The Effective Coarse-grained Hessian Dictates Allostery in IGPS

Peter T. Lake,[†] Russell B. Davidson,[†] Heidi Klem,[†] Glen M. Hocky,[‡] and Martin McCullagh[*,†]

†*Department of Chemistry, Colorado State University, Fort Collins, CO, USA*

‡*Department of Chemistry, New York University, New York, NY, USA*

E-mail: martin.mccullagh@colostate.edu

Abstract

The long-ranged coupling between sites that gives rise to allostery in a protein is built up from short-ranged physical interactions. Computational tools used to predict this coupling and its functional relevance have relied heavily on the application of graph theoretical metrics to residue-level correlations measured from all-atom molecular dynamics simulations. The short-ranged interactions that yield these residue-level correlations, and thus the appropriate graph Laplacian, are quantified by the effective coarse-grained Hessian. Here we compute an effective harmonic coarse-grained Hessian for a benchmark allosteric protein, IGPS, and demonstrate the improved locality of this Laplacian over two other connectivity matrices. Additionally, two centrality metrics are developed that indicate the direct and indirect importance of each residue at producing the covariance between the effector binding pocket and the active site. The results from this procedure are corroborated by previous mutagenesis experiments and lead to unique functional insights. In contrast to previous computational analyses, our results suggest that fP76-hK181, not fD98-hK181, is the most important contact for conveying direct allosteric paths across the HisF-HisH interface. fD98 is found to play a minor role in

paths and contribute greatly to indirect allostery between the effector binding pocket and the glutaminase active site.

Introduction

Allostery refers to the long-range functional coupling of sites in a macromolecule through networks of short-ranged interactions. This phenomenon can be a pivotal component of a protein's function[1] as demonstrated in GPCR signaling,[2] coupling of ATP binding and hydrolysis to mechanical work in motor proteins[3–6] and activation of oxygen binding in hemoglobin.[7–12] Many of these processes are initiated by the binding of an effector molecule that modulates the activity at a distal active site. This behavior has become an increasingly important target in the field of drug development due to possible improvements in selectivity over orthosteric sites.[13–17] Intriguingly, allostery can be incorporated into an enzyme's function even on the short time scales of directed evolution studies to produce a desired increase in activity.[18] Therefore, it is highly desirable to be able to identify and predict allostery, as well as the interactions ("allosteric pathways") involved in these processes.

There are many well established experimental techniques for characterizing allostery, including activity based assays to investigate non-Michaelis-Menten kinetic behavior,[19] H/D mass spectrometry,[20,21] as well as structural based approaches such as X-ray crystallography,[22] cryoEM,[23] and NMR.[24–32] Structural techniques can be used to identify residues that interact directly with or are structurally perturbed by the effector molecule by comparing *apo* and effector bound states of a protein. Identifying allosteric pathways is significantly more challenging. While pathways have been identified using experimental techniques in well-studied proteins such as hemoglobin,[10,33–35] the combination of NMR spectroscopy and computational techniques have allowed for the most robust description of allosteric pathways.[27,30,31,36,37]

Computational techniques used to investigate allostery rely on graph theoretical approaches to identify important residues or connections that convey information from the effector binding pocket to the active site.[29,38–40] A graph is constructed by populating a pairwise adjacency matrix, $\boldsymbol{A}$, that has finite values between connected residues. The closely related graph Lapla-

cian, $\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{A}$ where $\boldsymbol{D}$ is a diagonal matrix with elements $D_{ii} = \sum_j A_{ij}$, can be readily constructed the adjacency matrix. Once nodes and edge weights have been defined, graph theoretical analyses can be applied to identify paths in the graph, using Dijkstra's algorithm[41] or the Floyd-Warshall algorithm.[42] A variety of edge and node centrality metrics have been developed using these adjacency matrices to identify important amino acids contributing to the allosteric pathways in the protein.[29,37,43]

Two computational techniques have been used to generate these graphs: bioinformatics and molecular dynamics. Sequence-based bioinformatic methods, working under the assumption that allosteric pathways are functionally important and thus evolutionarily conserved, use residue pairwise sequence co-evolution to define a graph.[29,35,44] These methods do not explicitly take into account the 3D structure of a given protein as this information should be implicitly captured in sequence covariance. Additionally, structural databases can be used to build pairwise, knowledge-based potentials that describe the interactions observed in e.g. a protein's crystal structure.[45] None of these models, however, directly account for the ensemble nature of protein structure and the associated allosteric behavior.[11]

All-atom molecular dynamics (aaMD) is a well-established method to sample the configurational ensemble of a protein.[46–51] Analyses of such simulations have been used to identify allosteric importance based on residue interaction networks[29,52–54] or positional covariance-based metrics.[36,37,54–56] Both of these models are used to describe the average residue-residue couplings from the ensemble of protein configurations observed. Interestingly, these methods lead to dramatically different graphs: the residue interaction networks will be localized in space while the positional covariance will be delocalized. Thus, while aaMD provides an appealing measure of protein configurational space, the appropriate graph Laplacian to describe residue-level correlations is not well understood.

An alternative to using some form of the positional covariance as the adjacency matrix is to use the Hessian, which is a graph Laplacian constructed from the second spatial derivative of the Hamiltonian. Even though the Hessian is closely related to the covariance of the system, the difficulty of measuring it from simulation has made it relatively unused for the study of allostery. The separate field of bottom-up coarse-graining has previously approached the

problem of deriving an effective Hamiltonian of a protein in terms of a collection of springs, producing methods such as REACH and hENM.[57–59] We propose to connect these two fields of study, namely protein allostery and coarse-grained potentials, and use the resulting effective coarse-grained Hessian as the graph Laplacian to study protein allostery.

In the next section, we demonstrate that the effective coarse-grained Hessian is the appropriate graph Laplacian to quantify residue-level allostery from aaMD simulations. Motivated by the physical interpretation of the Hessian, we define two centrality metrics that indicate a Hessian element's importance at conveying covariance between a selected set of sources and sink residues. Finally, we apply this method to simulations of the imidazole glycerol phosphate synthase (IGPS) heterodimer, which is a well established benchmark allosteric system. *Our definition and resulting weights of allosteric pathways provide new insight into the physics underlying allostery, a key functional component of most proteins.*

Theoretical Framework

Structural allostery can be described as the positional change at one site due to the application of an external force at a different site. A linear response of a system to an external force, $\boldsymbol{f_{ex}}$, can be written as

$$\boldsymbol{\Delta x} = \frac{1}{T}\boldsymbol{C}\boldsymbol{f_{ex}}, \tag{1}$$

where $\boldsymbol{\Delta x}$ is the change in the equilibrium position due to the perturbation, $\boldsymbol{C}$ is the matrix of covariance of particle positions, and T is temperature in units of energy.[60] Equation 1 demonstrates the importance of positional covariance to predicting allostery, but does not indicate anything about the short-ranged physical interactions leading to this behavior.

Within a graph theoretical framework, the short-ranged interactions that lead to long-ranged correlations are what should populate the graph Laplacian of the system. The Hessian is the appropriate graph Laplacian for molecular systems because (1) Hessian elements are only finite for short-ranged physical interactions, (2) the Hessian is, by construction, a symmetric positive semi-definite matrix as all graph Laplacians are, and (3) the covariance is related to the generalized inverse of the Hessian. The final point is most well-known in elastic network

models (ENMs) for which the covariance and Hessian are related by the Moore-Penrose pseudoinverse ($\boldsymbol{H}^{+} \equiv \frac{1}{T}\boldsymbol{C}$). This relationship is consistent with the Gaussian Markov random field literature in which the graph Laplacian and the covariance are related by pseudoinverses.[61] In the context of allostery in proteins, the covariance is typically measured at a residue-level even if the underlying simulations have atomic resolution. Thus, the Hessian of interest is the second derivative of an effective coarse-grained Hamiltonian. Determining the effective Hessian from a measured covariance is a non-trivial problem due to the difficulties in converging a measured covariance.[62] Here we use a previously described coarse-graining procedure to generate an effective harmonic Hessian based on a measured covariance.

The Effective Harmonic Hessian from All-Atom Molecular Dynamics Simulations

Explicit solvent aaMD simulations provide a detailed and accurate structural ensemble picture of proteins under physiological conditions.[46–51] These simulations have been used to generate $3N \times 3N$ covariance matrices that provide the information necessary to predict $N-$residue level structural allostery. Motivated by the idea that the structural ensemble in a single free energy well from aaMD is well represented by a harmonic system,[63–67] effective harmonic Hamiltonians have been fit to mapped all-atom data such as the $N \times N$ covariance.[57–59] In this work, we construct an effective harmonic Hessian using a slightly modified heterogeneous-ENM (hENM) procedure[58] as described in the Supplementary Information (SI) Computational Methods section.

The result of the hENM procedure is a $N \times N$ force constant matrix, $\boldsymbol{k}$, that is optimized to reproduce pairwise particle variances. The 3×3 tensor element of the $3N \times 3N$ Hessian is defined as

$$\boldsymbol{H}_{ij} = -k_{ij}\hat{\boldsymbol{R}}_{ij} \otimes \hat{\boldsymbol{R}}_{ij}, \tag{2}$$

where $\hat{\boldsymbol{R}}_{ij} = \frac{\langle \boldsymbol{r}_i \rangle - \langle \boldsymbol{r}_j \rangle}{|\langle \boldsymbol{r}_i \rangle - \langle \boldsymbol{r}_j \rangle|}$ and $\langle \boldsymbol{r}_i \rangle$ is the average position of node i. The $3N \times 3N$ covariance can then be reconstructed as $\boldsymbol{C} = T\boldsymbol{H}^{+}$. The corresponding covariance calculated from the hENM Hessian demonstrates good correlation with the measured aaMD covariance (Figure S1).

Allosteric Paths in the Hessian

To investigate allosteric pathways, we start by dictating that the sum over these pathways yields the covariance between a selected pair of source and sink residues. This is motivated by the idea that the covariance is the physical observable that dictates the linear response between two residues. Employing a property from graph theory that, given a weighted adjacency matrix, $\boldsymbol{A}$, $(\boldsymbol{A}^{\ell})_{ij}$ is the weighted sum of all walks of length ℓ between nodes i and j, we can express the covariance in the following way,

$$\boldsymbol{C} = \sum_{\ell=0}^{\infty} \boldsymbol{A}^{\ell} = (\boldsymbol{I} - \boldsymbol{A})^{-1}\,, \tag{3}$$

where I is the identity matrix and the last equality arises from identifying the infinite sum as an example of the Neumann series. At this juncture, it is natural to write that $\boldsymbol{A} = \boldsymbol{I} - \boldsymbol{C}^{-1}$ but, in the context of molecular simulations, $\boldsymbol{C}$ is a singular matrix and, thus, not invertible due to the removal of center-of-mass translation and rotation.

If we consider $\boldsymbol{C}$ to be strictly invertible and map the infinite number of walks to a finite set of paths, P_{ij}, it can be shown that

$$\frac{1}{T}\boldsymbol{C}_{ij} = \sum_{n \in P_{ij}} \boldsymbol{H}_{ii}^{-1} \prod_{\langle \alpha,\beta \rangle \in n} -\boldsymbol{H}_{\alpha\beta} \times \left\{ \left[\boldsymbol{H}_{\{\alpha\}\mathrm{prev}} \right]^{-1} \right\}_{\beta\beta}\,, \tag{4}$$

where $\langle \alpha, \beta \rangle$ is an edge in the path and subscript $\{\alpha\}$prev denotes the principle submatrix of $\boldsymbol{H}$ obtained by removing all nodes previously visited in the path. The inverse terms in Equation 4 arise from mapping walks to paths; traversing all loops from α to itself results in terms related to the conditional variance of α. The only terms that couple two nodes together in Equation 4 are $\boldsymbol{H}_{\alpha\beta}$ which are the terms in the Hessian. We note that these terms in the Hessian are 3×3 tensors for a 3D system.

Paths are usually referenced in terms of a length ℓ as opposed to a weight. This distinction can be made by considering a path length as being the negative of a sum of the adjacency matrix on a logarithmic scale as opposed to the weight which is a product of elements in the adjacency matrix. In this way, the paths with the largest weights have the shortest lengths. Previous

studies of paths in protein allostery have relied upon exact evaluation of the shortest paths. Alternatively, we propose to study these paths by sampling them in a Monte Carlo scheme; even though the paths observed are not necessarily all the shortest paths within some number, the algorithm is much more efficient than the typically used Floyd-Warshall algorithm.[41,42] This method of sampling paths also leads us to consider these paths as part of an ensemble, all of which can contribute. The Monte Carlo sampling is also readily extended to account for path lengths which are not strictly a sum of pairwise interactions. The procedure is detailed in the SI Theory section.

A similar but more complicated procedure can be followed to determine paths that yield the covariance in a singular covariance matrix. The most crucial idea obtained from performing the full derivation is that fully connected paths, those with finite Hessian values connecting a given source and sink, are not the only contributions to the covariance. Additional terms, that can be described as broken paths, contribute to the covariance due to a coupling through the null-space of the singular matrix. The relevant outcome of this for the current work is that residues that contribute to direct paths are not the only residues that affect the covariance between a source and sink. This motivates the need for additional importance metrics to study allosteric interactions in a given graph.

Hessian Derivative as a Centrality Metric

The effective harmonic Hessian lends itself to another centrality metric. One can consider how changing a single spring constant alters the covariance between a given set of sources and sinks. This leads to an edge-based centrality metric we call the derivative metric, defined by

$$\delta_{edge}^{(ij)} = \frac{d\|\boldsymbol{C}_{mn}\|^2}{dk_{ij}}, \tag{5}$$

where $\|\boldsymbol{C}_{mn}\|^2$ is the squared Frobenius norm of the covariance tensor between source m and sink n. We map this metric down to a node-based centrality metric by

$$\delta_{node}^{(i)} = \sum_j k_{ij} \frac{d\|\boldsymbol{C}_{mn}\|^2}{dk_{ij}}. \tag{6}$$

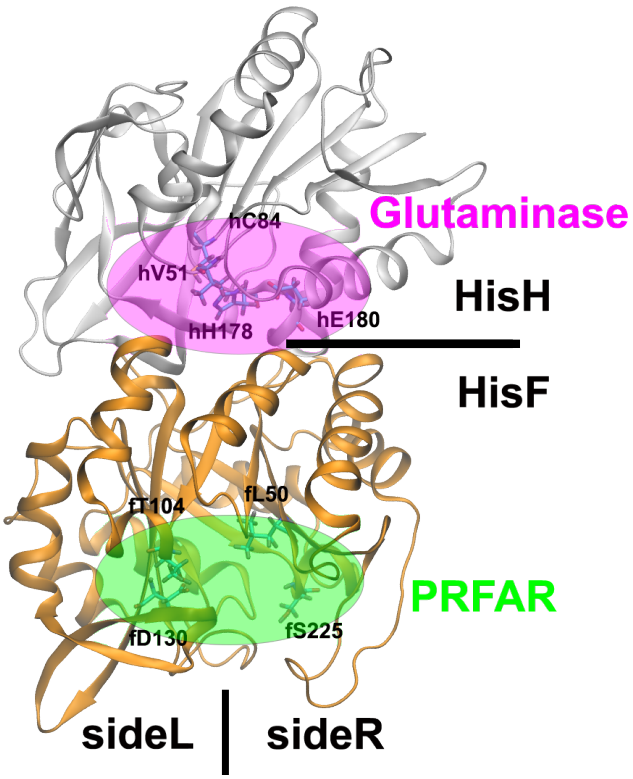

Figure 1: Structural depiction of the IGPS heterodimer. The protein is composed of HisH and HisF monomers. The PRFAR ligand binds in the pocket indicated by the green oval and source residues fL50, fT104, fD130 and fS225. The glutaminase active site, labeled in pink, is located in the HisH monomer near the interface. We chose four sink residues hV51, hC84, hH178 and hE180 to identify this pocket.

This mapping of the edge metric to a node metric is motivated by considering a fractional change of all the spring constants connecting a given node. These values can be positive and negative and vary by orders of magnitude, ranging from -5.57×10^{-3} Å^4 to 2.20×10^{-4} Å^4.

Results and Discussion

Model protein - IGPS

Imidazole glycerol phosphate synthase (IGPS) is an enzyme that functions in both the purine and histidine biosynthesis pathways of plants, fungi, archaea, and bacteria. In *Thermatoga maritima*, IGPS is a heterodimeric protein complex of HisH and HisF proteins (depicted in white and orange respectively in Figure 1). HisH catalyzes the hydrolysis of glutamine into

ammonia and glutamate. Nascent ammonia is then shuttled across the HisF–HisH interface and through the $(\beta/\alpha)_8$ barrel of HisF (an approximate distance of 25 Å). At the HisF cyclase active site, the ammonia reacts with N'-[(5'-phosphoribulosyl)formimino]-5-aminoimidazole-4-carboxamide ribonucleotide (PRFAR) to form imidazole glycerol phosphate (IGP) and 5-aminoimidazole-4-carboxamide (AICAR). These two reactions are strongly coupled with a 1:1 stoichiometry, despite the large distance that separates the two active sites.[68,69] In addition to the concerted mechanism between HisF and HisH, IGPS is classified as a V-type allosteric enzyme, such that the rate of the glutaminase reaction is critically dependent on the presence of the PRFAR ligand. Experimental assays have quantified this strong allosteric activation to be an approximate 4,900-fold increase in activity relative to basal levels.[70] Allosteric activation is also observed in the presence of the cyclization products, IGP and AICAR, but at reduced magnitudes.[70,71]

The allostery of the IGPS enzyme has been studied for nearly two decades using a broad range of experimental and theoretical methodologies. Mutational studies, coupled with biochemical assays and NMR experiments, have highlighted residues that are pivotal for the functionality of the protein complex as well as dynamic effects induced by the HisF ligand.[30,31,70–78] Additionally, MD simulations and graph-theoretic analyses have been applied to the IGPS system to study the allosteric effect of PRFAR and the allosteric paths that couple the two distal active sites.[30,36,37,54,55,73,79–82] For this work, we have performed aaMD simulations of *apo* IGPS as well as one mutant variant, totalling 1μs of trajectory for each system. Subsequently, an effective harmonic Hessian has been constructed and analyzed for both systems (SI Computational Methods).

Adjacency Matrix Comparison

Adjacency matrices based on the covariance and mutual information (MI) have been used to study allostery in IGPS.[54,55,81] The theoretical framework outlined above demonstrates the importance of a third adjacency matrix, one based on the Hessian. In order to compare and contrast between these three, we consider absolute values of normalized adjacency matrices. These matrices are limited to values between 0 and 1, with values near one indicating strong

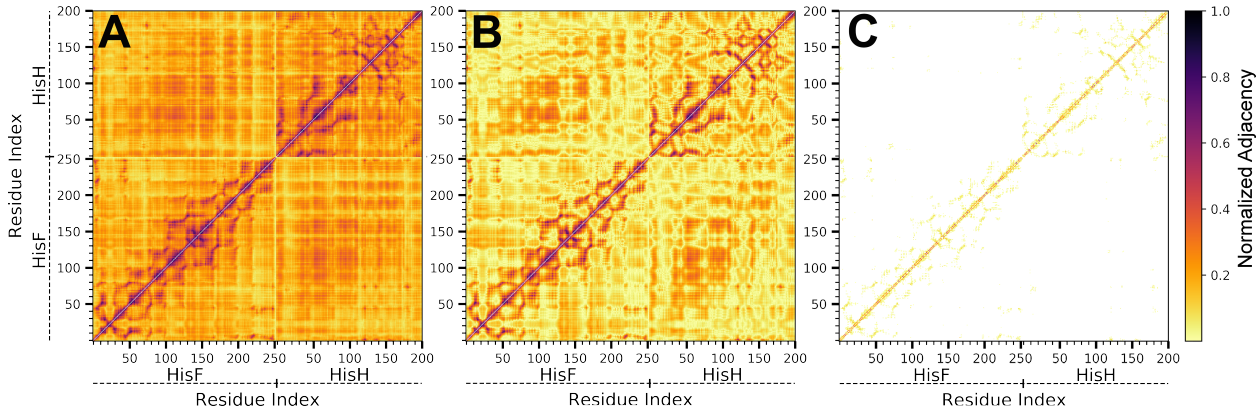

Figure 2: Normalized adjacency matrices computed from 4×250 ns all-atom molecular dynamics simulations of IGPS protein in its *apo* form. A) The linear mutual information adjacency, rMI,[83] B) the Pearson correlation adjacency and C) the effective harmonic Hessian.

connectivity and values near zero indicating weak connections. Residue-level normalized adjacency matrices based on our 1μs MD simulation of *apo* IGPS are shown in Figure 2.

Mutual information (MI) is a measure of correlation between two distributions based on the difference between marginal and joint Shannon entropies. The numeric values of this quantity range from zero to infinity, but can be "normalized" by computing $r_{MI} = \sqrt{1 - \exp(-\frac{2}{3}MI)}$ where this manipulation leads to pairs with large MI having $r_{MI} \approx 1$ and uncoupled pairs having $r_{MI} \approx 0$. We note, however, that neither the MI nor the rMI matrices are positive semi-definite, which differentiates it from the two other matrices discussed here. The motivation for using this "adjacency" matrix over, for example, the normalized 1D covariance matrix, is that it accounts for correlation along perpendicular degrees of freedom.[83] The residue-level r_{MI} matrix, computed from a linear MI, for our *apo* IGPS simulations is depicted in Figure 2A. This plot indicates strong correlations along the primary sequence as indicated by dark colors in the near off-diagonal elements of the matrix. Further off-diagonal dark coloring indicates secondary structural elements that also convey strong correlation. The break between the HisF and HisH proteins is also evident as there are clear dividing lines at residue 253. Despite stronger near diagonal correlation, the r_{MI} adjacency matrix demonstrates long-range contacts even in the HisF–HisH coupling sub-matrix in the bottom-right or top-left of the matrix.

An alternative to r_{MI} is the Pearson correlation matrix. This adjacency matrix is derived from the covariance matrix, which is positive semi-definite. In a 3D system of N particles,

the covariance matrix is a $3N \times 3N$ matrix or, equivalently, an $N \times N$ matrix of 3×3 tensor elements. A typical manipulation is to reduce the $3N \times 3N$ matrix to an $N \times N$ matrix by taking the trace of each tensor element. While this manipulation still yields a positive semi-definite matrix, we do note that it ignores couplings in orthogonal degrees of freedom upon mapping the system to a 1D system.[83] The resulting $N \times N$ matrix can be further manipulated to the Pearson correlation matrix with values between -1 and 1 by dividing by the square root of the diagonal elements of each corresponding row and column. Taking the absolute value of the Pearson correlation for the *apo* IGPS system yields the normalized adjacency matrix depicted in Figure 2B. While there are quantitative differences observed between r_{MI} and the Pearson correlation, the qualitative behavior remains the same; there are strong contacts along the primary sequence but there are finite contacts for almost all elements of the matrix.

The third adjacency matrix we consider is based on the effective harmonic Hessian. We compute this from the $3N \times 3N$ covariance matrix and the average structure from a simulation following the modified hENM procedure as discussed in the Theory section and further elaborated in the SI. The Hessian is a symmetric positive semi-definite matrix by construction, and the normalized adjacency matrix can be constructed in a manner analogous to what is done for the Pearson correlation. The Hessian-based normalized adjacency matrix (Figure 2C) shows strong primary sequence connectivity in agreement with the correlation and r_{MI} adjacency matrices. Secondary structure connectivity is observed in the near off-diagonal of Figure 2C in agreement with the strongly correlated off-diagonal regions of Figure 2A and B. These connections are distinctly weaker in the Hessian-based adjacency, however, and the farther off-diagonal elements are zero indicating little long-range connectivity in the graph. This adjacency matrix is consistent with the idea of short-ranged physical interactions that yield long-range correlation.

Direct Paths Comparison

Paths that convey the covariance between a set of sources and sinks are attractive physical interpretations of allostery. Direct paths, ones with complete connectivity in the adjacency matrix, are an incomplete picture of the covariance between any given source and sink but are

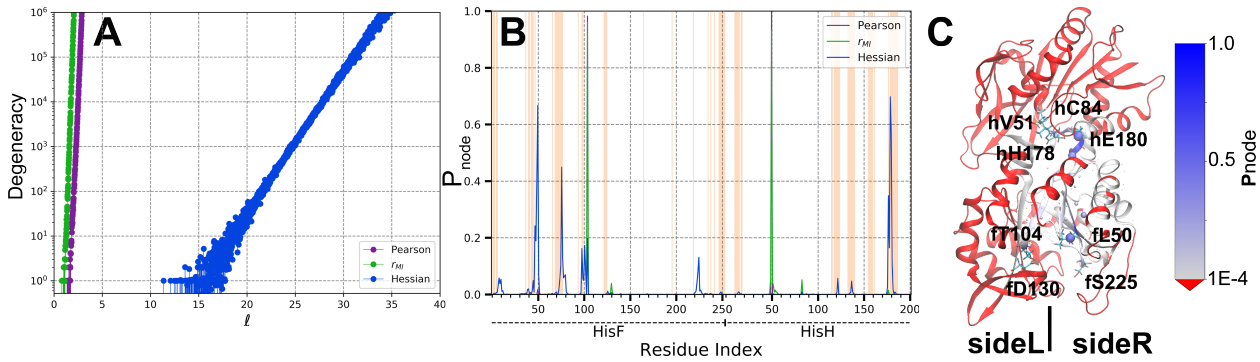

Figure 3: Comparison of paths and resulting centralities for different adjacency matrices for the IGPS protein. A) The path length degeneracy as a function of path length, ℓ. B) The probability of observing a given node in a sampled path, P_{node}. C) The probability of a given node in a sampled path for the Hessian-based adjacency matrix coloring residues in a structural representation of IGPS.

readily sampled. We use direct paths here to compare and contrast adjacency matrices. In this work, we sample paths between four source residues (fLeu50, fThr104, fAsp130, fSer225) that span the effector-binding pocket (green oval in Figure 1) and four sink residues (hVal51, hCys84, hHis178, hGlu180) in the glutaminase active site (pink oval in Figure 1 of IGPS that have been previously identified as important.[36,84,85] Path sampling is performed in a Monte-Carlo scheme as described in the Theoretical Framework section.

Figure 3A depicts the degeneracy of paths as a function of path length, ℓ, for the correlation, r_{MI} and Hessian paths. The paths in the Pearson correlation and r_{MI} adjacency matrices behave similarly; paths of extremely short length are found and the degeneracy grows rapidly with path length. This can be understood by observing finite values in all elements of these two matrices. The paths on the Hessian are longer and the degeneracy of paths grows much less rapidly than either the Pearson or r_{MI} paths. Again, this can be understood by the smaller number of connections in the adjacency matrix leading to both longer paths as well as more unique, short path lengths. These results demonstrate that the paths on the Hessian use a small number of short-ranged interactions to produce long-range correlation.

Paths in the Hessian sample different nodes than paths in the other two adjacency matrices. A residue centrality metric P_{node} can be defined as the probability of observing each residue in the sampled paths. This metric is plotted as a function of residue number in Figure 3B for

each adjacency matrix. The r_{MI} and Pearson adjacency matrices both have high probability of observing the source residue fThr104 in the paths ($P_{node} \sim 0.8$) but little probability of observing the other three source residues (fLeu50, fAsp130, fSer225). Similarly, the sink residue probabilities are dominated by hVal51 ($P_{node} \sim 0.8$) with some contribution from hCys84 ($P_{node} \sim 0.4$). This is due to the highly correlated nature of residues fThr104 and hVal51 yielding a direct path between the two that is much shorter than all other paths. The Hessian-based adjacency matrix, on the other hand, yields significant probability of observing source residue fLeu50 and sink residue hGlu180. There are, however, finite probabilities of observing the other source and sink residues in the Hessian paths.

The Hessian-based paths go through known important regions of IGPS. Figure 3C depicts these P_{node} results on the structure of the heterodimer, highlighting the localization of residues sampled in paths to sideR of HisF as indicated by the white to blue coloring. These paths mainly propagate from fLeu50 through a hydrophobic network that spans the interface of the beta barrel and surrounding alpha helices in HisF. This includes residue fVal48 ($P_{node} = 0.19$; to be discussed later). These results are consistent with previous results indicating the importance of sideR.[36,37]

Interestingly, the Hessian-based paths show significant probability of observing residues fPro76 ($P_{node} = 0.45$) and hLys181 ($P_{node} = 0.58$) as the connection between HisF and HisH. This is in contrast to some previous results indicating the importance of a salt-bridge at the interface between fAsp98 ($P_{node} = 0.16$) and hLys181. We leave further discussion of this to a subsequent section in which we compare to experimental results.

Derivative Centrality Metric of Hessian

Paths are an appealing approach to identify important residues between binding sites, but direct paths are not the only contributors to covariance between source and sink residues. How a mutation will ultimately affect the covariance between nodes via paths is not obvious. Using the Hessian as the adjacency matrix suggests another analysis to compare to mutagenesis experiments. Given the physical interpretation of the finite Hessian elements, we can consider how the covariance between a source and sink is affected by changes in the Hessian elements.

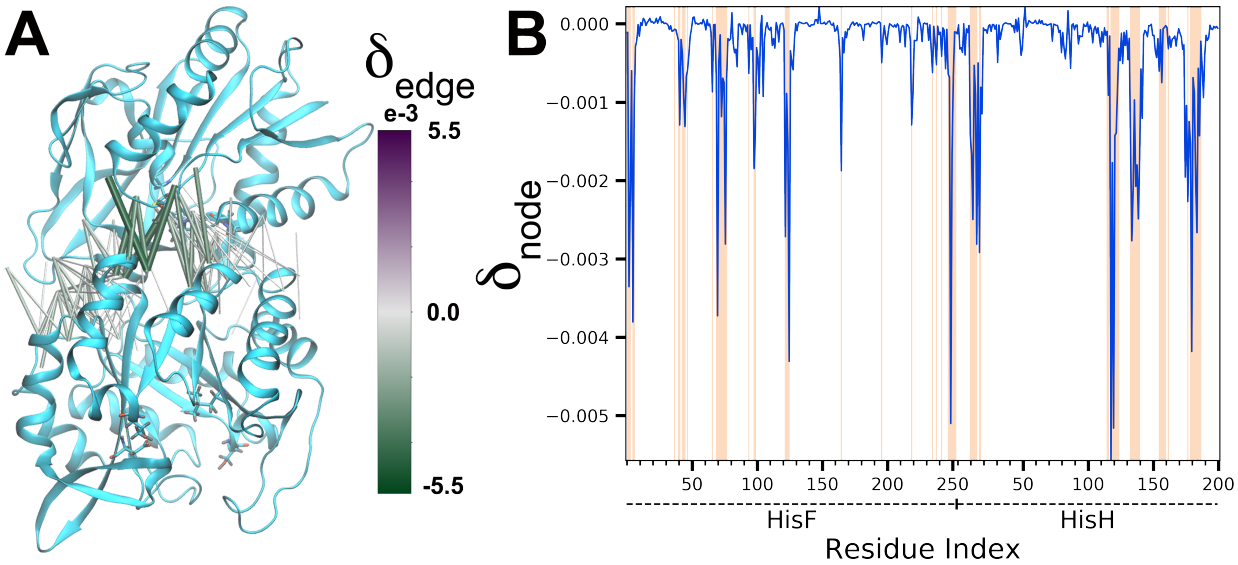

Figure 4: Hessian-based derivative centrality metric δ_{node} of IGPS apo. A) A structural representation edges with large derivative metric values B) Derivative node metric values as a function of residues. Residues that have connections across the HisF–HisH interface are colored by vertical orange lines.

The details of this method, termed the derivative centrality metric, are provided in the Theory section. This type of thinking parallels work from Rocks *et al.* but has not been applied directly to allostery in a protein.[86,87]

After applying a 12 Å distance cut-off to the derivative edge metric, contacts at the HisF–HisH interface are highlighted. A structural representation of the edges with large magnitude derivative values from Equation 5 is provided in Figure 4A. The highest density of large magnitude (green) edges span the heterodimer interface, specifically at a region identified previously by Amaro *et al.* and Rivalta *et al.* to be modulated by PRFAR-binding.[36,81] These past results are interpreted as a strengthening of the spring constants that span the interface in this region, leading to increased frequency of the reported "breathing motion". Results from the derivative centrality metric then suggest that the strengthening of these spring constants impacts the covariance between source and sink residues, thus indicating an indirect allosteric effect on the glutaminase active site due to PRFAR binding.

The derivative node metric (Equation 6) on *apo* IGPS highlights the importance of a second position at the interface between HisF and HisH. These values are plotted as a function of residue number in Figure 4B with interfacial residues highlighted in orange. A cluster of

three residues with large δ_{node} values are hP119 ($\delta_{node} = -5.6 \times 10^{-3}$Å^4), hM121 ($\delta_{node} = -5.2 \times 10^{-3}$Å^4) and fV125 ($\delta_{node} = -4.3 \times 10^{-3}$Å^4). These three residues sit at the interface between HisF and HisH, on SideL of the region discussed above. The interface is closer together at this position and the interfacial contacts in the Hessian matrix are observed to be stronger. The large magnitude δ_{node} values of these three residues suggest that the perturbation of their respective contacts will have a large impact on the covariance between sources and sinks. Due to the limited focus on this region of IGPS, we proffer this cluster of residues as potential targets for mutagenesis or inhibitory binding studies.

The path-based and derivative centrality metrics highlight different residues. This can be observed by noting that bridge residues, highlighted in orange, tend to be underrepresented in Figure 3B and highly represented in Figure 4B. The structural comparison also indicates that P_{node} highlights residues on SideR of the protein and goes somewhat linearly from the PRFAR binding site to the glutaminase active site. On the other hand, the derivative centrality metric highlights only the interfacial residues, which span both sides of the protein. The correlation between these metrics is plotted in Figure S2. As presented in that figure, the majority of residues highlighted in the derivative metric are not observed in the direct paths between sources and sinks. The few exceptions to this are interfacial residues such as hK181 that are both at the interface, as well as in the direct paths. The frequency of such residues is relatively rare, thus we conclude that these two analyses provide different and complementary information.

Comparison to Experimental Results for IGPS

The results from our simulations and centrality metrics on the effective harmonic coarse-grained Hessian compare favorably to kinetic experiments. To assess the mechanistic role of specific residues, we will focus on comparing our results to kinetic assay experiments of wild-type (WT) and mutant IGPS proteins. These assays compute the Michaelis-Menten enzymatic efficiency metric, k_{cat}/K_M, in the activated (effector bound) and basal states of the protein. The ratio of the enzymatic efficiency in these two states indicates the allosteric enhancement factor of the enzyme. The WT enzyme exhibits an allosteric enhancement factor of

Table 1: Residue importance for glutaminase related allostery in IGPS. A comprehensive list of published kinetic assay results for glutaminase activity of single point mutants of IGPS is included as well as our node centrality metrics, P_{node} and δ_{node}, for each of the mutated residues.

Residue	Mutation	Ratio of Activated/Basal k_{cat}/K_M	P_{node}	δ_{node} (10^{-3}Å^4)
fR5	fR5A[†,2]	870	0.0068	-3.8049
	fR5H[†,2]	71		
	fR5K[†,2]	440		
fV12	fV12A[1]	3095	0.0131	-0.0201
fK19	fK19A[1]	92	0.0000	0.0225
	fK19A[†,2]	45		
	fK19R[†,2]	110		
fV48	fV48A[1]	139	0.1891	-0.3954
fD98	fD98A[†,3]	2	0.1620	-1.8466
	fD98A[1]	54		
fK99	fK99A[†,2]	9100	0.0547	-1.4076
	fK99R[†,2]	700		
fT104	fT104A[†,4]	40	0.1973	0.1068
fQ123	fQ123A[†,3]	2000	0.0000	-0.8889
hN12	hN12A[†,3]	100	0.0004	-1.6660
hK181	hK181A[†,3]	2000	0.5797	-4.1806
WT[2,3]		4500 – 4900	–	–

Mutant residues are reported using *T. maritima* IGPS notation. [†] denotes mutations performed in *S. cerevisiae* IGPS. [1] is from Ref. 31, [2] is from Ref. 70, [3] is from Ref. 74 and [4] is from Ref. 81.

4,500 to 4,900[31,70] while most mutant proteins have a smaller enhancement factor relative to WT, indicating reduced allosteric activation by PRFAR (see Table 1).

The mutation of a residue found in an allosteric pathway is hypothesized to have an effect on the allosteric activation of the enzyme. Of the residues studied by mutagenesis (Table 1), we find that only residues hK181, fT104, fV48 and fD98 have a significant (above $\sim$ 6%) probability of being in the direct paths between the PRFAR binding pocket and glutaminase active site. Of these, single point mutation of fT104, fV48 and fD98 to alanines diminish allosteric enhancement factor by over an order of magnitude compared to WT. This is consistent with the picture that altering residues in the allosteric paths disrupts the ability of the enzyme to properly convey covariance between pockets.

Interestingly, a single point alanine mutation to hK181 only causes a factor of two decrease in allosteric enhancement factor when compared to WT. This result calls into question the importance of the salt-bridge between hK181 and fD98 that has been previously implicated as of extreme importance in IGPS allosteric paths. As was mentioned in the previous paths sections, we find that the residues fP76 and hK181 are more sampled in the paths than fD98. Additionally, in the IGPS *apo* Hessian, the fP76-hK181 and fD98-hK181 spring constants are 7.322 and 0.411 kcal mol^{-1}Å^{-2}, respectively. The interaction between fP76 and the aliphatic portion of hK181's side chain is stronger than the salt bridge observed between fD98-hK181. We hypothesized that these trends in interactions would be well maintained in an hK181A mutant. To test this, we performed a simulation of the mutant and found a comparable force constant for the fP76-hA181 edge (2.844 kcal mol^{-1}Å^{-2}) and no force constant between fD98-hA181. Therefore, the alanine mutation of hK181 constitutes a small perturbation to the direct allosteric paths, which explains the small decrease in observed allosteric enhancement factor.

Residues that are at the HisF–HisH interface but not in allosteric paths can also have a large effect on the covariance between sources and sinks as demonstrated by our derivative metric. In Table 1, these consist of residues fR5, fK99, fQ123 and hN12. The δ_{node} values for these residues are -3.8049×10^{-4} Å^4, -1.4076×10^{-4} Å^4, -0.8889×10^{-4} Å^4 and -1.6660×10^{-4} Å^4 respectively. Experimentally, mutations to these residues are all shown to decrease allosteric enhancement relative to WT. This suggests that the mutations have changed the interaction

network around the mutated residue, which, in turn, caused a change to the covariance between PRFAR binding pocket residues and glutaminase active site residues. Interesting, mutation to fK99 can either have a decrease (fK99R) or an increase (fK99A) in activity relative to WT. The derivative metric does not indicate how the change in covariance affects allosteric enhancement just that it will change. It will thus be of interest to study fK99 further to investigate how these two mutants affect the covariance.

If the path and derivative centrality metrics provide a complete picture of allostery in IGPS, then residues not in paths and not at the HisF–HisH interface will have little effect on the experimentally measured allosteric enhancement factor. Residue fV12 is observed in only 1% of paths and has a δ_{node} of only -0.0201×10^{-4} Å^4. A valine to alanine mutation at this position has little affect on the allosteric enhancement factor relatively to WT, suggesting that the P_{node} and δ_{node} results do provide a complete picture for this residue. In contrast, fK19 has $P_{node} = 0$ and $\delta_{node} = 0.0225 \times 10^{-4}$ Å^4 yet all three mutations listed in Table 1 have a large effect on allosteric enhancement factor. Amaro *et al.* performed aaMD simulations to model the unbinding of the PRFAR ligand from which fK19 was observed to play an important role in the recognition and binding of the effector molecule.[81] Thus, a mutation to the fK19 residue is likely to decrease the allosteric enhancement factor by simply reducing the binding affinity of PRFAR and this behavior would not perturb the allosteric pathways.

Conclusions

In this work, we provided evidence that the effective coarse-grained Hessian is the appropriate graph Laplacian to consider in the context of allostery. The Hessian only contains finite values for short-ranged physical interactions and can be rigorously tied to the covariance for harmonic systems. We use a previously developed coarse-graining protocol, hENM, to compute the best effective harmonic Hessian that captures the residue-level covariance of all-atom molecular dynamics systems.

With the Hessian as the graph Laplacian, we develop two centrality metrics to highlight important residues that contribute to allostery. Both of these metrics are applied to the IGPS

protein to investigate interactions between the PRFAR binding pocket and the glutaminase active site. The first of these metrics is based on direct path sampling and recapitulates the known importance of sideR in the allostery of IGPS. Paths in the Hessian-based adjacency matrix are found to be significantly longer and to be less diluted than paths found on two previously used adjacency matrices.

The second centrality metric we develop is based on the derivative of the covariance as a function of a given Hessian element. This metric is motivated by mutagenesis experiments in which one perturbs the interactions around a mutation site as compared to wildtype. This metric identifies residues at the interface between HisF and HisH as being important for the allosteric network between the PRFAR binding pocket and the glutaminase active site. The correlation between the derivative and path centrality metrics is found to be minimal, suggesting that these two metrics provide different yet important information about the covariance between the two pockets.

Results from the path and derivative centrality metrics on the effective Hessian corroborate and functionally explain previous mutagenesis experiments. Interestingly, the combination of experimental and simulation results suggest that fP76-hK181 is a more important HisF–HisH interfacial connection for allosteric paths than the previously implicated fD98-hK181. Additionally, we propose a novel sight for targeted mutational or inhibitory binding studies based on the results obtained from our two centrality metrics.

References

(1) Gunasekaran, K.; Ma, B.; Nussinov, R. Is allostery an intrinsic property of all dynamic proteins? *Proteins: Struct., Funct., Bioinf.* **2004**, *57*, 433–443.

(2) Canals, M.; Sexton, P. M.; Christopoulos, A. Allostery in GPCRs: 'MWC' revisited. *Trends Biochem. Sci.* **2011**, *36*, 663–672.

(3) Kelley, J. A.; Knight, K. L. Allosteric Regulation of RecA Protein Function Is Mediated by Gln194. *J. Biol. Chem.* **1997**, *272*, 25778–25782.

(4) Vale, R. D.; Milligan, R. A. The way things move: looking under the hood of molecular motor proteins. *Science* **2000**, *288*, 88–95.

(5) Hersch, G. L.; Burton, R. E.; Bolon, D. N.; Baker, T. A.; Sauer, R. T. Asymmetric interactions of ATP with the AAA+ ClpX6 unfoldase: allosteric control of a protein machine. *Cell* **2005**, *121*, 1017–1027.

(6) Davidson, R. B.; Hendrix, J.; Geiss, B. J.; McCullagh, M. Allostery in the dengue virus NS3 helicase: Insights into the NTPase cycle from molecular simulations. *PLoS Comput. Biol.* **2018**, *14*, e1006103–28.

(7) Bohr, C.; Hasselbalch, K.; Krogh, A. Ueber einen in biologischer Beziehung wichtigen Einfluss, den die Kohlensäurespannung des Blutes auf dessen Sauerstoffbindung übt1. *Skandinavisches Archiv Für Physiologie* **1904**, *16*, 402–412.

(8) Pauling, L. The Oxygen Equilibrium of Hemoglobin and Its Structural Interpretation. *Proc. Natl. Acad. Sci. U.S.A.* **1935**, *21*, 186–191.

(9) Cui, Q.; Karplus, M. Allostery and cooperativity revisited. *Protein Sci.* **2008**, *17*, 1295–1307.

(10) Motlagh, H. N.; Wrabl, J. O.; Li, J.; Hilser, V. J. The ensemble nature of allostery. *Nature* **2014**, *508*, 331–339.

(11) Nussinov, R. Introduction to Protein Ensembles and Allostery. *Chem. Rev.* **2016**, *116*, 6263–6266.

(12) Liu, J.; Nussinov, R. Allostery: An Overview of Its History, Concepts, Methods, and Applications. *PLoS Comput. Biol.* **2016**, *12*, e1004966.

(13) Nussinov, R.; Tsai, C.-J. Allostery in Disease and in Drug Discovery. *Cell* **2013**, *153*, 293–305.

(14) Lu, S.; Li, S.; Zhang, J. Harnessing Allostery: A Novel Approach to Drug Discovery. *Med. Res. Rev.* **2014**, *34*, 1242–1285.

(15) Cortina, G. A.; Kasson, P. M. Predicting allostery and microbial drug resistance with molecular simulations. *Curr. Opin. Struct. Biol.* **2018**, *52*, 80–86.

(16) Tautermann, C. S.; Binder, F.; Büttner, F. H.; Eickmeier, C.; Fiegen, D.; Gross, U.; Grundl, M. A.; Heilker, R.; Hobson, S.; Hoerer, S.; Luippold, A.; Mack, V.; Montel, F.; Peters, S.; Bhattacharya, S.; Vaidehi, N.; Schnapp, G.; Thamm, S.; Zeeb, M. Allosteric Activation of Striatal-Enriched Protein Tyrosine Phosphatase (STEP, PTPN5) by a Fragment-like Molecule. *J. Med. Chem.* **2019**, *62*, 306–316.

(17) An, X.; Lu, S.; Song, K.; Shen, Q.; Huang, M.; Yao, X.; Liu, H.; Zhang, J. Are the Apo Proteins Suitable for the Rational Discovery of Allosteric Drugs? *J. Chem. Inf. Model.* **2019**, *59*, 597–604.

(18) Buller, A. R.; van Roye, P.; Cahn, J. K. B.; Scheele, R. A.; Herger, M.; Arnold, F. H. Directed Evolution Mimics Allosteric Activation by Stepwise Tuning of the Conformational Ensemble. *J. Am. Chem. Soc.* **2018**, *140*, 7256–7266.

(19) Koshland, D. E.; Némethy, G.; Filmer, D. Comparison of experimental binding data and theoretical models in proteins containing subunits. *Biochemistry* **1966**, *5*, 365–385.

(20) Konermann, L. Heavy lessons in protein allostery. *Nat. Struct. Mol. Biol.* **2016**, *23*, 511–512.

(21) Berry, L.; Poudel, S.; Tokmina-Lukaszewska, M.; Colman, D. R.; Nguyen, D. M. N.; Schut, G. J.; Adams, M. W. W.; Peters, J. W.; Boyd, E. S.; Bothner, B. H/D exchange mass spectrometry and statistical coupling analysis reveal a role for allostery in a ferredoxin-dependent bifurcating transhydrogenase catalytic cycle. *Biochim. Biophys. Acta Gen. Subj.* **2018**, *1862*, 9–17.

(22) Daily, M. D.; Gray, J. J. Local motions in a benchmark of allosteric proteins. *Proteins: Struct., Funct., Bioinf.* **2007**, *67*, 385–399.

(23) Fan, G.; Baker, M. R.; Wang, Z.; Seryshev, A. B.; Ludtke, S. J.; Baker, M. L.; Serysheva, I. I. Cryo-EM reveals ligand induced allostery underlying InsP3R channel gating. *Cell Res.* **2018**, *28*, 1158–1170.

(24) Walma, T.; Spronk, C. A. E. M.; Tessari, M.; Aelen, J.; Schepens, J.; Hendriks, W.; Vuister, G. W. Structure, dynamics and binding characteristics of the second PDZ domain of PTP-BL. *J. Mol. Biol.* **2002**, *316*, 1101–1110.

(25) Fuentes, E. J.; Gilmore, S. A.; Mauldin, R. V.; Lee, A. L. Evaluation of energetic and dynamic coupling networks in a PDZ domain protein. *J. Mol. Biol.* **2006**, *364*, 337–351.

(26) Gianni, S.; Walma, T.; Arcovito, A.; Calosci, N.; Bellelli, A.; Engström, A.; Travaglini-Allocatelli, C.; Brunori, M.; Jemth, P.; Vuister, G. W. Demonstration of long-range interactions in a PDZ domain by NMR, kinetics, and protein engineering. *Structure* **2006**, *14*, 1801–1809.

(27) Manley, G.; Rivalta, I.; Loria, J. P. Solution NMR and computational methods for understanding protein allostery. *J. Phys. Chem. B* **2013**, *117*, 3063–3073.

(28) Grutsch, S.; Brüschweiler, S.; Tollinger, M. NMR Methods to Study Dynamic Allostery. *PLoS Comput. Biol.* **2016**, *12*, e1004620.

(29) O'Rourke, K. F.; Gorman, S. D.; Boehr, D. D. Biophysical and computational methods to analyze amino acid interaction networks in proteins. *Comput. Struct. Biotechnol. J.* **2016**, *14*, 245–251.

(30) Rivalta, I.; Lisi, G. P.; Snoeberger, N.-S.; Manley, G.; Loria, J. P.; Batista, V. S. Allosteric Communication Disrupted by a Small Molecule Binding to the Imidazole Glycerol Phosphate Synthase Protein-Protein Interface. *Biochemistry* **2016**, *55*, 6484–6494.

(31) Lisi, G. P.; East, K. W.; Batista, V. S.; Loria, J. P. Altering the allosteric pathway in IGPS suppresses millisecond motions and catalytic activity. *Proc. Natl. Acad. Sci. U.S.A.* **2017**, *114*, E3414–E3423.

(32) Xu, Y.; Zhang, D.; Rogawski, R.; Nimigean, C. M.; McDermott, A. E. Identifying coupled clusters of allostery participants through chemical shift perturbations. *Proc. Natl. Acad. Sci. U.S.A.* **2019**, *116*, 2078–2085.

(33) Perutz, M. F. Stereochemistry of cooperative effects in haemoglobin. *Nature* **1970**, *228*, 726–739.

(34) Perutz, M. F.; Wilkinson, A. J.; Paoli, M.; Dodson, G. G. The stereochemical mechanism of the cooperative effects in hemoglobin revisited. *Annu. Rev. Biophys. Biomol. Struct.* **1998**, *27*, 1–34.

(35) Süel, G. M.; Lockless, S. W.; Wall, M. A.; Ranganathan, R. Evolutionarily conserved networks of residues mediate allosteric communication in proteins. *Nat. Struct. Biol.* **2003**, *10*, 59–69.

(36) Rivalta, I.; Sultan, M. M.; Lee, N.-S.; Manley, G. A.; Loria, J. P.; Batista, V. S. Allosteric pathways in imidazole glycerol phosphate synthase. *Proc. Natl. Acad. Sci. U.S.A.* **2012**, *109*, E1428–36.

(37) Negre, C.; Morzan, U. N.; Hendrickson, H. P.; Pal, R.; Lisi, G. P.; Loria, J. P.; Rivalta, I.; Ho, J.; Batista, V. S. Eigenvector centrality for characterization of protein allosteric pathways. *Proc. Natl. Acad. Sci. U.S.A.* **2018**, *115*, E11201–E11208.

(38) Dokholyan, N. V. Controlling Allosteric Networks in Proteins. *Chem. Rev.* **2016**, *116*, 6463–6487.

(39) Schueler-Furman, O.; Wodak, S. J. Computational approaches to investigating allostery. *Curr. Opin. Struct. Biol.* **2016**, *41*, 159–171.

(40) Stolzenberg, S.; Michino, M.; LeVine, M. V.; Weinstein, H.; Shi, L. Computational approaches to detect allosteric pathways in transmembrane molecular machines. *Biochim. Biophys. Acta* **2016**, *1858*, 1652–1662.

(41) Dijkstra, E. W. A note on two problems in connexion with graphs. *Numer. Math.* **1959**, *1*, 269–271.

(42) Floyd, R. W. Algorithm 97: Shortest path. *Communications of the ACM* **1962**, *5*, 345.

(43) del Sol, A.; Fujihashi, H.; Amoros, D.; Nussinov, R. Residue centrality, functionally important residues, and active site shape: analysis of enzyme and non-enzyme families. *Protein Sci.* **2006**, *15*, 2120–2128.

(44) Lockless, S. W.; Ranganathan, R. Evolutionarily conserved pathways of energetic connectivity in protein families. *Science* **1999**, *286*, 295–299.

(45) Kaya, C.; Armutlulu, A.; Ekesan, S.; Haliloglu, T. MCPath: Monte Carlo path generation approach to predict likely allosteric pathways and functional residues. *Nucleic Acids Res.* **2013**, *41*, W249–55.

(46) Karplus, M.; Petsko, G. A. Molecular dynamics simulations in biology. *Nature* **1990**, *347*, 631–639.

(47) van Gunsteren, W. F. Molecular dynamics studies of proteins. *Curr. Opin. Struct. Biol.* **1993**, *3*, 277–281.

(48) Karplus, M.; McCammon, J. A. Molecular dynamics simulations of biomolecules. *Nat. Struct. Biol.* **2002**, *9*, 646–652.

(49) Karplus, M. Molecular Dynamics of Biological Macromolecules: A Brief History and Perspective. *Biopolymers* **2003**, *68*, 350–358.

(50) Karplus, M.; Kuriyan, J. Molecular dynamics and protein function. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 6679–6685.

(51) Perilla, J. R.; Goh, B. C.; Cassidy, C. K.; Liu, B.; Bernardi, R. C.; Rudack, T.; Yu, H.; Wu, Z.; Schulten, K. ScienceDirect Molecular dynamics simulations of large macromolecular complexes. *Curr. Opin. Struct. Biol.* **2015**, *31*, 64–74.

(52) Kong, Y.; Karplus, M. Signaling pathways of PDZ2 domain: a molecular dynamics interaction correlation analysis. *Proteins: Struct., Funct., Bioinf.* **2009**, *74*, 145–154.

(53) Vijayabaskar, M. S.; Vishveshwara, S. Interaction energy based protein structure networks. *Biophys. J.* **2010**, *99*, 3704–3715.

(54) Ribeiro, A. A.; Ortiz, V. Determination of signaling pathways in proteins through network theory: Importance of the topology. *J. Chem. Theory Comput.* **2014**, *10*, 1762–1769.

(55) Van Wart, A. T.; Durrant, J.; Votapka, L.; Amaro, R. E. Weighted Implementation of Suboptimal Paths (WISP): An Optimized Algorithm and Tool for Dynamical Network Analysis. *J. Chem. Theory Comput.* **2014**, *10*, 511–517.

(56) Proctor, E. A.; Kota, P.; Aleksandrov, A. A.; He, L.; Riordan, J. R.; Dokholyan, N. V. Rational Coupled Dynamics Network Manipulation Rescues Disease-Relevant Mutant Cystic Fibrosis Transmembrane Conductance Regulator. *Chem. Sci.* **2015**, *6*, 1237–1246.

(57) Moritsugu, K.; Smith, J. C. Coarse-grained biomolecular simulation with REACH: realistic extension algorithm via covariance Hessian. *Biophys. J.* **2007**, *93*, 3460–3469.

(58) Lyman, E.; Pfaendtner, J.; Voth, G. A. Systematic multiscale parameterization of heterogeneous elastic network models of proteins. *Biophys. J.* **2008**, *95*, 4183–4192.

(59) Moritsugu, K.; Smith, J. C. REACH: A program for coarse-grained biomolecular simulation. *Comput. Phys. Commun.* **2009**,

(60) Ikeguchi, M.; Ueno, J.; Sato, M.; Kidera, A. Protein Structural Change Upon Ligand Binding: Linear Response Theory. *Phys. Rev. Lett.* **2005**, *94*, 478–4.

(61) Besag, J. Spatial Interaction and the Statistical Analysis of Lattice Systems. *Journal of the Royal Statistical Society: Series B (Methodological)* **1974**, *36*, 192–225.

(62) Friedman, J.; Hastie, T.; Tibshirani, R. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* **2008**, *9*, 432–441.

(63) Brooks, B.; Karplus, M. Harmonic dynamics of proteins: Normal modes and fluctuations in bovinepancreatictrypsin. *Proc. Natl. Acad. Sci. USA* **1983**, *80*, 6571–6575.

(64) Atilgan, A. R.; Durell, S. R.; Jernigan, R. L.; Demirel, M. C.; Keskin, O.; Bahar, I. Anisotropy of Fluctuation Dynamics of Proteins with an Elastic Network Model. *Biophys. J.* **2001**, *80*, 505–515.

(65) Zheng, W.; Doniach, S. A comparative study of motor-protein motions by using a simple elastic-network model. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 13253–13258.

(66) Leioatts, N.; Romo, T. D.; Grossfield, A. Elastic Network Models Are Robust to Variations in Formalism. *J. Chem. Theory Comput.* **2012**, *8*, 2424–2434.

(67) Bastolla, U. Computing protein dynamics from protein structure with elastic network models. *WIREs Comput Mol Sci* **2014**, *4*, 488–503.

(68) Chaudhuri, B. N.; Lange, S. C.; Myers, R. S.; Chittur, S. V.; Davisson, V.; Smith, J. L. Crystal Structure of Imidazole Glycerol Phosphate Synthase: A Tunnel through a $(\beta/\alpha)8$ Barrel Joins Two Active Sites. *Structure* **2001**, *9*, 987–997.

(69) Douangamath, A.; Walker, M.; Beismann-Driemeyer, S.; Vega-Fernandez, M. C.; Sterner, R.; Wilmanns, M. Structural evidence for ammonia tunneling across the $(\beta\alpha)8$barrel of the imidazole glycerol phosphate synthase bienzyme complex. *Structure* **2002**, *10*, 185–193.

(70) Myers, R. S.; Jensen, J. R.; Deras, I. L.; Smith, J. L.; Davisson, V. J. Substrate-induced changes in the ammonia channel for imidazole glycerol phosphate synthase. *Biochemistry* **2003**, *42*, 7013–7022.

(71) Lisi, G. P. P.; Manley, G. A. A.; Hendrickson, H.; Rivalta, I.; Batista, V. S. S.; Loria, J. P. Dissecting Dynamic Allosteric Pathways Using Chemically Related Small-Molecule Activators. *Structure* **2016**, *24*, 1155–1166.

(72) Klem, T. J.; Chen, Y.; Davisson, V. J. Subunit interactions and glutamine utilization by Escherichia coli imidazole glycerol phosphate synthase. *J. Bacteriol.* **2001**, *183*, 989–996.

(73) Amaro, R. E.; Myers, R. S.; Davisson, V. J.; Luthey-Schulten, Z. A. Structural elements in IGP synthase exclude water to optimize ammonia transfer. *Biophysical Journal* **2005**, *89*, 475–487.

(74) Myers, R. S.; Amaro, R. E.; Luthey-Schulten, Z. A.; Davisson, V. J. Reaction coupling through interdomain contacts in imidazole glycerol phosphate synthase. *Biochemistry* **2005**, *44*, 11974–11985.

(75) Lipchock, J. M.; Loria, J. P. 1H, 13C, and 15N resonance assignment of imidazole glycerol phosphate (IGP) synthase protein HisF from Thermotoga maritima. *Biomol. NMR Assign.* **2008**, *2*, 219–221.

(76) Lipchock, J.; Loria, J. P. Millisecond dynamics in the allosteric enzyme imidazole glycerol phosphate synthase (IGPS) from Thermotoga maritima. *J. Biomol. NMR* **2009**, *45*, 73–84.

(77) Lipchock, J. M.; Loria, J. P. Nanometer propagation of millisecond motions in V-type allostery. *Structure* **2010**, *18*, 1596–1607.

(78) Lisi, G. P.; Currier, A. A.; Loria, J. P. Glutamine Hydrolysis by Imidazole Glycerol Phosphate Synthase Displays Temperature Dependent Allosteric Activation. *Front. Mol. Biosci.* **2018**, *5*, 4.

(79) Amaro, R.; Tajkhorshid, E.; Luthey-Schulten, Z. Developing an energy landscape for the novel function of a (β/α)8 barrel: Ammonia conduction through HisF. *Proceedings of the National Academy of Sciences* **2003**, *100*, 7599–7604.

(80) Amaro, R.; Luthey-Schulten, Z. Molecular dynamics simulations of substrate channeling through an $\alpha\beta$ barrel protein. *Chem. Phys.* **2004**, *307*, 147–155.

(81) Amaro, R. E.; Sethi, A.; Myers, R. S.; Davisson, V. J.; Luthey-Schulten, Z. A. A Network of Conserved Interactions Regulates the Allosteric Signal in a Glutamine Amidotransferase. *Biochemistry* **2007**, *46*, 2156–2173.

(82) VanWart, A. T.; Eargle, J.; Luthey-Schulten, Z.; Amaro, R. E. Exploring Residue Component Contributions to Dynamical Network Models of Allostery. *J. Chem. Theory Comput.* **2012**, *8*, 2949–2961.

(83) Lange, O. F.; Grubmüller, H. Generalized correlation for biomolecular dynamics. *Proteins: Struct., Funct., Bioinf.* **2006**, *62*, 1053–1061.

(84) Tesmer, J. J. G.; Klem, T. J.; Deras, M. L.; Davisson, V. J.; Smith, J. L. The crystal structure of GMP synthetase reveals a novel catalytic triad and is a structural paradigm for two enzyme families. *Nature Structural Biology* **1996**, *3*, 74–86.

(85) Zalkin, H.; Smith, J. L. Enzymes utilizing glutamine as an amide donor. *Advances in Enzymology and Related Areas of Molecular Biology* **1998**, *72*, 87–144.

(86) Rocks, J. W.; Pashine, N.; Bischofberger, I.; Goodrich, C. P.; Liu, A. J.; Nagel, S. R. Designing allostery-inspired response in mechanical networks. *Proc. Natl. Acad. Sci. U.S.A.* **2017**, *114*, 2520–2525.

(87) Rocks, J. W.; Ronellenfitsch, H.; Liu, A. J.; Nagel, S. R.; Katifori, E. Limits of multifunctionality in tunable networks. *Proc. Natl. Acad. Sci. U.S.A.* **2019**, *116*, 2506–2511.

Supporting Information for: The Effective Coarse-grained Hessian Dictates Allostery in IGPS

Peter T. Lake,[†] Russell B. Davidson,[†] Heidi Klem,[†] Glen M. Hocky,[‡] and Martin McCullagh[*,†]

†*Department of Chemistry, Colorado State University, Fort Collins, CO, USA*

‡*Department of Chemistry, New York University, New York, NY, USA*

E-mail: martin.mccullagh@colostate.edu

Computational Methods

Heterogeneous elastic network models

Here we will closely follow the procedure known as heterogenous ENM (hENM) to generate an effective harmonic Hessian that best reproduces the aaMD measured covariance.[1] Our adapted hENM algorithm is as follows:

1. Initiate a trial $N \times N$ force constant matrix, $\boldsymbol{k}$.

2. Create the $3N \times 3N$ Hessian matrix, $\boldsymbol{H}$ using the tensor matrix elements $\boldsymbol{H}_{ij} = -k_{ij} \cdot \hat{r}_{ij} \otimes \hat{r}_{ij}$ for $i \neq j$ and $\boldsymbol{H}_{ii} = \sum_{j \neq i} \boldsymbol{H}_{ij}$

3. Get the predicted $3N \times 3N$ covariance, $\boldsymbol{C}$, by performing the Moore-Penrose psuedo inverse of $\boldsymbol{H}$, $\boldsymbol{C} = T\boldsymbol{H}^{+}$.

4. Update the force constants based on

$$k_{ij}^{n+1} = k_{ij}^{n} - \alpha \left(\frac{1}{\sigma_{ij}^{n}} - \frac{1}{\sigma_{ij}^{target}} \right) \tag{SI1}$$

where $\sigma_{ij}^{n} = \hat{r}_{ij} \left[\boldsymbol{C}_{ii}^{n} + \boldsymbol{C}_{jj}^{n} - \boldsymbol{C}_{ij}^{n} - \boldsymbol{C}_{ji}^{n} \right] \hat{r}_{ij}$ is the variance of node i and j projected along the separation vector and α is a minimization step size parameter (we found $\alpha = 1 \times 10^{-2}$ to yield stable convergence).

5. Repeat steps 2-4 until a converged Hessian matrix has been produced.

All-Atom MD Simulations

All-atom, explicit solvent MD simulations are performed for the *apo*substrate state of *T. maritima* IGPS (PDB: 1GPW).[2] Specifically, chains C and D are used from this crystal structure. Crystallographic waters are maintained. The active site mutation, fN11D, present in this crystal structure was converted back to wild type.

The simulations are performed using the GPU-enabled AMBER18 software[3] and the ff14SB[4] parameters for proteins atoms. Using tleap, the starting structure is solvated in a TIP3P water box with at least a 12 Å buffer between the protein and periodic images. Sodium and chloride ions are added to neutralize charge and maintain a 0.10 M ionic concentration. Direct nonbonding interactions are calculated up to a 12 Å distance cutoff. The SHAKE algorithm is used to constrain covalent bonds that include hydrogen.[5] The particle-mesh Ewald method[6] is used to account for long-ranged electrostatic interactions. Before simulation began, a two stage minimization was performed: (1) 10,000 steps of conjugate gradient optimization were performed to minimize water positioning (substrate atoms restrained with a 75 kcal mol^{-1} Å^{-2} force constant) and (2) an additional 10,000 minimization steps with no restraints applied. The system was slowly heated from 25 K to 303 K over 1 ns. Additionally, 4 ns of NVT simulation was performed to equilibrate the cubic box volume. Finally, the simulations were run in the NTP ensemble, using the Langevin dynamics thermostat and Monte Carlo barostat to maintain the systems at 303 K and 1 bar. A 2 fs integration

time step is used, with energies and positions written every 5 ps. An initial 500 ns simulation was run, from which the structure at 250 ns was used to initialize three more independent trajectories. These new trajectories were given different random number seeds and were run for an additional 250 ns. We have a total of 1 μs of trajectory to analyze the *apo* state of IGPS.

Theory

Path Sampling

The emphasis of the ensemble of paths makes sampling the paths an apt way of studying them. For this purpose, we present a Markov chain simulation to sample the paths. This work is in contrast to previous studies, namely WISP,[7] that focus on finding the exact shortest paths up to some number. Instead, we statistically sample a distribution of paths. This algorithm is shown to be more efficient than using WISP to accomplish the same result.

The set of objects that are studied using the Markov chain simulation is the set of paths from a predefined source and sink nodes. Each path is then assigned an effective length. For the purpose of this work we consider some adjacency matrix $\boldsymbol{A}$ and functionalize it into a cost matrix $\boldsymbol{\chi}$ in an analogous way as the covariance is related to the Pearson correlation. Namely,

$$\chi_{ij} = -\log\left(\frac{|A_{ij}|}{\sqrt{A_{ii}A_{jj}}}\right) \tag{SI2}$$

In this study, we consider the paths of three adjacency matrices: the covariance, r_{MI}, and the Hessian. The length ℓ of a path is defined as a sum over values in the cost matrix corresponding to the edges in the path. The distribution of the paths that are studied is of the form

$$P_{path} \propto \exp\left[-\ell/\tau\right], \tag{SI3}$$

where the free parameter τ is an effective temperature of the simulation. The Markov chain sim-

ulation we present is not limited to these definitions of path lengths and path probabilities; more complex functions of the weights and the path distributions can readily be used.

A trial move in the Monte Carlo simulation consists of randomly choosing a node in the graph that is neither the source nor sink node from a uniform distribution, and add/remove the node to/from the path if it is currently/not in the present path. Any trial move that destroys the path, as will happen when an edge with a zero in the adjacency matrix is used, the move is rejected. The following values associated with the probability of adding the selected node n between each pair of connected nodes i and j in the path not containing n is then calculated

$$p_{ij} = \exp\left[-\frac{\chi_{in} + \chi_{nj} - \chi_{ij}}{\tau}\right]. \tag{SI4}$$

If the node is not in the current path, the trial move is a path with node n randomly inserted between two connected nodes with the probability proportional to p_{ij}. Detailed balance is achieved by using a Metropolis algorithm whereby the trial path is accepted with a probability P_{Met} of

$$P_{Met} = \min\left[1, \left(\sum_{\langle ij \rangle} p_{ij}\right)^{\pm 1}\right], \tag{SI5}$$

where the upper sign is for when the trial move removes a node from the path and the lower sign is for node addition.

The difference between using the Hessian as the adjacency matrix and the equation we derived relating the Hessian to the covaraince in terms of paths can be mocked in the sampling on a few levels. As a first pass, the Hessian can be treated as a $3N \times 3N$ object where an edge in a graph is described by a 3×3 submatrix. The nature of the Hamiltonian we are using to find spring constants of the system produces a 3×3 submatrix describing the interaction of node i and j of the form $k_{ij}\hat{\boldsymbol{r}}_{ij}\hat{\boldsymbol{r}}_{ij}^{\mathsf{T}}$. The weight w of a path can then be considered to be the product of these spring constants,

treated in the same fashion as above, as

$$
\begin{aligned}
w &= \prod_{\langle i,j\rangle} \frac{k_{ij}}{\sqrt{k_{ii}k_{jj}}} \hat{\boldsymbol{r}}_{ij}\hat{\boldsymbol{r}}_{ij}^{\mathsf{T}} \\
&= \hat{\boldsymbol{r}}_{i_1 i_2}\hat{\boldsymbol{r}}_{i_{M-1} i_M}^{\mathsf{T}} \prod_{\langle i,j,k\rangle} \cos\theta_{ijk} \prod_{\langle i,j\rangle} \frac{k_{ij}}{\sqrt{k_{ii}k_{jj}}},
\end{aligned}
\tag{SI6}
$$

where $\langle i, j, k\rangle$ is the set of all three consecutive nodes in the path. The two products in this equation can be interpreted as the scalar weight of the path and the tensor part is treated separately. Taking the length of the path to be equal to $\ell = -\log w$, these paths can similarly sampled with the following change to the p_{ij} defined above,

$$
p_{ij} = \left(\frac{\cos\theta_{i-1,i,n}\cos\theta_{i,n,j}\cos\theta_{n,j,j+1}}{\cos\theta_{i-1,i,j}\cos\theta_{i,j,j-1}}\right)^{1/\tau} \exp\left[-\frac{\chi_{in}+\chi_{nj}-\chi_{ij}}{\tau}\right]. \tag{SI7}
$$

Notice that this definition of path length is not trivially studied by exhaustive search algorithms such as that proposed by WISP, but it readily studied using this Markov chain simulation framework.

A simulation of the paths consists of N trial moves between writing the path to file, where N is the number of nodes in the graph, and then 10^6 paths are generated to sample the ensemble. The choice of τ is dependent on the desired ensemble of paths to sample. For the present study, τ is chosen such that the first thousand shortest paths are well sampled so that our results can be compared to that of WISP.

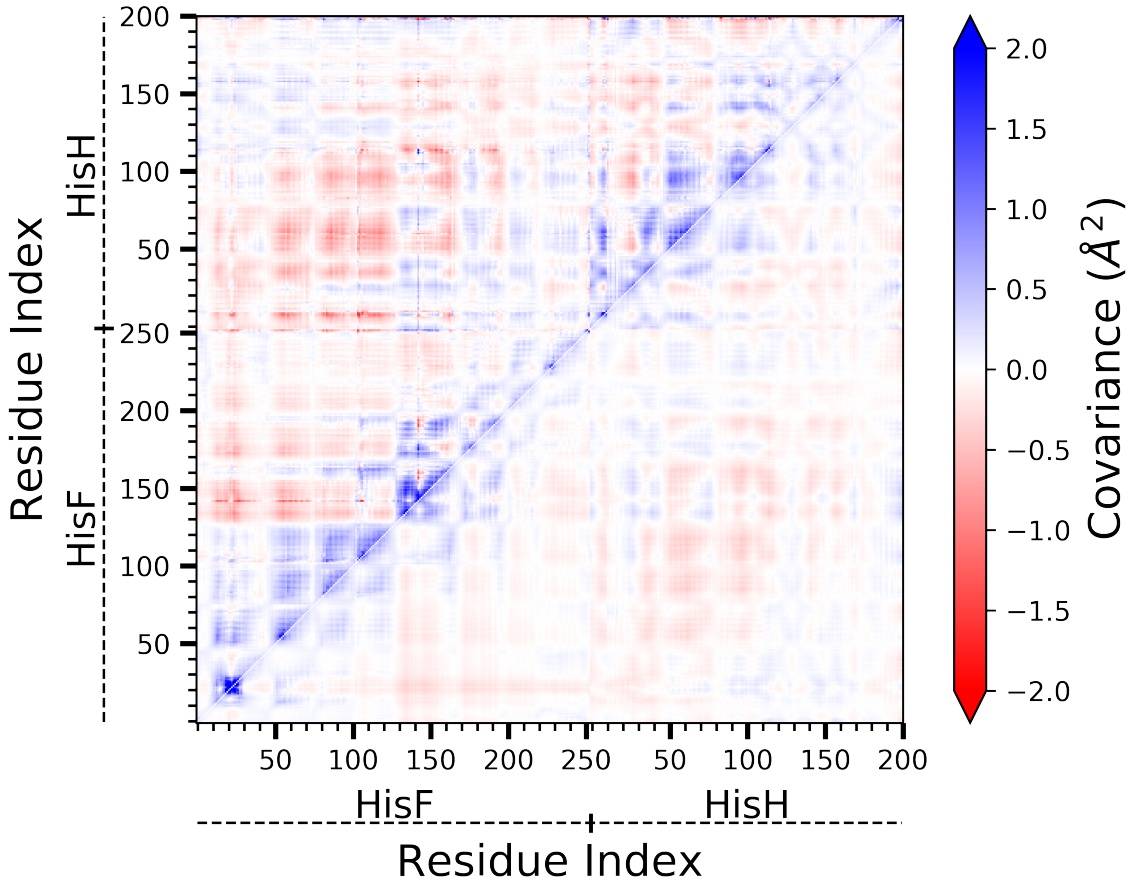

Figure SI1: Covariance matrices computed through the hENM procedure (bottom right) and raw simulation (top left) from 1 μs all-atom molecular dynamics of *apo* IGPS.

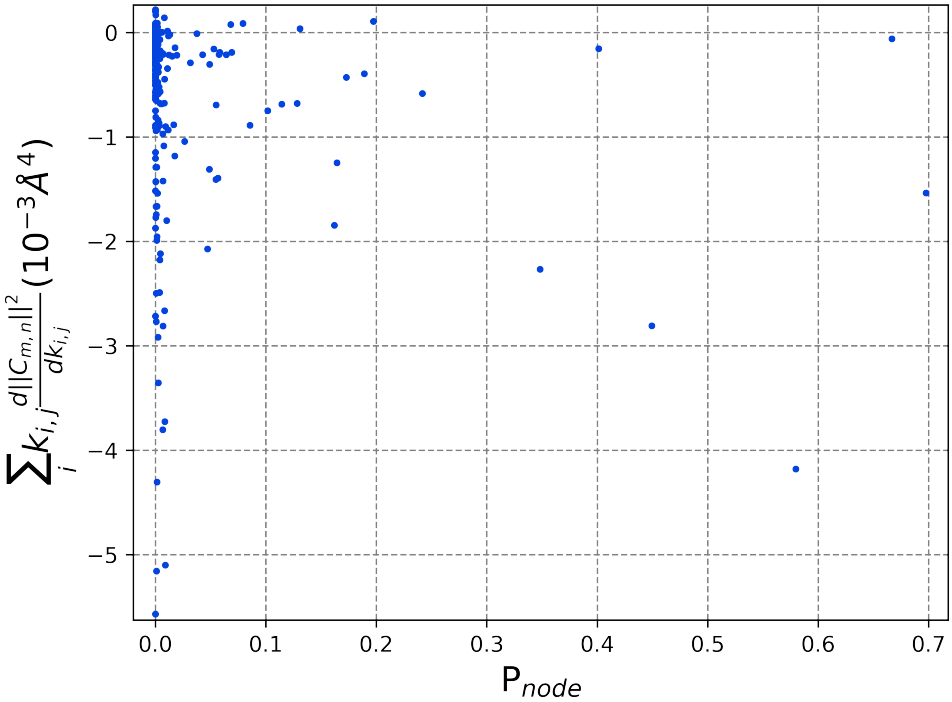

Figure SI2: Correlation of Hessian derivative metric and P_{node}.

References

(1) Lyman, E.; Pfaendtner, J.; Voth, G. A. Systematic multiscale parameterization of heterogeneous elastic network models of proteins. *Biophys. J.* **2008**, *95*, 4183–4192.

(2) Douangamath, A.; Walker, M.; Beismann-Driemeyer, S.; Vega-Fernandez, M. C.; Sterner, R.; Wilmanns, M. Structural evidence for ammonia tunneling across the ($\beta\alpha$)8barrel of the imidazole glycerol phosphate synthase bienzyme complex. *Structure* **2002**, *10*, 185–193.

(3) Case, D. A.; Ben-Shalom, I.; Brozell, S.; Cerutti, D.; Cheatham, T. I.; Cruzeiro, V.; Darden, T.; Duke, R.; Ghoreishi, D.; Gilson, M.; Gohlke, H.; Goetz, A.; Greene, D.; Harris, R.; Homeyer, N.; Izadi, S.; Kovalenko, A.; Kurtzman, T.; Lee, T.; LeGrand, S.; Li, P.; Lin, C.; Liu, J.; Luchko, T.; Luo, R.; Mermelstein, D.; Merz, K.; Miao, Y.; Monard, G.; Nguyen, C.; Nguyen, H.; Omelyan, I.; Onufriev, A.; Pan, F.; Qi, R.; Roe, D.; Roitberg, A.; Sagui, C.; Schott-Verdugo, S.; Shen, J.; Simmerling, C.; Smith, J.; Salomon-Ferrer, R.; Swails, J.;

Walker, R.; Wang, J.; Wei, H.; Wolf, R.; Wu, X.; Xiao, L.; York, D.; Kollman, P. AMBER18. 2018.

(4) Maier, J. A.; Martinez, C.; Kasavajhala, K.; Wickstrom, L.; Hauser, K. E.; Simmerling, C. ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *J. Chem. Theory Comput.* **2015**, *11*, 3696–3713.

(5) Ryckaert, J.-P.; Ciccotti, G.; Berendsen, H. J. Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *J. Comp. Phys.* **1977**, *23*, 327–341.

(6) Cheatham, T. E. I.; Miller, J. L.; Fox, T.; Darden, T. A.; Kollman, P. A. Molecular Dynamics Simulations on Solvated Biomolecular Systems: The Particle Mesh Ewald Method Leads to Stable Trajectories of DNA, RNA, and Proteins. *J. Am. Chem. Soc.* **1995**, *117*, 4193–4194.

(7) Van Wart, A. T.; Durrant, J.; Votapka, L.; Amaro, R. E. Weighted Implementation of Suboptimal Paths (WISP): An Optimized Algorithm and Tool for Dynamical Network Analysis. *J. Chem. Theory Comput.* **2014**, *10*, 511–517.