PLOS ONE

RESEARCH ARTICLE

SenseNet, a tool for analysis of protein structure networks obtained from molecular dynamics simulations

Markus Schneider*, Iris Antes†

TUM Center for functional Protein Assemblies and TUM School of Life Sciences, Technische Universität München, Freising, Germany

† Deceased.
* markusg.schneider@tum.de

OPEN ACCESS

Citation: Schneider M, Antes I (2022) SenseNet, a tool for analysis of protein structure networks obtained from molecular dynamics simulations. PLoS ONE 17(3): e0265194. https://doi.org/10.1371/journal.pone.0265194

Editor: Oscar Millet, CIC bioGUNE, SPAIN

Received: July 16, 2021

Accepted: February 25, 2022

Published: March 17, 2022

Data Availability Statement: Details concerning molecular dynamics simulations as well as initial structures, topologies and simulation input files are within the manuscript and its Supporting Information files. Molecular dynamics trajectories used for analysis are available from the Dryad database: https://doi.org/10.5061/dryad.3r2280gf2 (DOI:10.5061/dryad.3r2280gf2). The SenseNet Software is freely available at the Cytoscape App Store (https://cytoscape.org/) or at https://www.bioinformatics.wzw.tum.de/sensenet/method/. SenseNet is a software written and maintained by the authors, designed to be used as a plugin for the third-party network analysis tool Cytoscape 3, and

Abstract

Computational methods play a key role for investigating allosteric mechanisms in proteins, with the potential of generating valuable insights for innovative drug design. Here we present the SenseNet ("Structure ENSEmble NETworks") framework for analysis of protein structure networks, which differs from established network models by focusing on interaction timelines obtained by molecular dynamics simulations. This approach is evaluated by predicting allosteric residues reported by NMR experiments in the PDZ2 domain of hPTP1e, a reference system for which previous computational predictions have shown considerable variance. We applied two models based on the mutual information between interaction timelines to estimate the conformational influence of each residue on its local environment. In terms of accuracy our prediction model is comparable to the top performing model published for this system, but by contrast benefits from its independence from NMR structures. Our results are complementary to experimental data and the consensus of previous predictions, demonstrating the potential of our new analysis tool SenseNet. Biochemical interpretation of our model suggests that allosteric residues in the PDZ2 domain form two distinct clusters of contiguous sidechain surfaces. SenseNet is provided as a plugin for the network analysis software Cytoscape, allowing for ease of future application and contributing to a system of compatible tools bridging the fields of system and structural biology.

Introduction

Protein structure networks map atoms from a protein structure to nodes and define edges to represent atom interactions, e.g. contacts and hydrogen bonds. The resulting networks may be used to predict e.g. allosteric communication pathways [1–3] with potential applications in innovative drug design [4–8]. Most commonly, such analyses are based on individual crystal structures and rely on centrality measures such as betweenness centrality (BC) or characteristic path length centrality (CPLC) to identify functionally important residues [1–3,9]. However, application of these algorithms to experimental structures of e.g. the PDZ domain did not

distributed over the Cytoscape App Store. Cytoscape 3 can be downloaded for free at https://cytoscape.org/. The authors did not have special access privileges to Cytoscape or the Cytoscape App Store; all third-party software was used in a manner that is available to anyone. The source code for SenseNet and AIFgen is included in the java archive (.jar) files used to run these programs. To differentiate the distribution including source code from the original release, the new version was designated v1.0.1. There are otherwise no material differences to version 1.0.0 used in this study. Both versions remain available on our website and the Cytoscape app store.

Funding: This work was supported by the Deutsche Forschungsgemeinschaft (https://www.dfg.de/; SFB 1035/A10 and SFB749/C08 to IA). MS was supported by the TUM International Graduate School of Science and Engineering (IGSSE; https://www.igsse.gs.tum.de/). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

provide results consistent with experiment [10]. It has been generally recognized that highly dynamic effects such as allostery, which are not always associated with stable conformations, are difficult to study solely on the basis of individual experimentally obtained structures [5,11–13]. Computational methods for analyzing structure ensembles obtained from e.g. molecular dynamics simulations (MD), which capture the dynamic behavior of proteins, are therefore attractive for allosteric prediction [11,14–20]. Several tools exist for analysis of structure ensemble networks, among them xPyder [21], PyInteraph [22], MD-TASK [23], gRINN [24], PSN-Ensemble [25], NAPS [26,27], RIP-MD [28], Bio3D [29], MDN [30] and the Cytoscape plugin RINalyzer [31]. A common approach for network analysis of MD data is to define edges by correlation analysis of atomistic motions, which comes at the cost of losing structural and conformational details of the underlying interactions. In addition, many approaches use a rigid mapping of one node per residue, preventing the combination of different levels of resolution, e.g. to separate information flow between backbone and sidechain atoms. Finally, the majority of tools are provided as standalone programs or webservers, making it difficult to combine different algorithms within a single analysis session. To address these limitations, we developed SenseNet, a plugin for the free network analysis software Cytoscape [32]. SenseNet is based on an alternative strategy to scalar correlation coefficients, namely associating edges with MD-based timelines, which allow to track the evolution of interactions during a simulation by checking their existence at predefined timeslots. This representation allows for a larger variety of analyses than correlation-based approaches, like e.g. interaction averages, lifetime analysis, frame clustering, or shared information between timelines.

Ligand binding often modulates protein function by triggering conformational changes distant from the binding site. A major goal of computational allosteric prediction is to identify key residues sensing ligand binding events over long intramolecular distances; in the context of computational predictions, these residues are commonly labeled as "allosteric". For the purpose of evaluating these methods, PDZ domains are a well-established reference system. Members of this abundant domain class commonly bind C-terminal or short internal peptide sequences and participate in allosteric interactions with other domains [33,34], serving as initiators and mediators of protein assembly processes [35–37]. Although the domain is allosterically modulated by its peptide ligands, crystal and solution NMR structures of the PDZ2 domain of hPTP1e (human Protein-Tyrosine Phosphatase 1e) show no substantial conformational changes between apo and ligand bound states [38]. Therefore, the relationship between structure, dynamics, and allostery in the PDZ2 domain of hPTP1e was explored by Lee and coworkers, who identified a number of allosteric residues by probing the effects of ligand binding and point mutations on NMR backbone and methyl side chain dynamics [38–40]. However, open questions remain concerning the contribution of residues lacking methyl groups and how individual residues act together to form allosteric pathways, motivating structure-based computational prediction as a complementary strategy [41]. Methods previously applied to the PDZ2 system include interaction energy and correlation networks [42,43], elastic network models [44], hydrogen bond heat diffusion pathways [45], relative entropy networks of distance distributions (REDAN) [46], and coordinate fluctuations [47,48]. Furthermore, specialized simulation techniques were employed such as perturbation response scanning [49], rigid residue scan (RRS) [50], and NMR guided simulations [10,51]. However, results reported by computational studies have shown considerable variance, warranting efforts to consolidate and improve prediction models [41].

In this work, we present our network analysis software SenseNet and evaluate two of therein implemented, timeline-focused algorithms to find pathways of allosteric information transfer in the PDZ2 domain. By quantifying how much information the timelines of physical interactions provide about their environment, we obtained accurate models for

predicting allosteric residues in PDZ2. Finally, we propose a consolidated allosteric model combining our results with experimental data and the consensus of previous predictions, which suggests that PDZ2 contains two allosteric pathways formed by clusters of contiguous sidechain surfaces.

Materials & methods

Algorithms

Protein structure networks based on interaction timelines. In a structure network as implemented in SenseNet, each node (which together form the set of nodes N) represents a single atom or a group of atoms while edges represent interactions between nodes. If several interaction types (e.g. contacts or hydrogen bonds) are present, a node pair may be connected by more than one edge. Every interaction is associated with a timeline, representing the different states of the interaction in the analyzed ensemble of structures, e.g. simulation frames from an MD trajectory. We define an atomistic timeline as the vector

$$\boldsymbol{X}_{\alpha\beta k} = \left[\begin{cases} 1 & \text{if } \alpha \text{ and } \beta \text{ interact as type k in frame } t \\ 0 & \text{otherwise} \end{cases} \right]_t \tag{1}$$

where α, β are nodes representing single atoms, k is an interaction type and t is a simulation time frame (bold type face denotes matrices and vectors). Timelines of edges connecting two atom groups (e. g. residues) are calculated as

$$\boldsymbol{X}_{ijk} = \sum_{\alpha \in i} \sum_{\beta \in j} \boldsymbol{X}_{\alpha\beta k} \tag{2}$$

in which i, j are nodes representing atom groups. The connectivity between nodes is given by the symmetric adjacency matrix

$$\boldsymbol{A}_k = \left[\begin{cases} 1 & \text{if } i \text{ and } j \text{ are connected by an edge of type } k \\ 0 & \text{otherwise} \end{cases} \right]_{ij} \tag{3}$$

for each interaction type k. In combination, the sets of nodes and edges form a network which encodes both the structural topology of the protein system and the fluctuations between different conformational states through its interaction timelines. Those features can then be subjected to further analyses in order to gain insights into the dynamic behavior of the protein system. Note that in cases where the network is based on a single structure instead of an ensemble of structures, the network model reduces to a simple form where each timeline has a length of one and corresponds to the number of interactions between the connected nodes.

Allosteric prediction based on correlation between interaction timelines. We propose two novel algorithms, the node correlation factor (NCF) and difference node correlation factor (DNCF), to predict residues associated with allosteric function in proteins. Our model presupposes that in order for a residue to have an observable allosteric function, its conformations must be correlated to conformational changes in its immediate environment. The conformational states of all residues are encoded within the interaction timelines in the network. We define the immediate environment as the interactions represented by neighboring edges, i.e. edges which are separated by at most a single node. Hence, we begin by considering how each interaction is correlated to interactions in its immediate environment. By applying this

definition, we obtain the edge neighbor correlation factor (ECF) as

$$\mathrm{ECF}(i,j,k) = (\boldsymbol{A_k})_{ij} \cdot \sum_{l \in K} \sum_{n,m \in N} \mathrm{I}\left(\boldsymbol{X_{ijk}}; \boldsymbol{X_{nml}}\right) \cdot (\boldsymbol{A_l})_{nm} \cdot \chi_{ijk}(n,m,l) \tag{4}$$

with i, j belonging to the node set N, k and l being part of the interaction type set K, and I is the mutual information function

$$\mathrm{I}(\boldsymbol{X};\boldsymbol{Y}) = \sum_{x \in X} \sum_{y \in Y} p(x,y) \cdot \log_2\left(\frac{p(x,y)}{p(x)p(y)}\right) \tag{5}$$

in which $p(x, y)$ represents the joint probability of values x and y and $p(x)$ corresponds to the marginal probability of state x in timeline $\boldsymbol{X}$. The mutual information function is a non-linear measure of correlation quantifying the information shared between timelines, i.e. the increase of predictability of the states in timeline $\boldsymbol{X}$ if the other timeline $\boldsymbol{Y}$ is observed [52]. Furthermore, χ represents an indicator function selecting the neighboring edges of i, j, k and is defined as

$$\chi_{ijk}(n,m,l) = \delta_{in} + \delta_{jm} - \delta_{in}\delta_{jm}\,(\delta_{kl} + 1) \tag{6}$$

where δ is the Kronecker delta and the δ_{kl} term serves to exclude the self-information of edge i, j, k. The definition ECF score is intuitively illustrated using the network shown in Fig 1. The ECF score of the blue edge is calculated as the sum of mutual information contributions between the blue edge and all its neighboring edges, shown in green. Each contributing mutual information term indicates the strength of correlation between the interaction represented by

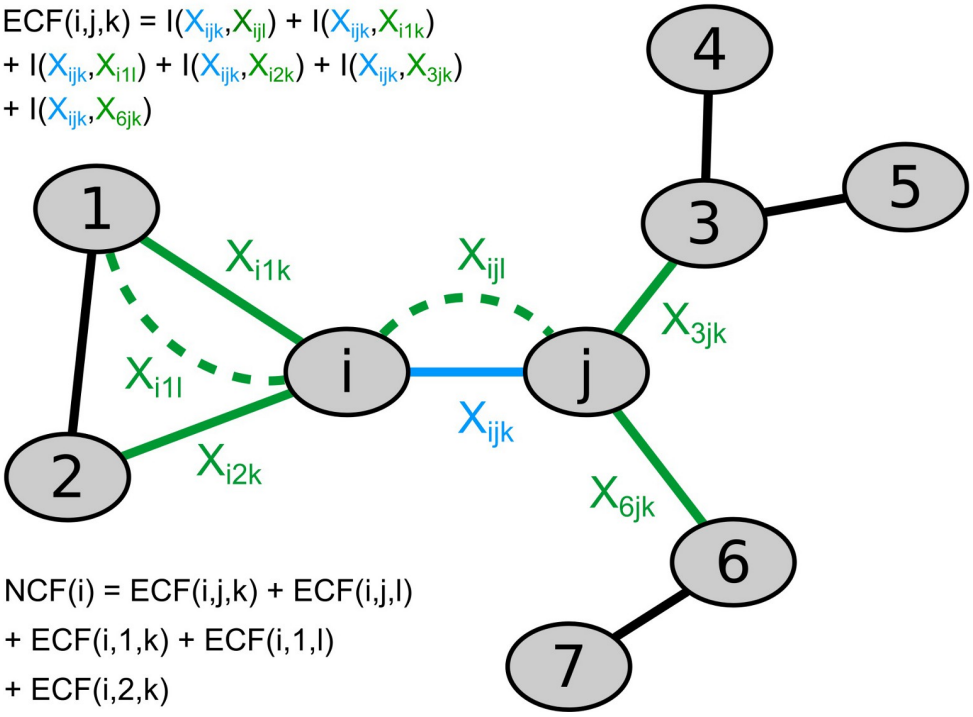

Fig 1. Example network demonstrating the calculation of edge correlation factor (ECF) and node correlation factor (NCF) scores. The ECF score of edge i, j, k (blue) is obtained by summing the mutual information of timeline X_{ijk} shared with the timelines of neighboring edges (green). The self-information $I(X_{ijk}, X_{ijk})$ is excluded. Subsequently, the NCF score of node i is calculated as the sum of ECF scores of all edges connected to i.

https://doi.org/10.1371/journal.pone.0265194.g001

the blue edge and the respective neighboring interaction. If the interaction states represented in the timeline of the blue edge are strongly correlated to the interaction states of its surrounding edges, it will lead to a high ECF score, suggesting that changes in one interaction may affect its immediate environment; In other words, information about conformational states could then potentially be transmitted via these strongly coupled interactions. Summing up the ECF scores of a node's adjacent edges gives the node correlation factor (NCF) which can be expressed as

$$\mathrm{NCF}(i) = \sum_{k \in K} \sum_{j \in N} \mathrm{ECF}(i, j, k) \tag{7}$$

and highlights residues with strong conformational coupling. These residues, as they participate in interactions that may transfer information to their environment, are thus likely candidates for showing behavior associated with protein allostery.

As an extension to the model, another aspect can be considered for the prediction of allosteric residues, namely the conformational differences between two states of a protein system, e.g. ligand bound and ligand free. The difference node correlation factor (DNCF) quantifies changes in timeline coupling between two networks, each created from a different MD trajectory simulating either the ligand bound or the ligand free state. After selecting one trajectory as the reference and the other as the target, the definition of Eq 5 is adjusted to

$$\mathrm{I}(\boldsymbol{X}; \boldsymbol{Y}) = \sum_{x \in \cup(\boldsymbol{X}, \hat{\boldsymbol{X}})} \sum_{y \in \cup(\boldsymbol{Y}, \hat{\boldsymbol{Y}})} \left| p(x,y) \cdot \log_2\left(\frac{p(x,y)}{p(x)p(y)}\right) - \hat{p}(x,y) \cdot \log_2\left(\frac{\hat{p}(x,y)}{\hat{p}(x)\hat{p}(y)}\right) \right| \tag{8}$$

with $\widehat{\boldsymbol{X}}$, $\widehat{\boldsymbol{Y}}$ denoting the timelines from the reference simulation matching the locations of $\boldsymbol{X}$ and $\boldsymbol{Y}$ of the target simulation and $\hat{p}$ representing the probabilities of the reference timelines. Note that edges which exist solely in the reference network do not contribute, therefore the score is not symmetric with respect to interchanging target and reference networks. Substitution of Eq 8 in Eq 4 yields the DNCF score. The DNCF score measures the change in shared information between equivalent interaction timelines in the target and reference systems. This can be illustrated with the following example: Suppose there are two neighboring interactions obtained from MD simulations of the system, and the timelines show that they are strongly correlated. Then the same system is simulated again, but now including a ligand bound to an allosteric site, which are sensed by residues associated with allosteric function. The binding of a ligand to an allosteric binding pocket is likely to change the nature and efficacy of information transfer within the protein, which can manifest stronger or weaker coupling between interaction timelines. The DNCF score is composed of the pointwise mutual information contributions of the allosterically activated system as encoded in timelines $\boldsymbol{X}$ and $\boldsymbol{Y}$, from which the contributions of the equivalent reference timelines $\widehat{\boldsymbol{X}}$ and $\widehat{\boldsymbol{Y}}$ are subtracted. Thus, high DNCF scores are expected from residues for which the coupling of interactions changes between the target and reference network, i.e. before and after binding of a ligand to an allosteric site.

An essential feature of our model emerges from the definitions of the ECF, NCF and DNCF scores, namely the explicit locality of network effects. By limiting our analysis on the shared information between adjacent residues in the network, the influence of spurious correlation is reduced. To illustrate, consider that any pair of residues in a protein, no matter how far apart, would be compared. This would lead to a drastic increase of evaluated correlation terms, and thus more residue pairs showing high correlation by pure chance. At the same time, the probability that two residues influence each other directly in a substantial manner (i.e. without

detectable changes in the residues between them) is lower if they are far apart, especially as the physical interactions included in our analysis, i.e. hydrogen bonds and carbon contacts, are of limited range. Adding up contributions of distant residues would thus substantially increase the noise introduced in the analysis. Instead, we propose that in most cases it is more productive to focus on the identification of neighboring residues directly exchanging information, and to analyze how they build chains of signaling residues. However, in instances of allosteric communication lacking this locality of effects, other methods may be more accurate.

Network node centrality methods for allosteric prediction. Measures of node centrality are commonly used to detect functional residues using protein structure networks [1,2,9]. When applying these methods to prediction of allosteric residues, it is postulated that residues important to transferring signals between functional sites are related to the most central nodes in the structure network, i.e. nodes that are essential when walking the shortest path between nodes along network edges. SenseNet implements two centrality functions for this purpose: Betweenness centrality (BC) finds those nodes which are located on the largest number of shortest paths over all possible node pairs [1,53]. It is defined as

$$BC(i) = \sum_{j,k \in N,\ i \neq j \neq k} \frac{\sigma_{jk|i}}{\sigma_{jk}} \tag{9}$$

where i, j, k belong to the set of nodes N, σ_{jk} is the number of shortest paths between j and k, and $\sigma_{jk|i}$ is the number of shortest paths between j and k passing through i. The second method implemented in SenseNet is characteristic path length centrality (CPLC) [9]. For this method, nodes that are crucial for maintaining the shortest paths are presumed to be key to communication, as measured by the robustness of shortest paths to the removal of individual nodes [9]. In order to determine the robustness of the network, the characteristic path length, i.e. the average length of shortest paths in the network is considered as

$$L = \frac{1}{N_p} \sum_{i,j \in N,\ i > j} d(i,j) \tag{10}$$

where N is the set of nodes, N_p is the number of node pairs in the network and $d(i, j)$ is the minimum number of edges to be traversed between i and j. The CPLC score corresponds to the effect of removing a node on the characteristic path length of the network, which can be expressed as

$$CPLC(i) = |L - L_i| \tag{11}$$

where L_i is the characteristic path length of the network after removal of node i.

The BC and CPLC algorithms are commonly applied to individual (crystal or NMR) structures and do not trivially transfer to structure ensembles from MD simulations. This is because the networks obtained from MD simulations contain a large number of additional spurious interactions in the network compared to a crystal structure. Since Eqs 9 and 10 utilize the shortest paths between nodes along a chain of edges without accounting for the stability of the interaction, an interaction present only in a tiny fraction of the simulation could be considered with the same importance as more long-lived, substantial interactions. In contrast, NCF and DNCF methods intrinsically limit the influence of spurious interactions due to the explicit locality of contributing interactions and by definition through the mutual information function. For this reason and the fact that BC and CPLC are most commonly used with individual structures, we applied these methods only to networks obtained from crystal and NMR structures.

Molecular dynamics simulations

MD simulations in this work are based on the crystal structures of hPTP1E-PDZ2 in the apo state (PDB-ID: 3LNX) and bound to the C-terminal peptide of RA-GEF-2 (PDB-ID: 3LNY) as well as the corresponding solution NMR structures 3PDZ and 1D5G, using the first model provided in the files. These NMR structures were chosen to allow for direct comparison with previous studies [10,39]. Protein and ligand residues missing in the crystal structures were added based on their NMR structure analogues using Modeller 9.18 [54], creating 100 candidate structures and selecting the model with the best DOPE score for simulations and network analyses. MD simulations were performed using the Amber16-AmberTools17 software suite [55] with the Amber14SB force field [56] and TIP3P water [57]. The system was solvated in a cubic water box using a minimum solute-face distance of 12 Å and 150 mM NaCl. For the nonbonded interactions a 12 Å direct space cutoff and PME summation for electrostatic interactions were applied. Energy minimization was performed until convergence to 0.01 kcal * mol^{-1} * $Å^{-1}$ was reached using the XMIN minimizer. Afterwards, the volume of the solvent box was adjusted to a solvent density of 1.00 kg * m^3. For all simulations a time step of 1 fs was applied and SHAKE [58] was used for hydrogen-containing bonds. Systems were gradually heated from 0 to 300 K over 1.7 ns using a variant of our published heatup protocol [59], restraining all heavy atoms by 2.39 kcal * mol^{-1} * $Å^{-2}$ until 20 K and all backbone atoms until 200 K. For the first 1.2 ns of the heatup a Langevin thermostat was used with a collision frequency of 4 ps^{-1} and for the last 0.5 ns a Berendsen barostat was employed with a relaxation time of 2 ps. Afterwards the NPT ensemble was used with a slow coupling Berendsen thermostat at 300 K (coupling time: 10 ps) in combination with a Berendsen barostat (relaxation time: 5 ps). For each system, ten independent simulations were performed for 1 μs each (based on separate heatup runs and different randomized Langevin seeds). The initial 100 ns of each replicon were removed before analysis to reduce bias towards initial structures. Trajectory post-processing was performed with CPPTRAJ [60], using the "nativecontacts" command for contact timelines of carbon atoms (saving both native and nonnative time series), and the "hbond" command for hydrogen bonds (distance cutoff 3.5 Å; angle cutoff 135˚). The data generated by CPPTRAJ provided the interaction timelines for all network analyses based on MD trajectories, i.e. for the NCF and DNCF methods. Interaction data for BC and CPLC analyses were extracted directly from the corresponding PDB files using AIFgen with equivalent settings for interactions and distance/angle cutoffs as detailed for CPPTRAJ (see example script in S2 File).

Protein structure networks

For analyses of protein structure networks and related quantities we used the SenseNet plugin (version 1.0.0) for Cytoscape (version 3.6.1) [32]. In order to create a network, SenseNet requires a list of atom-atom interaction timelines, where each interaction is defined by a minimum of one source atom, one target atom, an interaction type (e.g. hydrogen bond), and a timeline represented as a list of interaction values corresponding to each time frame (e.g. a list where 1 indicates presence of an interaction, while 0 indicates absence in each given frame). As a general input data format for SenseNet, we defined the AIF file format, which provides a list of interaction timelines as a structured text file that can be easily created, inspected and modified using a text editor (see S2 File for an example of the format). SenseNet provides tools for automatic generation of AIF files from multiple sources. Lists of interaction timelines as created by the CPPTRAJ "hbond" and "nativecontacts" analyses can be directly converted into AIF format using the SenseNet GUI or AIFgen, which provides a command line interface to the GUI functions available in SenseNet. Alternatively, SenseNet and AIFgen can extract

timelines of pairwise contacts or hydrogen bonds directly from PDB files using the same criteria as implemented in CPPTRAJ. Example scripts demonstrating the workflow for AIFgen for converting CPPTRAJ outputs and extraction of interactions from PDB files are given in S2 File. For this work, we converted CPPTRAJ outputs of contact and hydrogen bond analyses into AIF files using AIFgen (version 1.0.4).

ECF scores were calculated with SenseNet using the therein implemented "Correlation" function set to the "Mutual information" mode. Then, the "Degree" function was used to sum over the ECF scores calculated in the previous step. DNCF scores were calculated after importing first the reference and target systems (see Eq 8) as separate networks. As references in the context of DNCF calculations, we selected the network generated from the corresponding ligand bound simulation for the analysis of the network of the free protein, and vice versa. The DNCF scores were calculated using the "Correlation" function set to "Mutual information difference". The obtained edge scores were then summed up using the "Degree" function. Edges of the two networks were considered equivalent if they connected the same residues and were of the same interaction type (Edge mapping in SenseNet set to "Match Location"). Contact betweenness centralities (BC) [53] and characteristic path length centralities (CPLC) [9] were calculated using the respective modes within the "Centrality" function and normalized using the min-max procedure. For high throughput analyses, we used the CyREST interface of Cytoscape to call the corresponding SenseNet functions. Plots were generated using matplotlib (version 3.0.3) [61] with pictures of molecular structures by VMD (1.9.3) [62] and open-source PyMOL (version 1.8.4.0) [63].

Prediction of allosteric residues

Predictions were verified against methyl sidechain dynamics data [39], using classifications as allosterically active and inactive as defined by Cilia et al. ("NMR dataset", n = 25, see S1 Table) [10]. In that study, backbones of NMR structures and Monte Carlo sampling were used to find correlated side chain torsions. As this method was not applicable to alanine residues, the authors evaluated prediction performance using either the complete NMR dataset or a variant excluding alanine residues ("NMR-Ala dataset", n = 21). To be consistent with these former studies, we chose to adopt this scheme in this work. Receiver Operating Characteristic (ROC) curves were generated by plotting, for various prediction score thresholds, the corresponding False Positive Rates (FPR) and True Positive Rates (TPR) with False Positives (FP), True Positives (TP), False Negatives (FN) and True Negatives (TN) according to the NMR datasets. In addition, we generated Precision-Recall (PR) curves based on Precision (PPV) and Recall (equivalent to TPR) scores. The overall prediction performance was evaluated by calculating the area under the curve for both ROC (rocAUC) and PR plots (prAUC) using trapezoidal integration.

Results

Features and Implementation of SenseNet

SenseNet reads interaction data from structure ensemble files in PDB format or MD trajectory analysis outputs generated by CPPTRAJ [60]. By default, each node corresponds to a single amino acid and edges represent interactions on the amino acid level. SenseNet automatically determines the network topology from these timelines (Fig 2A), offering different adjustment options from removing rare interactions to considering only certain interaction types. Different levels of timeline analyses are possible, as users can either scroll through single time frames to investigate e.g. network evolution or time-dependent interactions, or analyze time-averaged networks. At any point during a running session, residue level nodes and associated

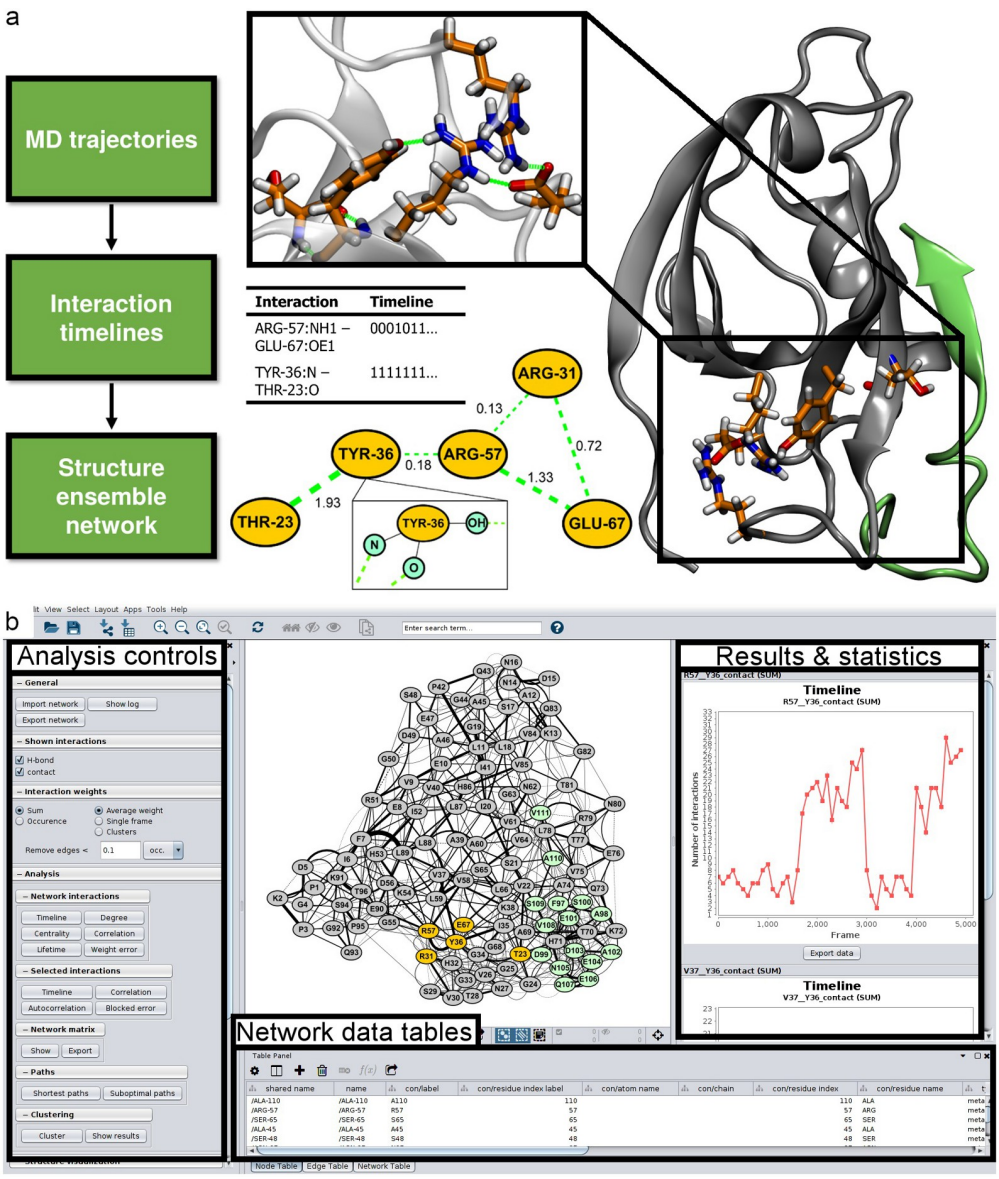

Fig 2. Example of parallel network and structure visualization using SenseNet. (a) Data representation, workflow and parallel representation of networks and molecular structures. (b) Example session showing the SenseNet GUI in Cytoscape.

https://doi.org/10.1371/journal.pone.0265194.g002

interactions can be split into individual atoms, allowing for system specific tailoring of different resolution levels. As an example application providing a detailed demonstration of this concept, we refer to our previous study analyzing the recognition of different DNA modifications by the protein UHRF1 [64]. SenseNet's user interface is separated into the main network and three control areas (Fig 2B). The left panel allows access to implemented analysis functions and displays visualization status information, such as the selected edge weighting scheme or a bar to scroll through different time frames of the network. Whenever an analysis is performed, a summary of obtained results appears on the right panel, either as tables or plots. In addition, results are written into the node and edge data tables in the bottom region, from where they can be utilized by other analysis functions, either by SenseNet or other tools. This workflow, in

combination with side-by-side network and structure visualization, allows for a rapid explorative cycle of performing quantitative analyses and intuitive exploration of the underlying structural details.

For quantitative analysis of timeline data, SenseNet offers functions for calculating timeline correlation, entropy, autocorrelation, lifetime, clustering, and network comparison. In addition, search algorithms for shortest paths as well as centrality measures are provided. Analysis results are presented as tables or plots and can be exported as raw data or images. For large scale workflows, analyses can be automated via batch script files or the CyREST interface. Network and structure visualization can be carried out in parallel by connecting SenseNet to the PyMOL [63], VMD [62], or UCSF Chimera [65] structure viewers, automatically highlighting selected nodes and edges from the network in the protein structure.

Evaluation of allosteric prediction methods using the PDZ2 domain

First, we reinvestigated the allosteric prediction performance of betweenness centralities (BC) and characteristic path length centralities (CPLC) based on networks generated from NMR and crystal structures, which had previously shown poor prediction performance for the PDZ2 system with CPLC as the best performing centrality model [10]. This allowed us to verify our implementation and to compare different network methods based on the same dataset. In line with the aforementioned work, we determined ROC and PR curves measuring the prediction accuracy of tested models with respect to the NMR dataset, which is composed of allosteric and non-allosteric residues based on methyl sidechain dynamics, and the corresponding NMR-Ala dataset variant excluding alanines [10,39] (S1 Table). In an attempt to replicate the network centrality predictions from Cilia et al. (NMR: 0.54, NMR-Ala: 0.59) [10], we calculated CPLC scores based on the crystal and NMR structures of the PDZ2-RA-GEF-2 complex using a carbon contact distance cutoff of 5 Å. For the NMR structure, resulting rocAUC scores were very close to the previously reported values (NMR: 0.55, NMR-Ala: 0.56) and only modestly higher for the crystal structure (NMR: 0.65, NMR-Ala: 0.69), indicating that the differences are only due to subtly differing details in network implementations.

In contrast to the centrality approach, interaction timelines generated from structure ensembles allow to additionally analyze the correlation between interactions, as quantified by the NCF and DNCF scores (see Materials & Methods). In general, residues with high NCF scores provide information, through linear and nonlinear correlation, about the interaction state of their environment. While the NCF estimates the information of residues within a single simulation, the DNCF score models the corresponding differences between two simulations, e.g. with and without a ligand. In order to obtain the structure ensembles necessary for calculation of these scores, we performed ten 1 μs MD simulations of the free PDZ2 domain and the PDZ2-RA-GEF-2 peptide complex. Timelines of contacts and hydrogen bonds were extracted and converted into protein structure networks using AIFgen and analyzed using SenseNet. First, we systematically evaluated all compared network methods (BC, CPLC, NCF, DNCF) using a grid search of 48 parameter combinations (S2 Table). These combinations were obtained by varying the contact distance cutoff from 4 to 9 Å, the interaction subset settings (all or only inter-sidechain interactions), and networks generated from different sources (apo- or peptide-bound structures; NMR or crystal structures). To understand which parameters are most important for prediction performance, we grouped all data points according to these categories followed by analysis of the obtained rocAUC score distributions. In the following, we focus predominantly on the results obtained for the NMR-Ala dataset, as alanine residues proved to be particularly difficult to predict for all methods tested here as well as those previously published. Fig 3A shows that average rocAUC scores over all combinations were

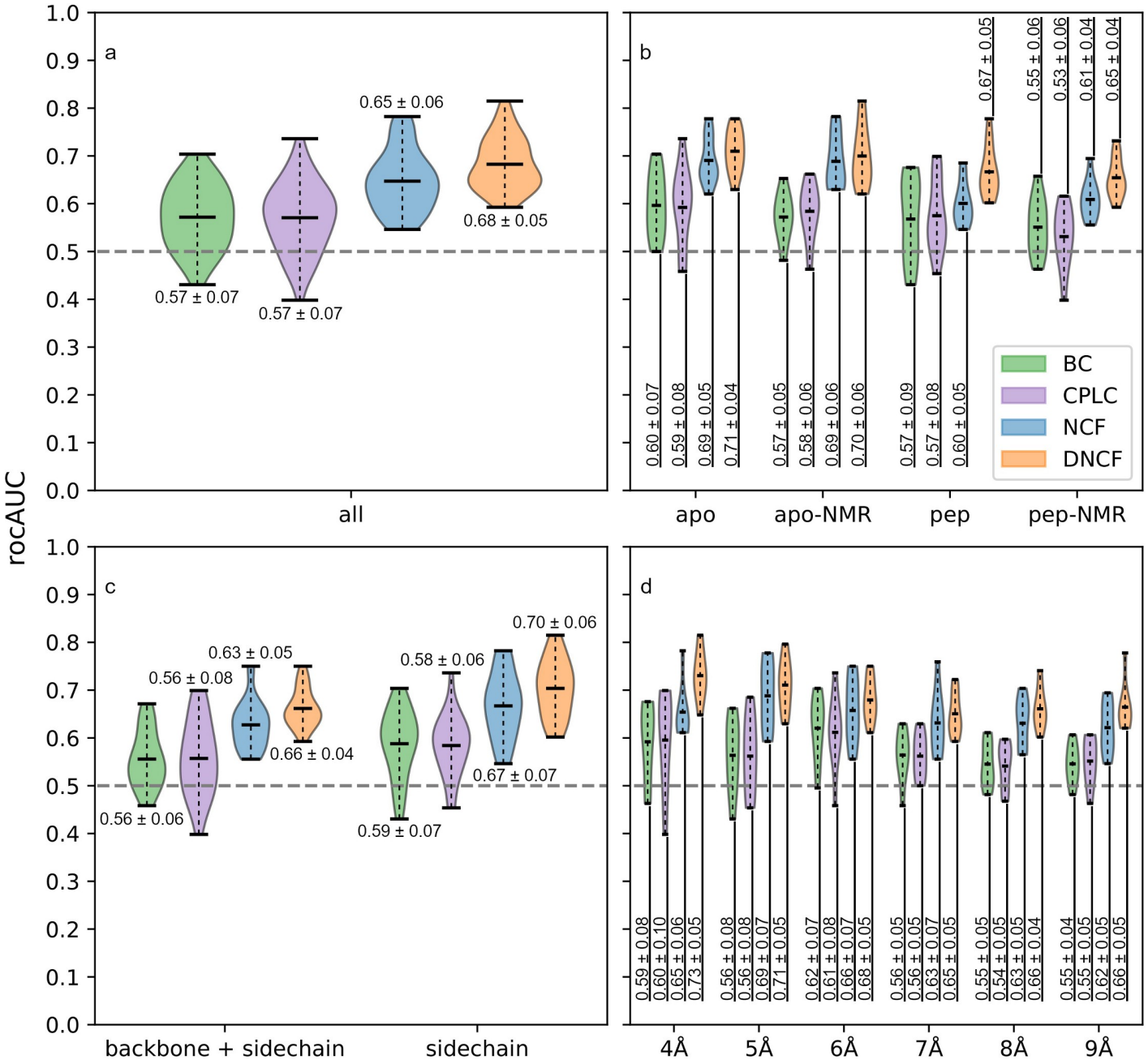

Fig 3. Influence of network parameters on prediction model performance based on the NMR-Ala reference set. Shaded areas show distribution estimates based on a gaussian kernel with added labels for mean and standard deviation. (a) Distributions including all parameter combinations. (b) Source of analyzed network data: Crystal structures (apo, pep) or NMR based structures (apo-NMR, pep-NMR). (c) Interaction subset: All interactions or sidechain-exclusive networks. (d) Distance cutoff for carbon-carbon contacts in the network.

https://doi.org/10.1371/journal.pone.0265194.g003

consistently highest for the DNCF method, followed by NCF and finally CPLC and BC, which registered 8–11% lower average AUC scores compared to the former methods. In a more detailed view (Fig 3B), we observed that on average, prediction performances improved if apo PDZ2 was used as starting structure compared to peptide bound systems, with relatively small differences for CPLC, BC, and DNCF (up to 5%), but more substantial improvements for NCF

(up to 9%). Interestingly, the NCF prediction performance based on the apo systems was almost as high as the DNCF scores although, in contrast to DNCF, they do not contain any information about the ligand. Regarding the set of included interactions in the network (Fig 3C), rocAUC scores increased on average by 2–4% if only inter-sidechain interactions were considered. Finally, analysis of contact cutoff distances shows that BC and CPLC method performances appear to peak at 6 Å, whereas a 4 to 5 Å cutoff worked best for the DNCF and NCF methods (Fig 3D). Observing the shape of rocAUC distributions and the lower performance limit for worst-case parameters can give an indication about the sensitivity of a method to choosing inappropriate network parameters. For BC and CPLC methods, several parameter combinations led to essentially random prediction performance (rocAUC ~ 0.5) (Fig 3), indicating a high sensitivity to parameter choices in order to achieve good accuracy. In contrast, NCF and even more so DNCF were consistently more robust, as they showed better performances even for suboptimal parameters over all categories (Fig 3). Many of the observed trends are reflected, to a lesser degree, on the full NMR reference set which includes alanine residues (S1 Fig). In conclusion, we first observe that all parameter categories follow consistent trends, highlighting the importance of parameter choice for prediction quality, which is particularly true for methods based on centrality. Second, this consistency is also observed if the different methods are compared, i.e. the favorable performances of NCF and DNCF models relative to centralities are reflected throughout all parameter settings.

The best performing CPLC model was obtained for the apo PDZ2 crystal structure and a carbon contact cutoff of 6 Å in a sidechain exclusive network, interestingly differing from the original evaluation discussed above (5 Å and including backbone interactions) [10]. Using the optimized parameters, the rocAUC score for the NMR-Ala dataset increased by 5% to 0.74, while performance for the NMR dataset degraded by 1% to 0.64, respectively (Table 1). The corresponding prAUC scores increased by 2% for the NMR dataset (0.75 to 0.77) and 5% for NMR-Ala (0.78 to 0.83). The BC method performed optimally with the same parameter set as CPLC, but with about 3 to 4% lower rocAUC scores (Table 1). Overall, only modest performance improvements could be achieved for the BC and CPLC methods by variation of network parameters.

For both DNCF and NCF models, the optimal parameter set consisted of a 4 Å contact cutoff in a sidechain exclusive network using simulations of the apo-NMR PDZ2 structure. Of all settings tested in the parameter search, DNCF was found to be the best overall predictor, achieving a rocAUC of 0.71 and prAUC of 0.82 on the full NMR set, which corresponds to a 5 to 7% improvement compared to the CPLC model. Accordingly, the performance on the NMR-Ala set was also higher than for the centrality methods with a rocAUC of 0.81 and a prAUC of 0.88. The best NCF model showed similar overall trends, but individual AUC scores were 1–5% lower (Table 1). In line with most published methods, rocAUC scores were

Table 1. Allosteric prediction performance of network-based models.

Reference set	Method	rocAUC	prAUC
NMR	NCF	0.66	0.79
NMR	DNCF	0.71	0.82
NMR	BC	0.61	0.74
NMR	CPLC	0.64	0.77
NMR-Ala	NCF	0.78	0.86
NMR-Ala	DNCF	0.81	0.88
NMR-Ala	BC	0.70	0.80
NMR-Ala	CPLC	0.74	0.83

https://doi.org/10.1371/journal.pone.0265194.t001

consistently 7–10% lower for the NMR dataset compared to NMR-Ala, which highlights the general difficulty for predicting this residue type (Table 2).

In order to obtain sufficient statistical sampling for the determination of optimal model parameters, we performed a total of 10 μs of simulations, which constitutes an increasingly common but still substantial computational effort at this time for a system the size of PDZ2. While such an effort is justified for evaluation studies, for practical and effective application a guideline as to what amounts to a reasonable simulation time should be established. To gain a rough estimate of this and the convergence of our model, we repeated our analysis using the DNCF model with optimal parameters, but with truncated trajectories for each replica. The first analysis was performed on trajectories shortened to contain only the first 100 ns (after removing the initial 100 ns to reduce replica bias towards the initial structure, as detailed above), yielding a cumulative simulation time of 1 μs (10 x 100 ns). Then, subsequent analyses were performed on the first 200 ns yielding a cumulative time of 2 μs, then 300 ns for 3 μs, and so on. This approach was chosen since it shows directly how our results would have changed had we chosen a shorter simulation time for our analysis. The obtained DNCF scores were compared to the NMR-Ala and NMR datasets and rocAUC and prAUC calculated accordingly (Fig 4A and 4C). These data indicate an improvement of prediction performance up until about 3 μs of cumulative simulation time, and remaining approximately constant past that point. Taking those 3 μs as the target time, we proceeded to determine whether it was more beneficial to use fewer replicas with longer individual simulations, or to use more replicas in combination with shorter simulation times. Thus, we compared predictions using between four and ten replicas, taking the appropriate amount of simulation frames from each replica to reach a total simulation time of 3 μs. For example, when using four replicas, each replica trajectory contributed 0.75 μs (total 3 μs from 4 x 0.75 μs), whereas for five replicas each contributed 0.6 μs, and so on. This analysis was performed for each possible combination of replicas, e.g. for four replicas we considered all ways to pick four replicas out of the total of ten replicas. Judging from both the means and standard deviations of rocAUC/prAUC results (Fig 4B and 4D), it is clearly beneficial to use up to 8 replicas, corresponding to 8 replica simulations of 375 ns each, to obtain a cumulative simulation time of 3 μs. With only two data points following after, it is unclear whether this trend would persist further, though we do not expect substantial improvements considering that the values observed at 9 and 10 replicas seem to indicate that a plateau was reached. Based on the totality of the data, we conclude that our DNCF model is adequately converged for the purpose of this study. It should be noted that our analysis constitutes a very rough estimate that is specifically limited to the PDZ2 system, whose allostery does not involve substantial conformational changes.

It has been pointed out that the allosteric residue sets from published computational predictions differ substantially for the PDZ2 system [41], fueling our interest determining how well

Table 2. Comparison of DNCF prediction performance with other published computational methods.

Reference set	Method	rocAUC	prAUC
NMR	DNCF	0.71	0.82
NMR	NMR/MC	0.74	0.82
NMR	RRS	0.65	0.75
NMR	REDAN	0.67	0.65
NMR-Ala	DNCF	0.81	0.88
NMR-Ala	NMR/MC	0.81	0.87
NMR-Ala	RRS	0.72	0.80
NMR-Ala	REDAN	0.62	0.61

https://doi.org/10.1371/journal.pone.0265194.t002

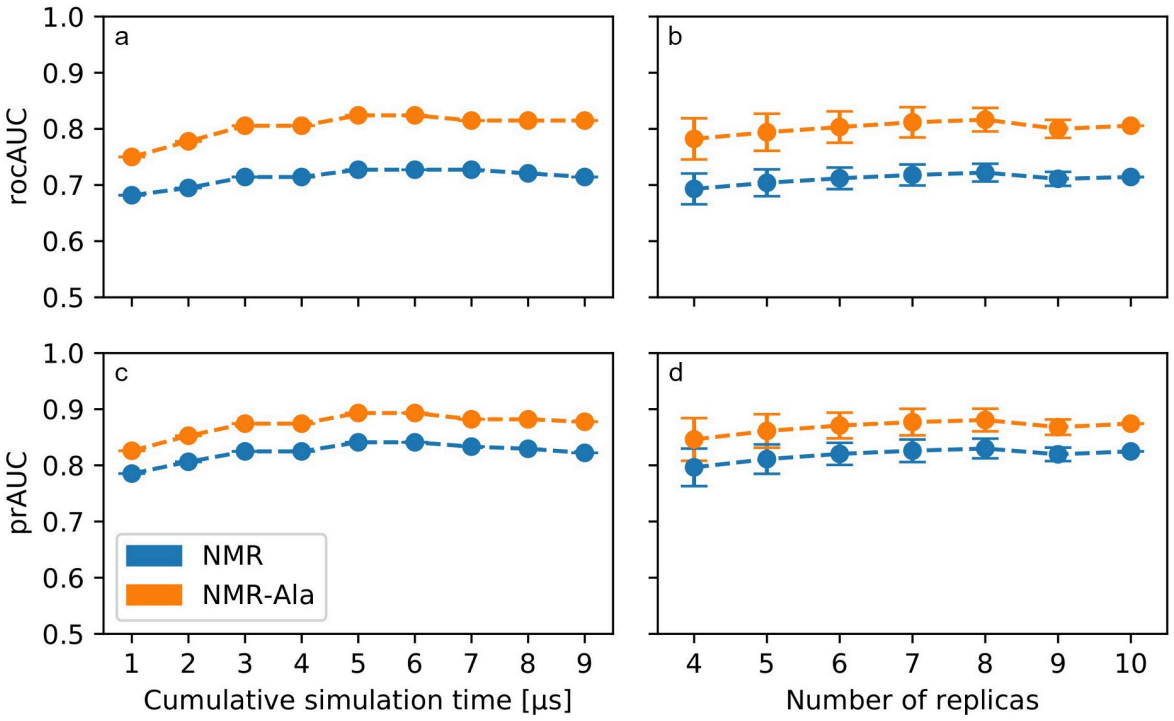

Fig 4. Effect of simulation time and number of replicas on prediction performance of the final DNCF model. (a,c) Timelines of all ten replicas were truncated, merged to the specified cumulative simulation time and analyzed successively. 1 μs of cumulative simulation time corresponds to a simulation time of 100 ns per replica (10 x 100 ns) after equilibration. (b,d) Cumulative simulation time of 3 μs was obtained from combining the appropriate amount for frames from the specified number of replicas. In the case of four replicas, each replica trajectory contributed 0.75 μs (total 3 μs from 4 x 0.75 μs), for five replicas each contributed 0.6 μs, and so on. Circles and bar handles represent the mean and standard deviation calculated over all possible replica combinations.

https://doi.org/10.1371/journal.pone.0265194.g004

these models agree with the NMR datasets. However, comparing models based on binary classifications alone can be misleading, since each classification relies on an implicit sensitivity threshold which might differ drastically between models. ROC and PR curves are more suitable for this task since they evaluate prediction performances at all possible thresholds, but require raw prediction scores, which are not always available. Fig 5 shows the ROC and PR curves for the models described above and those for which accompanying literature included the necessary scores. We observed comparably high performances for the DNCF and NMR/MC [10] models (Table 2, differences within 1–2%), followed by RRS [50] and REDAN [46]. As the NMR/MC model requires NMR structure data, the DNCF method offers a substantial advantage as the necessary simulations can be based on much more commonly available crystal structures. Thus, although these two methods show comparable accuracy, we expect that the DNCF method can applied to a wider range of systems. We also believe that the method has the potential to show improved results for systems for which induced fit phenomena are important, i.e. for which the conformational ensembles of the apo- and holo-structures differ considerably.

Application of allosteric predictions to the PDZ2 domain

Having established good agreement between DNCF scores and allosteric residues, we investigated the usefulness of these additional features for the biochemical interpretation of our predictions in the PDZ2 structure. Integrating the DNCF scores of the model described above into the structure network (Fig 6A and 6B) reveals two high scoring clusters of residues

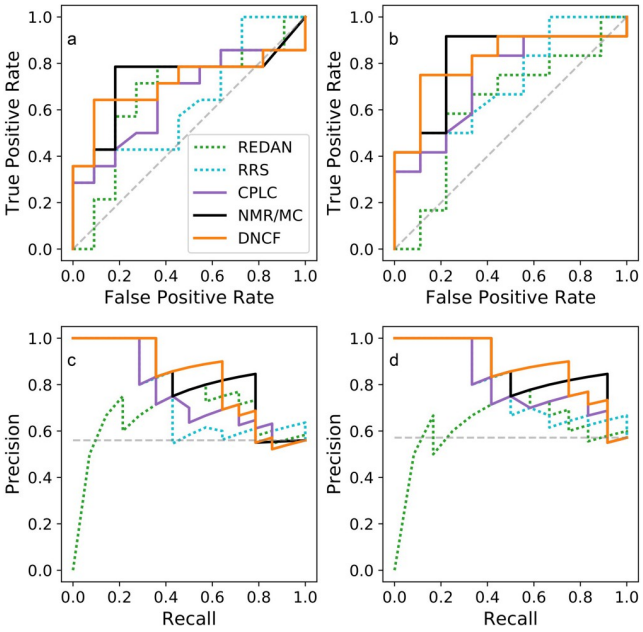

Fig 5. ROC and PR curves of selected prediction models. (a) ROC curve based on the NMR reference set. (b) ROC curve based on the NMR-Ala reference set. (c) PR curve based on the NMR reference set. (d) PR curve based on the NMR-Ala reference set.

https://doi.org/10.1371/journal.pone.0265194.g005

(clusters I and II). The majority of allosteric residues of the NMR dataset are located in cluster I, which stretches from the top region of the binding pocket towards helix α1 and sheet β1 (Fig 6B–6D). On the other hand, cluster II encompasses the lower part of the binding pocket surrounding the flexible loop L1 (residues 24–33), including the allosteric residues V26 and V30, furthermore its interaction partners R57, Y36, and finally the C-terminal region.

Comparing these observations to other network scoring methods, the NCF model shows a very similar cluster structure (Fig 7A), whereas for CPLC we observed increased scores for residues located next to the peptide binding groove, e.g. V22, L66, H71, A74, V75 and L78 (Fig 7B–7D). This can be explained directly by the definition of CPLC (see Algorithms section), which attributes high scores to residues bridging structural modules, e.g. binding grooves. On the other hand, centrality scores for loop L1 (specifically residues 30 to 32) in cluster II are substantially lower than in the timeline-based NCF and DNCF methods, which might be explained by the difficulties of a single structure network to represent the switching contacts of flexible regions. This indicates that centrality methods may fail to account for regions with intrinsic flexibility like the L1 loop, for which methods based on structure ensembles are potentially more appropriate.

Consensus model of allosteric information flow in PDZ2

Finally, we defined a new consensus model of allosteric information flow consolidating our and previous prediction models. For this we first determined a "consensus set" composed of residues predicted as allosteric in $\geq$ 50% from a selection of published studies (S3 Table) [10,42–45,47–51,66]. Next, we obtained a core set of allosteric candidates from our DNCF model, using the score threshold closest to the top left corner in Fig 5B (6.17 bits in S4 Table; TPR: 0.75; FPR: 0.11). This core prediction set (Fig 8 and S5 Table) contains 9 out of 14 residues from the NMR dataset and 11 of the 18 from the consensus set, while 14 residues are

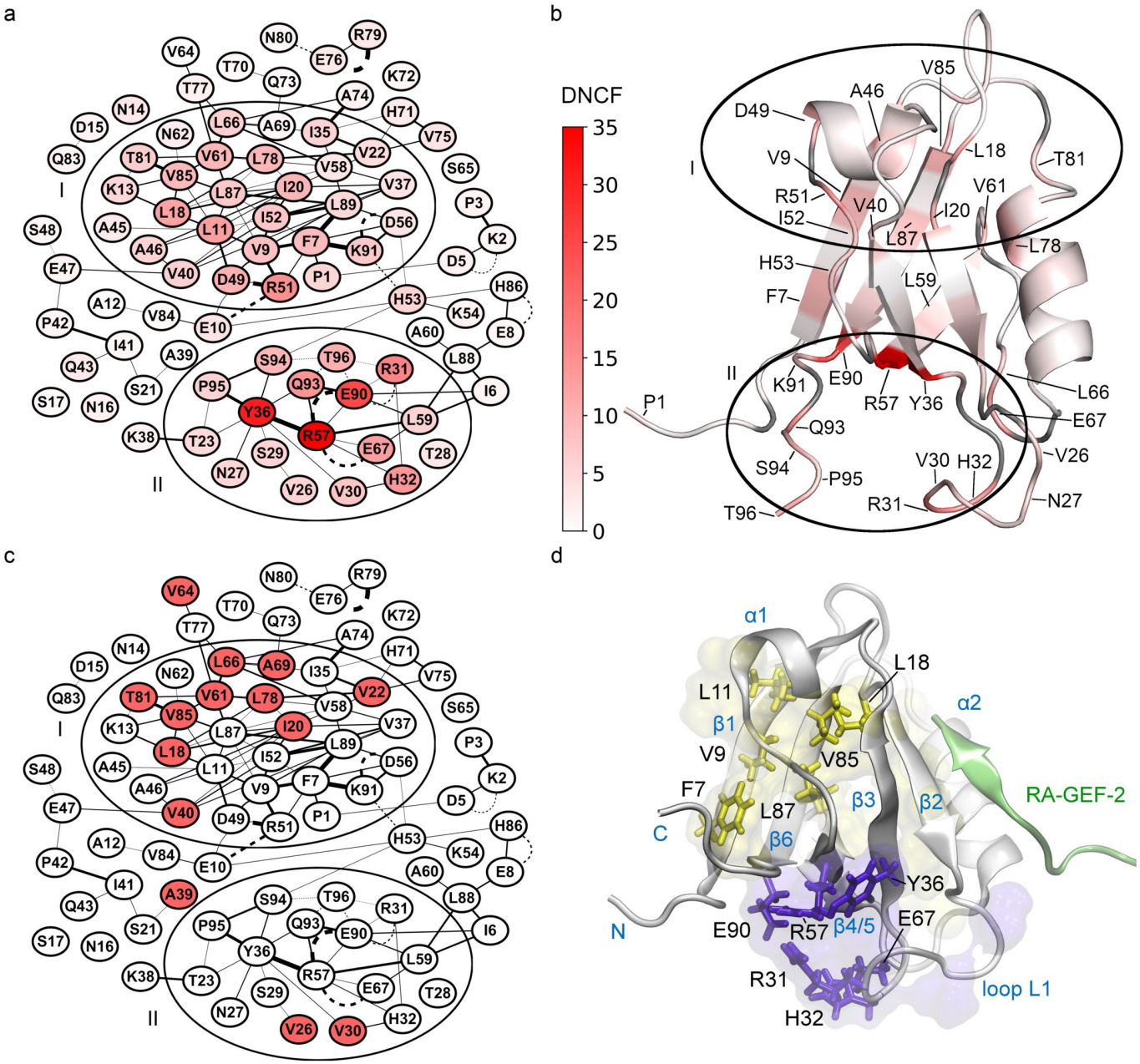

Fig 6. Allosteric predictions of the final DNCF model mapped to PDZ2 structures. For visual clarity, only edges occurring in ≥0.1% of simulation time are shown. (a) Network representation of DNCF predictions. Nodes are colored from low (white) to high (red) DNCF scores. (b) DNCF scores mapped to the apo PDZ2 structure (PDB-ID: 3PDZ). (c) Network showing experimentally determined allosteric residues (red) from the NMR dataset. (d) Allosteric clusters mapped to the RA-GEF-2 bound PDZ2 structure (PDB-ID: 1D5G): Cluster I (yellow surface) and Cluster II (purple surface). Specific residues discussed in the text are additionally shown as sticks.

https://doi.org/10.1371/journal.pone.0265194.g006

complementary predictions. Of these infrequently predicted residues, three form a contiguous surface located on the sheet β1 (F7, V9, L11), connected via L18, V85, and L87 to the peptide binding pocket (Fig 6D). In NMR experiments, V9 was shown to respond to the binding pocket I20F mutation with L11 and L87 as presumed linker residues [40], an interpretation supported by our model. Notably, the clusters surrounding V9 and Y36 agree very well with the DS3 and DS4 regions described previously [10]. Predictions of the C-terminal tail residues

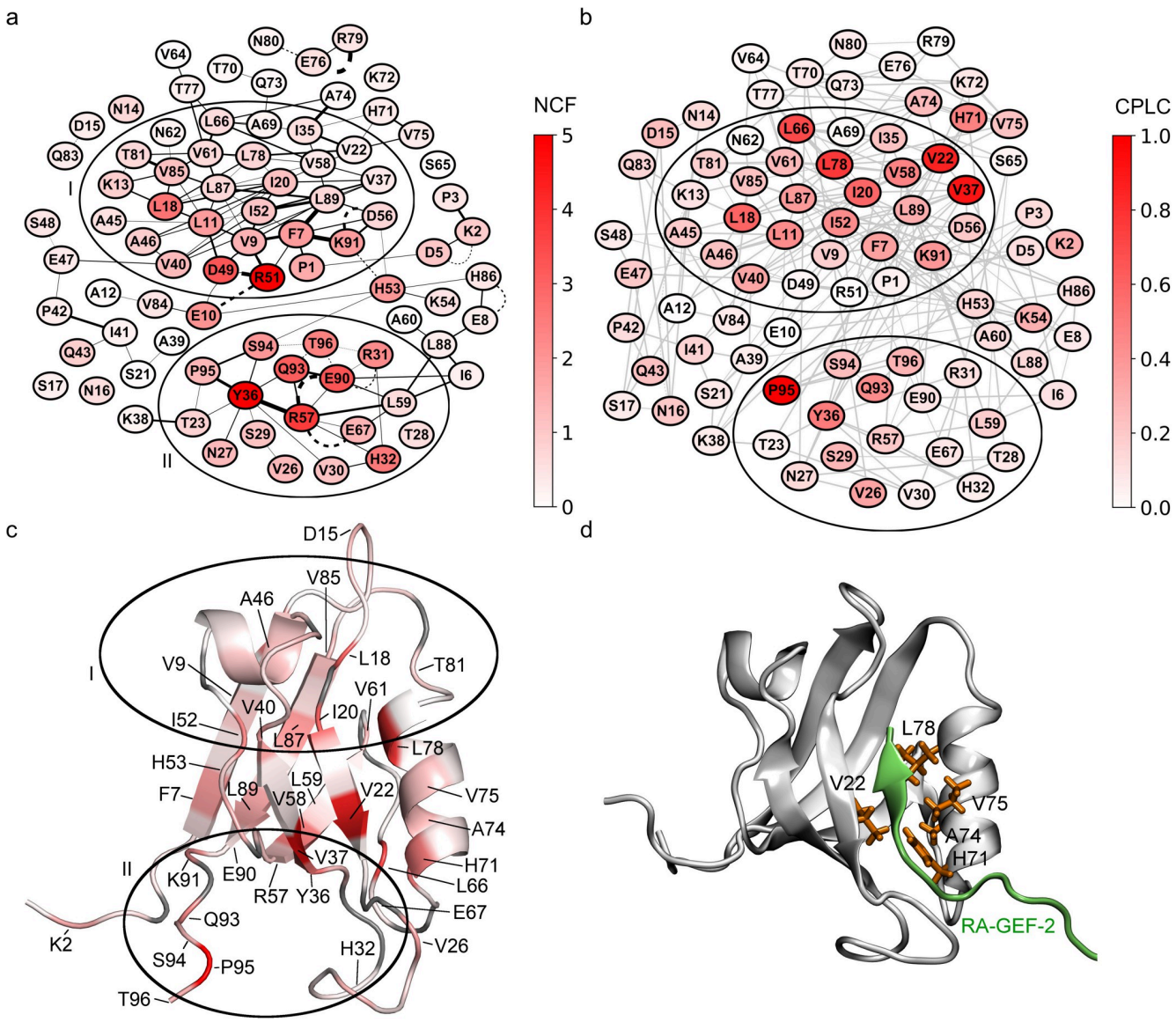

Fig 7. Allosteric predictions of the final NCF and CPLC models mapped to PDZ2 structures. Nodes colored from low (white) to high (red) scores. (a) Network representation of NCF predictions. For visual clarity, only edges occurring in ≥0.1% of simulation time are shown. (b) Network representation of CPLC predictions. Edge colors are shown in light grey to increase clarity. (c) CPLC scores mapped to the apo PDZ2 structure (PDB-ID: 3PDZ). (d) Notable residues predicted by CPLC mapped to the RA-GEF-2 bound PDZ2 structure (PDB-ID: 1D5G).

https://doi.org/10.1371/journal.pone.0265194.g007

(93 to 96) are difficult to assess as the high flexibility of free chain termini might not properly represent the common biological state, i.e. PDZ2 embedded in a multi-domain protein. Previous studies have formulated the idea of up to four separate distal sites (DS1—DS4) identified by following the interconnected surfaces of allosteric residues [10,39,66]. Our results suggest the existence of at least two allosteric clusters: Cluster I which encompasses DS1, DS2, and DS3, while cluster II corresponds to DS4.

Discussion

Integration of interaction timelines from molecular dynamics simulations into protein structure networks provides a promising framework for investigating dynamic effects in proteins

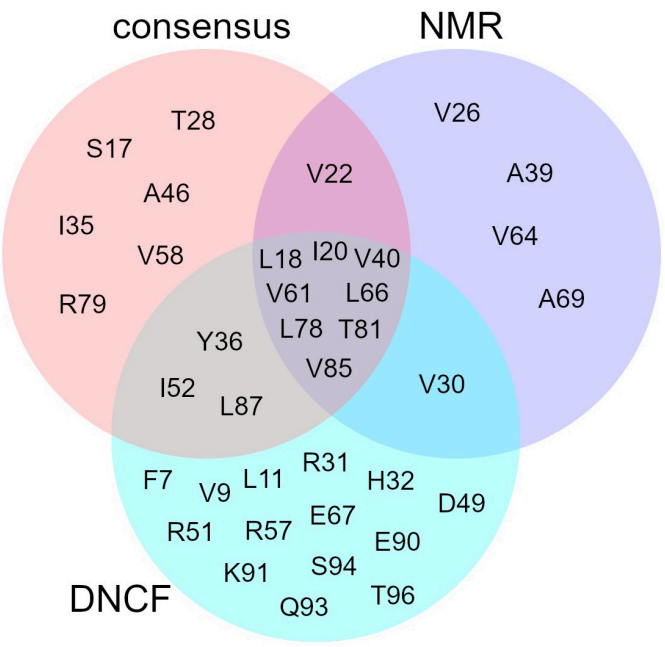

Fig 8. Intersection of the DNCF allosteric core set, NMR reference set, and the computational prediction consensus set.

https://doi.org/10.1371/journal.pone.0265194.g008

such as allostery. In this work, we introduce our network analysis tool SenseNet which builds on this theoretical foundation. Using the PDZ2 domain as a reference system, we evaluated four allosteric prediction models implemented in SenseNet, i.e. BC, CPLC, NCF and DNCF, and determined a set of network parameters optimizing their accuracy. Our results are consistent with literature data, as structure networks frequently use carbon contact cutoff distances between 4–6 Å [10,19,47,67,68], which corresponds approximately to the upper limit of attractive Van-der-Waals interactions. The trend for better prediction results using apo protein states might reflect the observed rigidification of the ligand binding site after binding [39] and is in line with previous suggestions that allosteric mechanisms may be intrinsic properties of apo structures [42,69]. Finally, the improvements observed in sidechain exclusive networks mirror the origins of the NMR dataset, which was obtained from methyl sidechain dynamics [39]. This also highlights an important caveat for comparing prediction models, as some methods might by design match certain types of experimental data more closely than others. Methods based on interaction timelines, i.e. NCF and DNCF, were consistently more accurate than the BC and CPLC methods based on network centrality. This highlights the benefits of using MD simulations to include protein dynamics in protein structure networks, which is achieved by application of methods utilizing interaction timelines. In contrast, centrality-based methods offer the advantage of requiring only a single structure, which makes them uniquely inexpensive in a situation where MD simulations are not feasible. Our data indicate that both BC and CPLC methods could achieve good prediction performances, but were sensitive to the choice of parameters used for network construction. For these in particular, further evaluation studies spanning multiple systems are needed to determine an optimal parameter set that performs well in a wide range of proteins. Of the methods tested, DNCF proved to be the most accurate and robust to changes in network parameters, followed by NCF. This reflects the DNCF method's ability to capture effects from two simulations representing different system states by comparing the changes in shared information. However, the NCF method appears to have

potential on its own for predictions based on apo structures alone, for example when there is no known structure of the investigated protein bound to the allosteric ligand.

The final allosteric model, based on the DNCF method, was found to be one of the models aligning most closely to experimental data out of those reported in literature, alongside NMR/MC. However, the DNCF approach offers three distinct advantages to NMR/MC: First, MD simulations for DNCF analyses can be started from only a single, e.g. X-ray, structure, while NMR/MC needs an NMR structure ensemble, which are far rarer and more limited to small proteins. Second, the DNCF method includes all residue types, while NMR/MC by definition cannot predict alanine residues. Third, the DNCF method has the potential to detect induced fit-based conformational changes, which are often not directly detectable in the structural ensembles of the apo-state alone. We determined that 3 μs of total simulation time, spread across 8 replicas and corresponding to 375 ns of simulation for each replica, approximated optimal prediction performance using the DNCF method in the PDZ2 system. These numbers are likely specific to the protein system under investigation and thus can only serve as a guideline for proteins of comparable size and with allosteric effects in the absence of large conformational changes. It should be noted, that fewer replicas and shorter simulation times could still achieve solid performance, which may be relevant when investigating larger proteins for which generating a comparable amount of simulation data may be infeasible. In these cases, additional validation with experimental data is indicated. Our numbers are in agreement with a previous study investigating the reproducibility between replicas in a 10 residue system as well as a 827 residue TCR-p-MHC complex, which recommended using between 5 to 10 replicas for simulations as a rule of thumb [70].

Mapping the results of our DNCF model to the structure of PDZ2 suggests the protein contains two distinct allosteric sites. Most of the experimentally verified allosteric residues from the NMR dataset are located in cluster I, while cluster II has little support from the experimental dataset as the region encompasses only four residues with methyl groups. To fill this gap, alternative experiments may be necessary such as mutational studies connected to changes in PDZ mediated activation. The locations of our observed clusters are matched by several other computational predictions [42,43,45]. Nevertheless, our data contrasts with studies reporting up to four distinct allosteric sites [10,39,66] by suggesting that these four sites are partially overlapping, leaving only two clearly separated allosteric regions. The variance in published allosteric predictions in the PDZ2 domain may be explained by the fact that the experimentally verified data in a single protein are naturally sparse, leading to potentially large error margins for validation. In addition, for many cases quantitative scores are not reported along binary classifications, impeding direct comparison of predictions. To improve prediction models, large scale studies including multiple proteins, computational methods, and experimental data sources will be necessary. With SenseNet we provide a network analysis tool offering considerable advantages over existing implementations: First, by defining edges via interaction timelines, all conformational states of a simulation are readily available for analysis, which is not possible if interactions are reduced to correlation coefficients. Second, adopting a multi-resolution approach via mapping of sub-structures of varying sizes to nodes (from atoms to residues) allows the creation of application-specific network topologies that reduce the underlying structural differences to the most informative level of details. Finally, integration of our tool into Cytoscape allows users to complement their analyses with the community driven ecosystem of biological network analysis plugins, e.g. by connecting structural analysis with system biological or sequence/evolutionary information. Based on these concepts, SenseNet provides an analysis platform implementing a range of well tested analysis algorithms, an easy-to-use UI driven implementation, and interactive side-by-side structure visualization. Together, these features serve as a potential foundation for wide application of timeline-based protein

structure networks, paving the way for comparative studies to improve model accuracies and aid experiments in unveiling detailed mechanisms of dynamic processes in biomolecules.

Supporting information

S1 Fig. Influence of network parameters on prediction model performance based on the NMR reference set. Shaded areas show distribution estimates based on a gaussian kernel with added labels for mean and standard deviation. (a) Distributions including all parameter combinations. (b) Source of analyzed network data: Crystal structures (apo, pep) or NMR based structures (apo-NMR, pep-NMR). (c) Interaction subset: All interactions or sidechain-exclusive networks. (d) Distance cutoff for carbon-carbon contacts in the network.
(TIF)

S1 Table. NMR reference set of experimentally verified allosteric and non-allosteric residues. Allosteric residues are represented by a value of 1, non-allosteric residues by a value of 0.
(XLSX)

S2 Table. Prediction model performances for all tested network parameter combinations.
(XLSX)

S3 Table. Computational predictions of allosteric residues including the DNCF model and previously published methods.
(XLSX)

S4 Table. Residue scores of final DNCF, NCF, and CPLC models.
(XLSX)

S5 Table. Comparison of the DNCF allosteric core set with the NMR reference and computational prediction consensus sets.
(XLSX)

S1 File. Initial structures, topologies, and input files for molecular dynamics simulations.
(ZIP)

S2 File. Scripts demonstrating an example workflow for the AIFgen tool.
(ZIP)

Acknowledgments

We thank Martin Zacharias for his constructive feedback during revision of the manuscript. This work is dedicated to the memory of Iris Antes, who passed away unexpectedly on the 4th of August 2021.

Author Contributions

Conceptualization: Markus Schneider, Iris Antes.

Data curation: Markus Schneider.

Formal analysis: Markus Schneider.

Funding acquisition: Iris Antes.

Investigation: Markus Schneider.

Methodology: Markus Schneider.

Project administration: Iris Antes.

Software: Markus Schneider.

Supervision: Iris Antes.

Validation: Markus Schneider.

Visualization: Markus Schneider.

Writing – original draft: Markus Schneider.

Writing – review & editing: Markus Schneider, Iris Antes.

References

1. O'Rourke KF, Gorman SD, Boehr DD. Biophysical and computational methods to analyze amino acid interaction networks in proteins. Comput Struct Biotechnol J. 2016; 14:245–51. https://doi.org/10.1016/j.csbj.2016.06.002 PMID: 27441044
2. Greene LH. Protein structure networks. Briefings in Functional Genomics. 2012; 11:469–78. https://doi.org/10.1093/bfgp/els039 PMID: 23042823
3. Di Paola L, Giuliani A. Protein contact network topology: a natural language for allostery. Current Opinion in Structural Biology. 2015; 31:43–8. https://doi.org/10.1016/j.sbi.2015.03.001 PMID: 25796032
4. Changeux JP. 50 years of allosteric interactions: the twists and turns of the models. Nat Rev Mol Cell Biol. 2013; 14(12):819–29. https://doi.org/10.1038/nrm3695 PMID: 24150612
5. Nussinov R, Tsai C-J. Allostery without a conformational change? Revisiting the paradigm. Current Opinion in Structural Biology. 2015; 30:17–24. https://doi.org/10.1016/j.sbi.2014.11.005 PMID: 25500675
6. Tsai C-J, Nussinov R. A Unified View of "How Allostery Works". PLoS Computational Biology. 2014; 10: e1003394. https://doi.org/10.1371/journal.pcbi.1003394 PMID: 24516370
7. Lu S, Li S, Zhang J. Harnessing allostery: a novel approach to drug discovery. Med Res Rev. 2014; 34 (6):1242–85. https://doi.org/10.1002/med.21317 PMID: 24827416
8. Nussinov R, Tsai CJ. Allostery in disease and in drug discovery. Cell. 2013; 153(2):293–305. https://doi.org/10.1016/j.cell.2013.03.034 PMID: 23582321
9. del Sol A, Fujihashi H, Amoros D, Nussinov R. Residues crucial for maintaining short paths in network communication mediate signaling in proteins. Mol Syst Biol. 2006; 2:2006 0019. https://doi.org/10.1038/msb4100063 PMID: 16738564
10. Cilia E, Vuister GW, Lenaerts T. Accurate prediction of the dynamical changes within the second PDZ domain of PTP1e. PLoS Comput Biol. 2012; 8(11):e1002794. https://doi.org/10.1371/journal.pcbi.1002794 PMID: 23209399
11. Popovych N, Sun S, Ebright RH, Kalodimos CG. Dynamically driven protein allostery. Nat Struct Mol Biol. 2006; 13(9):831–8. https://doi.org/10.1038/nsmb1132 PMID: 16906160
12. Schrank TP, Bolen DW, Hilser VJ. Rational modulation of conformational fluctuations in adenylate kinase reveals a local unfolding mechanism for allostery and functional adaptation in proteins. Proc Natl Acad Sci U S A. 2009; 106(40):16984–9. https://doi.org/10.1073/pnas.0906510106 PMID: 19805185
13. Motlagh HN, Wrabl JO, Li J, Hilser VJ. The ensemble nature of allostery. Nature. 2014; 508(7496):331–9. https://doi.org/10.1038/nature13001 PMID: 24740064
14. Feher VA, Durrant JD, Van Wart AT, Amaro RE. Computational approaches to mapping allosteric pathways. Curr Opin Struct Biol. 2014; 25:98–103. https://doi.org/10.1016/j.sbi.2014.02.004 PMID: 24667124
15. Greener JG, Sternberg MJ. Structure-based prediction of protein allostery. Curr Opin Struct Biol. 2018; 50:1–8. https://doi.org/10.1016/j.sbi.2017.10.002 PMID: 29080471
16. Hertig S, Latorraca NR, Dror RO. Revealing Atomic-Level Mechanisms of Protein Allostery with Molecular Dynamics Simulations. PLoS Comput Biol. 2016; 12(6):e1004746. https://doi.org/10.1371/journal.pcbi.1004746 PMID: 27285999
17. Wagner JR, Lee CT, Durrant JD, Malmstrom RD, Feher VA, Amaro RE. Emerging Computational Methods for the Rational Discovery of Allosteric Drugs. Chem Rev. 2016; 116(11):6370–90. https://doi.org/10.1021/acs.chemrev.5b00631 PMID: 27074285
18. Guo J, Zhou HX. Protein Allostery and Conformational Dynamics. Chem Rev. 2016; 116(11):6503–15. https://doi.org/10.1021/acs.chemrev.5b00590 PMID: 26876046

19. Daily MD, Gray JJ. Local motions in a benchmark of allosteric proteins. Proteins. 2007; 67(2):385–99. https://doi.org/10.1002/prot.21300 PMID: 17295319

20. Cooper A, Dryden DTF. Allostery without conformational change. European Biophysics Journal. 1984; 11(2):103–9. https://doi.org/10.1007/BF00276625 PMID: 6544679

21. Pasi M, Tiberti M, Arrigoni A, Papaleo E. xPyder: a PyMOL plugin to analyze coupled residues and their networks in protein structures. J Chem Inf Model. 2012; 52(7):1865–74. https://doi.org/10.1021/ci300213c PMID: 22721491

22. Tiberti M, Invernizzi G, Lambrughi M, Inbar Y, Schreiber G, Papaleo E. PyInteraph: a framework for the analysis of interaction networks in structural ensembles of proteins. J Chem Inf Model. 2014; 54 (5):1537–51. https://doi.org/10.1021/ci400639r PMID: 24702124

23. Brown DK, Penkler DL, Sheik Amamuddy O, Ross C, Atilgan AR, Atilgan C, et al. MD-TASK: a software suite for analyzing molecular dynamics trajectories. Bioinformatics (Oxford, England). 2017; 33 (17):2768–71. https://doi.org/10.1093/bioinformatics/btx349 PMID: 28575169

24. Sercinoglu O, Ozbek P. gRINN: a tool for calculation of residue interaction energies and protein energy network analysis of molecular dynamics simulations. Nucleic Acids Res. 2018; 46(W1):W554–W62. https://doi.org/10.1093/nar/gky381 PMID: 29800260

25. Bhattacharyya M, Bhat CR, Vishveshwara S. An automated approach to network features of protein structure ensembles. Protein science: a publication of the Protein Society. 2013; 22(10):1399–416. https://doi.org/10.1002/pro.2333 PMID: 23934896

26. Chakrabarty B, Parekh N. NAPS: Network Analysis of Protein Structures. Nucleic Acids Res. 2016; 44 (W1):W375–82. https://doi.org/10.1093/nar/gkw383 PMID: 27151201

27. Chakrabarty B, Naganathan V, Garg K, Agarwal Y, Parekh N. NAPS update: network analysis of molecular dynamics data and protein-nucleic acid complexes. Nucleic Acids Res. 2019. https://doi.org/10.1093/nar/gkz399 PMID: 31106363

28. Contreras-Riquelme S, Garate JA, Perez-Acle T, Martin AJM. RIP-MD: a tool to study residue interaction networks in protein molecular dynamics. PeerJ. 2018; 6:e5998. https://doi.org/10.7717/peerj.5998 PMID: 30568854

29. Grant BJ, Rodrigues AP, ElSawy KM, McCammon JA, Caves LS. Bio3d: an R package for the comparative analysis of protein structures. Bioinformatics. 2006; 22(21):2695–6. https://doi.org/10.1093/bioinformatics/btl461 PMID: 16940322

30. Ribeiro AA, Ortiz V. MDN: A Web Portal for Network Analysis of Molecular Dynamics Simulations. Biophys J. 2015; 109(6):1110–6. https://doi.org/10.1016/j.bpj.2015.06.013 PMID: 26143656

31. Doncheva NT, Klein K, Domingues FS, Albrecht M. Analyzing and visualizing residue networks of protein structures. Trends Biochem Sci. 2011; 36(4):179–82. https://doi.org/10.1016/j.tibs.2011.01.002 PMID: 21345680

32. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 2003; 13:2498–504. https://doi.org/10.1101/gr.1239303 PMID: 14597658

33. Petit CM, Zhang J, Sapienza PJ, Fuentes EJ, Lee AL. Hidden dynamic allostery in a PDZ domain. Proc Natl Acad Sci U S A. 2009; 106(43):18249–54. https://doi.org/10.1073/pnas.0904492106 PMID: 19828436

34. van den Berk LCJ, Landi E, Walma T, Vuister GW, Dente L, Hendriks WJAJ. An Allosteric Intramolecular PDZ−PDZ Interaction Modulates PTP-BL PDZ2 Binding Specificity. Biochemistry. 2007; 46 (47):13629–37. https://doi.org/10.1021/bi700954e PMID: 17979300

35. Harris BZ, Lim WA. Mechanism and role of PDZ domains in signaling complex assembly. Journal of Cell Science. 2001; 114(18):3219. https://doi.org/10.1242/jcs.114.18.3219 PMID: 11591811

36. Fan JS, Zhang M. Signaling complex organization by PDZ domain proteins. Neurosignals. 2002; 11 (6):315–21. https://doi.org/10.1159/000068256 PMID: 12566920

37. Hung AY, Sheng M. PDZ domains: structural modules for protein complex assembly. J Biol Chem. 2002; 277(8):5699–702. https://doi.org/10.1074/jbc.R100065200 PMID: 11741967

38. Zhang J, Sapienza PJ, Ke H, Chang A, Hengel SR, Wang H, et al. Crystallographic and nuclear magnetic resonance evaluation of the impact of peptide binding to the second PDZ domain of protein tyrosine phosphatase 1E. Biochemistry. 2010; 49(43):9280–91. https://doi.org/10.1021/bi101131f PMID: 20839809

39. Fuentes EJ, Der CJ, Lee AL. Ligand-dependent Dynamics and Intramolecular Signaling in a PDZ Domain. Journal of Molecular Biology. 2004; 335(4):1105–15. https://doi.org/10.1016/j.jmb.2003.11.010 PMID: 14698303

40. Fuentes EJ, Gilmore SA, Mauldin RV, Lee AL. Evaluation of energetic and dynamic coupling networks in a PDZ domain protein. J Mol Biol. 2006; 364(3):337–51. https://doi.org/10.1016/j.jmb.2006.08.076 PMID: 17011581

41. Gautier C, Laursen L, Jemth P, Gianni S. Seeking allosteric networks in PDZ domains. Protein Eng Des Sel. 2018; 31(10):367–73. https://doi.org/10.1093/protein/gzy033 PMID: 30690500

42. Kong Y, Karplus M. Signaling pathways of PDZ2 domain: a molecular dynamics interaction correlation analysis. Proteins. 2009; 74(1):145–54. https://doi.org/10.1002/prot.22139 PMID: 18618698

43. Vijayabaskar MS, Vishveshwara S. Interaction energy based protein structure networks. Biophys J. 2010; 99(11):3704–15. https://doi.org/10.1016/j.bpj.2010.08.079 PMID: 21112295

44. Raimondi F, Felline A, Seeber M, Mariani S, Fanelli F. A Mixed Protein Structure Network and Elastic Network Model Approach to Predict the Structural Communication in Biomolecular Systems: The PDZ2 Domain from Tyrosine Phosphatase 1E As a Case Study. J Chem Theory Comput. 2013; 9(5):2504–18. https://doi.org/10.1021/ct400096f PMID: 26583738

45. Mino-Galaz GA. Allosteric communication pathways and thermal rectification in PDZ-2 protein: a computational study. J Phys Chem B. 2015; 119(20):6179–89. https://doi.org/10.1021/acs.jpcb.5b02228 PMID: 25933631

46. Zhou H, Tao P. REDAN: Relative Entropy-Based Dynamical Allosteric Network Model. Mol Phys. 2019; 117(9–12):1334–43. https://doi.org/10.1080/00268976.2018.1543904 PMID: 31354173

47. Lu C, Knecht V, Stock G. Long-Range Conformational Response of a PDZ Domain to Ligand Binding and Release: A Molecular Dynamics Study. J Chem Theory Comput. 2016; 12(2):870–8. https://doi.org/10.1021/acs.jctc.5b01009 PMID: 26683494

48. Morra G, Genoni A, Colombo G. Mechanisms of Differential Allosteric Modulation in Homologous Proteins: Insights from the Analysis of Internal Dynamics and Energetics of PDZ Domains. Journal of Chemical Theory and Computation. 2014; 10(12):5677–89. https://doi.org/10.1021/ct500326g PMID: 26583250

49. Gerek ZN, Ozkan SB. Change in allosteric network affects binding affinities of PDZ domains: analysis through perturbation response scanning. PLoS Comput Biol. 2011; 7(10):e1002154. https://doi.org/10.1371/journal.pcbi.1002154 PMID: 21998559

50. Kalescky R, Zhou H, Liu J, Tao P. Rigid Residue Scan Simulations Systematically Reveal Residue Entropic Roles in Protein Allostery. PLoS Comput Biol. 2016; 12(4):e1004893. https://doi.org/10.1371/journal.pcbi.1004893 PMID: 27115535

51. Dhulesia A, Gsponer J, Vendruscolo M. Mapping of Two Networks of Residues That Exhibit Structural and Dynamical Changes upon Binding in a PDZ Domain Protein. Journal of the American Chemical Society. 2008; 130(28):8931–9. https://doi.org/10.1021/ja0752080 PMID: 18558679

52. Shannon CE. A mathematical theory of communication. Bell System Technical Journal. 1948; 27:379–423.

53. Freeman LC. A Set of Measures of Centrality Based on Betweenness. Sociometry. 1977; 40(1):35–41.

54. Šali A. Comparative protein modeling by satisfaction of spatial restraints. Molecular Medicine Today. 1995; 1(6):270–7. https://doi.org/10.1016/s1357-4310(95)91170-7 PMID: 9415161

55. Case DA, Berryman JT, Betz RM, Cerutti DS, Cheatham I, T.E., Darden TA, et al. AMBER 2015. University of California, San Francisco2015.

56. Maier JA, Martinez C, Kasavajhala K, Wickstrom L, Hauser KE, Simmerling C. ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. J Chem Theory Comput. 2015; 11(8):3696–713. https://doi.org/10.1021/acs.jctc.5b00255 PMID: 26574453

57. Jorgensen WL, Chandrasekhar J, Madura JD, Impey RW, Klein ML. Comparison of simple potential functions for simulating liquid water. The Journal of Chemical Physics. 1983; 79(2):926–35.

58. Miyamoto S, Kollman PA. Settle—an Analytical Version of the Shake and Rattle Algorithm for Rigid Water Models. Journal of Computational Chemistry. 1992; 13(8):952–62.

59. Duell ER, Glaser M, Le Chapelain C, Antes I, Groll M, Huber EM. Sequential Inactivation of Gliotoxin by the S-Methyltransferase TmtA. ACS Chem Biol. 2016; 11(4):1082–9. https://doi.org/10.1021/acschembio.5b00905 PMID: 26808594

60. Roe DR, Cheatham TE 3rd. PTRAJ and CPPTRAJ: Software for Processing and Analysis of Molecular Dynamics Trajectory Data. J Chem Theory Comput. 2013; 9(7):3084–95. https://doi.org/10.1021/ct400341p PMID: 26583988

61. Hunter JD. Matplotlib: A 2D Graphics Environment. Computing in Science & Engineering. 2007; 9(3):90–5.

62. Humphrey W, Dalke A, Schulten K. VMD: visual molecular dynamics. J Mol Graph. 1996; 14(1):33–8, 27–8. https://doi.org/10.1016/0263-7855(96)00018-5 PMID: 8744570

63. Schrodinger, LLC. The PyMOL Molecular Graphics System, Version 1.8. 2015.

64. Schneider M, Trummer C, Stengl A, Zhang P, Szwagierczak A, Cardoso MC, et al. Systematic analysis of the binding behaviour of UHRF1 towards different methyl- and carboxylcytosine modification patterns

at CpG dyads. PLOS ONE. 2020; 15(2):e0229144. https://doi.org/10.1371/journal.pone.0229144 PMID: 32084194

65. Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, et al. UCSF Chimera—A visualization system for exploratory research and analysis. Journal of Computational Chemistry. 2004; 25(13):1605–12. https://doi.org/10.1002/jcc.20084 PMID: 15264254

66. Lockless SW, Ranganathan R. Evolutionarily conserved pathways of energetic connectivity in protein families. Science. 1999; 286(5438):295–9. https://doi.org/10.1126/science.286.5438.295 PMID: 10514373

67. Taylor NR. Small world network strategies for studying protein structures and binding. Computational and structural biotechnology journal. 2013; 5:e201302006. https://doi.org/10.5936/csbj.201302006 PMID: 24688699

68. Yan W, Zhou J, Sun M, Chen J, Hu G, Shen B. The construction of an amino acid network for understanding protein structure and function. Amino Acids. 2014; 46(6):1419–39. https://doi.org/10.1007/s00726-014-1710-6 PMID: 24623120

69. del Sol A, Tsai CJ, Ma B, Nussinov R. The origin of allosteric functional modulation: multiple pre-existing pathways. Structure. 2009; 17(8):1042–50. https://doi.org/10.1016/j.str.2009.06.008 PMID: 19679084

70. Knapp B, Ospina L, Deane CM. Avoiding False Positive Conclusions in Molecular Simulation: The Importance of Replicas. J Chem Theory Comput. 2018; 14(12):6127–38. https://doi.org/10.1021/acs.jctc.8b00391 PMID: 30354113