NORTHWESTERN UNIVERSITY

Network Analysis of Protein Dynamics

A DISSERTATION

SUBMITTED TO THE GRADUATE SCHOOL
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

for the degree

DOCTOR OF PHILOSOPHY

Field of Mechanical Engineering

By

Jenny Liu

EVANSTON, ILLINOIS

September 2021

ABSTRACT

Proteins and many other systems are often conceptualized as networks to access analysis methods from the field of network science. Several approaches use molecular dynamics (MD) simulations of proteins to construct networks using correlational statistics. However, in the field of network science, a well-established approach for network construction is solving the inverse problem for a network that can produce the observed correlations.

We apply this inverse approach to three adhesion proteins - FimH and Siglec-8, and the SARS-CoV-2 spike protein - to identify networks that are distinct from correlation networks and instead resemble a contact map. Specifically, we use the inverse of the covariance matrix for backbone and dihedral angles. We select dihedral angles as a system of internal coordinates over external Cartesian coordinates to avoid potential distortions from structure alignment steps for proteins with hinged motion.

While more computationally expensive, solving the inverse problem can remove transitive correlations to produce networks that are robust among replicate MD simulations and that have physically interpretable interactions. In the inverse covariance networks, covalent interactions are stronger than hydrogen-bonds and non-bonding interactions. This pattern is not present in correlation networks. Moreover, backbone-backbone interactions dominate the inverse covariance networks, while interactions between sidechains dominate the correlation networks. Due to the differences in the networks constructed by correlation and by solving the inverse problem, there are also differences for comparing network edge strengths, topological properties, and community structures.

ACKNOWLEDGEMENTS

I would like to thank the teams of people who supported me during this process. First, I would like to thank my advisors, Luis Amaral and Sinan Keten, for their mentorship, teaching, and support during these years of research. Their scientific and non-scientific advice, generosity with their time, and experience in their fields made this research possible and a mostly fun experience. I am deeply grateful for their willingness to collaborate with each other. I also thank Michelle Driscoll for her guidance as part of the thesis committee, and Oluwaseyi Balogun for being on my qualifying committee and spending so much time trying to teach me basic mechanical engineering concepts.

Second, this research was done in the invigorating and exciting research environments produced by my labmates in both groups. I would like to thank Elizabeth, Martha, Kerim, Zhaoxu, Nitin, and Xinyan in the Keten group and Martin, Luiz, Helio, Julia, Meagan, Sophia, Sarah, Wei, Adam, and Peter in the Amaral group. In addition, I received a lot of administrative help from Justyna, Pat, Sinta, and Hillary. The MSTP Office has also provided a lot of help with writing, career development, and mentorship, especially Melissa, Jayms, Lindsay, and Joyce.

This project received years of generous support from the Northwestern Quest High Performance Computing Cluster and a lot of help from the NUIT RCS team. We also received support from the Department of Defense High Performance Computing Modernization Program. It was funded by the Paul and Daisy Soros Fellowship, the Achievement Rewards for College Scientists (ARCS) Illinois chapter, the NICO Data Science Initiative Intersectional Fellowship, and the National Institute of General Medical Sciences T32GM008152 for the MSTP program.

Lastly, and most importantly, I also thank my friends and family for their support and encouragement.

Glossary

FimH fimbrial adhesion protein H, protein found on the tip of hair-like structures that enable *Escherichia coli* bacteria to adhere to the sugar mannose on the urinary tract, a lectin.

graph-tool python library for analyzing networks, developed by Tiago Peixoto at Central European University.

Mathematics What mathematicians do.

MDAnalysis python library for analyzing molecular dynamics simulations, developed by Oliver Beckstein's group at Arizona State University.

MDTraj python library for analyzing molecular dynamics simulations, developed by Robert McGibbon in Vijay Pande's group at Stanford University.

Molecular dynamics simulations Computer simulations that take a collection of atoms and generates a trajectory of their movement over time, abbreviated MD simulations.

NAMD Nanoscale Molecular Dynamics, simulation engine developed by Klaus Schulten's group at the University of Illinois Urbana Champaign.

NetworkX python library for analyzing networks.

SARS-CoV-2 Severe Acute Respiratory Syndrome Coronavirus 2, originally 2019-nCoV, virus that caused the covid19 pandemic.

Siglec-8 sialic acid-binding immunoglobulin-like lectin, protein found on the surface of immune cells (specifically eosinophils and and mast cells) that recognize sialylated and sulfated sugars on the surface of human cells in order to prevent allergy to the self.

VMD Visual Molecular Dynamics, a software program for visualizing and analyzing molecular dynamics simulations, developed by Klaus Schulten's group at the University of Illinois Urbana Champaign.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1
INTRODUCTION AND BACKGROUND

1.1 Emergent properties of protein motion

How do the different parts of a protein move? Are some parts floppy, or more rigid? While many drugs target proteins based on static structure, such as the shape of the binding pocket, a better characterization of protein motion makes it possible to target regions based on dynamics [1], [2].

1.1.1 Molecular dynamics simulations

In order to characterize protein motion, scientists turn to computer simulations called *molecular dynamics (MD) simulations* [3], [4]. MD simulations provide spatiotemporal resolutions that would be difficult-to-impossible to obtain experimentally, depending on the modality [3]. It is possible to track the motion of every atom and make a movie of protein dynamics. Using MD simulations to characterize proteins is becoming increasingly more common as computational power grows [5]. Protein MD simulations have been performed for different types of proteins, including enzymes, receptors, and structural proteins. The proteins selected have ranged from oligopeptides to large systems, such as viral capsids. Simulation timescales range from picoseconds to about a millisecond [6].

In the Keten group, simulations are usually on the order of $\sim 10 - 100$ nanoseconds with $1 - 10$ picosecond resolution. The picosecond-to-nanosecond timescale is comparable to sidechain flipping and loop motions observed by nuclear magnetic resonance (NMR) experiments [4]. While our simulations are too short to sample biologically relevant timescales, such as conformational changes [4], we can still compare the difference in dynamics for different protein states. For example, we can see if the same protein region is floppier in one state than in another. Moreover, MD simulations do not completely recapitulate real protein dynamics, and the model assumptions are continuing to improve. For example, most simulations in the Keten group use a simplified model of water that introduces inaccuracy. (See Ch. 2 for a more in-depth discussion of how protein dynamics work.) As a result, we focus on methods to analyze MD trajectory data, which will hopefully remain relevant as models improve.

1.1.2 Analysis of MD simulations

Historically, a common framework for analyzing protein MD simulation was to focus on regions identified as important for protein function based on experimental work, such as studying the impact of mutations on function. However, it is increasingly possible to generate large MD simulation datasets for proteins with known structure, but poorly characterized function. While still valuable, human insight may not have the capacity to analyze the growing amount of MD trajectory data. The amount of data will likely continue to grow with computational power and the number of proteins with solved or accurately predicted protein structures. Excitingly, Google's AlphaFold can now predict protein structure with accuracy comparable to experimental approaches and can be combined with experimental data to solve particularly complex structures [7].

Developing automated approaches for analyzing MD trajectory data is also timely because increasing amounts of data by simulation groups are now available in depositories. Data depositories such as NOMAD and Dryad are growing due to efforts to improve reproducibility and open-source science [8]. There are also dedicated protein MD simulation depositories [8] such as MoDEL [9] and Dynameomics [10] that seek to identify motifs and phenotypes based on dynamics, in parallel with structural characterization. During the international crisis caused by the covid19 pandemic, MolSSI [6] provided a list of MD trajectories available for analysis. This included unprecedented millisecond-timescale data from distributed computing across people's personal devices on Folding@Home platform, greater than the combined computing power of the top 100 supercomputers [6].

As a result, there is a need for automation, as some MD datasets are already beyond the capabilities of individual humans to examine visually. Due to the increasing automation of protein structure determination and MD simulation production, this need for automated analysis methods will continue to grow. Moreover, automated approaches may be less biased and may identify important regions that were previously overlooked.

1.1.3 Network analysis of protein dynamics

Instead of focusing on one or two identified regions, what would we find if we considered the dynamics of the protein as a whole and looked for emergent properties? A hierarchy of properties resulting in complex behavior at higher levels has been found in a variety of biological systems [11], [12]. One approach that has been previously applied to proteins is network analysis. To abstract a protein as a network, each node is an atom (or group of atoms), and each edge is an interaction (e.g. within a cutoff distance to describe physical contact) [13].

It is useful to abstract proteins as networks because there are a variety of well-developed analysis techniques used in the field of network science [14] that may provide leverage for understanding proteins. Introducing mutations, chemical modifications to an amino acid, or a ligand may be analogous to modifying edge and node properties. The number, length, and redundancy of paths connecting one part of the network to another may be applicable to studying allostery, where one protein region regulates another, distant region. The difference in the pattern of edges, that are more interconnected in some regions and others, hints at the existence of communities. These communities may describe protein regions that move together, which is important both in allostery and in identifying the hierarchical building blocks for protein dynamics.

There is a decades-old historical basis for thinking of proteins as networks. Three commonly used methods used to construct networks from protein MD simulations are mutual information of dihedral angles, dynamic cross-correlation, and elastic network models. These methods are described in more detail in Ch. 2.

Looking at movies of protein MD simulations, some protein regions appear to move together, which has been described as correlated motion [15]. Thus, networks can be defined directly by setting the edges equal to correlative measures. The mutual information of dihedral angles captures nonlinear correlations of protein parts that can rotate around the axis of a bond [15]. Instead of focusing on bonds, it is also possible to use the Cartesian positions, focusing on the vector of fluctuations away from an equilibrium position. Dynamic cross-correlation measures the alignment of these fluctuations [16]. Instead of using correlative measures directly, it is also possible to solve for an underlying model that can give rise to the observed correlations. Elastic network models are a popular type of underlying model [17], [18].

While these methods have provided valuable, each has its own limitations that are described in more detail in Ch. 2. The elastic network model and the dynamic cross-correlation methods both use alignment procedures to calculate fluctuations relative to the equilibrium. The "artifacts" introduced by alignment has also been well-characterized [19]. In the mutual information method, the use of an internal coordinate system of dihedral angles neatly sidesteps the alignment step.

Conceptually, there is clearly a difference between edges identified by a correlative measure, i.e. mutual information or dynamic cross-correlation, and by an underlying model, i.e. an elastic network model. In the field of network science, solving the inverse problem for a system that can produce the observed dynamics is well-established, while the problems with using correlative

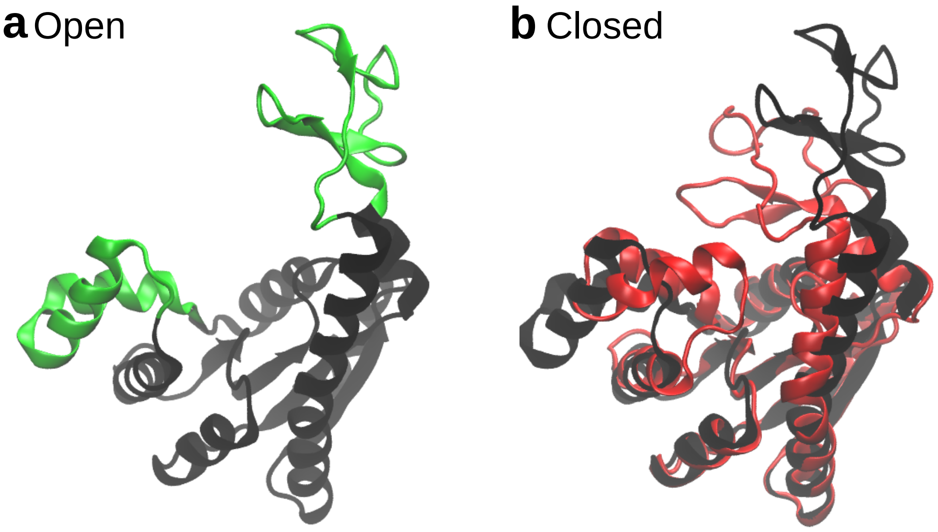

Figure 1.1: **Adenylate kinase can be described as sub-structures.** **a**, The core (black) and lids (greens) of adenylate kinase in the open state (PDBID 4AKE). **b**, Structure of the closed state (PDBID 1AKE, red) overlaid against the open state (black).

measures as a substitute for the structure are well-characterized [20].

As a result, these commonly-used methods often yield different network constructions from the same MD simulation dataset. It is not clear how to select a network for subsequent analysis.

1.1.4 Community detection on protein networks

Another analysis step that benefits from tools developed in network science is community detection, which abstracts proteins as regions larger than atoms, bonds, or amino acids. This is analogous to coarse-graining, an approach used to reduce degrees of freedom up speed up MD simulations [3]. Proteins have hierarchical structure: secondary structure motifs, such as helices and β-sheets, are combined in super-secondary structures and the “architecture” level of structure classification [21], such as β-sandwiches.

Identifying protein regions that move together is a useful step for characterizing protein MD simulations that relates dynamics back to structure and forwards to function. A model protein for studying dynamics is adenylate kinase 1.1 that resembles a box-like “core” with two flap-like “lids” [22]. The motion of each lid region relative to the core region describes ligand-binding [22]. In the protein folding process that starts from a chain of amino acids and ends with a structured, enzymatically active adenylate kinase protein, each lid region is thought to fold noncooperatively from the core region [23].

Moreover, the core is actually a collection of stiff regions connected by flexible hinges. The faster, picosecond-to-nanosecond timescale fluctuations of the hinge dynamics facilitate rearrangements that enable the slower, microsecond timescale motion of the flaps, by restricting or freeing space available for motion [22]. This indicates a hierarchy of protein structure related to dynamics at a hierarchy of timescales [22]. Functionally, these hinge motions are more restricted in adenylate kinase found in *Archaea* that grow at extremely high temperatures, explaining its high thermal stability [22]. The example of adenylate kinase illustrates the utility of using dynamics data to abstract the protein structure as a hierarchy of regions.

Similar to the problem of network construction, it is not clear how we should go about detecting protein regions using a network analysis approach called community detection [24]. A variety of approaches have been applied to proteins [25]–[27], and different approaches have identified different partitions for the same protein [26], [28], [29]. In Ch. 2, we describe a community detection method used in network science that avoids some of the pitfalls of methods commonly used in protein science. This approach, called stochastic block modeling, avoids user-specified parameters for the number and size of communities and can detect small communities on large networks [30], [31]. We choose this approach to begin our explorations of community detection in proteins in Ch. 5.

1.2 Scope and Objective

Can the process of constructing networks and detecting communities be made more parsimonious? We will strive for as little user-provided information as possible, avoiding information about previously-identified interactions, regions, secondary structure, domains, etc. Instead, we will focus on getting as much information as possible from automated analysis of MD trajectory data. This supports our goal to identify emergent properties from protein dynamics in an automated fashion. We will use develop and validate our tools on FimH and then apply them to two other proteins, Siglec-8 and the SARS-CoV-2 spike protein, which will be described in more detail below. For more details about our computational approach, see Ch. 2.

1.3 Protein systems

1.3.1 Bacterial FimH

We began with a well-characterized protein, FimH, found on bacteria causing urinary tract infections (Fig. 1.2a. When urine flows, these bacteria stick to human cells with hair-like structures called fimbriae, like the barbs of a cocklebur. The end of the fimbriae close to the bacteria is called FimA, and the part that sticks to human cells is called FimH. FimH has a lectin domain that binds a sugar (mannose) on the surface of human cells, and a pilin domain that connects to the rest of the fimbriae [32]. As a result, FimH is a drug target for preventing urinary tract infections, rather than treating with antibiotics afterwards to kill the bacteria [32].

Beyond its medical importance, FimH has interesting mechanical properties. When urine flows, shear flow pulls at the bacteria-human attachment, stretching FimH [33], [34]. Tensile force is thought to separate the two domains of FimH and then trigger a conformational change in the lectin domain. This changes the shape of the lectin domain's binding pocket, increasing

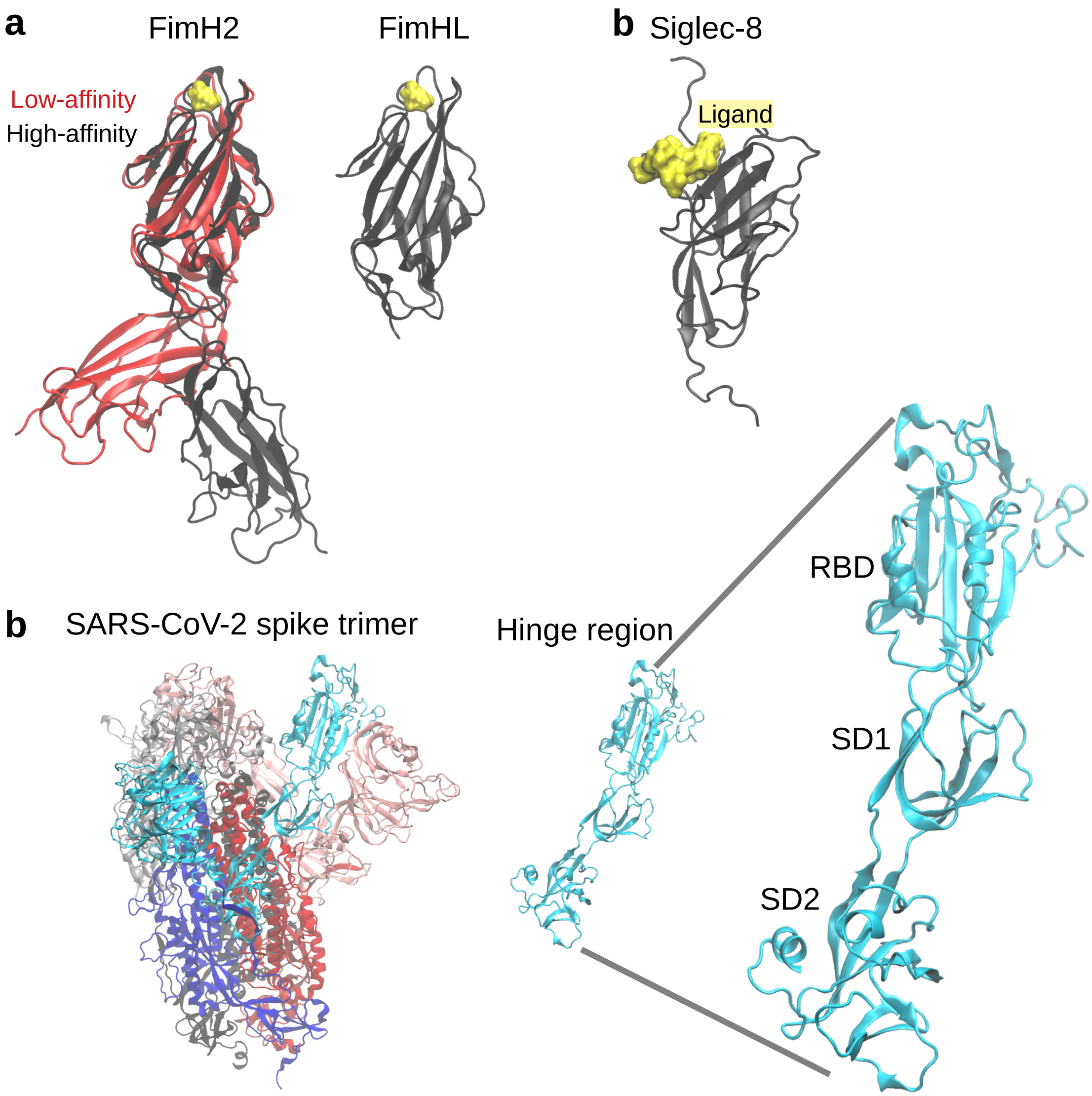

Figure 1.2: **Proteins studied. a**, The bacterial adhesin FimH2 in the full-length two-domain structure with low affinity for ligand (red) and with high affinity (black). The lectin domain of wild-type FimH has high affinity for ligand. **b**, The human immune adhesin Siglec-8 lectin domain in the holo (ligand-bound) state (PDBID 2N7B). **c**, The SARS-CoV-2 spike protein with three rotationally-symmetric protomers (red, blue, and black). The darker shades indicate the RBD involved in viral adhesion. The lighter shades indicate a stalk-like region responsible for viral entry into human cells. The medium shades indicate other protein regions. For one protomer, we show the RBD, SD1, and SD2.

its affinity for the mannose found on human cells [33]–[35]. Ultimately, the ability to bind with higher affinity in the presence of tensile force, or catch-bond behavior, allows the bacteria to stay bound during urine flow and retain mobility otherwise [32]. FimH's interesting mechanical ability is analogous to a children's toy, the Chinese finger trap, that wraps around two fingers and tightens when pulled [36], [37].

The biophysical mechanism underlying the catch-bond behavior is thought to be conformational allostery. Allostery means that a change in one part of the protein affects the function of another, distant part. In FimH, the interdomain region that is stretched is the allosteric site, which regulates the binding pocket. Conformational allostery means there is a marked change in protein structure, in contrast to dynamic allostery, where the structure does not change (much) [38]. Thus, FimH has interesting medical, mechanical, biophysical properties that makes it a model protein for studying conformational allostery [39].

1.3.2 Human Siglec-8

After studying FimH, we then search the Protein Data Bank and identified other proteins with similar structure to study, and one happened to be studied at Northwestern University. Like FimH, Siglec-8 is a lectin, a protein that binds sugar (Fig. 1.2b). Both have an immunoglobulin-like fold with two β-sheets arranged in a β-sandwich [33], [40]. However, Siglec-8 is a human protein found on immune cells and binds sugars on other human cells. Since these sugars indicate the cells are 'self' and not foreign intruders, Siglec-8 plays a role in the immune system's self-recognition and promotes immune tolerance [41]. Binding Siglec-8 causes the death or inactivation of inappropriately over-active immune cells [42], [43]. As a result, Siglec-8 is a potential drug target for treating allergic asthma and rhinitis

1.3.3 SARS-CoV-2 virus spike protein

During the covid19 pandemic, we also studied the spike protein of the SARS-CoV-2 virus. The spike protein allows the virus to adhere to and then enter human cells, which is the first step in infection. We focused on the adhesion step and the spike protein's receptor binding domain (RBD), which recognizes a protein receptor on human cells [44]. The RBD only has one central β-sheet [45], [46] unlike FimH and Siglec-8 which both have two. Moreover, the RBD could not be modeled alone, unlike the lectin domain of FimH or Siglec-8. MD simulations of the RBD by itself became unstable at the interdomain region. As a result, we modeled the RBD along with the domain below it, sub-domain 1 (SD1), as well as the one below that, sub-domain 2 (SD2) [45]. By choosing the places to truncate the model carefully to avoid instability, we were able to model this section of the spike protein separately.

We were also interested in the motion of the RBD relative to the entire spike protein. The spike protein is a trimer, with three rotationally-symmetric units, each with one RBD. In experimental studies, the spike proteins on the viral surface can have zero, one, two, or all three RBDs pointing 'up' and available for binding to the surface of human cells. It has been hypothesized that when the RBD is not available, the virus is more hidden from the immune system, since some antibodies only recognize the 'up' state[47]. However, other antibodies recognize the 'down' state, while others recognize different protein regions [47]. Since the spike protein is large, we were fortunate that the emergency nature of the covid19 pandemic led to the Bowman group [6] and others making their MD simulation data available online through MolSSI [6].

1.4 Thesis layout

Chapter 2 provides an overview of our computational approaches, with an emphasis on analyzing synthetic data generated from toy models. In these toy models, the network structures are known. We provide visualizations showing that the inverse of the covariance matrix can recover the original network, while correlative measures such as the covariance and the mutual information cannot. While useful as an illustration, it is not intended as a rigorous mathematical derivation or for benchmarking network construction methods. Instead, we build on the large body of work that has already been completed by others in the network science field.

In Chapter 2, we describe our rationale for using the inverse covariance matrix of dihedral angles, an internal coordinate system that requires circular statistics. In *Chapter 3*, we use circular statistics to describe the dihedral dynamics of a protein, FimH. We focus on differences in dynamics for FimH in different functional states.

In *Chapter 4*, we actually apply the inverse covariance matrix to dihedral dynamics. We further show the advantages of network construction using the inverse covariance of dihedral angles, rather than the covariance, correlation, or mutual information matrices. In addition to robustness and interpretability, the inverse covariance matrix shows a hierarchy of interaction strengths that resembles the qualitatively different types of interactions found in a protein. We also provide examples of identifying protein regions of interest for FimH, Siglec-8, and a component of the SARS-CoV-2 spike protein.

Chapter 5 summarizes our analysis pipeline, from dihedral dynamics to networks reconstruction to identifying protein regions of interest. Using the interactions we identified from dynamics in Ch. 4, we apply a community detection method from network science to automatically partition the protein into regions. Protein community detection may be useful for identifying emergent organizational structure, as well as reducing proteins into simpler components for coarse-grained simulations. We also summarize the limitations of our approach, and how some of these limitations may be addressed in the future.

CHAPTER 2
COMPUTATIONAL METHODS

In the Introduction, we described *molecular dynamics (MD) simulations* as a method to study protein dynamics. After briefly outlining how MD simulations are produced, we describe how we were inspired by the curse of dimensionality to develop simpler synthetic systems. We used the simpler systems to gain intuition about analyzing dynamics with tools from network analysis. We then returned to analyzing proteins using these tools. We focus on using dynamics to infer structural relationships, and then analyzing structure to identify regions that tend to move together.

2.1 MD simulations to study protein motion

2.1.1 MD simulations

There are many flavors of MD simulations that are suitable for answering different questions, such as extremely large systems, long timescales, and exploring rare events [3], [4], [48]. We focus on all-atomistic MD simulations in water, which means we look at every atom in the protein and a shell of water around it.

To perform MD simulations, the main ingredients are a protein structure, the simulation engine, and the force field. The protein structure gives the initial positions for every atom. The positions of atoms that could not be solved experimentally are guessed using tools such as CHARMM-GUI [49] or VMD [50]. The simulation engine performs the calculations that move the atoms over time. At every time step, it tabulates the forces each atom experiences from different types of interactions with other atoms, e.g. covalent, hydrogen bonding, van der Waals, etc. We use

NAMD [51] because it is free, has good parallelization, and others in the Keten lab had used it before. Other good options are GROMACS [52] and OPENMM [53], which are also free, and AMBER [54], for which Northwestern has an institutional license. The choice of force field defines the preferred shape of components (topologies) and the energetic penalty for moving away from those shapes (parameters). The force field also determines the options available for modeling water. The Keten group usually uses the CHARMM36 force field with the TIP3P water model [55]. While the commonly-used TIP3P water model's rigidity improves computational performance, this simplification does limit accuracy and affects protein dynamics [56].

This is an extremely simplified overview to highlight the ideas of a preferred structure, some energetic cost for less preferred structures, and using simulations to observe which structures can be realistically explored. For more detailed procedures see the Methods section of Chapter 3 and other papers by the Keten group.

2.1.2 *In silico* manipulations

The ~ 100 nanosecond timescale of protein MD simulations typically used in the Keten lab is on the order of loop and hinge motions [4]. While it is possible to study faster processes such as sidechain fluctuations, we cannot sample slower processes such as conformational changes. As a result, we only observe the impact of the manipulations we perform on these fast dynamics, but we cannot predict how these changes affect the global structure. However, these limitations have the benefit of keeping our *in silico* perturbations localized in space. For example, in reality, the removal of a disulfide bond may affect the conformation of the protein, but we can use it like a thought experiment to detect local changes in dynamics.

2.2 Network inference from protein dynamics

One difficulty with developing methods for protein MD simulations is the curse of dimensionality. The lectin domain of FimH has $N = 158$ amino acids, or 2,360 atoms. We can pretend each amino acid is only one atom on the stiff backbone, the Cα atom, a commonly-used simplification. If we want to study allosteric motion, where components move in concert, we would have to consider nN interactions, where $n \sim 8$ is the number of neighbors. This is similar to a type of elastic network model, called the Gaussian network model, that treats interactions between all neighbors as isotropic linear springs.

In another type of elastic network model, called the *anisotropic network model (ANM)*, the springs are directional [17], [18]. An added benefit of this model is that it only requires us to track the motion of each Cα atom, reducing the multiplier from n to the three dimensions of Cartesian space. After calculating the $3N \times 3N$ covariance matrix, its inverse yields a matrix of the directional stiffnesses. This inverse of the covariance matrix is also called the Hessian matrix.

$$\underline{\underline{C}}^{-1}[3N \times 3N] \sim \underline{\underline{H}}[3N \times 3N] \tag{2.1}$$

Another method to examine concerted Cα motion in Cartesian space is the *dynamical cross-correlation (DCC)*. For a pair of two atoms i, j, DCC detects correlative motion as the dot product of a pair of atom displacements, $\Delta r_i, \Delta r_j$, normalized by the magnitude of each displacement. These calculations are over the ensemble average, $\langle\rangle_t$, i.e. over all timesteps, t, for a single MD simulation.

$$DCC_{i,j} = \frac{\langle \Delta r_i \cdot \Delta r_j \rangle_t}{\sqrt{\langle |\Delta r_i| \rangle}_t \sqrt{\langle |\Delta r_i| \rangle}_t} \tag{2.2}$$

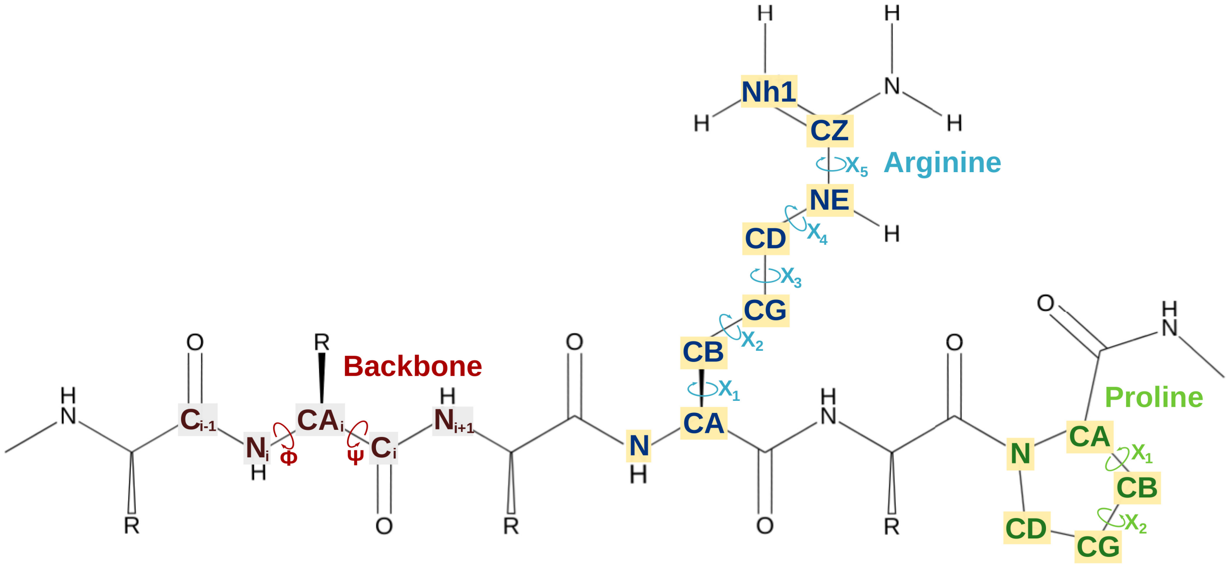

Figure 2.1: **Diagram of protein backbone and sidechain dihedral angles.** Atoms are labeled with the atom naming convention used by the Protein Data Bank, where $C\alpha$ is CA, $C\beta$ is CB, $C\gamma$ is CG, $C\delta$ is CD, etc. Backbone dihedral angles are defined using C_{i-1}-N_i-CA_i-C_i atoms for ϕ and N_i-CA_i-C_i-N_{i+1} atoms for ψ (red). Sidechain dihedral (or rotamer) angles χ_1-χ_5 are shown for Arg (blue). χ_1 is the closest rotamer to the backbone and is defined by N-CA-CB-CG. In contrast, the most distal rotamer, χ_5, is defined by CD-NE-NZ-Nh1. We further developed a system for reducing degrees of freedom by defining the custom pseudo-dihedral angle (υ) from N-CA and the last two atoms from the most distal dihedral angle For arginine (Arg), we use N-CA-Nz-Nh1. Similarly, we use N-CA-CG-CD for proline (Pro) because χ_1 is defined by N-CA-CB-CG, and χ_2 is defined by CA-CB-CG-CD (green).

Despite the advantages of ANM and DCC, 158×3 is still 474 degrees of freedom. Another way to simplify backbone motion is to describe each amino acid with two internal dihedral angles, ϕ and ψ on either side of the Cα (Fig. 2.1). This reduces the degrees of freedom to $158 \times 2 - 2 = 314$. We subtract two because each Cα on the two ends of the protein only has one neighbor. While this reduces the degrees of freedom, the periodicity of dihedral angles necessitates the use of circular statistics [19].

The methods to simplify protein motion mentioned so far have focused on the backbone and have completely neglected the floppy sidechains that branch off the stiff backbone. Analyzing sidechain motion is important because they interact with the ligand. Sidechains also interact with each other through strong covalent interactions like disulfide bonds, as well as non-covalent interactions like hydrogen bonds, salt bridges, polar interactions, and van der Waals interactions. Fortunately, like the stiff backbone, sidechains can also be simplified as one to five dihedral angles, depending on the length of the sidechain. Yet another approach to characterize protein motion uses mutual information to capture nonlinear correlations of both backbone and sidechain dihedrals [15].

2.2.1 How should we infer networks from protein dynamics?

Trying to develop methods to analyze the complicated dynamics of a system with hundreds of degrees of freedom is extremely challenging. It is even more challenging to try to develop network analysis methods when we are not sure what the network should look like, since different methods produce different networks.

Of the methods we briefly described, ANM and DCC yield qualitatively similar interaction patterns that resemble the residue-to-residue distance map of the protein [57]. However, both methods depend on displacements in Cartesian space, Δr_{ij}, which are calculated by aligning protein structures to a reference structure. This alignment step has been shown to produce dramatic artifacts for proteins with hinged motion, as well as more subtle artifacts for small, globular proteins [19], [58].

In contrast, the use of internal coordinates for dihedral angle dynamics avoids the alignment step and can more accurately identify motions observed experimentally [59]. Including both backbone and sidechain dihedral dynamics better reflects the complexity of real proteins, without increasing degrees of freedom as quickly as Cartesian descriptors. These beneficial features are found in the networks constructed from the mutual information of dihedral angles. However, network inference from correlative measures such as mutual information is also known to produce distorted views of the original network [20], [60].

Besides inferring network structure from dynamics, it is also possible to directly use the protein structures from the MD simulations. These usually take the form of residue-to-residue contact maps or interaction energy maps [13], [61]. Contact maps connect residues within a particular distance cutoff that can range from 4.5 [62] to 15 Å[17]. Similar, interaction energy maps connect residue pairs above a particular interaction energy threshold [62]. The enthalpic component of residue interaction energies is usually calculated without water, which affects Coulombic interactions and indirect couplings where residues are bridged by a water molecule [63], [64]. The entropic component is often neglected [65] or estimated from an elastic network approximation that uses an alignment step [63].

These considerations complicate network construction directly from protein structure. Fortunately, we know from ANMs that structural information is already encoded in the dynamics. The structural information can be retrieved from the inverse of the covariance matrix [17], [18]. The inverse of the covariance matrix is more accurate for network inference compared to using correlative measures [20], [60]. In addition, it is one of several approaches for network inference by solving the inverse problem for a network structure that can produce the observed dynamics [20], [60]. In the next section, we describe our efforts to build upon these advantages of using ANM for network inference, by replacing Cartesian coordinates with the internal coordinate system of dihedral angles.

2.3 Network inference from the inverse covariance matrix of dihedral angles

Using dihedral angles enables us to easily account for sidechain dynamics, and by using an internal coordinate system, we avoid the artifacts introduced by the alignment step. In real proteins, dihedral angles have steep energy wells (two for backbone dihedrals and three for sidechain dihedrals) in order to maintain the planarity of the peptide bond and to avoid steric hindrance by staggering bulky sidechains.

Let's start by making the inaccurate, but simplifying, assumption that each dihedral angle has a harmonic potential energy function centered around one preferred orientation. This is only reasonable for short, picosecond-timescales, which is usually faster than the timescale for backbone dihedral flipping [4]. Moreover, let's assume that dihedral angle pairs only have linear coupling, which neglects nonlinear interactions and many-body interactions. This might be reasonable because the covariance matrix of dihedral angles has been successfully used for principal component analysis to capture important motions involved in protein folding [58], [59]. Moreover, torsional network models (TNMs) of protein dynamics with coupled dihedral angles can recapitulate some aspects of protein motion [66]. However, unlike TNMs, we allow the coupling to vary among dihedral pairs.

We can describe these linearly coupled dihedral angles as the Hessian matrix, **H**, where H_{ij} is the coupling constant between a pair of dihedrals, i, j. A positive coupling indicates rotation in the same direction, like two windmills blown on by the same breeze. A negative coupling indicates rotation in opposite directions, like an eggbeater. Note that since almost all amino acids have two backbone dihedrals, and some have several sidechain dihedrals, the coupling between two amino acids is actually a submatrix. This makes sense because two amino acids may have a backbone hydrogen bond, and the side chains may also be involved in a salt bridge, disulfide bond, or other interactions.

The harmonic coupling assumption means that H_{ij} goes as $\frac{energy}{radians^2}$. Radians are a unitless measure, so H_{ij} has units $J = Nm$, but we find it easier to use J/rad^2. Some combination of angular displacements from equilibrium, θ, then produces the torque vector $f = -\mathbf{H}\theta$. The potential energy difference from that at equilibrium is $\Delta U = \int -f d\theta = \frac{1}{2}\theta^T\mathbf{H}\theta$.

We can also the use the Boltzmann relationship to describe the probability of a particular configuration in terms of angular displacements, θ, as $p[\theta] \sim exp[-\Delta U/k_bT]$. The exponent is $-\theta^T\mathbf{H}\theta/(2k_bT)$ and can be rearranged to $-\frac{1}{2}\theta^T(\frac{1}{k_bT}\mathbf{H})\theta$. This yields the probability distribution function in terms of **H** as

$$p[\theta] = exp[-\frac{1}{2}\theta^T(\frac{1}{k_bT}\mathbf{H})\theta] \tag{2.3}$$

By recognizing the form of a multivariate Gaussian distribution with covariance matrix **C** is

$$p[\theta] = exp[-\frac{1}{2}\theta^T\mathbf{C}^{-1}\theta] \tag{2.4}$$

we can see the inverse relationship between the Hessian and covariance matrices. This is the same as for ANMs (eq. 2.1), except that ANMs use Cartesian displacements.

$$\frac{1}{k_bT}\mathbf{H} = \mathbf{C}^{-1} \tag{2.5}$$

One advantage of this hand-wavy derivation is that we can set up a toy model to generate synthetic data by sampling a multi-variate Gaussian distribution with covariance matrix $\mathbf{C} = k_bT\mathbf{H}^{-1}$.

2.4 Toy models for studying network inference

Toy models that can generate synthetic data are useful for developing intuition for a system. As we are not mathematicians, these toy models helped us examine the covariance matrix and its inverse, which is well-established [20], [60], but new to us. This approach is also First, we use multi-variate Gaussians to study this relationship for non-circular statistics. Then, if we were mathematicians, we might have tried to do this for von Mises distributions, which are the circular random variable equivalent to Gaussians. Instead, we used spin models in order to explore continuous and discrete distributions of circular variables. It made sense at the time because we already knew that the dihedral angles did not follow Gaussian or von Mises distributions.

2.4.1 Gaussians

First, we study a network with edges indicate coupling by a linear spring in a model that does not account for the orientation of the edge. This network can be represented as an adjacency matrix, where each $A_{ij} > 0$ indicates a connecting edge (Fig. 2.2a). The adjacency matrix can be converted into a Laplacian matrix, $L_{i \neq j} = -A_{ij}$ and $L_{i=j} = \sum_{i<j} A_{ij}$. The Laplacian matrix can be inverted to generate the covariance matrix (Fig. 2.2a). This feature makes it easy to use the multi-variate Gaussian distribution to generate synthetic time-series data, where the displacement at each time-point is independent from the next. This synthetic data can be used to calculate sample covariance matrix, which can be inverted to recover the sample Laplacian and adjacency matrices (Fig. 2.2b).

To evaluate the accuracy of network inference, we compare the weights on the recovered adjacency matrix for true edges and non-edges (Fig. 2.2b). To quantify the comparison, we use the Kolmogorov-Smirnov (K-S) distance between the distributions of edges and non-edges. The

K-S distance between distributions with perfect separation is $KS = 1$. When $KS = 1$, we construct a modified K-S (mKS) distance by adding the distance between distributions, $mKS \doteq 1 + \frac{min[edge]-max[non-edge]}{mean[edge]-mean[non-edge]}$. Otherwise, $mKS \doteq KS$. Using the mKS distance, we show that the gap between true edges and non-edges narrows with smaller sample size (Fig. 2.2c). (Note that mKS is not a true distance. It is just helpful for monitoring gap size.)

Using this framework, we can also generate synthetic time-series from the same distribution, but with a temporal auto-correlation, τ, using a Monte Carlo process [67]. This makes it possible for us to decouple the effects of fluctuation size (k_bT) relative to system noise ($\zeta \sim N(0,1)$) and the effects of auto-correlation time (τ). The auto-correlation time constant is considered relative to the time-series length, $\mathcal{T}$. Using the same network, the example in Fig. 2.3 shows that lower k_bT/σ_ζ and shorter $\tau/\mathcal{T}$ decreases the accuracy of network inference. We then show the accuracy network inference by stepping through a grid of k_bT/σ_ζ and τ values in Fig. 2.4.

In molecular dynamics simulations, the the simulation temperature and the structural properties of the protein have complex interactions with each other to affect the fluctuation size, thermal noise, and auto-correlation. This can be seen as a natural consequence of temperature-dependent noise, instead of our contrived choice of constant σ_ζ. Nevertheless, this multivariate Gaussian toy model provides a sense of conditions where network inference is not possible. While not done here, it could be made more complicated by using multiple τ to make data quality more variable across the network, by adding a forcing function, or by substituting a more useful network structure.

2.4.2 Spin models

While the Gaussian multivariate model was useful for developing intuition, Gaussian random variables have domain $(-\infty, \infty)$, and dihedral angles are restricted to $(-\pi, \pi]$. As a result, we turned to Markov Chain Monte Carlo (MCMC) simulations of spin models on networks [68].

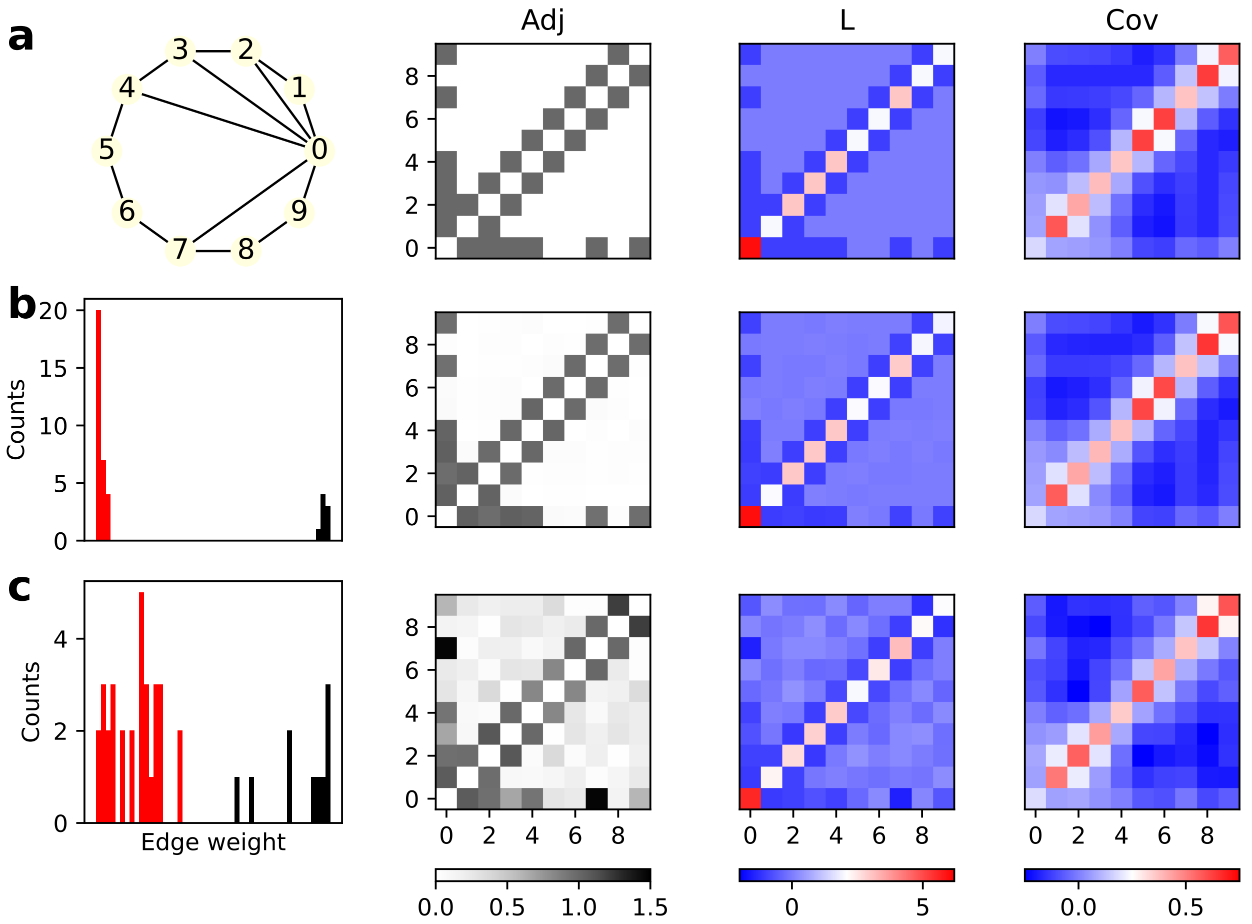

Figure 2.2: **Procedure for synthetic data generation and network inference using Gaussians.** **a**, The network with $N = 10$ nodes can be represented as an adjacency matrix and subsequently converted into an Laplacian matrix. The inverse of the Laplacian matrix is the covariance matrix, which is used to construct a multi-variate Gaussian distribution. **b**, We draw $\mathcal{T} = 100N^2$ independent samples from the multivariate Gaussian distribution. Data is used to construct a sample covariance matrix, which is inverted to recover the Laplacian and adjacency matrices. Using the recovered adjacency matrix, weights for true edges are shown in black, while those for non-edges are shown in red. $KS = 1$ because there is perfect separation, and $mKS = 1 + 0.9 = 1.9$. **c**, The same procedure for only $\mathcal{T} = N^2$ samples has $KS = 1$ and $mKS = 1.3$

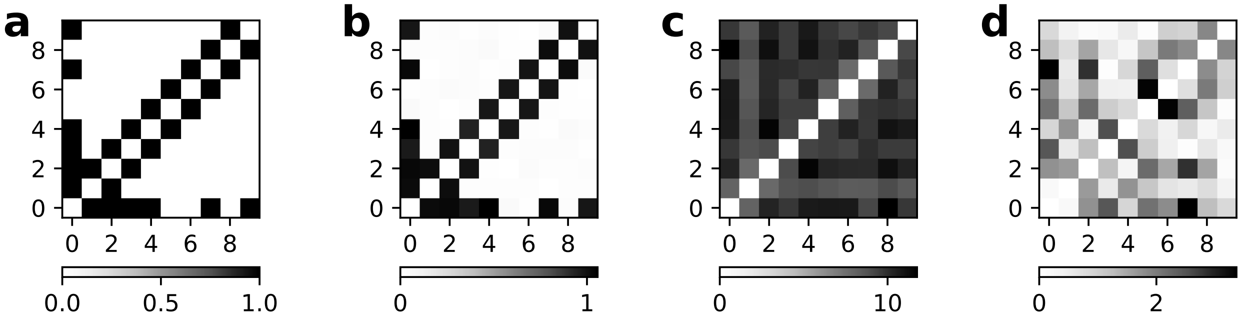

Figure 2.3: **Example of network inference with poor data quality.** Compared to the **a**, ground truth network, the **b**, network recovered under good conditions (no noise, no auto-correlation) is better than the networks recovered from data quality with **c**, low fluctuation size relative to noise ($k_bT/\sigma_\zeta = 10$ or **d**, less data relative to the auto-correlation length $\mathcal{T}/\tau = 10$. For ease of comparison, we use the same network and repeat panels a and b from Fig. 2.2. For each, we show the adjacency matrix.
.

The Ising model is an example of a discrete, 2-state spin model, where the spin can be in the "up" or "down" states. A pair of spins are in a low-energy state if they are aligned. In traditional Ising models of ferromagnetic materials, each spin is affected by neighboring spins on a lattice. Instead of the Ising model, we select for our spin models the Potts model with $q = 3$ discrete states and the XY model with continuous spin states from $(-\pi, \pi]$. We also replace the regularly-spaced lattice with a different network structure.

In MCMC simulations of spin models on networks, each node is a spin, ϕ. Each edge connects two nodes (i, j) and indicates the spins have interaction energy $E_{ij} = -J_{ij}cos[\phi_i, \phi_j$ where the edge weight $J_{ij} = 1$. At each MC step, we trialed a perturbation $|\phi| < \frac{\pi}{6}$ at a selected spin. A perturbation at node i would change the interaction energy with another node j to $E_{ij} =$

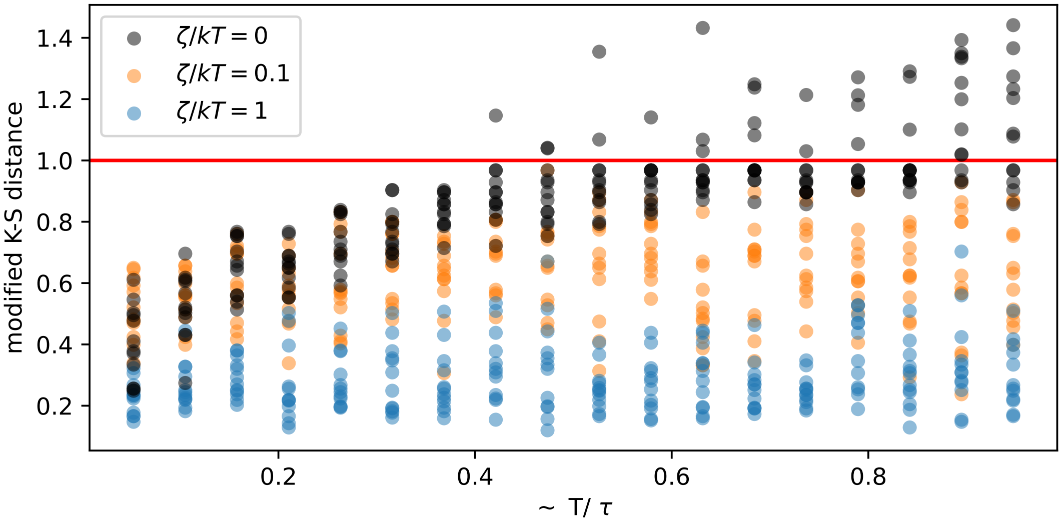

Figure 2.4: **Network inference accuracy depends on data quality.** We perform the network inference procedure for several replicates, varying noise relative to fluctuation size (σ_ζ/k_bT) and the amount of data relative to auto-correlation ($\mathcal{T}/\tau$). To assess accuracy, we use our modified K-S distance, which is higher than 1 (red line) when there is complete separation of the distribution of edges from non-edges. For incomplete separation, this is just the K-S distance. More noise reduces the accuracy of network inference, as shown by the lower mKS distances of the orange and blue data compared to the black data with no noise. Longer τ also limits accuracy by reducing the amount of data available, as shown by the lower mKS distances for low $\mathcal{T}/\tau$.

$-J_{ij}cos[\phi_i + \delta\phi_i - \phi_j]$. Thus, the change in energy would be

$$\begin{aligned} dE_{ij} &= -J_{ij}(cos[\phi_i + \delta\phi_i - \phi_j] - cos[\phi_i - \phi_j]) \\ &= 2J_{ij}sin[\frac{\delta\phi}{2}]sin[\phi_i + \frac{\delta\phi}{2} - \phi_j] \end{aligned} \tag{2.6}$$

The total change in energy after accounting for all nodes connected to i on the network is $dE_{tot} = 2\sum_j dE_{ij}$. The Metropolis Monte Carlo algorithm is a Markov Chain, where attempting a perturbation is equivalent to attempting to move from one configuration to another one. To ensure detailed balance in a Markov Chain, the probability of changing from configuration a to configuration b ($\Pi[a \rightarrow b]$), and the probability of being at a particular configuration (P_a), are related as $\Pi[a \rightarrow b]P_a = \Pi[b \rightarrow a]P_b$. This ensures the net flux is zero. This yields a calculatable value for the ratio of forward and backward transitions between the configuration states as related to the difference in energy, without requiring a calculation for the partition function $\mathcal{Z}$.

$$\begin{aligned} \frac{\Pi[b \rightarrow a]}{\Pi[a \rightarrow b]} &= \frac{P_a}{P_b} \\ &= \frac{exp[-E_a\beta]/\mathcal{Z}}{exp[-E_b\beta]/\mathcal{Z}} = exp[-(E_a - E_b)\beta] \end{aligned} \tag{2.7}$$

where the notation $\beta = (k_bT)^{-1}$ saves space. This means always accepting a favorable configuration change $a \rightarrow b$, i.e. $\Delta E = E_a - E_b \geq 0$. Unfavorable moves are accepted with probability $P[a \rightarrow b] = exp[-\Delta E\beta]$. This algorithm is biased towards configurations of the system near equilibrium, which are more common and relevant to calculating average behaviors, rather than rare configurations.

Our selection of nodes for applying perturbations is only semi-random. Every node is visited at an equal and consistent rate by forcing the selection to make “sweeps” of the whole network. Within each sweep, each node is visited only once, although in random order for each sweep. This seems to increase the equilibration speed of the system, although this may affect the calculation of auto-correlation. To reduce the effect, we only report spins every few sweeps when generating time-series data.

Using this synthetic data, we show that the inverse covariance matrix also recovers the network structure for discrete and continuous spin models (Fig. 2.5, even though the spin models differ from multivariate Gaussians. First, the discrete and continuous spins explore non-Gaussian distributions. The discrete model is multi-modal, and the discrete nature is analogous to extremely positive kurtosis. The continuous model is analogous to a uniform distribution. Moreover, unlike our multivariate Gaussian model, the coupling interactions are not harmonic, i.e. dE is not $\Delta\phi^2$. The relationship between the Boltzmann inversion using eq. 2.6 and the form of the multi-variate Gaussian probability distribution function in eq. 2.4 are not immediately obvious and beyond our current capabilities as non-mathematicians.

Lastly, our use of MCMC to generate synthetic data means that temperature affects the acceptance rate of a proposed perturbation, and thus fluctuation size. Temperature also affects auto-correlation, with an expected peak for τ at T_c, the critical point of the (high-order) phase transition. This is due to the long-range correlations expected during phase transitions. There is a temperature “Goldilocks zone” where network inference is most accurate (Fig. 2.6. At low temperatures, the fluctuations are too small to detect coupling. At higher temperatures, the fluctuations are too noisy. While higher temperatures $T \rightarrow T_c$ may lead to systems with long-range correlations, we find that the upper bound on temperature for accurate network reconstruction is quite a bit lower than T_c. On the other hand, we also find that the temperature range of the “Goldilocks zone” expands with longer simulation times and speculate that $\mathcal{T}/\tau$ may play a role in this system as well.

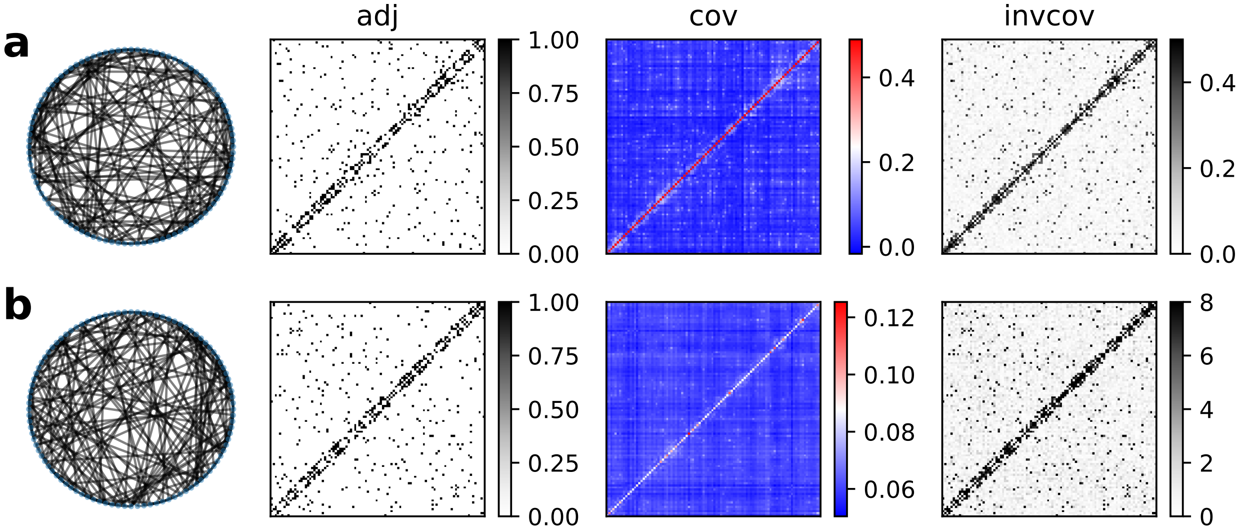

Figure 2.5: **Examples of network inference using synthetic data from spin models.** For **a**, a discrete $q = 3$ Potts model and for **b**, a continuous XY model, we show the adjacency matrix of the ground truth network, the covariance matrix, and the inverse of the covariance matrix. This demonstrates the inverse of the covariance matrix also applies to spin models.

Primarily, we used the spin model to test our code for network reconstruction using the inverse of the covariance matrix for circular variables. We also learned about the need for long simulations relative to the auto-correlation and the utility of trialing multiple temperatures. However, we did not quantify these effects due to time constraints and instead focused on analyzing proteins.

2.5 Limitations to the inverse covariance matrix

While the inverse of the covariance matrix is straightforward to compute using the Moore-Penrose pseudo-inverse, there are limitations due to the assumptions made and the amount of data required. Many systems do not follow Gaussian distributions and do not have spring-like coupling interactions. The number of unknowns scales as the square of the numbers of degrees of freedom. Nevertheless, the inverse of the covariance matrix is a useful, first-pass approach used in network science to infer structure [20].

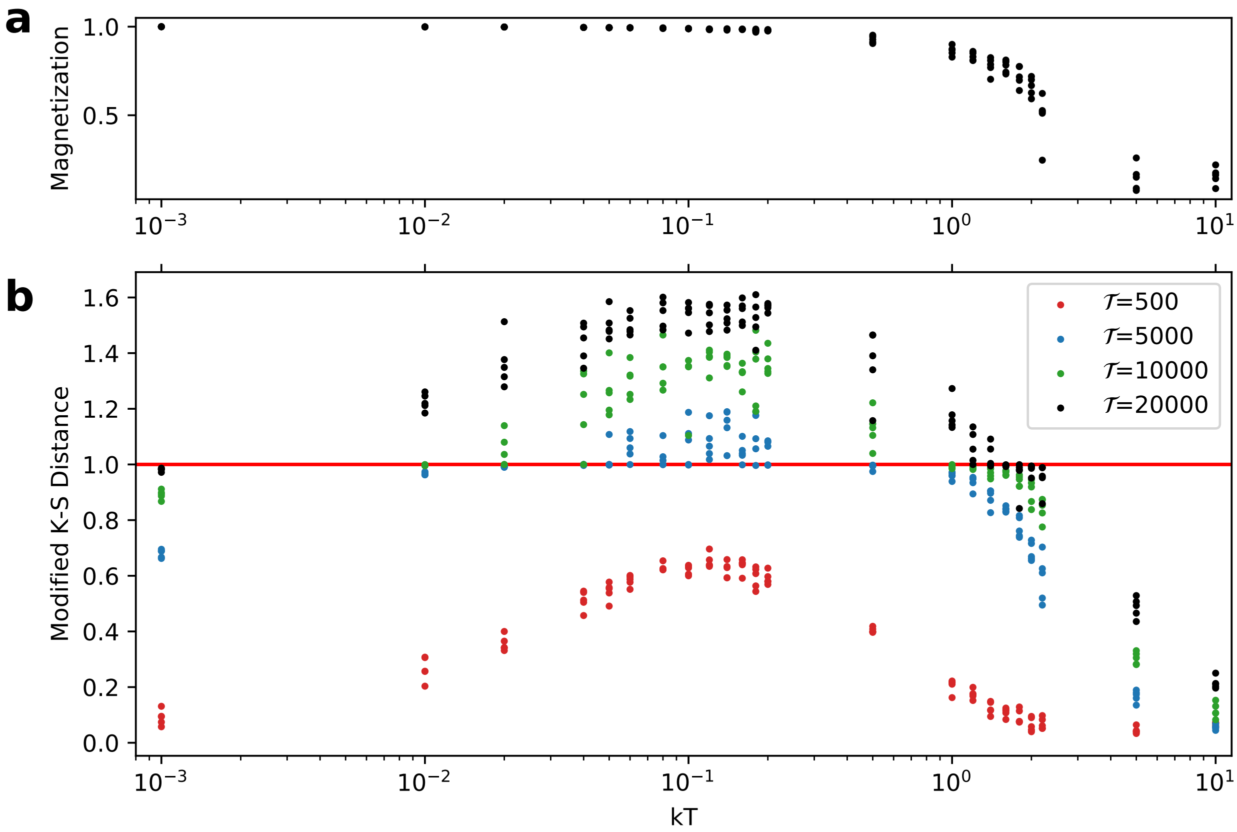

Figure 2.6: **Accuracy of network inference for spin models depends on temperature, which affects data quality.** WE show data for an XY spin model on a Watts-Strogatz small-world network produced by rewiring a nearest-neighbor network [68]. **a,** We performed MCMC simulations at multiple temperatures. We show magnetization as an observable describing the spin alignment. **b,** We next infer networks from the simulation data. To evaluate the accuracy of network inference, we calculate the modified K-S distance between the distributions of inverse covariance matrix weights for edges and for non-edges. We show the modified K-S distance as a function of temperature to approximate the "Goldilocks" zone for network reconstruction from the inverse of the covariance matrix. We draw a line where the modified K-S distance = 1, indicating perfect separation, or accurate network inference. At very low temperatures, fluctuations are too small; at very high temperatures, fluctuations are too big. The size of the "Goldilocks zone" also expands with larger sample size, shown in different colors. For very small sample sizes, network inference using the inverse covariance matrix is always inaccurate. We show data for an XY spin model.

We briefly mention a few alternative methods without an attempt at reviewing all methods described in more detail elsewhere [20], [60]. The approach used by Debbie Marks's group is based on the discrete states of the Potts model, requires less data by using a mean-field approximation, and has successfully been used to infer protein structure from sequence variations [69]. An approach developed by Jr-Shin Li's group avoids making assumptions about the nature of the coupling or the shape of the probability distribution function [65]. In a similar vein, reservoir computing is a machine learning approach that fits a model that is then used to predict the effects of perturbations at each input node, assuming the greatest impact will be felt at neighboring nodes [70]. Lastly, Tiago Peixoto uses a generative approach to identify a network structure that best describes the observed dynamics, i.e. the observed microstates for the Ising model on a network [60]. One consequence of Peixoto's choice of generative model as a type of stochastic block model (SBM) is that network inference can be "synergistically" combined with the detection of communities, or nodes with similar connectivity patterns. We will describe communities in more detail in the section below.

2.6 Network analysis

Our motivation for network inference from protein dynamics is the ability to apply network science tools. We will focus on strong interactions and on detecting communities, which are protein regions with coupled dynamics [26], [27], [71].

2.6.1 Edges

The direct output of network inference is the interaction strengths between two dihedral angles. We primarily care about the magnitude of interactions. The magnitude describes how strongly one protein component affects another. In contrast, the sign describes the direction of the spin coupling, or the direction the two dihedral angles move relative to each other. In the case of protein dynamics, there is no global rule that dihedral angles moving clockwise are beneficial, although there are likely specific applications, e.g. the direction of motion of a dihedral angle at a hinge affects closing and opening.

Since this is a fairly new method, we focus on identifying strong interactions. In addition, given two different states for a protein, we can compare each edge to identify interactions that are stronger in one state, or vice versa. Considering the difference for all edges also provides a way to compare the networks for each state. We discuss this in more detail in Ch. 4.

2.6.2 Community detection

Beyond individual edges, we are interested in emergent properties, such as protein regions with coupled motion. Identifying coupled regions is useful for understanding behavior such as allosteric regulation of a functional site from a distant site [27], [38], [65]

There are many algorithms available for community detection. Of these we select stochastic block modeling (SBM) [30], [31] as implemented by Peixoto [60]. SBM uses a generative model to produce networks according to rules defined by some latent variables. Peixoto's implementation uses Bayesian inference to relate the observed network to the ensemble of networks it is possible to generate from a set of parameters.

Given a network, A, that was generated from a parameter specifying there are $b = B$ communities and some other parameters, e, then Bayes' theorem for the probability of b given an observed network A is

$$
\begin{aligned}
P[b|A] = P[A|b,e]P[b,e]/P[A] &\sim P[A|b,e]P[b,e] \\
&= exp[-ln[A|b,e] - P[b,e]] \\
&= exp[-MDL]
\end{aligned}
\tag{2.8}
$$

where $P[b|A]$ is the posterior, $P[A|b,e]$ is the likelihood, and $P[b,e]$ is the prior. Maximizing the posterior is equal to minimizing the MDL, which is the minimum description length. The procedure is analogous to compressing the data required to describe the network. To minimize the description length, one can increase the likelihood, $P[A|b,e]$, or how likely it is to see the observed network, A, from the ensemble of networks generated with parameter b. This discourages very rare A. Since $P[A|b,e]$ improves as b, e become more complicated, the $-lnP[e,b]$ conveniently acts as a penalty term. For a much more mathematically correct derivation and useful details, such as why this approach is better than using information criteria to limit model complexity, please see Peixoto's book chapter [72].

As a result, it becomes unnecessary to specify the number and size of communities as parameters, or to step through a grid of parameter options and apply an information criterion to find the best fit. Moreover, selecting a hierarchical model also makes it possible to detect small communities on large networks [73]. This improves upon a problem with using non-hierarchical stochastic block modeling that is similar, but not exactly the same, as the "resolution limit" intrinsic to modularity optimization approaches for community detection [74].

CHAPTER 3
CONFORMATIONAL STABILITY OF THE BACTERIAL ADHESIN, FIMH, WITH AN INACTIVATING MUTATION

This was previously published: Liu J, Amaral LAN, Keten S. Conformational stability of the bacterial adhesin, FimH , with an inactivating mutation. Proteins. 2021;89(3):276-288.

3.1 Abstract

Allostery governing two conformational states is one of the proposed mechanisms for catch-bond behavior in adhesive proteins. In FimH, a catch-bond protein expressed by pathogenic bacteria, separation of two domains disrupts inhibition by the pilin domain. Thus, tensile force can induce a conformational change in the lectin domain, from an inactive state to an active state with high affinity. To better understand allosteric inhibition in two-domain FimH (H2 inactive), we use molecular dynamics simulations to study the lectin domain alone, which has high affinity (HL active), and also the lectin domain stabilized in the low-affinity conformation by an Arg-60-Pro mutation (HL mutant). Because ligand-binding induces an allostery-like conformational change in HL mutant, this more experimentally tractable version has been proposed as a "minimal model" for FimH. We find that HL mutant has larger backbone fluctuations than both H2 inactive and HL active, at the binding pocket and allosteric interdomain region. We use an internal coordinate system of dihedral angles to identify protein regions with differences in backbone and sidechain dynamics beyond the putative allosteric pathway sites. By characterizing HL mutant dynamics for the first time, we provide additional insight into the transmission of allosteric information across the lectin domain and build upon structural and thermodynamic data in the literature to further

support the use of HL mutant as a "minimal model." Understanding how to alter protein dynamics to prevent the allosteric conformational change may guide drug development to prevent infection by blocking FimH adhesion.

3.2 Introduction

The bacterial adhesin FimH is one of the most well-characterized model proteins for catch-bond behavior, in which tensile force paradoxically increases ligand affinity and lengthens the adhesion time [37], [75]. For many catch-bond proteins [37], including FimH, P-selectin [76], α-catenin [77] and the $\alpha\beta$ T-cell receptor [78], tensile force exposes previously buried regions and induces an allosteric conformational change [38], [79], [80]. A better understanding of the protein dynamics involved in the activation of FimH [81] could lead to treatments that target the allosteric site, which could prevent bacterial adhesion to the host during infections [82].

FimH is a 30 kDa two-domain bacterial adhesin found on the fimbrial tips of uropathogenic *E. coli* (UPEC) that binds to mannosylated ligands on urothelial cells [83]. Because it enables bacterial adhesion, which is one of the first steps in urinary tract infections, FimH is a critical virulence factor [83]. During urination, shear forces introduce tension into the FimH-mannose interaction to produce catch-bond behavior, causing FimH to bind more tightly and for longer duration [36], [37], [79], [81]. This catch-bond behavior distinguishes mannosylated ligands on the urinary tract from decoys and limits bacterial removal [83], [84]. Several glycomimetic drugs (mannosides) have been developed that competitively inhibit FimH and thus prevent bacterial adhesion [32], [85]. However, it is currently not known how to target the allosteric site(s) to block the conformational change.

The mechanism underlying catch-bond behavior involves an allosteric conformational change between an inactive state with low affinity and an active state with high affinity [37]. In the inactive state, the two domains of FimH are close together, which stabilizes the interdomain region between the ligand-binding lectin domain (HL) and the allosterically inhibitory pilin domain (HP) [34], [84], [86]. The transition to the active state is thought to occur after the ligand binds to the lectin domain and tensile force pulls the domains apart [87]. Exposing the interdomain region to water, which disrupts inhibition by the pilin domain, is thought to induce a conformational change in the lectin domain [35], [86]. The conformational change has been described by the width of the β-sandwich fold [33], [34], [84], as well as the putative allosteric pathway sites connecting the interdomain region to the binding pocket [35], [39], [79].

While two-domain FimH in the inactive conformation (H2 inactive) can undergo a force-induced, allosteric conformational change to the active state, a truncated protein consisting of the lectin domain alone (HL active) is constitutively active [37], [39], [86]. However, while the lectin domain with a single Arg-60-Pro amino acid mutation (HL mutant) is stabilized in the inactive conformation [79], HL mutant undergoes an allostery-like conformational change upon binding mannoside [39]. As a result, Rabbani *et al.*[39] have proposed HL mutant as a "minimal model" for FimH allostery that is smaller, consists of a single domain, and is more experimentally tractable than full-length FimH. Although the structure and function of the HL mutant have been well-characterized, its dynamics have not yet been investigated experimentally or computationally. Studying the dynamics of HL mutant will make it possible to identify protein regions with dynamical differences compared to H2 inactive, which can provide insight into the allosteric conformational change and help design additional mutations that lock HL mutant in the inactive state.

The impact of the Arg-60-Pro mutation on the structure and function of HL mutant has also been well-characterized. After Rodriguez *et al.*[79] selected the Arg-60-Pro mutation to energetically favor the inactive state from a set of trial mutations tested in RosettaDesign, Rabbani *et al.*[39]

then confirmed that the backbone structure of HL mutant matches the lectin domain of H2 inactive using crystallography and chemical shift mapping from ^{1}H-^{15}N-HSQC NMR spectroscopy.The structural similarity between HL mutant and H2 inactive suggests that the β-bulge has stronger allosteric coupling to the clamp segment than the interdomain regions, which are missing in HL mutant. Thus, different sites along the allosteric pathway may vary in their “coupling” strength to the binding pocket [35], [39]. It has also been demonstrated that HL mutant has more than 10 times higher ligand association and dissociation rate constants than H2 inactive, despite the structural similarity at the binding pocket[39]. Differences in binding affinity may arise from the fact that allosteric conformational changes from the pilin to the mannose binding site are thought to be propagated in a dynamic manner [34], [39], affecting the fluctuations of the clamp segment. This hypothesis suggests the need for a deeper investigation into the relative contributions of backbone and sidechain dynamics towards allosteric “information transfer” to the clamp segment, such as by comparing the fluctuations of HL mutant and H2 inactive. Rabbani *et al.*[39] also found that the addition of a mannoside ligand (n-heptyl α-D-mannopyranoside) induced a conformational change in HL mutant with NMR peaks that matched HL active at the binding pocket and the β-sandwich, but not near the mutation or the interdomain region. The conformational change of HL mutant was hypothesized to be similar to that of full-length FimH due to mannoside binding [34]. However, HL mutant has more than seven times higher affinity for mannoside than H2 inactive [39] The difference in binding affinity, despite the similarity in backbone structure, suggests further differences in the backbone dynamics or the sidechains.

The dynamics of HL mutant have not yet been studied using molecular dynamics (MD) simulations. However, Interlandi and Thomas[35] isolated the lectin domain from a crystal structure of H2 inactive and used nanosecond-timescale MD simulations in order to identify protein regions with different dynamics than HL active, focusing on the putative allosteric pathway.

We seek here to identify the protein regions with dynamical differences across HL mutant, H2 inactive, and HL active, for both the backbone and sidechains. In contrast to the top-down approach that starts with identifying landmarks along the putative allosteric pathway, we employ a bottom-up approach without defining sites *a priori* by directly comparing the structure and dynamics across the sequence of the lectin domain. Moreover, because using external Cartesian coordinates introduces artifacts from rigid body motion and relative domain motion, we use a system of internal coordinates based on the dihedral angles [19]. To this end, we perform 20 nanosecond (ns)-long all-atomistic MD simulations of HL mutant, H2 inactive, and HL active to identify protein regions with differences in dynamics.

We find that structural differences between HL mutant and H2 inactive are not limited to the clamp segment in the binding pocket, the interdomain loops, or the β-bulge region with the mutation. For the backbone dynamics, we find the greatest differences in the interdomain region and the binding pocket. There are also differences in sidechain orientation and dynamics, beyond the localized effect of having Arg or Pro at position 60. Our analyses using similarity matrices, a common tool in data science, confirms that the differences in backbone dynamics are more distinct than the differences in backbone structure or in sidechain orientation and dynamics.

The implications of our study extend beyond FimH. Our investigation demonstrates the advantages of a system of internal coordinates over Cartesian coordinates for quantifying dynamics. The regions we identify with dynamical differences across HL mutant, H2 inactive, and HL active using dihedral angles are not restricted to the identified landmarks on the putative allosteric pathway and further highlights the power of a bottom-up approach, which may be applicable to proteins where these landmarks are not yet identified.

3.3 Materials and Methods

3.3.1 FimH structures

We retrieved crystal structures for HL mutant, H2 inactive, and HL active from the Protein Data Bank (RRID:SCR_012820), as detailed in Table 3.1. We denote the full-length FimH protein in the inactive conformation as H2 inactive and denote the truncated structures that only include the lectin domain as HL mutant and HL active. For H2 inactive, we focus on two structures (4XOD and 5JQI) that were crystallized by separate groups. Due to natural sequence variations, different strains of *E. coli* produce FimH with slightly different amino acid sequences. To directly compare sidechain dynamics with matching sequences, we mutated 4XOD at three positions *in silico* to match HL mutant and HL active (Table 3.1). In contrast, we made no changes to 5JQI. Due to the sequence mismatch, sidechain comparisons involving 5JQI were limited, but backbone comparisons were not affected.

HL mutant has a single Arg-60-Pro mutation. To explore the impact of the residue identity alone, without otherwise changing the structure, we used MODELLER [88] (RRID:SCR_008395) to introduce a 'perturbation' *in silico* by changing back the Pro at position 60 to Arg for HL mutant; we made the opposite change for HL active (Supplementary Information). For all structures, we investigated systems with and without the *in silico* addition of mannose from 1KLF.

3.3.2 Equilibrium simulations

We performed all-atomistic MD simulaions using NAMD[51], with the CHARMM force field [55]. Our NAMD simulation parameters and system details are listed in Table 3.2. We prepared all simulation systems using VMD (Visual Molecular Dynamics) [50]. We solvated each protein with enough water molecules to prevent interactions with itself through the periodic boundary

Table 3.1: **PDB structures studied.** For H2 inactive, both 4XOD and 5JQI had differences in the primary sequence. For 4XOD, we mutated three positions to match the primary sequence for HL active.

Name	HL mutant[39]	HL active[89]	H2 inactive v1[34]	H2 inactive v2[84]
PDBID	5MCA	4AUU	4XOD	5JQI
Resolution (Å)	1.604	1.5	1.14	1.982
Residue 60	Pro	Arg	Arg	Arg
Residue 27	Val	Val	Ala-27-Val	Ala
Residue 70	Asn	Asn	Ser-70-Asn	Ser
Residue 78	Ser	Ser	Asn-78-Ser	Asn

conditions.

We performed six replicate simulations for each system. Because we found that replicates starting from a single solvated system were too similar to each other, we solvated, minimized, and equilibrated our replicates separately.

3.3.3 Backbone and sidechain dihedral angles

We focus on dihedral angles to compare protein structures and dynamics instead of Cartesian coordinates because dihedral angles are better suited for identifying the critical regions driving collective motion in Cartesian space, such as hinges that displace distal elements [58]. Moreover, using dihedral angles avoids artifacts from structural alignment and better captures angular motions that affect large segments in hinged proteins, such as FimH [19], [58], [91]. We describe backbone motion using the dihedral angles ϕ and ψ (see Fig. 2.1). We describe sidechain motion using both rotamer angles (χ) and custom-defined pseudo-dihedral angles (υ). Each amino acid has one to five rotamer angles, except for Ala and Gly. To compare the wild-type with the Arg-60-Pro mutation, we omit the last three rotamer angles in Arg. We define the pseudo-dihedral angles, υ, to compare sidechain dynamics when there are differences in the primary sequence, due to either the Arg-60-Pro mutation or the two versions of H2 inactive. We define one pseudo-dihedral angle

Table 3.2: **Details of molecular dynamics simulations.**

Setup and visualization	VMD[50] 1.9.3 (RRID:SCR_001820)
System dimensions	60 Åx 62 Åx 93 Åfor HL active
	60 Åx 72 Åx 123 Åfor H2 inactive
System sizes	>32,000 atoms for HL active
	>60,000 atoms for H2 inactive
Solvation	TIP3P water model with 16 Å padding on each side
Ionization	NaCl ions to neutralize and achieve 50 mM salt conc.
Simulation engine	NAMD[51] 2.13 (RRID:SCR_014894)
Ensemble	NPT
Temperature	300 K
Pressure	1 atm
Non-bonded interactions	Lennard-Jones potential (12 Å cutoff[90])
Electrostatic interactions	Particle-Mesh Ewald sum method[51]
Force field	CHARMM c36 July 2018 update
Timestep	1 fs
Coordinates saved every	1 ps
Energy-minimization	Conjugate gradient algorithm in NAMD
	10,000 steps with protein fixed
	10,000 steps with no atoms fixed
Equilibration	1 ns
Simulation	No ligand: 20 ns
	Ligand: disassociation or up to 20 ns

per amino acid using four atoms: N, CA, and the last two atoms from the most distal χ angle. For example, since Arg has χ_5 calculated by CD-NE-CZ-Nh1, we calculate υ from N-CA-CZ-Nh1 (Fig. 2.1). For Ala, we use N-CA-CB-H1, and we do not define υ for Gly.

We calculate the average, variance, and standard deviation using circular statistics, which account for the periodic wrapping from $-\pi$ to π [19]. We use the standard deviation to assess the range of backbone and sidechain fluctuations. We compare the standard deviation at each residue to identify protein regions with differences in fluctuation size among the starting structures. Because the backbone dynamics were similar in the presence or absence of mannose, as well as the amino acid at position 60, we pooled replicates for backbone dynamics to increase the statistical power. Specifically, we found no significant dynamical differences in the presence or absence of mannose, and extremely small, yet statistically significant, differences at isolated sites due to the amino acid at position 60. However, having Arg or Pro at position 60 affected the sidechain dynamics, so we did not pool replicates. For statistical significance testing, we use Mood's median test from the python package scipy (RRID:SCR_008058) [92]. To correct for multiple comparisons, we use the two-stage Benjamini-Krieger-Yekutieli procedure from the python package statsmodels (RRID:SCR_016074) [93], [94] with a false discovery rate of $\alpha = 0.05$. For robustness testing, we also implement the Bonferroni correction for multiple comparisons with $p = 0.05$.

3.3.4 Similarity matrix

After we identified differences between groups of HL mutant and H2 inactive trajectories, we quantify the extent of the difference by constructing similarity matrices that compare each pair of trajectories. As references, we choose one type of comparison where we expect low similarity and two types of comparisons where we expect high similarity. We expect low similarity, or large differences, when we compare HL active with both versions of H2 inactive. In contrast, we expect high similarity when we compare the backbone for both versions of H2 inactive, as well as when we compare replicates. Using this framework with the low- and high-similarity references, we use similarity matrices to compare the structure and dynamics of both the backbone and sidechains. Each similarity value is a non-parametric (rank-based) correlation. While we analyze dynamics with Spearman's correlation, we analyze structure using the equivalent correlation for circular variables described by Mardia and modified by Fisher and Lee [95]. Using these similarity matrices, we quantify the similarity between HL mutant, H2 inactive, and HL active and perform statistical testing using Mood's median test with a Bonferroni correction for multiple comparisons.

3.3.5 Dihedral principal component analysis

We applied principal component analysis to dihedral angle fluctuations (dPCA) to identify protein regions with coupled dynamics [19]. We chose to use dihedral angle dynamics because they have been shown to capture some motions missed by Cα fluctuations due to "artifacts" during the alignment step, particularly during hinged motion and for both large- and small-amplitude motions [19], [58].

To identify protein regions with coupled motion, we calculate the first through fourth lowest-frequency eigenmodes from the circular covariance matrix for each simulation. We then compare the eigenmodes of HL mutant, H2 inactive, and HL active. We focus on the amount of motion at a particular angle described by these eigenmodes, so we square the eigenvector value at each residue and sum across eigenmodes. Since the eigenvectors are unit vectors, we weight the summation by the eigenvalues, which measure the variance in the angle changes described by the eigenvectors. Moreover, we also calculate the unweighted average, using the eigenvectors that explained 70% of the variation, after finding similar patterns for 50-80% of the variation explained.

3.4 Results

3.4.1 Crystal structure comparisons

We use root mean squared displacement of the backbone Cα atoms (Cα-RMSD) to measure crystal structure differences and to broadly identify structural differences across HL mutant, H2 inactive, and HL active (Fig. 3.1 and SA.3B). As expected, our analyses show that the Cα-RMSD between HL mutant (5MCA) and the two versions of H2 inactive are quite small (0.76 Å to 4XOD and 0.93 Å to 5JQI). This similarity is consistent with the similarity between HL mutant and another version of H2 inactive (3JWN) reported by Rabbani *et al.*[39]. While HL mutant and H2 inactive have similar crystal structures, they are both quite different from HL active (4AUU). The Cα-RMSD between HL active and HL mutant is 3.15 Å and between HL active and the two versions of H2 inactive are 3.30 Å (4XOD) and 3.25 Å (5JQI).

Even though the structural differences between HL mutant and H2 inactive are small, we nonetheless examine the backbone regions with the greatest contributions to the Cα-RMSD differences (Fig. 3.1B and Fig. SA.3B). As described by Rabbani *et al.* [39], we find the largest difference at the clamp segment in the binding pocket and few differences within the "β-sandwich." We also detect some differences in an unnamed region on the opposite side of the binding pocket, as well as in the swing and insertion loops in the interdomain region (see Fig. 3.2 and SA.1 for named regions).

The difference between HL mutant and H2 inactive at the clamp segment has been attributed to its flexibility in the absence of ligand [39]. In particular, the angular differences found in the backbone dihedrals at the C-terminal end of the clamp segment could act like a "hinge" that displaces distal elements (arrow in Fig. 3.2). This hinge motion may cause the large displacements at the tip of the clamp identified by Cα-RMSD (Fig. 3.1B). Compared to the localized angular

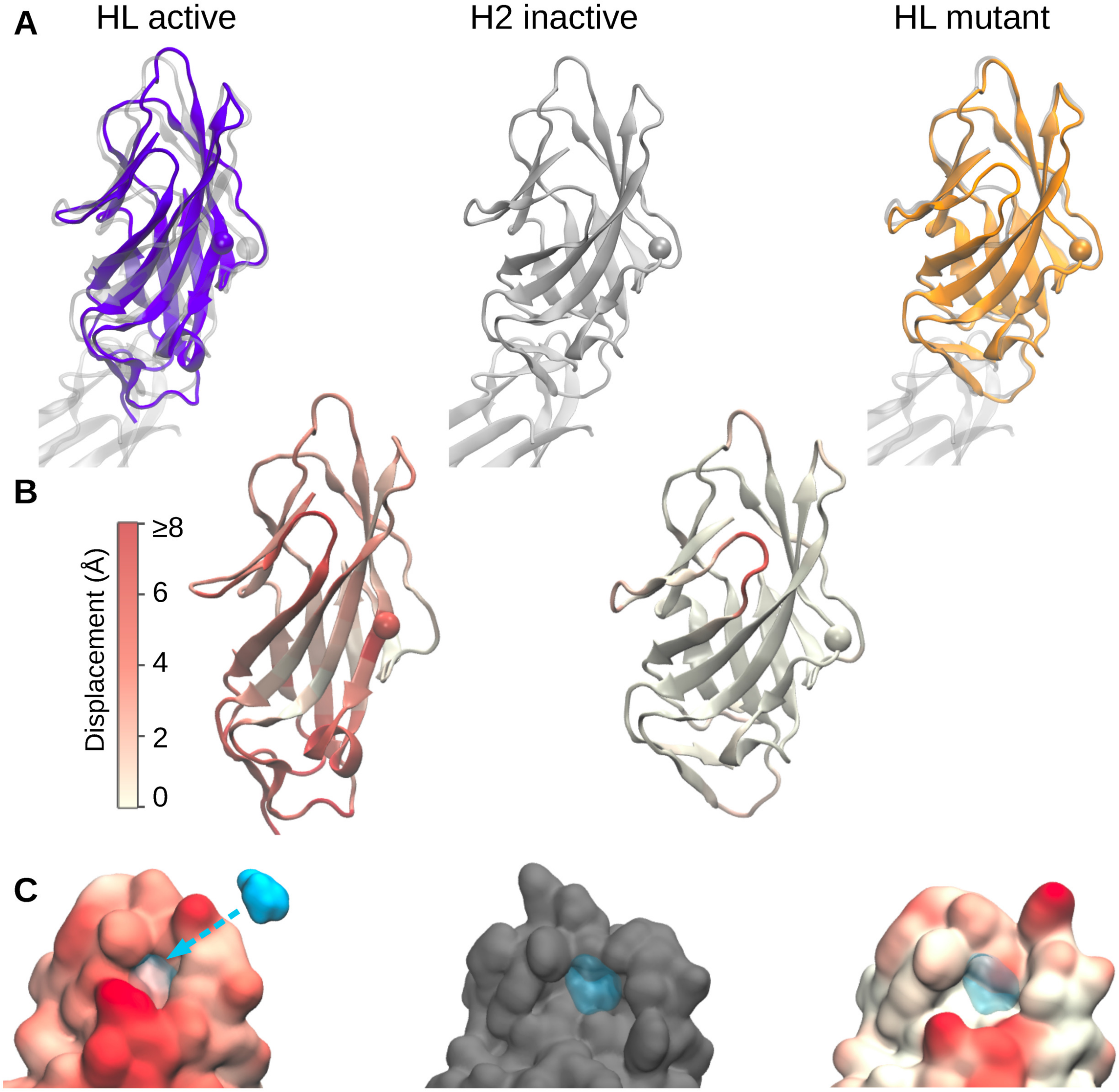

Figure 3.1: **The Arg-60-Pro mutant has a backbone structure similar to that of inactive FimH but with a partially closed binding pocket.** (A) NewCartoon representation of the backbone structures for H2 inactive with the pilin domain cropped (grey, PDBID 4XOD, low-affinity) overlaid by the lectin domain structures: HL active (PDBID 4AUU, orange, high-affinity) and HL mutant with Arg-60-Pro (PDBID 5MCA, purple). We highlight residue 60, the mutation site, with a bead. We use color to illustrate the displacement from H2 inactive for (B) the backbone and (C) the surface rendering of predicted sidechain orientation. The mannose ligand is shown in blue and was positioned after alignment with high-affinity, two-domain FimH (PDBID 1KLF).

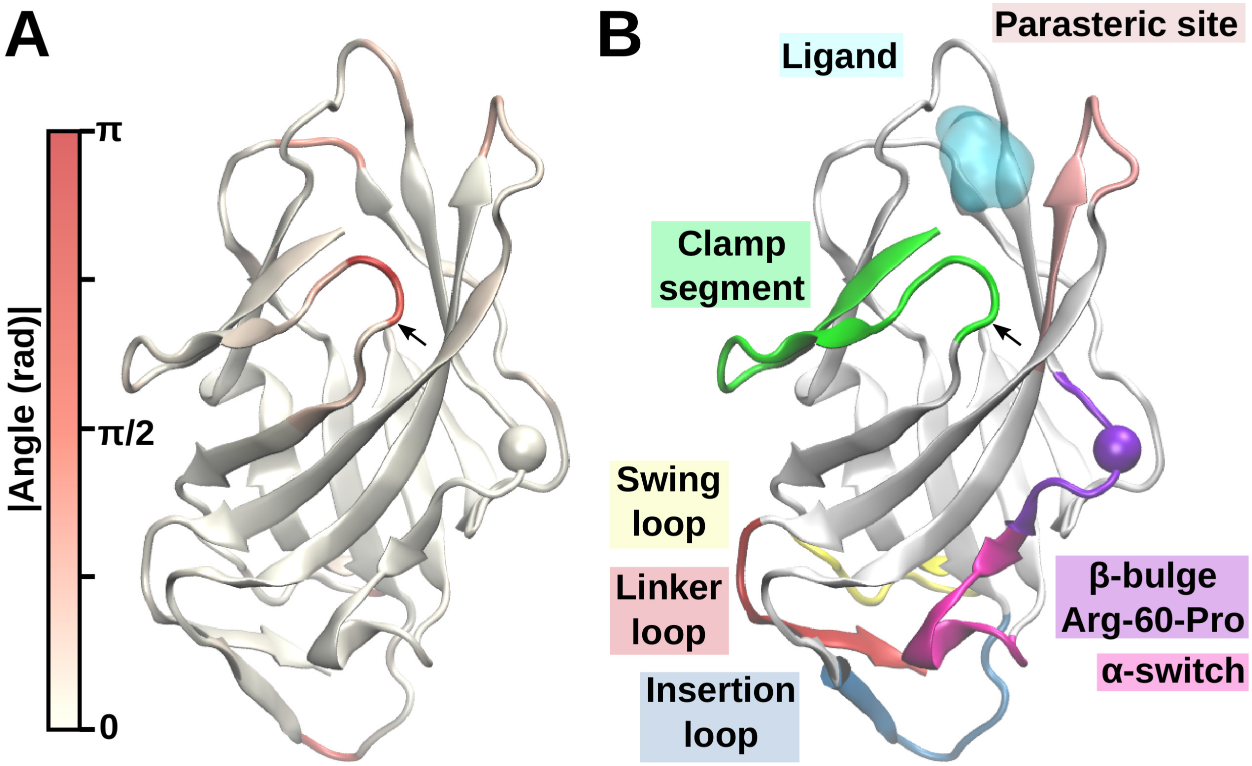

Figure 3.2: **Angle differences between HL mutant and H2 inactive.** (A) NewCartoon representation of the HL mutant backbone shows large changes in backbone dihedral angle differences at the C-terminal end of the clamp segment that may act like a hinge-joint (arrow). (B) Putative allosteric pathway sites[79], the parasteric site [82], and residue 60 (bead).

differences (Fig. SA.3 and SA.4), the overly broad protein regions identified by Cα-RMSD reveals the limitations of calculations involving structural alignment, which have previously been reported in the literature [19], [58] For this reason, residues with large dihedral angle motion may be more dynamically important and specific than the regions of several residues identified by large Cα-RMSD.

3.4.2 Backbone dynamics using internal coordinates

We thus consider an internal coordinate system that uses the average backbone dihedral angles[19] to identify differences between average structures obtained from the MD simulations. Using the dihedral angles to compare HL mutant and H2 inactive, we mostly find small differences in backbone angle orientation across the lectin domains (Fig. 3.3B). In contrast, we find large differences at the insertion loop and the hinge at the C-terminal end of the clamp segment (arrow label in Fig. 3.2). Because we also find differences at the hinge in our crystal structure comparisons, we think that the difference in the orientation of the clamp segment is not resolved within the 20-ns simulation timescale. In our simulations of longer than 200 ns, we find clamp segment conformations accessed by both HL mutant and H2 inactive, as well as conformations only found in our HL mutant simulations. However, we did not find interdomain region conformations shared by both HL mutant and H2 inactive.

Since we find the largest angle change at the hinge-joint and the largest contribution to Cα-RMSD at the center of the clamp segment, we interpret these results to demonstrate that our approach enables us to detect the joints of regions displaying hinge-like motion in an automated fashion. The hinges we identify are those relevant to the 20 ns timescale (Fig. 3.3C), and our smaller number of replicates at the 200 ns timescale suggests the existence of additional hinges (Fig. SA.11B).

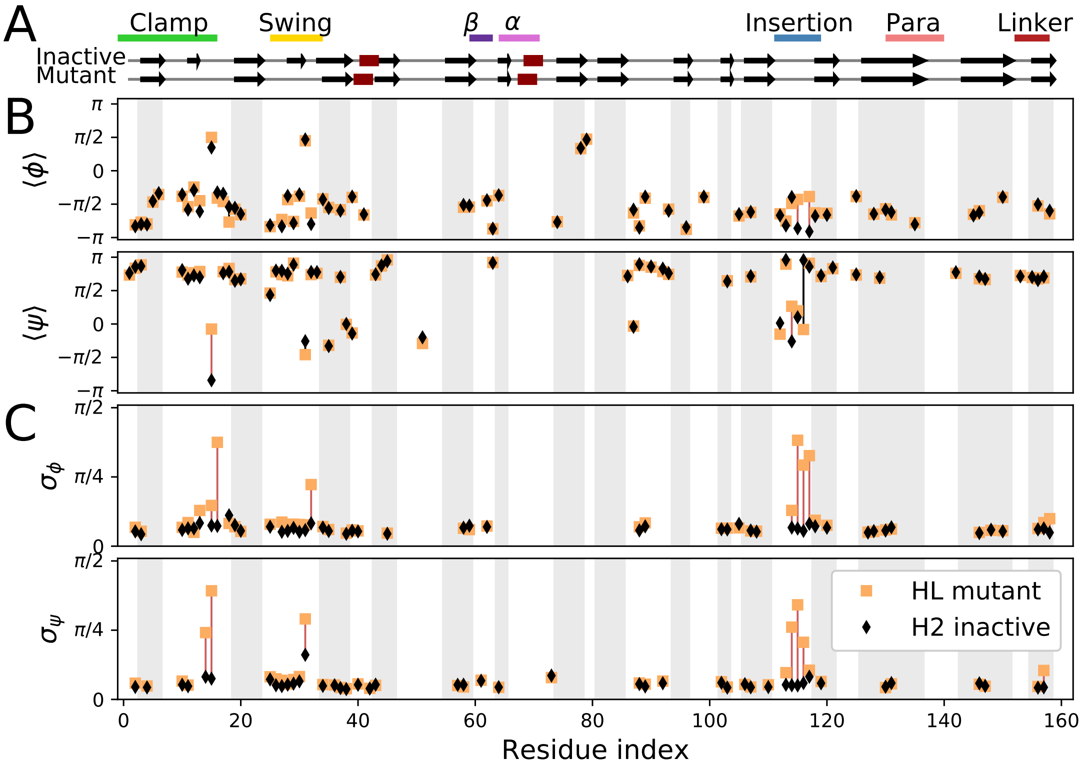

Figure 3.3: **Localized differences in the structure between HL mutant and H2 inactive cause differences in dynamics.** (A) Secondary structure of H2 inactive (top) and HL mutant (bottom) with features. (B) Protein structure comparison between HL mutant (orange squares) and H2 inactive (black diamonds) used average backbone dihedral angles ϕ and ψ, which provide an internal frame of reference. Structural differences are localized to the hinge of the clamp loop in the binding pocket and the insertion loop in the interdomain region, which is exposed to water in HL mutant. (C) Backbone dynamics comparison used the standard deviations of ϕ and ψ, quantifying the magnitude of angle fluctuations. HL mutant has larger fluctuations (red lines) at the clamp loop and the insertion loop. For statistically significant differences, the median values are shown.

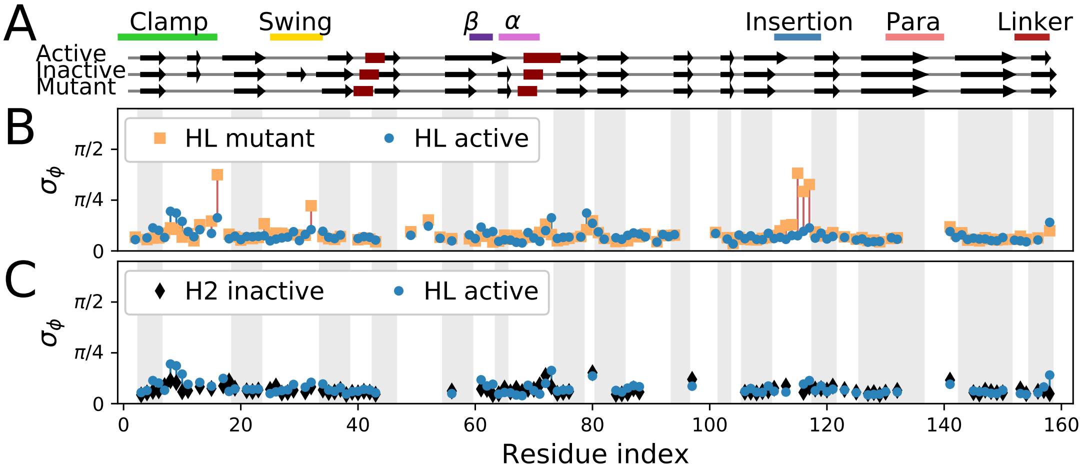

Figure 3.4: **Comparisons of backbone dynamics show HL mutant has larger fluctuations than H2 inactive and HL active.** Backbone dynamics were compared using the standard deviations of ϕ (See Fig. SA.10 for ψ). (A) Secondary structure of HL active (top), H2 inactive (middle), and HL mutant (bottom) with features. (B) HL mutant has larger backbone fluctuations in the insertion and swing loops than HL active (blue circle). (C) HL active and H2 inactive (black diamonds) have small differences in backbone dynamics throughout the lectin domain.

We next investigate backbone dynamics for the dihedral angles found in these hinge regions and use the standard deviation to compare the magnitude of the fluctuations (Fig. 3.3C and 3.4). We find that HL mutant has larger fluctuations than H2 inactive in the insertion and swing loops of the interdomain region and in the clamp segment of the binding pocket (Fig. 3.3C). Thus, the backbone regions with larger dynamical differences are also the regions with larger structural differences.

Since it has been hypothesized that the difference in dynamics is related to the exposed interdomain region of HL mutant [39], we next compare the dynamics of HL mutant and HL active, as both have the interdomain region exposed. We find that HL mutant also has larger fluctuations than HL active in the insertion and swing loops of the interdomain region (Fig. 3.4B). For the clamp segment, we find that HL mutant is more dynamic at the hinge-joint, while HL active is only slightly more dynamic at the tip of the clamp.

Surprisingly, we also find that HL active and H2 inactive display small differences in dynamics across the protein (Fig. 3.4C). HL active is only slightly more dynamic in the clamp segment and the C-terminus, with minute differences in dynamics across the protein. We interpret this finding to show that HL active and H2 inactive are more stable compared to HL mutant. The interdomain regions in H2 inactive and HL active may be stabilized by different mechanisms: contacts with the pilin domain for H2 inactive [34], [86], and a more rigid fold that reduces the solvent-accessibility of the hydrophobic interdomain region for HL active [35]. Our data supports the idea that different allosteric pathway sites have different coupling strengths to the binding pocket [39], [79].

The presence or absence of the mannose ligand does not result in statistically significant differences for the backbone structure or dynamics. We interpret these results to indicate that the initial protein structure has the largest impact on both structure and dynamics, at least for the up to 20 ns duration of our MD simulations.

3.4.3 Comparing sidechain orientations and fluctuations

For many proteins, including FimH, ligand-binding also depends on sidechain flexibility and optimal sidechain-ligand interactions [89]. To separate sidechain motion from displacements due to backbone motion, we again use internal angles. We quantify sidechain structure and dynamics using two alternative sets of angles: χ dihedral angles (or rotamer angles) and the custom-defined pseudo-dihedral angles υ. We define one υ angle, using N, Cα, and two distal atoms, for each amino acid except for Gly (see Methods and Fig. 2.1). Each amino acid, except for Gly and Ala, also has one to five χ rotamer angles. By construction, the υ angles have broader angle distributions than the χ angles and enable us to compare the two versions of H2 inactive in spite of their minor sequence variations.

While HL mutant and H2 inactive have similar backbone structures, sidechain orientations estimated from the crystal structures suggest that the geometries of their binding pockets are quite different (Fig. 3.1C). To separate sidechain motion from displacements due to backbone motion, we first compare the average sidechain orientations of HL mutant and H2 inactive using υ (Fig. 3.5B) and χ (Fig. SA.13B) angles. Our data show that the differences extend beyond the β-bulge segment with the Arg-60-Pro mutation. Indeed, in addition to the β-bulge segment, we find nine large differences in sidechain orientation across the protein, including the hinge-joint of the clamp segment, the swing loop, the parasteric site, and several unnamed regions. For example, we identify a difference in the sidechain orientation of Phe-144 in the β-sheet between the parasteric site and the linker loop. Phe-144 touches the tip of the clamp segment in HL mutant but not in H2 inactive. We find large differences in the orientation of Phe-43, Tyr-108, and Phe-144 using both υ and χ, but we only find a statistically significant difference at Arg-92 using υ, which combines all five χ angles for Arg into one measure between the base and the tip (Fig. 3.5B and SA.13B).

We also find differences in the sidechain orientation and dynamics between HL mutant and

H2 inactive. As expected, the longer Arg-60 in H2 inactive has larger fluctuations than Pro-60 in HL mutant. We also find regions with dynamical differences beyond the site of the Arg-60-Pro mutation. Using υ, the larger fluctuations in the swing loop of HL mutant (Fig. 3.5C) may be attributed to the missing interdomain contacts that are present in H2 inactive. Moreover, the swing loop of H2 inactive is structurally stabilized as a β-strand with the linker loop (Fig. 3.1 and SA.1). Using χ instead of υ, we find HL mutant also has larger fluctuations within the clamp segment (Fig. SA.13C). This suggests that HL mutant may have a more flexible clamp segment, which is consistent with the larger backbone fluctuations.

As expected, we find that HL active is different from both HL mutant and H2 inactive in both sidechain structure and dynamics at many positions across the protein (Fig. 3.5D and SA.13D and Table SA.1). In particular, many of these differences fall outside of the landmarks of the putative allosteric pathway[79] which are frequently the focus of attention and are primarily studied using backbone dynamics [35], [39]. Overall, the protein regions identified using χ and υ are very similar.

3.4.4 Robustness checking using a similarity matrix approach

In the previous section, we identified protein regions on HL mutant and H2 inactive with statistically significant differences in structure and dynamics by comparing groups of trajectories. Nonetheless, it remains difficult to quantify how different HL mutant is from H2 inactive relative to variations that may occur across replicates. To more fully assess the dissimilarity between the dynamical behavior of HL mutant and H2 inactive, we directly compare pairs of trajectories by calculating rank-based correlations and constructing similarity matrices (Fig. 3.6). We use the Spearman correlation for angle fluctuation size and an equivalent correlation for circular variables for average angle orientation. The similarity matrices allow us to compare each pair of trajectories,

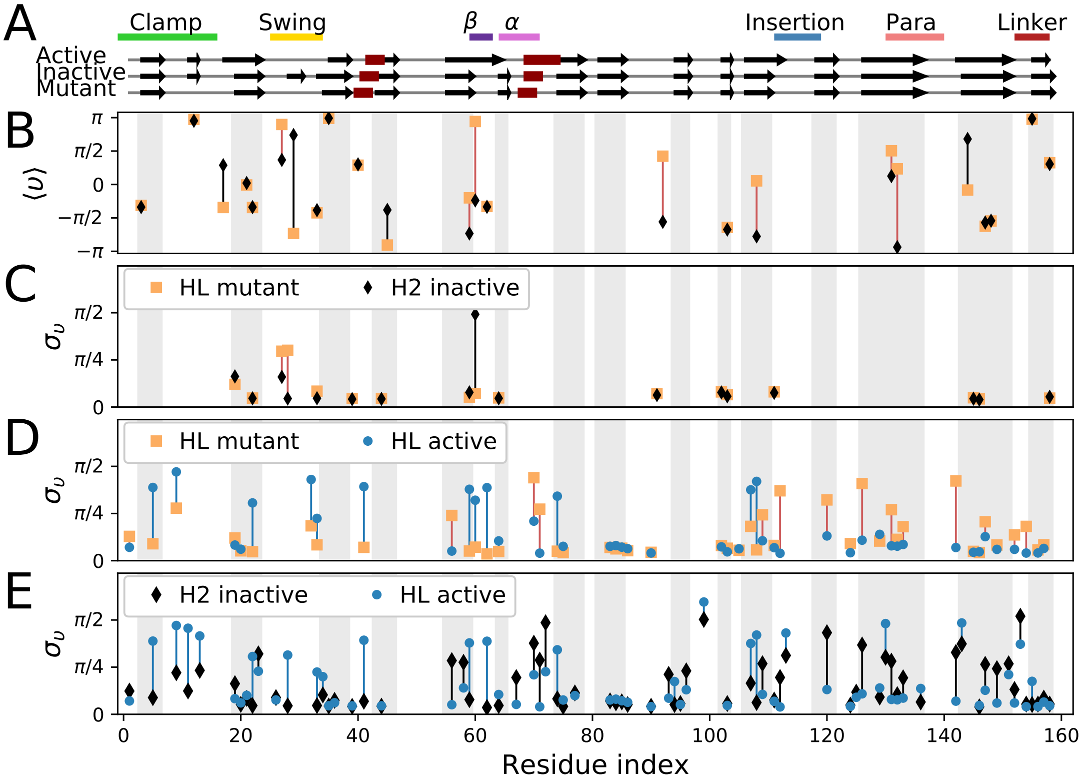

Figure 3.5: **Differences in the sidechain dynamics between HL mutant and H2 inactive affect the binding pocket, and both are different from HL active.** (A) Secondary structure of HL active (top), H2 inactive (middle), and HL mutant (bottom) with features. Sidechain structure and dynamics were compared using custom dihedral angles υ. (B) At the β-bulge, the hinge of the clamp segment, the parasteric region, the swing loop, and several other sites, the average pseudo-dihedral angle was different between HL mutant (orange squares) and H2 inactive (black diamonds). (C) Arg-60 in H2 inactive has larger fluctuations than Pro-60 in HL mutant. HL mutant also has larger fluctuations in the exposed swing loop. (D, E) Comparisons of HL active (blue circles) against HL mutant and H2 inactive show differences in sidechain dynamics beyond the allosteric pathway sites.

complementing our previous comparisons across groups.

We define the reference similarity values as: the low similarity expected for HL active vs H2 inactive, very high similarities for replicates, and high similarity for backbone comparisons of the two versions of H2 inactive. We use these reference values to compare the differences between HL mutant and H2 inactive.

As expected, we find that HL active is very different from both HL mutant and H2 inactive for all similarity matrices, and HL active replicates are relatively similar to each other. We find that the two H2 inactive structures are very similar to each other for backbone comparisons, but different for sidechain comparisons. These differences are smaller than the ones between H2 inactive and HL active, but are slightly larger than the ones between replicates of each version of H2 inactive (Fig. SA.14 and SA.15).

We find the replicates for HL mutant are similar to each other, except for the average backbone structure of HL mutant. One possible explanation is that the larger fluctuations found in HL mutant make the replicates less similar to each other. We find that having separate solvation, minimization, and equilibration preparations accentuated the difference between technical replicates, compared to having replicate production runs following a single preparation. Thus, for backbone structure, we find the lower similarity values for the comparisons of HL mutant vs H2 inactive is about the same as those for replicates of HL mutant (Fig. SA.14). In comparison, for backbone dynamics, we find clear differences between HL mutant and H2 inactive that are larger than the differences between replicates, yet smaller than the differences for HL active compared to both H2 inactive and HL mutant (Fig. SA.15).

We also find that average sidechain orientations had lower similarity overall for both the rotamer (χ) and the custom pseudo-dihedral (υ) angles. Despite overall lower similarities, we are still able to use the pseudo-dihedral angles to distinguish the two versions of H2 inactive, which

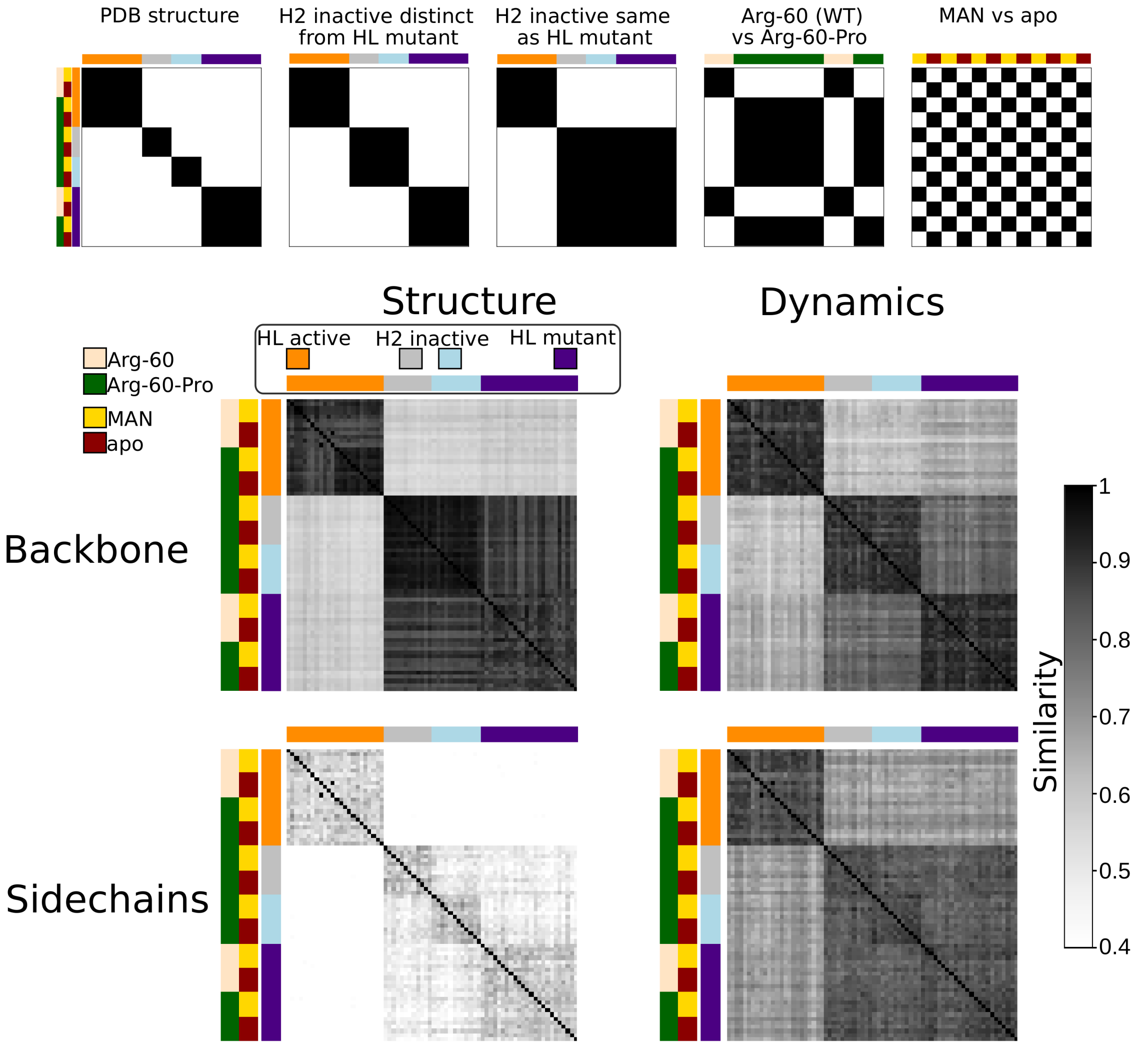

Figure 3.6: **Differences between HL mutant and H2 inactive in comparisons of backbone dynamics and sidechain orientation.** (top row, from left to right) Expected patterns based on the hypotheses that similarity depends on: the four PDB structures; the three protein states; grouping HL mutant with H2 inactive; the amino acid at position 60; and the presence of ligand. We constructed similarity matrices by comparing backbone dihedral (upper) and sidechain pseudo-dihedral (lower) angles, using the average angle for structure (left) and the standard deviation, for dynamics (right). The greyscale for similarity is the same for all four matrices. The system setup is described using a three-layer color code for the initial protein structure, mutation at position 60, and ligand. The purpose is to compare trajectories from different initial protein structures. HL active is distinct from both H2 inactive and HL mutant for all measures. For backbone dynamics, HL mutant is distinct from H2 inactive. For sidechain structure, the two versions of H2 inactive are different, and HL mutant is distinct from H2 inactive.

have differences in the amino acid sequence at three positions in the lectin domain (see Table 3.1). We also identify differences between HL mutant and H2 inactive (Fig. SA.14). For sidechain dynamics, we find slightly lower similarity values between HL mutant and H2 inactive relative to the replicates, which is comparable to the low similarity between the two versions of H2 inactive.

We find that the reduction in similarity for HL mutant and H2 inactive, relative to the high similarity between replicates, is smaller for sidechain dynamics than for backbone dynamics (Fig. SA.15). Across the four types of comparisons, we find that the differences primarily depend on whether the the initial protein structure is HL mutant, H2 inactive, or HL active (Fig. 3.6). We find larger differences due to the initial protein structure than those due to the amino acid at position 60 or to the presence of the ligand, which is consistent with our results in the previous section.

3.4.5 Coupled backbone motions

We apply dihedral principal component analysis (dPCA) [19] to to the backbone dynamics. To evaluate coupled backbone dynamics, or collective motion, we focus on the magnitude at each backbone dihedral by summing the eigenmodes weighted by the eigenvalues. Comparing HL mutant and H2 inactive for the first eigenmode, we find that HL mutant has more motion in the insertion loop and less motion at position 60 in the β-bulge segment. We find that after adding the second eigenmode, the difference in the amount of motion is larger (Fig. 3.7). After adding the third eigenmode, we find that HL mutant has more motion in the linker loop in the interdomain region and the clamp segment of the binding pocket, while there is less motion in the α-switch segment connected to the β-bulge segment. This is also magnified after adding the fourth eigenmode. We interpret these comparisons between HL mutant and H2 inactive to mean that there may be more coupled motion in HL mutant between the interdomain region and the clamp segment of the binding pocket. In contrast, we find decreased dynamics in HL mutant within the β-bulge/α-

switch segment, which may be due to Pro-60 limiting backbone dynamics compared to wild-type Arg-60 in H2 inactive.

In addition, we also calculate the unweighted average of the eigenvector magnitudes for the eigenmodes that explain 70% of the variance (Fig. 3.7). Similar to the comparison of backbone dynamics, we find that HL mutant has larger fluctuations than H2 inactive, primarily in the insertion, swing, and linker loops of the interdomain region and the clamp loop of the binding pocket. However, we find that HL mutant had smaller fluctuations for the hinge region of the clamp segment. We identified similar regions by averaging eigenmodes that explain 50-80% of the variance. We interpret the similar regions identified by both methods to indicate that HL mutant is missing stabilizing interactions in the interdomain region that are present in H2 inactive [86]. The more exposed HL mutant structure may lead to larger fluctuations that may be coupled to the binding pocket.

3.4.6 Conclusion

To compare the protein regions with differences in backbone and sidechain fluctuations, we show the regions identified with the variance as the volume of the sphere (Fig. 3.7). Different regions are identified for backbone and sidechain dynamics. For example, position 60 in the β-bulge and nearby regions showed a larger difference for Pro vs Arg in the sidechain than in the backbone. The regions where HL mutant had larger backbone fluctuations than H2 inactive include the interdomain region and the hinge of the clamp segment. Larger fluctuations at the binding and allosteric sites may explain why HL mutant can undergo an allostery-like conformational change without the tensile force required to separate the two domains of H2 inactive.

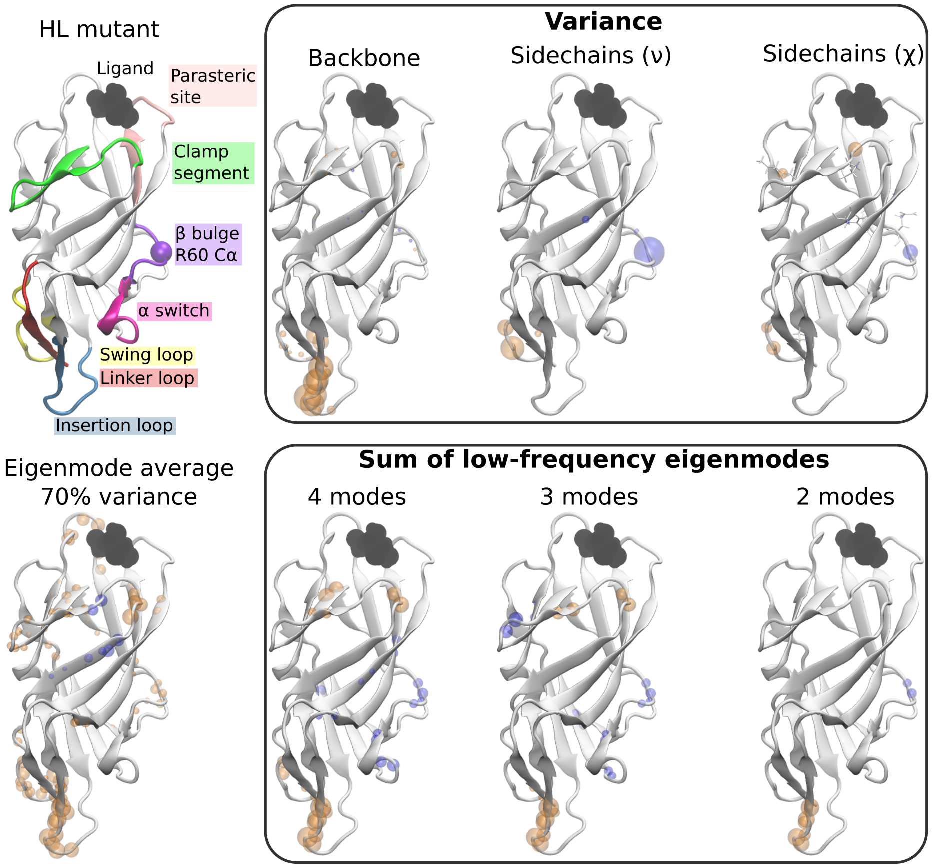

Figure 3.7: **HL mutant has larger fluctuations in the interdomain region and clamp segment.** (top) Backbone and sidechain angles with larger fluctuations for HL mutant (orange) or H2 inactive (blue) are shown as spheres on the protein backbone. The landmarks are shown on HL mutant, and the Arg-60-Pro mutation is indicated with a purple sphere at Cα. Fluctuation size, or dihedral variance, is shown as the sphere volume. For backbone dihedrals from Fig. 3.3, ϕ is shown at N, while ψ is shown at the carbonyl C. Sidechain pseudo-dihedrals (υ) from Fig. 3.5 are shown on Cα. Sidechain dihedrals χ are shown on the second atom defining the dihedral angle, which was usually Cα. The sidechain was drawn in the licorice representation for distal angles. (bottom) The backbone eigenmodes were compared two ways, using the eigenmodes responsible for 70% of the variance and also using the low-frequency eigenmodes.

3.5 Discussion

HL mutant (Arg-60-Pro) has been proposed as a "minimal model" of FimH allostery because the single-domain structure is more experimentally tractable than H2 inactive. Moreover, HL mutant undergoes an allostery-like conformational change upon binding mannoside ligand in contrast to the wild-type HL active [39]. The HL mutant crystal structure was shown to match H2 inactive, except at the clamp segment of the binding pocket, and this difference was attributed to the flexibility of the clamp segment [39]. However, the dynamics of HL mutant have not yet been studied experimentally, and this is the first computational characterization using MD simulations.

While HL mutant and H2 inactive overall share a similar structure, our analysis shows that differences in average backbone structure exist in the interdomain loops, in addition to the clamp segment of the binding pocket. In these regions, HL mutant also has larger backbone fluctuations than H2 inactive in our 20 ns simulations, which we also observed in the simulations longer than 200 ns. Over longer simulations, many clamp segment conformations are accessed by both HL mutant and H2 inactive, but we only found some conformations in HL mutant. The floppier clamp segment in HL mutant may contribute to the almost 14 times lower binding rate compared to H2 inactive, and possibly the unfavorable loss of conformational entropy upon binding mannoside [39]. We further used a technique from data science to show that these differences are robust compared to the variation among replicate simulations.

These structural and dynamical differences likely contribute to the more than seven times higher affinity to mannoside for HL mutant compared to H2 inactive [39], but their role in the mannoside-induced conformational change remains unknown. Rabbani *et al.*[39] has proposed studying additional mutations to HL mutant in order to prevent the conformational change. This would avoid the high-affinity state that strengthens bacterial adhesion during urination, which limits clearance and prolongs infections [32], [83]. Studies of single mutations suggest that the allosteric coupling to the binding pocket is stronger at position 60 than at the interdomain loops [35], [39], [79]. This is further supported by our finding that HL mutant has larger fluctuations at the binding pocket and the interdomain loops than H2 inactive. Thus, our dynamics characterization further supports the need to determine the structure of the mannoside-bound HL mutant [39]. In addition, greater backbone flexibility in the insertion and swing loops of the interdomain region may be useful for predicting additional mutations. Mutations that stabilize the interdomain region may inhibit the conformational change upon mannoside-binding. Comparing the dPCA eigenmodes for HL mutant and H2 inactive suggest these dynamical differences are related to the allosteric coupling between the interdomain region and the clamp segment. Longer simulations may provide additional insight into the mannoside-induced conformational change.

In addition to comparing HL mutant and H2 inactive, we also compared these structures to HL active. Interlandi and Thomas [35] found that differences between inactive and active FimH were isolated to the six sites defining the putative allosteric pathway in their study of backbone structure and dynamics using Cα-RMSD. In a later study, Rabbani *et al.* [39] focused on a subset of these sites when comparing inactive and active FimH using Cα-RMSD. However, the alignment step in Cα-RMSD calculations results in broader regions of interest since angular fluctuations displace distal elements [19], [58], [91]. Because we use an internal coordinate system of dihedral angles, we identify more localized differences. Thus, our backbone comparisons identify large differences at specific angles along the allosteric pathway.

We then extended our analyses of dihedral angles in the backbone to the sidechains (using both χ and υ) in order to compare the structure and dynamics. We find that HL active is dynamically different from HL mutant and H2 inactive at multiple regions beyond the putative allosteric pathway identified from the backbone. Further characterization of these sites may be useful for studying the role of sidechain dynamics in the allosteric mechanism in more detail. One goal may be to identify mutations that could be combined with Arg-60-Pro to produce a mutant locked in the inactive state despite the addition of mannoside ligand, a goal proposed by Rabbani *et al* [39]. Thus, our comparisons of HL mutant, H2 inactive, and HL active demonstrate that a bottom-up approach can identify differences in structure and dynamics that are not restricted to the putative allosteric pathway. Beyond our analyses of a “minimal model” for FimH allostery, our bottom-up approach and use of techniques from data science may be applicable to comparing the structure and dynamics of other proteins.

CHAPTER 4
A NEW APPROACH FOR EXTRACTING INFORMATION FROM PROTEIN DYNAMICS

4.1 Abstract

Advances in the characterization and prediction of protein structures have created increasing interest in the dynamics of folded proteins and their role on protein function, further increasing the importance of molecular dynamics simulations. To analyze these numerous datasets, proteins are often modeled as networks to take advantage of well-developed methods from network science. Protein networks are often constructed from correlative measures. However, in network science, it has been demonstrated that the inverse covariance matrix can identify network interactions. We apply this inverse approach to the dynamics of protein dihedral angles, a system of internal coordinates that avoids the structural alignment issue in hinge-like proteins. Using the well-characterized adhesion protein, FimH, we show that our method can detect differences between conformations, as well as local changes due to deletion of a disulfide bond. By applying our method to two other adhesion proteins – Siglec-8, an important protein for regulating the immune system, and the SARS-CoV-2 spike protein – we showed that the inverse approach identifies networks that resemble a contact map and are robust across replicates. Due to the differences in the networks constructed by correlation and by solving the inverse problem, there are also downstream differences in comparing networks, calculating network properties, and detecting communities. Extracting network structure from protein dynamics makes it tractable to apply techniques from network analysis, such as community detection, which may help identify collective motion in allostery.

4.2 Introduction

Advances in experimental structure determination [96], [97], computational structure prediction [98], and molecular dynamics simulations [3] have set the stage for high-throughput characterization of protein dynamics using molecular dynamics (MD) simulations. Combined with increased computational power, these advances have led to rapidly increasing numbers of longer MD simulations for larger macromolecular systems [5]. As a result, large datasets of MD trajectories are available from individual research labs [5] and repositories such as MoDEL [9], Dynameomics [99], Dryad, NoMaD, and MolSSI [8].

This wealth of MD trajectory data creates opportunities for expanding our understanding of protein dynamics and function. While selected snapshots from MD trajectories contain information about low energy states and can be used to identify conformational transitions, some physical phenomena, such as dynamic allostery in proteins, may be better captured by changes in dynamics over the course of the trajectory [38]. MD simulations can capture differences among protein variants, the impact of mutations, and modulation by small molecule binding at spatiotemporal resolutions that are difficult, or even impossible, to obtain experimentally [3].

Taking advantage of this growing wealth of MD trajectory data will require the development of robust methods for automated analysis. A common strategy for analyzing dynamics data involves creating a network by directly calculating contact times and interaction energies [13], [62], [64], [100]. An alternative strategy aims to identify the underlying interactions of multi-component systems by inferring a network structure from dynamics. Compared to interaction energy networks, building networks from dynamics is typically less computationally expensive and avoids less rigorous modeling of water or entropic contributions to the free energy [63]. Identifying a network structure makes it possible to apply network analysis tools in order to uncover emergent properties such as densely-connected communities [101], hotspots with many edges [102], and paths connecting active sites and allosteric regulatory sites in distant protein regions [13], [65].

In the study of proteins, the typical approach for constructing networks from protein MD simulations makes use of correlation measures that quantify how different protein regions ”move together.” These include a variety of methods that use linear [103] and non-linear [15], [27], [102], [104]–[106] correlation measures. Yet a rigorous mathematical analysis demonstrates that inferring the network structure by solving the inverse problem for a system that could produce the observed correlations is a more accurate approach than using the correlations directly [20]. The most straightforward form of solving the inverse problem is simply calculating the inverse of the covariance matrix [20]. Here we apply the inverse covariance approach to MD trajectory data.

A remaining challenge is how to define the nodes in such a network representation. An approach that has been used in proteins is to assign a node to each Cα atom in elastic network models [17], [18] . This choice has some appeal because the model describes “beads”, located in Cartesian coordinates, connected by linear springs [17], [18]. However, using Cartesian coordinates requires a structural alignment step that can introduce ”artifacts” during hinged motion for multi-domain proteins, and even for small, single-domain proteins [59]. Previously, we have demonstrated that an internal coordinate system using dihedral angles makes it possible to accurately localize motion that affects elements distal to the hinge [107].

Here, we show that network inference using inverse covariance analysis is robust across replicates and that it uncovers strong interactions among backbone dihedrals that form a contact-map pattern. While the contact-map pattern is also seen in elastic network models for single domains with high conformational stability, we continue to see this pattern even for multi-domain proteins with hinged motion when using inverse covariance analysis.

We demonstrate the value of the proposed approach by studying three physiologically significant proteins: the bacterial adhesion protein, $\mathrm{FimH_L}$, the human immune adhesion protein Siglec-8 [40], and two domains of the SARS-CoV-2 spike protein involved in adhesion to the human ACE2 receptor [46]. In addition to comparing the different structures of wild type and mutant $\mathrm{FimH_L}$, we are also able to detect localized structural changes due to breaking a disulfide bond *in silico*. For Siglec-8, we are able to detect differences between ”apo’ and ”holo’ states, despite their structural similarity [40]. For the SARS-CoV-2 spike protein, we examined the receptor binding domain (RBD) and its connecting subdomain 1 (SD1). While the hinge region connecting RBD-SD1 is open in the up state and closed in the down state, the individual domains remain structurally similar in the up and down states [46]. For Siglec-8 and spike RBD-SD1, which do not have large changes within protein domains, our approach enabled us to use protein dynamics to identify regions that experience changes in inferred interactions.

4.3 Methods

4.3.1 Protein structures

We retrieved crystal structure for $\mathrm{FimH_L}$ wild type and mutant, Siglec-8 apo and holo, and SARS-CoV-2 spike protein from the Protein Data Bank, as detailed in Table 4.1. For $\mathrm{FimH_L}$, we used crystal structures of the lectin domain without ligand for both the wild type and the mutant. To compare dynamics with and without the disulfide bond as a local perturbation made *in silico*, we used Visual Molcular Dynamics (VMD) to define the bond or two cysteins for $\mathrm{FimH_L}$. For Siglec-8, we simulated the ligand 6’S $\mathrm{sLe^x}$ without the 3-amino-propyl linker, which is not thought to interact with the binding pocket [40].

For the SARS-CoV-2 spike protein, we started from the refined structure on the CHARMM-GUI archive [46], [108]. To focus on one hinge system that is thought to be different between the down and up states, we isolated the receptor binding domain (RBD) and subdomain 1 (SD1)

Table 4.1: **PDB structures studied.** We studied FimH$_L$ in the active (wild type) and inactive (Arg60Pro mutant) states; the human immune-inhibitory protein Siglec-8 in the apo and holo (6'S sLe^x-bound)states; and the SARS-CoV-2 spike protein RBD-SD1 domains in the up and down states.

	FimH$_L$		**Siglec-8**		**Spike RBD-SD1**	
Method	Crystal		NMR		Cryo-EM	
Protein residues	158		145		276	
[NaCl] (mM)	50		150		150	
State	**Wild type**	**Mutant**	**Apo**	**Holo**	**Up**	**Down**
PDBID	4AUU	5MCA	2N7A	2N7B	6VSB	
Protein atoms	2360	2350	2290		4286	
System atoms	32292	31917	42684	50879	80383	89236
System size 1 (Å)	94	87	97	111	125	116
2 (Å)	61	67	70	72	94	95
3 (Å)	59	59	67	67	81	78

protein subunits without the glycans. For the up state, we used chain A where the RBD is accessible for binding the ACE2 receptor on human cells [46], and for the down state, we used chain B. We used the structure

We prepared all systems using VMD version 1.9.3 [50]. We solvated each protein with at least 16 Åof TIP3P water molecules on each side to prevent interactions with itself through the periodic boundary conditions. We added sodium and chloride ions to neutralize the system and achieve the desired salt concentration in Table 4.1.

4.3.2 Equilibrium simulations

We performed all-atomistic MD simulaions using NAMD [51], with the CHARMM force field.[55]. Our NAMD simulation parameters and system details are listed in Table 4.2.

After observing differences in correlated protein motions between replicates, we performed three replicates of over 200 ns each for wild type FimH$_L$. Due to the tradeoff between the number replicates and simulation length, we performed six replicates of 20 ns of FimH$_L$ to make compar-

Table 4.2: **Details of molecular dynamics simulations.** Software and parameters used.

Parameter	**Value**
Setup	VMD 1.9.3
Simulation engine	NAMD 2.13
Ensemble	NPT
Temperature	300 K
Pressure	1 atm
Non-bonded interactions	Lennard-Jones potential (cutoff)
Electrostatic interactions	Particle-Mesh Ewald sum method
Forcefield	CHARMM c36 July 2018 update
Timestep	1 fs
Coordinate saved every	1 ps
Energy minimization	Conjugate gradient algorithm in NAMD $\geq$ 10,000 steps with protein fixed $\geq$ 10,000 steps with protein free

isons of wild type and mutant FimH, as well as wild type FimH$_\mathrm{L}$ with and without the Cys3-Cys44 disulfide bond.

Since twenty lowest-energy structures are reported for Siglec-8, we performed a single replicate of 50 ns for each structure to compare apo and holo Siglec-8. For the spike RBD-SD1 domains, we performed six replicates of 60 ns.

4.3.3 Backbone and sidechain dihedral angle dynamics

We used dihedral angles to capture protein dynamics because dihedral angles identify localized regions responsible for the collective displacement of distal regions through space, such as in hinged motion [19]. Dihedral angles are also an internal coordinate system that avoids the structure alignment step when using Cartesian coordinates, which can introduce artifacts [59]. We use both backbone (ϕ, ψ and sidechain $\chi_1 - \chi_5$ dihedral angles.

4.3.4 Inverse of the covariance matrix

In the literature, the covariance matrix is one approach used to identify protein regions with motions that are related to the motions of many other regions; in particular it is used to identify correlated motions between distant regions in allostery [26], [27]. However, constructing networks from the covariance matrix, even with a threshold to remove weak correlations, is susceptible to induced correlations when two nodes (e.g. A and C) are not directly connected but share a connection with a third node (B) [20]. Borrowing from the field of network reconstruction, we use the inverse of the covariance matrix to identify the connections and weights, or edges, between nodes [20]. This approach is consistent with finding the inverse of a covariance matrix based on Cα positions, which fits the Hessian matrix describing an elastic spring network with anisotropy [17], [18]. We have found that the anisotropic elastic network model (ANM) has large errors when used to describe the motion of FimH, which is consistent with errors for hinge-motion described in literature [58]. As a result, we use dihedral angles. This approach is similar to the torsional network model (TNM) which uses equal spring constants to describe dihedral angles across the protein [66]. In contrast, the inverse of the covariance matrix uses the variances and the covariances of angles to calculate spring constants for a network of torsional springs. The nodes in our network are dihedral angles, the edges are like linearly coupled torsional springs, and the inverse of the covariance matrix is the Hessian matrix for a TNM.

To construct our network, we calculate the Moore-Penrose pseudo-inverse of the covariance matrix using both the backbone and sidechain dihedral angles. We use this approach to understand the relative contributions of backbone and sidechain dynamics to collective motion. Since the sign describes whether the angles turn in the same direction, we take the absolute value to get the interaction strength. We do not apply distance filters. While we use the 97th percentile as a value threshold for selecting strong interactions or visualizing the network on the protein, we do not use any thresholds for network comparisons. More generally, we recommend caution for applying thresholds to these networks for analysis.

4.3.5 Comparing networks of inferred interactions

To identify interactions that are stronger in one protein state than another, we compare each edge. We select for large differences between groups, relative to the variability within each group. To do this, we filter for differences larger than twice the standard deviation for each group. To compare an edge e between states a and b, each with an ensemble of m and n networks, this is $|\langle e_a \rangle_m - \langle e_b \rangle_n| > 2\sigma_{e_a}$ and $> \sigma_{e_b}$. We apply this rule without determining statistical significance with corrections for multiple comparisons, in order to see the full effects of comparing all interactions on the matrix. In Extended Data Fig. SB.14a, we also show an example of comparisons without filtering for large differences, in order to illustrate the persistence of the contact-map pattern. We perform the network comparisons in two ways: 1) for every edge on the network (see Extended Data Fig. SB.14b), 2) accounting for the multi-layer structure of the network by collapsing the backbone-backbone interactions into residue-residue interactions (Fig. 4.4).

4.4 Results

We present results below for these three proteins (Fig. 4.1a). We first focus on the well-characterized allosteric protein FimH$_\mathrm{L}$. Separation of the FimH$_\mathrm{L}$ domain from its connecting domain (bottom in all figures) is thought to induce an allosteric conformational change on the opposite end of the protein (top in figures) [33], [34]. This changes the binding pocket from a state with low affinity for the ligand to one with high affinity (Fig. 4.1a) [33]. While wild type FimH$_\mathrm{L}$ is trapped in the high-affinity state, a single-amino acid mutation (Arg60Pro) stabilizes FimH$_\mathrm{L}$ in the low-affinity state [39], [79]. The mutant FimH$_\mathrm{L}$ is of interest because it undergoes an allostery-like confor-

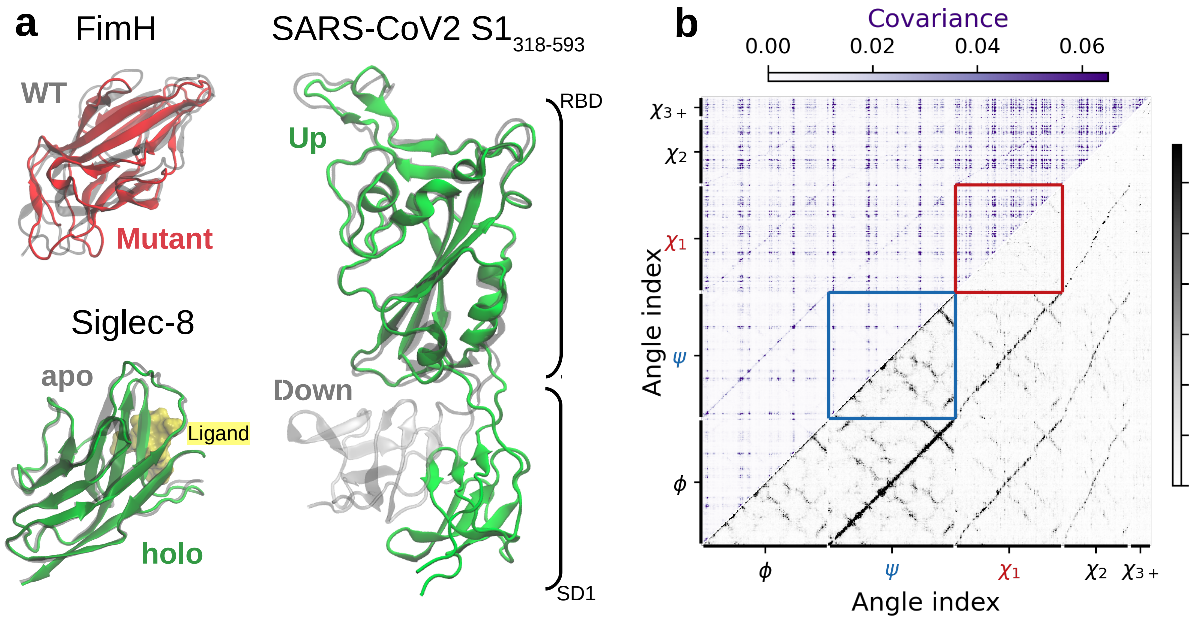

Figure 4.1: **Unraveling structural properties from protein conformational dynamics.** **a**, Cartoon illustrating the three adhesion proteins studied here. FimH$_L$ refers to the lectin domain of a bacterial adhesin found in uropathogenic *E. coli* that binds mannose and undergoes a conformational change under tensile force from urine flow. Siglec-8 refers to the lectin domain of a human immune-inhibitory protein found on eosinophils and mast cells. The SARS-CoV-2 RBD and SD1 domains are thought to undergo a down-to-up transition that makes the RBD available to bind ACE2. For each protein, we compare two states: FimH$_L$ wild type (PDB 4AUU) and mutant (PDB 5MCA), Siglec-8 with ligand 6'S-sLe^x (PDB 2N7B) and without (PDB 2N7A), and RBD-SD1 in down and up (PDB 6VSB). **b**, Comparison of covariance analysis of the dynamics (top left) versus the inverse covariance analysis (bottom right) from the dynamics of wild type FimH$_L$ (see Extended Data Fig. SB.1, B.2, and B.3 for the other proteins). While many studies rely on the analysis of the covariance matrix, our data clearly shows that the structure of the covariance matrix is dominated by artifacts (vertical and horizontal lines) which are stronger for side chain interactions (red square for $\chi_1 - \chi_1$). In contrast, the inverse covariance matrix clearly reveals a structure reminiscent of a contact map and is dominated by backbone interactions (blue square for $\psi - \psi$).

mational change upon binding mannoside ligands and has been proposed as a minimal model of allostery [39].

Like FimH_L, Siglec-8 binds a carbohydrate ligand and has an immunoglobulin-like fold with two β-sheets (Fig. 4.1a). However, Siglec-8 has a pre-formed binding pocket, leading to similar structures for the apo and holo states [40]. The SARS-CoV-2 spike protein RBD binds the human ACE2 receptor with a ”hook” region[45]. The hook becomes accessible in the up state when the hinge between the RBD and its connector opens. While the hook and interdomain hinge regions are flexible, the bulk of the RBD is structurally similar in the up and down states. Siglec-8 and the spike RBD-SD1 present an interesting challenge for detecting differences in inferred interactions where the protein state changes with small changes in structure within domains.

4.4.1 Define and validate network inference from inverse covariance analysis

Our approach for constructing a network representation of the dynamics of a given protein is comprised of three steps. In the first step, we obtain temporal dynamics for the nodes, which are the backbone (ϕ, ψ) and sidechain (χ_{1-5}) dihedral angles for each residue [15], [19]. In the second step, we calculate the circular covariance for dihedral angles [19], which can be thought of as the linearization of the interactions captured by mutual information. In the third step, we invert the covariance matrix using the Moore-Penrose pseudo-inverse to calculate the best fit for a linear coupling system that can give rise to the observed covariance matrix [20].

Similar to mutual information calculations conducted on other proteins [15], [102], we find that backbone-backbone interactions computed from the covariance are weak compared to sidechain-sidechain interactions (compare red and blue boxes in Fig. 4.1b and the mutual information in Extended Data Fig. SB.10). In these networks, some dihedral angles have a banding pattern, suggesting long-range interactions with many other dihedral angles (Fig. 4.1b). In contrast, the inverse covariance matrix has localized and specific interactions. In addition, the stronger backbone-backbone interactions have a repeating pattern that resembles the contact map of the protein, and this pattern appears to repeat more weakly in backbone-sidechain interactions.

The banding pattern in the covariance matrix is also widespread in mutual information matrices, where they been interpreted as long-range interactions important in protein allostery [15], [102]. In prior studies, the large number of long-distance edges produce ”hairball” networks, which led to the use of pruning algorithms [109], [110] or distance filters [26], [27], [111]. Thus, we wondered whether the long-range interactions are capturing a physical feature of the dynamics. To answer this question, we investigate the reproducibility of the covariance, correlation, and inverse covariance matrices extracted from different replicates of MD simulations.

In Fig. 4.2a, we contrast the matrix for one replicate in the upper-diagonal with the second replicate in the lower-diagonal and quantify the similarity in Fig. 4.2b. For replicate MD simulations, we used the same initial protein structure with randomized solvation and initial velocities. Both covariance matrices have banding patterns suggesting hotspots that interact with many residues across the protein. However, each replicate has its own banding pattern, with interaction strengths that are over ten times greater than those found in the other replicate, indicating high variability in the networks that one would construct from replicate simulations.

Since the banding pattern is associated with dihedral angles with high variance, we also consider the correlation matrix, which normalizes the covariance matrix by the variance of each dihedral angle. It is visually apparent that the banding in the covariance matrix is not simply due to high variance because there are still bands in the correlation matrix. While normalizing to the correlation matrix uncovers some interactions in a contact map pattern, they are weak compared to the banding pattern (Extended Data Fig. SB.4 for matrix with lower maximum value (95th percentile) for the color map scale).

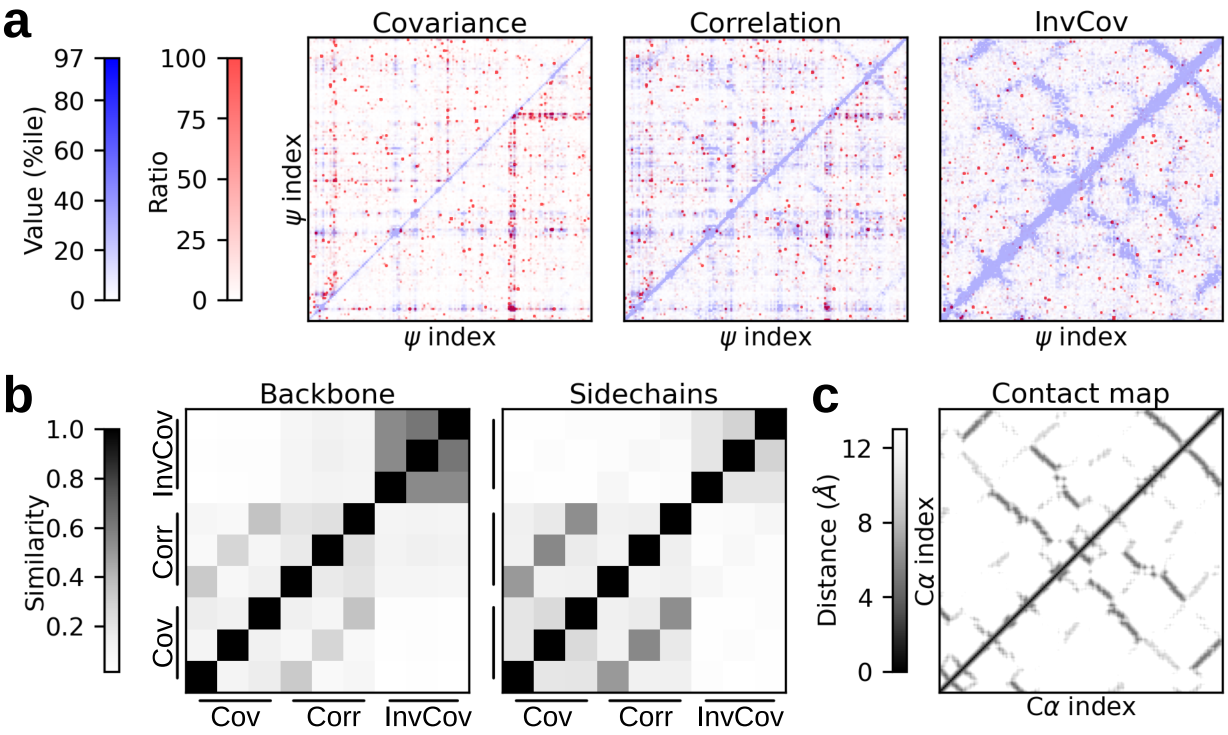

Figure 4.2: **Inverse covariance matrix is robust across replicates whereas covariance and correlation matrices are not.** **a**, Triangular regions above and below the matrix diagonal show results from two replicates of wild type FimH$_L$ starting from the same protein structure. We show $\psi - \psi$ interactions (see Extended Data Fig. SB.5 for $\chi_1 - \chi_1$ interactions). We show interaction strength in blue with a normalization for each triangular region made based on the 97th percentile of observed strengths. (See Extended Data Fig. SB.4 for weaker interactions visible when normalized to the 95th percentile.) In red, we show the ratio for interaction strength between the two replicates. Purple indicates strong interactions that are not reproduced in the other replicates. For covariance and correlation matrices, we find that backbone-backbone interactions are mostly quite weak, but the strong interactions ¿97th percentile vary drastically between replicates. In contrast, for the inverse covariance matrix, the strongest backbone-backbone interactions resemble the contact map and are symmetric across the diagonal (panel c). **b**, To evaluate the robustness of network inference, we calculate the Jaccard similarity coefficient for the covariance, correlation, and inverse covariance analyses methods across three simulation replicates. We define edges above the threshold of $\geq$97th percentile. (See Extended Data. Fig. SB.6 for other thresholds.) In grey scale, we show similarity separately for $\psi - \psi$ and $\chi_1 - \chi_1$ interactions. Darker grey indicates results are similar across replicates for the inverse covariance approach and much less similar for the other two methods. **c**, Cα inter-residue distance from the crystal structure. Darker grey indicates shorter distance.

In contrast to the irreproducible results obtained with the covariance matrix, for the inverse of the covariance matrix, we find a pattern that is visually similar to the 12Å contact map (Fig. 4.2c). The diagonally symmetric contact map pattern in blue indicates similarly strong interactions for two replicates (Fig. 4.2a). After quantifying the robustness across three replicates using the Jaccard similarity index, we find the inverse covariance has higher similarity (59-72% shared edges) than the covariance (8-10%) or the correlation (13-16%) for $\psi - \psi$ backbone interactions (Fig. 4.2b). The $\chi_1 - \chi_1$ similarity values for inverse covariance are lower, but still higher than for the other two methods. These data clearly demonstrates that networks inferred from inverse covariance analysis are more robust than networks constructed from correlational measures.

4.4.2 Inverse covariance analysis yields structural networks

Prompted by the strong visual resemblance between the inverse covariance network and the 12Å contact map and the complete absence of this pattern in the covariance network, we wondered if the inverse covariance matrix could be used to identify specific physical interactions. To answer this question, we overlaid the strongest edges ($\geq$ 98th percentile) on the 12Å contact map, highlighting the backbone-backbone edges as blue dots and the sidechain-sidechain edges as red crosses (Fig. 4.3a). To correct for the high variability across replicates we previously found for the covariance networks, we averaged networks across the three replicates. Despite the averaging, the covariance network shows strong interactions across distant protein regions and are dominated by $\chi_1 - \chi_1$ interactions. In contrast, the inverse covariance network is almost entirely restricted to the contact map. To understand how these contrasting patterns affect interpretation, we next visualize strong interactions as edges drawn on the protein structure (Fig. 4.3b). Since drawing all edges would make the covariance network indecipherable, we only show the edges originating from Lys4. These edges are difficult to justify physically because they are farther apart than the van der Waals cutoff

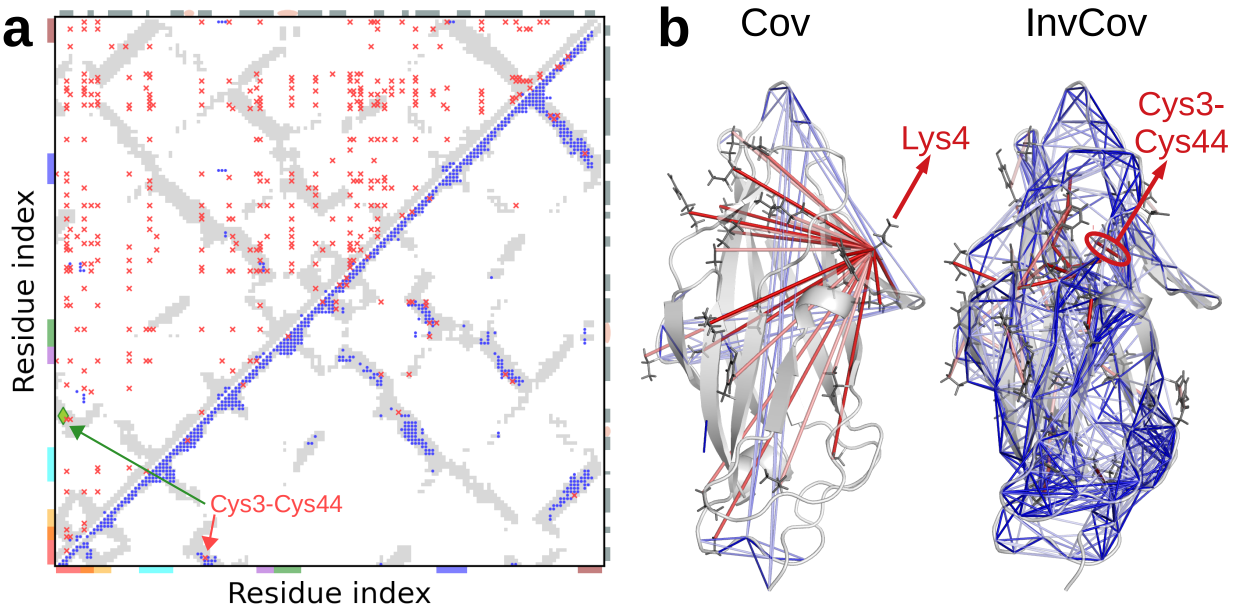

Figure 4.3: **The inverse covariance matrix enables us to extract a "contact map"-like network from the protein dynamics.** **a**, Comparison of strong interactions identified for the covariance matrix (top left) and for the inverse covariance matrix (bottom right) of wild type FimH. To provide context for our data, we plot the 12 Å contact map in grey within the matrix. On the top and right axes we show helix (pink) and strand (teal) secondary structures from DSSP. On the left and bottom axes we show putative allosteric pathway landmarks [79]: clamp loop (red), pocket zipper (yellow), swing loop (cyan), β-bulge (purple), α-switch (green), insertion loop (blue), and linker loop (dark red). The edges we show are the strongest dihedral interactions, with the threshold set at the 97th percentile of all dihedral interactions. See Extended Data Fig. SB.11 for other cutoffs. We averaged edge weight across three replicates. The blue dots represent the average of backbone-backbone interactions by residue. The red crosses represent sidechain-sidechain interactions ($\chi_1 - \chi_1$). The inverse covariance network are predominantly backbone interactions that fall within the 12 Å contact map. There is a $\geq$ 99th percentile $\chi_1 - \chi_1$ interaction the Cys3-Cys44 disulfide bond (pink arrow). However, this interaction is only 80th percentile in strength (green diamond) for the covariance matrix, which is dominated by other $\chi_1 - \chi_1$ interactions. **b**, Since the covariance matrix has many long-range interactions, we only show the backbone interactions and the sidechain interactions for Lys4. In contrast, the inverse covariance network has mostly short-range interactions, including the disulfide bond. We show backbone interactions on the Cα atoms and sidechain interactions on the the 4th χ_1 atoms. We show backbone interactions in blue and sidechain interactions in red. Darker colors indicate stronger interactions.

distance for MD simulations, and electrostatic interactions are largely screened by the solvent.

In contrast, the inverse covariance network uncovers edges that mostly connect physically close residues. Specifically, we find a $\chi_1 - \chi_1$ edge between Cys3-Cys44 for the sole disulfide bond in $\mathrm{FimH_L}$, and that this edge is missing from the covariance network (red vs green arrow in Fig. 4.3a).

Examining the contact map pattern of the inverse covariance network in more detail, we compare edge weight with the distance between Cα atoms (Extended Data Fig. SB.12). We find that for backbone-backbone interactions, the strongest interactions are between residues connected by a peptide bond, followed by hydrogen bonds within β sheets, and then non-bonding interactions.

After examining backbone-backbone interactions, we next looked at the progressively weaker interactions involving sidechains distal from the backbone (Fig. 4.1b and Extended Data Fig. SB.13). We find the contact map pattern is still apparent for $\phi - \chi_1$ or $\psi - \chi_1$ interactions, but becomes very weak for $\phi - \chi_2$ or $\psi - \chi_2$ interactions, and becomes indistinguishable from noise for interactions between proximal and distal sidechain dihedrals. The inverse covariance analysis thus suggests that backbone dihedral motion is most strongly coupled to nearby backbone dihedrals and has more dissipated effects on sidechain dihedrals. This relationship is consistent with how backbone motions can sterically trap or free sidechains, whereas sidechain motions are more limited in their impact on backbone motion [112].

The different strengths of interactions for backbone-backbone and backbone-sidechain edges suggests that qualitatively different types of interactions have different properties. For two residues *i* and *j*, the backbone-backbone edge $\phi - \psi[i, j]$ are larger than the backbone-sidechain edge $\chi_1 - \chi_1[i, j]$, which is consistent with the physical differences between these two edge types. Moreover, most backbone-backbone edges within the same residue, $\phi - \psi[i, i]$, are stronger than backbone edges connecting to other residues, $\phi - \phi$[i, j] and $\psi - \psi$[i, j] (Extended Data Fig. B.12).

4.4.3 Detecting both large and small structural changes in $\mathrm{FimH_L}$

4.4.3 Conformational differences between wild type and mutant

As a way to validate our approach, we next test if we are able to identify the well-characterized differences between wild type $\mathrm{FimH_L}$ and the Arg60Pro mutant. To compare inferred networks for the wild type and mutant proteins, we identified edges where the average difference was larger than two times the standard deviation across each group of replicates. We performed this analysis once with the entire set of edges (Extended Data Fig. SB.14), and again with only the backbone interactions collapsed into a residue interaction network. For our comparisons and the matrix visualization of the differences, we do not apply a distance filter or a threshold for the edge-strength. However, for the visualization on the protein, we only show differences with magnitude larger than the 97th percentile.

The visualization of the differences enables us to identify several interactions stronger in either the mutant or the wild type proteins (red or blue patches, respectively, in Fig. 4.4a). This is consistent with the difference in initial structure (RMSD=3.15Å) and in dihedral dynamics [107].

For concreteness, we focus on two regions at the edge of the protein structure that are easier to visualize: the binding pocket zipper at the top of $\mathrm{FimH_L}$, and the insertion loop at the bottom. In the pocket zipper, we found much stronger interactions for the mutant protein (median:2.8-fold, IQR:2.0–5.2-fold), which correspond to smaller dihedral fluctuations [107]. On the other hand, in the insertion loop, we identified changes in interaction that were stronger in the wild type than the mutant protein (3.2-fold, 2.2–3.7-fold). Structurally, this is consistent with how the insertion loop is stabilized in the wild type structure and exposed to solvent in the mutant protein [35], [39]. Dynamically, stronger interactions within the insertion loop is consistent with smaller dihedral fluctuations in the wild type protein [107].

We further identify differences at the β-bulge, α-switch, and swing loop regions of the allosteric pathway, consistent with structural differences between wild type and mutant proteins. The mutant protein has stronger interactions within the loop formed by the β-bulge (2.0-fold, 1.2–2.3-fold) and also with a nearby loop. In the wild type protein, the loop is smoothed out into a β-strand. The wild type protein has stronger interactions (1.8-fold, 1.5–2.6-fold) in the α-helix, compared to the 3_{10}-helix in the mutant protein, which is probably due to different hydrogen bonding patterns. In the swing loop, we again find stronger interactions in the wild type protein (2.0-fold, 1.6–2.7-fold).

For these allosteric pathway landmarks, it is visually apparent that we detect large differences in inferred interactions when structures are closer in one state and stretched apart in the other state. Beyond these regions, there are several other regions with similarly large changes in interaction between the wild type and mutant proteins, shown in blue and red patches (Fig. 4.4a).

4.4.3 Disulfide bond

We next used wild type FimH_L to explore the impact of removing the single disulfide bond between Cys3-Cys44 *in silico* on fast, nanosecond-timescale dynamics. Using the inverse covariance analysis, we correctly identified the 6-fold stronger χ_1-χ_1 interactions in the presence of the disulfide bond, which was the largest difference detected (Fig. 4.4d). This matches our expectations because the covalent bond between the most distal atoms forming the χ_1 rotamer angle directly couples χ_1 dynamics. Together, these analyses show that the inverse covariance analysis method is sensitive to both local differences and conformational differences.

4.4.4 Network rearrangement in the conserved region of Siglec-8 in the ligand-bound state

Like FimH_L, the human immune cell adhesion protein, Siglec-8, is also a lectin with an immunoglobulin-like fold with a single disulfide bond. For Siglec-8, we compare the apo (no ligand) and holo

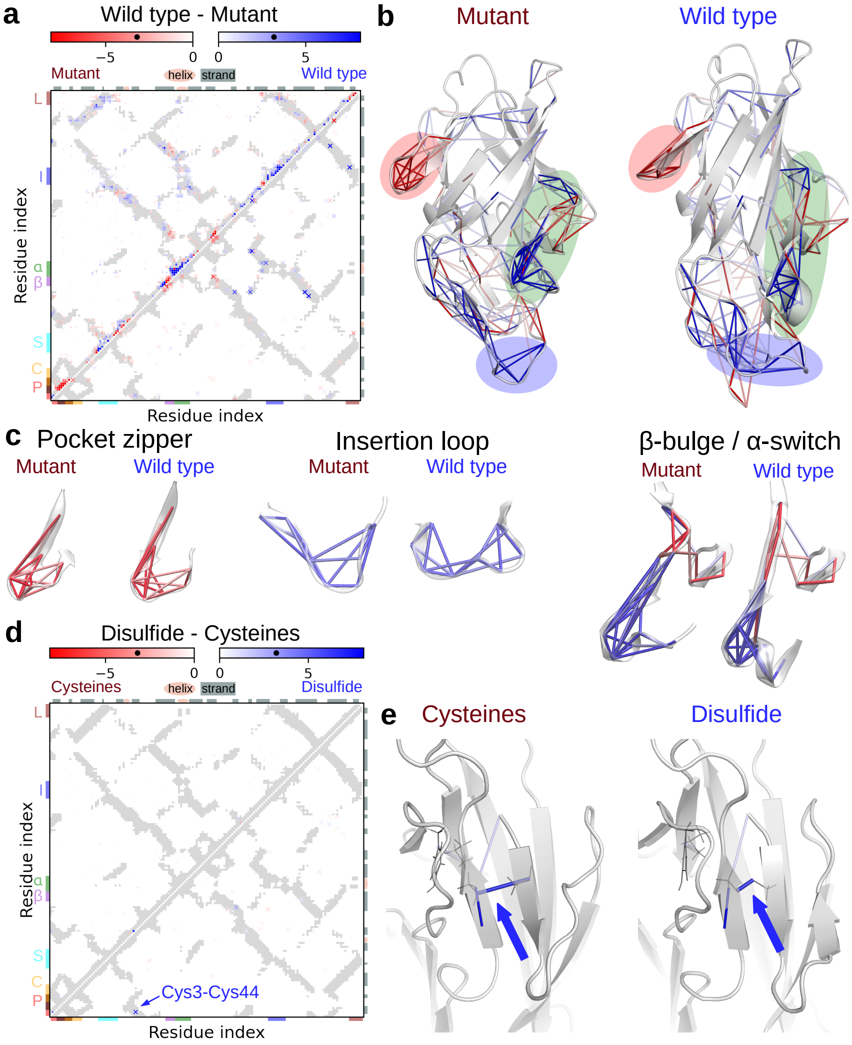

Figure 4.4: **Inverse covariance analysis can detect both large and small structural changes in FimH$_L$.** **a**, Comparing inferred networks for wild type and mutant proteins, we show differences in the backbone (top left in dots) and $\chi_1 - \chi_1$ (bottom right in crosses) over the 12 Å distance cutoff in grey. Red indicates stronger interactions for the mutant protein; blue for the wild type protein. The borders show secondary structure and landmarks as described in Fig. 4.3a. The black dots on the colorbar describe the 97$^{\text{th}}$ percentile in magnitude. Larger differences are drawn on the protein, while all differences greater than 2σ are plotted on the adjacency matrix. **b**, The pocket zipper (red), insertion loop (blue), and β-bulge/α-switch (green) are highlighted in **c**. **d**, Comparing wild type FimH$_L$ with the Cys3-Cys44 disulfide bond intact or broken *in silico*. **e**, The blue lines shows that the Cys3-Cys44 $\chi_1\chi_1$ (arrow) and the Phe43-Cys44 backbone-backbone interactions are stronger when the disulfide bond is intact.

(bound to the native 6'S-sLe$_x$ ligand) states. Apo and holo Siglec-8 have extremely similar structures due to the rigid binding pocket loops that only differ by a few sidechain rearrangements [40]. The rigidity of the CC' binding pocket loop in apo Siglec-8 occurs in the absence of stabilizing secondary structure motifs [40].

One hypothesized mechanism is that the Arg70 sidechain forms hydrogen bonds with the backbone of Pro57 and Asp60 within the CC' loop [40] (Fig. 4.5a). Using our inferred interaction networks, we were unable to detect these proposed hydrogen bonds. We only detected a moderate interaction between Arg70 and Pro62 (Extended Data Fig. SB.15. However, we were able to identify other strong interactions, such as the Cys31-Cys91 disulfide bond (7.1-fold stronger than the Arg70-Pro62 interaction) and the Arg79-Asp102 salt bridge (3.7-fold stronger). Analysis of the MD trajectories makes clear that Arg70, Pro57, and Asp60 have large fluctuations relative to each other, while the disulfide bond and salt bridge only have small fluctuations (Fig. 4.5b).

Moreover, the CC' loop backbone also has surprisingly small fluctuations for an unstructured loop (Fig. 4.5b). Using our inferred interaction approach, we identified strong interactions within the CC' loop, as well as interactions from outside the loop to its hinges at Ala53 and Pro62 (Extended Data Fig. SB.15). Taken together, these observations lead us to propose an alternative mechanism to hydrogen bonds between Arg 70 and the CC' loop. In our view, the CC' loop is stabilized internally and externally at the hinge edges.

Comparing apo and holo Siglec-8, the holo state has stronger χ_1-χ_1 interactions corresponding to the Cys31-Cys91 disulfide bond (Fig. 4.5c). The disulfide bond is conserved in the siglec family and is located on the sheet of the β-sandwich opposite the binding pocket [40]. Nearby, we also observe other changes in interaction strength involving Asp90. Although distant from the binding site, the Asp90-Cys91-Ser92 motif in Siglec-8 is a variant of the Asn-Cys-Ser or -Thr motif that is conserved in the rest of the siglec family [113]. Differences in interaction strength between apo and holo states identifies changes in the dynamics of this evolutionarily conserved region during ligand-binding.

In contrast, removing the Cys31-Cys91 disulfide bond from the holo state *in silico* has a different pattern (Extended Data Fig. SB.16). Both ligand-binding and the presence of the disulfide bond have a stronger χ_1-χ_1 interaction at Cys31-Cys91. Both conditions also stabilize Cys31, indicated by decreased backbone dihedral fluctuations and increased duration within an extended secondary structure as assigned by the Dictionary of Secondary Structure of Proteins (DSSP) algorithm, and shorter Cys31-Cys91 Cα distance. Taken together, we find that the network rearrangement that occurs with ligand-binding increases Siglec-8 conformational stability near an evolutionarily conserved disulfide bond, even though the region is not near the binding site, and the protein has similar structure in the apo and holo states.

4.4.5 Comparison of the spike protein RBD in the up and down states shows network rearrangement without large structural differences

Next, we investigated whether there are differences in the networks inferred from the 'up' and 'down' states of the RBD-SD1 domains of the SARS-CoV-2 spike protein (Fig. 4.1a). The RBD connects to SD1 (Fig 4.5e) via two hinge-like loops that are more flexible in the up state than the down state [114]. While the down and up states have different orientations around the hinge, they have similar SD1 structures (Cα-RMSD=0.64Å) and somewhat similar RBD structures (Cα-RMSD=1.55Å).

Opening the hinge angle in the down-to-up transition is thought to make the binding site on the RBD domain available to attach to human ACE2 [45], [46]. To focus on the RBD-SD1 hinge, we isolated these domains from the rest of the spike protein and ignored glycosylated sugars. While these simplifications limit the strength of our conclusions into the function of the spike protein, it

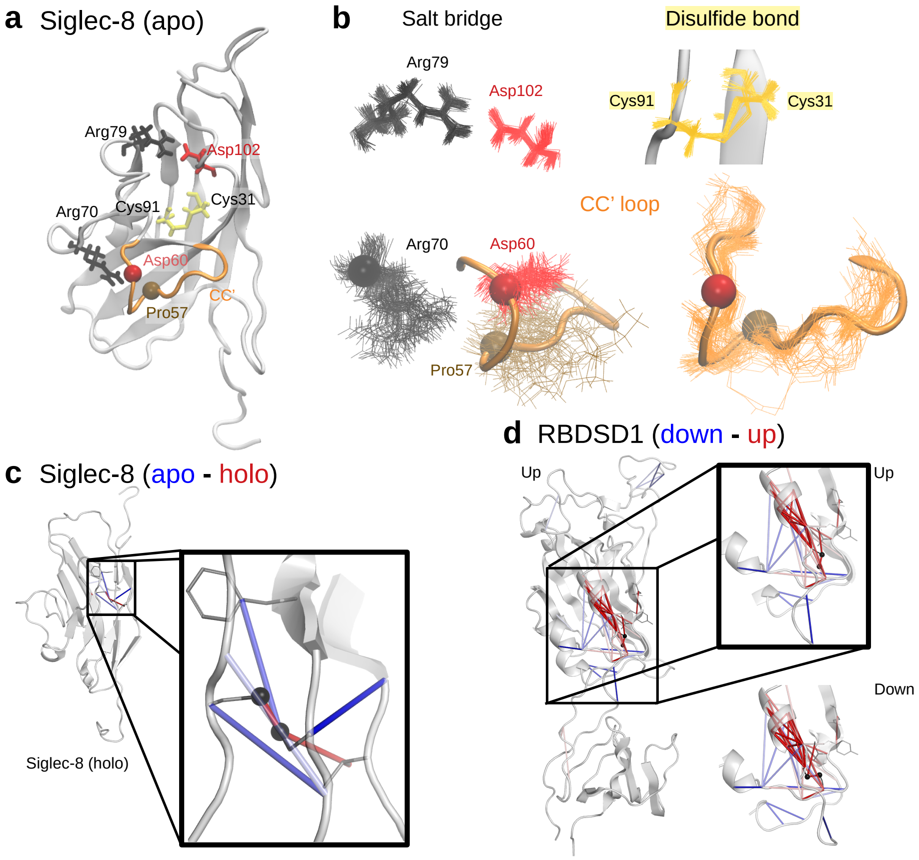

Figure 4.5: **Inferred networks identify strong interactions and changes in interaction strength.** **a**, Illustration of Siglec-8 highlighting the CC' loop of the binding pocket in orange and the wire diagrams of the residues involved in the Arg79-Asp102 salt-bridge, the Cys31-Cys91 disulfide bond, and Arg70, which is hypothesized to form hydrogen bonds with the CC' loop. It has been recently hypothesized that the CC' loop is stabilized by hydrogen bonds from Arg70 to Pro57 and Asp60, shown as spheres at their Cα positions. **b**, Comparison of the dynamics for the interactions highlighted in a. We show the superposition of aligned snapshots taken every 6 ns. **c**, Our approach also identifies rearrangements of strong interactions without large structural changes. Comparing Siglec-8 with (holo) and without (apo) ligand reveals rearrangement of strong interactions in a region opposite the binding pocket, including the Cys31-Cys91 disulfide bond (black spheres). We show interactions stronger in holo (red) and apo (blue) on the holo structure. **d**, For the SARS-CoV-2 spike protein, rotation around the RBD-SD1 hinge changes the structure in the down and up states, while the individual domains have similar structure. We detect differences near the hinge, including the Cys336-Cys361 disulfide bond (black spheres) and nearby α-helix. We show interactions stronger in the down (blue) and up (red) states on both structures.

nonetheless provides a useful system for comparing dynamics in a system that initially resembles rigid-body motion around a hinge.

As seen in Fig 4.5e, several strong network interactions are rearranged for the up and down states near the α-helix at residues 338-344. For example, the down state has stronger interactions at several edges connecting Cys336 to residues within the helix. However, those residues in the helix have different interactions in the up state, such as with Leu335 or with Asp364 on a neighboring loop. In this region, we also find that the down state has stronger interactions at the Cys336-Cys361 disulfide bond. In the down state, we also find that Cys331 has smaller backbone dihedral fluctuations and stays longer in the β-sheet secondary structure.

We also compared both the up and down states with the ‘off’ state, from a spike protein where all the RBD are hidden. We find that the up and down states are more similar to each other than to the off state (Extended Data Fig. SB.17. This suggests that after one protomer reaches the up state, there may be a smaller set of rearrangements in the down-to-up transition than the larger set of arrangements required for the first protomer to reach the up state.

As a proof-of-concept, we investigated whether it is possible to infer an interaction network for the trimeric SARS-CoV-2 spike protein, which is much larger than the RBD-SD1 domains. However, the amount of data required for network inference using the inverse covariance method scales faster than the number of residues. As a result, our approach required more data than was available from the two sets of publicly-accessible simulations for which the up and down states are labeled [115], [116]. Instead, we chose the longer simulations that explore the transition among the down, up, and open states published by the Bowman lab [6]. Using 100,000 snapshots, we were able to infer an interaction network for the 3,363 residues of the spike protein (Extended Data Fig. SB.18). The number of snapshots is an order of magnitude larger than those currently available for labeled states. Thus, it may be feasible to compare states for the entire S1/S2 complex of the spike protein if there is sufficient data, or by choosing a less data-intensive method for solving the inverse problem.

4.5 Discussion

We identify some of the shortcomings of correlation-based approaches for network inference from protein dynamics using the covariance, correlation, and mutual information matrices. These networks have low reproducibility among replicates and exhibit long-range connections that are difficult to tie to physical explanations. To address these shortcomings, we use the inverse of the covariance matrix. This is a well-established technique from network inference[20] for solving the inverse problem for a system that can produce the observed correlated dynamics. Our approach builds networks where each node is a dihedral angle, including both the backbone (ϕ, ψ) and the sidechains (χ_{1-5}), and edges are inferred from the coupling interactions between angles. We chose the internal coordinate system of dihedral angles[19] to easily include sidechain dynamics, localize hinges that drive distal dynamics[107], and to avoid the alignment step in Cartesian coordinates that introduces ‘artifacts’ in hinged motion [58].

Using the inverse covariance approach, we detected differences in conformation, subtle differences between protein states without large conformational changes, and localized perturbations in the structure of biomedically important proteins. The inverse covariance networks capture a hierarchy of interactions that resemble that qualitatively different types of interactions, suggesting a multi-layer network structure. The strongest edges connect dihedral angles with covalently-bonded atoms, with weaker interactions for greater distances. Moreover, the contact map-like pattern found in backbone-backbone interactions are repeated more weakly in backbone-sidechain and sidechain-sidechain interactions. This hierarchy of inferred interactions is consistent with the smaller backbone rearrangements related to larger sidechain motions [112].

Our results suggest that solving the inverse problem uncovers the underlying interactions that ultimately drive protein dynamics, but are not well-captured by cataloguing the observed correlated motions or comparing static structures. However, inverting the covariance matrix is the simplest of a variety of tools available for network inference from dynamics used in the network science field [20], [72]. While the simplicity of inverting the covariance matrix increases accessibility, there are some obvious limitations [20]. We calculate the circular covariance matrix on dihedral angle distributions that are multi-modal. The inverse of the covariance matrix is analogous to linearly coupled torsional springs, which do not represent the complexity of atomic interactions within a protein. Moreover, network inference by inverting the covariance matrix requires a large amount of data [20]. Our work establishes a baseline approach, which can be easily built upon by incorporating more sophisticated [20], [65] - and yet more involved - approaches that better describe dihedral distributions [117] or account for nonlinear interactions [70].

Despite these limitations, our approach yielded insights for three adhesion proteins. Comparing the networks inferred for two protein states at a time, we were able to tie differences in inferred network structure to structural and dynamical differences. For $\mathrm{FimH_L}$, a comparison of inferred networks for the wild type and mutant proteins identifies protein regions with conformational changes consistent with the allosteric pathway sites [79]. For Siglec-8 and the SARS-CoV-2 RBD-SD1 construct, we were able to detect network rearrangements despite the similar structures of Siglec-8 in the apo and holo states, and of the individual RBD and SD1 domains in the up and down states. In Siglec-8, we were also able to use strong interactions identified by the network to propose a mechanism for stabilizing the unstructured, yet rigid, binding pocket. Thus, we show that the inverse problem approach can identify protein regions of interest by comparing dynamical differences, even in the absence of large conformational differences.

CHAPTER 5
CONCLUSION AND FUTURE WORK

5.1 Summary of key findings and significance

Protein structure and dynamics are essential for function. However, some changes in protein function are better captured by changes in dynamics, without an obviously identifiable change in structure, such as in the case of dynamical allostery [1], [38]. To capture dynamical differences, molecular dynamics (MD) simulations are becoming an increasingly common tool for analyzing protein dynamics. As a result, the amount of MD trajectory data available to the scientific community is also growing [8], and will likely continue to grow with advances in computing [5] and developments in protein structure determination and prediction [7].

There is a correspondingly growing need for automated analysis of MD trajectory data. Well-characterized proteins have experimentally-identified protein regions that are known to be important for function and regulation. These regions serve as guideposts for focusing the analysis of MD trajectories. However, many proteins are not well-characterized, and MD simulations provide an inexpensive and high-throughput tool for protein characterization [3], [4]. In the absence of previously-identified regions of interest, analyzing the dynamics of the protein in its entirety to identify emergent properties and regions of interest may provide an avenue for automated and unbiased characterization.

One framework for identifying emergent properties that has previously been applied to protein structure and dynamics is network analysis [13], [17], [26], [62], [118]. Constructing a network for a protein makes it possible to apply a variety of analysis tools from the field of network science.

These tools can potentially identify protein regions that are vulnerable, important in communication, or a smaller part of a larger region. However, it is not yet clear how to select from the variety of methods available for network construction. Moreover, it is also not well understood how to select the subsequent network analysis tools to study a particular function.

In this work, we describe a potential approach for constructing and analyzing networks from MD trajectories. We focus on the dynamics of the bacterial adhesion protein, FimH, as a well-characterized model protein for allostery [39]. The allosteric site of FimH is located between its two domains and acts like a force-sensor [81]. Domain separation exposes the allosteric site to water and is thought to trigger a conformational change in the domain with the binding pocket [33], [79], [86], increasing affinity for ligand. This regulation of the binding function is allosteric because the force-sensing site is distant from the binding pocket [38]. In FimH, the protein regions of interest along the allosteric pathway were previously identified experimentally and computationally [35], [79].

Chapter 3: Intriguingly, isolating the ligand-binding, lectin domain keeps FimH in the high-affinity state, but a single mutation in one of the allosteric pathway sites keeps the isolated lectin domain in the low-affinity state. As a result, the single lectin domain of FimH was proposed as a model system for allostery [39]. This work provided the first use of MD simulations to characterize the dynamics of the inactivating Arg-60-Pro mutation. Analysis of MD trajectories for FimH is complicated by the hinged motion between the two domains. The structural alignment step used in previous analyses of FimH can introduce ‘artifacts’ [58] that can be avoided using an internal coordinate system [19]. As a result, we performed our analyses with an internal coordinate system of dihedral angles. In addition to the backbone dihedral angles, we also included an analyses of the sidechain dihedral angles. Since we wanted to directly compare sidechains with different lengths due to the mutation, we introduced a pseudo-dihedral angle that made it possible to compare sidechain dynamics for different amino acids (except for Gly). Our analysis was able to identify new regions of interest based on dynamics and localize the hinge regions that drive the motion of distal protein regions. This demonstrated the advantages of using an internal coordinate system to characterize FimH dynamics.

Chapter 4: Using the data from *Chapter 3*, we next sought to construct networks from protein dynamics. Based on the ideas sketched in *Chapter 2*, we decided to construct networks from the inverse of the covariance matrix of dihedral angles. The procedure is analogous to constructing anisotropic elastic networks in Cartesian coordinates [17], [18], except we used the internal coordinate system of dihedral angles. To assess our choice, we compared the inverse covariance matrix against correlative measures such as the covariance, correlation, and mutual information matrices. We found that the inverse covariance matrix was more robust across technical replicates and physically interpretable. In particular, the backbone-backbone interactions are much stronger than most sidechain-sidechain interactions. The network also captures a hierarchy of interaction strengths that suggest qualitatively different types of interactions.

After constructing networks, we then compared the networks for the lectin domain of FimH, with and without the mutation. This procedure identified protein regions of interest that matched the landmarks along the allosteric pathway [79]. Next, we extended our analyses to two other adhesion proteins, the human immune protein Siglec-8, and the binding domain of the SARS-CoV-2 spike protein. For Siglec-8, we used network analysis to propose a mechanism to explain the dynamics of the binding pocket. We also identified differences in dynamics at an evolutionarily conserved protein region on the opposite of the protein from the binding pocket. Similarly, for the SARS-CoV-2 spike protein, we identified differences in dynamics at a protein region thought to be involved in a hinged motion that makes the binding site available for viral attachment [45], [46].

5.2 Limitations

5.2.1 Molecular dynamics simulations

The use of MD simulations and our selection of network construction and analysis techniques have provided insights into analyzing protein dynamics. However, these techniques have limitations, and we are still refining our analyses.

While MD simulations have provided data about protein dynamics at small spatiotemporal resolutions, our simulations that go up to several hundred nanoseconds are not long enough to capture the FimH allosteric conformational change [4]. As a result, we could only compare the dynamics of FimH in different states, instead of directly observing the dynamics involved in the conformational change. Our simulations also depend on the accuracy of the experimental procedures used to solve the protein structure. While Siglec-8 was solved using NMR, we used crystal structures for FimH, and a combination of crystal and Cryo-EM structures for the spike protein.

Moreover, protein dynamics from MD simulations deviate from those observed experimentally [119]. In particular, protein dynamics are influenced by the dynamics of the surrounding water. The TIP3P model of water used in this work may limit the observation of coupled protein dynamics that might be relevant to studying allostery [56]. In the future, as simulation engines and force fields improve, it may be more less problematic to apply network inference to more accurate MD trajectory data.

5.2.2 Network construction

For network inference from protein dynamics, we selected a very simple approach that is often used in the field of network science as a first-pass analysis for inferring the underlying structure [20]. We first capture correlated dynamics using the (circular) covariance matrix, and then calculate the pseudo-inverse matrix [20], [60]. The first step does not account for the multi-modal distribution of dihedral angles. The second step does not account for nonlinear coupling interactions. Our network inference simplifies the protein into a system of coupled torsional springs (see *Chapter 2*), which does not capture the full complexity of interactions within a protein. Lastly, the pseudo-inverse can fail dramatically with insufficient or very noisy data. Many of these issues can be avoided with more sophisticated algorithms for solving the inverse problem for network inference [20], [60], [70].

5.3 Opportunities for future research

5.3.1 Understanding protein dynamics

This exploratory work into one method for network inference from protein dynamics adds to the available approaches for abstracting proteins as networks. It opens the door for applying a variety of more sophisticated and computationally efficient methods from the network and data science fields to MD datasets. Some, but not all, protein regions in physical proximity are encoded as strong interactions in the inferred network. Thus, network inference may provide another tool on the way to exploring the fundamental relationship between protein structure and dynamics. It may help answer questions such as, how do we predict the impact of small variations in structure on dynamics? The ability to accurately predict the impact of structural perturbations on dynamics adds another dimension to protein engineering and drug design.

In the nearer future, the network inference approach may provide a way for comparing protein states, beyond those studied here. For example, there are multiple sequence variations of FimH among different strains of *Escherichia coli* and in another genus, *Klebsiella*. There are also multiple crystal structures available for FimH, which are further modified by different ligands. While comparative studies of variations in sequence [120] and structure [121] have provided insights into function, the dynamics from MD simulations have not yet been systematically compared. In a similar vein, FimH has similar structure to other bacterial proteins, some of which bind to sugars like FimH. Others bind protein ligands at a different protein region, using a different mechanism. The dynamics of these proteins, variants, and mutated forms, have not been systematically compared either [122].

This can be extended to other proteins to study how dynamics are regulated by variations in structure and sequence - including variations found in nature or introduced in the laboratory through deep mutational scanning or random mutagenesis. In addition to protein engineering and drug design, this may one day provide a biophysical link between variations of unknown significance and potential disease states.

5.3.2 Network analysis to study emergent properties

From a network science perspective, constructing a network is merely the first step. Among the options for analyzing networks, two interesting paths for proteins are community detection to identify protein regions and path analysis.

Proteins have a hierarchy of structure that affects dynamics, and our work suggests this hierarchy can be captured via network inference from protein dynamics. Using inferred networks avoids the need to introduce structural information, such as distance masks or secondary structure assignments. Thus, community detection on these inferred networks may provide a more unbiased way to identify protein regions that move in a coordinated fashion. Community detection also enables the comparison of protein states at a higher level than individual edges, which can better capture differences in emergent dynamical behavior (see Fig. 5.1. It is not yet established how to select a method for community detection on protein networks, or how to test the validity of the resulting partition.

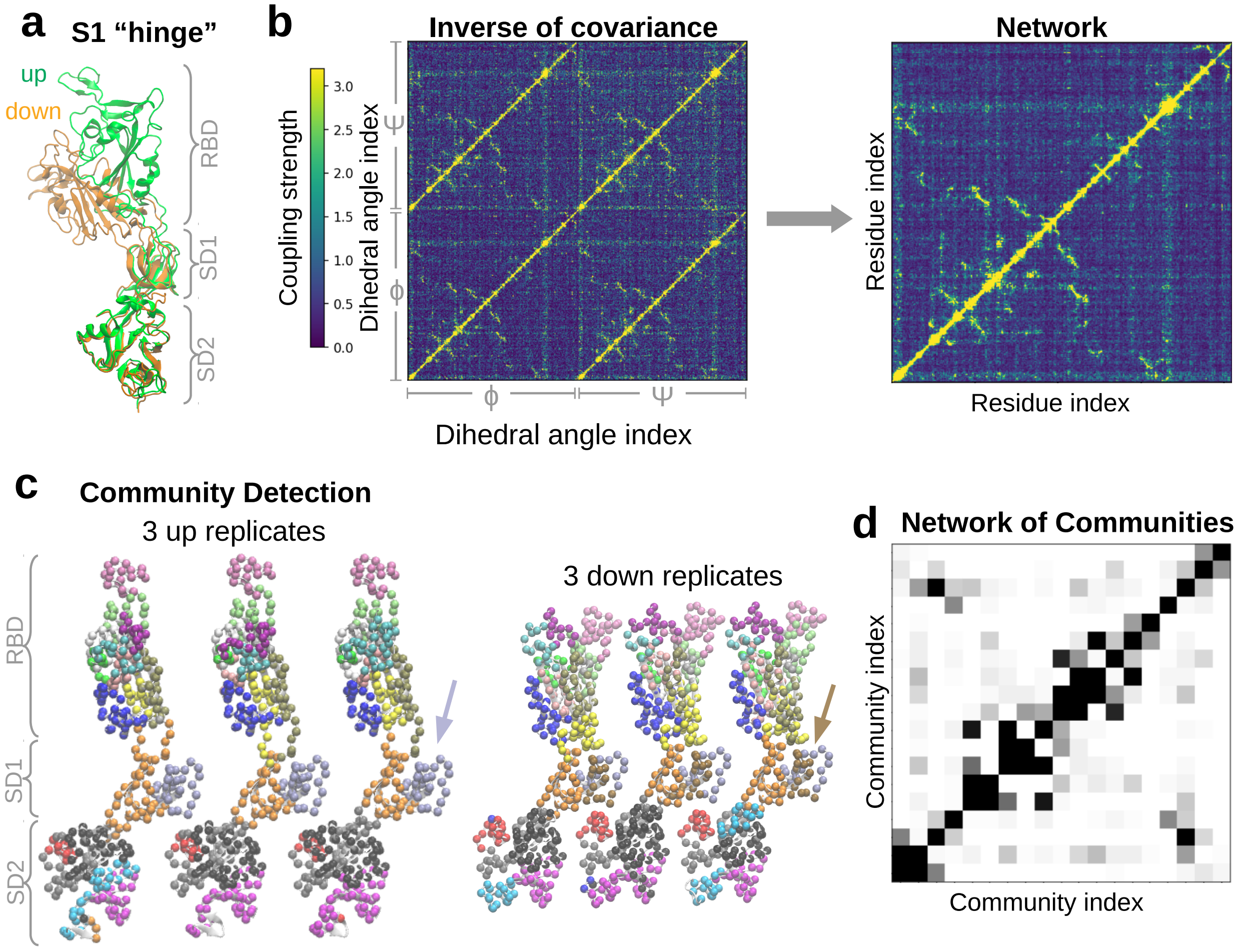

Figure 5.1: **Procedure for one approach to community detection from protein networks.** **a**, Spike protein ”hinge” region involved in the down-to-up transition consists of the receptor binding domain (RBD), sub-domain 1 (SD1), and SD2. **b**, From all-atomistic MD simulations, we use dynamics data to infer an interaction network. We then average the coupling strengths by residue to get a residue interaction network. We have shown that this network resembles the protein contact map for this protein and others (see *Chapter 4*). **c**, We perform network analysis for community structure using degree-corrected stochastic block modeling. Communities can be thought of as inter-connected regions of the protein with coupled motion. We find differences in community structure between the up and down states. For example, in SD1, the grey community in the up state is split into two communities (grey and brown) in the down state. This suggests that coupled motion in the SD1 domain of the up state is disrupted in the down state. **d**, In the future, we will also compare differences in coupling among communities. Here we show the network of communities.

The organization of individual edges into paths provides another approach for characterizing networks. One model for protein allostery focuses on paths connecting the (regulatory) allosteric and the (regulated) functional sites [65], [123]. The biophysical explanation for these allosteric pathways is the transmission of perturbations across the protein. The network analogy is particularly useful in allostery because it does not require a direct connection between the allsoteric and functional sites. Instead, the network structure can give rise to long-range effects as an emergent property of the combination of short-range interactions. It may be possible to validate the idea that methods to predict allostery from structure are indeed using structure to infer dynamics [65], by examining the networks inferred from dynamics. Moreover, this might also provide a way to integrate structural and dynamical information to study allostery that requires a change in structure, such as in FimH.

REFERENCES

[1] K. Henzler-Wildman and D. Kern, “Dynamic personalities of proteins,” *Nature*, vol. 450, no. 7172, pp. 964–972, Dec. 2007.

[2] W. Pitsawong, V. Buosi, R. Otten, R. V. Agafonov, A. Zorba, N. Kern, S. Kutter, G. Kern, R. A. Pádua, X. Meniche, and D. Kern, “Dynamics of human protein kinase aurora a linked to drug selectivity,” *eLife*, vol. 7, Jun. 2018.

[3] S. A. Hollingsworth and R. O. Dror, “Molecular Dynamics Simulation for All,” *Neuron*, vol. 99, no. 6, pp. 1129–1143, 2018.

[4] M. C. Zwier and L. T. Chong, “Reaching biological timescales with all-atom molecular dynamics simulations,” *Current Opinion in Pharmacology*, vol. 10, no. 6, pp. 745–752, Dec. 2010.

[5] C. T. Lee and R. E. Amaro, “Exascale computing: A new dawn for computational biology,” *Computing in Science and Engineering*, vol. 20, no. 5, Sep. 2018.

[6] M. I. Zimmerman, J. R. Porter, M. D. Ward, S. Singh, N. Vithani, A. Meller, U. L. Mallimadugula, C. E. Kuhn, J. H. Borowsky, R. P. Wiewiora, M. F. Hurley, A. M. Harbison, C. A. Fogarty, J. E. Coffland, E. Fadda, V. A. Voelz, J. D. Chodera, and G. R. Bowman, “SARS-CoV-2 simulations go exascale to predict dramatic spike opening and cryptic pockets across the proteome,” *Nature Chemistry*, vol. 13, no. 7, pp. 651–659, May 2021.

[7] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A. W. Senior, K. Kavukcuoglu, P. Kohli, and D. Hassabis, “Highly accurate protein structure prediction with AlphaFold,” *Nature*, pp. 1–11, Jul. 2021.

[8] A. Elofsson, B. Hess, E. Lindahl, A. Onufriev, D. v. d. Spoel, and A. Wallqvist, “Ten simple rules on how to create open access and reproducible molecular simulations of biological systems,” *PLOS Computational Biology*, vol. 15, no. 1, e1006649, 2019.

[9] T. Meyer, M. D'Abramo, A. Hospital, M. Rueda, C. Ferrer-Costa, A. Pérez, O. Carrillo, J. Camps, C. Fenollosa, D. Repchevsky, J. L. Gelpí, and M. Orozco, "MoDEL (Molecular Dynamics Extended Library): A Database of Atomistic Molecular Dynamics Trajectories," *Structure*, vol. 18, no. 11, pp. 1399–1409, Nov. 2010.

[10] M. W. van der Kamp, R. D. Schaeffer, A. L. Jonsson, A. D. Scouras, A. M. Simms, R. D. Toofanny, N. C. Benson, P. C. Anderson, E. D. Merkley, S. Rysavy, D. Bromley, D. A. Beck, and V. Daggett, "Dynameomics: A Comprehensive Database of Protein Dynamics," *Structure*, vol. 18, no. 4, pp. 423–435, Mar. 2010.

[11] U. Alon, "Biological networks: The tinkerer as an engineer," *Science*, vol. 301, no. 5641, pp. 1866–1867, Sep. 2003.

[12] S. Itzkovitz, R. Levitt, N. Kashtan, R. Milo, M. Itzkovitz, and U. Alon, "Coarse-graining and self-dissimilarity of complex networks," *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, vol. 71, no. 1, p. 016 127, Jan. 2005.

[13] L. Di Paola, M. De Ruvo, P. Paci, D. Santoni, and A. Giuliani, "Protein contact networks: An emerging paradigm in chemistry," *Chemical Reviews*, vol. 113, no. 3, pp. 1598–1613, Mar. 2013.

[14] P. Holme, *Form and Function of Complex Networks — Enhanced Reader*. Umea: Print & Media, Umea, 2004, ISBN: 91-7305-629-4.

[15] K. H. DuBay, J. P. Bothma, and P. L. Geissler, "Long-range intra-protein communication can be transmitted by correlated side-chain fluctuations alone," *PLoS Computational Biology*, vol. 7, no. 9, Sep. 2011.

[16] K. Kasahara, I. Fukuda, and H. Nakamura, "A novel approach of dynamic cross correlation analysis on molecular dynamics simulations and its application to Ets1 dimer-DNA complex," *PLoS ONE*, vol. 9, no. 11, e112419, Nov. 2014.

[17] A. R. Atilgan, S. R. Durell, R. L. Jernigan, M. C. Demirel, O. Keskin, and I. Bahar, "Anisotropy of fluctuation dynamics of proteins with an elastic network model," *Biophysical Journal*, vol. 80, no. 1, pp. 505–515, 2001.

[18] K. Moritsugu and J. C. Smith, "Coarse-grained biomolecular simulation with REACH: Realistic extension algorithm via covariance hessian," *Biophysical Journal*, vol. 93, no. 10, pp. 3460–3469, Nov. 2007.

[19] A. Altis, P. H. Nguyen, R. Hegger, and G. Stock, “Dihedral angle principal component analysis of molecular dynamics simulations,” *Journal of Chemical Physics*, vol. 126, no. 24, p. 244 111, Jun. 2007.

[20] H. C. Nguyen, R. Zecchina, and J. Berg, “Inverse statistical problems: from the inverse Ising problem to data science,” *Advances in Physics*, vol. 66, no. 3, pp. 197–261, Jul. 2017.

[21] I. Sillitoe, N. Bordin, N. Dawson, V. P. Waman, P. Ashford, H. M. Scholes, C. S. Pang, L. Woodridge, C. Rauer, N. Sen, M. Abbasian, S. Le Cornu, S. D. Lam, K. Berka, I. H. Varekova, R. Svobodova, J. Lees, and C. A. Orengo, “CATH: Increased structural coverage of functional space,” *Nucleic Acids Research*, vol. 49, no. D1, pp. D266–D273, Jan. 2021.

[22] K. A. Henzler-Wildman, M. Lei, V. Thai, S. J. Kerns, M. Karplus, and D. Kern, “A hierarchy of timescales in protein dynamics is linked to enzyme catalysis,” *Nature*, vol. 450, no. 7171, pp. 913–916, Nov. 2007.

[23] L. Rundqvist, J. Ådén, T. Sparrman, M. Wallgren, U. Olsson, and M. Wolf-Watz, “Noncooperative folding of subdomains in adenylate kinase,” *Biochemistry*, vol. 48, no. 9, pp. 1911–1927, Mar. 2009.

[24] S. Fortunato and D. Hric, “Community detection in networks: A user guide,” *Physics Reports*, vol. 659, pp. 1–44, Nov. 2016.

[25] J. Eargle and Z. Luthey-Schulten, “NetworkView 3D display and analysis of protein RNA interaction networks,” *Bioinformatics*, vol. 28, no. 22, pp. 3000–3001, Nov. 2012.

[26] Y. Karami, T. Bitard-Feildel, E. Laine, and A. Carbone, “”Infostery” analysis of short molecular dynamics simulations identifies highly sensitive residues and predicts deleterious mutations,” *Scientific Reports*, vol. 8, no. 1, p. 16 126, Dec. 2018.

[27] M. C. R. Melo, R. C. Bernardi, C. de la Fuente-Nunez, and Z. Luthey-Schulten, “Generalized correlation-based dynamical network analysis: a new high-performance approach for identifying allosteric communications in molecular dynamics trajectories,” *The Journal of Chemical Physics*, vol. 153, no. 13, p. 134 104, Oct. 2020.

[28] A. P. Kornev and S. S. Taylor, “Dynamics Driven Allostery in Protein Kinases,” *Trends in Biochemical Sciences*, vol. 40, no. 11, pp. 628–647, Nov. 2015.

[29] T. F. Guclu, A. R. Atilgan, and C. Atilgan, “Dynamic Community Composition Unravels Allosteric Communication in PDZ3,” *The Journal of Physical Chemistry B*, vol. 125, no. 9, pp. 2266–2276, Mar. 2021.

[30] C. Lee and D. J. Wilkinson, “A review of stochastic block models and extensions for graph clustering,” *Applied Network Science*, vol. 4, no. 1, pp. 1–50, Dec. 2019.

[31] T. Funke and T. Becker, “Stochastic block models: A comparison of variants and inference methods,” *PLoS ONE*, vol. 14, no. 4, e0215296, Apr. 2019.

[32] L. K. Mydock-McGrane, T. J. Hannan, and J. W. Janetka, “Rational design strategies for FimH antagonists: new drugs on the horizon for urinary tract infection and Crohn’s disease,” *Expert Opinion on Drug Discovery*, vol. 12, no. 7, pp. 711–731, Jul. 2017.

[33] I. L. Trong, P. Aprikian, B. A. Kidd, M. Forero-Shelton, V. Tchesnokova, P. Rajagopal, V. Rodriguez, G. Interlandi, R. Klevit, V. Vogel, R. E. Stenkamp, E. V. Sokurenko, and W. E. Thomas, “Structural basis for mechanical force regulation of the adhesin FimH via finger trap-like beta sheet twisting,” *Cell*, vol. 141, no. 4, pp. 645–655, 2010.

[34] M. M. Sauer, R. P. Jakob, J. Eras, S. Baday, D. Eriş, G. Navarra, S. Bernèche, B. Ernst, T. Maier, and R. Glockshuber, “Catch-bond mechanism of the bacterial adhesin FimH,” *Nature Communications*, vol. 7, no. 1, p. 10 738, Apr. 2016.

[35] G. Interlandi, W. E. Thomas, I. G, and T. WE, “Mechanism of allosteric propagation across a β-sheet structure investigated by molecular dynamics simulations,” *Proteins: Structure, Function and Bioinformatics*, vol. 84, no. 7, pp. 990–1008, Jul. 2016.

[36] W. E. Thomas, V. Vogel, and E. Sokurenko, *Biophysics of catch bonds*, Jun. 2008.

[37] E. V. Sokurenko, V. Vogel, and W. E. Thomas, “Catch-Bond Mechanism of Force-Enhanced Adhesion: Counterintuitive, Elusive, but ... Widespread?” *Cell Host and Microbe*, vol. 4, no. 4, pp. 314–323, Oct. 2008.

[38] S. J. Wodak, E. Paci, N. V. Dokholyan, I. N. Berezovsky, A. Horovitz, J. Li, V. J. Hilser, I. Bahar, J. Karanicolas, G. Stock, P. Hamm, R. H. Stote, J. Eberhardt, Y. Chebaro, A. Dejaegere, M. Cecchini, J. P. Changeux, P. G. Bolhuis, J. Vreede, P. Faccioli, S. Orioli, R. Ravasio, L. Yan, C. Brito, M. Wyart, P. Gkeka, I. Rivalta, G. Palermo, J. A. McCammon, J. Panecka-Hofman, R. C. Wade, A. Di Pizio, M. Y. Niv, R. Nussinov, C. J. Tsai, H. Jang, D. Padhorny, D. Kozakov, and T. McLeish, “Allostery in Its Many Disguises: From Theory to Applications,” *Structure*, vol. 27, no. 4, pp. 566–578, Apr. 2019.

[39] S. Rabbani, B. Fiege, D. Eris, M. Silbermann, R. P. Jakob, G. Navarra, T. Maier, and B. Ernst, "Conformational switch of the bacterial adhesin FimH in the absence of the regulatory domain: Engineering a minimalistic allosteric system," *Journal of Biological Chemistry*, vol. 293, no. 5, pp. 1835–1849, Feb. 2018.

[40] J. M. Pröpster, F. Yang, S. Rabbani, B. Ernst, F. H.-T. Allain, and M. Schubert, "Structural basis for sulfation-dependent self-glycan recognition by the human immune-inhibitory receptor Siglec-8," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 113, no. 29, E4170–E4179, Jul. 2016.

[41] T. Kiwamoto, N. Kawasaki, J. C. Paulson, and B. S. Bochner, "Siglec-8 as a drugable target to treat eosinophil and mast cell-associated conditions," *Pharmacology and Therapeutics*, vol. 135, no. 3, pp. 327–336, 2012.

[42] D. J. Carroll, J. A. O'Sullivan, D. B. Nix, Y. Cao, M. Tiemeyer, and B. S. Bochner, "Sialic acid–binding immunoglobulin-like lectin 8 (Siglec-8) is an activating receptor mediating β2-integrin–dependent function in human eosinophils," *Journal of Allergy and Clinical Immunology*, vol. 141, no. 6, pp. 2196–2207, Jun. 2018.

[43] S. A. Hudson, N. V. Bovin, R. L. Schnaar, P. R. Crocker, and B. S. Bochner, "Eosinophil-selective binding and proapoptotic effect in vitro of a synthetic siglec-8 ligand, polymeric 6'-sulfated sialyl lewis X," *Journal of Pharmacology and Experimental Therapeutics*, vol. 330, no. 2, pp. 608–612, 2009.

[44] P. V'kovski, A. Kratzel, S. Steiner, H. Stalder, and V. Thiel, *Coronavirus biology and replication: implications for SARS-CoV-2*, Mar. 2021.

[45] R. Henderson, R. J. Edwards, K. Mansouri, K. Janowska, V. Stalls, S. M. Gobeil, M. Kopp, D. Li, R. Parks, A. L. Hsu, M. J. Borgnia, B. F. Haynes, and P. Acharya, "Controlling the SARS-CoV-2 spike glycoprotein conformation," *Nature Structural and Molecular Biology*, vol. 27, no. 10, pp. 925–933, Oct. 2020.

[46] D. Wrapp, N. Wang, K. S. Corbett, J. A. Goldsmith, C.-L. Hsieh, O. Abiona, B. S. Graham, and J. S. McLellan, "Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation," *Science*, vol. 367, no. 6483, pp. 1260–1263, Mar. 2020.

[47] C. O. Barnes, C. A. Jette, M. E. Abernathy, K. M. A. Dam, S. R. Esswein, H. B. Gristick, A. G. Malyutin, N. G. Sharaf, K. E. Huey-Tubman, Y. E. Lee, D. F. Robbiani, M. C. Nussenzweig, A. P. West, and P. J. Bjorkman, "SARS-CoV-2 neutralizing antibody structures inform therapeutic strategies," *Nature*, vol. 588, no. 7839, pp. 682–687, Dec. 2020.

[48] T. Maximova, R. Moffatt, B. Ma, R. Nussinov, and A. Shehu, “Principles and Overview of Sampling Methods for Modeling Macromolecular Structure and Dynamics,” *PLOS Computational Biology*, vol. 12, no. 4, e1004619, Apr. 2016.

[49] S. Jo, T. Kim, V. G. Iyer, and W. Im, “CHARMM-GUI: A web-based graphical user interface for CHARMM,” *Journal of Computational Chemistry*, vol. 29, no. 11, pp. 1859–1865, Aug. 2008.

[50] W. Humphrey, A. Dalke, and K. Schulten, “VMD: Visual molecular dynamics,” *Journal of Molecular Graphics*, vol. 14, no. 1, pp. 33–38, Feb. 1996.

[51] J. C. Phillips, R. Braun, W. Wang, J. Gumbart, E. Tajkhorshid, E. Villa, C. Chipot, R. D. Skeel, L. Kalé, and K. Schulten, “Scalable molecular dynamics with NAMD,” *Journal of Computational Chemistry*, vol. 26, no. 16, pp. 1781–1802, Dec. 2005.

[52] M. J. Abraham, T. Murtola, R. Schulz, S. Páll, J. C. Smith, B. Hess, and E. Lindah, “GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers,” *SoftwareX*, vol. 1-2, pp. 19–25, Sep. 2015.

[53] P. Eastman, J. Swails, J. D. Chodera, R. T. McGibbon, Y. Zhao, K. A. Beauchamp, L. P. Wang, A. C. Simmonett, M. P. Harrigan, C. D. Stern, R. P. Wiewiora, B. R. Brooks, and V. S. Pande, “OpenMM 7: Rapid development of high performance algorithms for molecular dynamics,” *PLoS Computational Biology*, vol. 13, no. 7, e1005659, Jul. 2017.

[54] D. A. Case, T. E. Cheatham, T. Darden, H. Gohlke, R. Luo, K. M. Merz, A. Onufriev, C. Simmerling, B. Wang, and R. J. Woods, “The Amber biomolecular simulation programs,” *Journal of Computational Chemistry*, vol. 26, no. 16, pp. 1668–1688, Dec. 2005.

[55] J. Huang, S. Rauscher, G. Nawrocki, T. Ran, M. Feig, B. L. de Groot, H. Grubmüller, and A. D. MacKerell, “CHARMM36m: an improved force field for folded and intrinsically disordered proteins,” *Nature Methods*, vol. 14, no. 1, pp. 71–73, 2017.

[56] E. J. Haddadian, H. Zhang, K. F. Freed, and J. F. Douglas, “Comparative Study of the Collective Dynamics of Proteins and Inorganic Nanoparticles,” *Scientific Reports*, vol. 7, no. 1, pp. 1–18, Feb. 2017.

[57] S. K. Mishra and R. L. Jernigan, “Protein dynamic communities from elastic network models align closely to the communities defined by molecular dynamics,” *PLoS ONE*, vol. 13, no. 6, Y. Zhang, Ed., e0199225, Jun. 2018.

[58] F. Sittel, A. Jain, and G. Stock, “Principal component analysis of molecular dynamics: On the use of Cartesian vs. internal coordinates,” *Journal of Chemical Physics*, vol. 141, no. 1, p. 014 111, Jul. 2014.

[59] A. Altis, M. Otten, P. H. Nguyen, R. Hegger, and G. Stock, “Construction of the free energy landscape of biomolecules via dihedral angle principal component analysis,” *Journal of Chemical Physics*, vol. 128, no. 24, p. 245 102, Jun. 2008.

[60] T. P. Peixoto, “Network Reconstruction and Community Detection from Dynamics,” *Physical Review Letters*, vol. 123, no. 12, Sep. 2019.

[61] M. Ernst, F. Sittel, and G. Stock, “Contact- and distance-based principal component analysis of protein dynamics,” *Journal of Chemical Physics*, vol. 143, no. 24, p. 244 114, Dec. 2015.

[62] X. Q. Yao, M. Momin, and D. Hamelberg, “Establishing a Framework of Using Residue-Residue Interactions in Protein Difference Network Analysis,” *Journal of Chemical Information and Modeling*, vol. 59, no. 7, pp. 3222–3228, Jul. 2019.

[63] E. Wang, H. Sun, J. Wang, Z. Wang, H. Liu, J. Z. Zhang, and T. Hou, “End-Point Binding Free Energy Calculation with MM/PBSA and MM/GBSA: Strategies and Applications in Drug Design,” *Chemical Reviews*, vol. 119, no. 16, pp. 9478–9508, Aug. 2019.

[64] O. Serçinoğlu and P. Ozbek, “GRINN: A tool for calculation of residue interaction energies and protein energy network analysis of molecular dynamics simulations,” *Nucleic Acids Research*, vol. 46, no. W1, W554–W562, Jul. 2018.

[65] S. Wang, E. D. Herzog, I. Z. Kiss, W. J. Schwartz, G. Bloch, M. Sebek, D. Granados-Fuentes, L. Wang, and J. S. Li, “Inferring dynamic topology for decoding spatiotemporal structures in complex heterogeneous networks,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 115, no. 37, pp. 9300–9305, Sep. 2018.

[66] R. Mendez and U. Bastolla, “Torsional Network Model: Normal Modes in Torsion Angle Space Better Correlate with Conformation Changes in Proteins,” *Physical Review Letters*, vol. 104, no. 22, p. 228 103, Jun. 2010.

[67] C. Roberts and G. Casella, *Monte Carlo statistical methods*, 2nd ed. Berlin: Springer texts in statistics, 2005, ISBN: 978-1-4757-4145-2.

[68] B. J. Kim, H. Hong, P. Holme, G. S. Jeon, P. Minnhagen, and M. Y. Choi, “XY model in small-world networks,” *Physical Review E - Statistical Physics, Plasmas, Fluids, and Related Interdisciplinary Topics*, vol. 64, no. 5, p. 5, Oct. 2001.

[69] F. Morcos, A. Pagnani, B. Lunt, A. Bertolino, D. S. Marks, C. Sander, R. Zecchina, J. N. Onuchic, T. Hwa, and M. Weigt, “Direct-coupling analysis of residue coevolution captures native contacts across many protein families,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 108, no. 49, E1293–E1301, Dec. 2011.

[70] A. Banerjee, J. Pathak, R. Roy, J. G. Restrepo, and E. Ott, “Using machine learning to assess short term causal dependence and infer network links,” *Chaos*, vol. 29, no. 12, p. 121 104, Dec. 2019.

[71] W. P. Grant and S. E. Ahnert, “Modular decomposition of protein structure using community detection,” *Journal of Complex Networks*, vol. 7, no. 1, E. Estrada, Ed., pp. 101–113, Feb. 2019.

[72] T. P. Peixoto, “Bayesian stochastic blockmodeling,” in *Advances in Network Clustering and Blockmodeling*, John Wiley & Sons, Ltd, Dec. 2019, pp. 289–332, ISBN: 9781119483298.

[73] T. P. Peixoto, “Hierarchical block structures and high-resolution model selection in large networks,” *Physical Review X*, vol. 4, no. 1, p. 011 047, Mar. 2014.

[74] T. P. Peixoto, “Parsimonious Module Inference in Large Networks,” *Physical Review Letters*, vol. 110, no. 14, p. 148 701, Apr. 2013.

[75] K. C. Dansuk and S. Keten, “A Simple Mechanical Model for Synthetic Catch Bonds,” *Matter*, vol. 1, no. 4, pp. 911–925, 2019.

[76] T. T. Waldron and T. A. Springer, “Transmission of allostery through the lectin domain in selectin-mediated cell adhesion,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 106, no. 1, pp. 85–90, Jan. 2009.

[77] N. Ishiyama, R. Sarpal, M. N. Wood, S. K. Barrick, T. Nishikawa, H. Hayashi, A. B. Kobb, A. S. Flozak, A. Yemelyanov, R. Fernandez-Gonzalez, S. Yonemura, D. E. Leckband, C. J. Gottardi, U. Tepass, and M. Ikura, “Force-dependent allostery of the α-catenin actin-binding domain controls adherens junction dynamics and functions,” *Nature Communications*, vol. 9, no. 1, Dec. 2018.

[78] D. K. Das, Y. Feng, R. J. Mallis, X. Li, D. B. Keskin, R. E. Hussey, S. K. Brady, J. H. Wang, G. Wagner, E. L. Reinherz, and M. J. Lang, "Force-dependent transition in the T-cell receptor β-subunit allosterically regulates peptide discrimination and pMHC bond lifetime," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 112, no. 5, pp. 1517–1522, Feb. 2015.

[79] V. B. Rodriguez, B. A. Kidd, G. Interlandi, V. Tchesnokova, E. V. Sokurenko, and W. E. Thomas, "Allosteric coupling in the bacterial adhesive protein FimH," *Journal of Biological Chemistry*, vol. 288, no. 33, pp. 24 128–24 139, Aug. 2013.

[80] N. V. Dokholyan, "Controlling Allosteric Networks in Proteins," *Chemical Reviews*, vol. 116, no. 11, pp. 6463–6487, Jun. 2016.

[81] P. Aprikian, G. Interlandi, B. A. Kidd, I. Le Trong, V. Tchesnokova, O. Yakovenko, M. J. Whitfield, E. Bullitt, R. E. Stenkamp, W. E. Thomas, and E. V. Sokurenko, "The bacterial fimbrial tip acts as a mechanical force sensor," *PLoS Biology*, vol. 9, no. 5, May 2011.

[82] D. I. Kisiela, H. Avagyan, D. Friend, A. Jalan, S. Gupta, G. Interlandi, Y. Liu, V. Tchesnokova, V. B. Rodriguez, J. P. Sumida, R. K. Strong, X. R. Wu, W. E. Thomas, and E. V. Sokurenko, "Inhibition and Reversal of Microbial Attachment by an Antibody with Parasteric Activity against the FimH Adhesin of Uropathogenic E. coli," *PLoS Pathogens*, vol. 11, no. 5, May 2015.

[83] C. N. Spaulding and S. J. Hultgren, *Adhesive Pili in UTI pathogenesis and drug development*, Mar. 2016.

[84] V. Kalas, J. S. Pinkner, T. J. Hannan, M. E. Hibbing, K. W. Dodson, A. S. Holehouse, H. Zhang, N. H. Tolia, M. L. Gross, R. V. Pappu, J. Janetka, and S. J. Hultgren, "Evolutionary fine-tuning of conformational ensembles in FimH during host-pathogen interactions," *Science Advances*, vol. 3, no. 2, e1601944, Feb. 2017.

[85] R. Hevey, "Strategies for the development of glycomimetic drug candidates," *Pharmaceuticals*, vol. 12, no. 2, p. 55, Apr. 2019.

[86] P. Aprikian, V. Tchesnokova, B. Kidd, O. Yakovenko, V. Yarov-Yarovoy, E. Trinchina, V. Vogel, W. Thomas, and E. Sokurenko, "Interdomain interaction in the FimH adhesin of Escherichia coli regulates the affinity to mannose," *Journal of Biological Chemistry*, vol. 282, no. 32, pp. 23 437–23 446, Aug. 2007.

[87] O. Yakovenko, V. Tchesnokova, E. V. Sokurenko, W. E. Thomas, and T. A. Springer, "Inactive conformation enhances binding function in physiological conditions," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 112, no. 32, pp. 9884–9889, Aug. 2015.

[88] N. Eswar, B. John, N. Mirkovic, A. Fiser, V. A. Ilyin, U. Pieper, A. C. Stuart, M. A. Marti-Renom, M. S. Madhusudhan, B. Yerkovich, and A. Sali, "Tools for comparative protein structure modeling and analysis," *Nucleic Acids Research*, vol. 31, no. 13, pp. 3375–3380, Jul. 2003.

[89] A. Wellens, M. Lahmann, M. Touaibia, J. Vaucher, S. Oscarson, R. Roy, H. Remaut, and J. Bouckaert, "The tyrosine gate as a potential entropic lever in the receptor-binding site of the bacterial adhesin FimH," *Biochemistry*, vol. 51, no. 24, pp. 4790–4799, Jun. 2012.

[90] R. B. Best, X. Zhu, J. Shim, P. E. Lopes, J. Mittal, M. Feig, A. D. MacKerell, B. RB, Z. X, S. J, L. PE, M. J, F. M, and M. AD, "Optimization of the additive CHARMM all-atom protein force field targeting improved sampling of the backbone ϕ, ψ and side-chain $\chi(1)$ and $\chi(2)$ dihedral angles," *Journal of Chemical Theory and Computation*, vol. 8, no. 9, pp. 3257–3273, Sep. 2012.

[91] Y. Mu, P. H. Nguyen, and G. Stock, "Energy landscape of a small peptide revealed by dihedral angle principal component analysis," *Proteins: Structure, Function and Genetics*, vol. 58, no. 1, pp. 45–52, Jan. 2005.

[92] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, İ. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, A. Vijaykumar, A. P. Bardelli, A. Rothberg, A. Hilboll, A. Kloeckner, A. Scopatz, A. Lee, A. Rokem, C. N. Woods, C. Fulton, C. Masson, C. Häggström, C. Fitzgerald, D. A. Nicholson, D. R. Hagen, D. V. Pasechnik, E. Olivetti, E. Martin, E. Wieser, F. Silva, F. Lenders, F. Wilhelm, G. Young, G. A. Price, G. L. Ingold, G. E. Allen, G. R. Lee, H. Audren, I. Probst, J. P. Dietrich, J. Silterra, J. T. Webber, J. Slavič, J. Nothman, J. Buchner, J. Kulick, J. L. Schönberger, J. V. de Miranda Cardoso, J. Reimer, J. Harrington, J. L. C. Rodríguez, J. Nunez-Iglesias, J. Kuczynski, K. Tritz, M. Thoma, M. Newville, M. Kümmerer, M. Bolingbroke, M. Tartre, M. Pak, N. J. Smith, N. Nowaczyk, N. Shebanov, O. Pavlyk, P. A. Brodtkorb, P. Lee, R. T. McGibbon, R. Feldbauer, S. Lewis, S. Tygier, S. Sievert, S. Vigna, S. Peterson, S. More, T. Pudlik, T. Oshima, T. J. Pingel, T. P. Robitaille, T. Spura, T. R. Jones, T. Cera, T. Leslie, T. Zito, T. Krauss, U. Upadhyay, Y. O. Halchenko, and Y.

Vázquez-Baeza, “SciPy 1.0: fundamental algorithms for scientific computing in Python,” *Nature Methods*, vol. 17, no. 3, pp. 261–272, Mar. 2020.

[93] S. Seabold and J. Perktold, “statsmodels: Econometric and statistical modeling with python,” in *9th Python in Science Conference*, 2010.

[94] Y. Benjamini, A. M. Krieger, and D. Yekutieli, “Adaptive linear step-up procedures that control the false discovery rate,” *Biometrika*, vol. 93, no. 3, pp. 491–507, Sep. 2006.

[95] N. I. Fisher and A. J. Lee, “A correlation coefficient for circular data,” *Biometrika*, vol. 70, no. 2, pp. 327–332, Aug. 1983.

[96] M. Levitt, “Growth of novel protein structural data,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 104, no. 9, pp. 3183–3188, Feb. 2007.

[97] T. C. Terwilliger, D. Stuart, and S. Yokoyama, “Lessons from structural genomics,” *Annual Review of Biophysics*, vol. 38, no. 1, pp. 371–383, Jun. 2009.

[98] E. Callaway, *’It will change everything’: DeepMind’s AI makes gigantic leap in solving protein structures*, Dec. 2020.

[99] S. J. Rysavy, D. A. Beck, and V. Daggett, “Dynameomics: Data-driven methods and models for utilizing large-scale protein structure repositories for improving fragment-based loop prediction,” *Protein Science*, vol. 23, no. 11, pp. 1584–1595, Nov. 2014.

[100] B. R. Amor, M. T. Schaub, S. N. Yaliraki, and M. Barahona, “Prediction of allosteric sites and mediating interactions through bond-to-bond propensities,” *Nature Communications*, vol. 7, no. 1, pp. 1–13, Aug. 2016.

[101] A. Di Nola, H. J. Berendsen, and O. Edholm, “Free Energy Determination of Polypeptide Conformations Generated by Molecular Dynamics,” *Macromolecules*, vol. 17, no. 10, pp. 2044–2050, 1984.

[102] S. Singh and G. R. Bowman, “Quantifying Allosteric Communication via Both Concerted Structural Changes and Conformational Disorder with CARDS,” *Journal of Chemical Theory and Computation*, vol. 13, no. 4, pp. 1509–1517, Apr. 2017.

[103] S. Bowerman and J. Wereszczynski, “Detecting Allosteric Networks Using Molecular Dynamics Simulation,” in *Methods in Enzymology*, vol. 578, Academic Press Inc., Jan. 2016, pp. 429–447.

[104] A. Hacisuleyman and B. Erman, “Entropy Transfer between Residue Pairs and Allostery in Proteins: Quantifying Allosteric Communication in Ubiquitin,” *PLoS Computational Biology*, vol. 13, no. 1, J. M. Briggs, Ed., e1005319, Jan. 2017.

[105] P. M. Gasper, B. Fuglestad, E. A. Komives, P. R. Markwick, and J. A. McCammon, “Allosteric networks in thrombin distinguish procoagulant vs. anticoagulant activities,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 109, no. 52, pp. 21 216–21 222, Dec. 2012.

[106] O. F. Lange and H. Grubmüller, “Generalized correlation for biomolecular dynamics,” *Proteins: Structure, Function and Genetics*, vol. 62, no. 4, pp. 1053–1061, Mar. 2006.

[107] J. Liu, L. A. N. Amaral, and S. Keten, “Conformational stability of the bacterial adhesin, ¡scp¿FimH¡/scp¿ , with an inactivating mutation,” *Proteins: Structure, Function, and Bioinformatics*, vol. 89, no. 3, pp. 276–288, Mar. 2021.

[108] H. Woo, S. J. Park, Y. K. Choi, T. Park, M. Tanveer, Y. Cao, N. R. Kern, J. Lee, M. S. Yeom, T. I. Croll, C. Seok, and W. Im, “Developing a fully glycosylated full-length SARS-COV-2 spike protein model in a viral membrane,” *Journal of Physical Chemistry B*, vol. 124, no. 33, pp. 7128–7137, Aug. 2020.

[109] M. Cruz, T. Frederick, S. Singh, N. Vithani, M. Zimmerman, J. Porter, K. Moeder, G. Amarasinghe, and G. Bowman, “Discovery of a cryptic allosteric site in Ebola’s ’undruggable’ VP35 protein using simulations and experiments,” *bioRxiv*, p. 2020.02.09.940510, Feb. 2020.

[110] C. R. Knoverek, U. L. Mallimadugula, S. Singh, E. Rennella, T. E. Frederick, T. Yuwen, S. Raavicharla, L. E. Kay, and G. R. Bowman, “Opening of a Cryptic Pocket in β-lactamase Increases Penicillinase Activity,” *bioRxiv*, p. 2021.04.14.439842, Apr. 2021.

[111] B. J. Grant, L. Skjærven, and X.-Q. Yao, “The Bio3D packages for structural bioinformatics,” *Protein Science*, vol. 30, no. 1, pp. 20–30, Jan. 2021.

[112] D. A. Keedy, I. Georgiev, E. B. Triplett, B. R. Donald, D. C. Richardson, and J. S. Richardson, “The Role of Local Backrub Motions in Evolved and Designed Mutations,” *PLoS Computational Biology*, vol. 8, no. 8, Y. Ofran, Ed., e1002629, Aug. 2012.

[113] S. Freeman, H. C. Birrell, K. D’Alessio, C. Erickson-Miller, K. Kikly, and P. Camilleri, “A comparative study of the asparagine-linked oligosaccharides on siglec-5, siglec-7 and

siglec-8, expressed in a CHO cell line, and their contribution to ligand recognition," *European Journal of Biochemistry*, vol. 268, no. 5, pp. 1228–1237, Mar. 2001.

[114] Z. Ke, J. Oton, K. Qu, M. Cortese, V. Zila, L. McKeane, T. Nakane, J. Zivanov, C. J. Neufeldt, B. Cerikan, J. M. Lu, J. Peukes, X. Xiong, H.-G. Kräusslich, S. H. W. Scheres, R. Bartenschlager, and J. A. G. Briggs, "Structures and distributions of SARS-CoV-2 spike proteins on intact virions," *Nature*, vol. 588, no. 7838, pp. 498–502, Dec. 2020.

[115] L. Casalino, Z. Gaieb, J. A. Goldsmith, C. K. Hjorth, A. C. Dommer, A. M. Harbison, C. A. Fogarty, E. P. Barros, B. C. Taylor, J. S. McLellan, E. Fadda, and R. E. Amaro, "Beyond Shielding: The Roles of Glycans in the SARS-CoV-2 Spike Protein," *ACS Central Science*, vol. 6, no. 10, pp. 1722–1734, Oct. 2020.

[116] D. E. S. Research, *Molecular Dynamics Simulations Related to SARS-CoV-2*.

[117] D. S. Marks, L. J. Colwell, R. Sheridan, T. A. Hopf, A. Pagnani, R. Zecchina, and C. Sander, "Protein 3D Structure Computed from Evolutionary Sequence Variation," *PLoS ONE*, vol. 6, no. 12, A. Sali, Ed., e28766, Dec. 2011.

[118] J. Wang, A. Jain, L. R. McDonald, C. Gambogi, A. L. Lee, and N. V. Dokholyan, "Mapping allosteric communications within individual proteins," *Nature Communications*, vol. 11, no. 1, pp. 1–13, Jul. 2020.

[119] M. C. Childers and V. Daggett, "Validating Molecular Dynamics Simulations against Experimental Observables in Light of Underlying Conformational Ensembles," *Journal of Physical Chemistry B*, vol. 122, no. 26, pp. 6673–6689, 2018.

[120] S. L. Chen, C. S. Hung, J. S. Pinkner, J. N. Walker, C. K. Cusumano, Z. Li, J. Bouckaert, J. I. Gordon, and S. J. Hultgren, "Positive selection identifies an in vivo role for FimH during urinary tract infection in addition to mannose binding," *Proceedings of the National Academy of Sciences*, vol. 106, no. 52, pp. 22 439–22 444, Dec. 2009.

[121] P. Magala, R. E. Klevit, W. E. Thomas, E. V. Sokurenko, and R. E. Stenkamp, "RMSD analysis of structures of the bacterial protein FimH identifies five conformations of its lectin domain," *Proteins: Structure, Function, and Bioinformatics*, vol. 88, no. 4, pp. 593–603, Apr. 2020.

[122] D. I. Kisiela, P. Magala, G. Interlandi, L. A. Carlucci, A. Ramos, V. Tchesnokova, B. Basanta, V. Yarov-Yarovoy, H. Avagyan, A. Hovhannisyan, W. E. Thomas, R. E. Stenkamp, R. E. Klevit, and E. V. Sokurenko, "Toggle switch residues control allosteric transitions

in bacterial adhesins by participating in a concerted repacking of the protein core," *PLOS Pathogens*, vol. 17, no. 4, O. Francetic, Ed., e1009440, Apr. 2021.

[123] R. Nussinov and C. J. Tsai, *Allostery in disease and in drug discovery*, Apr. 2013.

NETWORK ANALYSIS OF PROTEIN DYNAMICS

Approved by:

Dr. Sinan Keten, Advisor
Civil Engineering and Mechanical Engineering Depts.
Northwestern University

Dr. Luis A. N. Amaral
Chemical and Biological Engineering Dept., Northwestern Institute of Complex Systems (NICO)
Northwestern University

Dr. Michelle Driscoll
Physics Dept.
Northwestern University

Date Approved: August 1, 2021

APPENDIX A
CONFORMATIONAL STABILITY OF THE BACTERIAL ADHESIN, FIMH, WITH AN INACTIVATING MUTATION

A.1 Choice of structures and notation

H2 refers to both the lectin and pilin domains, while HL refers to the lectin domain alone. H2 inactive refers to the “low-affinity structure” [35] that corresponds with the “tense” state in the Monod-Wyman-Changeux (MWC) model notation of Kalas *et al.* [84]. Structurally, the lectin and pilin domains are “associated” in H2 inactive using the notation of Sauer *et al.* [34]. In contrast, H2 active has high affinity for ligand, corresponds to the “relaxed” state in the MWC model [84], and has “separated” domains [34]. At this time, crystal structures for H2 active have relatively poor resolution due to the flexibility of the interdomain region (PDBID 1KLF, 2.79 Å, 2001 used Fig. SA.1 and SA.2; PDBID 4XOB, 3.003 Å, 2016; PDBID 5JR4, 2.596 Å, 2017). Moreover, MD simulations of H2 active require large and computationally expensive simulation boxes to prevent self-interactions during hinged motion [84]. As a result, H2 active was not studied here. Instead, we studied the dynamics of the high-affinity conformation of the lectin domain using HL active, which is constitutively active (PDBID 4AUU, 1.6 Å, 2012). Similarly, we focused on two versions of H2 inactive with higher resolution, and another version of H2 inactive with lower resolution (PDBID 3JWN, 2.69 Å, 2009) was also not studied here.

A.2 Putative allosteric pathway

Several landmarks have been identified in the putative allosteric pathway through a combination of experimental and computational approaches [79]. These landmarks are illustrated on the New-Cartoon representations of the protein backbone in Fig. 3.2B, SA.1, and SA.2 and are also used in the headers of plots describing dynamical or structural differences (e.g. Fig. 3.3A). The landmarks in the binding pocket are the pocket zipper (1-11) and the clamp (8-16), which are collectively referred to as the clamp segment [39], [79]. To these, we also add the parasteric segment (132-140), which interacts with the ligand in the active state but is exposed in the inactive state, so that antibody binding stabilizes the inactive conformation of the binding pocket [82]. Next, the landmarks in the interdomain region are the swing loop (25-34), the insertion loop (111-119), and the linker loop (152-158), which covalently connects the lectin and pilin domains [79]. Lastly, the Arg-60-Pro substitution of HL mutant is in the β-bulge (59-63), which is adjacent to the α-switch (64-71), and mutations in these regions stabilize the inactive or active conformations of the lectin domain [39], [79].

A.3 The impact of the Arg or Pro at position 60

We examined the effect of the Arg-60-Pro mutation alone on HL mutant by changing Pro to the wild-type Arg *in silico*. Analogously, we modified HL active by changing the Arg to Pro at position 60 *in silico* to mimic the mutation. We find that the differences in backbone structure and dynamics due to the Arg-60-Pro mutation are small and isolated for HL mutant and HL active. For the sidechain dynamics, the cyclic Pro at residue 60 in HL mutant has smaller fluctuations than the longer Arg, and this is the only statistically significant difference.

Table S A.1: **Residues with differences in sidechain dynamics.** For the comparisons of HL mutant vs H2 inactive, HL mutant vs HL active, and H2 inactive vs HL active, we report the residue index. In the parentheses, we report the difference in the median value of the standard deviation for υ if $\Delta\sigma > 0.2$ radians or about 11.5 degrees.

Comparison	Residues
HL mutant - H2 inactive	7 (0.43), 28 (0.80), 60 (-1.32)
HL mutant - HL active	5 (-0.94), 9 (-0.60), 22 (-0.81), 32 (-0.77), 33 (-0.44), 41 (-1.01), 56 (0.59), 59 (-1.03), 60 (-0.78), 62 (-1.10), 70 (0.72), 71 (0.73), 74 (-0.92), 107 (-0.61), 108 (-1.14), 109 (0.43), 112 (1.04), 120 (0.60), 126 (0.94), 131 (0.60), 133 (0.30), 142 (1.11), 147 (0.25), 152 (0.24), 154 (0.44)
H2 inactive - HL active	5 (-0.94), 9 (-0.79), 11 (-1.04), 13 (-0.57), 19 (0.25), 22 (-0.82), 23 (0.29), 28 (-0.84), 33 (-0.56), 34 (-0.30), 41 (-1.02), 56 (0.73), 58 (0.43), 59 (-0.94), 62 (-1.10), 67 (0.45), 70 (0.53), 71 (0.78), 72 (0.82), 74 (-0.82), 93 (0.40), 94 (-0.38), 96 (0.31), 99 (-0.29), 107 (-0.66), 108 (-1.12), 109 (0.51), 112 (0.49), 113 (-0.37), 120 (0.95), 126 (0.81), 130 (-0.56), 131 (0.64), 133 (0.34), 136 (-0.23), 142 (0.81), 143 (-0.35), 147 (0.43), 149 (0.57), 152 (0.23), 153 (0.47), 155 (-0.39)

A.4 Differences in sidechain dynamics

Protein residues with different sidechain dynamics are listed in Table SA.1.

A.5 Longer simulations

The fluctuations observed over the 20 ns simulations are only relevant to the fluctuations of single dihedrals, rather than the large conformational changes between the inactive and active states, which may require much longer simulations, on the order of a microsecond to observe [35], [84]. While we focus on local dihedral angle fluctuations, the ability of hinged motion to displace distal elements does allow us to observe opening and closing of the clamp segment at the binding pocket and the interdomain motions of H2 inactive.

We performed long simulations of up to 200ns for 3 replicates each of HL mutant, H2 inactive,

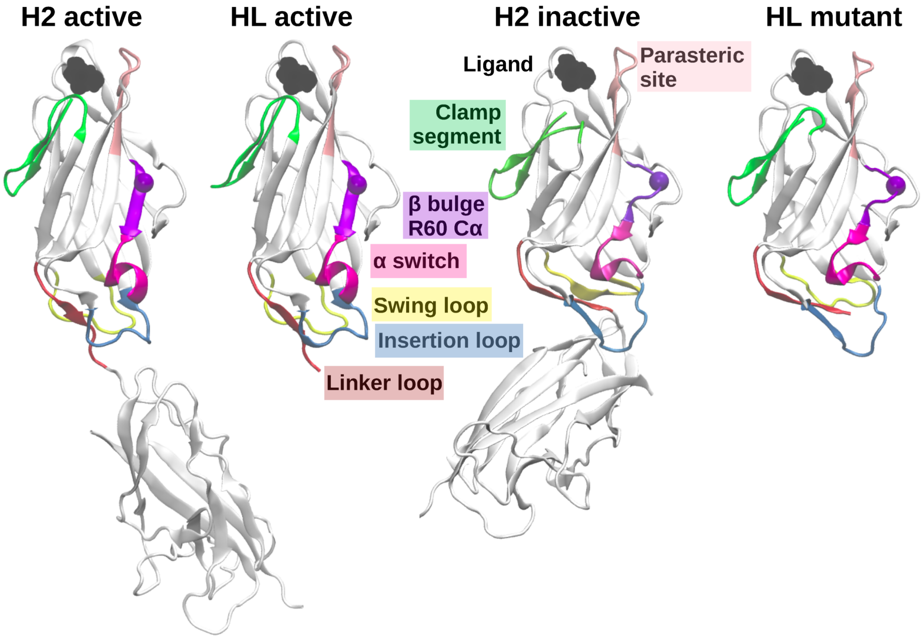

Figure S A.1: **Structural differences for active and inactive conformations (front side).** Backbone structures are shown as NewCartoon representations. H2 refers to both the lectin and pilin domains, while HL refers to the lectin domain alone. HL active (PDBID 4AUU) matches the structure of H2 active (PDBID 1KLF) in the lectin domain; HL mutant (PDBID 5MCA) matches H2 inactive (PDBID 4XOD). The mutation at residue 60 is indicated by a bead at Cα. The putative allosteric pathway sites and the parasteric site are shown on the cartoon structure.

and HL active. We observed that the Cα-RMSD remained within about 2 Angstroms (Fig. SA.5, SA.6, and SA.7). In order to gain greater insight into the variability of the protein dynamics, we also analyzed our long simulations by breaking them down into overlapping windows of 20 ns. We found that the average standard deviation across all angles also remained stable over most of the stimulation (Fig. SA.8), with occasional spikes that corresponded to localized changes at specific residues (Fig. SA.9). The protein regions with dynamics differences identified from the shorter simulations are qualitatively similar to those identified by comparing the 3 replicates of longer simulations (Fig. SA.10 and SA.11). While the longer simulations identified additional regions with dynamical differences, it is limited by fewer replicates.

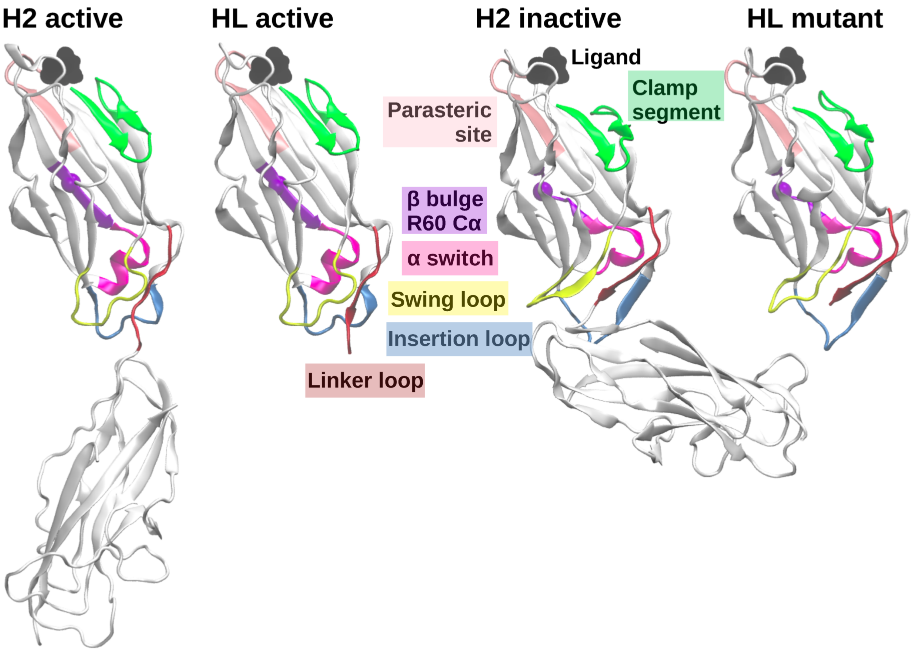

Figure S A.2: **Structural differences for active and inactive conformations (back side).** Structures in Fig. SA.1 rotated 180 degrees to show the view from the back.

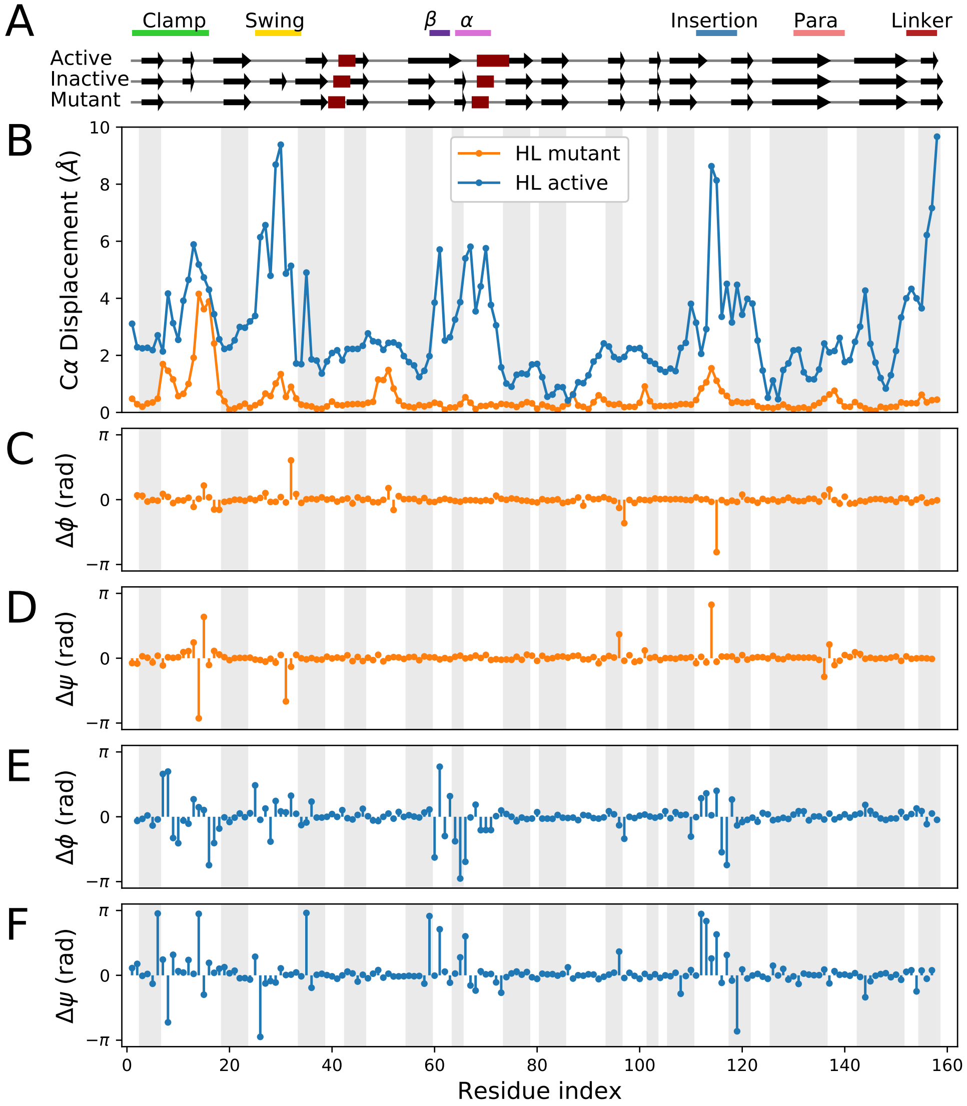

Figure S A.3: **Structural comparisons of HL mutant and HL active vs H2 inactive v1.** (A) Secondary structure of HL active (top), H2 inactive (middle), and HL mutant (bottom) with features. (B) Cα-displacement in Å describes the contribution of each residue to the RMSD. Angular displacements of backbone dihedrals ϕ and ψ for HL mutant show localized structural changes (C, D). HL mutant has fewer angular displacements than HL active (E, F). See Fig. SA.4 for difference represented on the protein structure.

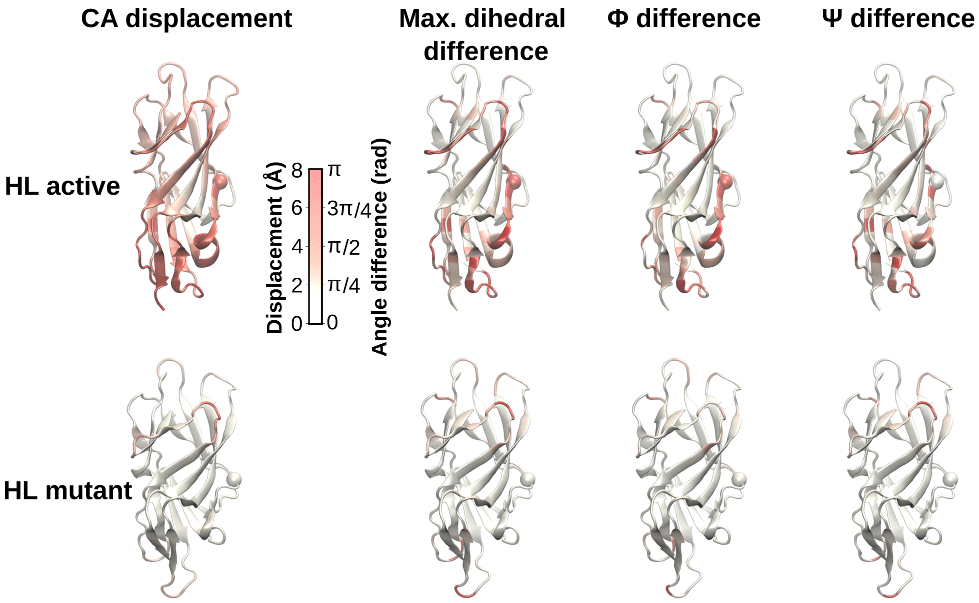

Figure S A.4: **Visualizing differences in crystal structure with RMSD and angle differences.** HL active and HL mutant are compared against H2 inactive using (left) $C\alpha$-RMSD and (right) backbone angle differences from Fig. SA.3. Color is scaled to the angle difference as $|\phi|$, $|\psi|$, and the maximum of $[|\phi|, |\psi|]$. Angles are shown on the corresponding residue of the NewCartoon representation.

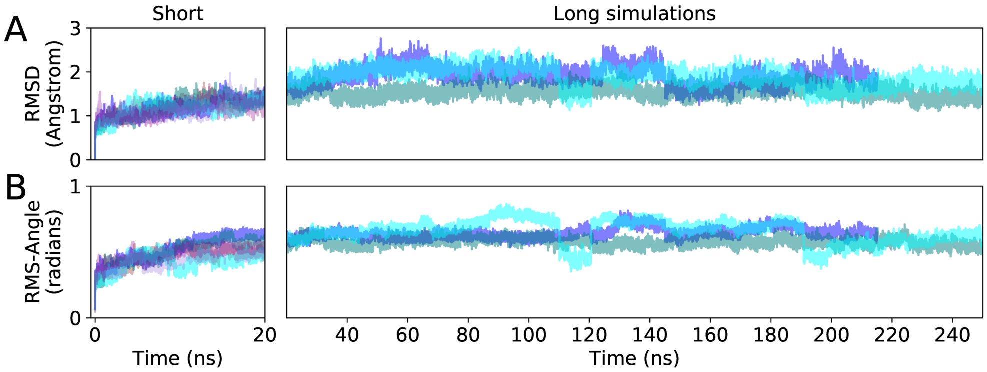

Figure S A.5: **Short vs long simulations for HL mutant** (A) Cα root-mean-squared deviations (RMSD) and (B) the root-mean-square of dihedral angle deviations for 3 replicates

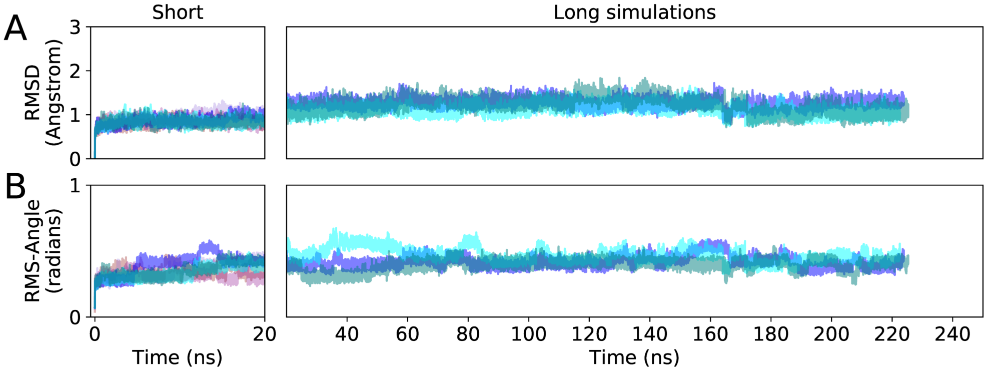

Figure S A.6: **Short vs long simulations for H2 inactive** (A) Cα root-mean-squared deviations (RMSD) and (B) the root-mean-square of dihedral angle deviations for 3 replicates

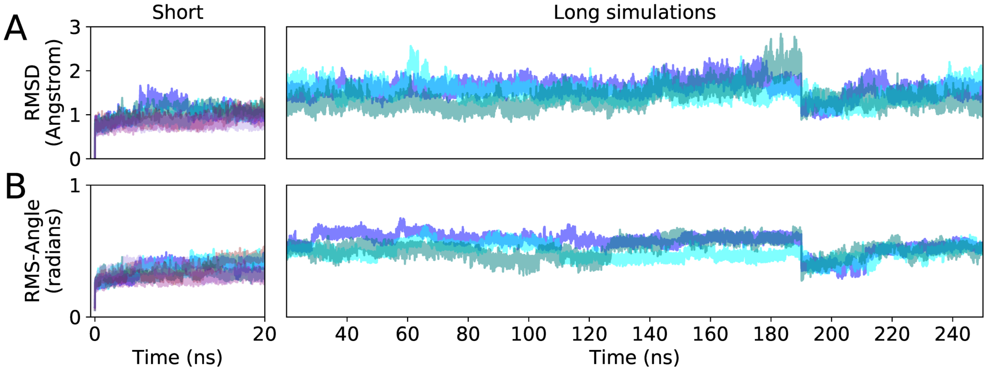

Figure S A.7: **Short vs long simulations for HL active** (A) Cα root-mean-squared deviations (RMSD) and (B) the root-mean-square of dihedral angle deviations for 3 replicates

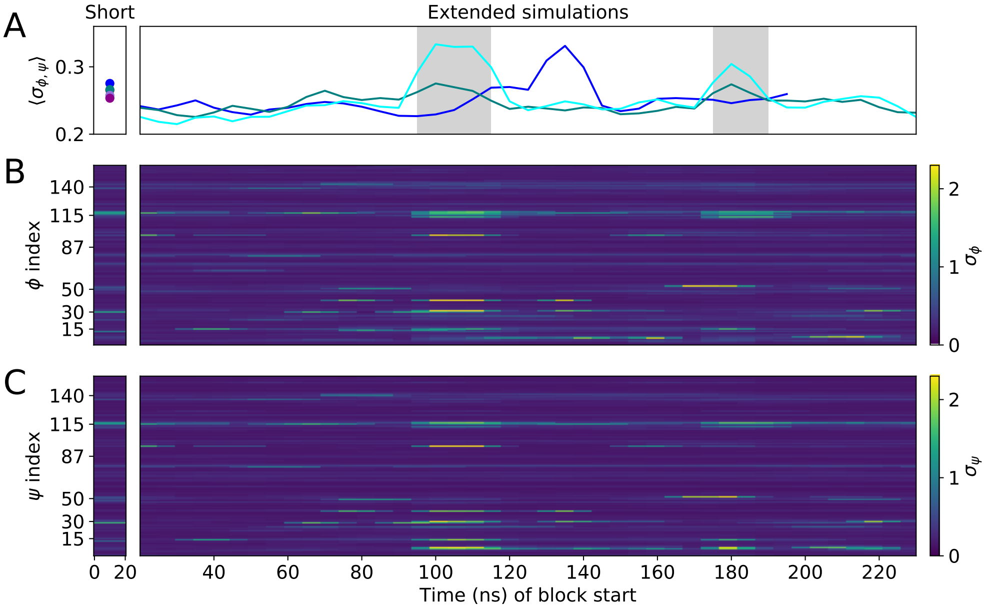

Figure S A.8: **Long simulations divided into 20 ns blocks show spikes in the standard deviation** (grey patches) (A) The standard deviation at each backbone dihedral angle was calculated for each block. The average standard deviation is shown at each block for three replicates. (B, C) The ϕ and ψ standard deviations are shown for the cyan replicate.

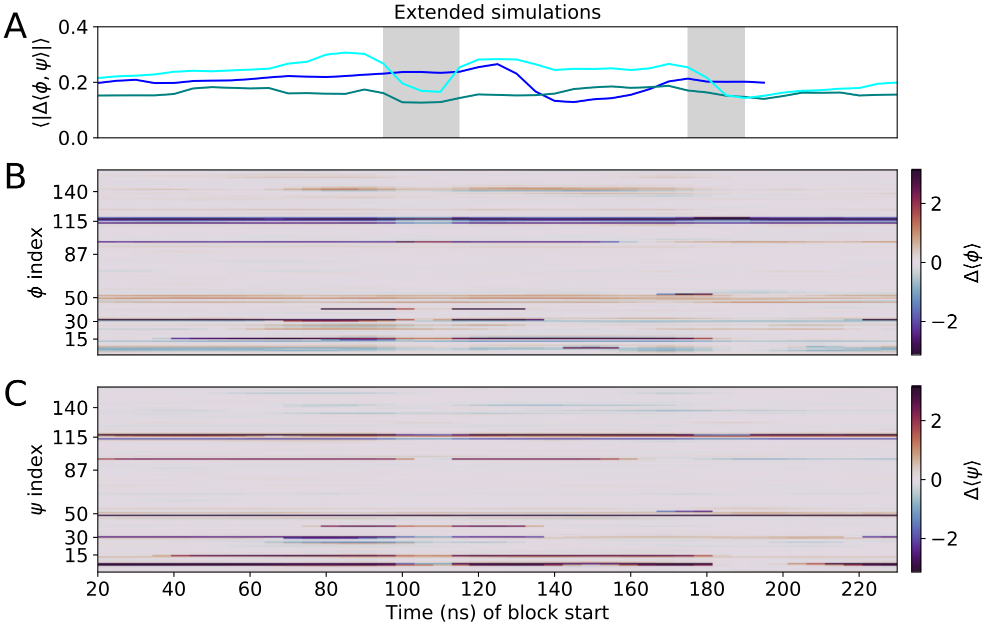

Figure S A.9: **Transient changes in average structure correspond to time periods with spikes in standard deviation** (grey patches from Fig. A.8) (A) Average structure deviation from the short 20 ns simulation. (B, C) Deviations in ϕ and ψ structure are shown for the cyan replicate.

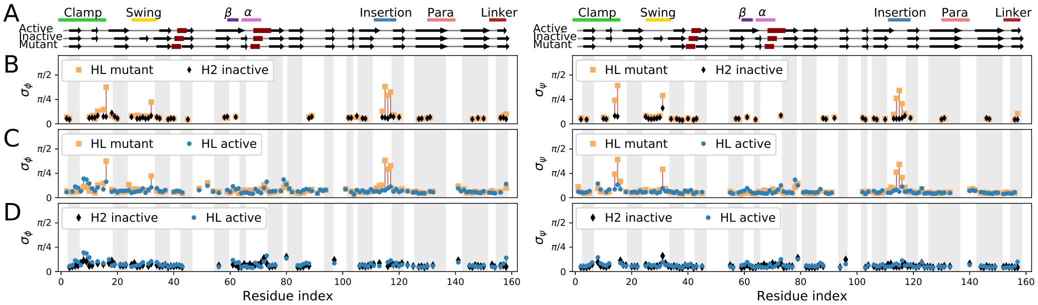

Figure S A.10: **Comparisons of backbone dynamics show HL mutant has larger fluctuations than H2 inactive and HL active.** Backbone dynamics were compared using the standard deviations of ϕ (left) and ψ (right). (A) Secondary structure of HL active (top), H2 inactive (middle), and HL mutant (bottom) with features. (B) HL mutant (orange squares) has larger backbone fluctuations in the clamp segment and the insertion and swing loops than H2 inactive (black diamonds). (C) HL mutant also has larger fluctuations than HL active (blue circles) in the swing and insertion loops. HL active has slightly larger fluctuations at the tip of the clamp segment than both HL mutant and H2 inactive. (D) HL active and H2 inactive have smaller differences in backbone dynamics throughout the lectin domain. The comparisons between HL mutant and H2 inactive are repeated from Fig. 3.3C to show the alignment. The ϕ fluctuations are repeated from Fig. 3.4 to compare with those for ψ.

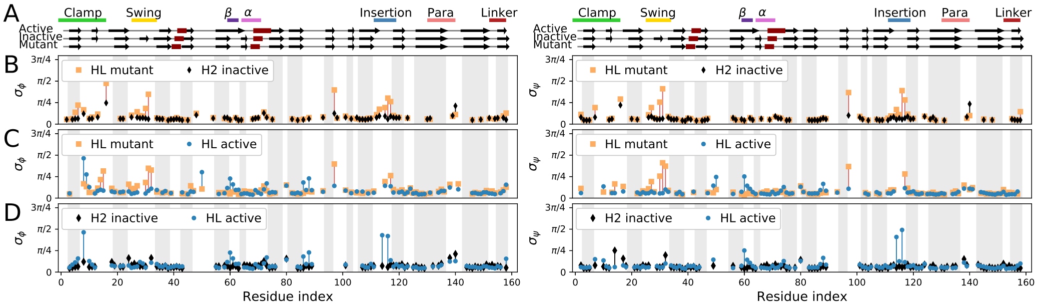

Figure S A.11: **For longer simulations, comparisons of backbone dynamics also show HL mutant has larger fluctuations than H2 inactive and HL active.** Backbone dynamics were compared using the standard deviations of ϕ (left) and ψ (right). (A) Secondary structure of HL active (top), H2 inactive (middle), and HL mutant (bottom) with features. (B) Similar to the shorter simulations, HL mutant (orange squares) has larger backbone fluctuations in the clamp segment and the insertion and swing loops than H2 inactive (black diamonds). The longer simulations show HL mutant also has larger fluctuations than both H2 inactive and HL active in an unnamed loop near the binding pocket (Ser-97) and slightly larger fluctuations in the parasteric site. (C) Similar to the shorter simulations, HL mutant has larger backbone fluctuations in the swing and insertion loops, while HL active has larger fluctuations at the tip of the clamp segment. Longer simulations show HL active has larger fluctuations in an unnamed region in the binding pocket (Glu-50). (D) The longer simulations show that HL active has larger fluctuations than H2 inactive at the tip of the clamp segment and the insertion loop. HL active also has slightly larger fluctuations than both H2 inactive and HL mutant in the β-bulge/α-switch region and its neighboring loop (Thr-86 and Ser-88).

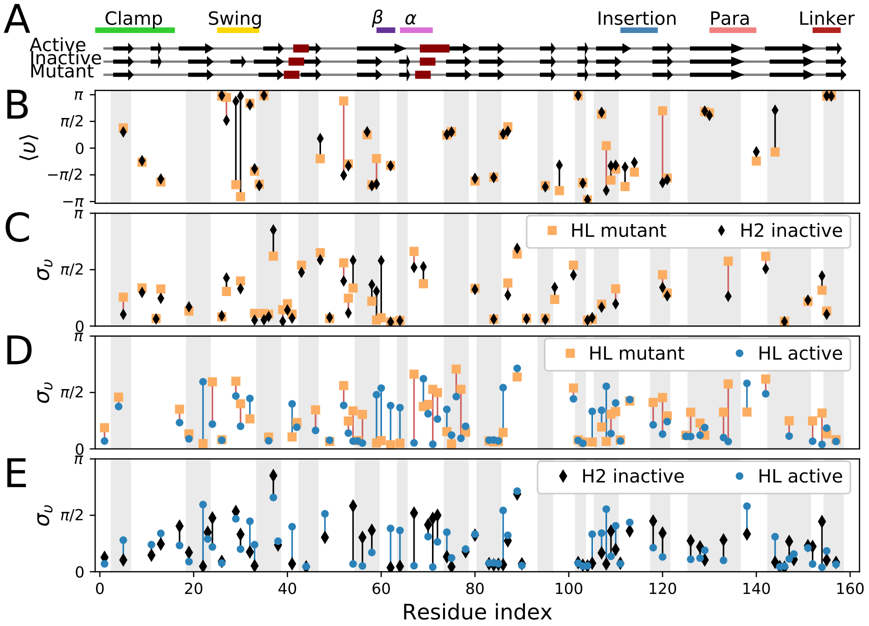

Figure S A.12: **For longer simulations, comparisons of pseudo-dihedral angles suggests differences in side chain dynamics across HL mutant, H2 inactive, and HL active.** (A) Secondary structure of HL active (top), H2 inactive (middle), and HL mutant (bottom) with features. Sidechain structure and dynamics were compared using rotamer angles υ as in Fig. 3.5. Comparing HL mutant (orange squares) and H2 inactive (black diamonds), (B) the average sidechain structure shows differences in the swing and insertion loops, but the differences in the clamp segment are small. (C) The dynamics show larger fluctuations for Arg-60 in H2 inactive compared to Pro-60 in HL mutant, as well as several smaller differences across the protein. (D, E) Comparisons of HL active (black circles) against HL mutant and H2 inactive show differences in sidechain dynamics beyond the allosteric pathway sites.

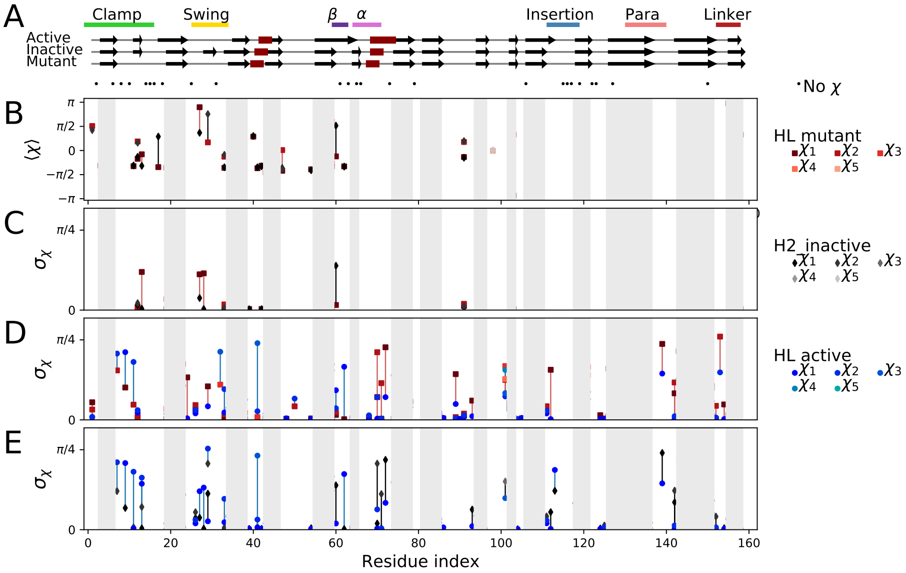

Figure S A.13: **Using rotamer angles to compare the sidechain dynamics of HL mutant, H2 inactive, and HL active.** (A) Secondary structure of HL active (top), H2 inactive (middle), and HL mutant (bottom) with features. (B) Sidechain structure and dynamics were compared using rotamer angles χ, which are defined for each amino acid except for Gly and Ala (black dots indicate no χ). At the β-bulge, the hinge of the clamp segment, the parasteric region, the swing loop, and several other sites, the average dihedral angle was different between HL mutant (warm-colored squares) and H2 inactive (grey-scale diamonds). (C) Compared to H2 inactive, HL mutant has larger fluctuations in the clamp segment and swing loop. The wild-type Arg-60 in H2 inactive has larger fluctuations for χ_2 than the mutated Pro-60 in HL mutant. (D, E) Comparisons of HL active (cool-colored circles) against HL mutant and H2 inactive show differences in sidechain dynamics across the lectin domain.

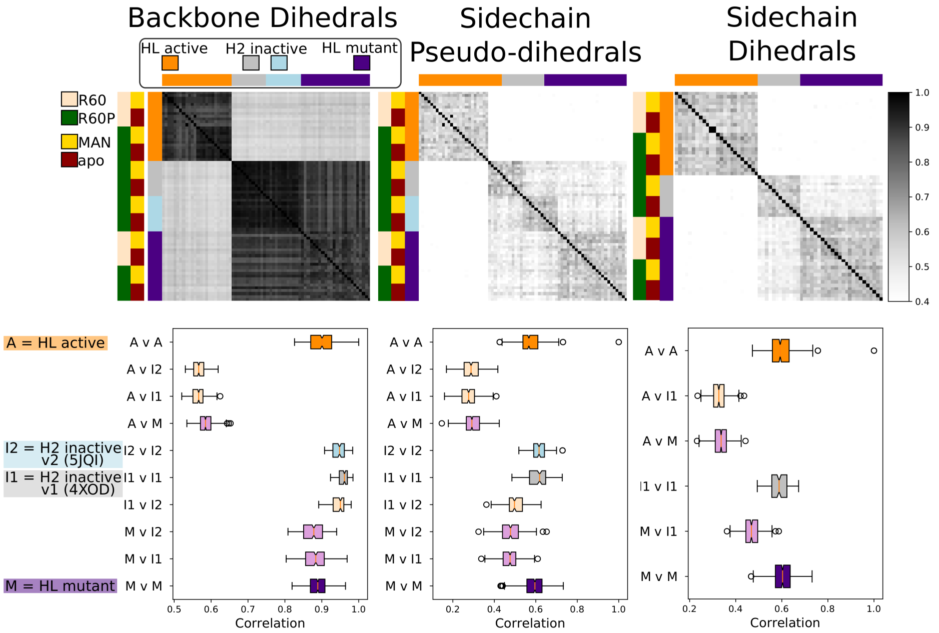

Figure S A.14: **Similarity matrices with backbone and sidechain structure.** Comparing HL active, H2 inactive, and HL mutant using backbone dihedral angles (left), sidechain pseudo-dihedral angles (middle), and sidechain dihedral angles χ (right). For χ, only one version of H2 inactive is used because the amino acid sequence sequence matches HL active and HL mutant. The similarity measure is a non-parametric (rank based) correlation, using Mardia's method modified by Fisher and Lee, to account for the circularity of the angles [95]. The system setup is described using a three-layer color code for the initial protein structure, mutation at position 60, and ligand. For the similarity matrices (top), the distributions of the correlations are described as box plots (bottom). The purpose is to compare trajectories from different initial protein structures. The first set of reference comparisons come from the same initial protein structures: orange for HL active, blue and grey for the two versions of H2 inactive, and purple for HL mutant. Another reference is the comparison between the two versions of H2 inactive. The last reference is the low similarity when HL active is compared against H2 inactive and HL mutant, for all measures. For backbone structure, the reference comparison for H2 inactive has higher similarity than the one for HL mutant, which has values close to the comparisons between HL mutant and H2 inactive. For sidechain pseudo-dihedral angles, all comparisons have lower similarity, HL mutant is different from H2 inactive, and the two versions of H2 inactive are also different from each other, due to the difference in amino acid sequence. In addition, for sidechain dihedral angles χ, the comparison between HL mutant and H2 inactive also has lower similarity than the reference comparisons.

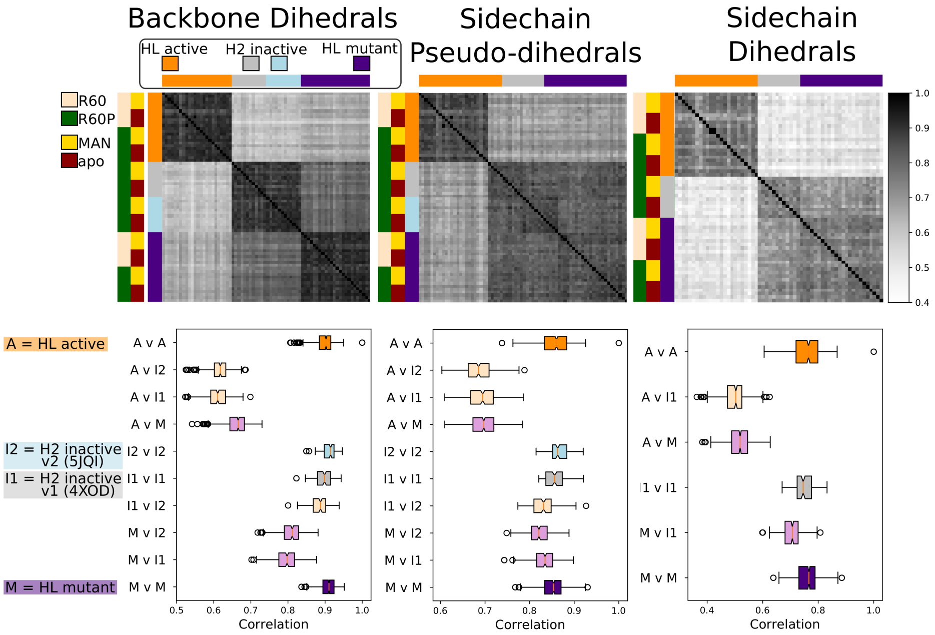

Figure S A.15: **Similarity matrices with backbone and sidechain dynamics.** Comparing HL active, H2 inactive, and HL mutant using backbone dihedral angles (left), sidechain pseudo-dihedral angles (middle), and sidechain dihedral angles χ (right) as described in Fig. SA.14. HL active is different from both H2 inactive and HL mutant for all measures. For backbone dynamics, the comparison between HL mutant and H2 inactive has significantly lower similarity than the reference comparisons. For sidechain dynamics using pseudo-dihedral angles, there are small differences between the two versions of H2 inactive that are about the same as the low similarity between HL mutant and H2 inactive. Moreover, using sidechain dihedral angles χ, there are differences between HL mutant and H2 inactive.

APPENDIX B

A NEW APPROACH FOR EXTRACTING INFORMATION FROM PROTEIN DYNAMICS

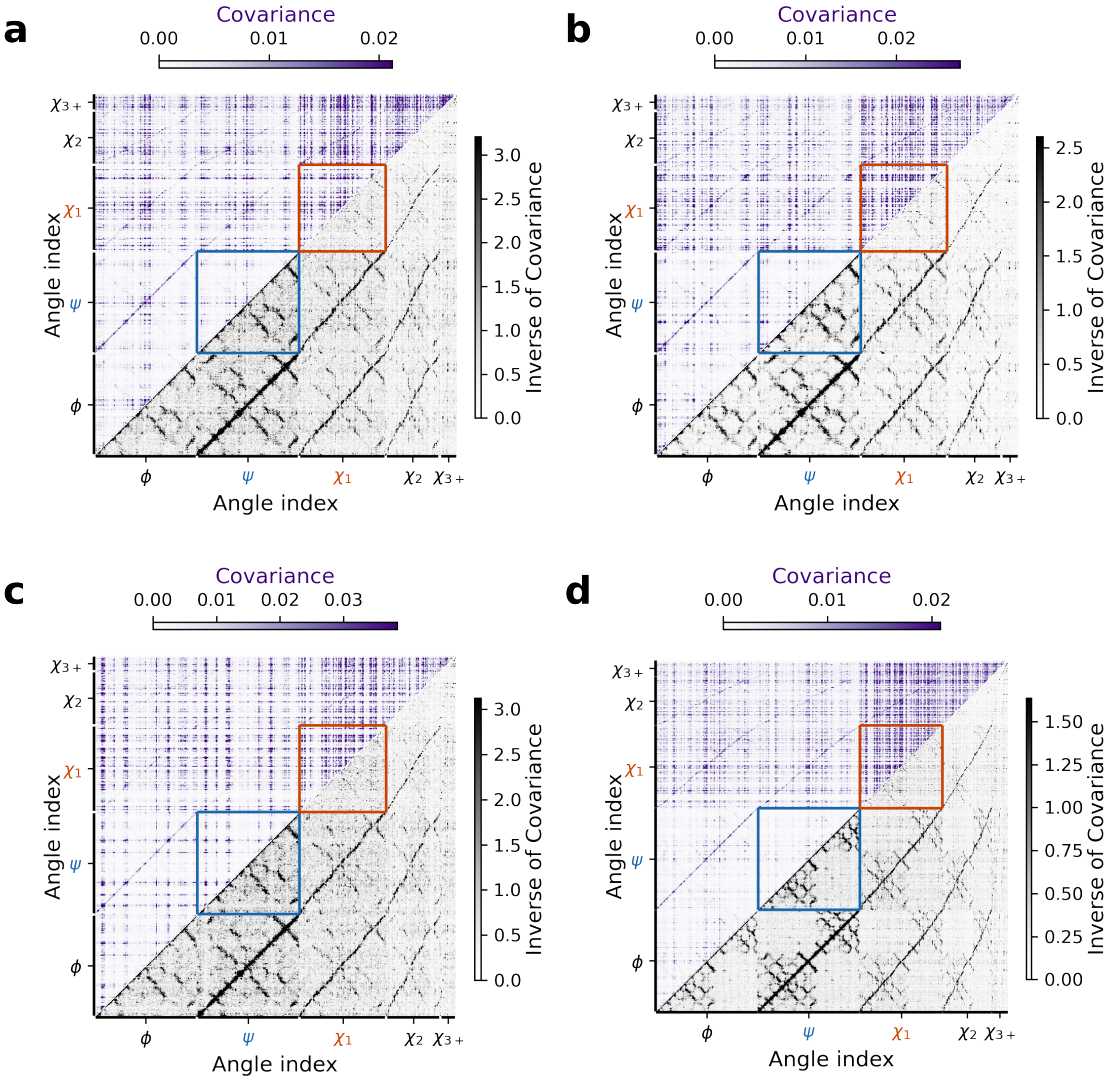

Figure S B.1: **The covariance matrix and its inverse show different patterns for FimH.** We show that the inverse covariance matrix resembles the contact map for four representative examples. **a**, Wild-type FimH$_L$ (PDBID 4AUU) in 20ns and **b**, 200ns simulations for a different replicate than the one in Fig. 4.1b. **c**, Arg-60-Pro mutant FimH$_L$ (PDBID 5MCA) in a 20ns simulation. **d**, FimH$_2$ (PDBID 4XOD), which has both the lectin and pilin domains in a 200ns simulation. The two-domain structure is visually apparent in the inverse covariance matrix. As in Fig. 4.1b, we show the covariance matrix in the top left triangle and the inverse of the covariance matrix in the bottom right triangle. For each dataset, we set the colorscale maximum to the 97th percentile.

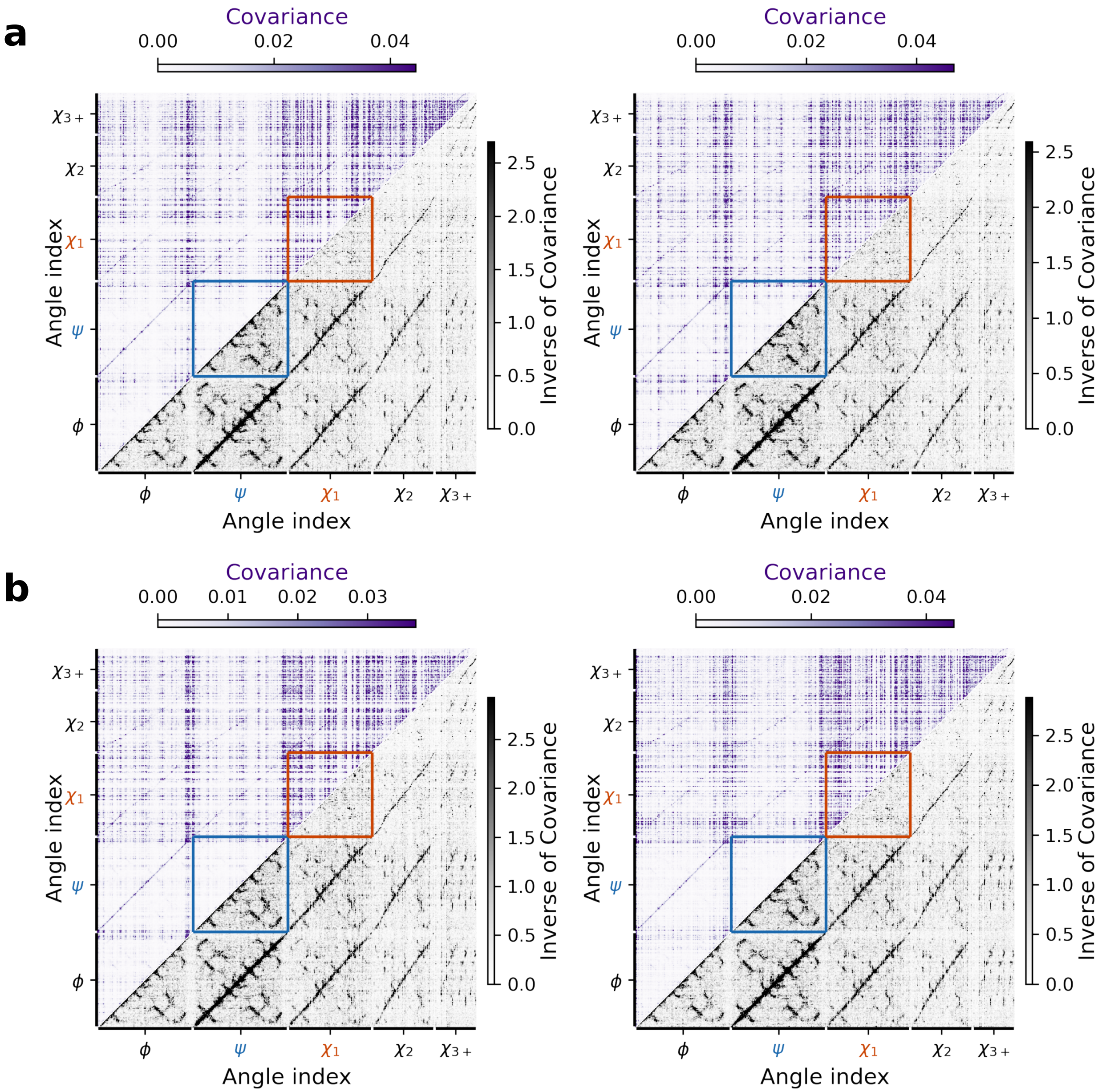

Figure S B.2: **The covariance matrix and its inverse show different patterns for Siglec-8.** We show that the inverse covariance matrix resembles the contact map for representative examples in the apo and holo states. Two replicates of Siglec-8 **a**, apo (without ligand, PDBID 2N7A) and **b**, holo (bound to 6'S-sLe^x, PDBID 2N7B) See Extended Data Fig. B.1 for description.

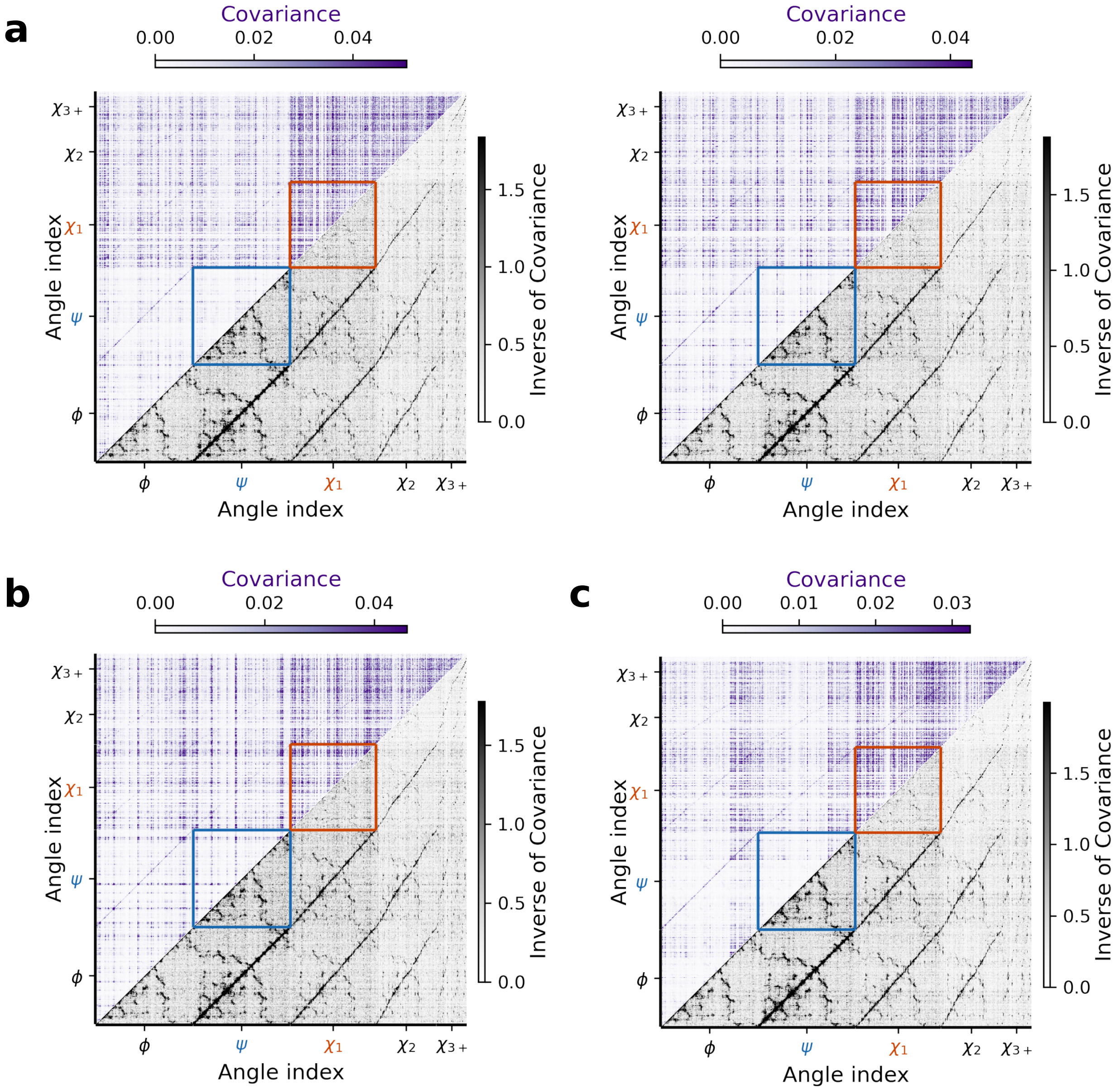

Figure S B.3: **The covariance matrix and its inverse show different patterns for SARS-CoV-2.** We show that the inverse covariance matrix resembles the contact map for representative examples of RBD-SD1 domains in the **a**, up, **b**, down, **c**, and off states. RBD-SD1 domains in the up and down states are from chains A and B of PDBID 6VSB. The off state is chain A of PDBID 6VXX. See Extended Data Fig. B.1 for description.

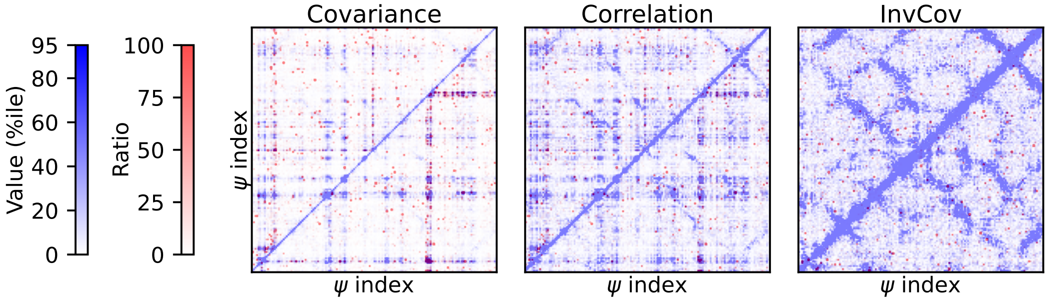

Figure S B.4: **Inverse covariance matrix has strong interactions in a pattern that is much weaker in the correlation matrix.** We show the data from Fig. 4.2 with the colormap set to the 95th percentile to show weaker interactions in the correlation matrix.

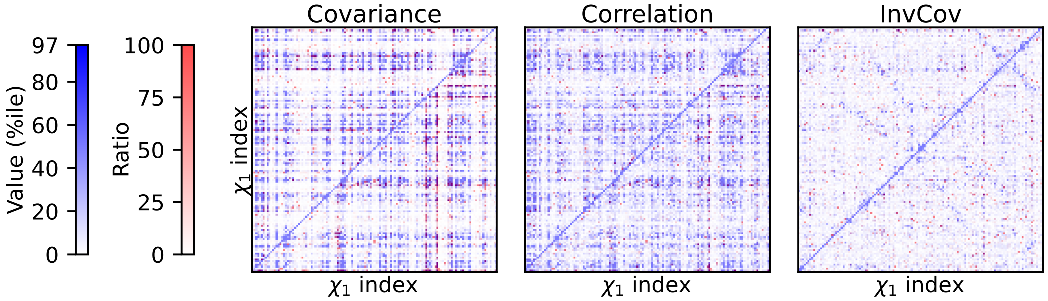

Figure S B.5: **Covariance, correlation, and inverse covariance for $\chi_1 - \chi_1$ interactions.** We show corresponding $\chi_1 - \chi_1$ interactions from Fig. 4.2. The covariance and correlation matrices show banded patterns. Bands with purple indicate strong interactions that were much weaker in the other replicate. The inverse covariance matrix shows a faint contact map pattern.

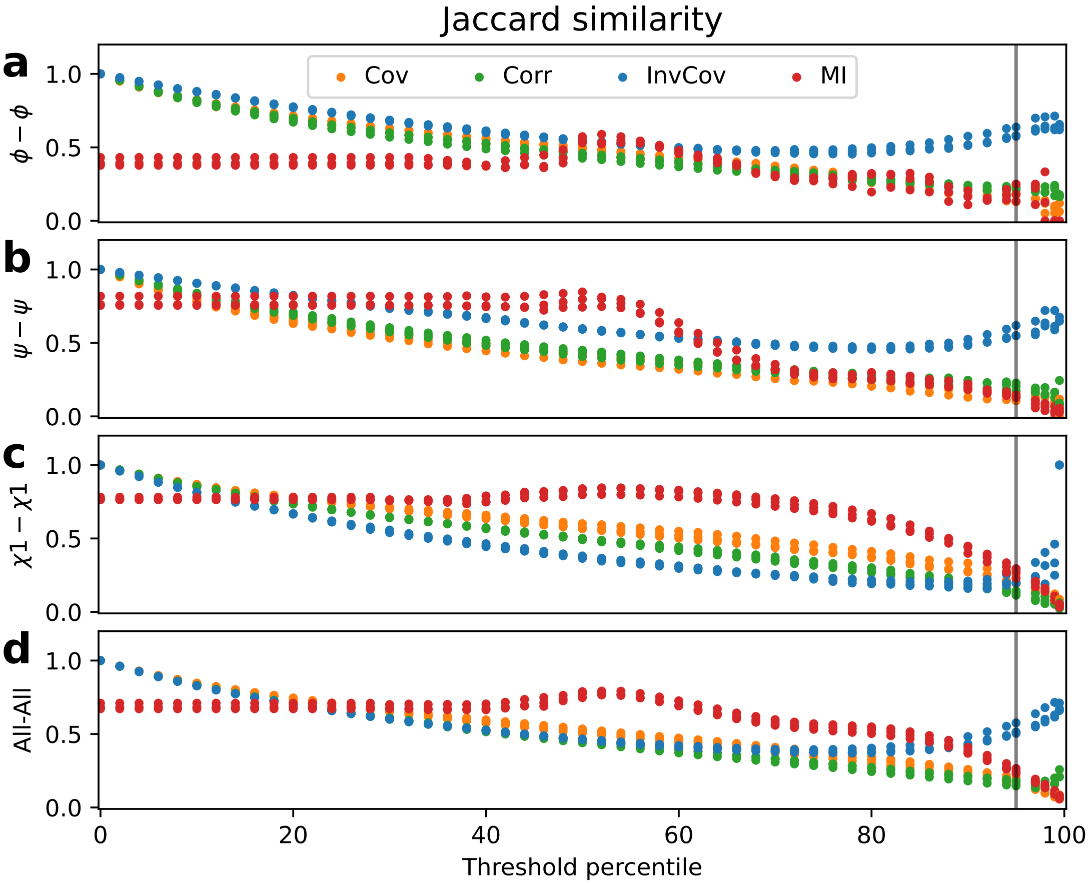

Figure S B.6: **Similarity for covariance, correlation, inverse covariance, and mutual information.** We calculate the Jaccard similarity for the networks defined by these four methods at different thresholds. We do this for the backbone-backbone and sidechain-sidechain interactions, and also for the entire protein. See Extended Data Fig. B.7-B.9 for adjacency matrix representations at representative thresholds.

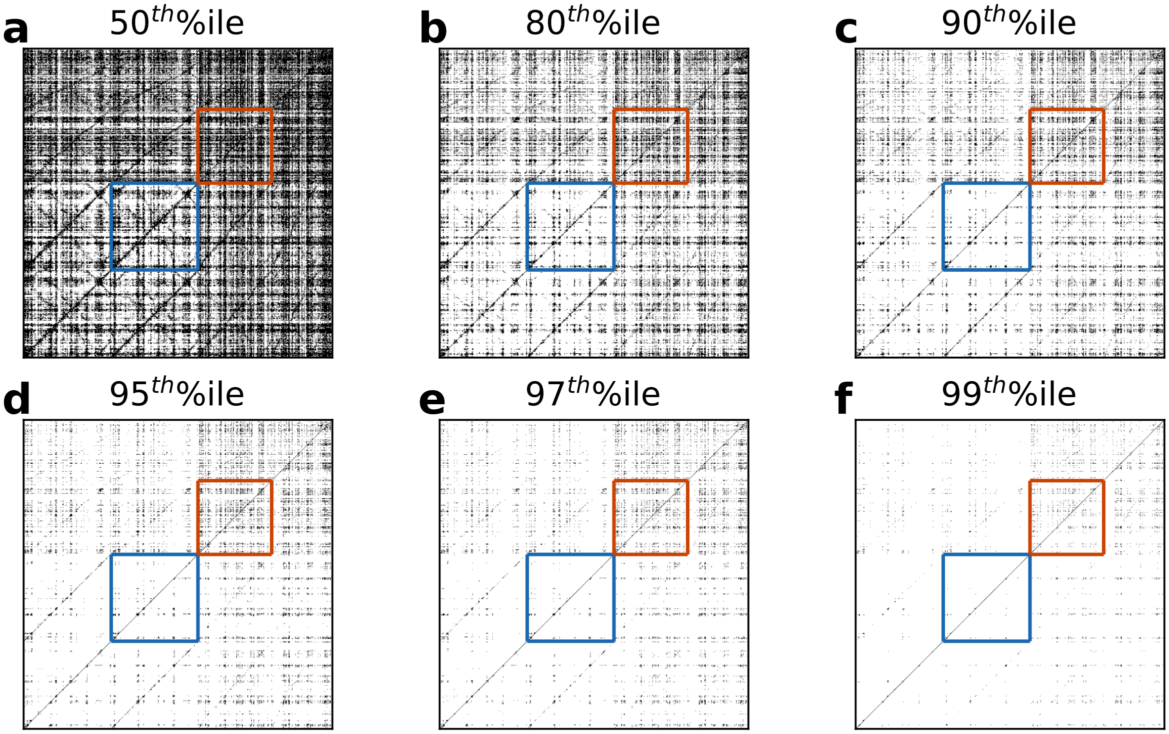

Figure S B.7: **Representative thresholds for defining networks from the covariance matrix.**

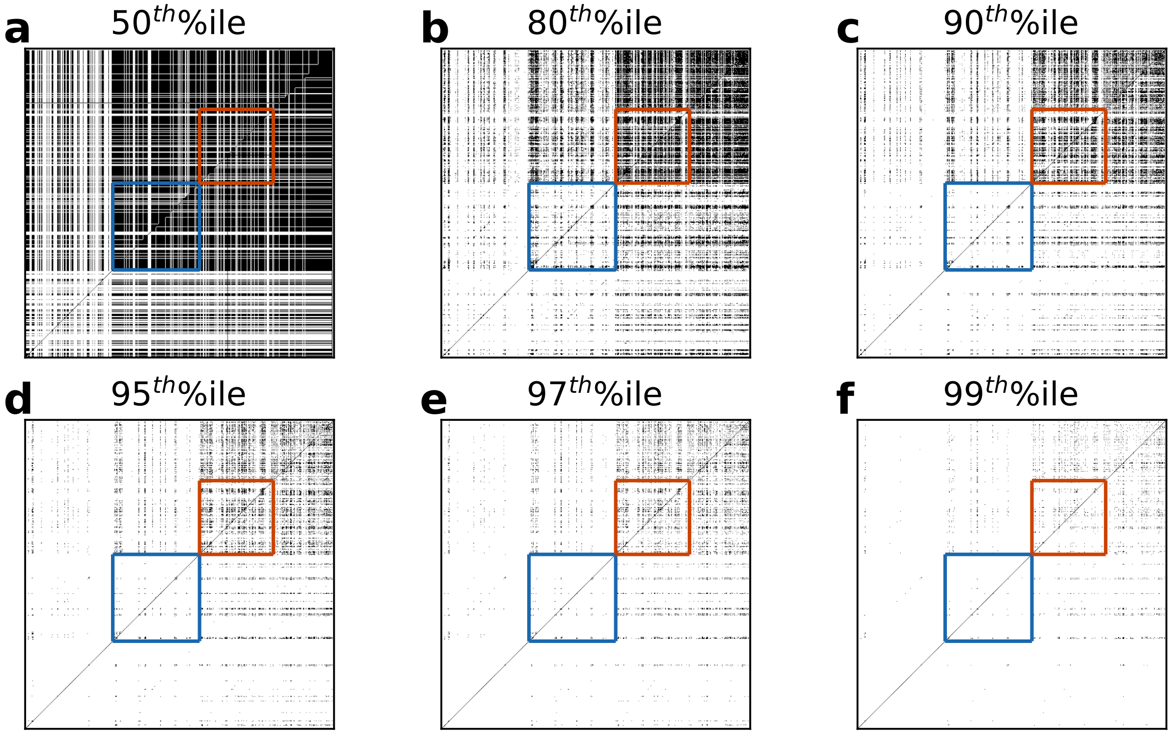

Figure S B.8: **Representative thresholds for defining networks from the mutual information matrix.**

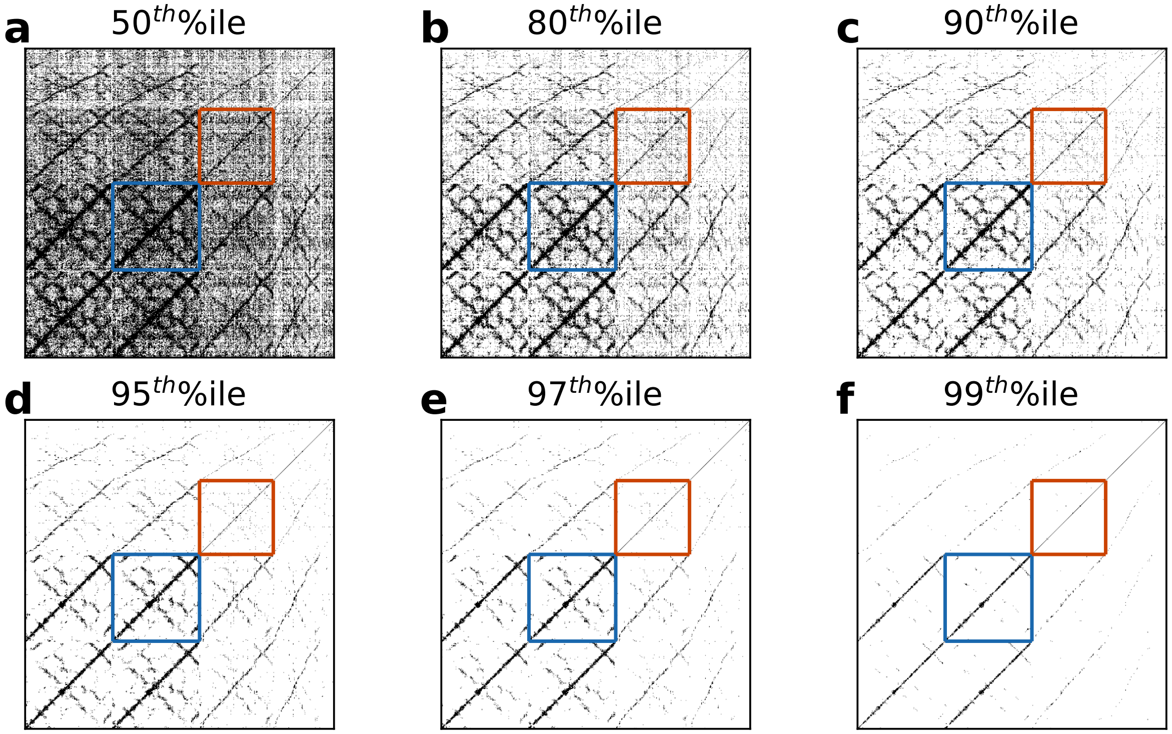

Figure S B.9: **Representative thresholds for defining networks from the inverse covariance matrix.**

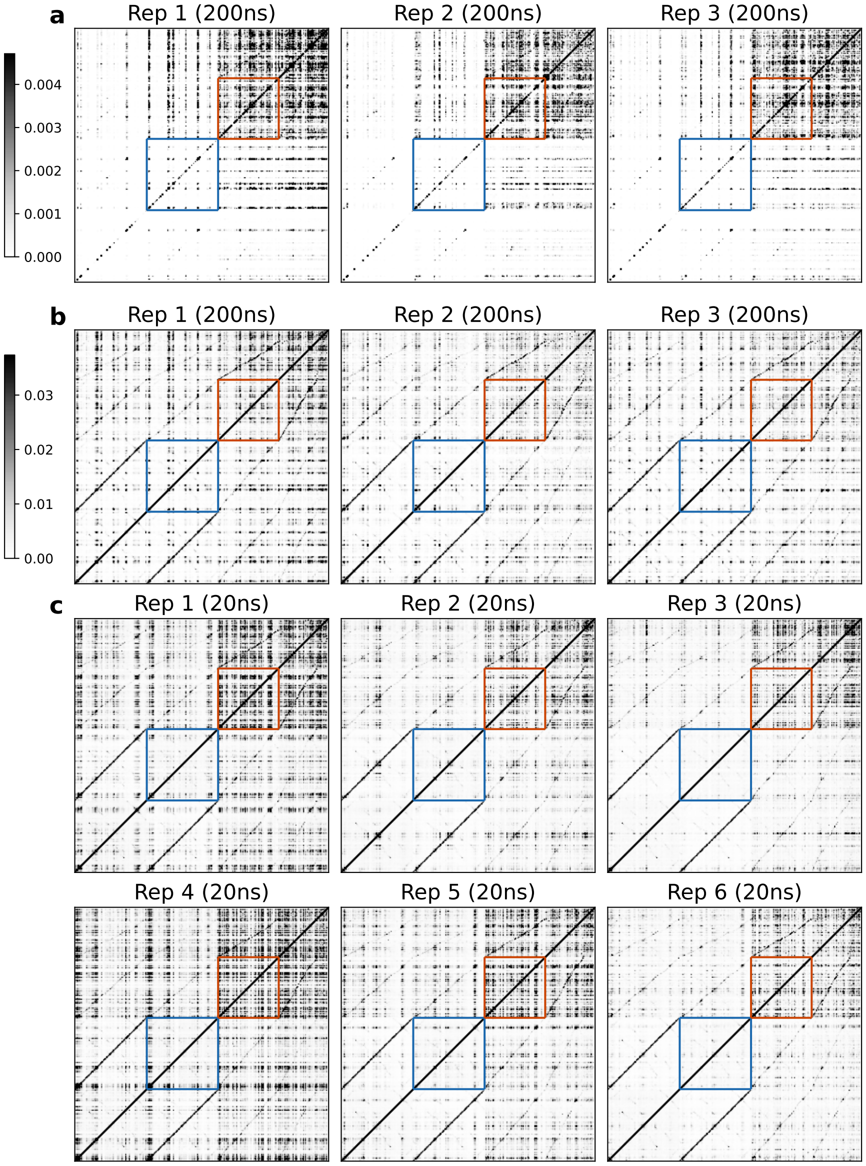

Figure S B.10: **Mutual information matrix shows sidechain-sidechain interactions are stronger than backbone-backbone interactions.** We show data for wild-type FimH$_L$. Starting with three replicate simulations of 200ns each, we show the mutual information calculated from **a**, transition state analysis [15] with a core of 90 degrees and **b**, from a histogram based approach. For each pair of angles, we define two sets of unequally spaced bins at the deciles for each individual angle. These two sets of bins are used to construct a 2 dimensional histogram. **c**, We then show six replicate simulations of 20ns. The first three are truncated versions of the longer simulations.

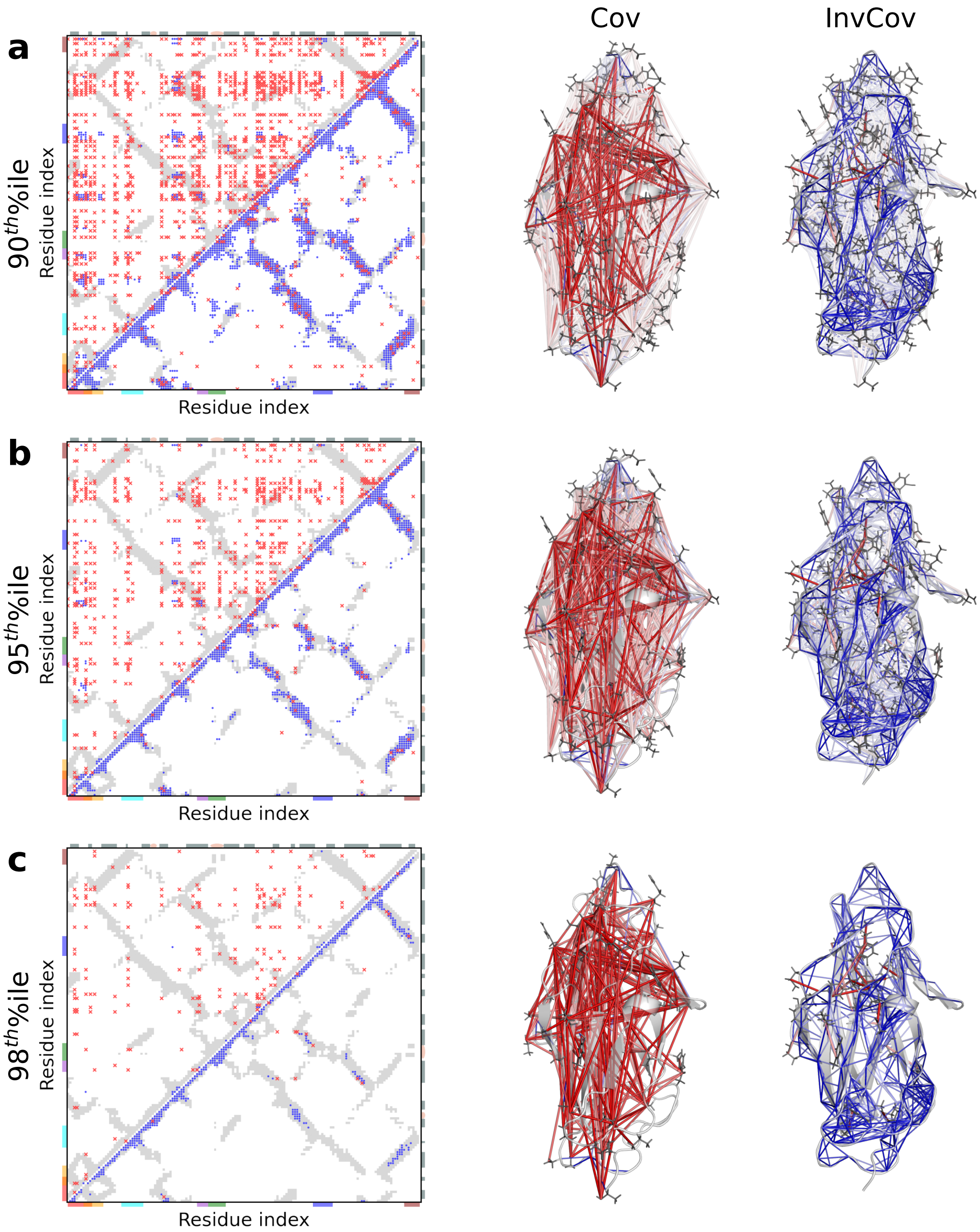

Figure S B.11: **The inverse covariance matrix has a 'contact map"-like network at multiple thresholds.** Using the same schematic as Fig. 4.3, with the threshold set at the 97_{th} percentile, we show the covariance (top left) and inverse covariance (bottom right) at various thresholds for visualizing the network: **a**, 90^{th} percentile, **b**, 95^{th} percentile, and **c**, 98^{th} percentile.

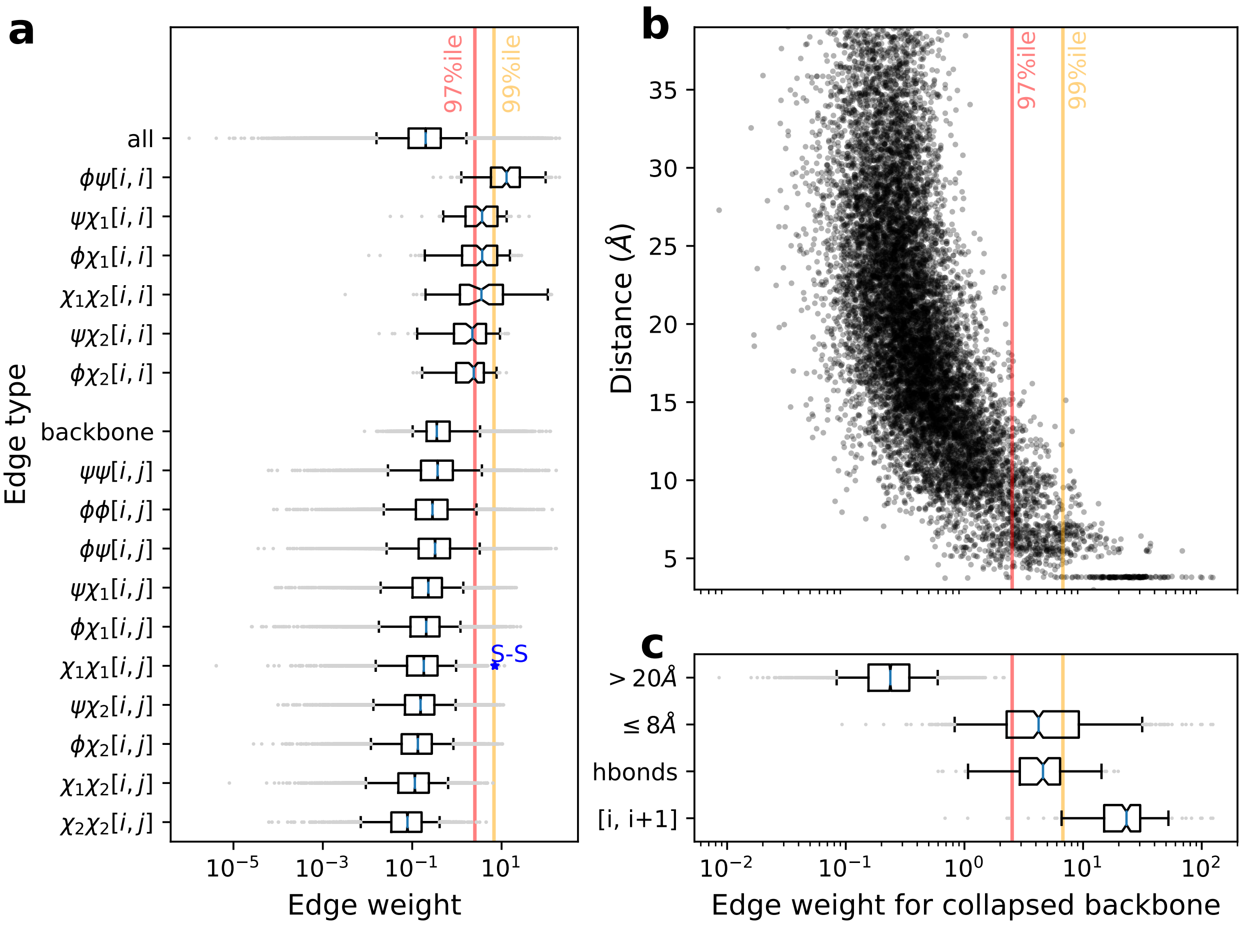

Figure S B.12: **Hierarchy of interaction strengths suggest a multilayer network.** **a**, For qualitatively different interaction types, we show the distribution of interaction strengths. We compare interactions within the same residue for different dihedral angles [i, i], as well as interactions between different residues for the same dihedral angles [i, j]. The backbone label indicates the collapse of backbone layers, as calculated from the mean interaction strength for $\phi - \phi$, $\psi - \psi$, $\phi - \psi$, and $\psi - \phi$ interactions. The box-and-whisker plots mark the 5, 25, 50, 75, and 95th percentiles, with the median in blue, and outliers in grey. The Cys3-Cys-44 disulfide bond is shown with a star in dark blue. **b**, For the collapsed backbone interactions, we show the relationship between edge strength and Cα-Cα distance. Stronger interactions are associated with smaller distance. **c**, We show distributions of interaction strengths for residue pairs with Cα-Cα distances that are far apart ($\geq$ 20Å) and close together ($\leq$ 8Å), residue pairs with backbone hydrogen bonds, and neighbors on the primary sequence (i, i+1).

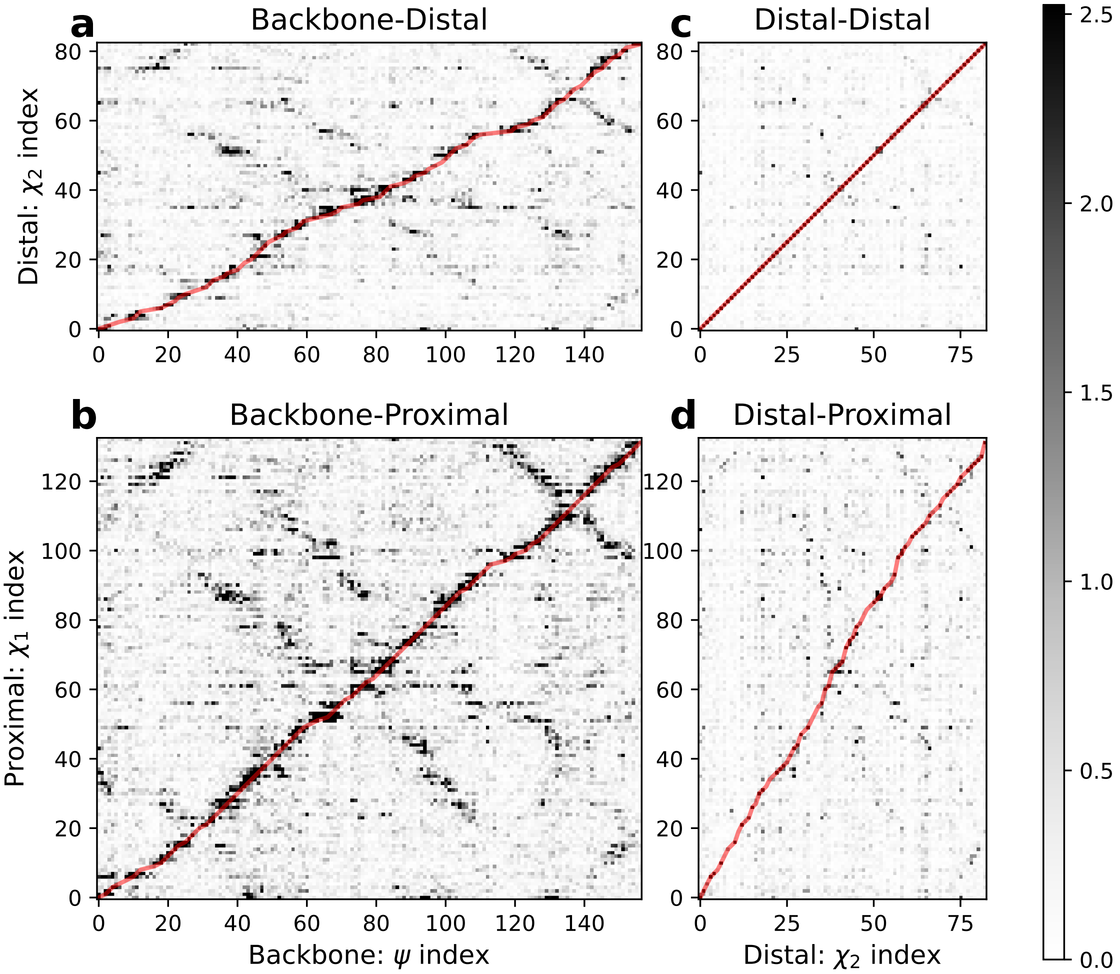

Figure S B.13: **Sidechain interactions become weaker further away from the backbone.** Grey scale indicates absolute value of the interaction strength for the inverse covariance for one replicate of wild type FimH$_\mathrm{L}$. **a,** Backbone dihedral (ψ) interactions with a distal sidechain dihedral (χ_2) is weaker than **b,** interactions with a more proximal one (χ_1). $\psi - \chi_1$ interactions still retains the contact map pattern. **c,** While interactions between distal sidechains $\chi_2 - \chi_2$ do not have a clear pattern, yet, **d,** between proximal and distal sidechains ($\chi_1 - \chi_2$), there is still a slight pattern.

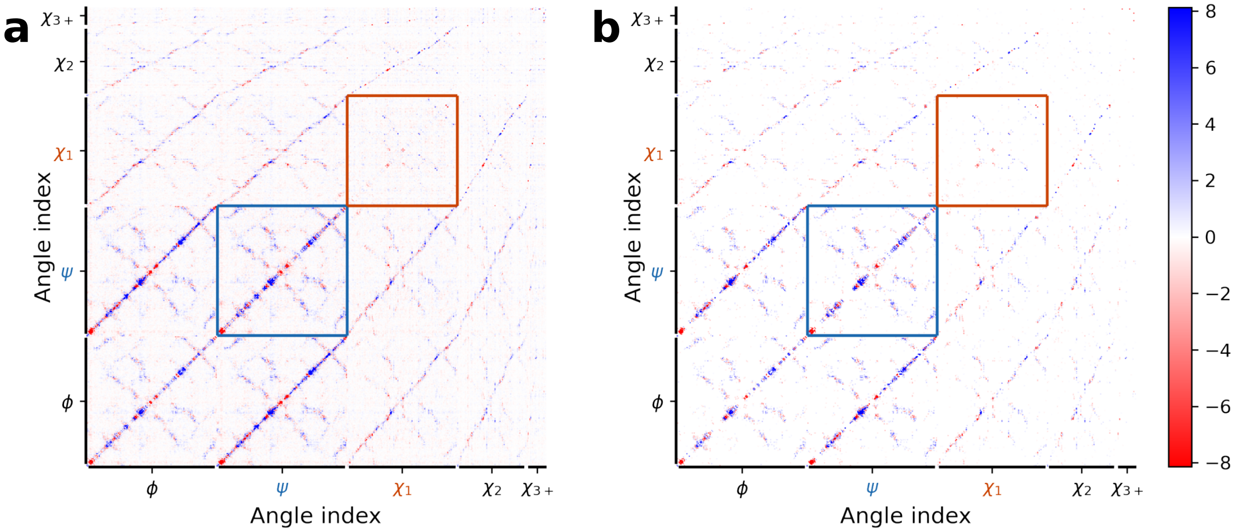

Figure S B.14: **Network comparison for wild type and mutant FimH$_L$ for all inferred edges.** For wild type - mutant, we show edges stronger in the wild type in blue, and those for the mutant in red. We show two versions: **a**, without thresholding and **b**, requiring differences to be larger than the 97$_{th}$ percentile of edge interaction strengths. We do not apply any filters based on distance, secondary structure, or other structural information.

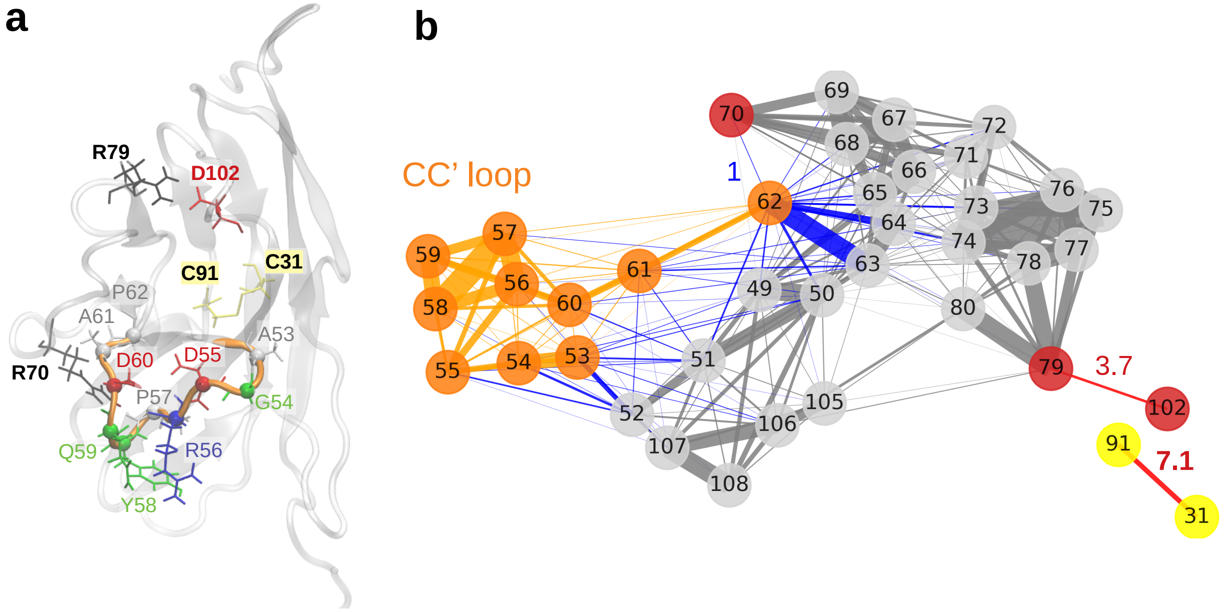

Figure S B.15: **Using networks of inferred interactions to identify stabilizing interactions with the Siglec-8 CC' binding pocket loop.** **a**, Apo Siglec-8 with the residues of the CC' loop (Ala53-Pro62) and the proposed interaction with Arg70 with Pro57 and Asp60. For reference, we also label the Arg79-Asp102 salt bridge and the Cys31-Cys91 disulfide. **b**, The section of the inferred network that includes the disulfide bond, Asp102, and all the strong interactions with and within the CC' loop. Nodes are labeled with the residue index. Edges that indicate strong interactions above the 97_{th} percentile are shown in orange for those within the loop, in grey for those outside the loop, and in blue for outside connections to the loop. Thicker edges indicate stronger interactions. Arg70 only has a strong interaction with Pro62 at the hinge of the loop. Relative to the Arg70-Pro62 edge, the interactions for Arg79-Asp102 and Cys31-Cys91 are 3.7 and 7.1-fold stronger. The strongest interactions involving the CC' loop are internal interactions and external interactions with Pro62 and Ala53. This suggests a different mechanism for the rigidity of the CC' loop.

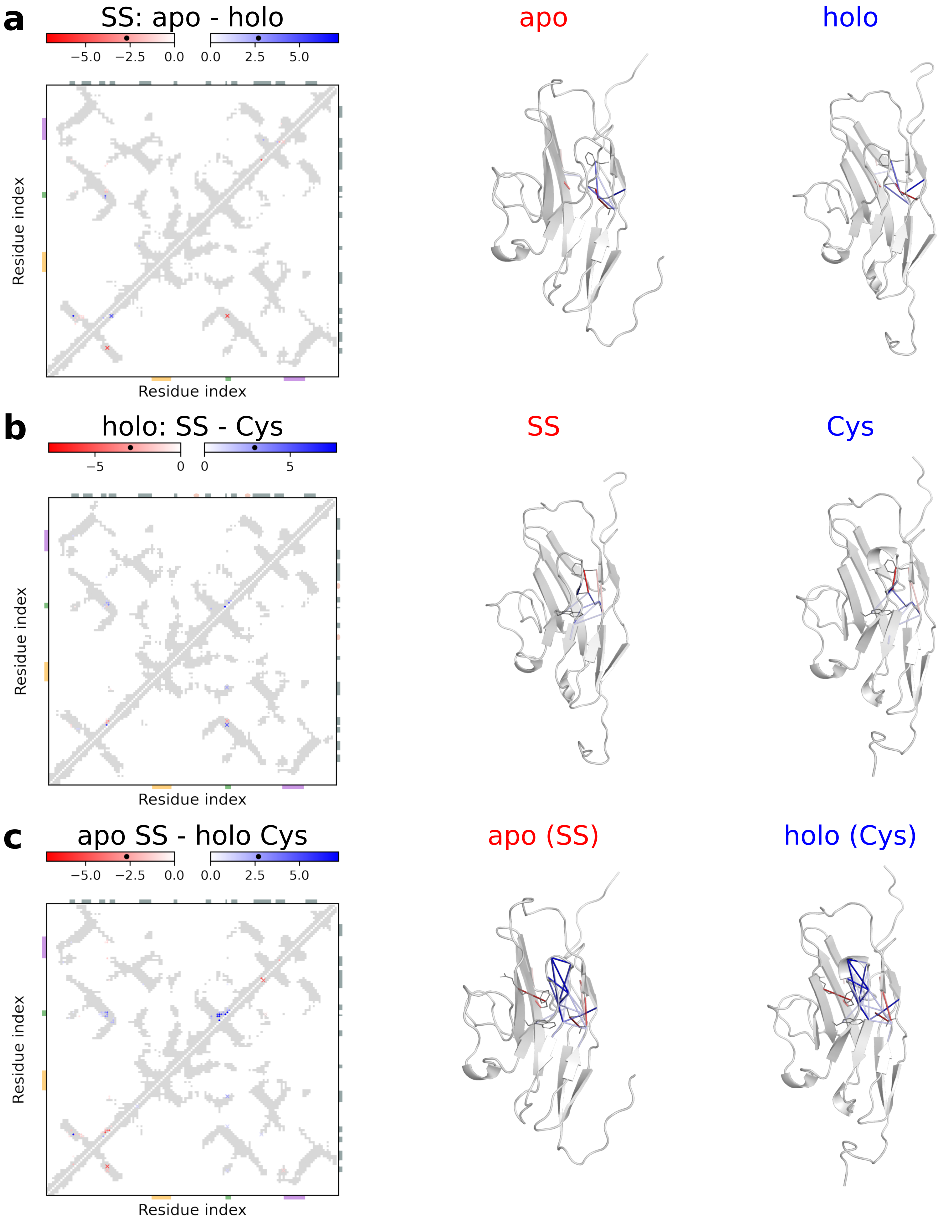

Figure S B.16: **Impact of ligand-binding and removing the Cys31-Cys91 disulfide bond on Siglec-8.** Difference in inferred network interactions for **a**, apo vs holo states, as well as **b**, the holo state with the disulfide bond vs with the bond broken *in silico*. Due to the difference in inferred interaction strength at the disulfide bond between apo and holo Siglec-8, we also compared **c**, the apo state with the disulfide bond intact vs the holo state with the bond broken. We use the same representation scheme as in Fig. 4.4. We show landmarks on the bottom and left borders: CC' loop (orange), GG' loop (purple), and the evolutionarily conserved Asp90-Cys91-Ser92 motif (green).

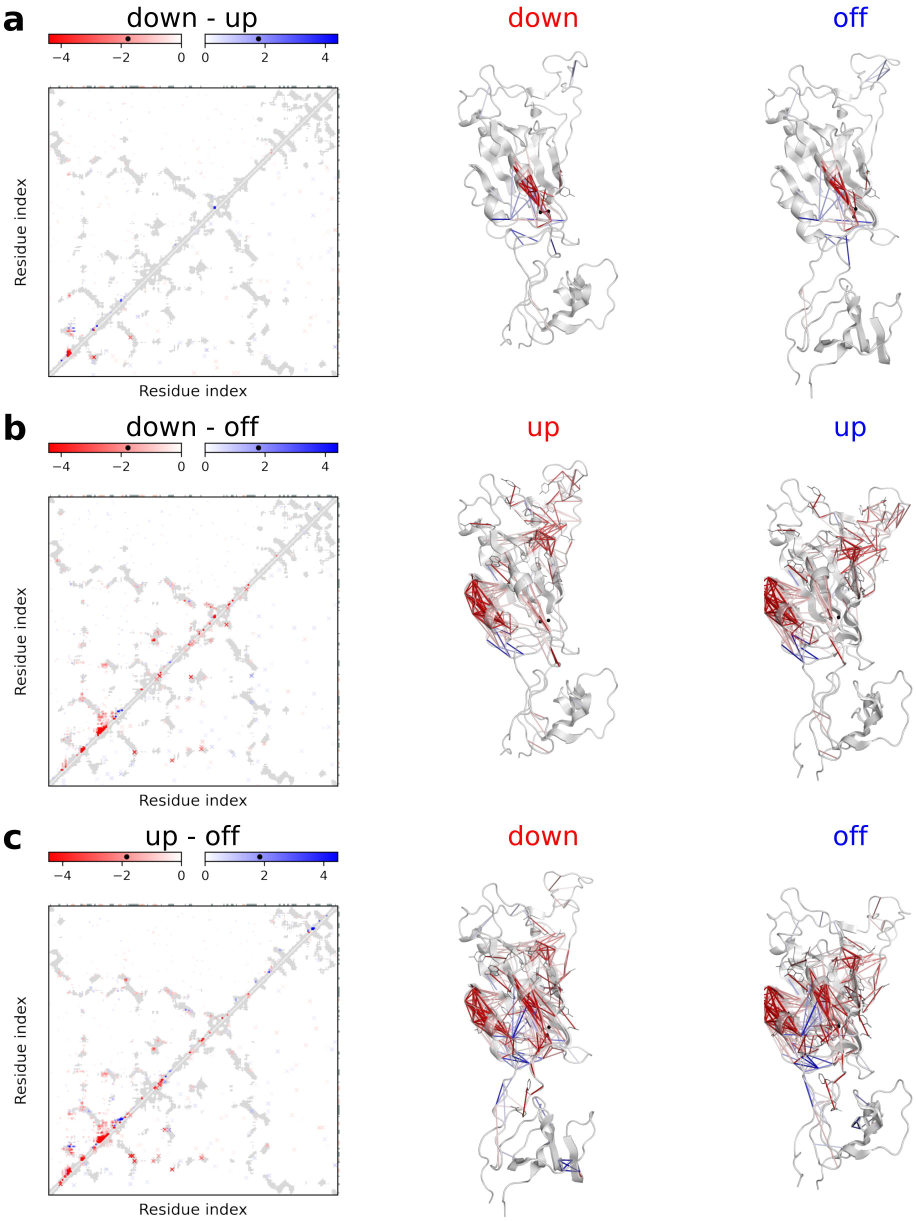

Figure S B.17: **Impact of up, down, or off state for the SARS-CoV2-spike protein RBD-SD1 domains.** In the up state (PDBID 6VSB, chain A), the RBD is accessible to bind human ACE2 for viral attachment. In the neighboring protomer's down state (PDBID 6VSB, chain B), the RBD is hidden. In the off state (PDBID 6VXX, chain A), all RBDs are hidden. Difference in inferred network interactions for **a**, down vs up states, **b**, down vs closed states, **c**, and up vs closed states. We use the same representation scheme as in Fig. 4.4 and Extended Data Fig. SB.16.

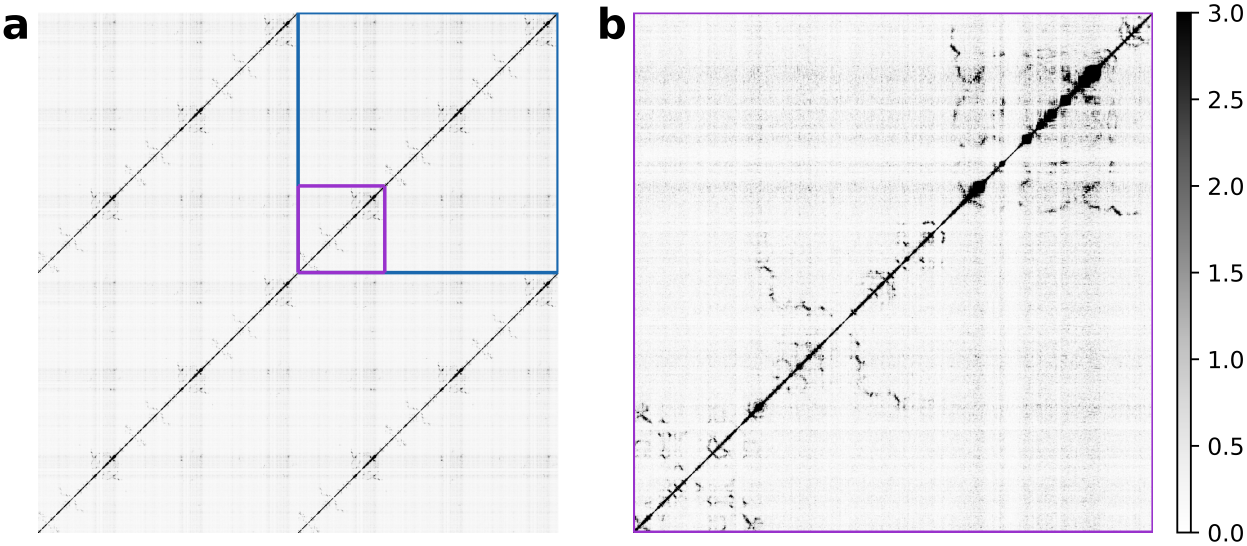

Figure S B.18: **Network inference for the trimeric spike protein.** **a**, Backbone-backbone interactions for the trimer, with ψ-ψ interactions highlighted by the blue box. The first protomer is shown with a purple box. **b**, Zoomed-in view of the first protomer's ψ-ψ interactions.

VITA

Jenny Liu grew up in Mayfield Heights near Cleveland, Ohio. She went to Washington University in St. Louis, MO for college, studying electrical and biomedical engineering between 2009-2013. During this time, she took no mechanical engineering classes. After attempting some wet lab research, she decided computational research might be more suitable. In 2015, she entered medical scientist training program at Northwestern University in Chicago, IL. After two years of medical school, she started studying mechanical engineering.