{L1}. The values normalized so that maxima are equal to 1 for better comparison. Structural segments are signified by purple coloring and labelled on top of corresponding residue range. **b** Residue by residue average community composition for Ω = 4 in WT-PDZ3. Residue index range is arranged according to sectors calculation. **c** SCA weighted correlation matrix. Red lines signify the residue range of α_2.

4. CONCLUSIONS

The biological function of PDZ3 is measured by binding affinity experiments,[16, 17] and functional complexes such as WT_{L1}, $G330T_{L1}$, $G330T_{L2}$, $H372A_{L2}$ and DM_{L2} are revealed

pointing to the allosteric character of this domain. However, the underlying mechanism deciding the binding fate cannot be clarified by the thermodynamic measurements alone. Since the conformational differences between these forms are minimal, a dynamical assessment is needed to further interpret the allosteric mechanism involved in communication. In a recent study on PDZ2, it was discussed that the mean structures do not necessarily differ between favorable and unfavorable complexes, whereby having different fraction of substates leads to ligand binding.[86] Similarly, studies on PDZ3 have shown that while the conformational change is not essential for allostery,[87, 88] N/C termini and α_3 play an important role on its function;[5, 7-10, 23, 24] e.g., α_3 exhibits significantly different impact on ligand-binding within the temperature limits of 10 - 40º C.[11] The affinity experiments demonstrate that the binding free energy difference between the functional and dysfunctional PDZ3 complexes is around ~2.5 kcal/mol,[16] whereby overall energy contribution of a non-bonded interaction, including hydrogen bonds, is ~1.0 kcal/mol.[80, 89-91] Therefore, assessing conformational changes resulting from these energy differences is challenging.

We perform MD and FEP simulations to show the energetic, conformational and dynamical bases for the ligand specificity of the mutations in PDZ domains. One novelty of this study is that we suggest a reverse approach to use FEP calculations to validate the adequacy of ensembles collected in classical MD simulations (Figure 3.2). Then, we propose a simple model to predict binding free energies to distinguish the ligand selected by various PDZ3 mutants from classical MD simulations of affordable length (Figure 3.8).

Another useful feature of using FEP calculations has been to discuss properties of the system that could not have been deduced otherwise. These are, (i) changes in the tautomeric states of key residues in apo forms; (ii) the cost of the individual mutations in the ligand free and ligand bound forms. We find the computational $\Delta\Delta G$ values are in agreement with the experimentally determined adaptive pathway, and the *in silico* $\Delta\Delta\Delta G$ results corroborate the experiments.[16, 23] Moreover, the multiple MD simulations uncover intriguing fluctuation patterns of the glutamic acid rich N-terminus region (Figure 3.4) which provide the -4 net charge of the protein. Accordingly, to predict binding free energies from MD trajectories, the

proposed model accounts for this electrostatic effect along with the change in the hydrogen bond contributions due to different mutations (Table 3.1). Our model fits well the results of binding experiments; furthermore, its regression parameters are in the biophysically relevant range (Equation 3.8).

Our simple model demonstrates that the charged residues of the N-terminus have a decisive role in mutation – binding partner matches for functional activity, even though the main contributions to binding free energy come from the hydrogen bonds formed in the binding pocket (Table 3.3). Namely, (i) the H372A mutant prefers L_2 because of the Born solvation energy of the N-terminus, which is negligible for the L_1 bound complex; and (ii) while the G330T prefers to bind both L_1 and L_2, the preference to L_1 is strengthened by the electrostatic contributions of the complex, similar to the choice of L_1 in lieu of L_2 for the wild type.

These observations lead to the following fundamental question: Would mutations select the same ligands functionally, if we were to intervene with the proposed communication between the N-terminus and the rest of the protein? To answer, we simply remove the 12 N-terminus residues from all of the ligand-bound proteins, repeat the calculations (Figure 3.9), and monitor the changes in the binding energies. We find that the costs of the G→T transition in the presence of L_1 in cycle ① and the binding pocket H→A transition in the presence of L_2 in cycle ④ are 3.3 and 2.6 kcal/mol larger, respectively. This finding clearly summarizes the moderating role of the N-terminus for selecting the functional ligand for the PDZ3 domain. Our further analysis via the proposed simple binding free energy model indicates that the FEP differences in the truncated PDZ3 are mainly due to modified hydrogen bond interactions in the binding pocket (Table 3.1).

Nevertheless, we develop a methodology to decipher the hidden states governing favorable binding and specificity. We set out to show that the slight variations occurring during the dynamical motions provide information on the functionality of PDZ3 complexes, and the community composition of underlying states dictates the allosteric communication without

undertaking significant conformational changes.[87, 88] For this purpose, MD-simulated trajectories of PDZ3 complexes are investigated by using community detection tools from graph theory.[38, 39] To understand the modularity of PDZ3, the communication between the N/C termini, α_2, α_3 and the ligand is assessed.[7, 9, 23, 25] Although, node centrality measures are informative to the extent of pinpointing residues whose centrality are shifted depending on the variant studied (Figure 3.11), our community composition analysis is more sensitive. We find that PDZ3 complex variants have diverse community configurations, and the fraction of these changes modulates binding preferences.

For a successful binding event, the following conditions must be met: (i) The N-terminus must share a community with the C-terminus and α_3 (blue communities in Figure 3.17); (ii) ligand must not only be a part of the binding site community (green communities in Figure 3.17), but it must also share communities with the N/C termini and α_3 in (i). Moreover, if the N-terminus co-inhabits community with the β_1 helix, the ligand specificity is further reinforced (orange region in Figure 3.17, DM_{L2}). To the best of our knowledge, while there are several studies that use communities in proteins to determine a collection of residues that act in concert,[92-96] this is the first study where the dynamical exchanges between communities have been woven into a narrative for protein functionality.

In sum, while the N-terminus confers the specificity, C-terminus and α_3 are the essential vehicles for the formation of the PDZ complex. In fact, α_3 acts as a hub for the whole protein by sustaining the communication with all structural segments. We propose the method developed in this work as a general methodology to study protein structures where the mechanism of action is not readily disclosed by conformational changes. For the particular case of PDZ domains, the behavior of communities put forth in this study is due to residue network members having a high number of redundant edges and common neighbors. These redundancies translate into highly sensitive shifts in allosteric networks depending on the external conditions imposed, leading to the measured binding affinity differences. It is apparent that point mutations perturb a plethora of interactions that have entropic[97, 98] and electrostatic[9, 26] origins. Our coarse-grained approach in this work along with our previous

all-atom approach[23] paves the way for further analysis of entropy-enthalpy compensation. Moreover, the supertertiary structures of PDZ domains have recently gained closer attention, e.g., via the experimental studies on PDZ1-PDZ2 tandem[2] and PDZ3-SH3-GK tandem.[3] In particular, it was shown that the binding affinities to peptides differ substantially depending on PDZ3 being isolated or as part of a supertertiary structure and the shifts in the allosteric network of PDZ3 was identified as the reason for the differences.[3] Our results indicate that the arrangement of the allosteric network will be sensitive to the external conditions imposed, which might include the packing arrangements assumed by the supertertiary structure. PDZ3 domain operates as a bridge with its N-terminus bound to PDZ tandem, while the C-terminus is attached to SH3-GK. The ~50 residue-long linker that precedes the N-terminus of PDZ3 may be presumed not to constrain the mobility of this highly charged region while it may be a tool to communicate signals. It is possible that the supertertiary structure controlled through the termini of PDZ3 might play a role in function, a hypothesis which may be tested in future studies, with the emergence of new PDB structures for the supramolecular assembly.

Finally, we note that owing to their centrality in protein-protein interaction pathways,[99-101] PDZ domains may be interesting drug targets against resistance conferring mutants in cancer. Our results about the terminal regions, such as N/C termini and α_3 might act as basis for more specific targets, considering the inadequate information belonging to PDZ3 in the literature.

5. EPILOGUE AND FUTURE WORK

The community composition analysis display that there are underlying states of conformational dynamics belonging to a protein, and the composition changes, even though size of the communities stays same through the MD trajectory. However, a physics-based interpretation of this behavior is yet to be provided. To this end, we compute common neighbors between the structural segments, and results show that N-terminus and ligand do not share neighbors. We further focus on shortest path lengths and find that N-

terminus/ligand have at least three edges-long paths in between, taking nodes of two segments as sources and targets. Similar to the shortest path length scrutiny, communicability[59] analyses, where we assess 'walks', do not lead to any differentiating results. On the other hand, number of different shortest paths between the protein segments might be informative about the occurrence of community structures, since it is possible to detect the nature of redundancies that makes communities close-knit by assessing this measure.

There are different ways to construct graphs from protein structures/trajectories. First, Laplacian matrix (Γ)[102] is computed by using the inverse of cross-correlation matrix from the MD trajectories. Also, it is possible to compute the pairwise distance of all residues and construct one graph per MD simulation.[39] However, sampled conformations may be misleading in cross-correlation matrices,[49] and an average graph loses the information about the underlying slight changes. In this thesis, we study time evolution of graphs by using a physical cut-off stemming from the radial distribution function, so that these graphs could be assessed by employing 'temporal network' theory and measures[103] in a future study. Additionally, deep learning pipelines are available to construct networks from correlated motions (such as skeletal motions and Kuramoto oscillators)[104, 105] which might be feasible to use for future studies.

Deep mutational scan or deep mutagenesis (DMS) is a method to investigate change of binding affinity for single and double mutations of a protein, and this method explains biological function of a protein experimentally. There are studies about inferring the three dimensional structures of proteins by using DMS,[106, 107] in addition to other investigations about high-order mutations[18] and analyses/perspectives about DMS results.[108-111] Nonetheless, DMS data of PDZ3 that uses a well-known experimental pipeline[16, 17, 25] (Rama Ranganathan scheme) has been published recently.[112] In the light of these studies, double mutation matrix of a protein might be utilized to construct a graph, since it is possible to decipher a correlation like 'coupling' information from the DMS matrix. Further, single and double mutations matrices may be employed as weights for nodes and/or edges. In sum, it is

essential to understand the underlying biophysics of binding affinity changes that are manifested by mutations for future work.

In this thesis, analyses of MD simulations are based on both external (RMSD, RMSF and cross-correlation) and internal (SASA, hydrogen-bonds and network analyses) degrees of freedom, where a caveat of the former is ‘superposition’ of a dynamic protein trajectory. In the process of superposition, slight changes that lead functional shifts diminish, especially for a highly mobile protein like PDZ3. To tackle this sampling problem, analysis pipelines have been established by using collective variables, machine/deep learning and reaction coordinates in the literature,[113-120] and these methods may be used for future studies. Additionally, the problem of timescale, where we investigate phenomena on the order of milliseconds[121] by using nanoseconds-long simulations might be solved by using deep learning-based simulation methods.[122-124]

REFERENCES

(1) Liu, X.; Fuentes, E. J., Emerging Themes in PDZ Domain Signaling: Structure, Function, and Inhibition. Elsevier Ltd: 2019; Vol. 343, pp 129-218.

(2) Rodzli, N. A.; Lockhart-Cairns, M. P.; Levy, C. W.; Chipperfield, J.; Bird, L.; Baldock, C.; Prince, S. M., The Dual PDZ Domain from Postsynaptic Density Protein 95 Forms a Scaffold with Peptide Ligand. *Biophys J* **2020,** *119*, 667-689.

(3) Laursen, L.; Kliche, J.; Gianni, S.; Jemth, P., Supertertiary Protein Structure Affects an Allosteric Network. *Proc Natl Acad Sci U S A* **2020,** *117*, 24294-24304.

(4) Mostarda, S.; Gfeller, D.; Rao, F., Beyond the Binding Site: The Role of the β2 – β3 Loop and Extra-Domain Structures in PDZ Domains. *PLoS Comput Biol* **2012,** *8*, e1002429.

(5) Gerek, Z. N.; Ozkan, S. B., Change in Allosteric Network Affects Binding Affinities of PDZ Domains: Analysis through Perturbation Response Scanning. *PLoS Comput Biol* **2011,** *7*, e1002154.

(6) Morra, G.; Genoni, A.; Colombo, G., Mechanisms of Differential Allosteric Modulation in Homologous Proteins: Insights from the Analysis of Internal Dynamics and Energetics of PDZ Domains. *J Chem Theory Comput* **2014,** *10*, 5677-5689.

(7) Camara-Artigas, A.; Murciano-Calles, J.; Martinez, J. C., Conformational Changes in the Third PDZ Domain of the Neuronal Postsynaptic Density Protein 95. *Acta Crystallogr D Struct Biol* **2019,** *75*, 381-391.

(8) Gautier, C.; Visconti, L.; Jemth, P.; Gianni, S., Addressing the Role of the α-helical Extension in the Folding of the Third PDZ Domain From PSD-95. *Sci Rep* **2017,** *7*, 12593.

(9) Kumawat, A.; Chakrabarty, S., Hidden Electrostatic Basis of Dynamic Allostery in a PDZ Domain. *Proc Natl Acad Sci U S A* **2017,** *114*, E5825-E5834.

(10) Petit, C. M.; Zhang, J.; Sapienza, P. J.; Fuentes, E. J.; Lee, A. L., Hidden Dynamic Allostery in a PDZ Domain. *Proc Natl Acad Sci U S A* **2009,** *106*, 18249-18254.

(11) Bozovic, O.; Jankovic, B.; Hamm, P., Sensing the Allosteric Force. *Nat Commun* **2020,** *11*, 5841.

(12) Doyle, D. A.; Lee, A.; Lewis, J.; Kim, E.; Sheng, M.; MacKinnon, R., Crystal Structures of a Complexed and Peptide-Free Membrane Protein–Binding Domain: Molecular Basis of Peptide Recognition by PDZ. *Cell* **1996,** *85*, 1067-1076.

(13) Niethammer, M.; Valtschanoff, J. G.; Kapoor, T. M.; Allison, D. W.; Weinberg, R. J.; Craig, A. M.; Sheng, M., CRIPT, a Novel Postsynaptic Protein that Binds to the Third PDZ Domain of PSD-95/SAP90. *Neuron* **1998,** *20*, 693-707.

(14) Tonikian, R., et al., A Specificity Map for the PDZ Domain Family. *PLoS Biology* **2008,** *6*, e239.

(15) Lockless, S. W.; Ranganathan, R., Evolutionarily Conserved Pathways of Energetic Connectivity in Protein Families. *Science* **1999,** *286*, 295-299.

(16) McLaughlin, R. N., Jr.; Poelwijk, F. J.; Raman, A.; Gosal, W. S.; Ranganathan, R., The Spatial Architecture of Protein Function and Adaptation. *Nature* **2012,** *491*, 138-142.

(17) Raman, A. S.; White, K. I.; Ranganathan, R., Origins of Allostery and Evolvability in Proteins: A Case Study. *Cell* **2016,** *166*, 468-480.

(18) Poelwijk, F. J.; Socolich, M.; Ranganathan, R., Learning the Pattern of Epistasis Linking Genotype and Phenotype in a Protein. *Nat Commun* **2019,** *10*, 4213.

(19) Stiffler, M. A.; Chen, J. R.; Grantcharova, V. P.; Lei, Y.; Fuchs, D.; Allen, J. E.; Zaslavskaia, L. A.; MacBeath, G., PDZ Domain Binding Selectivity is Optimized Across the Mouse Proteome. *Science* **2007,** *317*, 364-369.
(20) Ozbaykal, G.; Atilgan, A. R.; Atilgan, C., In Silico Mutational Studies of Hsp70 Disclose Sites With Distinct Functional Attributes. *Proteins* **2015,** *83*, 2077-2090.
(21) Gumbart, J. C.; Roux, B.; Chipot, C., Standard binding free energies from computer simulations: What is the best strategy? *J Chem Theory Comput* **2013,** *9*, 794-802.
(22) Dixit, S. B.; Chipot, C., Can Absolute Free Energies of Association Be Estimated from Molecular Mechanical Simulations? The Biotin−Streptavidin System Revisited. *J Phys Chem A* **2001,** *105*, 9795-9799.
(23) Guclu, T. F.; Kocatug, N.; Atilgan, A. R.; Atilgan, C., N-Terminus of the Third PDZ Domain of PSD-95 Orchestrates Allosteric Communication for Selective Ligand Binding. *J Chem Inf Model* **2021,** *61*, 347-357.
(24) Liu, J.; Nussinov, R., Energetic Redistribution in Allostery to Execute Protein Function. *Proc Natl Acad Sci U S A* **2017,** *114*, 7480-7482.
(25) Salinas, V. H.; Ranganathan, R., Coevolution-Based Inference of Amino Acid Interactions Underlying Protein Function. *eLife* **2018,** *7*, 1-20.
(26) Kumawat, A.; Chakrabarty, S., Protonation-Induced Dynamic Allostery in PDZ Domain: Evidence of Perturbation-Independent Universal Response Network. *J Phys Chem Lett* **2020,** *11*, 9026-9031.
(27) Atilgan, A. R.; Akan, P.; Baysal, C., Small-world communication of residues and significance for protein dynamics. *Biophys J* **2004,** *86*, 85-91.
(28) Turgut, D.; Atilgan, A. R.; Atilgan, C., Assortative Mixing in Close-Packed Spatial Networks. *PLoS One* **2010,** *5*, e15551.
(29) Atilgan, A. R.; Turgut, D.; Atilgan, C., Screened Nonbonded Interactions in Native Proteins Manipulate Optimal Paths for Robust Residue Communication. *Biophys J* **2007,** *92*, 3052-3062.
(30) Brinda, K. V.; Vishveshwara, S., Oligomeric Protein Structure Networks: Insights Into Protein-Protein Interactions. *BMC Bioinformatics* **2005,** *6*, 296.
(31) Vendruscolo, M.; Dokholyan, N. V.; Paci, E.; Karplus, M., Small-World View of the Amino Acids That Play a Key Role in Protein Folding. *Phys Rev E Stat Nonlin Soft Matter Phys* **2002,** *65*, 061910.
(32) Greene, L. H.; Higman, V. A., Uncovering Network Systems Within Protein Structures. *J Mol Biol* **2003,** *334*, 781-791.
(33) del Sol, A.; Fujihashi, H.; Amoros, D.; Nussinov, R., Residue Centrality, Functionally Important Residues, and Active Site Shape: Analysis of Enzyme and Non-Enzyme Families. *Protein Sci* **2006,** *15*, 2120-2128.
(34) Sheik Amamuddy, O.; Verkhivker, G. M.; Tastan Bishop, O., Impact of Early Pandemic Stage Mutations on Molecular Dynamics of SARS-CoV-2 M(pro). *J Chem Inf Model* **2020,** *60*, 5080-5102.
(35) Penkler, D. L.; Atilgan, C.; Tastan Bishop, O., Allosteric Modulation of Human Hsp90α Conformational Dynamics. *J Chem Inf Model* **2018,** *58*, 383-404.
(36) Atilgan, C., Chapter Two - Computational Methods for Efficient Sampling of Protein Landscapes and Disclosing Allosteric Regions. In *Advances in Protein Chemistry and Structural Biology*, Karabencheva-Christova, T. G.; Christov, C. Z., Eds. Academic Press: 2018; Vol. 113, pp 33-63.

(37) Girvan, M.; Newman, M. E., Community Structure in Social and Biological Networks. *Proc Natl Acad Sci U S A* **2002,** *99*, 7821-7826.
(38) Bowerman, S.; Wereszczynski, J., Detecting Allosteric Networks Using Molecular Dynamics Simulation. 1 ed.; Elsevier Inc.: 2016; Vol. 578, pp 429-447.
(39) Melo, M. C. R.; Bernardi, R. C.; de la Fuente-Nunez, C.; Luthey-Schulten, Z., Generalized Correlation-Based Dynamical Network Analysis: a New High-Performance Approach for Identifying Allosteric Communications in Molecular Dynamics Trajectories. *J Chem Phys* **2020,** *153*, 134104.
(40) Brown, D. K.; Penkler, D. L.; Sheik Amamuddy, O.; Ross, C.; Atilgan, A. R.; Atilgan, C.; Tastan Bishop, O., MD-TASK: A Software Suite for Analyzing Molecular Dynamics Trajectories. *Bioinformatics* **2017,** *33*, 2768-2771.
(41) Guclu, T. F.; Atilgan, A. R.; Atilgan, C., Dynamic Community Composition Unravels Allosteric Communication in PDZ3. *J Phys Chem B* **2021,** *125*, 2266-2276.
(42) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E., The Protein Data Bank. *Nucleic Acids Res* **2000,** *28*, 235-42.
(43) Waterhouse, A., et al., SWISS-MODEL: Homology Modelling of Protein Structures and Complexes. *Nucleic Acids Res* **2018,** *46*, W296-W303.
(44) Phillips, J. C.; Braun, R.; Wang, W.; Gumbart, J.; Tajkhorshid, E.; Villa, E.; Chipot, C.; Skeel, R. D.; Kale, L.; Schulten, K., Scalable Molecular Dynamics With NAMD. *J Comput Chem* **2005,** *26*, 1781-1802.
(45) Kalé, L.; Skeel, R.; Bhandarkar, M.; Brunner, R.; Gursoy, A.; Krawetz, N.; Phillips, J.; Shinozaki, A.; Varadarajan, K.; Schulten, K., NAMD2: Greater Scalability for Parallel Molecular Dynamics. *J Comput Phys* **1999,** *151*, 283-312.
(46) Best, R. B.; Zhu, X.; Shim, J.; Lopes, P. E.; Mittal, J.; Feig, M.; Mackerell, A. D., Jr., Optimization of the Additive CHARMM All-Atom Protein Force Field Targeting Improved Sampling of the Backbone ϕ, ψ and Side-Chain χ_1 and χ_2 Dihedral Angles. *J Chem Theory Comput* **2012,** *8*, 3257-3273.
(47) Humphrey, W.; Dalke, A.; Schulten, K., VMD: visual molecular dynamics. *J Mol Graph* **1996,** *14*, 33-8, 27-8.
(48) Darden, T.; Perera, L.; Li, L.; Pedersen, L., New tricks for modelers from the crystallography toolkit: the particle mesh Ewald algorithm and its use in nucleic acid simulations. *Structure* **1999,** *7*, R55-R60.
(49) Atilgan, C.; Okan, O. B.; Atilgan, A. R., Network-based models as tools hinting at nonevident protein functionality. *Annu Rev Biophys* **2012,** *41*, 205-25.
(50) Bakan, A.; Meireles, L. M.; Bahar, I., ProDy: protein dynamics inferred from theory and experiments. *Bioinformatics* **2011,** *27*, 1575-7.
(51) Lu, N.; Kofke, D. A.; Woolf, T. B., Improving the efficiency and reliability of free energy perturbation calculations using overlap sampling methods. *J Comput Chem* **2004,** *25*, 28-39.
(52) Zwanzig, R. W., High-Temperature Equation of State by a Perturbation Method. I. Nonpolar Gases. *J. Chem. Phys.* **1954,** *22*, 1420-1426.
(53) Liu, P.; Dehez, F.; Cai, W.; Chipot, C., A Toolkit for the Analysis of Free-Energy Perturbation Calculations. *J Chem Theory Comput* **2012,** *8*, 2606-16.
(54) Bennett, C. H., Efficient estimation of free energy differences from Monte Carlo data. *J. Comput. Phys.* **1976,** *22*, 245-268.

(55) Okan, O. B.; Turgut, D.; Atilgan, C.; Atilgan, A. R.; Ozisik, R., Could Network Structures Generated With Simple Rules Imposed on a Cubic Lattice Reproduce the Structural Descriptors of Globular Proteins? *bioRxiv* **2020**, 2020.10.01.321992.
(56) Freeman, L. C., A Set of Measures of Centrality Based on Betweenness. *Sociometry* **1977,** *40*, 35.
(57) Newman, M. E. J., A measure of betweenness centrality based on random walks. *Social Networks* **2005,** *27*, 39-54.
(58) Estrada, E.; Higham, D. J.; Hatano, N., Communicability betweenness in complex networks. *Physica A: Statistical Mechanics and its Applications* **2009,** *388*, 764-774.
(59) Estrada, E.; Hatano, N., Communicability in complex networks. *Phys Rev E Stat Nonlin Soft Matter Phys* **2008,** *77*, 036111.
(60) Brandes, U., A Faster Algorithm for Betweenness Centrality. *J Math Sociol* **2001,** *25*, 163-177.
(61) Hagberg, A. A.; Schult, D. A.; Swart, P. J. In *Exploring Network Structure, Dynamics, and Function using NetworkX*, Pasadena, CA USA, 2008; Varoquaux, G.; Vaught, T.; Millman, J., Eds. Pasadena, CA USA, 2008; pp 11-15.
(62) Ashkenazy, H.; Abadi, S.; Martz, E.; Chay, O.; Mayrose, I.; Pupko, T.; Ben-Tal, N., ConSurf 2016: an improved methodology to estimate and visualize evolutionary conservation in macromolecules. *Nucleic Acids Res* **2016,** *44*, W344-50.
(63) Tan, K. P.; Varadarajan, R.; Madhusudhan, M. S., DEPTH: a web server to compute depth and predict small-molecule binding cavities in proteins. *Nucleic Acids Res* **2011,** *39*, W242-8.
(64) Bai, Q.; Tan, S.; Xu, T.; Liu, H.; Huang, J.; Yao, X., MolAICal: a soft tool for 3D drug design of protein targets by artificial intelligence and classical algorithm. *Brief Bioinform* **2020**.
(65) El Khoury, L.; Santos-Martins, D.; Sasmal, S.; Eberhardt, J.; Bianco, G.; Ambrosio, F. A.; Solis-Vasquez, L.; Koch, A.; Forli, S.; Mobley, D. L., Comparison of affinity ranking using AutoDock-GPU and MM-GBSA scores for BACE-1 inhibitors in the D3R Grand Challenge 4. *J Comput Aided Mol Des* **2019,** *33*, 1011-1020.
(66) Chodera, J. D.; Mobley, D. L.; Shirts, M. R.; Dixon, R. W.; Branson, K.; Pande, V. S., Alchemical free energy methods for drug discovery: progress and challenges. *Curr Opin Struct Biol* **2011,** *21*, 150-60.
(67) Aleksandrov, A.; Proft, J.; Hinrichs, W.; Simonson, T., Protonation patterns in tetracycline:tet repressor recognition: simulations and experiments. *Chembiochem* **2007,** *8*, 675-85.
(68) Mondal, D.; Florian, J.; Warshel, A., Exploring the Effectiveness of Binding Free Energy Calculations. *J Phys Chem B* **2019,** *123*, 8910-8915.
(69) Chen, D.; Oezguen, N.; Urvil, P.; Ferguson, C.; Dann, S. M.; Savidge, T. C., Regulation of protein-ligand binding affinity by hydrogen bond pairing. *Sci Adv* **2016,** *2*, e1501240.
(70) Still, W. C.; Tempczyk, A.; Hawley, R. C.; Hendrickson, T., Semianalytical treatment of solvation for molecular mechanics and dynamics. *J Am Chem Soc* **1990,** *112*, 6127-6129.
(71) Warshel, A.; Sharma, P. K.; Kato, M.; Parson, W. W., Modeling electrostatic effects in proteins. *Biochim Biophys Acta* **2006,** *1764*, 1647-76.
(72) Baysal, C.; Meirovitch, H., Novel Procedure for Developing Implicit Solvation Models for Peptides and Proteins. *J Phys Chem B* **1997,** *101*, 7368-7370.

(73) Wesson, L.; Eisenberg, D., Atomic solvation parameters applied to molecular dynamics of proteins in solution. *Protein Sci* **1992,** *1*, 227-35.
(74) Smith, K. C.; Honig, B., Evaluation of the conformational free energies of loops in proteins. *Proteins* **1994,** *18*, 119-32.
(75) Williams, R. L.; Vila, J.; Perrot, G.; Scheraga, H. A., Empirical solvation models in the context of conformational energy searches: application to bovine pancreatic trypsin inhibitor. *Proteins* **1992,** *14*, 110-9.
(76) Kang, Y. K.; Nemethy, G.; Scheraga, H. A., Free energies of hydration of solute molecules. 1. Improvement of the hydration shell model by exact computations of overlapping volumes. *J Phys Chem* **1987,** *91*, 4105-4109.
(77) Stouten, P. F. W.; Frömmel, C.; Nakamura, H.; Sander, C., An Effective Solvation Term Based on Atomic Occupancies for Use in Protein Simulations. *Mol Simul* **1993,** *10*, 97-120.
(78) Vila, J.; Williams, R. L.; Vasquez, M.; Scheraga, H. A., Empirical solvation models can be used to differentiate native from near-native conformations of bovine pancreatic trypsin inhibitor. *Proteins* **1991,** *10*, 199-218.
(79) Ooi, T.; Oobatake, M.; Nemethy, G.; Scheraga, H. A., Accessible surface areas as a measure of the thermodynamic parameters of hydration of peptides. *Proc Natl Acad Sci U S A* **1987,** *84*, 3086-90.
(80) Sheu, S. Y.; Yang, D. Y.; Selzle, H. L.; Schlag, E. W., Energetics of Hydrogen Bonds in Peptides. *Proc Natl Acad Sci U S A* **2003,** *100*, 12683-12687.
(81) Williams, D. H.; Searle, M. S.; Mackay, J. P.; Gerhard, U.; Maplestone, R. A., Toward an estimation of binding constants in aqueous solution: studies of associations of vancomycin group antibiotics. *Proc Natl Acad Sci U S A* **1993,** *90*, 1172-8.
(82) Fersht, A. R.; Shi, J. P.; Knill-Jones, J.; Lowe, D. M.; Wilkinson, A. J.; Blow, D. M.; Brick, P.; Carter, P.; Waye, M. M.; Winter, G., Hydrogen bonding and biological specificity analysed by protein engineering. *Nature* **1985,** *314*, 235-8.
(83) Newman, M. E., Fast Algorithm for Detecting Community Structure in Networks. *Phys Rev E Stat Nonlin Soft Matter Phys* **2004,** *69*, 066133.
(84) Rivoire, O.; Reynolds, K. A.; Ranganathan, R., Evolution-Based Functional Decomposition of Proteins. *PLoS Comput Biol* **2016,** *12*, e1004817.
(85) Tesileanu, T.; Colwell, L. J.; Leibler, S., Protein sectors: statistical coupling analysis versus conservation. *PLoS Comput Biol* **2015,** *11*, e1004091.
(86) Bozovic, O.; Zanobini, C.; Gulzar, A.; Jankovic, B.; Buhrke, D.; Post, M.; Wolf, S.; Stock, G.; Hamm, P., Real-Time observation of Ligand-Induced Allosteric Transitions in a PDZ Domain. *Proc Natl Acad Sci U S A* **2020,** *117*, 26031-26039.
(87) Nussinov, R.; Tsai, C. J., Allostery Without a Conformational Change? Revisiting the Paradigm. *Curr Opin Struct Biol* **2015,** *30*, 17-24.
(88) Tsai, C. J.; del Sol, A.; Nussinov, R., Allostery: Absence of a Change in Shape Does Not Imply That Allostery Is Not at Play. *J Mol Biol* **2008,** *378*, 1-11.
(89) Li, J.; Wang, Y.; An, L.; Chen, J.; Yao, L., Direct Observation of CH/CH van der Waals Interactions in Proteins by NMR. *J Am Chem Soc* **2018,** *140*, 3194-3197.
(90) Bartlett, G. J.; Newberry, R. W.; VanVeller, B.; Raines, R. T.; Woolfson, D. N., Interplay of Hydrogen Bonds and n→π* Interactions in Proteins. *J Am Chem Soc* **2013,** *135*, 18682-18688.

(91) Baker, E. G.; Williams, C.; Hudson, K. L.; Bartlett, G. J.; Heal, J. W.; Porter Goff, K. L.; Sessions, R. B.; Crump, M. P.; Woolfson, D. N., Engineering Protein Stability With Atomic Precision in a Monomeric Miniprotein. *Nat Chem Biol* **2017,** *13*, 764-770.
(92) Stetz, G.; Astl, L.; Verkhivker, G. M., Exploring Mechanisms of Communication Switching in the Hsp90-Cdc37 Regulatory Complexes with Client Kinases through Allosteric Coupling of Phosphorylation Sites: Perturbation-Based Modeling and Hierarchical Community Analysis of Residue Interaction Networks. *J Chem Theory Comput* **2020,** *16*, 4706-4725.
(93) Mhashal, A. R.; Romero-Rivera, A.; Mydy, L. S.; Cristobal, J. R.; Gulick, A. M.; Richard, J. P.; Kamerlin, S. C. L., Modeling the Role of a Flexible Loop and Active Site Side Chains in Hydride Transfer Catalyzed by Glycerol-3-phosphate Dehydrogenase. *ACS Catal* **2020,** *10*, 11253-11267.
(94) Astl, L.; Verkhivker, G. M., Dynamic View of Allosteric Regulation in the Hsp70 Chaperones by J-Domain Cochaperone and Post-Translational Modifications: Computational Analysis of Hsp70 Mechanisms by Exploring Conformational Landscapes and Residue Interaction Networks. *J Chem Inf Model* **2020,** *60*, 1614-1631.
(95) Kuenze, G.; Vanoye, C. G.; Desai, R. R.; Adusumilli, S.; Brewer, K. R.; Woods, H.; McDonald, E. F.; Sanders, C. R.; George, A. L., Jr.; Meiler, J., Allosteric mechanism for KCNE1 modulation of KCNQ1 potassium channel activation. *eLife* **2020,** *9*, e57680.
(96) Leander, M.; Yuan, Y.; Meger, A.; Cui, Q.; Raman, S., Functional plasticity and evolutionary adaptation of allosteric regulation. *Proc Natl Acad Sci U S A* **2020,** *117*, 25445-25454.
(97) Kalescky, R.; Zhou, H.; Liu, J.; Tao, P., Rigid Residue Scan Simulations Systematically Reveal Residue Entropic Roles in Protein Allostery. *PLoS Comput Biol* **2016,** *12*, e1004893.
(98) Kalescky, R.; Liu, J.; Tao, P., Identifying Key Residues for Protein Allostery through Rigid Residue Scan. *J Phys Chem A* **2015,** *119*, 1689-700.
(99) Christensen, N. R.; Calyseva, J.; Fernandes, E. F. A.; Luchow, S.; Clemmensen, L. S.; Haugaard-Kedstrom, L. M.; Stromgaard, K., PDZ Domains as Drug Targets. *Adv Ther* **2019,** *2*, 1800143.
(100) Udugamasooriya, D. G.; Sharma, S. C.; Spaller, M. R., A chemical library approach to organic-modified peptide ligands for PDZ domain proteins: a synthetic, thermodynamic and structural investigation. *Chembiochem* **2008,** *9*, 1587-9.
(101) Aljameeli, A.; Thakkar, A.; Thomas, S.; Lakshmikanthan, V.; Iczkowski, K. A.; Shah, G. V., Calcitonin Receptor-Zonula Occludens-1 Interaction Is Critical for Calcitonin-Stimulated Prostate Cancer Metastasis. *PLoS One* **2016,** *11*, e0150090.
(102) Doruker, P.; Atilgan, A. R.; Bahar, I., Dynamics of proteins predicted by molecular dynamics simulations and analytical approaches: application to alpha-amylase inhibitor. *Proteins* **2000,** *40*, 512-24.
(103) Holme, P.; Saramäki, J., Temporal networks. *Physics Reports* **2012,** *519*, 97-125.
(104) Li, M.; Chen, S.; Chen, X.; Zhang, Y.; Wang, Y.; Tian, Q., Actional-Structural Graph Convolutional Networks for Skeleton-based Action Recognition. *arXiv e-prints* **2019**, arXiv:1904.12659.
(105) Kipf, T.; Fetaya, E.; Wang, K.; Welling, M.; Zemel, R., Neural Relational Inference for Interacting Systems. *arXiv e-prints* **2018**, arXiv:1802.04687.
(106) Schmiedel, J. M.; Lehner, B., Determining protein structures using deep mutagenesis. *Nat Genet* **2019,** *51*, 1177-1186.

(107) Rollins, N. J.; Brock, K. P.; Poelwijk, F. J.; Stiffler, M. A.; Gauthier, N. P.; Sander, C.; Marks, D. S., Inferring protein 3D structure from deep mutation scans. *Nat Genet* **2019,** *51*, 1170-1176.
(108) Li, X.; Lehner, B., Biophysical ambiguities prevent accurate genetic prediction. *Nat Commun* **2020,** *11*, 4923.
(109) Bolognesi, B.; Faure, A. J.; Seuma, M.; Schmiedel, J. M.; Tartaglia, G. G.; Lehner, B., The mutational landscape of a prion-like domain. *Nat Commun* **2019,** *10*, 4162.
(110) Horovitz, A.; Fleisher, R. C.; Mondal, T., Double-mutant cycles: new directions and applications. *Curr Opin Struct Biol* **2019,** *58*, 10-17.
(111) Werner, M.; Gapsys, V.; de Groot, B. L., One Plus One Makes Three: Triangular Coupling of Correlated Amino Acid Mutations. *J Phys Chem Lett* **2021,** *12*, 3195-3201.
(112) Nedrud, D.; Coyote-Maestas, W.; Schmidt, D., A large-scale survey of pairwise epistasis reveals a mechanism for evolutionary expansion and specialization of PDZ domains. *Proteins* **2021**.
(113) Bonati, L.; Rizzi, V.; Parrinello, M., Data-Driven Collective Variables for Enhanced Sampling. *J Phys Chem Lett* **2020,** *11*, 2998-3004.
(114) Allison, J. R., Computational methods for exploring protein conformations. *Biochem Soc Trans* **2020,** *48*, 1707-1724.
(115) Sidky, H.; Chen, W.; Ferguson, A. L., Machine learning for collective variable discovery and enhanced sampling in biomolecular simulation. *Mol Phys* **2020,** *118*, 1-21.
(116) Wehmeyer, C.; Noe, F., Time-lagged autoencoders: Deep learning of slow collective variables for molecular kinetics. *J Chem Phys* **2018,** *148*, 241703.
(117) Sittel, F.; Stock, G., Perspective: Identification of collective variables and metastable states of protein dynamics. *J Chem Phys* **2018,** *149*, 150901.
(118) Noe, F.; Clementi, C., Collective variables for the study of long-time kinetics from molecular trajectories: theory and methods. *Curr Opin Struct Biol* **2017,** *43*, 141-147.
(119) Nagel, D.; Weber, A.; Stock, G., MSMPathfinder: Identification of Pathways in Markov State Models. *J Chem Theory Comput* **2020,** *16*, 7874-7882.
(120) Brandt, S.; Sittel, F.; Ernst, M.; Stock, G., Machine Learning of Biomolecular Reaction Coordinates. *J Phys Chem Lett* **2018,** *9*, 2144-2150.
(121) Gianni, S.; Engstrom, A.; Larsson, M.; Calosci, N.; Malatesta, F.; Eklund, L.; Ngang, C. C.; Travaglini-Allocatelli, C.; Jemth, P., The kinetics of PDZ domain-ligand interactions and implications for the binding mechanism. *J Biol Chem* **2005,** *280*, 34805-12.
(122) Doerr, S.; Majewski, M.; Perez, A.; Kramer, A.; Clementi, C.; Noe, F.; Giorgino, T.; De Fabritiis, G., TorchMD: A Deep Learning Framework for Molecular Simulations. *J Chem Theory Comput* **2021**.
(123) Gao, X.; Ramezanghorbani, F.; Isayev, O.; Smith, J. S.; Roitberg, A. E., TorchANI: A Free and Open Source PyTorch-Based Deep Learning Implementation of the ANI Neural Network Potentials. *J Chem Inf Model* **2020,** *60*, 3408-3415.
(124) Noe, F.; Olsson, S.; Kohler, J.; Wu, H., Boltzmann generators: Sampling equilibrium states of many-body systems with deep learning. *Science* **2019,** *365*, 1147.