Unveiling Conserved Allosteric Hot Spots in Protein Domains from Sequences

Aysima Hacisuleyman[1*] and Dirk Fasshauer[1]

[1]Department of Computational Biology, University of Lausanne, CH-1015 Lausanne, Switzerland

[*] Corresponding author: aisima.chatzisouleiman@unil.ch

Keywords:

allosteric mechanisms, entropy transfer, DHFR, PDZ, SH3, S100, pattern matching, dGNM.

Abbreviations:

Gaussian network model (GNM), molecular dynamics (MD), multiple sequence alignment (MSA), hidden Markov models (HMM)

Abstract

The amino acid sequence determines the structure, function, and dynamics of a protein. In recent years, enormous progress has been made in translating sequence information into 3D structural information using artificial intelligence. However, because of the underlying methodology, it is an immense computational challenge to extract this information from the ever-increasing number of sequences. In the present study, we show that it is possible to create 2D contact maps from sequences, for which only a few exemplary structures are available on a laptop without the need for GPUs. This is achieved by using a pattern-matching approach. The resulting contact maps largely reflect the interactions in the 3D structures. This approach was used to explore the evolutionarily conserved allosteric mechanisms and identify the source–sink (driver-driven) relationships by using an established method that combines Schreiber's concept of entropy transfer with a simple Gaussian network model. The validity of our method was tested on the DHFR, PDZ, SH3, and S100 domains, with our predictions consistently aligning with the experimental findings.

Introduction

Proteins play a crucial role in cellular functions, often working in collaboration with other molecules to carry out tasks, and they are subjected to certain regulatory mechanisms, including allostery. Allostery refers to the regulation of a protein's function by the binding of a regulatory molecule at a site other than the protein's active site. The first allosterically regulated enzymes were discovered and characterized about half a century ago. It was discovered that allosteric binding can cause a change in conformation. The concept of a two-state conformational change was later extended, as reviewed by Kalodimos and Edelstein [1, 2], so it is now assumed that allostery is best described as an ensemble of dynamic states.

Allosteric activation serves as a robust switch, toggling between an inactive and an active state. Typically, for allosteric mechanisms, multiple lower energy regions exist within a free energy landscape [3]. Mutations often disrupt this equilibrium, diverting the course of an allosteric communication pathway [4]. Moreover, allosteric sites often serve as binding sites for drugs and are therefore of high pharmacological interest. The concept of "dynamic allostery," as proposed by Cooper and Dryden, explains how allosteric communication pathways revolve around a mean conformational state, involving correlated thermal fluctuations and highlighting the entropic nature of allostery [5].

Dynamic allostery is a fundamental concept in the regulation of proteins' function, where proteins undergo conformational changes at remote sites to respond to environmental cues. These remote interactions are essential for fine-tuning the proteins' activity. Examples include hemoglobin's ability to switch between states of binding and releasing oxygen [6], allosteric enzymes such as aspartate carbamoyltransferase adjusting their conformation depending on the availability of specific molecules [7], G protein-coupled receptors transmitting signals across cell membranes through long-range conformational changes [8], and cyclic nucleotide-gated channels in neurons undergoing conformational changes to transduce chemical signals into electrical signals [9]. These examples highlight the ever-changing and adaptive nature of proteins, showcasing how dynamic allostery is central to their biological functions by involving conformational fluctuations driven by thermal motion and entropic effects.

A tool used to explore conformational fluctuations in proteins arose from the concept of transfer entropy, which was introduced by Schreiber. This is a measure of the flow of

information between two time series [10]. It has been useful for identifying the causal inferences of correlated thermal fluctuations in proteins from the trajectories of molecular dynamics simulation [11–15] or those produced by coarse-grained approaches such as elastic network models, specifically the Gaussian network model (GNM) [13, 16–19]. It has also been applied in a broad range of fields to estimate the causal relationships in finance, neuroscience, social networks, and climate science. The GNM provides a simplified yet effective way to model the dynamics of protein structures [20–23]. It is the simplest elastic network model focusing on the inter-residue contact topology of the proteins and capturing the collective motions and fluctuations with significantly reduced demand for computer processing. We recently formulated a fast and efficient method for calculating the entropy transferred between pairs of residues by combining Schreiber's concept of transfer entropy with the dynamic version of the GNM (dGNM) [18].

Finding allosteric sites in a protein usually requires time-consuming biochemical, biophysical, and structural experiments. In recent decades, computer simulations of protein dynamics have played an increasingly important role. So far, only the dynamics of existing structures have been investigated using computer simulations. Such studies have been carried out selectively, and it has remained unclear to what extent an allosteric mechanism has been conserved in a given protein family. Compared with the enormous number of protein sequences available in Uniprot [24], the total number of high-resolution protein structures in Protein Data Bank (PDB) [25] is negligible. This bottleneck has made it difficult to gain insights into structural diversity. Methods of predicting structures have improved drastically over the years, shifting from the idea of identifying co-varying positions in multiple sequence alignments (MSAs) as directly interacting residues [26] to using deep neural networks. This bottleneck has widened substantially in the last few years, as programs are now available, such as Alphafold [27] and ESMFold [28], with which the 3D structures of proteins can be predicted. To predict the structures, these programs make use of physical, geometric, and evolutionary constraints. The evolutionary constraints used are based on the fact that homologous sequences fold into similar structures. MSAs show that protein families frequently possess extended segments of conserved residues, which play a critical role in the protein's function. For instance, cell surface receptor proteins exhibit diverse amino acid sequences at their binding sites for extracellular cues, yet they display greater similarity in their amino acid sequences that engage with common intracellular signaling proteins.

To comparatively investigate the dynamics of entire protein families, a variety of predicted structures can now be included in analyses. However, examining an entire protein family could prove overly resource-intensive in terms of both the computational and memory requirements, as each structure would have to be calculated first. A more straightforward and computationally efficient approach for deriving structural attributes from sequences involves generating their contact maps. A contact map serves as a 2D depiction of the relative closeness of residues within the protein's spatial arrangement. This, in turn, can be used to investigate the structural dynamics of the protein using GNMs. As mentioned, MSAs also contain information on the residues that are in direct contact. This information can therefore also be used to create contact maps without having to calculate the complete 3D structure of the protein. It is, in principle, possible to go directly from the sequences to the functional dynamics of a protein using direct coupling analysis (DCA), for example, as has recently been shown [29]. Such shortcuts mean an immense saving in computational power, given the huge volume of protein sequences available.

In this study, we have taken a different approach to tap into the vast amount of sequence information available to generate contact maps, providing a simple yet effective means to investigate the complex landscape of proteins' functions and allosteric regulation. We developed a pattern-matching approach, which identifies short, conserved sequence motifs from homologous 3D structures from PDB [25] or AlphafoldDB [30, 31]. These sequence motifs are used to generate contact maps. The contact maps are then used to apply the entropy transfer concept of Schreiber [10] in combination with dGNM [18] to identify sources and sinks of information, as previously described. This methodology offers a computationally efficient means to pinpoint potential allosteric regulation sites, opening new avenues for understanding the conserved allosteric regulation mechanisms in protein domains and relating them to the proteins' functions. We considered four protein domains, namely DHFR, PDZ, S100, and SH3, which are widely studied for their allosteric behavior, and determined a common sink–source behavior for each domain.

Methods

Collecting sequences

First, we extracted the sequence datasets for the following protein domains from the Pfam database [32] using the representative Proteome 75 [33] dataset: dihydrofolate reductase (DHFR) (Pfam ID: PF00186), PSD-95/Discs-large/ZO-1 (PDZ) (Pfam ID: PF00595), SRC Homology 3 (SH3) (Pfam ID: PF100018), and S100 (Pfam ID: PF01023). We selected these four widely studied protein domains in order to establish our method. For each of the domains, high-resolution structures are available. From the initial set of sequences from Pfam, we removed any duplicate sequences. For each sequence, the conserved domain was extracted using the appropriate Hidden Markov model (HMM) [34] from Pfam [32]. Only the sequences with a high alignment score (sequences with bit-scores higher than the overall mean bit-score) and a length greater than or equal to the respective HMM motif were selected, which covered the entire sequence region for the specific domain. Initially, we downloaded 43,000 DHFR, 173,969 PDZ, 93,857 SH3, and 5,082 S100 domain sequences. After the selective filtering step, we established a set of high-quality sequences. The number of sequences per domain was 5,895 for DHFR, 5,070 for PDZ, 4,674 for SH3, and 2,507 for the S100 domain.

Generation of contact maps: pattern search and matching

For generating contact maps for the set of sequences, we searched the PDB [21] using the JackHMMER [34] tool for the structures of each of the four studied domains. Note that for sequences that do not have homologous structures, AlphaFoldDB [30] could be utilized as an alternative resource. For each of the four protein families, we curated a list of homologous PDB structures. In these structures, we subsequently identified contact patterns, namely the center of mass of residues of up to four amino acids that were located at a maximum distance of 8 Å from each other. Conserved contact patterns were successful in identifying novel motifs in a specific set of protein domains in an earlier study [35], as they represented a specific fold. All sequences in a set were then searched for homologous structures, and the contact patterns were extracted and saved for each domain.

For each pattern, residues that were located 8 Å from each other were denoted in uppercase letters and the residues in between are denoted in lowercase letters. To generate the contact maps, we used MATLAB's (MATLAB (R2023b)) *localalign* [36] function and locally aligned the uppercase elements within a pattern to the query sequence with a gap-opening penalty of 10, which was set to 8 by default. An empty N × N contact matrix, C, was generated where N is the length of the sequence. If three uppercase strings within the pattern aligned with the residues i, j, and k of the query sequence, a value of 1 was added to each corresponding pair within the contact matrix, namely, C(i,j), C(i,k), and C(j,k).

For each domain, the contact maps established by the pattern-matching approach were compared with contact maps generated directly from the established structures.

dGNM and calculation of transfer entropy

GNM is a simplified representation of proteins' dynamics and structural fluctuations. It is widely used to study the collective motions and essential dynamics of proteins. GNM provides insights into the low-frequency global motions of proteins by considering only the connectivity between atoms, without explicitly simulating the atomic interactions. The connectivity between the atoms is represented by the Kirchhoff matrix, Γ. The contact map can be used to define the Kirchhoff matrix. The Γ matrix and the contact map (CM) have the following format:

$$\Gamma_{ij} = \begin{cases} 0, if\ i \neq j\ and\ R_{ij} > r_c \\ -1, if\ i \neq j\ and\ R_{ij} \leq r_c \\ -\sum_{k=1,k\neq i}^{N} \Gamma_{ik}\ , if\ i = j \end{cases} \quad (1)$$

$$\mathrm{CM}_{ij} = \begin{cases} 0, R_{ij} > r_c \\ 1, R_{ij} \leq r_c \end{cases} \quad (2)$$

where i, j, and k represent the residue indices, R_{ij} is the distance between the Cα atoms, and r_c is the cutoff distance used to determine the first interaction shell, taken as 8 Å for this study. The Γ matrix relates the correlations of residues to their connectivity.

$$\langle \Delta R_i\, , \Delta R_j \rangle = \left(\Gamma_{ij}\right)^{-1} \quad (3)$$

The inverse of the Γ matrix can be computed by an eigenvalue decomposition since the determinant is zero.

Schreiber's transfer entropy formulation helps us understand how the past behavior of one process influences the future behavior of another process; therefore, it requires a time delay, τ, and describes the causal relationships between processes. The dynamic version of the GNM describes the entropy transferred from residue i to residue j with the following equation:

$$T_{i\to j}(\tau) = H\left(\Delta R_j(t+\tau)|\Delta R_j(t)\right) - H\left(\Delta R_j(t+\tau)|\Delta R_i(t), \Delta R_j(t)\right) \qquad (4)$$

where the first term, $H\left(\Delta R_j(t+\tau)|\Delta R_j(t)\right)$, is the conditional entropy of residue j at time t+τ, given its values at time t. The second term, $H\left(\Delta R_j(t+\tau)|\Delta R_i(t), \Delta R_j(t)\right)$, is the conditional entropy of residue j at time t+τ, given the values of residue i and j at time t. The result $T_{i\to j}(\tau)$ indicates the extent of the reduction in the entropy in the future fluctuations of residue j, given the previous values of residues i and j. When the conditional entropy terms are expressed in terms of probabilities, commonly known as Shannon's entropy formula, the transfer entropy equation becomes:

$$T_{i\to j}(\tau) = -\langle \ln p\left(\Delta R_j(t+\tau), \Delta R_j(t)\right)\rangle + \langle \ln p\left(\Delta R_j(t)\right)\rangle + \langle \ln p\left(\Delta R_j(t+\tau), \Delta R_i(t), \Delta R_j(t)\right)\rangle - \langle \ln p\left(\Delta R_i(t), \Delta R_j(t)\right)\rangle \qquad (5)$$

The time correlations of the residues' fluctuations are given by the dGNM as:

$$\langle \Delta R_i(t), \Delta R_j(t+\tau)\rangle = \sum_k A_{ij}(k) exp\left\{-\lambda_k \frac{\tau}{\tau_0}\right\} \qquad (6)$$

$$A_{ij}(k) = -\lambda_k{}^{-1} u_i^{(k)} u_j^{(k)} \qquad (7)$$

where $\lambda_k$is the kth eigenvalue and $u_i^{(k)}$ is the ith component of the kth eigenvector of the Γ matrix. When Eqn. 6 is substituted into Eqn. 5, the transfer entropy formula becomes:

$$T_{i\to j}(\tau)= 0.5\ \ln\left[\left(\sum_k A_{jj}(k)\right)^2 - \left(\sum_k A_{jj}(k)\ exp\left\{-\lambda_k\frac{\tau}{\tau_0}\right\}\right)^2\right] -$$
$$0.5\ ln\left[\left(\sum_k A_{ii}(k)\right)\left(\sum_k A_{jj}(k)\right)^2 + 2\left(\sum_k A_{ij}(k)\right)\sum_k A_{jj}(k)\ exp\left\{-\lambda_k\frac{\tau}{\tau_0}\right\}\sum_k A_{ij}(k)\ exp\left\{-\lambda_k\frac{\tau}{\tau_0}\right\} -\right.$$
$$\left.\left\{\left(\sum_k A_{ij}(k)\ exp\left\{-\lambda_k\frac{\tau}{\tau_0}\right\}\right)^2 + \left(\sum_k A_{ij}(k)\right)^2\right\}\left(\sum_k A_{jj}(k)\right) -\right.$$
$$\left.\left(\sum_k A_{jj}(k)\ exp\left\{-\lambda_k\frac{\tau}{\tau_0}\right\}\right)^2\left(\sum_k A_{ii}(k)\right)\right] - 0.5\ \ln\left(\sum_k A_{jj}(k)\right) + 0.5\ \ln\left[\left(\sum_k A_{ii}(k)\right)\left(\sum_k A_{jj}(k)\right) -\right.$$
$$\left.\left(\sum_k A_{ij}(k)\right)^2\right] \qquad (8)$$

The details of the formulation can be found in previous articles [18, 37, 38]. We can use the net entropy transfer that accounts for a causal relationship to understand which residue is the driver and which residue is the driven. The net entropy transfer is the difference between the entropy transferred from residue i to j and the entropy transferred from residue j to i. We define net transfer entropy as:

$$Net\ T_{i\to j}(\tau) = T_{i\to j}(\tau) - T_{j\to i}(\tau) \qquad (9)$$

A positive net transfer entropy value indicates that residue i exerts an influence on the fluctuations of j, and vice versa. The net entropy transfer equals zero under two conditions: (1) when residue i and residue j are not correlated or (2) the transfer from residue i to j is the same as that from residue j to residue i, indicating no causality. When we combine the net entropy transfer from residue i to all available residues j, we can determine the overall behavior of residue i, particularly whether it acts as an entropy source or an entropy sink.

$$Net\ T_{i\to\odot}(\tau) = \sum_j T_{i\to j}(\tau) - T_{j\to i}(\tau) \qquad (10)$$

If the cumulative net entropy of residue i has a positive value, residue i is the source of entropy for the system, where information flows from residue i to the rest of the system. Conversely, if the cumulative net entropy of residue i has a negative value, information flows from the source residues to residue i, designating it as an entropy sink.

We performed a dGNM analysis on the predicted contact maps derived from protein sequences. Subsequently, we calculated the net transfer entropy for each protein and used Eqn. 10 to plot the cumulative net entropy transfer values and identified local maxima on the positive side and local minima on the negative side of the plot.

$$List_{Source} = local\ max(Net\ T_{i\to\odot}(\tau) > 0)|_{All\ sequences} \quad (11)$$
$$List_{Sink} = local\ min(Net\ T_{i\to\odot}(\tau) < 0)|_{All\ sequences}$$

We stored the cumulative net entropy transfer values for these local maxima in a list of "sources" and those for the local minima in a list of "sinks" with their cumulative net entropy values.

After generating the "source" and "sink" lists for a protein family, we summed the collected cumulative net values of entropy transfer over all sequences. By summing these values across all sequences, we gained insights into each residue's role in the allosteric communication within the domain. This method helped us find conserved entropy source and sink regions in the protein domains. To be able to identify the exact source and sink residues for protein families, we set thresholds for positive and negative values. We used a quantile-based approach. The threshold for the sources was set at the 85th percentile, and the threshold for the sinks was set at the 15th percentile.

Results and discussion

Proteins are classified into protein families on the basis of their sequences and structural similarities, which have a common evolutionary origin. This can be used to study their function and allosteric regulation. We have developed a method to quickly convert the sequences of protein families into contact maps without having to model a 3D structure for each sequence. To test our approach, we selected domains with different folds. We used four protein domains for which high-resolution structures are available and whose allosteric behavior has been sufficiently studied, namely the DHFR, PDZ, SH3, and S100 domains. This allowed us to use high-quality sequences from the Pfam database. To use the available high-resolution 3D structures for the domains, we recursively searched the PDB database using JackHMMER. The characteristic folds of these protein domains are often established through characteristic conserved sequence motifs that are in direct contact in the 3D structure. We applied a contact threshold of 8 Å and selected characteristic patterns of up to four amino acids. Contact patterns have been used successfully in identifying novel motifs in a specific set of protein domains [22]. We have extended this strategy by using the contact patterns found this way to establish the contact map. The contact maps were then transformed into the Kirchhoff matrix and used in the dGNM for further analysis. For each protein, we compared a contact map established from the available PDB structures with the contact map established solely with our pattern-matching approach. In all cases, the distant interactions in the 3D structure were clearly recognized by the pattern-matching approach, which allowed us to study the protein domain in detail without having the high-resolution structures available. Our approach is illustrated in Fig. 1.

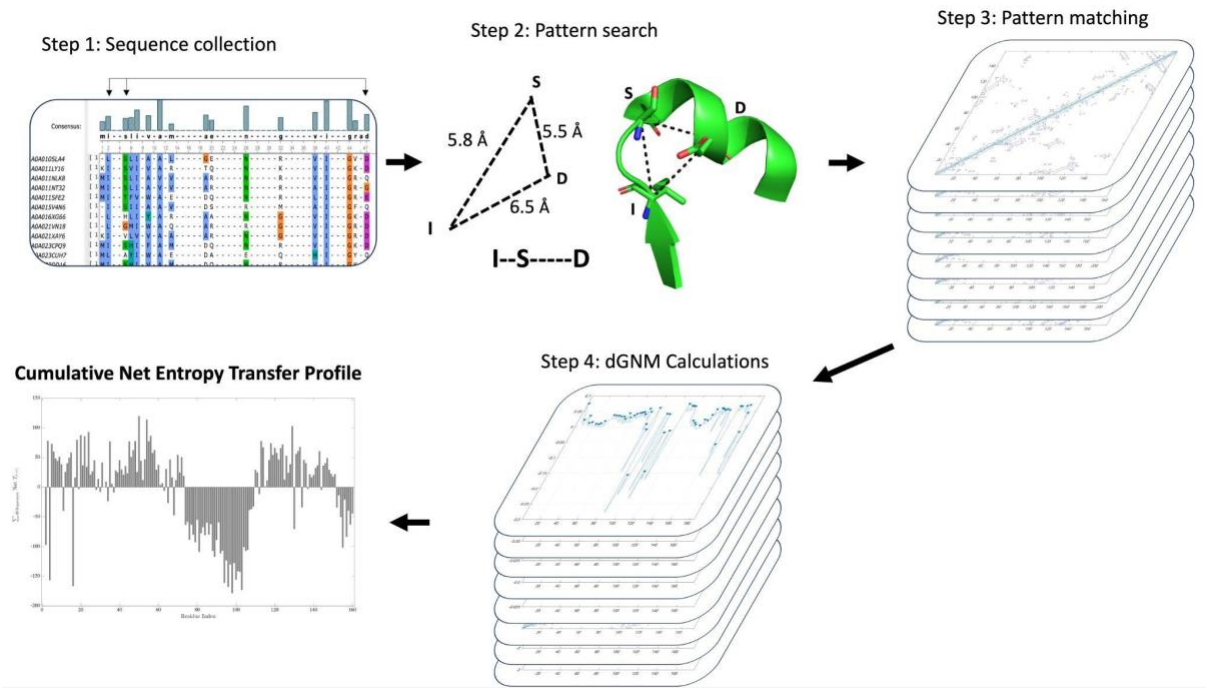

Figure 1. Workflow illustrating the procedure for finding the net entropy transfer profiles for protein domains without their high-resolution structures being available.
In the first step, the sequences of a protein family were searched for with already established HMMs in the Pfam database, and a non-redundant set was generated. In Step 2, contact patterns were established with the help of homologous protein structures. In Step 3, these contact patterns were searched for in the individual sequences, and a contact map was created. These contact maps were then used for dGNM calculations and finding the net entropy transfer profile for specific protein domains (Step 4).

DHFR

The enzyme DHFR converts 7,8-dihydrofolate (DHF) into 5,6,7,8-tetrahydrofolate, an important cofactor in many enzymatic reactions, by using nicotinamide adenine dinucleotide (NADPH) as a hydride donor. Its allosteric mechanism has been studied extensively via a variety of experimental and computational approaches [39]. The backbone of the DHFR domain, which is approximately 160 residues long, consists of a central beta-pleated sheet composed of eight strands. Of these strands, seven align in a parallel, whereas the eighth adopts an antiparallel orientation. These beta strands are joined together by four alpha helices. Three loops play a significant role in the function: the Met20 loop (Residues 9–23), the βF-βG loop (FG loop; Residues 116–132), and the βG-βH loop (GH loop; Residues 135–149). The active site is located within the Met20 loop, which features a conserved Pro-Trp

dipeptide at Positions 21 and 22, respectively. The tryptophan within this dipeptide plays a crucial role in the substrate-binding of the enzyme. DHFR switches its conformation during its catalytic cycle, going between an open and a closed state, depending on whether its active site is available or blocked by the Met20 loop. The residues that are involved in the allosteric communication pathway have been identified by nuclear magnetic resonance experiments, molecular dynamics simulations, and mutational studies (reviewed in [39]).

To determine the side chains involved in allosteric regulation using GNM, we started by generating the contact maps. To start, we downloaded DHFR's sequences from Pfam with the accession number PF00186, removed the duplicate sequences, and retained the sequences with a high HMM score for the DHFR HMM profile for further analysis. A higher HMM score than the mean HMM score served as the criterion for inclusion.

A comparison of the predicted contact maps using our pattern-matching approach and contact maps extracted from the high-resolution structures is shown in Fig. 2. Overall, the predicted contact maps follow the contacts of the true contact map. Note that for some of the domains studied (e.g., DHFR, as mentioned above), PDB structures exist in different conformations that reflect the dynamic process of their activity. These structural dynamics are also visible in the contact maps, both the one that takes the mean structure as a basis and also the one that is based on the pattern-matching approach.

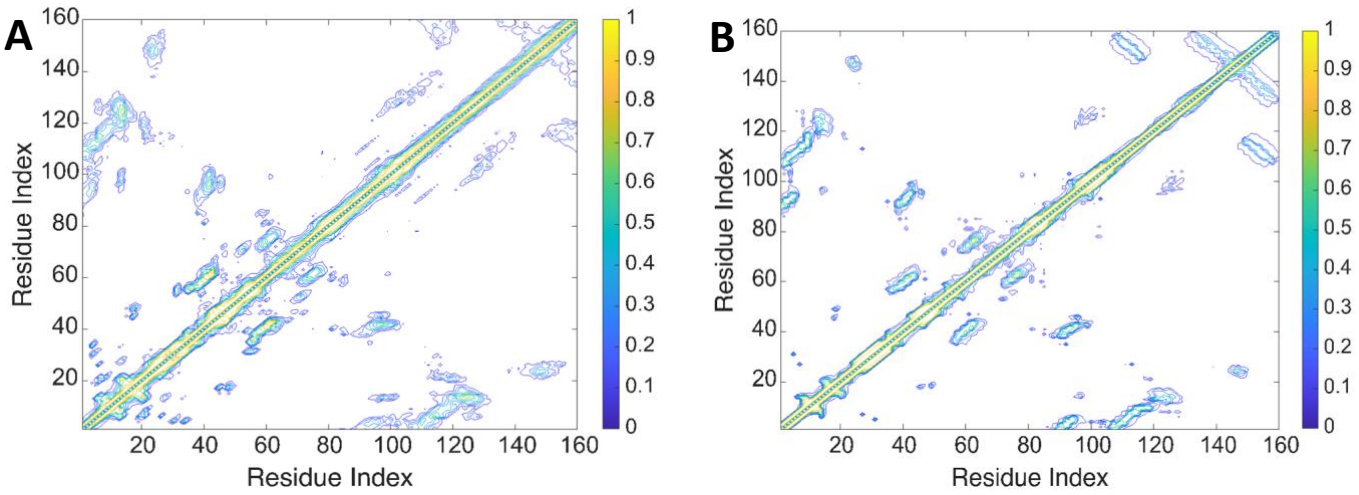

Figure 2: Comparison of the mean predicted and real contact maps for DHFR.
(A) Mean contact maps predicted for 5895 sequences using the pattern-matching method. (B) Mean contact maps for 550 DHFR structures in the PDB.

Subsequently, we performed a dGNM analysis of the predicted contact maps. We calculated the cumulative net transfer entropy for each sequence by using Eqn. 10. For each sequence, we identified the local maximum and local minimum values for the net transfer entropy. As an example, a plot of the cumulative net entropy transfer calculated from the predicted contact map is shown in Fig. 3 for just one DHFR sequence. Generally, negative values are labeled as entropy sinks, whereas positive values are labeled as entropy sources.

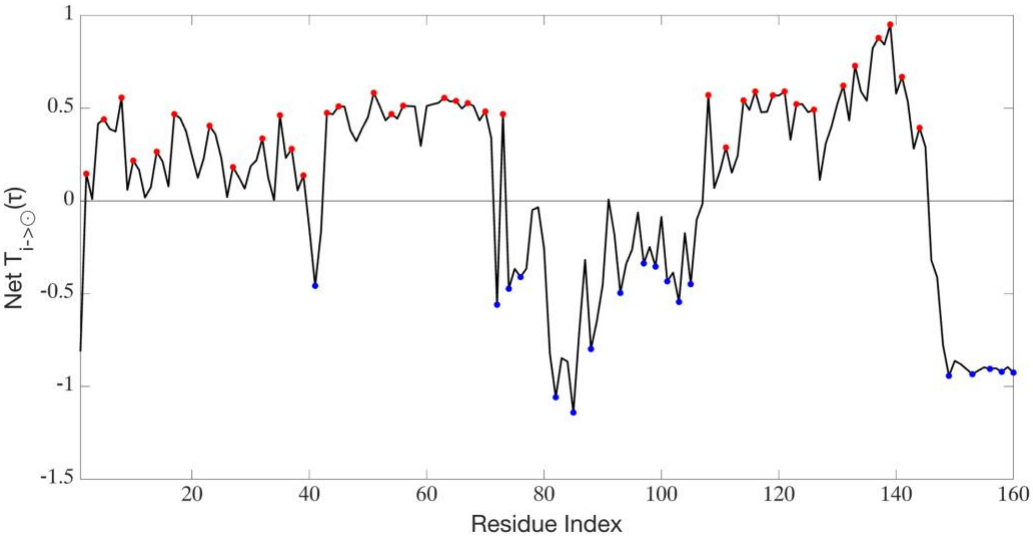

Figure 3: Cumulative net entropy transfer of residues calculated by using Eqn. 10 for a sample DHFR protein.

Entropy source residues (local maxima) are labeled in red; entropy sink residues (local minima) are labeled in blue for the DHFR protein of the ploughfish (Gymnodraco acuticeps) (UniProt ID A0A6P8URK9).

For all DHFR sequences, we then summed up the cumulative net entropy transfer values for each source and sink residue we had detected as the local maxima and minima. We set a threshold on the total cumulative net entropy values to detect the most probable source and sink residues for the entire DHFR family. The threshold values for the total cumulative net entropy transfer for the DHFR family are indicated in Fig 3. The residues with total cumulative net entropy transfer values in the upper 15% were considered to be conserved source residues and the values in the lower 15% were considered to be conserved sink residues and thus very probably represent the residues conserved throughout the entire protein family.

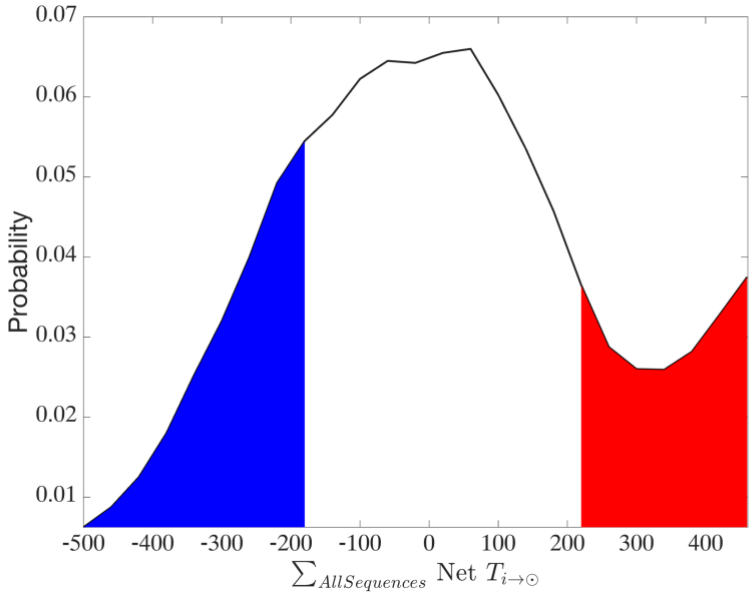

Figure 4: Example of a distribution plot showing the selection of thresholds for the total cumulative net entropy transfer for entropy sources and sinks for the DHFR family.
The threshold for selecting the entropy sources was set at the 85th percentile and the threshold for entropy sinks was set to the 15th percentile.

Next, we viewed the position of the cumulative source and sink residues identified by our analysis, which revealed that loops that are critical for catalytic activity [40] exhibit notably high cumulative net entropy values (Fig. 4A), suggesting that a signal is transmitted from these regions to the rest of the protein. We mapped the residues with high and low cumulative net entropy values onto the structure, providing a visual representation of how this network is distributed across the domain (Fig. 4B). The majority of the information sources originate from the active site, known for its ligand-binding role. In a previous study, it was established that removing the nicotinamide adenine dinucleotide phosphate (NADP) cofactor disrupts the coordination between the FG loop and the adenosine binding domain (residues 40–50) [41]. Our observations align with this finding, revealing that the binding of NADP induces stronger signals than DHF binding, whereby information flows from the ligand-binding sites to the Met20, FG, and GH loops, causing them to undergo conformational rearrangements and attenuation. From our analysis, we identified a network of residues that play an active role in the enzyme's function, suggesting that the hydride transfer mechanism has been preserved throughout the course of evolution [42, 43].

Notably, nuclear magnetic resonance studies have shown that different mutations of a conserved glycine residue in the FG loop, G121, significantly slowed the transfer of hydride, except in the G121V mutant, where gating becomes kinetically significant, with three-fold faster rates of transferring hydride. Mutations of S148 in the GH loop modulate the ligand's off-rates [40, 44]. Removing the central residues (residues between 16–19) of the Met-20 loop and substituting a single glycine residue results in a 500-fold reduction in the rate of transferring hydride and an increase in the rate of NADPH dissociation [44]. X-ray studies [45] and molecular dynamics simulations [46] have suggested that the interaction of these three important loops affects how well the enzyme binds to its ligands and how fast it performs its function.

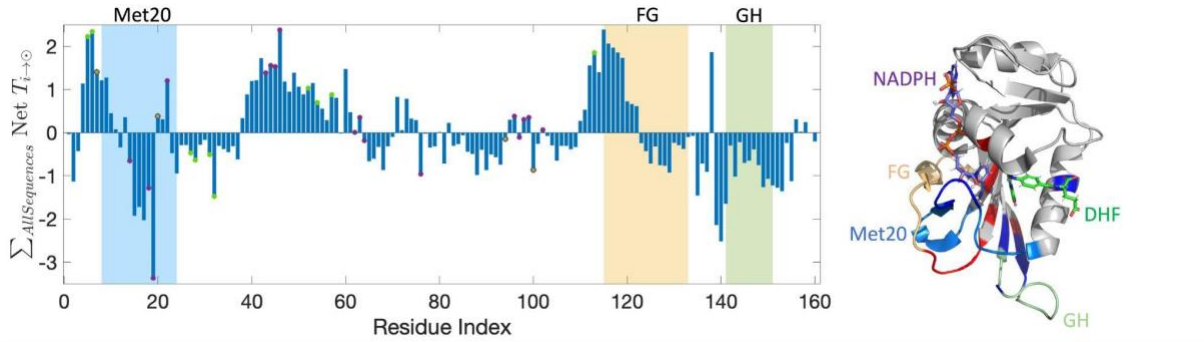

Figure 5: Normalized cumulative net entropy profile of DHFR.
(A) Normalized summation of the cumulative net entropy values of the entropy sources (local maxima) and entropy sinks (local minima) over all selected sequences of DHFR. DHF-binding residues are labeled in green, NADP-binding residues are labeled in purple, and residues that bind both are labeled in purple-lined green dots. The regions of the Met20, FG, and GH loops are shaded on the plot with blue, yellow, and green, respectively. (B) Residues with high cumulative net entropy values (>1.72) and low cumulative net entropy values (<–1.18) are colored in red and blue, respectively, on the structure of DHFR with PDB ID 4PDJ [47]. NADP+ is colored purple and DHF is colored green. The regions of the Met20, FG, and GH loops are indicated.

PDZ

PDZ domains are versatile protein modules that play essential roles in mediating protein–protein interactions, particularly in the context of cellular signaling and organization. PDZ domains interact with a wide range of protein partners, including receptors, ion channels, enzymes, and scaffolding proteins. These interactions often occur within the cell membrane

or at specialized cellular junctions, such as the synapse in neurons [48, 49]. PDZ domains consist of approximately 90 residues that fold into a compact globular structure [50]. They recognize short amino acid sequences found at the C-terminus of the target proteins. Generally, PDZ domains consist of six β strands arranged in two sheets and two α helices. The peptide binds into the hydrophobic cavity between βB and αB. PDZ domains are grouped into three types on the basis of their peptide-binding specificity: PDZ1, PDZ2, and PDZ3. PDZ2 and PDZ3 have an additional α helix at their C-terminus, which stabilizes the protein and is suggested to have a role in ligand-binding [51]. Binding ligands into the cavity of PDZ domains leads to alterations in the motions of more distant side chains. Ligand-binding can be regulated by other proteins, as shown, for example, for the cell polarity protein Par-6. Binding of the Rho GTPase Cdc42 drastically increases the ligand-binding affinity [52]. PDZ domains have been extensively studied for their allosteric mechanism (reviewed in [51] and [53]).

To investigate PDZ domains, we first generated contact maps for the PDZ domain's sequences. Note that the analysis includes only the conserved core region of all PDZ domains and does not include the third helix, which also plays a role in the allosteric regulation of the domain. It has been shown that the removal of this helix decreases the ligand-binding affinity [54]. In Fig. 6, a comparison of the predicted and real contact maps is shown. Overall, the predicted contact maps (Fig. 6a) retrieved the populated contact regions of the real contact map (Fig. 6B).

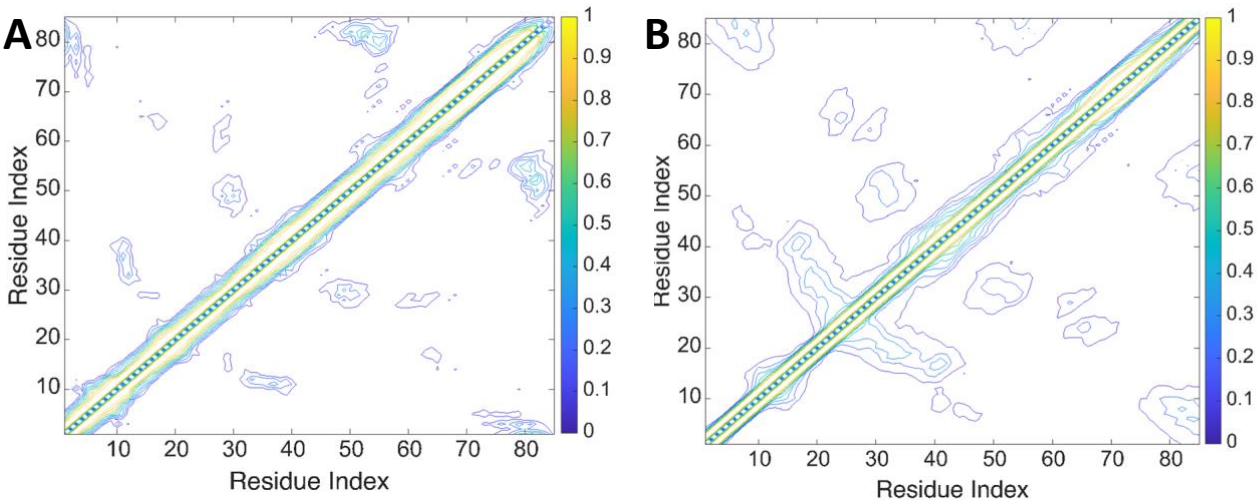

Figure 6: Comparison of the mean predicted and real contact maps for PDZ domains.
Mean contact maps predicted for 5,070 sequences using the pattern-matching method. (B) Mean contact maps for 500 PDZ structures in PDB.

As described above in detail for DHFR, we performed a dGNM analysis using the predicted contact map for the PDZ domain. We identified the conserved sources and sinks for the canonical PDZ domain through an analysis of the cumulative net transfer entropy. The sources and sinks detected this way (Fig. 7) are in agreement with previous studies. The majority of the signals came from βB, αA, βC, and βE, and were directed towards βA and αB, where the ligand binds the shallow pocket between βB and αB. The pattern of sources and sinks detected here matches the experimental findings [53, 55] and probably reflects the molecular events occurring during ligand-binding in terms of allostery. This suggests that the distal residue dynamics change upon ligand-binding. In particular, the αA helix was repeatedly found as an allosteric region that induces changes in the ligand-binding site in various PDZ domains [50, 51, 53, 55–57].

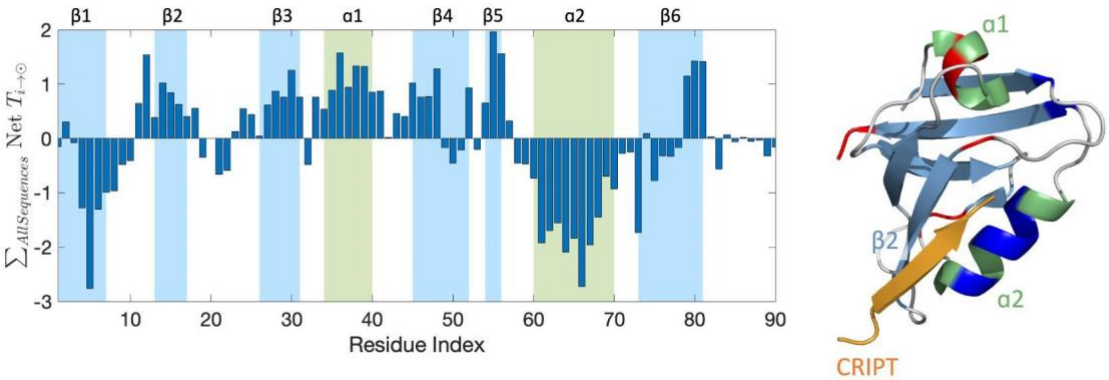

Figure 7: Normalized cumulative net entropy profile of PDZ.

(A) Normalized summation of the cumulative net entropy of entropy sources (local maxima) and entropy sinks (local minima) over all selected sequences of PDZ. Secondary structures are colored to show β sheets in blue and a helices in green. (B) Residues with high cumulative net entropy values (>1.32) and low cumulative net entropy values (<–1.73) are colored in red and blue, respectively on the structure of PDZ3 with PDB ID 5HFF [58]. Note that the third a helix is not displayed. Secondary structures are colored according to Panel A. The peptide CRIPT is colored in orange.

SH3

SRC homology 3 (SH3) domains are approximately 50–60 amino acids long and are involved in substrate recognition, signal transduction, cytoskeletal modification, and proliferation. They recognize proline-rich motifs with hydrophobic pockets formed by aromatic residues. Their binding site involves the region between the RT loop (between β1 and β2), characterized by conserved arginine (R) and threonine (T) amino acids, and the SRC loop (between β2 and β3). Another variable loop, the distal loop, is located on the opposite face of the domain (reviewed in [59–62]). Although SH3 domains are highly conserved, relatively small, and simple, they are highly specific in their interactions. A potential approach for achieving this specificity involves adjusting the binding strength by allosteric effects, where distant residues from the binding site can influence the binding energetics for a particular substrate.

As outlined above for the other domains, we used individual sequences to generate a contact map for the SH3 domains. Overall, the populated regions of the predictions (Fig. 8A) matched the real contact map (Fig. 8B).

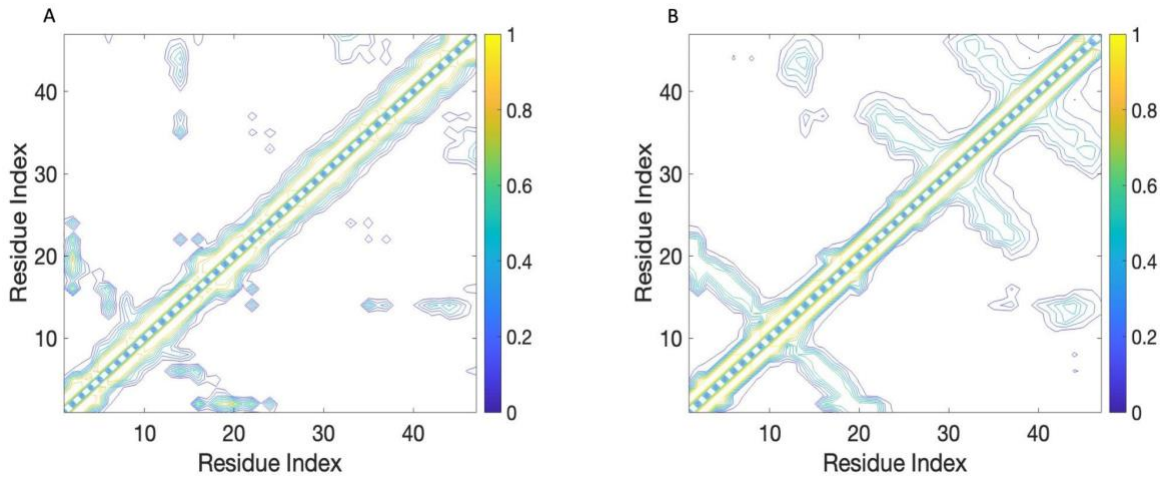

Figure 8: Comparison of the mean predicted and real contact maps for SH3.

(A) Mean contact maps predicted for 4,674 sequences using the pattern-matching method. (B) Mean contact maps for 357 structures of SH3 in PDB. Note that some SH3 domains have slightly elongated β3 and β4 regions, which are not considered in the HMM of Pfam. Only the regions of the HMM were used for this presentation.

The normalized cumulative net entropy profile of the SH3 domains (Fig. 9) showed that the source residues are located at the ligand-binding regions, which are distributed among the RT and SRC loops [63]. Among them, two distinct residues have been proven to be the major

allosteric sites for the SH3 domain (Residues 15 and 45) [64]. The distal loop, which helps stabilize the tertiary contacts, is among the regions with a strong signal source. In Fig. 9B, we demonstrate how the allosteric network's residues are distributed through the SH3 domain and how the distant regions and the ligand-binding region connect. Our results agree with the network residues identified in previous studies [63, 64].

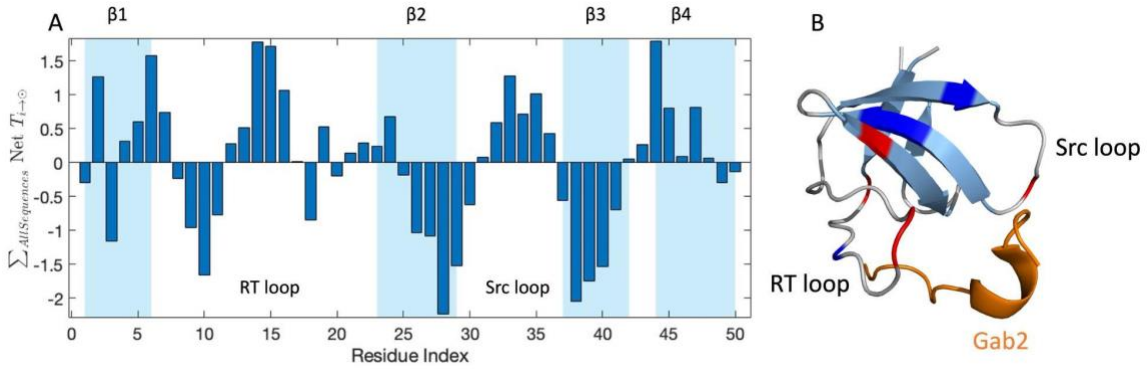

Figure 9: Normalized cumulative net entropy profile of SH3.
(A) Normalized summation of the cumulative net entropy of entropy sources (local maxima) and entropy sinks (local minima) over all selected sequences of the SH3 domain. Secondary structures are labeled in the plot. (B) Residues with high cumulative net entropy values (>1.32) and low cumulative net entropy values (<–1.68) are colored in red and blue, respectively, on the structure of the SH3 domain with the PDB ID 2VWF [65]. The proline-rich peptide Gab2 is colored in orange.

S100

S100 proteins are a family of calcium-binding proteins approximately 90 residues long that are involved in a diverse set of functions such as enzyme regulation, cytoskeletal dynamics, signal transduction, the immune response, cell differentiation, proliferation, and migration (reviewed in [66–68]). Most of them act as homodimers. Each subunit has two Ca^{2+}-binding sites formed by helix–loop–helix motifs ("EF-hand type"), the C-terminal canonical EF loop is similar to the one found in other proteins such as calmodulin, and the N-terminal "pseudo-EF" is specific to S100 proteins. The two EF-hands are connected by a loop, which is critical for the interactions with the target proteins [66]. Normally the Ca^{+2}-binding affinity of most S100 proteins is moderate [69]. Interestingly, the Ca^{2+}-binding affinities of both loops are different [70]. Upon binding the effector proteins, the binding, the Ca^{+2}-binding affinity increases through heterotropic allostery [71, 72]. Upon binding Ca^{2+}, the domain undergoes

a conformational change, exposing hydrophobic regions that facilitate interactions with other proteins [55]. The contribution of each of the two different calcium-binding loops to this cooperativity is not fully understood. Fig. 10 illustrates the underlying conformational changes caused by binding Ca^{+2}.

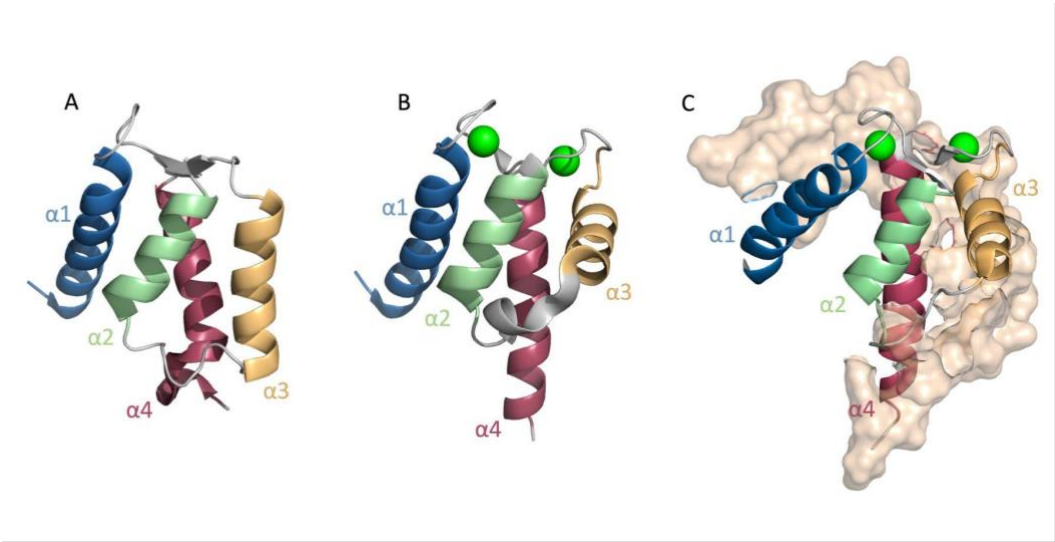

Figure 10: Cartoon representation of the conformational changes in the S100 domain upon binding Ca^{+2}.

(A) Ca^{+2}-free form (apo state) of an S100 domain (PDB ID 1K9P [73]). (B) S100 domain with two Ca^{+2} ions bound (holo state, PDB ID 1K9K [73]). The bound Ca^{+2} ions are shown as green spheres. Ca^{+2}-binding changes the conformation of the α3 and exposes the hydrophobic pocket for ligand-binding. (C) S100 domains with two bound Ca^{+2} ions and the possible ligand-binding regions shown as the surface, the ligands from the S100 domains of S100A4 (PDB ID 4CFR [74]), and S100B (PDB ID 1MWN [75]).

As for the investigations above, we first established a contact map (Fig. 11C) using our pattern-matching approach and compared this with the contact maps based on high-resolution structures. As outlined above, S100 proteins undergo a larger conformational change when binding Ca^{+2} (Fig. 10). This conformational change alters the interaction pattern of the domain, as can be seen for the different contact maps of the apo (Fig. 11A) and holo states (Fig. 11B) of the protein. Remarkably, the predicted contact map (Fig. 11C) showed additional contacts between the α2 and α4 helices, and appears to represent the two conformational states, apo and holo, reflecting the dynamic nature of the allosteric regulation of S100 proteins.

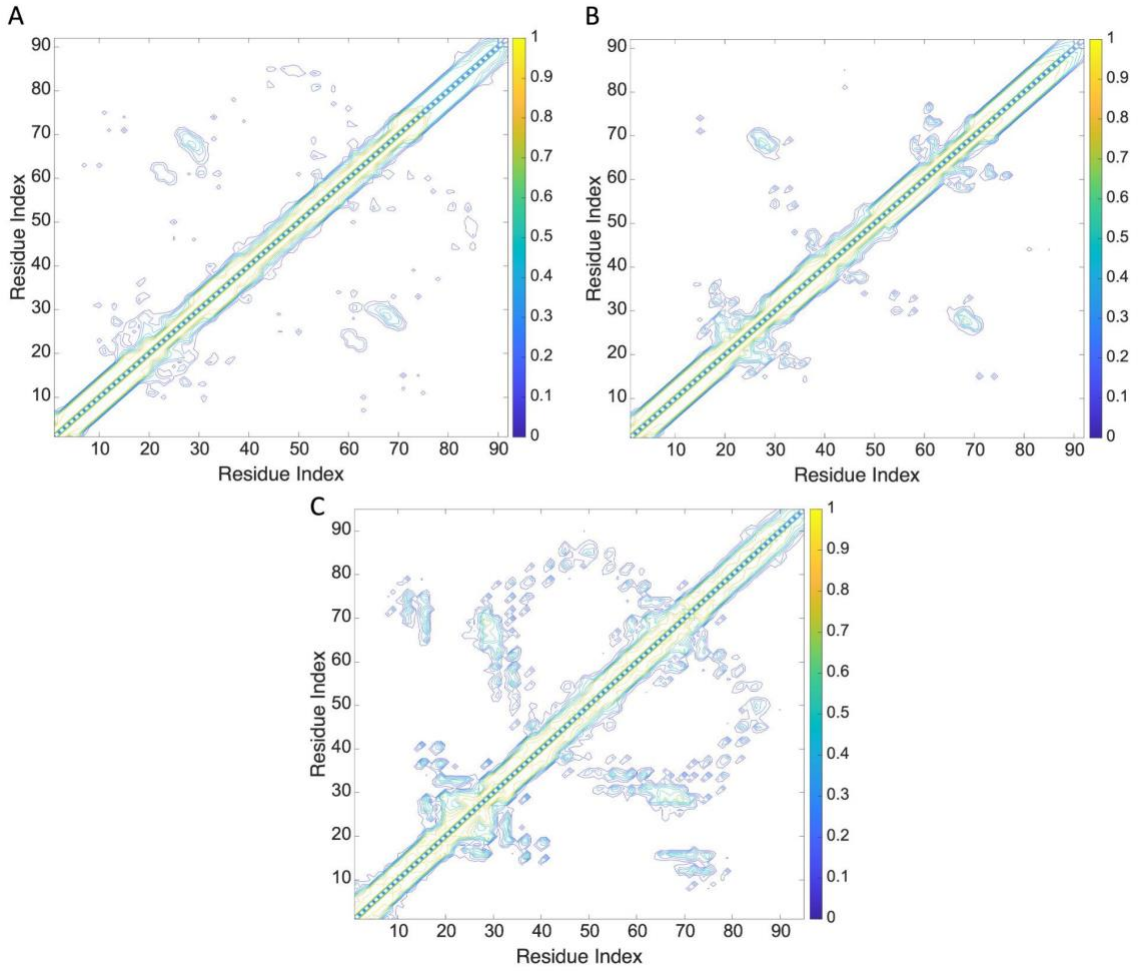

Figure 11: Comparison of the apo and the holo conformations of the PDB structures and the predicted contact map for S100.

Mean contact maps for (A) 57 apo and (B) 108 holo structures of the S100 dimer in PDB. (C) Mean contact map predicted for 2,507 sequences using the pattern-matching method.

The normalized cumulative net entropy profile of S100 domains (Fig. 12A) showed that the binding of each Ca^{+2} induced a different type of signal. It has been discussed in previous studies that binding of Ca^{+2} to the second EF-hand is a principal trigger for the conformational switch [76]. In Fig. 12A, we can see how calcium-binding induces changes in the functionally relevant domains responsible for target recognition, such as the hinge domain between the EF-hands and the C-terminal residues. We can see how EF-I affects the overall communication

in the protein after binding calcium in the EF-II. Our findings agree with the experimental relaxation rates and molecular dynamics simulations [76].

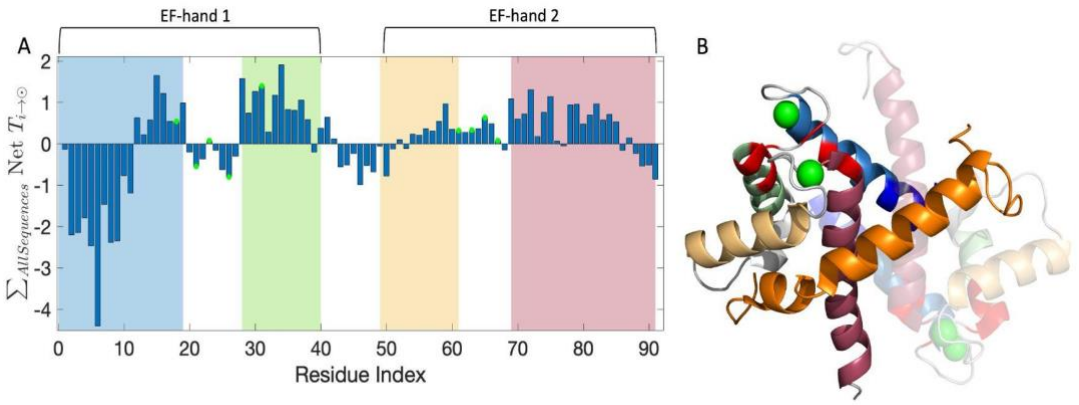

Figure 12: Normalized cumulative net entropy profile of S100.
(A) Normalized summation of the cumulative net entropy of entropy sources (local maxima) and entropy sinks (local minima) over all selected sequences of the S100 domain. The four helices are colored in the same way as in Fig. 10: α1, blue; α2, green; α3, yellow; and α4, red. The Ca^{+2}-binding residues in EF-hands 1 and 2 are marked with green circles. (B) Ribbon diagram of the structure of S100 (PDB ID 4CFR [58]). Residues with high cumulative net entropy values (>1.14) and low cumulative net entropy values (<–1.75) are colored in red and blue, respectively. The bound Ca^{+2} ions are shown as green spheres. The interaction partner Myosin-9 is colored orange.

Conclusion

In this study, we present a method of exploring the conservation of allosteric mechanisms in a diverse set of protein domains such as DHFR, PDZ, SH3, and S100. To test this method, we drew on already researched protein domains for which extensive MSAs with HMMs were available in the Pfam database and for which at least one high-resolution structure was available. We introduced pattern-matching to generate a simple 2D representation of the contact topologies of proteins directly from their sequences. This approach makes it possible to produce 2D contact maps for a large protein family without first having to create a model of the 3D structure for each sequence using Alphafold or ESMFold, which significantly minimizes the computing time required. This approach helped us analyze a large number of sequences for a given domain within a short amount of time without the need to download

structures from the database or rely on GPUs on a personal computer. The contact maps created in the process largely reflect the contacts of existing 3D structures. This is a significant improvement on methods that calculate the contacts from coevolutionary correlations.

Building on Schreiber's concept of transfer entropy, we identified the causal relationships and key residues in protein–protein interactions and explained the conformational changes that occur during mutation or ligand-binding. Furthermore, our findings have important implications for the analysis of networks in newly discovered proteins, as the allosteric communication pathway and specific residues contributing to this process are preserved throughout evolution within protein domains. How well this method works for investigating other, less well-studied protein families for allosteric mechanisms needs to be shown.

This approach unveils the structural and functional intricacies of these domains, shedding light on protein–ligand interactions and allosteric communication networks, which are consistent with previous experimental findings. These insights serve to deepen our understanding of molecular mechanisms while paving the way for the development of targeted drugs and novel therapeutic interventions.

Availability of data

The data that support the findings of this study are openly available in the Zenodo repository https://zenodo.org/records/11185696

Acknowledgments

This work was supported by the Swiss National Science Foundation (Grant 31003A_182732 and 310030_219549 to D.F.). We thank the Division de Calcul et Soutien à la Recherche of the UNIL for access to the university's computer infrastructure. We thank all members of the Fasshauer Laboratory for helpful discussions.

Author contributions

A.H. designed the study, performed the experiments, and analyzed the data; A.H. and D.F. wrote the paper.

Competing interests

The authors declare no competing interests.

Reference list

1. Kalodimos C, Edelstein S: **Allosteric interactions and biological regulation (Part I)**. *Journal of Molecular Biology* 2013, May, **425** (9):1391–1592.

2. Kalodimos C, Edelstein S: **Allosteric interactions and biological regulation (Part II)**. *Journal of Molecular Biology* 2013, July, **425** (13):2277–2392.

3. Tsai C-J, Nussinov R: **A unified view of "how allostery works"**. *PLoS Computational Biology* 2014, **10** (2):e1003394.

4. Liu J, Nussinov R: **Allostery: an overview of its history, concepts, methods, and applications**. *PLoS Computational Biology* 2016, **12** (6):e1004966.

5. Cooper A, Dryden D: **Allostery without conformational change: a plausible model**. *European Biophysics Journal* 1984, **11** :103–109.

6. Ahmed MH, Ghatge MS, Safo MK: **Hemoglobin: structure, function and allostery**. In *Vertebrate and Invertebrate Respiratory Proteins, Lipoproteins and Other Body Fluid Proteins* (Hoeger U, Harris JR, eds) 2020:345–382.

7. Helmstaedt K, Krappmann S, Braus GH: **Allosteric regulation of catalytic activity: *Escherichia coli* aspartate transcarbamoylase versus yeast chorismate mutase**. *Microbiology and Molecular Biology Reviews* 2001, **65** (3):404–421.

8. Li M, Bao Y, Li M, Guo J: **GPCR allostery: a view from computational biology**. *Current Medicinal Chemistry* 2023, **30** (40):4533–4553.

9. Hu Z, Zheng X, Yang J: **Conformational trajectory of allosteric gating of the human cone photoreceptor cyclic nucleotide-gated channel**. *Nature Communications* 2023, **14** (1):4284.

10. Schreiber T: **Measuring information transfer**. *Physical Review Letters* 2000, **85** (2):461–464.

11. Barr D, Oashi T, Burkhard K, Lucius S, Samadani R, Zhang J, Shapiro P, MacKerell AD Jr, van der Vaart A: **Importance of domain closure for the autoactivation of ERK2**. *Biochemistry* 2011, **50** (37):8038–8048.

12. Corrada D, Morra G, Colombo G: **Investigating allostery in molecular recognition: insights from a computational study of multiple antibody–antigen complexes**. *The*

Journal of Physical Chemistry B 2013, **117** (2):535–552.

13. Garcia Michel LR, Keirns CD, Ahlbrecht BC, Barr DA: **Calculating transfer entropy from variance–covariance matrices provides insight into allosteric communication in ERK2**. *Journal of Chemical Theory and Computation* 2021, **17** (5):3168–3177.

14. Hacisuleyman A, Erman B: **Entropy transfer between residue pairs and allostery in proteins: quantifying allosteric communication in ubiquitin**. *Plos Computational Biology* 2017, **13** (1): e1005319.

15. Kamberaj H, van der Vaart A: **Extracting the causality of correlated motions from molecular dynamics simulations**. *Biophysical Journal* 2009, **97** (6):1747–1755.

16. Acar B, Rose J, Fas BA, Ben-Tal N, Lewinson O, Haliloglu T: **Distinct allosteric networks underlie mechanistic speciation of ABC transporters**. *Structure* 2020, **28** (6):651–663.

17. Altintel B, Acar B, Erman B, Haliloglu T: **Subsets of slow dynamic modes reveal global information sources as allosteric sites**. *Journal of Molecular Biology* 2022, **434** (17):167644.

18. Hacisuleyman A, Erman B: **Causality, transfer entropy, and allosteric communication landscapes in proteins with harmonic interactions**. *Proteins: Structure, Function, and Bioinformatics* 2017, **85** (6):1056–1064.

19. Hu D, Hu D, Liu L, Barr D, Liu Y, Balderrabano-Saucedo N, Wang B, Zhu F, Xue Y, Wu S: **Identification, clinical manifestation and structural mechanisms of mutations in AMPK associated cardiac glycogen storage disease**. *EBioMedicine* 2020, **54:** 102723.

20. Sen TZ, Feng Y, Garcia JV, Kloczkowski A, Jernigan RL: **The extent of cooperativity of protein motions observed with elastic network models is similar for atomic and coarser-grained models**. *Journal of Chemical Theory and Computation* 2006, **2** (3):696–704.

21. Bahar I, Atilgan AR, Erman B: **Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential**. *Folding and Design* 1997, **2** (3):173–181.

22. Yang L, Song G, Jernigan RL: **How well can we understand large-scale protein motions using normal modes of elastic network models?** *Biophysical Journal* 2007, **93** (3):920–929.

23. Katebi AR, Sankar K, Jia K, Jernigan RL: **The use of experimental structures to model protein dynamics**. *Molecular Modeling of Proteins* 2015, **1215**:213–236.

24. The UniProt Consortium: **UniProt: the universal protein knowledgebase in 2023**. *Nucleic Acids Research* 2023, **51** (D1):D523–D531.

25. wwPDB Consortium: Protein Data Bank: the single global archive for 3D macromolecular structure data. *Nucleic Acids Research* 2019, **47** (D1):D520–D528.

26. Morcos F, Pagnani A, Lunt B, Bertolino A, Marks DS, Sander C, Zecchina R, Onuchic JN, Hwa T, Weigt M: **Direct-coupling analysis of residue coevolution captures native contacts across many protein families**. *Proceedings of the National Academy of Sciences* 2011, **108** (49):E1293–E1301.

27. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Žídek A, Potapenko A: **Highly accurate protein structure prediction with AlphaFold**. *Nature* 2021, **596** (7873):583–589.

28. Lin Z, Akin H, Rao R, Hie B, Zhu Z, Lu W, Smetanin N, Verkuil R, Kabeli O, Shmueli Y: **Evolutionary-scale prediction of atomic-level protein structure with a language model**. *Science* 2023, **379** (6637):1123–1130.

29. Jia K, Kilinc M, Jernigan RL: **Functional protein dynamics directly from sequences**. *The Journal of Physical Chemistry B* 2023, **127** (9):1914–1921.

30. Varadi M, Anyango S, Deshpande M, Nair S, Natassia C, Yordanova G, Yuan D, Stroe O, Wood G, Laydon A: **AlphaFold protein structure database: massively expanding the structural coverage of protein-sequence space with high-accuracy models**. *Nucleic Acids Research* 2022, **50** (D1):D439–D444.

31. Schaeffer RD, Daggett V: **Protein folds and protein folding**. *Protein Engineering, Design & Selection* 2011, **24** (1–2):11–19.

32. Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar GA, Sonnhammer EL, Tosatto SC, Paladin L, Raj S, Richardson LJ: **Pfam: the protein families database in 2021**. *Nucleic Acids Research* 2021, **49** (D1):D412–D419.

33. Chen C, Natale DA, Finn RD, Huang H, Zhang J, Wu CH, Mazumder R: **Representative proteomes: a stable, scalable and unbiased proteome set for sequence analysis and functional annotation**. *PloS One* 2011, **6** (4):e18910.

34. Johnson LS, Eddy SR, Portugaly E: **Hidden Markov model speed heuristic and iterative HMM search procedure**. *BMC Bioinformatics* 2010, **11**:1–8.

35. Bradley P, Kim PS, Berger B: **TRILOGY: Discovery of sequence-structure patterns across diverse proteins**. *Proc Natl Acad Sci U S A* 2002, 99(13): 8500–8505.

36. Barton GJ: **An efficient algorithm to locate all locally optimal alignments between two sequences allowing for gaps**. *Bioinformatics* 1993, **9** (6):729–734.

37. Erkip A, Erman B: **Dynamics of large-scale fluctuations in native proteins. Analysis based on harmonic inter-residue potentials and random external noise**. *Polymer* 2004, **45** (2):641–648.

38. Haliloglu T, Bahar I, Erman B: **Gaussian dynamics of folded proteins**. *Physical Review Letters* 1997, **79** (16):3090.

39. Goldstein M, Goodey NM: **Distal regions regulate dihydrofolate reductase-ligand interactions**. *Allostery: Methods and Protocols* 2021:185–219.

40. Osborne MJ, Schnell J, Benkovic SJ, Dyson HJ, Wright PE: **Backbone dynamics in dihydrofolate reductase complexes: role of loop flexibility in the catalytic mechanism**. *Biochemistry* 2001, **40** (33):9846–9859.

41. Van Den Bedem H, Bhabha G, Yang K, Wright PE, Fraser JS: **Automated identification of functional dynamic contact networks from X-ray crystallography**. *Nature Methods* 2013, **10** (9):896–902.

42. Li J, Fortunato G, Lin J, Agarwal PK, Kohen A, Singh P, Cheatum CM: **Evolution conserves the network of coupled residues in dihydrofolate reductase**. *Biochemistry* 2019, **58** (37):3861–3868.

43. Li J, Lin J, Kohen A, Singh P, Francis K, Cheatum CM: **Evolution of optimized hydride transfer reaction and overall enzyme turnover in human dihydrofolate reductase**. *Biochemistry* 2021, **60** (50):3822–3828.

44. McElheny D, Schnell JR, Lansing JC, Dyson HJ, Wright PE: **Defining the role of active-site loop fluctuations in dihydrofolate reductase catalysis**. *Proceedings of the National Academy of Sciences* 2005, **102** (14):5032–5037.

45. Sawaya MR, Kraut J: **Loop and subdomain movements in the mechanism of**

***Escherichia coli* dihydrofolate reductase: crystallographic evidence**. *Biochemistry* 1997, **36** (3):586–603.

46. Radkiewicz JL, Brooks CL: **Protein dynamics in enzymatic catalysis: exploration of dihydrofolate reductase**. *Journal of the American Chemical Society* 2000, **122** (2):225–231.

47. Wan Q, Bennett BC, Wilson MA, Kovalevsky A, Langan P, Howell EE, Dealwis C: **Toward resolving the catalytic mechanism of dihydrofolate reductase using neutron and ultrahigh-resolution X-ray crystallography**. *Proceedings of the National Academy of Sciences* 2014, **111** (51):18225–18230.

48. Amacher JF, Brooks L, Hampton TH, Madden DR: **Specificity in PDZ-peptide interaction networks: computational analysis and review**. *Journal of Structural Biology: X* 2020, **4** :100022.

49. Sheng M, Sala C: **PDZ domains and the organization of supramolecular complexes**. *Annual Review of Neuroscience* 2001, **24** (1):1–29.

50. Ye F, Zhang M: **Structures and target recognition modes of PDZ domains: recurring themes and emerging pictures**. *Biochemical Journal* 2013, **455** (1):1–14.

51. Gautier C, Laursen L, Jemth P, Gianni S: **Seeking allosteric networks in PDZ domains**. *Protein Engineering, Design and Selection* 2018, **31** (10):367–373.

52. Whitney DS, Peterson FC, Kittell AW, Egner JM, Prehoda KE, Volkman BF: **Binding of crumbs to the Par-6 CRIB-PDZ module is regulated by Cdc42**. *Biochemistry* 2016, **55** (10):1455–1461.

53. Stevens AO, He Y: **Allosterism in the PDZ family**. *International Journal of Molecular Sciences* 2022, **23** (3):1454.

54. Chi CN, Haq SR, Rinaldo S, Dogan J, Cutruzzola F, Engström Å, Gianni S, Lundstrom P, Jemth P: **Interactions outside the boundaries of the canonical binding groove of a PDZ domain influence ligand binding**. *Biochemistry* 2012, **51** (44):8971–8979.

55. Petit CM, Zhang J, Sapienza PJ, Fuentes EJ, Lee AL: **Hidden dynamic allostery in a PDZ domain**. *Proceedings of the National Academy of Sciences of the United States of America* 2009, **106** (43):18249–18254.

56. Stevens AO, Kazan IC, Ozkan B, He Y: **Investigating the allosteric response of the**

PICK1 PDZ domain to different ligands with all-atom simulations. *Protein Science* 2022, **31** (12):e4474.

57. Peterson FC, Penkert RR, Volkman BF, Prehoda KE: **Cdc42 regulates the Par-6 PDZ domain through an allosteric CRIB–PDZ transition**. *Molecular cell* 2004, **13** (5):665–676.

58. Raman AS, White KI, Ranganathan R: **Origins of allostery and evolvability in proteins: a case study**. *Cell* 2016, **166** (2):468–480.

59. Mehrabipour M, Jasemi NSK, Dvorsky R, Ahmadian MR: **A systematic compilation of human SH3 domains: a versatile superfamily in cellular signaling**. *Cells* 2023, **12** (16):2054.

60. Mayer BJ, Eck MJ: **SH3 domains: minding your p's and q's.** *Current Biology* 1995, **5** (4):364–367.

61. Kurochkina N, Guha U: **SH3 domains: modules of protein–protein interactions**. *Biophysical Reviews* 2013, **5**:29–39.

62. Klimov D, Thirumalai D: **Stiffness of the distal loop restricts the structural heterogeneity of the transition state ensemble in SH3 domains**. *Journal of Molecular Biology* 2002, **317** (5):721–737.

63. Malagrinò F, Troilo F, Bonetti D, Toto A, Gianni S: **Mapping the allosteric network within a SH3 domain**. *Scientific Reports* 2019, **9** (1):8279.

64. Faure AJ, Domingo J, Schmiedel JM, Hidalgo-Carcedo C, Diss G, Lehner B: **Global mapping of the energetic and allosteric landscapes of protein binding domains**. *bioRxiv* 2021:2021.2009. 2014.460249.

65. Harkiolaki M, Tsirka T, Lewitzky M, Simister PC, Joshi D, Bird LE, Jones EY, O'Reilly N, Feller SM: **Distinct binding modes of two epitopes in Gab2 that interact with the SH3C domain of Grb2**. *Structure* 2009, **17** (6):809–822.

66. Marenholz I, Heizmann CW, Fritz G: **S100 proteins in mouse and man: from evolution to function and pathology (including an update of the nomenclature)**. *Biochemical and Biophysical Research Communications* 2004, **322** (4):1111–1122.

67. Fritz G, Botelho HM, Morozova-Roche LA, Gomes CM: **Natural and amyloid self-assembly of S100 proteins: structural basis of functional diversity**. *The FEBS Journal*

2010, **277** (22):4578–4590.

68. Bresnick AR: **S100 proteins as therapeutic targets**. *Biophysical Reviews* 2018, **10** (6):1617–1629.

69. Pietzsch J, Hoppmann S: **Human S100A12: a novel key player in inflammation?** *Amino Acids* 2009, **36** :381–389.

70. Sivaraja V, Kumar TKS, Prudovsky I, Yu C: **Three-dimensional solution structure of a unique S100 protein**. *Biochemical and Biophysical Research Communications* 2005, **335** (4):1140–1148.

71. Young BD, Cook ME, Costabile BK, Samanta R, Zhuang X, Sevdalis SE, Varney KM, Mancia F, Matysiak S, Lattman E: **Binding and functional folding (BFF): a physiological framework for studying biomolecular interactions and allostery**. *Journal of Molecular Biology* 2022, **434** (23):167872.

72. Donato R, Heizmann CW: **S100B protein in the nervous system and cardiovascular apparatus in normal and pathological conditions**. *Cardiovascular Psychiatry and Neurology* 2010, **2010: 929712**.

73. Otterbein LR, Kordowska J, Witte-Hoffmann C, Wang C-LA, Dominguez R: **Crystal structures of S100A6 in the Ca^{2+}-free and Ca^{2+}-bound states: the calcium sensor mechanism of S100 proteins revealed at atomic resolution**. *Structure* 2002, **10** (4):557–567.

74. Duelli A, Kiss B, Lundholm I, Bodor A, Petoukhov MV, Svergun DI, Nyitray L, Katona G: **The C-terminal random coil region tunes the Ca^{2+}-binding affinity of S100A4 through conformational activation**. *PloS One* 2014, **9** (5):e97654.

75. Inman KG, Yang R, Rustandi RR, Miller KE, Baldisseri DM, Weber DJ: **Solution NMR structure of S100B bound to the high-affinity target peptide TRTK-12**. *Journal of Molecular Biology* 2002, **324** (5):1003–1014.

76. Wang Q, DiForte C, Aleshintsev A, Elci G, Bhattacharya S, Bongiorno A, Gupta R: **Calcium mediated static and dynamic allostery in S100A12: implications for target recognition by S100 proteins**. *Protein Science* 2024, **33** (4):e4955.