Chapter 1

Introduction

1.1 Protein conformational landscapes encode functional information.

Proteins are the machines that power cellular function and life. They allow us to see, smell, think, and carry out many of the basic functions required for us to live. However, when they malfunction or misbehavior (usually through mutation), they can also result in diseases like cancer or heart disease among others. Proteins are also utilized by viruses and bacteria to infect host cells, replicate, or even break down the drugs we give them.

Understanding protein behaviors relevant to health and disease depends on being able to model them at them at an atomic detail. Knowing atomic scale behaviors allows us to infer things like mechanism, thermodynamic behaviors, kinetic behaviors. This level of detail can provide predictive models and explanations for why certain mutations may cause disease, together which can be useful for targeting proteins using drug design methods. Indeed, knowledge of chemical interactions allows for design of chemical groups against pockets to jam them open [1]. Atomic-scale knowledge may even provide guiding principles upon which proteins can be designed to perform novel functions, which has implications for therapeutic design and industrial applications [2].

Structural biology methods have been transformative in allowing us to learn about structure of proteins and their behaviors. Indeed, the first view of protein structures was done using X-Ray crystallography [3]. However, static structures do not provide the complete picture of protein behavior. These static structures may not be able to provide complete information about a proteins stability [4], ligand affinities or specificities [5], or how different mutations would impact function [6]. Indeed, it has been often observed that crystal structures of the same protein families are too similar to explain the difference in their measure physiological parameters [7].

As the atoms of a protein move around relative to one another, a protein is able to shift between an enormous number of different structures. Indeed, even small proteins of $<$100 amino acids have $\sim$200 rotatable bonds along its backbone alone, granting access to more than 10^{60} backbone conformations alone [8]. Many of these structures however are never accessed by the protein in physiologically relevant timescales, but of the fraction that are accessed, some of them may have relevance to a protein's mechanism and biophysical behavior. Each of these structures that a protein may shape-shift has an associated energy that characterizes the atomic interactions within the protein and between the protein and its environment. Given that the probability of a protein adopting any one structure is proportional to the exponential of the energy of that structure, we are able to characterize how likely a protein is to adopt some states over others. The phase space of a protein's energies (or probabilities) and their corresponding structures are often referred to as an "energy landscape", with most likely states (such as those observed by crystallographic methods) are named as "ground" states due to being energy minima.

There are also less likely "excited" states that the protein can transition to from these ground states, some of which may contain key functional information. NMR and HDX have provided unique functional insight into the conformational heterogeneity a protein can have [9, 10]. Indeed, work on DHFR has been critical in understanding the complete catalytic cycle and the residue level conformations (and dynamics) of each stage in the cycle [11]. NMR experiments on DHFR that was arrested in one stage of its catalytic cycle provided evidence that DHFR was also adopting stages in the latter part of its cycle [11]. Mutational experiments also put forth a correlation between dynamics and catalytic ability [10–12]. Powerful combinations of NMR, crystallography, and computer simulations

have been rationalized cofactor- and mutational-effects on kinases [13,14]. With advances in structural methodologies, it is even possible to resolve structural information about excited states [11, 12]. Other EPR experiments have granted unique insight into the conformational landscape of proteins [15]. NMR also has the additional power to measure the degree of conformational entropy of residues [16]. Recently, Cryo-EM structures of even large complexes have revealed the degree of conformational heterogeneity in physiologically relevant systems [17–20]. It is important to note that each of these methods have tradeoffs between resolution, labelling strategies that may perturb the system, or other limitations (system size, material requirements, etc.). Altogether, a large body of work studying excited states of folded proteins suggests that the equilibrium motions of a protein encode all of its functionally relevant states.

Molecular dynamics (MD) simulations have the potential to provide atomistic detail to explain complex biological processes. These simulations, by iterating Newton's laws of motion over each atom to compute their movements over time, act as "computational microscopes" allowing us to observe the different conformations a protein adopts [21, 22]. A perfect simulation, run at equilibrium conditions of a protein, would completely describe a protein's thermodynamic and kinetic behaviors. However, there are major limitations to MD: (i) The accuracy of atomic parameters (aka "force fields") that are used to describe the interaction energies in a system (A topic that is discussed extensively elsewhere [23]). (ii) Simulations utilize femtosecond-sized timesteps, making it expensive to gather data at biologically relevant timescales (microsecond to milliseconds). (iii) Interpreting these large datasets with contain thousands of unique structures to generate biologically meaningful predictions remains a daunting task from both scientific and engineering standpoints.

1.2 The Folding@home platform allows access to protein motions at biologically relevant timescales.

1.2.1 Folding@home distributes simulations across thousands of computers at once.

A myriad of methods have been developed to improve sampling of protein motions out to biologically relevant timescales [24–27]. However, many of them observe these events by perturbing the thermodynamics or kinetics of a system via the Hamiltonian or other means. In turn this affects the predictions these simulations make about the system at equilibrium. The advent of GPUs allowed for us to access even longer timescales with equilibrium simulations, but they are also bounded by an upper limit of possible parallelism in computing architectures [28–31]. Recently new and novel adaptive sampling techniques have come up to allow for sampling of proteins at equilibrium [32] but require knowledge of a system or a directed order parameter along which sampling is improved. Specialized hardware has also been developed (such as the ANTON supercomputer)] [33], and these have allowed computational biophysicsts to study processes at unprecedented timescales. However utilizing and maintaining this kind of specialized hardware can be a costly endeavor.

To sample unbiased simulations at biologically relevant timescales, The Folding@home platform was developed in 2000 by the Pande group [34]. Folding@home, now headquartered at the Bowman lab at WUSTL, is a distributed computing network that works to run MD simulations around the world. The platform runs off of the donations of thousands of citizen-scientists who have downloaded the app onto their computers, clusters, and hardware around the world. Folding@home works by taking trajectories of simulations and running them as smaller "work units" that are each small simulation on the order of single nanoseconds. The starting file for a work-unit is generated server-side, which is sent to a client software somewhere in the world (Fig. 1.1). Each work unit is a single chunk of time in a single MD trajectory. The client runs the work-unit and returns it, which becomes one chunk of a single trajectory. This is used to generate the subsequent work-unit (which represents the next chunk of time in a simulation trajectory). These work units can be stitched together to generate a single trajectory. Eventually you end up with dataset of multiple trajectories, that together massively sample a large amount of the energy landscape of a protein. Indeed, interpreting datasets of this scale is a unique "big data" challenge, and requires both the development of new methods and new software [35, 36].

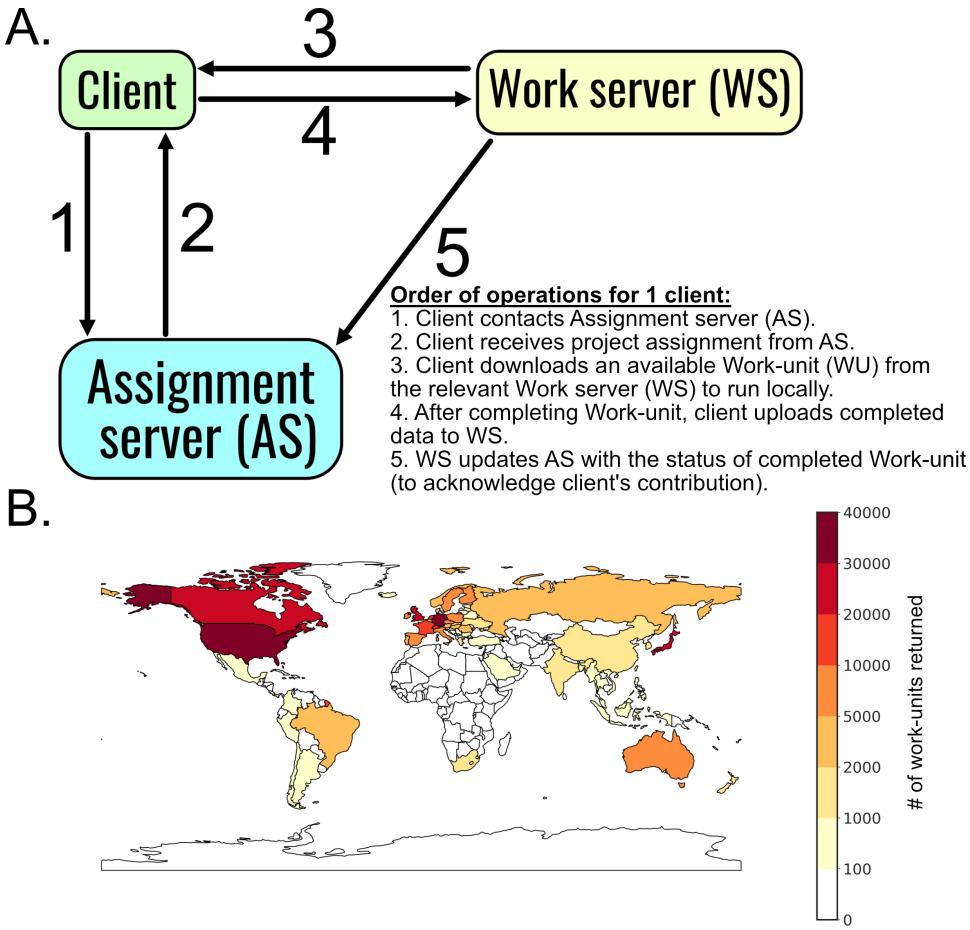

Figure 1.1: Folding@home distributes work units worldwide to run simulations **A.** Workflow schematic detailing the steps a citizen-scientist's computer ("client", green) takes to receive, run, and return a work-unit successfully by communicating with the Assignment Server (blue) and the Work server (yellow). **B.** Heatmap of completed work-units returned from each country in the world in a representative 48-hour period. The number of returned work units is indicated by the color scale (right).

1.2.2 Markov State Models allow for the construction of unified models from large simulation datasets.

Markov state models (MSMs) are network representations of a protein's free-energy landscape, providing map representations of protein conformational space with thermodynamic and kinetic properties taken from equilibrium simulations (Fig. 1.2). Rather than depending on a single long simulation that would eventually gather statistics on every state and transition between states, MSMs are capable of stitching together multiple short trajectories into a single unified landscape. In doing so, MSMs capture slow events that are far beyond the reach of any individual simulation. Thanks to distributed networks like Folding@home, gathering such large numbers of short trajectories is tractable in a reasonable time-frame. Many reviews have been dedicated to providing accessible and in-depth explanations of MSM technology [35, 37, 38].

MSMs require two components to describe the dynamics of a biomolecular system: (i): A discretization of a high dimensional state space into n conformational states, and (ii) A model of the stochoastic transitions between each

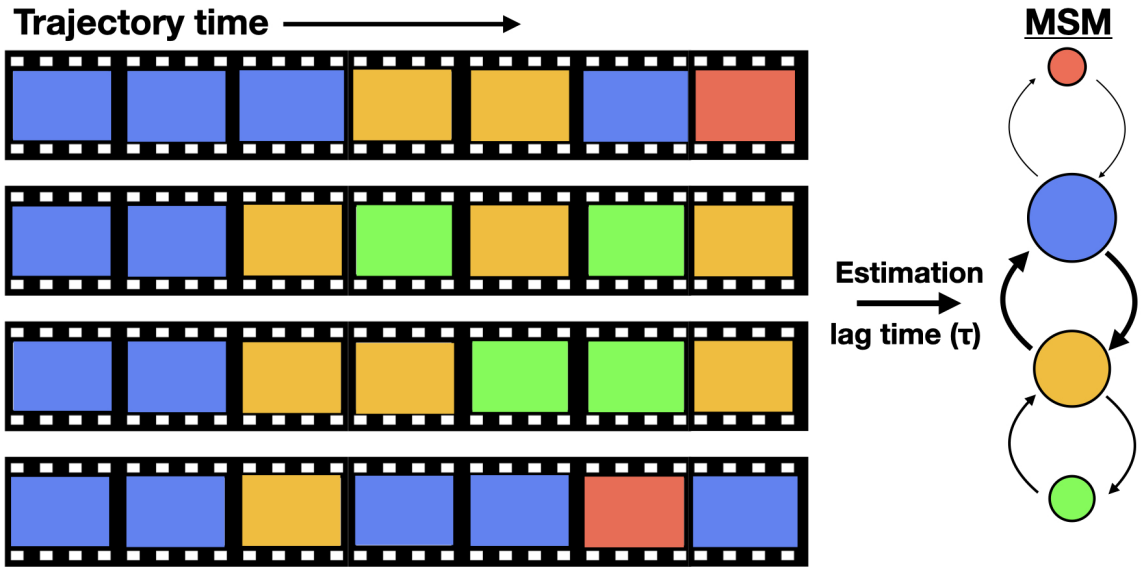

Figure 1.2: Diagram highlighting the conversion of a simulation dataset into a Markov State Model (MSM). Simulation trajectories (left) are parsed into discrete states on a per-frame basis (red, blue, green, yellow), and estimation methods are used to convert the trajectories into a MSM (right). The spheres radii is proportional to the population of that state, while arrow thickness denotes transition probability.

states, represented as a Transition Matrix **P**. The probability of transitioning from states i to state j (P_{ij})is:

$$P_{ij}(\tau) = Prob(x_{t+\tau} \in S_j | x_t \in S_i) \tag{1.1}$$

where τ is a lag-time parameter across which transitions between states are observed. These transition matrices **P** give rise to a stationary distribution π as a result of the eigenvalue equation:

$$\pi^T \mathbf{P} = \pi^T \tag{1.2}$$

where T represents a single time-step. In an MSM, the stationary distribution π represents the equilibrium probabilities for each state. Therefore, assuming sufficient statistics are collected to observe transitions between states, MSMs are able to appropriately capture the equilibrium thermodynamic and kinetic behaviors of a system. It is important to note that underlying MSM construction is the intuition that the discretized dynamics of biomolecules is memoryless (aka Markovian). That is, the probability of a transition from state i to state j at time t is only dependent on state i and not any of the previously visited states.

For any constructed MSM, the Markovian assumption is tested by looking at implied timescales plots of a Markov State Model. That is, for a given state decomposition, the molecular relaxation timescales for eigenvalues λ and eigenvectors r_i are computed for a series of lag-times τ:

$$t_i = -\frac{\tau}{\ln |\lambda_i(\tau)|} \tag{1.3}$$

These timescales t_i are plotted for a series of lag-times τ (**??**). If the timescales remain relatively unchanged for a series of τ values and higher, then the models constructed for those values of τ can be considered Markovian. The values of P_{ij} can then be estimated using a variety of estimation methods [37, 39].

A critical component of high-quality MSM construction is the features selected to be used in the discretization of state space. It is important to balance statistical error against systematic bias in the choice of features used to discretize states, as features that exist in narrow ranges may result in poor statistics, while broadly-ranged features may have large systematic errors due to convolution of multiple motions into a single state. Choice of appropriate feature is challenging often due to the sheer size of simulation datasets and the system-specific knowledge often need to identify appropriate features.

Much success has previously been seen using geometric features, such as cartesian coordinates or dihedral angles [40–42]. Other features such as solvent accessible surface area (SASA) or ligand-residence-time have also been useful for construction of predictive MSMs [exposons paper, Sun and Singh, eLife; APM paper]. One major breakthrough involved the use of time-lagged independent component analysis (tICA) [43], which transforms input coordinates into collective coordinates to identify the rarest (which are assumed to be the slowest) motions. For protein folding simulations, this can provide excellent dimensionality reduction, since the slowest coordinate is often the rarest and most valuable [44, 45]. However, in simulations of folded proteins, the rarest motions observed in simulations may be less valuable or interesting, or artifacts of sampling due to the size of many disease-causing biological systems.

Once appropriate features, such as atomic cartesian coordinates of the backbone, are selected, there are a myriad of approaches to cluster them into discrete states. One major method is a hybrid k-centers/k-medoids approach [46,47]. In brief, for each feature a distance metric, such as the Root Mean Square Deviation (RMSD) is computed between every pair. (i) The k-centers algorithm chooses an initial cluster center either as a predetermined point or randomly chosen, and then all points are assigned to this initial cluster. (ii) The distance to between every point and its assigned cluster center is then computed. (iii) the point with the largest distance to the assigned cluster center is then labelled as a new cluster center. (iv) The distances between all points and all cluster centers are recalculated, and points are reassigned to their closest cluster center based on these new added labels. Steps $(ii - iv)$ are repeated until the maximum distance from any point to its assigned cluster center goes below a specified threshold, or a maximum number of cluster centers is reached.

To refine the cluster center assignments, and ensure that the "center" of each cluster is truly equidistant from all assigned points, a k-medoids algorithm known as Partitioning around Medoids (PAM) [47]. PAM proceeds by iterating through each cluster and choosing a new center from one of the points currently assigned. All points and states are reassigned based on this new proposed cluster center. From this proposed center, a cost is calculated (the sum of distances from each point to their respective center), and the proposed center placement is accepted if the cost is minimized.

Once states are discretized and assigned across a simulation trajectory, the transition probability matrix is estimated. As mentioned above, the matrix of transitions between states is counted based on some lag time τ. This could be as simple as counting the number of transitions from states i to j between all states, and divide by the number of states i observed. To maintain ergodicity, some methods only estimate over the largest connected subset of states [48] or using maximum likelihood estimators that respect detailed balance [49]. Simple methods to average the count matrix with the average of itself

$$C_{ij}^{transpose} = \frac{C_{ij} + Cji}{2} \tag{1.4}$$

where C_{ij} is the observed number of transitions from state i to state j, and C_{ji} represents the number of transitions in the reverse direction. Subsequent row normalization is used to calculate the equilibrium probabilities:

$$\pi_i = \frac{\sum_j C_{ij}^{transpose}}{\sum_{k,j} C_{k,j}^{transpose}} \tag{1.5}$$

where $C_{ij}^{transpose}$ is the averaged number of transitions between states i and state j, and $C_{k,j}^{transpose}$ is the number of transitions between states k and j. Recent success has been observed by simply adding a pseudocount $\tilde{C}$ to serve as an estimate of the system in absence of data [39, 50]. This pseudocount is computed as a single observed transition that is divided up across all states

$$\tilde{C} = \frac{1}{N} \tag{1.6}$$

where N is the number of states.

1.2.3 The scalable power of Folding@home has generated insights into protein behaviors.

There are many success stories that have come out of the usage of the Folding@home platform. Success stories of Folding@home include the observation of a millisecond folding in 2010 [40]. Markov state models have been shown to quantitatively agree with experimental measurements [29,30]. Particularly, there is excellent agreement between microsecond scale simulations and the properties of systems measured with NMR and room-temperature crystallography [51, 52]. Indeed, simulations have been able to interpret the impact of mutations on diseases such as phenylketonuria [53] and characterize the landscapes of a myriad of targets [54]. Folding@home and MSMs have allowed for the assessment of families of protein homologs [7].

Folding@home has made significant progress in developing and supplementing experimental work with predictive models. For example, Folding@home was used to identify novel cryptic pockets in TEM-1 β-lactmase [55], the enzyme most directly involved in antibiotic breakdown and microbial resistance. Indeed, accounting for the dynamics within the active site of TEM-1 β-lactamase substantially improved the predictive ability of modern virtual screening technologies [6]. Similar approaches have yielded valuable insights into the pH dependence of protein-protein interactions [56].

1.3 Allosteric communication is critical for protein function, but difficult to infer.

1.3.1 Allosteric communication is universal and critical for biological function.

One phenomena in biology that happens at long-timescales is communication between distant structural elements of protein, referred to as allostery [57]. First recognized in hemoglobin [58] where the binding of oxygen to a single subunit increases the oxygen-affinity for the other three subunits. Since then, the importance of allostery has been recognize in a myriad of cellular functions, such as transcription factors [59, 60] or cellular signaling [61].

A famous protein (and drug target) is the G protein coupled receptor, which transmit informations from outside the cell to inside the cell based on a stimulus. This stimulus can be anything from ligand binding [62] to membrane deformations [63]. Structural methods revealed rearrangements of transmembrane helices that converts the GPCR into an "active" form [64]. However, recent data of different GPCR-effector complexes have highlighted the conformational heterogeneity can exist in the allostery and activation of GPCRs [17–20].

This idea that a protein's conformational landscape can impact its allosteric behavior leads one to speculate if all proteins have some degree of allosteric coupling [65]. Indeed, the ubiquity of allostery has been acknowledge in studies of natural and directed evolution [66, 67], where mutations distant from the active site can impact measured properties. Given the potential universal nature of allostery, it is worth speculating if mutations and ligands both act by tapping into existing allosteric networks to modulate the distribution of structures and dynamics a residue can adopt [38]. Thus, capturing and understanding allosteric coupling in proteins could present new opportunities for modulating biological processes, designing therapeutics, or even designing new proteins.

Mounting evidence shows that leveraging allostery can be used to modulate proteins. There are a multitude of biological systems where allostery is leveraged to inhibit or activate protein-protein interactions, and so it may be possible to identify small molecules that could achieve the same objective of modulating protein behaviors. An allosteric drug that could modulate protein behaviors could play a huge role in restoring lost functions or reducing overactive protein behaviors. However, modern drug-design methods often require the presence of a "pocket" to successfully design small molecule hits. Many surfaces involved in protein-protein interactions are often too flat for a small molecule to bind tightly [68], and targeting known ligand binding sites of critical signaling proteins like GPCRs often creates the risk of off-target effects. Identifying distant pockets that are not as conserved between homologs could be a means to achieve specificity [69].

Hidden allosteric targets known as 'cryptic pockets' could be promising targets for drug design methods. The shape-shifting nature of proteins' implies the existence of states that contain new pockets that are not observed in existing experimental structures. The hidden pockets may also be cryptic allosteric sites that are connected to key functional sites via the underlying allosteric network of a protein [70]. Successful methods have emerged to identify novel allosteric sites, some of which have been verified by experiments [55]. Indeed, the value of cryptic pockets has

been supported by the discovery of small molecule inhibitors that are shown to bind an allosteric pocket and modulate a protein's function [6, 71–73]. Computer simulations provide promising avenues to hunt and target cryptic pockets, an effort which has yield promising results [74, 75], but remains a challenge to apply to a wide variety of symptoms. Furthermore, the discovery of a distant pocket in a protein does not imply it is "useful" as a drug target, because it is difficult to measure the degree of coupling between the cryptic pocket and a protein's functional regions (like active site). Understanding the degree (and residue-level detail) of how a cryptic pocket couples to functional regions could further supplement drug design strategies and rational drug design approaches. However, it remains a challenge to obtain a complete picture of the allosteric network, of a protein, due to the complex nature of allosteric coupling.

1.3.2 Inferring allosteric communication in proteins remains a non-trivial task.

Methods for inferring allostery typically rely on observing concerted structural changes. A system with two distant sites may jump between alternative configurations in some coupled fashion. That is, the structure of site A may be coupled to the configuration of site B, and vice versa. This extreme example of conformational selection could be inferred by comparing structures of proteins before and after some perturbation to one of the sites is introduced (such as ligand binding). Indeed, crystallography and HDX methods have provide useful in revealing residues involved in allosteric networks of TIM-barrels and their catalytic domains [76]. Multiple NMR methods have also proven useful in studying the nature of allosteric communication between proteins [77].

Likewise, computational methods measure concerted structural changes using a variety of metrics on a myriad of features of proteins. While some methods utilize sequence coevolution to group proteins regions in to "sectors" that are coupled to one another [78], there is growing recognition that molecular simulations can capture the atomic detail of allosteric coupling between sites. The underlying assumption that a proteins functional states are encoded in the equilibrium simulations implies that observing correlated motions in MD simulations would be representative of the degree of coupling between residues. Indeed, a number of algorithms use a myriad of features and metrics to quantify coupling [79, 80]. Some features used could be the backbone carbon-alpha atoms of proteins, and measuring the degree of covariance in pairs of C alpha atoms [81]. Other methods utilize mutual information methods on dihedral angles to quantify how much better one residue's dihedral angle predicts the dihedral angle of another residue [82]. However, there has been growing recognition that allostery via concerted structural changes is not the only mechanism through which two sites may be coupled.

In recent years the role of conformational entropy in allosteric communication has become increasingly acknowledge. The importance of conformational entropy was first theoretically described in 1984 by Cooper and Dryden [45] [83]. Since this, experimental evidence for this "dynamical allostery" has grown, particularly via NMR data demonstrated two sites on a transcription factor were coupled with no discernible structural changes [84]. Furthermore, intrinsically disordered regions can also play a role in allosteric coupling, as normally ordered regions of a protein may transition locally into a higher entropy excited states, rapidly hopping between multiple thermodynamic minima. This effective flattening of the free energy surface for a set of residues distinguishes dynamic allostery from the previously discussed mechanism of concerted structural changes, which only describes a pair of alternating conformations. Indeed, there is growing recognition for the need to measure both concerted structural changes and changes in conformational entropy from MD simulations to more completely measure allosteric coupling in proteins [85]. The ability to construct allosteric networks measuring both structure and disorder, has the potential to explain the mechanism of coupling in many complex biological process, and may present opportunities to identify promising druggable pockets and the role they play in modulating protein function.

1.4 Scope of thesis.

It remains an important, but critical, challenge to understand allostery to more completely describe biological behavior. The potential power of MD simulations to explain complex biological processes in atomistic detail presents a promising avenue to achieve these goals, but the tools to do so remain limited in scope, and generating simulation datasets that capture slow allosterically driven processes remains a challenge. Thus, this thesis seeks to push forward our understanding of allosteric communication and the conformational landscape of proteins, and leverage these insights towards understanding fundamental biological phenomena or supplementing drug design efforts.

In this thesis I will describe a method to infer allosteric coupling in MD simulations both concerted structural changes and conformational entropy. This will be done by measuring the Correlation of all Rotameric and Dynamical States (CARDS) – a novel method presented in chapter **??**. This algorithm builds upon previous work that infers allostery through structural changes by seeking to capture allostery through changes in conformational entropy. It parses dihedral angles into dynamical states, capturing whether a rotamer is ordered (remaining in a single basin) or disordered (rapidly hopping between basis). We then describe our framework to measure coupling between every pair of residues by computing coupling between rotameric states, between dynamical states, as well as cross-correlations between them. We apply the CARDS methods to a system with known dynamic allostery, the Catabolite Activator Protein (CAP) a transcription factor whose allosteric behavior was previously measured in experimental and ITC studies.

Chapter **??** describes the application of CARDS and other MD/MSM methods to a known allosteric system of importance, the heterotrimeric G proteins. Heterotrimeric G proteins are molecular switches that regulate everything including vision, smell, and neurotransmission. Malfunctions in G protein activation implicated in cancers such as Uveal Melanoma. While considerable work has studied the process of G protein thermodynamics and kinetics, a complete mechanism of activation remains unclear; in part including the allosteric network coupling the receptor and nucleotide binding sites. Here we describe in atomistic detail for the first time a complete mechanism of G protein activation, GDP release, and the conformational and dynamical changes driving this process.

Chapter **??** describes the discovery of a hidden cryptic pocket in the previously 'undruggable' ebolavirus protein VP35. The seeding strategy described in chapter **??** is applied to the RNA-binding VP35 protein. A cryptic pocket is identified from a Folding@home dataset using existing methods, and CARDS identifies the degree of coupling between the cryptic pocket and residues known to be important for Protein-Protein and Protein-Nucleic-Acid interactions (PPIs and PNIs, respectively). The existience of this cryptic pocket is then tested using experiments and supported, and the functional importance of this distance allosteric site is supported using experimental techniques that demonstrate VP35 inhibition by stabilizing the pocket open state.

Chapters **??** and **??** describe recent efforts utilizing Folding@home to study SARS-CoV-2 the virus behind the covid19 pandemic. In chapter **??**, we rapidly generate conformations of the Nucleocapsid protein folded domains, both RNA-binding- and Dimerization-Domains, and use them in conjunction with further simulations and experiments. This study, described in chapter **??**, shows that the nucleocapsid protein is dynamic, disordered, and undergoes Liquid Liquid Phase Separation (LLPS) behavior. Folding@home simulations of the folded domain are used to seed Monte Carlo simulations of the intrinsically disordered regions of the Nucleocapsid protein to obtain a complete picture of the free energy landscape, which would not have been achieved using a single method alone. These simulations are integrated with experimental data to describe a symmetry-breaking model that explains how the N protein may package the genome.

In Chapter **??**, we describe a recent effort where Folding@home shifted focus to simulating potential drug targets in the proteome of the SARS-CoV-2 virus, the cause of the COVID-19 pandemic. Many citizen-scientists around the world rallied to the cause, downloading the app and helping us run simulations. We describe simulations of the almost every protein possible from SARS-CoV-2, generating a historic 0.1 seconds of equilibrium simulation data using Folding@home. Included is a description of how this was achieved through generations donations, partnerships and contributions to Folding@home that allowed the platform to surpass the exascale barrier in computing speed, a feat never before achieved in human history. We utilize this monumental computational power to simulate the viral proteome to study how the viral Spike protein uses conformational masking to evade an immune response, and describe efforts to identify cryptic pockets that are not present in existing experimental snapshots. The data generated by Folding@home presents new potential targets for drug design efforts, and new structural and mechanistic insights that may supplement the design of therapeutics.

Chapter **??** demonstrates how MD simulations can be integrated with standard structural biology techniques to explain mechanisms of antibiotic resistance. The cefotaximase enzyme CTXM is responsible for the breakdown of many modern cephalosporins, which can result in microbial resistance and sustained infections. With each new drug generated, such as Ceftazidime (CAZ), CTXM has been shown to accrue mutations that grant it enhanced resistance against these new inhibitors. This chapter will focus on two mutations, D240G and P167S. Counterintuitively, these mutations are not additive in their behavior, and that the presence of both of these mutations abrogates the CAZ resistance profile in CTXM. Combining MD and crystallographic methods, along with biochemical approaches, we demonstrate how these mutations uniquely alter the acyl-enzyme complex and modulate a key-region in CTXM known

as the Ω-loop. We show that while each mutation uniquely modulates the Omega-Loop to better accomdate CAZ, both mutations revert the conformational behavior of the Omega-loop back to it's non-resistant state. This study demonstrates the unique ability of combining MD with structural biology methods to better understand fundamental behaviors and biological phenomena.

Lastly, in chapter **??** I summarize the main advancements presented within this thesis. While previous chapters describe present advancements that have been made, this chapter explores the general implications of these findings and how allosteric coupling may be a universal phenomena. I expand by discussing future projects, questions that present the next steps, and speculate on the prospect of further integrating simulations with experiments to better answer fundamental biological questions.

Bibliography

[1] Bryan L Roth, John J Irwin, and Brian K Shoichet. Discovery of new GPCR ligands to illuminate new biology. *Nature chemical biology*, 13(11):1143–1151, November 2017.

[2] Brian D. Weitzner, Yakov Kipnis, A. Gerard Daniel, Donald Hilvert, and David Baker. A computational method for design of connected catalytic networks in proteins. *Protein Science*, 28(12):2036–2041, December 2019.

[3] J. C. Kendrew, G. Bodo, H. M. Dintzis, R. G. Parrish, H. Wyckoff, and D. C. Phillips. A three-dimensional model of the myoglobin molecule obtained by x-ray analysis. *Nature*, 181(4610):662–666, March 1958.

[4] Sofia Khan and Mauno Vihinen. Performance of protein stability predictors. *Human Mutation*, 31(6):675–684, March 2010.

[5] Jian Yin, Niel M. Henriksen, David R. Slochower, Michael R. Shirts, Michael W. Chiu, David L. Mobley, and Michael K. Gilson. Overview of the SAMPL5 host–guest challenge: Are we doing better? *Journal of Computer-Aided Molecular Design*, 31(1):1–19, January 2017.

[6] Kathryn M Hart, Chris M W Ho, Supratik Dutta, Michael L Gross, and Gregory R Bowman. Modelling proteins' hidden conformations to predict antibiotic resistance. *Nature communications*, 7:12965, October 2016.

[7] Justin R Porter, Artur Meller, Maxwell I Zimmerman, Michael J Greenberg, and Gregory R Bowman. Conformational distributions of isolated myosin motor domains encode their mechanochemical properties. *eLife*, 9:e55132, May 2020.

[8] C Levinthal. How to Fold Graciously. *Topics in mossbauer spectroscopy*, 20(1):25–44, October 1969.

[9] Robert L. Baldwin. Early days of protein hydrogen exchange: 1954-1972. *Proteins: Structure, Function, and Bioinformatics*, 79(7):2021–2026, July 2011.

[10] Katherine A Henzler-Wildman, Ming Lei, Vu Thai, S Jordan Kerns, Martin Karplus, and Dorothee Kern. A hierarchy of timescales in protein dynamics is linked to enzyme catalysis. *Nature*, 450(7171):913–916, December 2007.

[11] David D Boehr, Dan McElheny, H Jane Dyson, and P E Wright. The dynamic energy landscape of dihydrofolate reductase catalysis. *Science*, 313(5793):1638–1642, September 2006.

[12] David D Boehr, Ruth Nussinov, and P E Wright. The role of dynamic conformational ensembles in biomolecular recognition. *Nature chemical biology*, 5(11):789–796, November 2009.

[13] Adelajda Zorba, Vanessa Buosi, Steffen Kutter, Nadja Kern, Francesco Pontiggia, Young-Jin Cho, and Dorothee Kern. Molecular mechanism of Aurora A kinase autophosphorylation and its allosteric activation by TPX2. *eLife*, 3:e02667, May 2014.

[14] S Jordan Kerns, Roman V Agafonov, Young-Jin Cho, Francesco Pontiggia, Renee Otten, Dimitar V Pachov, Steffen Kutter, Lien A Phung, Padraig N Murphy, Vu Thai, Tom Alber, Michael F Hagan, and Dorothee Kern. The energy landscape of adenylate kinase during catalysis. *Nature Structural & Molecular Biology*, 22(2):124–131, February 2015.

[15] Ned Van Eps, Lori L Anderson, Oleg G Kisselev, Thomas J Baranski, Wayne L Hubbell, and Garland R Marshall. Electron paramagnetic resonance studies of functionally active, nitroxide spin-labeled peptide analogues of the C-terminus of a G-protein alpha subunit. *Biochemistry*, 49(32):6877–6886, August 2010.

[16] A Joshua Wand. The dark energy of proteins comes to light: conformational entropy and its role in protein function revealed by NMR relaxation. 23(1):75–81, February 2013.

[17] Antoine Koehl, Hongli Hu, Shoji Maeda, Yan Zhang, Qianhui Qu, Joseph M Paggi, Naomi R Latorraca, Daniel Hilger, Roger Dawson, Hugues Matile, Gebhard F X Schertler, Sébastien Granier, William I Weis, Ron O Dror, Aashish Manglik, Georgios Skiniotis, and Brian K Kobilka. Structure of the μ-opioid receptor–G i protein complex. *Nature*, 383:1, June 2018.

[18] Christopher J Draper-Joyce, Maryam Khoshouei, David M Thal, Yi-Lynn Liang, Anh T N Nguyen, Sebastian G B Furness, Hariprasad Venugopal, Jo-Anne Baltos, Jürgen M Plitzko, Radostin Danev, Wolfgang Baumeister, Lauren T May, Denise Wootten, Patrick M Sexton, Alisa Glukhova, and Arthur Christopoulos. Structure of the adenosine-bound human adenosine A1 receptor-Gi complex. *Nature*, 63:1, June 2018.

[19] Javier García-Nafría, Rony Nehmé, Patricia C Edwards, and Christopher G Tate. Cryo-EM structure of the serotonin 5-HT1B receptor coupled to heterotrimeric Go. *Nature*, 7:118, June 2018.

[20] Yanyong Kang, Oleg Kuybeda, Parker W de Waal, Somnath Mukherjee, Ned Van Eps, Przemyslaw Dutka, X Edward Zhou, Alberto Bartesaghi, Satchal Erramilli, Takefumi Morizumi, Xin Gu, Yanting Yin, Ping Liu, Yi Jiang, Xing Meng, Gongpu Zhao, Karsten Melcher, Oliver P Ernst, Anthony A Kossiakoff, Sriram Subramaniam, and H Eric Xu. Cryo-EM structure of human rhodopsin bound to an inhibitory G protein. *Nature*, 63(Suppl. 1):1256, June 2018.

[21] Ron O Dror, Robert M Dirks, J P Grossman, Huafeng Xu, and David E Shaw. Biomolecular Simulation: A Computational Microscope for Molecular Biology. *dx.doi.org*, 41(1):429–452, May 2012.

[22] Eric H. Lee, Jen Hsin, Marcos Sotomayor, Gemma Comellas, and Klaus Schulten. Discovery through the computational microscope. *Structure*, 17(10):1295–1306, October 2009.

[23] Pedro E. M. Lopes, Olgun Guvench, and Alexander D. MacKerell. Current status of protein force fields for molecular dynamics simulations. In Andreas Kukol, editor, *Molecular Modeling of Proteins*, volume 1215, pages 47–71. Springer New York, New York, NY, 2015.

[24] Weinan E and Eric Vanden-Eijnden. Transition-Path Theory and Path-Finding Algorithms for the Study of Rare Events. *dx.doi.org*, 61(1):391–420, March 2010.

[25] Donald Hamelberg, John Mongan, and J. Andrew McCammon. Accelerated molecular dynamics: A promising and efficient simulation method for biomolecules. *The Journal of Chemical Physics*, 120(24):11919–11929, June 2004.

[26] Dietmar Paschek, Hugh Nymeyer, and Angel E. García. Replica exchange simulation of reversible folding/unfolding of the Trp-cage miniprotein in explicit solvent: On the structure and possible role of internal water. *Journal of Structural Biology*, 157(3):524–533, March 2007.

[27] Stefano Piana and Alessandro Laio. A bias-exchange approach to protein folding. *The Journal of Physical Chemistry B*, 111(17):4553–4559, May 2007.

[28] John E. Stone, James C. Phillips, Peter L. Freddolino, David J. Hardy, Leonardo G. Trabuco, and Klaus Schulten. Accelerating molecular modeling applications with graphics processors. *Journal of Computational Chemistry*, 28(16):2618–2640, December 2007.

[29] Mark S. Friedrichs, Peter Eastman, Vishal Vaidyanathan, Mike Houston, Scott Legrand, Adam L. Beberg, Daniel L. Ensign, Christopher M. Bruns, and Vijay S. Pande. Accelerating molecular dynamic simulation on graphics processing units. *Journal of Computational Chemistry*, 30(6):864–872, April 2009.

[30] Peter Eastman and Vijay Pande. Openmm: a hardware-independent framework for molecular simulations. *Computing in Science & Engineering*, 12(4):34–39, July 2010.

[31] Peter Eastman, Jason Swails, John D. Chodera, Robert T. McGibbon, Yutong Zhao, Kyle A. Beauchamp, Lee-Ping Wang, Andrew C. Simmonett, Matthew P. Harrigan, Chaya D. Stern, Rafal P. Wiewiora, Bernard R. Brooks, and Vijay S. Pande. OpenMM 7: Rapid development of high performance algorithms for molecular dynamics. *PLOS Computational Biology*, 13(7):e1005659, July 2017.

[32] Maxwell I. Zimmerman and Gregory R. Bowman. Fast conformational searches by balancing exploration/exploitation trade-offs. *Journal of Chemical Theory and Computation*, 11(12):5747–5757, December 2015.

[33] David E. Shaw, Martin M. Deneroff, Ron O. Dror, Jeffrey S. Kuskin, Richard H. Larson, John K. Salmon, Cliff Young, Brannon Batson, Kevin J. Bowers, Jack C. Chao, Michael P. Eastwood, Joseph Gagliardo, J. P. Grossman, C. Richard Ho, Douglas J. Ierardi, István Kolossváry, John L. Klepeis, Timothy Layman, Christine McLeavey, Mark A. Moraes, Rolf Mueller, Edward C. Priest, Yibing Shan, Jochen Spengler, Michael Theobald, Brian Towles, and Stanley C. Wang. Anton, a special-purpose machine for molecular dynamics simulation. *Communications of the ACM*, 51(7):91–97, July 2008.

[34] M Shirts and V S Pande. COMPUTING: Screen Savers of the World Unite! *Science*, 290(5498):1903–1904, December 2000.

[35] Brooke E Husic and Vijay S Pande. Markov State Models: From an Art to a Science. *Journal of the American Chemical Society*, page jacs.7b12191, January 2018.

[36] J. R. Porter, M. I. Zimmerman, and G. R. Bowman. Enspara: Modeling molecular ensembles with scalable data structures and parallel computing. *The Journal of Chemical Physics*, 150(4):044108, 2019.

[37] John D Chodera and Frank Noé. Markov state models of biomolecular conformational dynamics. *Current opinion in structural biology*, 25:135–144, April 2014.

[38] Catherine R Knoverek, Gaya K Amarasinghe, and Gregory R Bowman. Advanced Methods for Accessing Protein Shape-Shifting Present New Therapeutic Opportunities. *Trends in biochemical sciences*, December 2018.

[39] Maxwell I Zimmerman, Justin R Porter, Xianqiang Sun, Roseane R Silva, and Gregory R Bowman. Choice of Adaptive Sampling Strategy Impacts State Discovery, Transition Probabilities, and the Apparent Mechanism of Conformational Changes. *Journal of Chemical Theory and Computation*, q-bio.BM:acs.jctc.8b00500, October 2018.

[40] Vincent A Voelz, Gregory R Bowman, Kyle Beauchamp, and Vijay S Pande. Molecular simulation of ab initio protein folding for a millisecond folder NTL9(1-39). *Journal of the American Chemical Society*, 132(5):1526–1528, February 2010.

[41] Gregory R Bowman and Phillip L Geissler. Equilibrium fluctuations of a single folded protein reveal a multitude of potential cryptic allosteric sites. *Proceedings of the National Academy of Sciences*, 109(29):11681–11686, July 2012.

[42] Xianqiang Sun, Sukrit Singh, Kendall Blumer, and Gregory R Bowman. Simulation of spontaneous G protein activation reveals a new intermediate driving GDP unbinding. *eLife*, 7, October 2018.

[43] L Molgedey and HG Schuster. Separation of a mixture of independent signals using time delayed correlations. *Physical review letters*, 72(23):3634–3637, June 1994.

[44] Mohammad M Sultan and Vijay S Pande. tICA-Metadynamics: Accelerating Metadynamics by Using Kinetically Selected Collective Variables. *Journal of Chemical Theory and Computation*, 13(6):2440–2447, June 2017.

[45] Guillermo Pérez-Hernández, Fabian Paul, Toni Giorgino, Gianni De Fabritiis, and Frank Noé. Identification of slow molecular order parameters for Markov model construction. *arXiv.org*, (1):015102, February 2013.

[46] Teofilo F. Gonzalez. Clustering to minimize the maximum intercluster distance. *Theoretical Computer Science*, 38:293–306, 1985.

[47] Hae-Sang Park and Chi-Hyuck Jun. A simple and fast algorithm for K-medoids clustering. *Expert Systems with Applications*, 36(2):3336–3341, March 2009.

[48] Jan-Hendrik Prinz, John D. Chodera, Vijay S. Pande, William C. Swope, Jeremy C. Smith, and Frank Noé. Optimal use of data in parallel tempering simulations for the construction of discrete-state Markov models of biomolecular dynamics. *The Journal of Chemical Physics*, 134(24):244108, June 2011.

[49] Jan-Hendrik Prinz, Hao Wu, Marco Sarich, Bettina Keller, Martin Senne, Martin Held, John D. Chodera, Christof Schütte, and Frank Noé. Markov models of molecular kinetics: Generation and validation. *The Journal of Chemical Physics*, 134(17):174105, May 2011.

[50] Matthew A. Cruz, Thomas E. Frederick, Sukrit Singh, Neha Vithani, Maxwell I. Zimmerman, Justin R. Porter, Katelyn E. Moeder, Gaya K. Amarasinghe, and Gregory R. Bowman. Discovery of a cryptic allosteric site in ebola's 'undruggable' vp35 protein using simulations and experiments. *bioRxiv*, 2020.

[51] Gregory R Bowman. Accurately modeling nanosecond protein dynamics requires at least microseconds of simulation. *Journal of computational chemistry*, pages n/a–n/a, June 2015.

[52] Gregory R Bowman and Phillip L Geissler. Extensive conformational heterogeneity within protein cores. *The Journal of Physical Chemistry B*, 118(24):6417–6423, 2014.

[53] Yunhui Ge, Elias Borne, Shannon Stewart, Michael R. Hansen, Emilia C. Arturo, Eileen K. Jaffe, and Vincent A. Voelz. Simulations of the regulatory ACT domain of human phenylalanine hydroxylase (Pah) unveil its mechanism of phenylalanine binding. *Journal of Biological Chemistry*, 293(51):19532–19543, December 2018.

[54] Shi Chen, Rafal P Wiewiora, Fanwang Meng, Nicolas Babault, Anqi Ma, Wenyu Yu, Kun Qian, Hao Hu, Hua Zou, Junyi Wang, Shijie Fan, Gil Blum, Fabio Pittella-Silva, Kyle A Beauchamp, Wolfram Tempel, Hualiang Jiang, Kaixian Chen, Robert J Skene, Yujun George Zheng, Peter J Brown, Jian Jin, Cheng Luo, John D Chodera, and Minkui Luo. The dynamic conformational landscape of the protein methyltransferase SETD8. *eLife*, 8:e45403, May 2019.

[55] Justin R Porter, Katelyn E Moeder, Carrie A Sibbald, Maxwell I Zimmerman, Kathryn M Hart, Michael J Greenberg, and Gregory R Bowman. Cooperative Changes in Solvent Exposure Identify Cryptic Pockets, Switches, and Allosteric Coupling. *Biophysical Journal*, 116(5):818–830, March 2019.

[56] Roberta Pascolutti, Xianqiang Sun, Joseph Kao, Roy L. Maute, Aaron M. Ring, Gregory R. Bowman, and Andrew C. Kruse. Structure and dynamics of pd-l1 and an ultra-high-affinity pd-1 receptor mutant. *Structure*, 24(10):1719–1728, October 2016.

[57] V. J. Hilser. An ensemble view of allostery. *Science*, 327(5966):653–654, February 2010.

[58] M F Perutz. Stereochemistry of cooperative effects in haemoglobin. *Nature*, 228(5273):726–739, November 1970.

[59] Shiou-Ru Tzeng and Charalampos G Kalodimos. Dynamic activation of an allosteric regulatory protein. *Nature*, 462(7271):368–372, November 2009.

[60] Shiou-Ru Tzeng and Charalampos G Kalodimos. Protein dynamics and allostery: an NMR view. 21(1):62–67, February 2011.

[61] William I Weis and Brian K Kobilka. The Molecular Basis of G Protein–Coupled Receptor Activation. *Annual review of biochemistry*, 87(1):897–919, June 2018.

[62] Daniel Wacker, Raymond C. Stevens, and Bryan L. Roth. How ligands illuminate gpcr molecular pharmacology. *Cell*, 170(3):414–427, July 2017.

[63] Kadla R Rosholm, Natascha Leijnse, Anna Mantsiou, Vadym Tkach, Søren L Pedersen, Volker F Wirth, Lene B Oddershede, Knud J Jensen, Karen L Martinez, Nikos S Hatzakis, Poul Martin Bendix, Andrew Callan-Jones, and Dimitrios Stamou. Membrane curvature regulates ligand-specific membrane sorting of GPCRs in living cells. *Nature Chemical Biology*, 13(7):724–729, July 2017.

[64] Søren G F Rasmussen, Brian T DeVree, Yaozhong Zou, Andrew C Kruse, Ka Young Chung, Tong Sun Kobilka, Foon Sun Thian, Pil Seok Chae, Els Pardon, Diane Calinski, Jesper M Mathiesen, Syed T A Shah, Joseph A Lyons, Martin Caffrey, Samuel H Gellman, Jan Steyaert, Georgios Skiniotis, William I Weis, Roger K Sunahara, and Brian K Kobilka. Crystal structure of the $\beta 2$ adrenergic receptor-Gs protein complex. *Nature*, 477(7366):549–555, July 2011.

[65] K Gunasekaran, Buyong Ma, and Ruth Nussinov. Is allostery an intrinsic property of all dynamic proteins? *Proteins*, 57(3):433–443, November 2004.

[66] Philip A. Romero and Frances H. Arnold. Exploring protein fitness landscapes by directed evolution. *Nature Reviews Molecular Cell Biology*, 10(12):866–876, December 2009.

[67] Merijn L M Salverda, J Arjan G M De Visser, and Miriam Barlow. Natural evolution of TEM-1 β-lactamase: experimental reconstruction and clinical relevance. *FEMS microbiology reviews*, 34(6):1015–1036, November 2010.

[68] Michelle R. Arkin and James A. Wells. Small-molecule inhibitors of protein–protein interactions: progressing towards the dream. *Nature Reviews Drug Discovery*, 3(4):301–317, April 2004.

[69] Anthony Ivetac and J. Andrew McCammon. Mapping the druggable allosteric space of g-protein coupled receptors: a fragment-based molecular dynamics approach: computational mapping of novel druggable sites on gpcrs. *Chemical Biology & Drug Design*, pages no–no, July 2010.

[70] Jeanne A Hardy and James A Wells. Searching for new allosteric sites in enzymes. *Current Opinion in Structural Biology*, 14(6):706–715, December 2004.

[71] James R Horn and Brian K Shoichet. Allosteric Inhibition Through Core Disruption. *Journal of Molecular Biology*, 336(5):1283–1291, March 2004.

[72] M. R. Arkin, M. Randal, W. L. DeLano, J. Hyde, T. N. Luong, J. D. Oslob, D. R. Raphael, L. Taylor, J. Wang, R. S. McDowell, J. A. Wells, and A. C. Braisted. Binding of small molecules to an adaptive protein-protein interface. *Proceedings of the National Academy of Sciences*, 100(4):1603–1608, February 2003.

[73] Jonathan M. Ostrem, Ulf Peters, Martin L. Sos, James A. Wells, and Kevan M. Shokat. K-Ras(G12c) inhibitors allosterically control GTP affinity and effector interactions. *Nature*, 503(7477):548–551, November 2013.

[74] Sandor Vajda, Dmitri Beglov, Amanda E Wakefield, Megan Egbert, and Adrian Whitty. Cryptic binding sites on proteins: definition, detection, and druggability. *Current opinion in chemical biology*, 44:1–8, May 2018.

[75] Julie R. Schames, Richard H. Henchman, Jay S. Siegel, Christoph A. Sotriffer, Haihong Ni, and J. Andrew McCammon. Discovery of a novel binding trench in hiv integrase. *Journal of Medicinal Chemistry*, 47(8):1879–1881, April 2004.

[76] Patrick A. Frantom, Hui-Min Zhang, Mark R. Emmett, Alan G. Marshall, and John S. Blanchard. Mapping of the allosteric network in the regulation of -isopropylmalate synthase from *mycobacterium tuberculosis* by the feedback inhibitor l -leucine: solution-phase h/d exchange monitored by ft-icr mass spectrometry. *Biochemistry*, 48(31):7457–7464, August 2009.

[77] Gregory Manley and J. Patrick Loria. NMR insights into protein allostery. *Archives of Biochemistry and Biophysics*, 519(2):223–231, March 2012.

[78] S W Lockless and R Ranganathan. Evolutionarily conserved pathways of energetic connectivity in protein families. *Science*, 286(5438):295–299, October 1999.

[79] Victoria A Feher, Jacob D Durrant, Adam T Van Wart, and Rommie E Amaro. Computational approaches to mapping allosteric pathways. 25:98–103, April 2014.

[80] Joe G Greener and Michael Je Sternberg. Structure-based prediction of protein allostery. 50:1–8, October 2017.

[81] Toshiko Ichiye and Martin Karplus. Collective motions in proteins: A covariance analysis of atomic fluctuations in molecular dynamics and normal mode simulations. *Proteins*, 11(3):205–217, November 1991.

[82] Christopher L McClendon, Gregory Friedland, David L Mobley, Homeira Amirkhani, and Matthew P Jacobson. Quantifying Correlations Between Allosteric Sites in Thermodynamic Ensembles. *Journal of Chemical Theory and Computation*, 5(9):2486–2502, September 2009.

[83] A Cooper and D T Dryden. Allostery without conformational change. A plausible model. *European biophysics journal : EBJ*, 11(2):103–109, 1984.

[84] Nataliya Popovych, Shangjin Sun, Richard H Ebright, and Charalampos G Kalodimos. Dynamically driven protein allostery. *Nature Structural & Molecular Biology*, 13(9):831–838, September 2006.

[85] Milo M Lin. Timing Correlations in Proteins Predict Functional Modules and Dynamic Allostery. *Journal of the American Chemical Society*, page jacs.5b08814, April 2016.