Technische Universität München

TUM School of Life Sciences

# Prediction of long-range conformational coupling and allostery in proteins using structure networks

Markus Schneider, M. Sc.

Vollständiger Abdruck der von der TUM School of Life Sciences der Technischen Universität München zur Erlangung eines

Doktors der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

| | |
|---|---|
| Vorsitz: | Prof. Dr. Aphrodite Kapurniotu |
| Prüfende der Dissertation: | 1. Prof. Dr. Dmitrij Frischmann |
| | 2. Prof. Dr. Martin Zacharias |
| | 3. Prof. Dr. Giorgio Colombo |

Die Dissertation wurde am 23.10.2023 bei der Technischen Universität München eingereicht und durch die TUM School of Life Sciences am 25.03.2024 angenommen

*To my wife Lena, for her love, support and understanding,*

*As well as my parents Erika and Gerhard, for showing me the way,*

*To my sister Michaela, for inspiring me with her strength,*

*And to my son Simon, who gives me boundless joy and purpose.*

MARKUS SCHNEIDER

# ABSTRACT

Drug discovery and design aims to devise compounds with the ability to precisely regulate the activity of specific target biomolecules. Insufficient selectivity of drugs presents a substantial challenge in the process, leading to many side effects, including severe toxicity. Allosteric drugs have emerged as a promising approach to address difficult protein targets. In contrast to conventional drugs binding directly to protein active sites, allosteric drugs bind to separate effector sites, modulating protein function at the active site indirectly over long distances. However, the location of these effector sites and their mechanism of action is a priori unknown for many proteins, impeding targeted drug design workflows. Various computational methods aimed at predicting allosteric effector sites have been developed, but so far, no definite approach has been established. This work investigates the effects of ligand binding on the dynamics of four protein systems, as observed during molecular dynamics simulations. Analysis of ligand binding patterns in two nucleotide binding pockets of UHRF1 revealed conformational coupling between distant protein regions, which could be captured effectively using a protein structure network model. This specific form of network follows the formation and dissolution of atom interactions during simulations, like hydrogen bonds or hydrophobic contacts, allowing to correlate protein conformations by the dynamics of their interaction states. The concept was further developed on PDZ2, a commonly used protein system for testing computational methods aimed at predicting conformational coupling and allostery, and formalized into the SenseNet analysis framework. Two novel scores are proposed for prediction of allosteric regions, based on mutual information between interaction timelines obtained from molecular dynamics simulations, and evaluated using a validation set assembled from NMR data. Comparing these results with other published predictions for the PDZ2 system revealed systematic problems with current approaches, emphasizing the need for larger studies and unbiased evaluation standards. Within these limitations, the proposed scores showed good agreement with known allosteric protein regions. Finally, the SenseNet model was applied to study the differential effects of ligand binding in two Hsp70 chaperones, DnaK and BiP, using networks generated from molecular dynamics simulations approximating different phases of the chaperone conformational cycle. This revealed a conserved structural core of residues with allosteric roles, complemented by variant specific residues in marginal regions, which may help explain biochemical differences between these related proteins. Beyond providing a single prediction model, SenseNet serves as an open-source platform integrating different network analysis approaches. Based on this foundation, future developments could yield advanced techniques for more accurate prediction of allosteric communication and improving the sampling efficiency of molecular dynamics simulations.

# ZUSAMMENFASSUNG

Drug discovery und -design zielt darauf ab, Medikamente zu entwickeln welche die Aktivität von Zielmolekülen präzise regulieren. Eine unzureichende Selektivität von Wirkstoffen kann zu einer Vielfalt von Nebenwirkungen führen, bis hin zu schwerer Toxizität, und stellt deshalb eine erhebliche Herausforderung dar. Allosterische Medikamente haben sich als ein vielversprechender Ansatz für besonders schwierige Zielproteine herauskristallisiert. Im Gegensatz zu konventionellen Medikamenten, die direkt an das aktive Zentrum von Proteinen binden, zielen allosterische Wirkstoffe auf separate Effektor-Bindetaschen, welche die Funktion des Proteins indirekt und über große Distanzen modulieren. Die genaue Lage der Effektorregionen und ihr Wirkmechanismus sind jedoch für viele Proteine a priori unbekannt, was eine gezielte Wirkstoffentwicklung erschwert. Unter den verschiedenen bioinformatischen Methoden zur Vorhersage von allosterischen Effektorregionen hat sich noch kein klar überlegener Ansatz etabliert. Im Rahmen dieser Arbeit wurden die Auswirkungen der Ligandenbindung auf die Dynamik von vier Proteinsystemen mittels Molekulardynamik-Simulationen untersucht. Eine Analyse der Interaktionsmuster in zwei Nukleotid-Bindetaschen von UHRF1 zeigte eine Konformationskopplung zwischen den Aminosäuren zweier getrennter Proteinregionen, welche effektiv durch ein Proteinstrukturnetzwerk dargestellt werden kann. Das entwickelte Netzwerkmodell kodiert die Bildung und Auflösung von Atominteraktionen während der Simulationen, wie zum Beispiel Wasserstoffbrückenbindungen oder hydrophobe Kontakte; hierdurch wird es ermöglicht, Proteinkonformationen basierend auf der Dynamik ihrer Interaktionszustände zu korrelieren. Eine Folgestudie nutzte PDZ2, ein beliebtes Testsystem für bioinformatische Methoden zur Vorhersage von Konformationskopplung und Allosterie, um diese Idee weiterzuentwickeln und schlussendlich als SenseNet Analyseframework zu formalisieren. Es werden zwei neue Scores zur Vorhersage von allosterischen Regionen vorgeschlagen, basierend auf der Mutual Information zwischen Interaktions-Timelines während Molekulardynamik-Simulationen; dieser Ansatz wurde mithilfe eines NMR-Datensatzes gegen experimentelle Daten validiert. Ein Vergleich dieser Ergebnisse mit anderen publizierten Vorhersagen für das PDZ2-System deutet auf systematische Probleme in aktuellen Modellen hin. Diese Beobachtungen unterstreichen die Notwendigkeit weitreichenderer Studien sowie besserer Standards für Daten und Implementationen, um Modelle vergleichbarer zu machen. Trotz dieser Limitationen zeigen die vorgeschlagenen Scores eine gute Übereinstimmung mit bekannten allosterischen Proteinregionen. Schlussendlich wurde das SenseNet Modell auf zwei Hsp70 Chaperon Varianten, DnaK und BiP, angewandt, um die differentiellen Auswirkungen von Ligandenbindung in diesen Systemen zu untersuchen. Hierzu wurden Proteinstrukturnetzwerke eingesetzt, welche aus Molekulardynamik-Simulationen generiert wurden und verschiedene Phasen des Chaperon-Konformationszyklus approximieren. Die Netzwerke zeigten einen konservierten strukturellen Kern von Aminosäuren mit allosterischen Funktionen, der durch variantenspezifische Randregionen ergänzt wird. Zusätzlich zur Bereitstellung seines Vorhersagemodells öffnet SenseNet die Möglichkeit, als Open-Source Plattform verschiedene Ansätze zur Netzwerkanalyse zu integrieren. Möglichkeiten zur weiteren Entwicklung dieser Plattform beinhalten neue Methoden zur präziseren Vorhersage allosterischer Kommunikation und Verbesserung der Effizienz von Molekulardynamik-Simulationen.

# TABLE OF CONTENTS

# 1. INTRODUCTION

## 1.1. Motivation for predicting long range conformational coupling between protein regions

Protein structure analysis can be used to identify key regions regulating biochemical functions. Insights gained from structure-based approaches have various applications in protein engineering (1-5) and drug design (6-8). Conventional drugs are typically optimized to bind specific active sites, also known as orthosteric sites, within proteins. Within such sites, drug binding competes with endogenous ligands to block a specific aspect of protein function (6, 9). Ensuring that custom designed ligands bind specifically to the targeted active sites, but not other structurally similar proteins, remains a major challenge for drug development. As an emerging alternative approach, allosteric drugs forego the orthosteric site by targeting instead a spatially distinct effector site. Regulation of protein function is thus achieved by long range allosteric effects between the effector site and the orthosteric site (9-11). Effector sites associated with allostery have been observed to be less evolutionary conserved than active sites (12), which opens possibilities for difficult protein targets as this could reduce off-target binding to structurally similar binding pockets. There are numerous potential advantages to allosteric drugs, including increased specificity, fewer side effects and lower toxicity than orthosteric drugs (10, 11). In practice, the search for allosteric drugs has proven challenging, with only 19 FDA approved allosteric drugs as of 2020 (13). While some of the emerging difficulties are readily identified, such as higher hydrophobicity and lower binding affinities of allosteric modulators (14), progress is also impeded by our incomplete understanding of the nature and mechanism of allosteric interactions (13). Establishing an accurate model for the prediction of effector regions associated with allosteric control could greatly increase the efficiency of allosteric drug development (9, 11, 13, 15).

# INTRODUCTION

Allostery encompasses the conditions by which the activity of a protein's functional site is regulated via perturbation of a separate effector site in the same protein. Models attempting to link this functional coupling to a structural mechanism have evolved drastically over the sixty years since the concept was introduced (15-18). The mid-1960's Monod–Wyman–Changeux (MWC) (19) and Koshland–Nemethy–Filmer (KNF) (20) models explained allostery as a structural change between two functionally distinct protein states, induced by binding of an effector molecule. To understand the underlying allosteric mechanism, crystallographic structures of different protein states are commonly analyzed with the goal of tracking the conformational changes of key residues. This approach became well known for its application to explaining cooperative binding in hemoglobin (21-23) and has been successful in explaining the allosteric behavior of many proteins (24). However, reports of allostery without apparent conformational changes (25-28), which may only be traceable through transient intermediate structures (29-31), are challenging to unify with an approach focusing solely on a limited number of energetically favorable protein conformations. Based on a model of energy landscapes used for investigation of protein folding, in 1999 the Nussinov group proposed a mechanism of "conformational selection" (32). In this view, perturbation of a protein at the effector binding site shifts the balance of populated conformations within the structure ensemble; this shift can influence the functional site, like a ligand binding site, by increasing or decreasing the free energy of the binding competent state (Figure 1). This constitutes a notable shift in perspective, from observation of conformational changes between individual protein structures, towards focusing on an ensemble encompassing a variety of stable and transient conformations. While the revision and refinement of allosteric paradigms is still a very active process, most recent efforts concur on the essential role of the protein's conformational ensemble (15, 17, 33).

Computational methods in conventional drug design workflows are commonly used to select the most promising leads from a large list of drug candidates, reducing the required number of in vitro experiments to be performed in high throughput screening (34). Techniques employed for this purpose are often based on molecular docking (8, 35) or MMGBSA/MMPBSA (36, 37), with a focus on optimizing the interactions between an orthosteric functional site in the protein and one or more candidate ligands. In contrast to orthosteric binding sites which are already known from endogenous ligands, the location of allosteric sites and the structure of its ligands are rarely

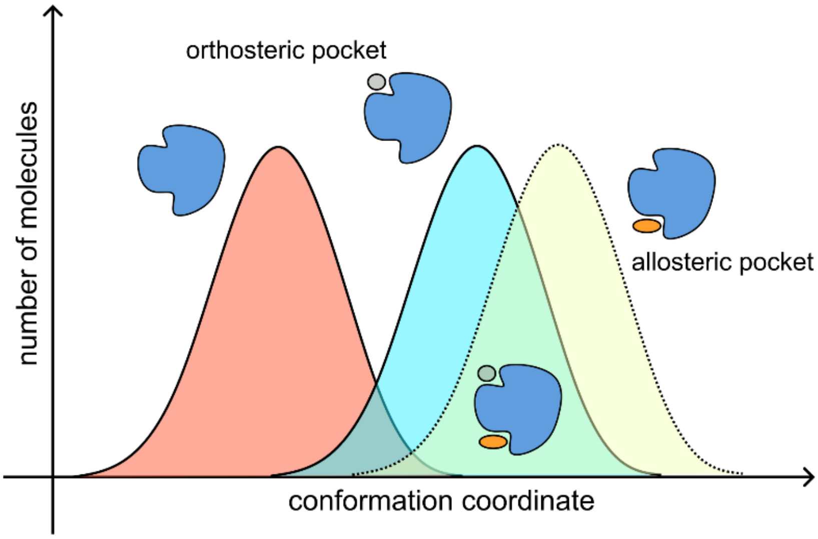


**Figure 1. Conformational selection and population shift model of allostery.** In this model, binding of an allosteric ligand (orange) shifts the conformational ensemble of a protein (blue) to enhance binding of an orthosteric ligand (grey). In the presence of an allosteric ligand, proteins shift from the unbound conformation (red) towards a complex with the allosteric ligand bound (yellow). This complex has conformational overlap with the light blue conformation, shifting the conformational equilibrium towards binding of the orthosteric ligand. Figure recreated and adapted from ref. (18).

observed in crystal structures (13). This problem is compounded when considering hidden, or cryptic allosteric sites, which have been shown to form under specific conditions like the presence of stabilizing compounds (13, 38, 39). Such sites are thus difficult to detect by investigating a limited number of structures, which represent only the most probable conformations. With no previous knowledge of the location of the effector site or its structure, methods designed to evaluate the strength of protein-ligand binding tend to lack an effective starting point for allosteric sites and ligands. Various methods have been developed (13) to detect allosteric sites based on scanning for cavities within a protein structure, such as AlloSite (40), AlloPred (41), PARS (42, 43), AlloSigMA (44), CavityPlus (45), DynOmics (46), SPACER (47) and STRESS (48). The common approach pursued by these methods is a combination of extracting geometric features from a single protein structure, as in binding pocket detection, combined with an approximate modelling of correlated modes using normal mode analysis (NMA) and elastic network models (ENM). These choices emphasize that simply detecting a possible binding pocket is not sufficient to find effective allosteric ligands, as any putative effector site must be able to effectively influence activity at the targeted protein active site. While computationally inexpensive, NMA and ENM models offer only a greatly simplified and incomplete view of protein dynamics. With increasing protein size and distance between functional and putative effector sites, conformational coupling becomes more difficult to predict, due to the inherent complexity of conformational shifts underlying allostery (15, 49). It seems likely that adequate modelling of long-range allosteric effects requires accounting for complex protein dynamics, specifically determining residues which can sense conformational perturbations at a distance. To make use of these opportunities, established approaches used for protein and drug design can be supplemented by methods describing complex forms of conformational coupling between distant protein regions. Molecular Dynamics (MD) simulations represent a viable approach for sampling conformations of a protein more accurately than NMA and ENM, but are much more expensive to calculate to sufficient statistical precision, which limits the accuracy of observations and conclusions based on them (50, 51).

Accounting for a large ensemble of protein conformations, including combinations of different conformations at allosteric and functional sites, complicates a mechanistic understanding of allostery. Observation of conformational coupling presents one possible strategy to make this problem more tractable: Instead of enumerating the conformational details of individual protein conformations, one can ask which residues have the largest conformational impact on their environment. This approach can be inferred from the conformational selection model (24), which proposes that allosteric effects arise when the conformational shift of binding an effector to the allosteric site impacts the conformational equilibrium at the protein's functional site. In this view, conformational coupling becomes a prerequisite to allostery, and thus has the potential to be a useful predictor. Furthermore, tracking the interdependence of residue conformations allows to include information about both the energetically favored states, equivalent or close to structures obtained via crystallography, and the hidden transient structures contributing to the structure ensemble. This view does not consider conformational coupling and allostery as fully equivalent; while allostery implies conformational coupling, the reverse is not necessarily true. A true allosteric effect requires a functional component that is observable via experiments. Therefore, any predictions of allostery must be carefully analyzed in the context of available experimental data, and ideally should always be complemented with follow up experiments, which may be informed by predictions.

## 1.2. Computational methods to estimate conformational coupling

A sizable number of tools have been developed to predict conformational coupling and protein allostery (11, 13, 33, 52-54). Although this variety may appear redundant at first glance, prediction models are still in an early stage (52) as their relative strengths and limitations are not yet fully realized. The following section outlines a selection of key methods demonstrating common principles and how their specific tradeoffs affect their applicability to certain systems and problems. To date, it is often difficult to comprehensively compare the performances of individual algorithms and determine their accuracy for predicting experimental data. The challenges begin with choosing an experimental dataset for validating predictions; for example, NMR measurements are among the most popular data sources for this purpose. However, Fuglestad et al. (55) showed that different computational prediction methods can match different sets of NMR-derived parameters, depending on the timescales of investigated dynamics. Evaluating accuracy of a computational prediction model must therefore consider the specific nature of experimental data it was designed to predict. In turn, the appropriate prediction method must be chosen with respect to the timescales of relevant conformational coupling, which may differ between proteins or even between different modes of motion within the same protein (55).

Observing the underlying ideas of notable prediction methods can explain how they may be able to probe different timescales and dynamics. Methods relying on individual single protein structures include structure networks utilizing shortest paths and centrality measures (56-58), pathway tracing through alternative conformations in crystallography structures (59), pairwise energetic contribution of residue pairs compared to mutated variants (“frustration”) (60) and conformational rearrangements between multiple structures (61). Due to their limitation to a few or even just one structure, these methods only describe conformational coupling indirectly and through extrapolation from a static network topology. Their accuracy in predicting allostery depends on the validity of their topological assumptions, e.g., that allosteric communication follows the shortest path in a network, or the availability of multiple structures with relevant conformational changes. The discovery of hidden (“cryptic”) allosteric sites, forming only sporadically in high energy states and invisible in crystal structures, highlights the limitations of static structure analysis (38, 39).

The class of coevolution methods is distinguished by their emphasis on protein sequences rather than structures (13, 53, 62). Given a sufficiently deep, high-quality Multiple Sequence Alignment (MSA), these methods extract coevolution patterns and score the strength of evolutionary coupling between residues. Residues with strong coevolution are predicted to have a functional role in the protein, as the preservation of function constrains viable mutations. Different variants of this analysis have been developed (62), among them statistical coupling analysis (63, 64) (SCA), corrected MI

(65) (MIp), observed minus expected squared (OMES) (66), direct coupling analysis (DI) (67) and Protein Sparse Inverse COVariance (PSICOV) (68). Among many applications, coevolution couplings have been shown to be connected to protein folding (69), protein dynamics involving allostery (70), and allosteric effects in protein complexes (71). Their unique use of sequence data is particularly attractive as they can be combined with structural methods like Molecular Dynamics for a complementary approach (70, 72). However, it has been noted that coevolution signals bear a strong correlation to protein contact networks (62, 67), which may put an effective upper limit to the gains of combined sequence and structure approaches.

Beyond analyses of protein sequences and individual structures, Elastic Network Models (ENMs) offer an efficient way to explore simple protein dynamics (33, 41, 43, 73, 74). In an ENM, the protein structure is treated as a three-dimensional network of coarse-grained nodes, commonly representing the Cα atom of each residue, which are pairwise connected by elastic springs. This simplified physical model can then be subjected to by Normal Mode Analysis (NMA) (33, 41, 43, 73, 74), which diagonalizes the (simplified harmonic) potential energy landscape and yields the (linearly) independent low frequency motions (“modes”) of the system. From these modes, residues with coupled motion are extracted to predict allosteric effects. Low computational cost makes analyses based on ENM and NMA attractive for high-throughput workflows, albeit subject to the requirement that the relevant dynamics are captured by harmonic fluctuations around an energy minimum. This assumption neglects more complex unfolding or rigid-body motions, as the approximation breaks down once even minor conformational changes away from the original structure are considered (33).

The limitations imposed by ENM methods emphasize the attractivity of Molecular Dynamics (MD), as conformational coupling can be observed directly, extracting correlations from the motions observed during simulation (75-78). Conventional, or equilibrium MD aims to imitate the motions of a protein by modelling relevant forces between atoms and allowing the atom positions to move according to these forces, recording the visited conformations after discrete time steps. Given a sufficient number of simulation steps, the trajectory of sampled conformation snapshots is expected to correspond to the conformational states of a system under the simulated conditions. MD simulations allow to estimate the timings of the observed conformational coupling, enabling comparison with experimental data of similar time scales. However, in order to obtain sufficient statistical precision, the timescale of observed coupling should be 100 – 1000 times faster than the total simulation time. The now widespread use of GPUs in MD has greatly increased achievable simulation time scales, and there are ongoing efforts to improve computational efficacy. Notable developments in the field include coarse grained models (79-81), which replace atomistic systems with lower resolution models that are easier to sample, and advanced sampling methods (82-84) aiming to exploit more cost-effective algorithms to evolve a system given a set of starting conformations. Despite significant advances in the field in recent years, for the time being conventional MD remains the standard method for probing complex motions in proteins. With current standard simulation times in the µs – ms range, analyses based on conventional MD are largely limited to coupling events occurring in magnitudes between ~ 100 ns – 1 ms. Consequently, investigation of conformational coupling using MD simulations places a requirement on analysis methods which can extrapolate within the constraints of high statistical noise caused by limited simulation times. Whereas conventional MD and advanced sampling variants focus on obtaining a comprehensive structure ensemble describing all relevant protein conformations, specialized MD variants have been developed to trace the propagation of artificial dynamics through the system. In Pump-probe MD, an oscillating force is added to excite a specific protein region (85). The simulation is then given time to propagate these fluctuations, which are subsequently analyzed by extracting the fluctuation power spectrum from the atomic coordinates. Finally, this power spectrum is compared to spectra obtained from baseline simulations or simulations with modified forces; residues affected by the propagation of the oscillating forces are highlighted as conformationally coupled. Perturbation Response Scanning follows a similar strategy by applying random forces to individual residues and measuring the magnitude and direction of atomic displacements of the whole protein (86). A correlation consensus is then formed to measure the influence of applying force on each residue on the rest of the structure (13). Other forms of specialized simulation protocols are available, though generally they share the same principle of trading generalizability of simulations, which target the full dynamics of the system and can be subjected to different analyses, in exchange for potentially higher prediction specificity.

Given the complexity of analyzing correlated motion in MD trajectories between pairs of thousands of atoms, network models have received increased attention in recent years. A well-chosen network model can help to abstract away irrelevant atomistic details, reduce noise, and allow an intuitive exploration of the essential motions observed during MD simulation. Common models can be broadly categorized into two classes, depending on whether nodes in the network represent topological features within the protein (protein structure networks) or distinct conformational states in the structure ensemble, as for example in Markov State Models (MSM). Protein structure networks consist of nodes, representing atoms or residues, which are connected by edges corresponding to interactions between atoms (57, 58, 87-94). Interactions are most often defined by inter-residue contacts within a distance cutoff chosen at any limit between 4 and 8 Å, i.e., the edges represent short-range interactions in the approximate range of noncovalent forces between atoms (87-89). If a residue pair in the structure fulfills the selected criteria for an interaction, an edge is drawn between the corresponding nodes; otherwise, they are considered unconnected. The resulting network of nodes and edges can then be subjected to analysis methods inspired by graph theory, in order to determine interactions and residues with noteworthy characteristics with respect to their role in the network (56, 57, 88, 89). In their original and most widely used form, protein structure networks are based on single structures obtained by crystallography, but some extensions to molecular dynamics trajectories have been proposed (see section 1.5 for an overview). The second kind of network model does not map atoms to nodes, but instead uses nodes to represent a distinct conformational state of the protein, with connecting edges indicating transition probabilities between these states. Markov State Models use this approach to detect patterns within the dynamic transitions of the system and provide an intuitive model for interpretation based on states and transition rates, akin to reaction equations commonly utilized in biochemistry (13, 95). The proposal offered by MSMs make them applicable for a variety of tasks, including the modelling of protein folding, MD advanced sampling, and allosteric regulation (38, 96-98). However, in practice the creation of MSMs can be quite challenging, as it involves several nontrivial tasks, such as the appropriate way of defining appropriate metastable protein states which captures the relevant kinetics of the protein (95, 99, 100) or the selection of appropriate lag times during coarse graining (101).

Among available methods for investigating conformational coupling, choosing an appropriate approach for a specific problem requires a detailed understanding of the respective advantages and drawbacks of each method, including an analysis of the aspects of (i) accuracy for predicting a particular type of experimental data, (ii) computational cost, (iii) opportunity cost of performing specialized simulations versus nonbiased simulations that can be reused for other purposes. Over the course of this dissertation, the need for detailed analyses of strongly localized protein-ligand interaction patterns drove us towards a protein structure network model, which was adapted for MD simulations to account for ligand induced flexibility. Noting both the intuitive strength of this approach and its apparent connection to conformational coupling, we found that the integration of MD data in protein structure networks still had not been explored in the same depth as, for example, Markov state models.

## 1.3. Prediction of conformational coupling using topological features from protein structure networks

Protein structures obtained by crystallography or NMR provide information about the most dominant conformations of the system. Given only a small number of structures, it is difficult to measure conformational coupling directly by observing correlated motion. Instead, a model must be provided to predict conformational coupling from the static topology of atoms and residues within the system. The most common assumption is that residue conformations, in the context of a stable folded protein, are most effectively modulated by their direct environment. This is justified by the short effective range of the dominant inter-atomic forces, i.e., hydrophobic contacts, hydrogen bonds and salt bridges (102). Most structure network models follow this premise by analyzing the neighborhood topology of residues and their connections formed by short-range interactions. In this model, conformational coupling can be predicted by determining residues located on “shortest paths” between protein regions of interest.

Within the domain of predicting functional residues in proteins, the concept of analyzing the shortest paths in a network has been applied extensively (54, 56, 77, 78, 87, 88, 103-108). In this context, the shortest path is the one that requires the least steps when traversing between two nodes along the edges of the network. This approach views long range conformational coupling as a signal transmitted via chains of connected residues in the protein (56). This signal is sometimes visualized as a cascade of residues flipping between well- defined conformations, where each residue the chain triggers the next like a series of switches, although the statistical nature of allostery should serve as warning against assuming an overly simplistic view (17). Other possible forms of signaling include modulation of frequencies and amplitudes of residue-level fluctuations, rotation and packing, or even domain-level oscillations and hinge-bending motions (27). Shortest path models do not explicitly presume a specific mechanism of action, suggesting instead that the signal is likely to travel along the shortest, and thus most effective, path through the protein structure.

There is no agreement on a singular measure used to quantify nodes and edges relevant to shortest paths, with variations of “Betweenness Centrality” (BC) and “Characteristic Path Length Centrality” (CPLC) among the most widely used (56-58, 87). Betweenness Centrality (57, 109) evaluates the fraction of shortest paths passing through a specific node, with respect to the total number of shortest paths connecting all possible node pairs. It can be expressed as

$$BC(i) = \sum_{j,k \in N,\ \ i \neq j \neq k} \frac{\sigma_{jk|i}}{\sigma_{jk}} \tag{1.1}$$

where $i, j, k$ are part of the set of nodes $N$, $\sigma_{jk}$ 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 $i$.

Residues located on many shortest paths receive the highest scores, as they represent bottlenecks in the topology of the network (Figure 2) and are thus presumed to play an important role in communicating signals. In contrast, Characteristic Path Length Centrality (CPLC) investigates the robustness of the shortest paths in the network with respect to the removal of individual nodes (56). It is calculated by evaluating how the characteristic path length, i.e., the average length $L$ of shortest paths in the networks changes when a specific node is removed, expressed as

$$CPLC(i) = |L - L_i| \tag{1.2}$$

where $L_i$ is the characteristic path length of the network after removal of node $i$. The characteristic path length can be calculated as

$$L = \frac{1}{N_p} \sum_{i,j \in N,\ i>j} d(i,j) \tag{1.3}$$

where $N$ is the set of nodes, $N_p$ is the number of node pairs in the network and $d(i,j)$ is the shortest path between $i$ and $j$. Taken together, residues which cause substantial growth in the average length of shortest paths after removal from the network show the highest CPLC scores.

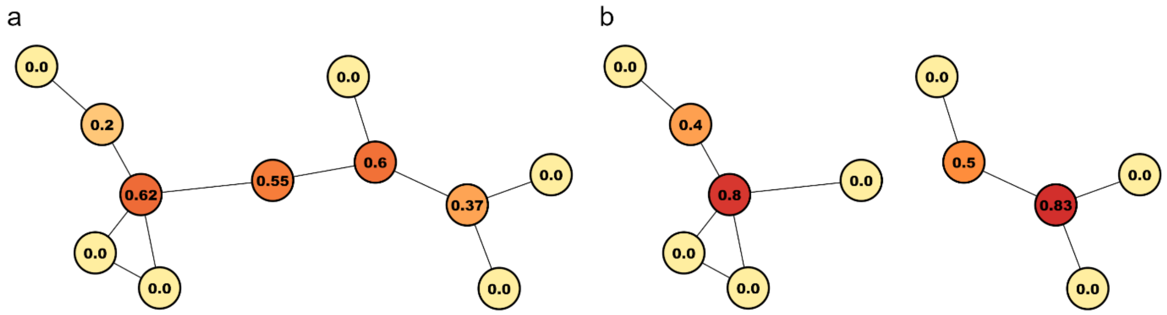


**Figure 2. Effect of spurious edges on centrality measures.** Networks (a) and (b) show changes in Node Betweenness Centrality before and after removal of a single edge.

Due to their tendency to attribute high scores to bottlenecks within networks, centrality scores are highly sensitive to the presence of individual edges. This poses a substantial problem for application of centrality methods to networks determined from structure ensembles. The introduction of transient residue interactions leads to a large number of spurious edges, even if they are only present in a tiny fraction of the ensemble. Generally, algorithms can be adapted to add appropriate weights to interactions (110). However, it is not obvious how the stability of an interaction should scale the shortest path measures as shown in eqs. 1.1 to 1.3. For example, simply scaling the length of a path by the stability of interactions would favor routing of signals through rigid secondary structures or hydrophobic cores over residues with – arguably more interesting - semi-stable “switch-like” conformational changes. These problems are commonly avoided by applying centrality algorithms only to individual structures, usually those obtained either by X-Ray crystallography or NMR.

## 1.4. Conformational coupling in structure ensembles

Having access to a representative portion of the protein's structure ensemble allows direct observation of conformational coupling between residues. Such an ensemble can be provided by NMR or calculated by Molecular Dynamics simulations. Once a representative subset of the ensemble is obtained, conformational coupling is determined by evaluating which residues change their conformation in a pattern that suggests systemic positional interdependence. One of the most straightforward strategies to quantify such a pattern is the Pearson correlation coefficient between atom coordinates (commonly one atom per residue, e.g. Cα or Cβ atoms) as

$$C_{ij} = \frac{\langle \Delta\vec{r_i}(t) \cdot \Delta\vec{r_j}(t) \rangle}{\sqrt{\langle \Delta\vec{r_i}(t)^2 \rangle \langle \Delta\vec{r_j}(t)^2 \rangle}} \tag{1.4}$$

where angle brackets denote the average over all structures in the ensemble and $\Delta\vec{r_i}(t) = \vec{r_i}(t) - \langle\, \vec{r_i}(t) \rangle$ is the deviation of the residue's center of mass coordinate vector from its average (77, 78). The $C_{ij}$ values can then be analyzed in a matrix or in a network-based approach, e.g. tracing a path of strongly correlated residues using shortest path methods (78). However, in these applications the Pearson coefficient suffers from two weaknesses (75): First, as is evident from the dot product in eq. 1.4, it only works for collinear vectors, which means that the coefficient is effectively blind to any correlated lateral motion between two residues (111). Second, Pearson's coefficient measures the strength of a linear relationship between variables and may give misleading results for non-linear correlations (75). To address both problems, correlation measures based on Mutual Information (MI) have been suggested as alternatives (75, 76). The MI is defined as

$$I(X,Y) = \iint dxdy\, p(x,y) \log\left(\frac{p(x,y)}{p_x(x)p_y(y)}\right) \tag{1.5}$$

with $X, Y$ as random variables realized by the observations $x, y$ with joint probability density $p(x, y)$ and marginal densities $p_x$ and $p_y$. The densities are generally unknown and must be estimated from the sets of observed values. Applied to atom coordinate vectors of eq. 1.4, the densities would correspond to the estimated probability distributions of $\Delta\vec{r_i}$ and $\Delta\vec{r_j}$. MI does not impose a specific model to the relationship between the two variables, thus avoiding the collinearity issue while capturing both linear and non-linear correlation (75). The fundamental challenge for calculating correlation using MI is to obtain reliable estimates for the probability distributions, which is a nontrivial task (112, 113). However, as an initial starting point for developing prediction models, a substantially simplified form can be used. Instead of calculating correlation between raw atom coordinates, one can attempt to find high-level structural features in the trajectory that encode the conformational interdependence between residues. These features offer several advantages: First, they can be chosen to tailor the correlation to intuitively meaningful interaction states. For example, if hydrogen bonds or hydrophobic interactions are chosen as features, correlation can then be measured between high-level interaction instead of low-level atom coordinates. This makes it easier to develop and evaluate scoring algorithms based on correlation, since the resulting scores are can be reasoned about more easily. Second, high-level interaction representations can be chosen to be easily discretized, which allows a simplified calculation of mutual information (MI)

$$I(X, Y) = \sum_{x \in X} \sum_{y \in Y} p(x, y) \log\left(\frac{p(x, y)}{p_x(x) p_y(y)}\right) \tag{1.6}$$

with $p$ now representing the probability distributions of discrete conformations $x$ and $y$. These distributions can be straightforwardly estimated by counting the frequency of each conformation occurring in the structure ensemble. This can be illustrated with a simple example: Suppose that $X$ and $Y$ each represent a hydrogen bond between atoms belonging to different residues in the structure ensemble. In each structure, a hydrogen bond may either be in conformation 0 (absent) or conformation 1 (present), which can vary between structures as residues move and engage in different interaction patterns. The first term of the MI would evaluate the probability of both hydrogen bonds occurring together in the same structure $p(1,1)$ in relation to the expected probability if both were to occur independently from each other, i.e. $p_x(1)p_y(1)$. The remaining terms cover the remaining cases of $p(0,1)$,$p(1,0)$ and $p(0,0)$. The probabilities can be obtained simply by counting the structures in the ensemble for which the hydrogen bonds are present. Considering these advantages, we saw potential for investigating conformational coupling by combining discrete Mutual Information with the previously described structure networks, as a way of quantifying the interdependence of distinct structural conformations encoded by short-range physical interactions.

## 1.5. Available software for analysis of protein structure networks obtained from MD simulations

Several tools have been published for analysis of structure ensembles, most often in the form of MD trajectories, using protein structure networks. The most common strategy is to evaluate one or more types of inter-atom contacts (hydrophobic, hydrogen bonds or salt bridges) for each frame of the trajectory and create a time-averaged protein structure network. In this type of network, an edge between nodes (atoms) represents a contact within a predefined cutoff distance, usually between 4 and 8 Å, occurring for a minimum fraction of the total simulation time. The MD-TASK package (114) offers a set of python scripts to create protein structure networks, calculate various centrality measures and correlation between atom coordinates. Wordom (115) is a general purpose MD analysis tool for the command line, which also offers creation of average protein structure networks and network path analysis. NAPS (116), RINalyzer (117), PyInteraph and RIP-MD (118) similarly transform MD trajectories into an averaged network of contacts (hydrophobic, hydrogen bonds or salt bridges) for analyses and provide visualization as well as analysis tools centered around centrality measures. MDN (119) generates a network from pairwise residue interaction energies and predicts allosteric coupling using a centrality based measure. The gRINN tool (120) has similar capabilities to those mentioned, but adds functions to analyze correlations between pairwise residue energies. Another category of tools focuses on networks in which edges are not defined by direct contacts; instead, an edge is defined between any two atoms showing substantial cross-correlation between atom coordinates of the corresponding nodes. In this network variant, edges can occur between distant protein regions, if their atoms move in a correlated pattern. As in networks based on interactions, centrality measures are the most dominant approach to detect pathways of conformationally coupled residues. Methods based on cross-correlation networks include NetworkView (121) and xPyder (122). The COMMA (123) and COMMA2 (124) tools use a mix of atom coordinate correlation, minimum inter-atom distances, distance variance, non-covalent interaction strengths, and secondary structures to identify coupled cliques and communication pathways.

Despite the abundance of computational tools, their underlying principles are very similar. MD trajectories are transformed into a time-averaged network, with residues as nodes and edges representing either atom-atom contacts or the cross-correlation of atom coordinates. In most cases, this is followed by calculation of network centrality measures to predict conformationally coupled residues, which are proposed to act as regulators of allostery. Based on the available tools, we identified three specific shortcomings of conventional approaches to analyse protein structure networks from MD simulations: First, the use of centrality methods with their high sensitivity to spurious edges, which occur frequently in networks obtained from structure ensembles. Second, the correlation measures employed are severely limited in their ability to accurately reflect conformational dynamics, most notably by their restriction to a linear model and blindness to lateral motion. Third, network resolution of available tools was rigid, as they focused either on residue or single atom level, with little support to treat certain protein regions as residues while zooming into specific regions of interest, e.g., individual atoms of small molecule ligands

# 1.6. Investigated protein systems

## 1.6.1. PDZ2

PDZ domains are a ubiquitous class of protein domains involved in the formation of protein complexes (125-127). They recognize C-terminal or short internal peptides (128, 129), which activates recruitment of other proteins and mediates complex assembly (125-127). In contrast to classical models of allostery, ligand-induced activation of PDZ2 does not lead to substantial conformational differences observed in crystal structures (130). The PDZ2 domain is a commonly used benchmark system to evaluate allosteric prediction models using a comprehensive set of experimental data. The dataset was the result of a series of studies performed by Lee and coworkers, who investigated the effects of ligand binding and residue mutations on backbone and sidechain dynamics of individual residues, as measured by NMR spin relaxation (130-132). As the dataset contains mainly residues containing methyl groups, insights derived from it remains open for expansion by computational predictions. However, a review comparing the results of several prediction models, reported by separate publications, found substantial disagreement between those methods (133). This problem was exacerbated by the fact that many of these methods reported only lists of predicted residues instead of raw prediction scores, making a direct comparison difficult, since each method was set to a different sensitivity. Moreover, most of the compared methods only described algorithms while lacking readily available implementations, impeding repetition of the same prediction with different thresholds or on other protein systems. The open questions concerning PDZ2 biology, divergence of previous prediction results and the lack of systematic comparative analyses between models justified further investigations into PDZ2. The PDZ2 system was therefore chosen as the first system to evaluate the network-based allosteric prediction models in this thesis, with particular attention given to establishing a clear connection between predictions and experimental data, as well as consolidating our results with previous models.

### 1.6.2. UHRF1

The protein UHRF1 is a regulator of DNA methylation maintenance, acting as a probe for hemi-methylated CpG sites (134-136). It is composed of a ubiquitin-like domain, a Tandem-Tudor domain, a PHD domain, a SRA domain, and a RING domain (Figure 3). After binding to a hemi-methylated CpG site, i.e., a site where one DNA strand is methylated but the other is unmodified, the SRA domain of UHRF1 recruits the methyltransferase DNMT1, which subsequently methylates the second strand (134-136). In addition to its role in targeting DNMT1 action, UHRF1 recognizes specific patterns of post-translational histone modifications and acts as an E3 ubiquitin ligase (137). Its multitude of functions highlight the central role UHRF1 plays in cell maintenance, DNA damage repair and genetic regulation (138).

The DNA modification pattern is the result of a dynamic balance between methylation and demethylation processes. Proteins of the TET family (TET1, TET2 and TET3) are able to remove the methyl group of 5-methylcytosine (5mC) via step-wise oxidation with the intermediates 5-hydroxymethylcytosine (5hmC), 5-formylcytosine (5fC) and 5-carboxylcytosine (5caC) (139). However, it has been suggested that these intermediates may also serve a functional role as specialized epigenetic markers (140). Genome mapping studies found detectable levels and accumulation of intermediates in distinct DNA regions (141-143), whereas specific cell types and conditions could be associated with increased concentrations of these modifications, such as hmC in neuronal cells (144) and caC in tumor cells (145). The subtle chemical differences between the mC and its oxidized derivatives hmC, fC and caC raised the question whether UHRF1, as an established mC reader, would be able to recognize those variants as well.

Investigations into the specific recognition mechanism showed that UHRF1 binds to the DNA helix and flips the hemi-methylated DNA base out of its strand, pushing it into a deep-seated binding pocket (135). In parallel, UHRF1 inserts a flexible loop called the NKR finger into the major groove of the DNA, forming hydrogen bonds with the DNA backbone. The NKR finger is in direct contact to the CpG site’s cytosine on the opposite strand, i.e., the position that is methylated after DNMT1 recruitment. A CpG site in which both strands are methylated leads to a steric clash with the NKR finger, and binding studies have demonstrated that UHRF1 shows greatly diminished binding affinity to these fully methylated CpG sites (146, 147). Allosteric effects were reported for the UHRF1 system in connection to its histone binding activity, but not within the limited context of the single SRA domain (148-150). We had received preliminary experimental data that showed an intriguing increase of binding affinity for a fully carboxylated CpG site. This prompted further investigations, during which we detected signs of conformational coupling between the DNA base binding pocket and the NKR finger. The analyses we performed would form the basis for the network models developed in this thesis. An earlier analysis as part of the author's Master's Thesis (151) had shown promising trends indicating potential for a network-based analysis; however, attempts to replicate these results failed. A follow-up investigation showed flaws in the original models and simulation setup. Hence, the project was started completely from scratch, with no models or data shared with the originally reported results. The new setup used newly derived computational models, force field parameter sets, a different simulation engine, much more extensive simulations, multiple replicas, and substantially improved analysis methods. All results presented here are derived from the new setup.

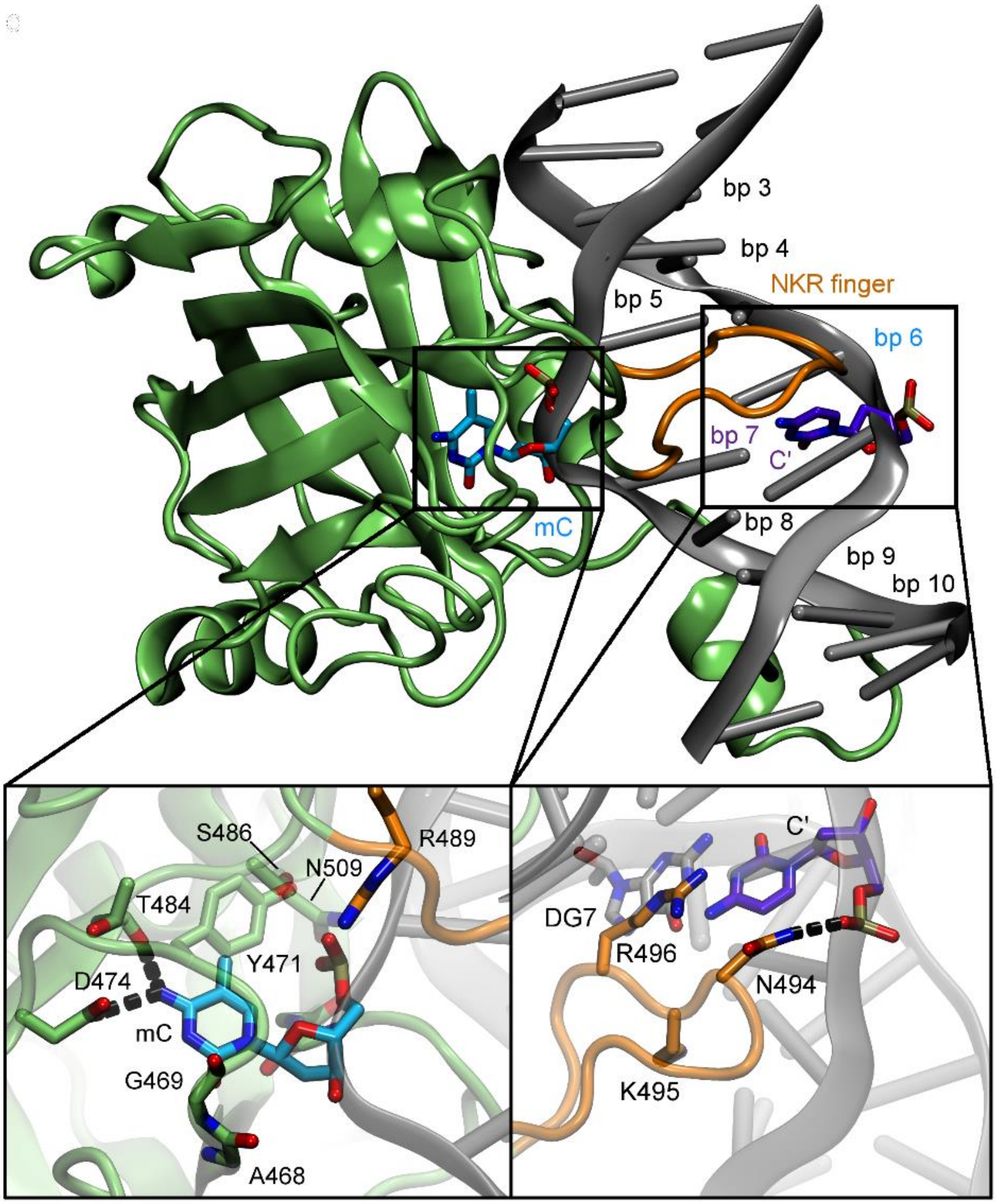


**Figure 3. Structure of the UHRF1—DNA complex.** Molecular dynamics structure of the SRA domain of UHRF1 bound to hemi-methylated DNA. Insets show a magnification of the nucleotide binding pocket and NKR finger regions. Figure cropped from ref. (193).

## 1.6.3. Hsp70 chaperones

Hsp70 (Heat shock protein 70 kDa) chaperones are a protein class whose function relies on a complex conformational cycle, regulated by an extensive allosteric network (152-157). Advancement of the conformational cycle is achieved by binding different ligands and cochaperones, which is sensed and communicated through the protein by a network of allosteric residues. As chaperones, Hsp70's fulfill many roles, including the support of correct protein folding, regulation of apoptosis, as well as supporting membrane translocation and de-aggregation of misfolded complexes (153, 154, 157-160). The Hsp70 chaperone achieves this by binding hydrophobic stretches within proteins, which are exposed when a protein is partially or fully unfolded. The binding of the chaperones prevents further folding at those protein stretches until they are released. In concert with its cochaperones, Hsp70 activity can guide its protein

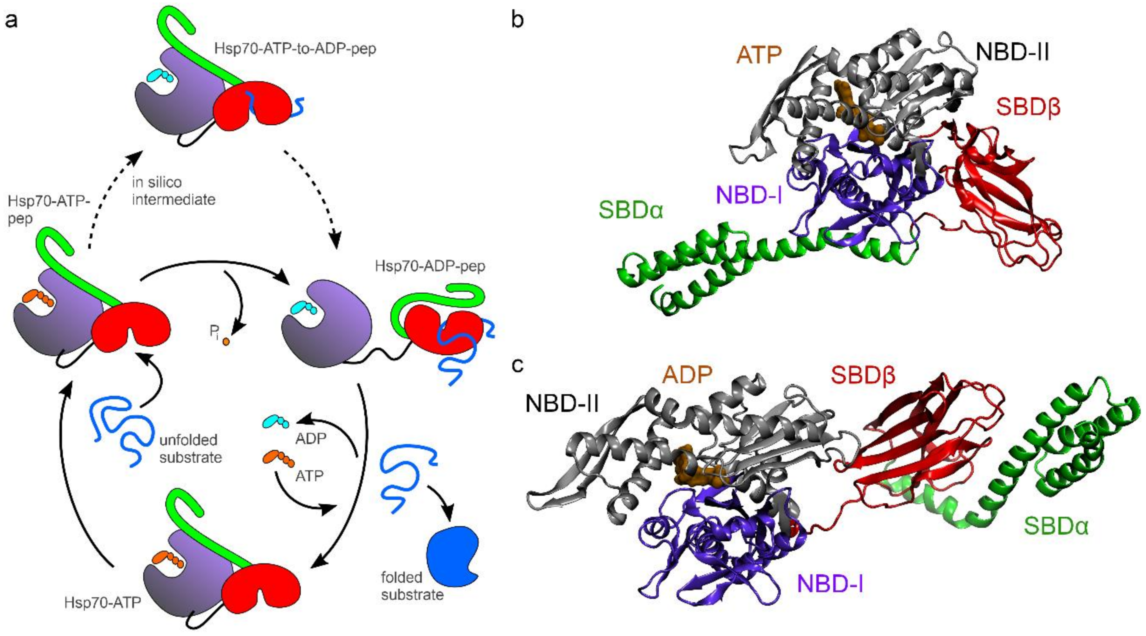


**Figure 4. Structural organization and conformations of Hsp70 chaperones.** (a) Simplified representation of the Hsp70 conformational cycle. (b,c) Representative structures of DnaK-ATP (b) and DnaK-ADP (c) extracted from molecular dynamics simulations. Figure adapted from ref. (195).

substrates towards different processes, from refolding to degradation (154). Due to their ubiquity, including several isoforms in humans, and their association with cell repair and neurogenerative diseases, Hsp70s have been proposed as potential therapeutic targets (153, 159, 161-163).

The conformational cycle of Hsp70s is characterized by a series of major domain rearrangements, most strikingly the complete docking and undocking of its nucleotide binding (NBD) and substrate binding (SBD) domains (152-155, 164). The NBD is further subdivided into two rotatable lobes (NBD-I and NBD-II), while the SBD consists of a core subdomain housing the substrate binding pocket (SBDβ) and an α-helical lid domain that can close over the binding pocket, locking a bound substrate in place (SBDα). In Hsp70's "open" conformation, the NBD lobes envelop an ATP molecule and the SBDβ binding pocket is empty (Figure 4). The NBD and SBDβ domains are docked onto each other, with the SBDα lid open and bound to the NBD-I lobe. The binding of a substrate, either a peptide or short internal protein stretch, to the SBDβ binding pocket triggers a partial undocking of the NBD-SBDβ interface (165). This allows the NBD lobes to rotate into a conformation favoring hydrolysis of the ATP molecule, followed by completion of the NBD-SBD undocking and closing of the SBDα lid over the SBDβ binding pocket. This stable, closed conformation features a fully undocked NBD-SBD, ADP bound to the NBD, and a peptide substrate bound to the SBDβ. The cycle is completed by release of the substrate, exchange of ADP to ATP and domain re-docking (152-156). The conformational changes are facilitated through a network of allosteric residues, which have been studied primarily through mutation of individual residues and detecting associated effects on protein activity in the E.coli Hsp70 variant DnaK (155). These residues sense and adapt their conformations both as a response to the different ligand binding states, but also due to binding and interactions with a number of cochaperones, e.g. DnaJ and the nucleotide exchange factor grpE for DnaK (152-154, 156, 166-168). However, due to the large size of Hsp70 proteins (over 600 amino acids), assembling a complete description of the allosteric network by individual mutations is a laborious and unfinished process. Furthermore, evolutionary adaptation of Hsp70 variants to different organisms or organelles has led to a high sequential variance within the protein family, and it is unclear how insights obtained for the well-studied E. coli chaperone DnaK are transferable to these specialized variants.

The Hsp70 BiP is a human chaperone variant native to the endoplasmatic reticulum, specialized to folding proteins passing through membranes and the secretion pathway (169, 170). In addition, cellular surface BiP (csBiP) can be detected in some cell types under stress conditions, along with proliferation of BiP to the cytosol, nucleus and mitochondria (171-173). The occurrence of csBiP in tumor cells opens the possibility for applications in cancer treatment (174-178) or as an antiviral target, notably including for the SARS-CoV-2 virus (179-182). As is typical for Hsp70's, DnaK and BiP share a highly similar structure, but their sequence homology is below 50 %. While the proteins fulfill similar core functions, a number of differences have been reported with respect to their biochemical behavior: BiP interacts with a different and more expansive set of cochaperones, important for allosteric regulation, and showed substantial differences in NBD-SBD docking and SBDα lid dynamics (168, 183) as well as their functional behavior (184).

Binding of the peptide substrate is an essential step in the conformational cycle of Hsp70 and its allosteric regulation. In accordance to its wide-ranging chaperone functions, Hsp70s can bind to a large variety of peptides, with a notable bias towards sequences containing hydrophobic residues. The substrate binding site, located in the SBDβ subdomain, can fit a peptide stretch of five to seven amino acids (185-188). Extensive studies revealed that Hsp70 variants showed different preferences in substrate recognition (189, 190). The residue located in the center of this binding pattern is the one most deeply enveloped by the binding pocket. For example, it was observed that while DnaK favored hydrophobic but smaller residues, BiP would also accept more bulky aromatic amino acids like tryptophane in its binding pocket (191). This indicates substantial differences in the binding pocket structure between DnaK and BiP, with potential implications for the mechanisms of allosteric activation. As the development of small molecule Hsp70 allosteric modulators progresses (162), applications targeting specific proteins like BiP will require a detailed understanding of the allosteric mechanisms and the evolutionary differences between Hsp70 variants. Despite the reported differences between DnaK and BiP with respect to substrate recognition, dynamics and function, the underlying residue mutations giving rise to these behaviors are not well explored, aspects which could prove important for potential therapeutic applications.

## 1.7. Objectives and thesis structure

The goal of this thesis was to accurately predict and characterize protein regions associated with conformational coupling and allosteric regulation within Hsp70 chaperones. Relevant structural and dynamical differences between two members of the Hsp70 family, DnaK and BiP, were to be analyzed to deepen understanding of the influence of individual mutations on biochemical function and the effects of evolutional divergence. Insights gained from these investigations were expected to help elucidate the mechanisms of conformational coupling and ligand induced conformational changes, enabling applications in protein engineering and allosteric drug design. In the following, we describe in detail the three phases of the project: Exploratory analyses of Hsp70 chaperone systems including characterization of their structural differences enabling specific ligand binding preferences, development and validation of a novel allosteric prediction model based on protein structure networks, and finally application of this model to investigate evolutionary differences in Hsp70 allosteric regulation following ligand binding. The results of our investigations were reported in four published manuscripts. In the first publication, we combined experimental peptide array data, molecular docking, and statistical learning to predict peptide sequences recognized by the human Hsp70 chaperone BiP. Our structural analysis distinguished the binding pockets of BiP in contrast to the E. coli variant DnaK and provided a quantitative model to predict likely BiP-binding regions in proteins (192). Next, we investigated the structural coupling between two key protein regions in UHRF1, employing network analyses of ligand interaction patterns to explain the observed conformational rearrangements (193). These ideas were further developed in the third publication, which formalized our network model and provided a comparative analysis of different algorithms for allosteric prediction, while evaluating their accuracy using the PDZ2 benchmark system (194). Finally, we utilized our previously established network models to investigate the transmission of ligand binding signals and subsequent allosteric regulation in the Hsp70 chaperones DnaK and BiP (195). We detect pathways of conformationally coupled residues in alignment to experimentally verified allosteric residues, and predict several new allosteric candidates. Comparing prediction results of DnaK and BiP, we were able to suggest several specific residues with specific roles in either Hsp70 variant, including a secondary allosteric pathway unique to BiP. Our predictions have the potential to contribute to the understanding of evolutionary adaptation of proteins to specific environments and the development of allosteric drugs targeting specific protein variants. The ability to use our reference implementation SenseNet together with a variety of data sources, including NMR ensembles and existing MD simulations, allows for easy application of our prediction models on other protein systems.

# 2. GENERAL METHODS

## 2.1. Conformational sampling using Molecular Dynamics

Molecular Dynamics (MD) encompasses a set of simulation techniques to generate structural ensembles of a molecular system by iterative calculation of new conformations based on the forces acting on its atoms. The forces between atoms are described using a force field, a parameterized function to calculate the potential energy of a system. The force field combines energy functions and parameters aimed at reproducing quantum mechanical potentials and/or experimental observations from small molecules. Popular force fields for all-atom simulations of proteins include the most recent iterations of the AMBER force field family Amber-ff14SB (196) and Amber-ff19SB (197), CHARMM36 (198), or OPLS-AA (199) with the updated parameter sets OPLS-AA/L (200) and OPLS-AA/M (201). In a typical force field (202), covalent bonds are represented as harmonic spring potentials, with a sinusoidal function term modelling the torsion potential (eq. 2.1). The noncovalent interactions are composed of an electrostatic term and a Lennard-Jones potential term. Further terms can be added for correction (e.g., enforcing planar aromatic rings) or to reflect different dynamics (e.g., hydrogen bonding terms). Given a starting structure of atom positions and a force field as in eq. 2.1, the potential energy of the system can be determined as

$$\begin{aligned} E_{pot}(\vec{r_i}) = \sum_{bonds} k_d(d - d_0)^2 + \sum_{angles} k_\theta(\theta - \theta_0)^2 + \sum_{dihedrals} V_n(1 + \cos(n\phi - \gamma)) \\ + \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \left[\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}}\right] + \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \left[\frac{q_i q_j}{\varepsilon_i r_{ij}}\right] \end{aligned} \qquad (2.1)$$

where $\vec{r_i}$ is the configuration vector of atom coordinates, $d, \theta, \phi$ are bond length, bond angle and dihedral angles of the current configuration and the zero subscripts denote respective reference parameters. Furthermore, $k_d, k_\theta, V_n$ are force constants for bonds, angles, and dihedrals, while $n, \gamma$ parameterize the shape of the dihedral potential. The first three terms represent the covalent interactions of the potential, while the fourth and fifth term corresponds to the non-bonded Lennard-Jones (LJ) and electrostatic interactions, respectively. The $A_{ij}, B_{ij}$ parameters determine the strength of LJ interactions, $r_{ij}$ corresponds to the distance between atoms $i$ and $j$, $q_i$ is the partial atomic charge of atom $i$, and $\varepsilon_l$ is the effective dielectric constant.

The equations of motion used to model the dynamics of the system require a vector of atom coordinates and at least its first two derivatives, namely velocity and acceleration. The force field equation (eq. 2.1) yields the acceleration via Newton's second law

$$\vec{F_\iota} = -\frac{dE_{pot}}{d\vec{r_\iota}} = \frac{d\vec{p_\iota}}{dt} = m_i \frac{d^2\vec{r_\iota}(t)}{dt^2} \tag{2.2}$$

which, given a snapshot of current atom coordinates and velocities, is sufficient for numerical simulation. To obtain the next snapshot of the simulation, corresponding to advancing the simulation by a small timestep, the equations of motion are integrated numerically. Common options include for example the Leapfrog or Verlet integrators (203).

The environment of a protein is an essential factor during simulation. In our simulations, protein and DNA molecules were placed into a cubic box of water molecules and ions to achieve a predefined salt concentration chosen to match experimental buffer conditions. To avoid creating a border region between solvent and what would effectively constitute effective vacuum at the faces of the box, the entire system was mirrored into all directions (periodic boundary conditions). The simulation of an approximately infinite environment significantly increases the computational cost as many more atom interactions must be considered. For interaction terms with a short range, i.e., LJ interactions, cutoffs are used to minimize the number of required calculations. On the other hand, electrostatic interactions have longer range and affect protein stability. Direct summation of these interactions with cutoffs is error prone and suffers from convergence issues. Therefore, electrostatic interactions were treated using particle mesh Ewald (PME) summation, which allows efficient computation of interactions in infinite periodic systems. In PME, the interatomic potential energies are separated in a short-range part, which is calculated in real space, and a long-range part calculated in Fourier space (204). It can be shown that in their respective domains, both the real and Fourier contributions converge quickly and can therefore be truncated without loss of accuracy. In addition, Fast Fourier transformation is used for even more efficient computation of the Fourier term (205).

MD simulations which evolve solely based on the equations of motion approximate the NVE (or “microcanonical”) ensemble, that is a structure ensemble in which the number of particles N, the volume V and the total Energy E are conserved within algorithmic and numerical limits. However, this allows both the temperature T and pressure P of the system to fluctuate, which are constant in typical experimental conditions. Therefore, simulations typically use the NVT (“canonical”) ensemble, in which temperature is controlled using a thermostat, or the NPT (“isobaric-isothermal”) ensemble, which uses a thermostat in addition to a barostat regulating pressure. Within this thesis, two different thermostats were used. The Langevin thermostat is a popular option as it is straightforward and can help overcoming small energy barriers which improves efficiency of the simulation. The equations of motion are modified to

$$m_i \frac{d^2\vec{r_i}}{dt^2} = \vec{F_i}(\vec{r_i}) - \xi \frac{d\vec{r_i}}{dt} + \vec{R_i}(t) \qquad (2.3)$$

Where $\vec{F_i}$ is the inter-atomic force according to eq. 2.2, $\xi$ is a frictional coefficient, and $R$ is a random force which averages to 0. The targeted temperature can be maintained by adjusting $\xi$ (slowing down atoms) and $R$ (accelerating atoms) accordingly. The second thermostat used in our simulations is the Berendsen or “weak coupling” thermostat. It is one of the oldest strategies for temperature control, based on rescaling the temperature of the system gradually using

$$\frac{dT}{dt} = \frac{T_0 - T(t)}{\tau} \qquad (2.4)$$

where $T_0$ is the reference temperature, and $\tau$ is the weak time coupling constant. A known weakness of this thermostat is that it does not strictly generate the canonical ensemble, although this error is expected to be small in large systems, i.e., proteins including a solvent box. More substantially, it has been noted that the thermostat may unphysically shift kinetic energy towards slowly fluctuating degrees of freedom, leading to the “flying ice cube” effect (206). This has led to sometimes very strong recommendations against the use of the Berendsen thermostat (207), as other alternatives are often favored. While the problems of the Berendsen thermostat are plentiful and well appreciated, it should be noted that it is unclear whether they have a significant impact on systems of average protein size (207). Lagging availability of modern thermostats in MD software still make the Berendsen thermostat a possible (though hopefully soon obsolete) choice, especially since alternative implementations are not always fully understood and come with their own drawbacks. In contrast to e.g., the popular Langevin thermostat, the Berendsen thermostat does not introduce a random force component, which makes it in principle better suited for studies of correlated motion. It was reported that the flying icecube effect can be partially counteracted by setting a large time constant in eq. 2.4, which gives the system more time to equilibrate and prevents the detrimental accumulation of kinetic energy (207). To control pressure during simulations, an analogue of Berendsen’s thermostat was used, which rescales the volume of the box to set the desired pressure using the formula

$$\frac{dp}{dt} = \frac{p_0 - p(t)}{\tau_p} \tag{2.5}$$

where $p_0$ and $\tau_p$ are the pressure equivalents to eq. 2.4. In practice, volume fluctuations due to pressure control tend to be very small for protein systems. For this reason, it is not expected that the Berendsen barostat suffers from the same problems as its thermostat equivalent, leaving it a viable option for simulations for the time being.

## 2.2. Combined Sequence- and Structure Based Prediction of Ligand Binding

Hsp70 chaperones are promiscuous binders which recognize a wide range of protein substrates (189, 190). Despite this flexibility, high-throughput peptide binding experiments have been able to discern characteristic binding motifs. As Hsp70s are known to bind an elongated, unfolded peptide stretch of about 5 to 7 amino acids, many prediction models have been built on the premise that the binding affinity can be approximated by combining the contributions of individual independent subpockets (208). These methods commonly start from a set of strongly binding substrate peptides, determined by experiments, followed by generating a Position-specific scoring matrices (PSSM). A PSSM encodes the probability of finding each amino acid (rows) in each of the subpockets (columns) among the strong binders. Predictions for peptides are performed from the sequence of the candidate peptide, looking up the scores for the matching amino acids from the PSSM columns, and summing them to obtain a final score.

In our work, we diverge from this approach, by instead calculating a PSSM based on structures of the protein-substrate complex. This structure-based position-specific scoring matrix (SB-PSSM) uses interaction energies calculated from a force field to estimate the contributions of each position to the overall binding affinity. First, a structure model of the Hsp70 protein bound to a seven amino-acid substrate was created, providing a template for the peptide backbone. In our study, we chose a homology model of BiP created using MODELLER (209) from a structure of DnaK bound to the HTFPAVL peptide. Then, the substrate peptide was mutated in silico using IRECS (210) to the AAAPAAA peptide as the baseline structure. Next, we used this baseline structure to create variants placing each of the 20 canonical amino acids at each of the 7 positions using IRECS, while keeping the other positions constant. This systematic amino acid scan corresponded to a mutation pattern of [XAAPAAA, AXAPAAA, AAXPAAA, …], where X was the position to be mutated. After creating a total of 20 x 7 structures, where each amino acid was placed once into each position, an energy minimization was performed to ensure that the interaction energies were obtained from a relaxed state. Finally, individual energy terms between protein and substrate are extracted from the minimized structure. Energy minimization and interaction energy calculations are based on the OPLS-AA force field (199) in order to allow for the use of Pepscore coefficients (211). The Pepscore was designed as a scoring function for molecular docking, providing weighting coefficients for transforming force field energy terms into a composite score to rank docking conformations. For each position in the SB-PSSM matrix, the structure with the corresponding protein-substrate mutant was prepared, energy minimized and the score obtained as the Pepscore-weighted sum of the Coulomb and Lennard-Jones interaction energy terms.

During our studies, it was found that the SB-PSSM required corrections for certain edge cases, for example when a residue was not placed properly in the sub-pocket due to limited space. In such cases, values were adjusted manually following an analysis of the binding pocket. As a method to improve the predictive performance of the SB-PSSM, we implemented an integrated approach combining aspects of sequence-based und structure-based prediction. It was apparent from the pattern of strongly binding sequences that central residues of the substrate peptide were stronger predictors than those at the termini of the peptide; however, this was not well reflected by taking the unweighted sum of SB-PSSM column values. We therefore set out to optimize the coefficients of the SB-PSSM column scores using logistic regression. First, we obtained a set of experimental binding data, based on Fluorescence Anisotropy Spectroscopy measurements of peptide arrays and literature sources (183, 191, 212, 213). Based on these measurements, the peptides were categorized as binders/non-binders and collected in the dataset $(x_i, y_i),\ i = 1, \dots, l$ where $l$ is the total number of peptides. Each peptide $i$ in the dataset is described by a training vector $x_i \in R^n$ (the SB-PSSM scores for each position within the peptide) and a binder/non-binder class label $y_i = [1, -1]$. The prediction model is then given by the logistic function

$$p(x_i) = \frac{1}{1 + e^{-w^T x_i}} \tag{2.6}$$

as the vector of weight coefficients $w \in R^n$, which is determined numerically by minimizing

$$\min_{w} \frac{1}{2} w^T w + C \sum_{i=1}^{l} \log\left(1 + \ e^{-y_i w^T x_i}\right) \tag{2.7}$$

where $C > 0$ is a cost parameter and the first term providing for L2 regularization. Implementation details for solving eq. 2.7 can be found in ref. (214); in our work, the implementation provided by the python package scikit-learn (215) (version 0.16.1) was used . The regularization cost parameter $C$ was optimized separately by three-fold cross-validation for the best area-under-curve (AUC) value in the corresponding receiver operating characteristic (ROC) curve.

## 2.3. Analysis of conformational coupling in protein structure networks

The SenseNet network model was designed to extend on previous implementations of protein structure networks by including data from structure ensembles obtained from MD trajectories. The most common strategy to define a protein structure network is to map protein residues to nodes and to connect these nodes with edges representing close range interactions, such has hydrophobic contacts or hydrogen bonds. As we intended to investigate conformational coupling using protein structure networks, we considered several programs which allow network analyses of MD trajectories (see section 1.5). However, our evaluation revealed that they lacked one or more capabilities which we required for our project: First, an improved measure of correlation over Pearson's coefficient; second, avoiding reliance on centrality measures due to their instability problems when applied to networks obtained from MD trajectories; third, the model should allow analyses on different levels of resolution, from residues to atoms; and finally, the software should be publicly available and open source to be readily reproducible. As none of the available tools provided all desired features, we decided to define a custom model and implementation (Figure 5).

The fundamental building block of the SenseNet network model is provided by the interaction timeline. For each structure in the ensemble, all pairs of atoms are scanned to determine whether an interaction is present in that structure (e.g., a hydrogen bond, salt bridge or a contact between nonpolar atoms). An interaction is counted if the atom pair fulfills a set of criteria, e.g., specific atom types and/or within a maximum distance. If the criteria are fulfilled, the timeline will show a 1 at the structure's position in the trajectory; otherwise, it will be 0. The full atomistic interaction timeline is then given as

$$\boldsymbol{X}_{\alpha\beta k} = \left[\begin{cases} 1 & \text{if } \alpha \text{ and } \beta \text{ interact as type } k \text{ in trajectory frame } t \\ 0 & \text{otherwise} \end{cases}\right]_t \tag{2.8}$$

where $\alpha, \beta$ are atom nodes, $k$ is an interaction type and $t$ is a simulation snapshot at a given timepoint (trajectory frame). At the atom level, the resolution of interactions is very high but too fine-grained for intuitive analyses. Thus, timelines can be combined to represent a coarser but more intuitive description of the system, e.g., in terms of interactions between residues. To obtain a combined timeline, the atomistic timelines are summed element-wise as

$$\boldsymbol{X}_{\boldsymbol{ijk}} = \sum_{\alpha \in i} \sum_{\beta \in j} \boldsymbol{X}_{\boldsymbol{\alpha\beta k}} \tag{2.9}$$

with $i, j$ as nodes representing residues encompassing one or more atoms. Multiple edges can connect the same node pair if they represent different interaction types, e.g., carbon-carbon contacts and hydrogen bonds. The connectivity between node pairs is described for each interaction type by the symmetric adjacency matrix

$$\boldsymbol{A}_{\boldsymbol{k}} = \left[\begin{cases} 1 & \text{if } i \text{ and } j \text{ are connected by an edge of type } k \\ 0 & \text{otherwise} \end{cases}\right]_{ij} \tag{2.10}$$

for each interaction type $k$. These structures are the central feature of the network model and offer a wide range of possibilities for further analyses. In contrast to networks where edges are based on correlation coefficients, the SenseNet method

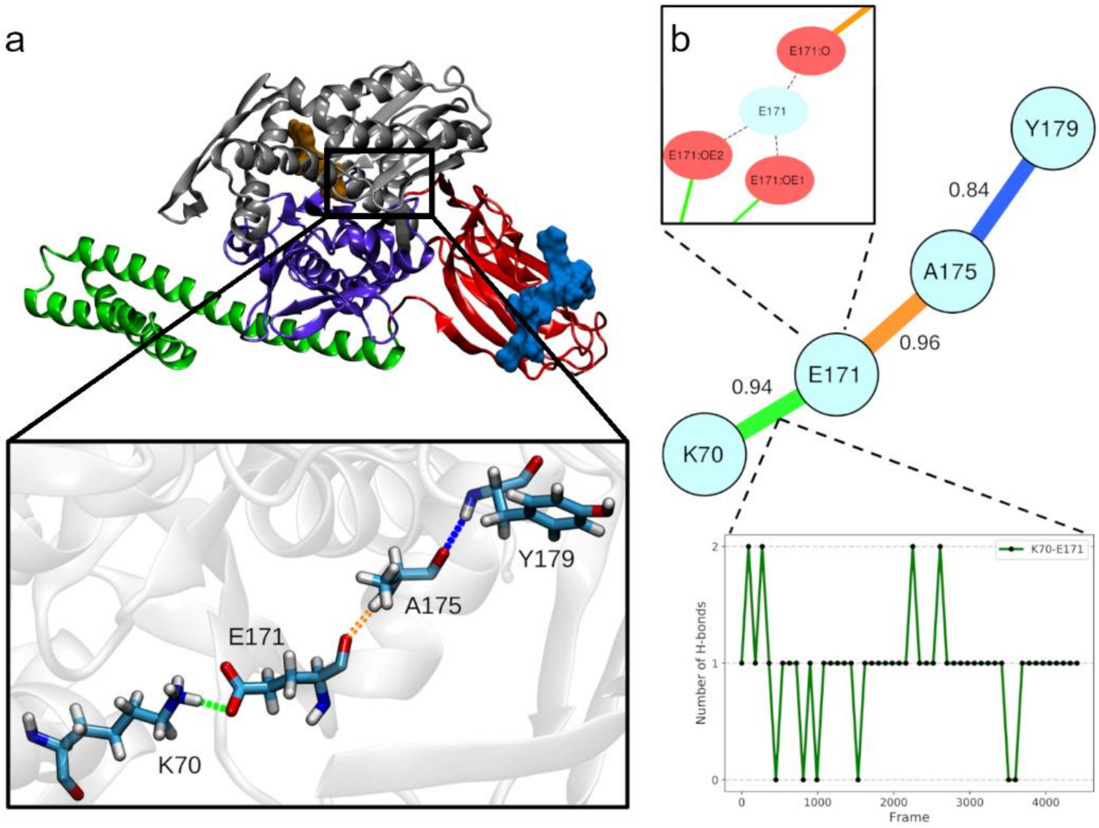


**Figure 5. The SenseNet network model.** (a) Networks are extracted from protein structure ensembles obtained from molecular dynamics. Residues make up the nodes of the network, while atom interactions, like hydrogen bonds, are modelled as edges. (b) Dynamic time- and spatial resolution in structure ensemble networks. Residue nodes can be split up into individual atom nodes, whereas timelines track the dynamic evolution of interactions within the structure ensemble.

treats structure ensembles as a natural extension of a single structure model. This can be easily verified by observing that the timeline-based model reduces to the conventional single structure network if the length of the timeline is one. This allows to use algorithms established for single structure networks by applying them on individual trajectory frames of the timeline. Alternatively, network analyses can be performed on a network representing the average of all structures. In this form, timelines are commonly weighted by their time average

$$avg(X_{\alpha\beta k}) = \frac{1}{T}\sum_{t} X_{\alpha\beta k}\,(t) \tag{2.11}$$

where $T$ is the total number of frames in the timeline. This value corresponds to the average number of interactions between two residues or atoms in the structure ensemble. Due to the flexibility of residues, ensembles generated by molecular dynamics contain many spurious interactions which continuously form and dissolve during simulation. A common approach to minimize the effect of those spurious interactions is to limit analyses to interactions which were present for a minimum fraction of the simulation

$$occ(X_{\alpha\beta k}) = \frac{1}{T}\sum_{t} \min\,(1, X_{\alpha\beta k}\,(t)) \tag{2.12}$$

which we call the occurrence of an interaction.

The network model can be understood as a low-resolution description of the structure ensemble, transforming atom coordinates into distinct interaction states. For example, in a network consisting entirely of hydrogen bond edges, the conformations of the ensemble are distinguished by the number of hydrogen bonds between the residues at any snapshot of the trajectory. Focusing analyses on these interactions allows for an intuitive encoding of conformational state and more sophisticated methods of quantifying correlation, avoiding the collinearity limitations of measures based on atom coordinates and Pearson’s coefficient. In addition, interaction timelines (eqs. 2.8 and 2.9) can be used with arbitrary network resolution, i.e., from atom to residue level, or a mix between the two. The remainder of this section elaborates on how this network model can be applied to measure conformational correlation in MD trajectories.

Using the discrete mutual information from eq. 1.6 as $\mathrm{I}(\boldsymbol{X_{ijk}};\boldsymbol{X_{nml}})$, we define an edge correlation factor (ECF) as

$$\mathrm{ECF}(i,j,k) = (\boldsymbol{A_k})_{ij} \cdot \sum_{l \in K} \sum_{n,m \in N} \mathrm{I}\left(\boldsymbol{X_{ijk}}; \boldsymbol{X_{nml}}\right) \cdot (\boldsymbol{A_l})_{nm} \cdot \chi_{ijk}(n,m,l) \tag{2.13}$$

where $k$ and $l$ are interaction types, $i, j, n, m$ are elements of the node set $N$, and $\chi$ is an indicator function which yields 1 if the edge corresponding to timeline $X_{nml}$ is adjacent to the edge of $X_{ijk}$ in the network, and 0 otherwise. The concept is intuitively visualized in Figure 6.

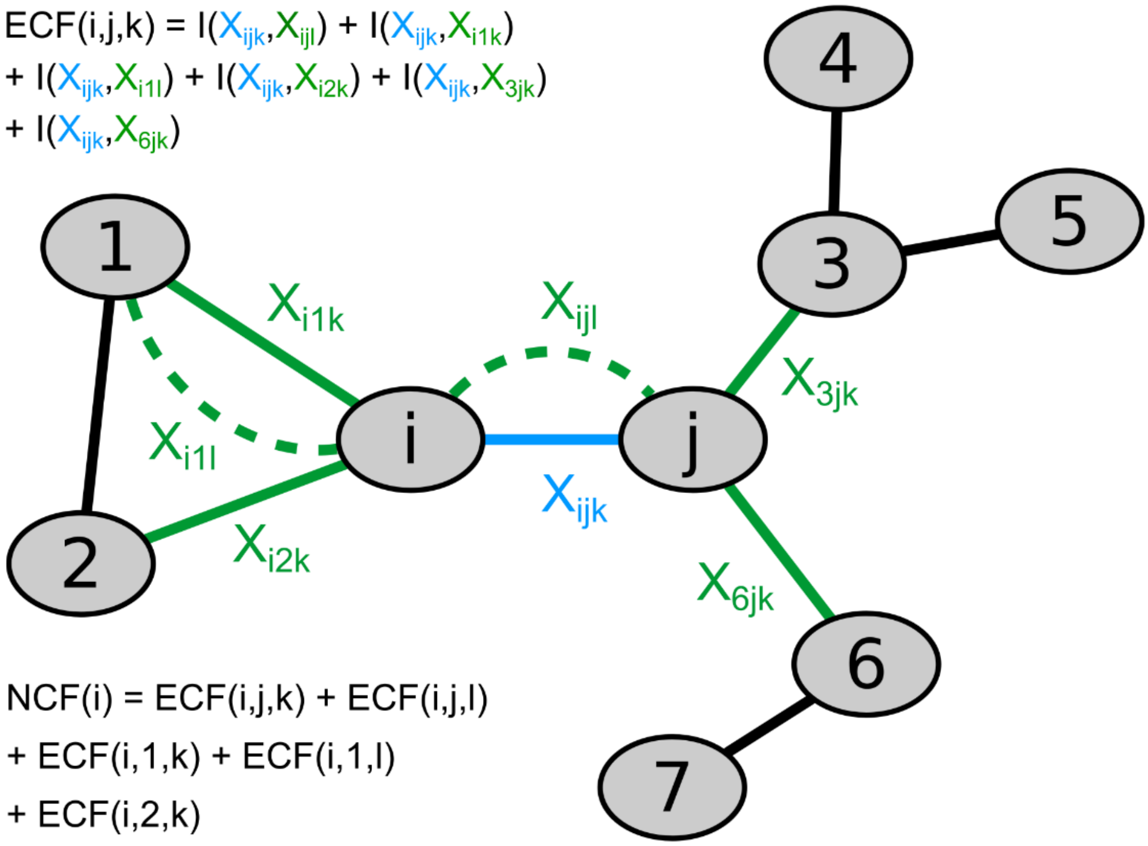

ECF(i,j,k) = I(X_ijk,X_ijl) + I(X_ijk,X_i1k)
+ I(X_ijk,X_i1l) + I(X_ijk,X_i2k) + I(X_ijk,X_3jk)
+ I(X_ijk,X_6jk)
1
2
3
4
5
6
7
i
j
X_i1k
X_ijl
X_3jk
X_i1l
X_ijk
X_i2k
X_6jk
NCF(i) = ECF(i,j,k) + ECF(i,j,l)
+ ECF(i,1,k) + ECF(i,1,l)
+ ECF(i,2,k)

**Figure 6. 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*. Figure from ref. (194).

Limiting evaluated correlation to adjacent edges in the network emphasizes the local effects of interactions on their immediate network, where long range effects are propagated consecutively through clusters or chains of localized conformational changes. This ensures that the contributions to the ECF are dominated by local motions, remaining unaffected by spurious correlation between distant protein regions, most of which are unlikely to be coupled in a functional manner. On the downside, should any coupling between distant residues occur that does not manifest in any way within the conformations of intermediate residues, it would not be detectable with this approach. High ECF scores indicate strong correlation of an interaction with other interactions in its direct environment; it is assumed in the model that a change in the interaction pattern, for example by binding of a ligand, would then cause conformational changes in these interactions due to conformational coupling. In other words, the ECF score conveys how much information an interaction provides about the interaction states in its environment. For the purpose of computational predictions, scoring residues is usually more desirable than individual interactions. This description corresponds more closely to commonly reported residue-based experimental data, like mutagenesis experiments or NMR spin couplings. In our work, we achieved this by summing ECF scores of each node's adjacent edges to node correlation factors (NCF)

$$\mathrm{NCF}(i) = \sum_{k \in K} \sum_{j \in N} \mathrm{ECF}(i, j, k) \tag{2.14}$$

Where $K$ represents the set of interaction types in the network.

ECF and NCF scores provide a model of conformational correlation between residues within a single structure ensemble, as obtained e.g., from a MD simulation. It is a well appreciated limitation of conventional MD simulations that they tend to cover only a limited set of possible system states. For example, simulating the dynamic binding and unbinding of ligand molecules requires prohibitively large computational effort using conventional MD. To address this problem, we modified the definition of the ECF (eq 2.13) so it could account for the differences observed from two simulations representing different states of the protein. First, one of the simulations was assigned as the reference system. The mutual information term of eq. 2.13 was then replaced by

$$\mathrm{I}(\boldsymbol{X};\boldsymbol{Y}) = \sum_{x \in \cup(\boldsymbol{X},\widehat{\boldsymbol{X}})} \sum_{y \in \cup(\boldsymbol{Y},\widehat{\boldsymbol{Y}})} \left| p(x,y) \cdot \log_2\left(\frac{p(x,y)}{p(x)p(y)}\right) - \hat{p}(x,y) \cdot \log_2\left(\frac{\hat{p}(x,y)}{\hat{p}(x)\hat{p}(y)}\right) \right| \tag{2.15}$$

where $\boldsymbol{X}, \boldsymbol{Y}, p$ correspond to a timeline in the analyzed network (e.g., a protein bound to an allosteric ligand), and $\widehat{\boldsymbol{X}}, \widehat{\boldsymbol{Y}}, \hat{p}$ correspond to the equivalent timeline of the reference simulation (e.g., the same protein without a ligand). Eq. 2.15 calculates the composite score by subtracting the pointwise mutual information of co-occurring timeline states. In contrast to the base ECF score, which is based on a single simulation, the updated formula tracks the differences between the two structural ensembles. In structural terms, this could occur for example when the binding of an allosteric ligand causes rigidification of a protein region or otherwise changes the dynamics of involved residues. To distinguish scores calculated with the modified formula eq. 2.15, we define them as “difference edge correlation factor” (DECF) and “difference node correlation factor” (DNCF) for the NCF analogue, respectively.

Due to the limited time scale of MD simulations, differences in residue dynamics induced by ligands are strongest close to binding sites, but may not be detectable at longer distances. As an alternative to performing more and longer simulations, we developed a variant combining the DNCF method with ideas from shortest path centralities (see section 1.3). The centrality model presumes that information travels along the shortest, i.e., the most efficient path along the network. This idea can be implemented by performing a random walk through the network. In addition, the DNCF scores give an estimate for the strength of conformational coupling between neighboring residues. Combining these approaches yields a random walk directed by DNCF weighted probabilities

$$p(i) = \frac{DNCF(i)}{\sum_{n \in N} DNCF(n)} \tag{2.16}$$

where $N$ is the set of neighbors of the currently visited node, and $p$ is the probability for candidate node $i \in N$ to be picked for the next step of the random walk. To sample different paths, the random walk is repeated several thousand times with different random seeds and results accumulated. The final score of the DNCF-RW (“DNCF random walk”) is given by the number of times each node was visited during the walks. This analysis can be performed untargeted, that is for a specified number of steps after choosing a starting node, simulating the diffusion of a signal through the network. In the targeted variant, the random walk is stopped and counted only when it arrives at a predefined target node. This allows to scan two network regions for connecting pathways of conformationally coupled residues. To improve the efficiency of calculations and limit the influence of random walks getting lost in distant parts in the network, a probability to restart the random walks at each step can be set. This probability should be chosen to allow a substantial fraction for paths longer than the shortest path between source and target nodes. While conceptually similar to centralities, the DNCF-RW method focuses on the transmission of signals between individual nodes, instead of an average signal between all possible node pairs. This feature allows to investigate specific processes, like the spreading of a signal from a ligand binding pocket or the communication between two selected protein regions.

It must be noted that in the context of these models, the notions of signaling, information transfer or communication pathway does not imply a literal transfer of bits or necessarily require a "switch-like" sequence of orchestrated conformational changes. Instead, the DNCF and DNCF-RW scores respond to adjacent residues whose conformations affect each other as observed from correlation in their interaction states, without suggesting a specific mechanism. These could manifest in various ways, from switch type conformational changes to complex modulation of protein flexibility, which can be visualized like "breathing" motions of the protein (216). Consequently, our models do not presume pathways as a minimal chain of residues, but rather a diffuse and plastic cluster of interconnected residues connecting key protein regions. This mirrors views that have been expressed when discussing allosteric mechanisms, which reject the simplistic model of allostery as defined sequences of conformational changes along minimalistic pathways, suggesting instead an ensemble model of multiple contributing allosteric pathways (15, 17, 33) .

## 2.4. Implementation of the SenseNet framework for analysis of protein structure networks

We released the network analysis implementations used in this thesis as the SenseNet software (194), a plugin for the network analysis and visualization tool Cytoscape 3 (217). Both SenseNet and Cytoscape 3 are open-source and freely available from www.bioinformatics.wzw.tum.de and www.cytoscape.org. The plugin is written in Java 8 and released under the GNU Lesser General Public License (LGPL), with source code distributed in the JAR archive files used to run the program, and at https://gitlab.com/sensenet-md/sensenet.

SenseNet constructs networks from lists of atom-atom interaction timelines. These timelines contain the number or strength of different interactions, like carbon contacts, hydrogen bonds, or salt bridges, for each frame of the trajectory. These data can be readily calculated by many tools designed for analyses of MD trajectories, such as CPPTRAJ (218). Once interaction timeline data has been imported, nodes of atoms and residues are automatically inferred from atoms participating in interactions. To provide an alternative to reading CPPTRAJ output directly, we defined the custom AIF file format as a standardized data format for SenseNet input. Each line in the UTF-8 encoded AIF file corresponds to a data record with multiple comma-separated fields. The first field defines the type of record, determining how the rest of the line is parsed. A list of records and their associated data fields (for the relevant TIMELINE and DIFFERENCE_TIMELINE records) is given in Table 1.

**Table 1. Structure of the AIF file format for creating protein structure networks using SenseNet.**

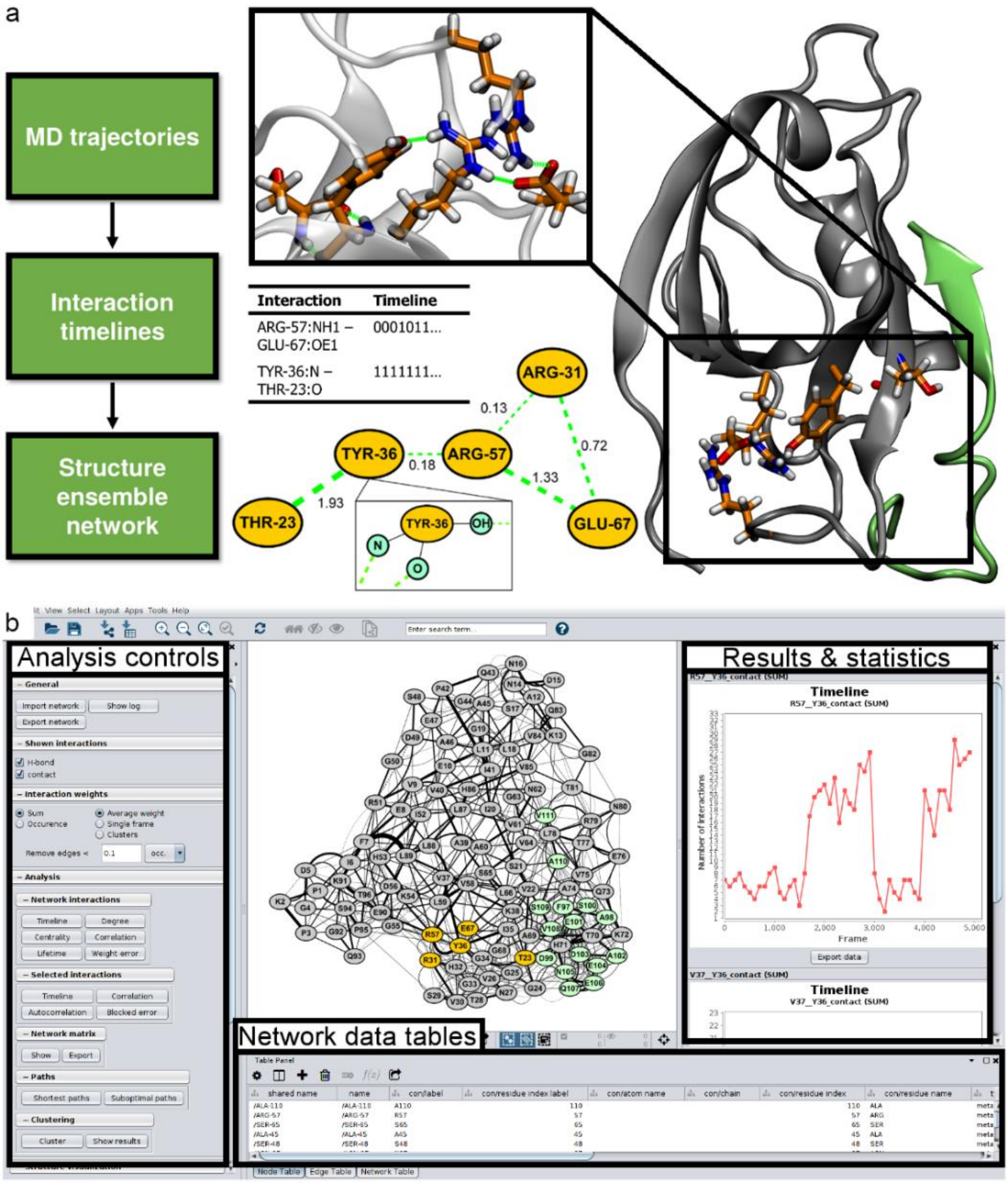
| Field index | Field name | Type |
|---|---|---|
| 0 | record type[a] | String |
| 1 | Source atom name | String |
| 2 | Target atom name | String |
| 3 | Source residue index | Integer |
| 4 | Target residue index | Integer |
| 5 | Source residue name | String |
| 6 | Source residue insert | String |
| 7 | Target residue insert | String |
| 8 | Source alternative location | String |
| 9 | Target alternative location | String |
| 10 | Source chain | String |
| 11 | Target chain | String |

| | | |
|---|---|---|
| 12 | Bridge atom names | String list[b] |
| 13 | Timeline | Float list[b] |

[a] Field structure differs between records. The table represents the TIMELINE and DIFFERENCE_TIMELINE record types.

[b] List elements delimited by whitespace

AIF files can be generated using our command line tool AIFgen, which is distributed alongside SenseNet. AIFgen calculates contacts, hydrogen bonds or salt bridges from PDB files or read analysis outputs from the MD analysis software CPPTRAJ (218). In addition, SenseNet includes AIFgen to import from any valid data source on the fly.



| Interaction | Timeline |
|---|---|
| ARG-57:NH1 – GLU-67:OE1 | 0001011... |
| TYR-36:N – THR-23:O | 1111111... |

Shown interactions

☑ H-bond

☑ contact

Interaction weights

● Sum ○ Occurence

● Average weight ○ Single frame ○ Clusters

Remove edges < 0.1 occ.

| shared name | name | con/label | con/residue index label | con/atom name | con/chain | con/residue index | con/residue name | t |
|---|---|---|---|---|---|---|---|---|
| /ALA-110 | /ALA-110 | A110 | 110 | | | 110 | ALA | meta |
| /ARG-57 | /ARG-57 | R57 | 57 | | | 57 | ARG | meta |
| /SER-65 | /SER-65 | S65 | 65 | | | 65 | SER | meta |
| /ALA-45 | /ALA-45 | A45 | 45 | | | 45 | ALA | meta |
| /SER-48 | /SER-48 | S48 | 48 | | | 48 | SER | meta |

**Figure 7. Example of parallel network and structure visualization using SenseNet.** (a) Parallel representation of networks, data tables and molecular structures. (b) Example session showing the SenseNet GUI in Cytoscape. Figure from ref. (194)

After reading a dataset of interaction timelines, SenseNet creates a new network and fills it with nodes and edges representing the atoms and atom-atom interactions from the dataset. Next, nodes are created for each unique residue position, each corresponding to a group of atoms ("group node"). When a residue node is added to the network, the corresponding atom nodes of the group are removed. At any point of an analysis session, the user can contract atom nodes into a single residue node, or expand a residue node into its individual atom nodes. This allows to dynamically change the network topology between residue and atom levels. For example, a network can be analyzed in which the protein is represented by residue nodes, but with nodes for each individual ligand atom. SenseNet automatically keeps track of interactions between group nodes, creating combined timelines as necessary (eq. 2.9) to represent interactions between atoms and residues (Figure 7).

In addition to standard analysis methods such as network centralities, SenseNet provides implementations of the ECF, NCF, DECF and DNCF methods as outlined in section 2.3. ECF scores are calculated by using SenseNet's "Correlation" function set to the "Neighbour" and "Mutual Information" modes. The "Frame Weight" option is set to "Sum" for these calculations, which causes atom-atom interactions to be combined into residue-level interactions by summing its elements. In the next step, the NCF score is obtained by using the "Degree" analysis function, which allows calculation of weighted node degree scores (110). In the analysis' settings, the "Degree weight" is set to "Edge weight sum" to calculate a weighted score according to values from an edge table column, and the "sen/correlation" column (containing the previously calculated ECF scores) is selected for the "Weight column" option. The NCF scores are automatically written into the "sen/degree" node table column. Similarly, DECF scores were calculated with the same settings as ECF, but using the "Mutual Information Difference" mode and selecting a previously imported reference network as required by eq. 2.15. The "Edge mapping mode", controlling how equivalent edges are determined between the active and the reference networks, was set to "Match Location"; in this setting, edges are considered equivalent if all residue attributes (index, insert, alternative location, as per the RCSB PDB standard) and all interaction attributes (interaction type, bridge atoms) match, with the only exception of residue names, which may differ. This option allows to calculate DECF scores of mutated residues. Once DECF scores are calculated, the DNCF calculation proceeds with the “Degree” function the same as for NCF. The ”Centrality” function of SenseNet is used to calculate of BC and CPLC centralities, which implement Dijkstra’s algorithm (219) using pseudocode of Brandes (220) and modifications of del Sol et al. (56). A full description of all algorithms implemented in SenseNet is provided in the manual (see Appendix).

# 3. RESULTS

## 3.1. Publication 1: BiPPred: Combined sequence- and structure-based prediction of peptide binding to the Hsp70 chaperone BiP

**Markus Schneider**[1], Mathias Rosam[1], Manuel Glaser, Atanas Patronov, Harpreet Shah, Katrin Christiane Back, Marina Angelika Daake, Johannes Buchner, and Iris Antes

[1] contributed equally

This study was first published online on June 10, 2016 in the peer reviewed journal *Proteins: Structure, Function and Bioinformatics.*

Citation

**Schneider, M.**, Rosam, M., Glaser, M., Patronov, A., Shah, H., Back, K. C., Daake, M. A., Buchner, J., & Antes, I. (2016). BiPPred: Combined sequence- and structure-based prediction of peptide binding to the Hsp70 chaperone BiP. *Proteins*, *84*(10), 1390–1407. https://doi.org/10.1002/prot.25084

Contributions of the author

Data curation, Investigation (computational), Method development (BiPPred), Software (BiPPred webserver), Visualization, Manuscript preparation and editing

Follow-up and related publications

Rosam, M., Krader, D., Nickels, C., Hochmair, J., Back, K. C., Agam, G., Barth, A., Zeymer, C., Hendrix, J., **Schneider, M.**, Antes, I., Reinstein, J., Lamb, D. C., & Buchner, J. (2018). Bap (Sil1) regulates the molecular chaperone BiP by coupling release of nucleotide and substrate. *Nature structural & molecular biology*, *25*(1), 90–100. https://doi.org/10.1038/s41594-017-0012-6

## Summary

Hsp70 chaperones are marked by their promiscuous binding of many protein substrates. This chaperone family of chaperones is ubiquitous in a wide range of cell types and organisms, fulfilling critical roles in protein folding and preventing aggregation. The pattern of preferentially bound protein regions, a generally hydrophobic peptide stretch between five to seven amino acids in length, differs between Hsp70 variants. Wanting to understand the evolutionary differences between these variants, we developed BiPPred to predict peptide binders for the Hsp70 chaperone BiP from their amino acid sequence, based on a multi-scale pipeline integrating experimental data, molecular docking and MMPBSA calculations. The experimental dataset was assembled from both literature and peptide array fluorescence anisotropy assays. We encountered difficulties detecting strongly binding 7mers using peptide arrays, which measured 15meric peptides, and all attempts at mapping down to smaller subsequences showed high signal variance. However, as we found several 7meric subsequences which never elicited a strong fluorescence signal, we decided to use this data for the nonbinding class of the dataset, while the binder class was composed of literature data. A structural model of BiP was created using homology modelling with a AAAAAAA peptide placed in the binding pocket as a baseline ligand. Then, each position of the peptide was systematically mutated to the other 19 canonical amino acids, using a combination of IRECS and energy minimization to predict the lowest energy conformation of the mutant. The interaction energy of this conformation was extracted and transformed into Pepscores, yielding a 7x20 matrix of interaction energy scores for all position and amino acid combinations. This structure-based position specific scoring matrix (SB-PSSM) formed the basis of our predictions. Although the matrix alone was not very predictive of peptide binding, additional statistical learning of the 7 matrix column weights using logistic regression, based on the collected experimental and literature data, yielded a model with decent capability to separate peptide binders and nonbinders. We further performed molecular dynamics simulations of the 7x20 conformations used to create the matrix, analysing the stability of modelled conformations. We found several instances for which the docking generated conformation proved unstable or otherwise inappropriate, for

example when the residue could not be placed in the binding pocket. After adjusting the SB-PSSM to correctly reflect these cases, predictive performance of the matrix improved even further to a rocAUC (receiver operator characteristic area under the curve) of 0.85 after training. To validate our model, we followed up on a selection of peptides using DynaDock, MMPBSA simulations and experiments to calculate binding affinities. We found that using this pipeline, binding peptides could not only be correctly identified, but we were also able to predict the correct orientation of the bound peptide. Our results showed that, under the constraints of limited experimental data, sequence prediction models could be generated for specific Hsp70 variants using an integrated approach of statistical learning and structural modelling.

## 3.2. Publication 2: Systematic analysis of the binding behaviour of UHRF1 towards different methyl- and carboxylcytosine modification patterns at CpG dyads

**Markus Schneider[1]**, Carina Trummer[1], Andreas Stengl, Peng Zhang, Aleksandra Szwagierczak, M. Cristina Cardoso, Heinrich Leonhardt, Christina Bauer, Iris Antes

[1] contributed equally

This study was first published online on February 21, 2020 in the peer reviewed journal *PLOS ONE.*

Citation

**Schneider, M.**, Trummer, C., Stengl, A., Zhang, P., Szwagierczak, A., Cardoso, M. C., Leonhardt, H., Bauer, C., & Antes, I. (2020). Systematic analysis of the binding behaviour of UHRF1 towards different methyl- and carboxylcytosine modification patterns at CpG dyads. *PloS one*, *15*(2), e0229144. https://doi.org/10.1371/journal.pone.0229144

Contributions of the author

Conceptualization, Data curation, Formal analysis, Investigation (computational), Method development (computational), Software, Validation, Visualization, Manuscript preparation and editing

## Summary

In this study, we investigated the binding of UHRF1, an essential protein for maintaining DNA methylation pattern, to DNA carrying different epigenetic modifications. Using electrophoretic mobility shift assays (EMSAs) and Microscale Thermophoresis (MST), we measured binding affinities of UHRF1 to CpG sites containing combinations of 5-methylcytosine (mC), 5-hydroxymethylcytosine (hmC), 5-

formylcytosine (fC) or 5-carboxylcytosine (caC). UHRF1 binds to DNA by flipping one cytosine (which carries one epigenetic modification, either mC, hmC, fC, or caC) out of its helix and enveloping it in its nucleotide binding pocket. In parallel, UHRF1 places the NKR finger into the major groove of the DNA, where it is in contact with a second potentially modified cytosine; if only the flipped base carries a modification, the configuration is termed hemi-modified, whereas if both carry one, it is called fully modified. In addition to its well-known role as a selective binder of hemi-methylated CpG sites, these experiments revealed that UHRF1 strongly bound to fully carboxylated or hybrid methylated-carboxylated CpG sites. We performed a series of Molecular Dynamics simulations to analyse the binding modes of the UHRF1-DNA complex in different modification contexts. Modification configurations were distinguished based on the modifications present on the flipped-out DNA strand (xC) or on the distal strand contacting the NKR finger (xC'); for example, caC-caC' indicated a fully carboxylated configuration. Conformational changes close to the two sites harboring DNA modifications were tracked by extracting the pattern of polar interactions, i.e., hydrogen bonds and salt bridges, from the MD trajectories. These patterns were visualized as protein structure networks, using in-house software package that would later become SenseNet. The networks showed that the presence of carboxylcytosine instead of methylcytosine lead to substantial conformational changes in two protein regions involved in DNA binding, namely the nucleotide binding pocket and the NKR finger. While the hemi-modified mC-C' configuration was overall stable in the nucleotide binding pocket, the caC-C' configuration showed strong salt bridges between the carboxylated site and and the R489 residue. This caused the caC base to rotate partially out of the binding pocket, as indicated by the weakening of its surrounding hydrogen bonding pattern. On the other hand, on the distal DNA strand, our networks indicated that mC' configurations pushed the NKR finger out of the major groove by steric repulsion, whereas caC' was able to bind to the NKR finger via salt bridges. Notably, whereas the hemi-carboxylated caC'-C' and fully methylated mC-mC' configurations were instable, the combined caC-caC' and mC-caC' modification configurations showed a strong and stable binding pattern. We investigated this apparent conformational coupling in detail, tracing its different manifestations within interaction networks, protein-DNA distances, DNA flexibility, and DNA groove dynamics. Our computational analyses provided a rationale for our experimental observations, showing how certain modification configurations could stabilize or destabilize UHRF1-DNA binding by conformational changes between coupled regions. This work contributed both as an integrated experimental-computational investigation of unexplored aspects of UHRF1, and as a pioneering application of the future SenseNet analysis framework.

## 3.3. Publication 3: SenseNet, a tool for analysis of protein structure networks obtained from molecular dynamics simulations

**Markus Schneider** and Iris Antes

This study was first published online on March 17, 2022 in the peer reviewed journal *PLOS ONE.*

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

Contributions of the author

Conceptualization, Data curation, Formal analysis, Investigation, Method development, Software, Validation, Visualization, Manuscript preparation and editing

Follow-up and related publications

Dultz, G., Shimakami, T., **Schneider, M.**, Murai, K., Yamane, D., Marion, A., Zeitler, T. M., Stross, C., Grimm, C., Richter, R. M., Bäumer, K., Yi, M., Biondi, R. M., Zeuzem, S., Tampé, R., Antes, I., Lange, C. M., & Welsch, C. (2020). Extended interaction networks with HCV protease NS3-4A substrates explain the lack of adaptive capability against protease inhibitors. *The Journal of biological chemistry*, *295*(40), 13862–13874. https://doi.org/10.1074/jbc.RA120.013898

Zheng, C., **Schneider, M.**, Marion, A., & Antes, I. (2022). The Q41R mutation in the HCV-protease enhances the reactivity towards MAVS by suppressing non-reactive pathways. *Physical chemistry chemical physics : PCCP*, *24*(4), 2126–2138. https://doi.org/10.1039/d1cp05002h

## Summary

Building on our previous work utilizing protein structure networks to trace the dynamics of conformationally coupled protein regions, we aimed for a deeper exploration of the potential within this approach. In this study, we presented the SenseNet analysis framework and software as a general-purpose tool for studying protein structure networks obtained from structure ensembles, for example a molecular dynamics trajectory. Whereas earlier iterations of this model had been focusing on establishing the fundamental network model, network visualization and basic statistics, we now turned our focus towards finding quantitative measures of computational coupling, with potential applications connected to biochemical features like protein allostery. SenseNet takes an abstract view of a protein, describing its conformations in terms of interaction states between atoms. These interaction states are encoded in a binary timeline, where each frame denotes whether two atoms interact in this particular time frame of the structure ensemble. An interaction is defined according to geometric criteria, i.e., distances or angles, to model hydrophobic interactions or hydrogen bonds. Together, these interactions make up the protein structure network, wherein nodes correspond to atoms or residues and edges represent interactions, each associated with its timeline of states. Based on the network of interaction timelines, we developed two models of conformational coupling, termed the node correlation factor (NCF) and difference node correlation factor (DNCF). To calculate the NCF of a residue, all interactions of that residue are collected and their mutual information with respect to its surrounding interactions in the network calculated; the NCF is then obtained as the sum of all these contributions. Thus, the NCF is a measure of the amount of information revealed by knowing the conformation of a residue, indicating conformational coupling in the local environment. The DNCF score is a variant of the NCF, designed to evaluate how conformational coupling changes between two simulations of different system configurations, for example with and without a bound ligand. To validate our approach, we predicted residues with known allosteric roles in the PDZ2 domain of hPTP1e, a well-known reference system for testing this class of prediction models. Compared with predictions published previously for this system, our model was more accurate than network approaches based on individual structures, and performed on par with the top

performing models using NMR data. We further compared our models to network centrality methods, the most common approach to predict conformational coupling in networks, showing that our approach was both more accurate and reliable over a wide range of network parameters. Our results complement available experimental data and consolidates the efforts of previously published predictions by establishing a consensus model. Together, these results suggest two distinct residue clusters in PDZ2 with potential allosteric roles. We present this model concurrently with the release of SenseNet as a plugin for the free network analysis software Cytoscape, demonstrating its functions, capabilities, and potential as a comprehensive and flexible tool for protein structure network analysis and visualization.

## 3.4. Publication 4: Comparison of allosteric signaling in DnaK and BiP using mutual information between simulated residue conformations

**Markus Schneider** and Iris Antes

This study was first published online on September 13, 2022 in the peer reviewed journal *Proteins: Structure, Function and Bioinformatics.*

Citation

**Schneider, M.**, & Antes, I. (2023). Comparison of allosteric signaling in DnaK and BiP using mutual information between simulated residue conformations. *Proteins*, *91*(2), 237–255. https://doi.org/10.1002/prot.26425

Contributions of the author

Conceptualization, Data curation, Formal analysis, Investigation, Method development, Software, Validation, Visualization, Manuscript preparation and editing

### Summary

The ubiquity of Hsp70 chaperones, prevalent in many pro- and eukaryotic organisms, highlights their essential role in protein folding and homeostasis. While the general structure of Hsp70 is well conserved, i.e., its domain organization including a nucleotide binding (NBD), substrate binding (SBDβ) and lid domain (SBDα), protein homologues can show substantial differences in terms of sequence identity and allosteric regulation. One of the most important mechanisms of Hsp70 function is the allosteric activation of ATP hydrolysis after binding of a protein substrate, facilitated by a network of conformationally coupled residues. The human Hsp70 homologue BiP has been considered a potential target for various medical applications, including in the fields of cancer research, neurodegenerative diseases, and viral infection. However, most knowledge of allosteric control in Hsp70 is based on the E. coli variant

DnaK, a protein with lower than 50 % sequence homolgy to BiP. Using the SenseNet analysis framework we established before, we predicted residues with potential roles in allosteric control in DnaK and BiP, specifically with focus on the coupling between protein substrate binding and ATP hydrolysis. Based on crystal structures of DnaK and BiP, we created structural models representing different phases of the Hsp70 conformational cycle. The modelled phases were Hsp70-ATP with an empty substrate binding pocket, Hsp70-ATP with a bound peptide substrate, and Hsp70-ATP-to-ADP approximating the conformation directly after hydrolysis. We performed Molecular Dynamics simulations of these configurations for DnaK and BiP, extracted protein structure networks from atom interactions, and used the DNCF method to analyze the effect of ligand binding and ATP hydrolysis on the conformational coupling between residues. Comparing the DNCF scores with a list of residues known to fulfill allosteric roles in DnaK; we observed strong agreement between our predictions and experiments, with several new allosteric candidates predicted by the model. Mapping predictions to structures revealed that these candidates were loosely associated with three clusters, collocating with the NBD-SBD domain interface and the binding sites of the DnaJ/ERdJ3 and GrpE cochaperones. In addition to structural clusters, predicted candidates also show a tendency to collocate in the protein sequence, indicating organization of allosteric roles into sequential and structural modules. Comparing results between simulated Hsp70 variants, although most predictions overlapped between DnaK and BiP, we observed unique candidates in either protein variant, which were in 40 % of cases directly related to differences in the amino acid sequence. Structural mapping of these differences indicated underlying mechanisms for previously known trends in BiP, particularly rigidification of the NBD-SBD interface, increased SBD flexibility, and differing modes of action between the grpE/BAP cochaperone class of nucleotide exchange factors. Finally, by combining the DNCF scores with a search for shortest network paths, we predicted a BiP specific pathway of conformationally coupled residues, with a potential role in regulating allosteric effects between substrate binding and nucleotide hydrolysis. Our study shows substantial differences in allosteric regulation between Hsp70 homologues, highlighting potential to engineer therapeutics tailored towards specific variants.

# 4. DISCUSSION

We conducted several studies investigating conformational coupling in proteins, focusing on the consequences of ligand binding, from short range adaptation in binding pockets to long range allosteric effects. All these processes occur and act in concert in Hsp70 chaperones, which were therefore of major interest for our work. Deepening insights into the intricate mechanisms of Hsp70 could aid discoveries in wider areas such as allosteric drug design, fine tuning protein activity, and evolution of protein function to specific organisms or organelles.

## 4.1. BiPPred: Combined sequence- and structure-based prediction of peptide binding to the Hsp70 chaperone BiP

Our first study was aimed towards an understanding of evolutionary differences in substrate recognition and development of a prediction model for high-throughput detection of BiP-binding protein regions (192). We chose a prediction model based on logistic regression, using both sequence and structure derived features to compensate for a limited number of data points. We attempted to train on experimental data from fluorescence anisotropy peptide arrays, but encountered difficulties with respect to the reproducibility of peptide array intensities and verification fluorescence anisotropy spectroscopy experiments in solution. The suspected reasons were artifacts induced by peptide mobilization on the array and the possibility of interfering formation of short helical secondary structures on the 15-meric peptides, which may reduce the binding propensity of these regions (221). The technical limitation of peptide arrays to longer peptides than the substrate recognition pattern, estimated between five to seven amino acids, has been an ongoing challenge in creating accurate prediction models (188, 191, 208, 212, 222). Despite the low reproducibility of peptide sequences with strong fluorescence, indicating strong binding to BiP, we determined that the data was still useful to provide negative data, i.e., peptide regions which never evoked a strong binding signal in the peptide array. We composed our training data set from literature data of known BiP binding sequences and added non-binding sequences from peptide array data. Our observations have implications for the use of peptide array data, particularly in the context of statistical learning, as those are commonly applied for such purposes (208). It is not clear whether the problems were within our specific setup or reflect general issues of using peptide arrays in this specific manner; however, we recommend that for applications sufficiently similar to ours, such data should be validated using an orthogonal method.

The difficulties encountered with respect to the experimental dataset motivated us to attempt to minimize the impact of statistical learning on the prediction model. It had been demonstrated that binder/non-binder sequence motifs extracted from experimental binding assays could be effectively complemented by structural models, which could make up for biases or gaps in experimental datasets (208, 223, 224). Contrasting to these previous approaches, our prediction model did not use sequence motifs as an input feature for training. Instead, all training features were based exclusively on the structure based position-specific scoring matrix (SB-PSSM) obtained from force field derived interaction scores within our structural models. Experimental data was used exclusively for determining the relative weights of each amino acid position in the matrix, i.e., one weight for each matrix column for a total of five weights. The process still required manual adjustments, as we observed that the interaction scores in the SB-PSSM were not able to effectively penalize sterically implausible protein-peptide complex models, for example when a peptide amino acid could not be placed inside its respecting binding subpocket due to its shape. In such cases, the amino acid was placed close to the protein surface, but in a superficial position outside the binding pocket that did not confer specific binding. Using a series of molecular dynamics simulations, we investigated problematic amino acid – subpocket combinations and modified selected SB-PSSM positions to reflect their apparent binding stability in the trajectories. Using this strategy, we were able to derive a SB-PSSM that showed, without training, the same prediction performance as the unmodified SB-PSSM after training. Furthermore, additional training could not further improve the prediction accuracy of this new SB-PSSM.

We successfully derived a structure-based model for high throughput prediction of BiP binding sequences in proteins. By combining statistical learning on experimental binding data with a workflow based on molecular docking and molecular dynamics, we obtained a model that could obtain accurate results despite having too few datapoints for large scale machine learning approaches, which have been successful for predicting MHC binding peptides (223). Our results highlighted that despite the overall conservation of Hsp70 structure, the detailed structural differences between the well-studied DnaK and specific variants like BiP had functional implications, like different substrate recognition preferences. General representations of Hsp70 cycles and function are often based primarily on the few most well studied model proteins like DnaK. While this is currently a necessary simplification, it should be noted that differential behavior of specific protein variants is not negligible in the context of specific application like drug design, for which BiP has been identified as a promising target (153, 159, 161-163). Consequently, we became interested in studying how differences in protein sequence and structure could affect other functional aspects of Hsp70 proteins, like the network of allosteric control that had been proposed to regulate and advance the chaperone’s conformational cycle (155).

Following our publication of BiPPred, recent studies have reported advanced strategies of integrating structural data and statistical learning based on peptide arrays (225-227). ChaperISM (227) closely followed the earlier LIMBO approach (208) using DnaK peptide arrays and FoldX interaction energies, but diverged by building a position independent scoring matrix (PISM) which showed a modest increase in accuracy compared to previous models targeting DnaK. Interestingly, they performed also slightly more accurately than BiPPred on the BiPPred dataset. This indicates potential in PISM models for building more transferable models covering multiple Hsp70 variants; however, it should be noted that large parts of the BiPPred dataset were assembled from DnaK data, which may partly explain the apparent transferability of ChaperISM. Overall, PSSMs like BiPPred or LIMBO were shown to perform strongly, but in each case were strictly limited to the proteins they were trained on. A recent DnaK-specific model called Paladin (226) begins with a similar premise to BiPPred. The authors calculate a PSSM from interaction energies based on short MD simulations, but elect to train a custom linear model for weighting these energy terms instead of general-purpose interaction scores such as the PepScore used by BiPPred. Although Paladin did not report higher overall model accuracy than previous models, it is the first DnaK model to predict forward and reverse binding orientations, a feature that had so far been only explicitly considered by BiPPred. In total, these studies have drawn similar conclusions with respect to the binding motifs of Hsp70 chaperones, and by necessity use simple linear models in combination with structural analysis to make up for the lack of high-quality binding data. Notably, the accuracy of models has not been shown to improve significantly in recent iterations, indicating that a performance ceiling may have been reached with current approaches. Lacking the abundance of binding data that has been crucial to the analogous problem of binder predicting in MHC (223), increasing prediction accuracy in Hsp70 systems may necessitate to look for alternative solutions.

## 4.2. Systematic analysis of the binding behaviour of UHRF1 towards different methyl- and carboxylcytosine modification patterns at CpG dyads

In this study, we investigated differences in the binding of UHRF1 to DNA carrying different epigenetic modifications, namely methyl- and carboxylcytosine. Our network-based analysis of interaction patterns in the nucleotide binding pocket UHRF1, combined with the propagation of conformational changes from and towards the NKR finger, would later grow into the SenseNet analysis framework. Our work expanded the picture of UHRF1, initially regarded as an exclusive reader of hemi-methylated CpG sites, towards a more multi-faceted protein able to recognize different combinations of modifications, specifically fully carboxylated and hybrid methylated-carboxylated DNA. We further showed that binding constants of UHRF1-DNA complexes as reported in literature and by us depend on the length of the respective DNA constructs used, presumably because other groups used constructs with a higher density of methylated sites (134, 147, 148, 228-230). Our observation that fully carboxylated DNA was a preferred binder to UHRF1 compared to hemi-methylated DNA was surprising due to the well-established steric clashes, found in crystal structures and molecular dynamics simulations, that prevented binding of fully methylated DNA (134, 135, 146, 147, 231). This motivated an in-depth structural study as an additional line of evidence and to find an explanation for this divergent behavior.

Our molecular dynamics simulations of UHRF1, bound to different combinations of methylated and carboxylated CpG sites, revealed a substantial effect of these modifications on interactions within the UHRF1 binding pocket. Due to the high number of interactions within the binding pocket, we chose to represent the interaction pattern as a protein structure network, with the atoms of the bound DNA base as a set of nodes which interacted with its surrounding protein residue nodes via hydrogen bonds and salt bridges. While the methyl group acted like a hydrophobic anchor to stabilize the flipped-out modified base in its canonical conformation, the carboxyl group engaged in hydrogen bonds and salt bridges in a different part of the binding pocket, causing the whole base to shift into an alternative binding conformation. Interestingly, this alternative conformation was observed to be instable if it occurred in a hemi-carboxylated context, but stable if the CpG site on the opposite DNA strand was carboxylated as well. The reason for this behavior became clear when we observed the protein structure network of the NKR finger, which was in contact with the CpG site on the opposite DNA strand. If the opposite DNA strand featured a methylated CpG site, the NKR finger was pushed away by steric repulsion. However, a carboxylated CpG site was able to establish salt bridges with the NKR finger, which consequently shifted, but was stabilized by the attractive electrostatic forces. The observation that the modification on the opposite DNA strand was able to influence the conformation of the nucleotide binding pocket indicated the presence of conformational coupling, mediated via the NKR finger. Using simulations of different combinations of DNA modifications, both in the binding pocket and at the opposite DNA strand, we were thus able to explain the molecular mechanism for the binding preference of UHRF1 to fully carboxylated CpG sites.

The analysis strategy of UHRF1 binding to DNA modifications using protein structure networks proved to be insightful and efficient in describing the conformational shifts between the simulated systems. Focusing on the interaction pattern of hydrogen bonds and salt bridges to visualize conformations allowed for intuitive reasoning about the differential conformational effects of additional methylated and carboxylated groups. By calculating the number of average interactions from a molecular dynamics trajectory, we were better able to assess the stability of interactions in the structure ensemble, compared to static structures. However, in this form our analysis approach was difficult to transfer generally to other protein systems. The differences in charge and polarity between methylated and carboxylated CpG sites allowed to limit analyses to a relatively low number of hydrogen bonds and salt bridges and neglect hydrophobic interactions. While hydrogen bonds and salt bridges are readily evaluated, due to their localization to individual polar atoms and strong electrostatic forces, hydrophobic interactions are more difficult to capture accurately. In addition, the larger number of apolar atoms and thus potential interaction partners leads to a significant increase in network complexity, making direct visual analyses of networks unfeasible. Finally, though in this work we were able to trace conformational coupling using these networks, it had to be inferred based on comparing conformations which were readily apparent from inspecting the trajectory. The same approach, however, was unlikely to succeed in cases with more subtle coupling. Therefore, in order to apply the same ideas to different systems, it was clear that the network model required an automated analysis and scoring approach to determine conformationally coupled residues.

The observed binding preference of UHRF1 for fully carboxylated CpG was unexpected, though well reproduced using different binding assays and was suggested as a possibility in earlier DNA pull-down experiments (232). Both our experiments and simulations detected this preference, although strictly in an in vitro context. The low abundance of detected carboxylcytosine in most cell types suggests that it may be unlikely for UHRF1 to encounter it, much less on both DNA strands at once (233). However, elevated levels of carboxylated DNA were observed in neuronal and tumor cells, indicating that its regulation is dynamic and dependent on the cellular context (144, 145). Furthermore, the dioxygenase TET3, facilitating demethylation of methylcytosine via a carboxylcytosine intermediate, has also been shown to recognize fully carboxylated CpG sites (234-236). The biological role of carboxylcytosine has not yet been explored comprehensively, though mounting evidence suggests that it may play a role in genome maintenance and and regulation (141-143, 237-239). In this context, an expanded picture of UHRF1 as a more universal reader of epigenetic marker combinations appears attractive, especially considering its importance for activating DNA damage repair processes (240, 241).

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

The results of our previous work indicated a relationship between conformationally coupled protein regions and our analyses based on protein structure networks. This motivated us to improve on our approach to expand its applicability, most notably by including hydrophobic interactions in our networks and evaluating different scoring functions to quantify conformational coupling between residues. We chose a simple model to approximate hydrophobic interactions by counting the contacts between nonpolar, i.e. carbon atoms within a short range, generally between 4 – 6 Å. This range aligns with common choices for protein structure networks (26, 88, 242-244) and corresponds to the effective upper limit of Van-der-Waals interactions. The attraction of this model is derived from its analogy to the definition of hydrogen bonds or salt bridges, and its accessibility for calculation and visualization. For scoring, we first evaluated the widely used node centrality approach, specifically betweenness centrality (eq. 1.1) and central path length centrality (eq. 1.2). These methods are often reported to yield good results for finding essential residues, such as catalytic sites or residues with presumed allosteric roles (56-58, 87, 103). We conducted molecular dynamics simulations of the PDZ2 domain, a common model system for evaluating algorithms designed for prediction of residues associated with allostery (133). From the MD trajectories, as well as crystallographic and NMR structures, we constructed protein structure networks and performed several analyses to predict residues with known allosteric roles. However, we as well as others found that centrality-based analyses of protein structure networks could not reliably predict allosteric residues in the PDZ2 domain (242). This was verified using a systematic evaluation of parameters used for network construction, including the cutoff for nonpolar interactions, the evaluation of all interactions versus only sidechain interactions, and different (crystal or NMR) structures used for network analysis. Most critically, centrality-based methods exhibited a large variance in model performance dependent on the network parameters, with some combinations dropping to random model performance. We did not present centrality calculations based on networks generated from molecular dynamics trajectories, as in our hands their predictive performance was generally inferior to single structure equivalents (data not shown). We attributed this behavior primarily to the addition of many edges corresponding to spurious interactions, which substantially altered the shortest paths of the network despite their low overall occurrence (see chapter 1.3). Similarly, simply removing edges below a certain cutoff of overall occurrence in the simulation, led to highly divergent results depending on the lower occurrence limit. In total, centrality methods for analysis of protein structure networks failed to produce reliable predictions of conformational coupling, regardless of whether they were based on individual structures or molecular dynamics trajectories. Moreover, it is doubtful whether the information of a structure ensemble can be reasonably included into the framework of shortest paths, due to the problem of edges introduced through spurious interactions.

Interaction timelines represent a natural extension for protein structure networks from individual structures to structure ensembles. Each frame of the timeline corresponds to the protein structure network created from the matching snapshot of the trajectory. This model translates atom coordinates into intuitively tractable conformation states (e.g., the number of contacts between residues), while preserving more information about the underlying conformational changes than creating a single network from conventional correlation coefficients. We then proceeded to build on that model by evaluating whether the Mutual Information between interaction timelines , corresponding to edges in the network, was predictive of conformational coupling. We evaluated two MI-based scores in our work: The first, termed Node Correlation Factor

(NCF), represents the aggregated mutual information the interactions of a network node convey about their closely neighboring interactions. Atom or residue nodes with high NCF scores show high interdependence of their conformation with their environment. One easily visualized example would be a “switch”-type conformational change of a residue with a bulky sidechain, modulating the conformations of surrounding residues depending on its position. However, in contrast to models based on Pearson’s coefficient, there is no assumed geometry or linearity in the model, as the generality of MI can accurately reflect arbitrarily complex relationships. Our approach distinguishes itself from alternative approaches using MI in several aspects. Protein structure networks combined with MI provides a consistent and intuitive model for capturing residue dynamics in the context of the surrounding residues; in other words, it allows to distinguish residues located in central positions of interest as reflected by the topology of the network. The resulting implicit locality of effects is a unique characteristic of the NCF score, as only correlation between neighboring residues in the network are considered. This provides a natural filtering of functionally relevant correlation, as opposed to calculating full pairwise residue matrices where most residue pairs are too distant to impact each other meaningfully (245, 246). Our data model proposes that MI based on atom interaction timelines may align more closely to biochemical function than MI models calculated from conformational entropy from atom coordinate covariance (76, 106, 247) or dihedral angles (242, 245, 248-250). However, this claim has yet to be verified in comparative studies, and it seems possible that different encodings of residue conformations could be used in a complementary fashion, e.g., by calculating MI from interactions and dihedrals. Finally, many models of conformational coupling are focused on analysis of a single biochemical system, whereas often researchers are interested in the differences between two related systems, e.g., two homologous proteins or the same protein in different ligand configurations. For these purposes, we developed the Difference Node Correlation Factor (DNCF), which further expands on the NCF by allowing to measure how information transfer between interactions changes when comparing two different simulations, for example one simulation of a protein-ligand complex and another of an unbound protein. This serves to shift the focus of analysis from correlated motion towards the modulation of correlations.

We evaluated the capacity of our NCF and DNCF scores to predict conformational coupling in the PDZ2 domain, using a dataset derived from NMR spin relaxation (130-132) to verify our analysis. Both scores achieved superior prediction accuracy compared to common centrality-based approaches while being much less sensitive to the set of network parameters used. The DNCF score, capturing the conformational shifts between ligand bound and unbound states in addition to conformational coupling, performed better over a wide range of network parameters than NCF. We further compared our results with previously published predictions by other groups using different computational methods, including one study based on Rigid Residue Scan (251), another on network centralities weighted by relative entropy of simulated Cα distances (252), and finally a network approach based on mutual information between simulated side chain rotamers (242). Among the methods published with sufficient data to be amenable to quantitative comparison, the NCF and DNCF models were among the most accurate. Interestingly, the only other method showing comparable performance was the conceptually closest, namely the model utilizing mutual information between side chain rotamers by Cilia et al. (242). However, there are notable differences which make our network model more widely applicable: First, the Cilia et al. model relies on Monte Carlo simulations in NMR structure ensembles, which are relatively rare compared to structures from other sources. In contrast, our model uses standard molecular dynamics simulations which can be based on any structure obtained from crystallography or NMR experiments, or even from a computational model like AlphaFold (253, 254). Second, while the Cilia et al. model is based on side-chain dihedrals of protein residues and thus unable to analyze alanine and glycine, our model is applicable to all protein residues and can even be used for non-protein ligands. Moreover, due to the ability of SenseNet to dynamically switch between different levels of structural resolution, the influence of non-protein compounds can be traced down to its individual atoms. Thus, the SenseNet framework provides a general approach to analyze structure ensembles which is not limited to proteins, enabling applications for all classes of chemical structures.

The NCF and DNCF scores are two of numerous methods to predict conformational coupling and allostery in proteins. Based on our results for PDZ2, we expect NCF and

DNCF to provide higher accuracy of predictions than centrality-based methods, while not requiring specialized simulation setups as necessary e.g. for Rigid Residue Scan (251, 255) or pump-probe MD (85) approaches. While the predictions accuracies we have determined for our methods are promising, they still need to be re-evaluated on a much larger dataset including a variety of different proteins and allosteric mechanisms. Initiatives like the Allosteric Database (256, 257) could be used to generate a set of proteins with allosteric residues to validate predictions. However, a comprehensive study of more than ten proteins will require substantial effort, even accounting only for the simulations required for SenseNet. Our analysis revealed that a minimum simulation time of 3 µs was required, spread over multiple replicas, in order to obtain accurate predictions. Furthermore, as prediction algorithms tend to allow a variety of tunable parameters, particularly when specialized simulations are involved, it is arguably difficult to compare prediction algorithms fairly unless one is an expert in applying all tested algorithms. This problem is compounded by the fact that published prediction algorithms often do not provide a reference implementation. Consequently, cases like PDZ2 with substantial divergence of published predictions make it difficult to reconcile models with what can be observed in experiments (133). Similar challenges have been encountered in prediction of protein structures, which have led to competitions like the biannual Critical Assessment of Methods of Protein Structure Prediction (CASP) (258). An analogous initiative, in which different groups compete with predictions on a dataset while being blind to the true solution beforehand, may become necessary for the field of allosteric predictions. This could serve to reduce the risk of over tuning parameters, while allowing an equal playing field for competing algorithms.

## 4.4. Comparison of allosteric signaling in DnaK and BiP using mutual information between simulated residue conformations

Having established a novel method to predict conformational coupling using protein structure networks, we proceeded to apply this knowledge to study allosteric regulation in Hsp70 chaperones. In our earlier investigations, we analysed how structural differences between the binding pockets of DnaK and BiP translated to different preferences in protein substrates (192). Considering the extent of functional and regulatory differences between these two homologues (168, 183, 184), we wondered to which degree the mechanisms of allosteric regulation would be conserved between DnaK and BiP.

We performed MD simulations of DnaK and BiP, extracted protein structure networks using SenseNet and determined conformationally coupled residues by calculating DNCF scores from different nucleotide and substrate binding states. Residues with known roles in allosteric regulation, as determined by mutagenesis experiments in DnaK, were found to be associated with high DNCF scores. On this basis, we predicted several additional residues with potential allosteric function in DnaK and BiP, the majority of which collocated to known allosteric sites. These sites, both previously known and newly predicted, cluster into three distinct regions in the protein structure: The shortest path between the nucleotide and protein substrate binding pockets, the cochaperone binding site of DnaJ/ErdJ3, and the binding site of the GrpE cochaperone. These regions form clusters or pathways of conformationally coupled residues, in which each residue is thought to be a link within a chain, with the proposition that perturbations may ripple through the chain via a cascade of localized movements. The initiating motion of this cascade is provided by a conformational trigger, i.e., ATP hydrolysis or binding of ligands and cochaperones. This view is compatible with the commonly invoked idea of signaling pathways in Hsp70 (155, 259), although we use the term “signal” not in a literal manner, but implying conformational coupling in a strictly statistical sense. While it is possible that the transmission of a signal through the pathway may manifest itself as a series of switch-like conformational changes, e.g., a series of residues rotating from one to another conformation, the motions captured by the DNCF score may be arbitrarily complex. For example, protein regions may rigidify or relax upon ligand binding (260) or perform “breathing” type motions (216) manifesting changes in residue fluctuations instead of easily identifiable conformation switches. As the DNCF score is purely derived from conformational coupling statistics, i.e., the changes of joint probabilities of residue conformations occurring together, it can pick up changes in coupling strength regardless of the shape of the underlying motion. In other words, if it is observed in MD simulations that residue 1 in state A correlates to state B in residue 2, the DNCF score only reflects how the strength of this relation changes between different simulations (e.g., with and without a ligand); the nature of the motion that leads to this change is undefined. This statistical view of signaling pathways thus aligns with modern descriptions of allostery by emphasizing the statistical relations between conformations in the ensemble (15, 17, 32, 33).

Most residues with high DNCF scores were found within close range to the nucleotide and substrate binding sites. This is expected, as the simulations compared for the mutual information differences induced structural changes in exactly these regions, i.e., ATP exchanged to ADP and bound vs. unbound NRLLLTG peptide in the substrate binding pocket. It is important to note that these structural changes were introduced artificially, as no experimental structures of these states were available. Due to the limited time range of MD simulations, structural changes are unlikely to have propagated through the entire structure; thus, the accuracy of predictions is likely to decline with increasing distance from the sites where structural changes were induced. In other words, the DNCF scores for DnaK and BiP are likely to be lower for residues if they are further away from the nucleotide and substrate binding sites. Potential allosteric candidates could thus be missed if residue scores are considered in isolation. However, by adding more context, additional candidates can be inferred by their position relative to other high-scoring residues. In our work, we achieved this by performing random walks in the networks, weighted by residue DNCF scores, to find pathways between the nucleotide and protein substrate binding site (DNCF with random walk, or “DNCF-RW”). Intuitively, this approach models how a conformational change originating from a source site, if it propagated between nodes with a probability proportional to the residue’s DNCF score, would travel through the network until it reached the target site. This way, residues which are located on the path between allosteric sites and have high DNCF scores relative to their surroundings are highlighted, even if their absolute DNCF scores are low due to insufficient simulation time. Combining DNCF with random walks combines two orthogonal ideas, namely the localized coupling between residue conformations and the detection of shortest network paths, which is the essence of prediction methods based on network centralities (56-58, 87). Using the DNCF-RW method, we were able to detect residues which were not within the top bracket of individual DNCF scores, but which could be inferred to play an allosteric role from their topological context, i.e., bottlenecks connecting clusters of high DNCF residues.

By comparing the differences between residues with top DNCF and DNCF-RW scores, we were able to predict residues with specific roles in DnaK and BiP. While the majority of predicted allosteric residues in both proteins were conserved, we found significant differences, about 40 % of which were directly related to mutations in the protein sequences. We found that these differences aligned with phenomena observed in previous experiments, like the increase of SBD dynamics in BiP (261) or the fundamental differences in GrpE/BAP cochaperone mechanisms (262). Most of our current knowledge of Hsp70 mechanism is based on DnaK as the most accessible model system (155). Our data highlights how individual protein regions were shaped by evolution to fulfill protein functions in the different environments encountered by DnaK and BiP, respectively. Particularly in the development and application of allosteric drugs or biomarkers, small differences between evolutionary variants have the potential to prove highly significant.

# 5. OUTLOOK

In this work, we developed a framework to analyze conformational coupling using timelines of atom interactions in structure ensembles, based on protein structure networks. Initially designed for investigation of interactions in ligand binding pockets, which had served as the foundation of our BiPPred algorithm, we expanded the methodology in a study involving the binding of UHRF1 to DNA modifications, and finally established the final model including two mutual information-based analysis algorithms with SenseNet. We then applied these prediction algorithms to determine additional candidates with allosteric roles in Hsp70 proteins, focusing specifically on residues with specific roles in the Hsp70 variants DnaK and BiP. Our results demonstrate the potential of our SenseNet model in two areas: First, it can be used for characterization and visualization of interactions in structural regions of interest, such as binding pockets and interfaces associated with ligand binding, in the context of structure ensembles. Second, we showed that the NCF and DNCF scores, based on mutual information between interaction timelines obtained from structure ensembles, can accurately predict residues playing regulatory roles in protein allostery. We have provided case studies for successful applications covering both aspects; still, much work remains to be done to improve on weak points and ensure that the method is generalizable to a wide range of systems.

Novel computational methods for prediction of allosteric residues are often presented and evaluated on just a handful of systems, at least in their initial publication; our own work is no exception (248, 252, 255, 263-269). Going forward, extensive efforts are required to compare different prediction approaches while minimizing bias towards specific datasets. First and foremost, there is a great need for a commonly accepted gold standard dataset for validating predictions of allosteric residues, spanning a comprehensive range of protein systems. Ideally, a similar initiative to the CASP competition (258), adapted for allosteric prediction, could help to produce high quality and unbiased evaluations of methods; however, as long as datasets high quality annotations of allosteric residues cannot be produced with similar efficacy as protein structures, comparisons against gold standard datasets are the next viable option. As of 2023, the Allosteric Database (256, 257) features annotations of short of 2000 allosteric proteins, which could provide the basis for generating a curated validation standard. Additional though smaller datasets have been brought forward by Greener et al. (41) and Panjkovich et al. (43); nevertheless, assembly and curation of these datasets remains laborious and limited to small scale. Publication of novel prediction methods should be accompanied with well documented and easy to use reference software implementations, as we have attempted by publishing the SenseNet software. This not only serves to make the prediction accessible to the wider research community, but also allows to set up studies comparing a range of different methods to learn about the specific strengths and weaknesses of each. In addition, there is potential for consensus predictions combining different methods to increase the overall accuracy. Even when a gold standard evaluation set for allosteric residue prediction is chosen and reference implementations are provided, the computational cost even for a modest set of about ten proteins could be substantial. Still, when faced with situations as the glaring divergence of computational predictions for allostery in PDZ domains (133), the importance of better method validation efforts cannot be overstated, justifying their significant cost.

With the SenseNet framework, we have provided an extensible foundation for easy to use and effective application of protein structure networks based on structure ensembles. Implemented as a plugin for the free network analysis software Cytoscape (217), it is accessible to a broad audience of researchers and can be easily combined with complementary network analyses available from Cytoscape's ecosystem of user-contributed plugins. The advantages of this approach become apparent when considering the most common distribution channels for network analysis tools, which are dominated by stand-alone webservers or libraries for Python and R programming languages. These implementations, while appropriate for batch analysis, commonly offer only basic visualization features and are difficult to use for exploratory analyses. In contrast, the modular plugin architecture of the Cytoscape platform allows independent but interoperable tools within a single GUI environment, providing extensive support for interactive network visualization. SenseNet is published under the and published open source under the lesser GNU General Public License (LGPL) and is designed to be easily extensible even without source code modifications; for example, the flexibility of the AIF input format can easily accommodate different types of interactions or even timelines of continuous values, like interaction energies or atom distances.

The full potential of the data model underlying SenseNet, namely the distinction of conformations in a structure ensemble in terms of interaction timelines, is not yet fully explored. In our work, interaction timelines were generated by simple geometric criteria, i.e., carbon-carbon contacts to model hydrophobic interactions and combined distance/angle constraints for hydrogen bonds. Additional interactions might be considered, like π-π interactions, cation–π interactions, or dihedral configurations as used in other methods (242, 250). Further improvements may involve the discretization of interaction states in residue – residue interactions: Whereas currently the count of interactions between atoms of both residues is used to provide an aggregate timeline, this approach can suffer when some interaction states (i.e., interaction counts) are underrepresented in the structure ensemble. This makes obtaining accurate statistics for mutual information more difficult, and several proposed strategies for obtaining unbiased estimators could be evaluated (112, 113). We expect that any improvements in interactions statistics would straightforwardly improve the accuracy and precision of allosteric predictions.

Existing applications of SenseNet have so far focused on analysis and visualization of simple interaction statistics and a mutual-information based prediction of residues associated with allosteric control. Alternative analyses could make use of the timeline data model: For example, convergence of network statistics between simulation replicas could be useful to determine whether a MD simulation is undersampled. Conventional MD simulations have a well-documented tendency to get stuck in locally optimal potential energy minima, necessitating long simulations and the usage of independent replicas to compensate. It is notoriously difficult to determine whether a MD simulation has run long enough to sample the relevant configuration space sufficiently (50, 51). Generally, this requires an informed decision on both the number of replicas (independent runs of the same simulation diverging due to stochastic elements in the dynamics or numerical inaccuracies) and the simulation time of each individual replica. How to determine these parameters to ensure reproducibility and accuracy for an arbitrary simulation setup, i.e., absolute convergence, is a matter of active discussion (50, 51). However, it is commonly agreed that as a minimum, the results of MD analyses should be consistent both between replicas and between time slices of an individual simulation (relative convergence) (50). SenseNet already implements several functions which can help to determine relative convergence of interaction timelines, such as calculation of timeline autocorrelation or blocked averages (50, 270). Both methods implement a different approach to estimate the true number of statistical samples within a MD trajectory. We have also performed initial work to compare the distribution of interaction timeline elements, tracking distribution differences between timeline blocks using e.g. Jensen-Shannon and Kullback-Leibler divergences (271, 272). Interestingly, initial unpublished results indicate that the divergence between replicas decays exponentially with increasing simulation length per replica, suggesting possibilities to optimize the tradeoff of simulating more replicas vs. individually longer replica simulations. As problems from undersampling belong to the most critical concerns for the accuracy of any simulation, the potential of SenseNet to provide sanity checking and mitigation strategies should be explored. Due to the inherent inefficiencies of sampling by conventional Molecular Dynamics, there are ongoing efforts to improve sampling using alternative dynamics and force fields. Alternative schemes for generating structure ensembles, commonly subsumed under term “Advanced Sampling”, implement acceleration strategies for a simulation to escape locally optimal potential energy wells (83, 273, 274). As a complementary strategy, coarse-grained force fields have been developed to reduce the number of possible conformations, effectively blurring out atomistic details in order to simplify exploration of the relevant configuration space during simulation (79, 80). Both approaches are orthogonal, can be combined with each other, and even intersect or be supplemented by modern Machine Learning approaches like Deep Neural Networks (275). As the SenseNet data model was designed to be as agnostic as possible with respect to its input data, we expect that our analysis methods can be applied straightforwardly to all kinds of advanced sampling methods for detection of highly coupled system modes and conformational barriers. Once identified, protein regions stuck in long-lived conformations could be specifically targeted during advanced sampling, forcing the system out of stale conformations by adding forces driving exploration based on Collective Variables (CV). Combined with tracking of relative simulation convergence, an iterative scheme of CV selection and exploitation could prove a useful tool to increase sampling efficiencies in molecular dynamics.

The relationship between molecular simulation and experiment has both synergistic and competitive aspects. While there are numerous examples where simulation and experiment have been used to complement each other in a true interdisciplinary fashion, computational methods are under pressure to ensure that their predictions align with what is observed in biochemical experiments. Due to the significant cost of simulations in researcher expertise, computational resources and time, computations requiring complex workflows can end up being less efficient than upscaling experiments in the lab. This pressure is felt even in areas where computational methods were originally thought to have an advantage, like drug design and binding optimization. This work, among others, is tailored towards a problem for which computational methods are poised to play a unique role, namely the study of protein allostery and its applications in drug design. As a priori unknown binding sites are difficult to probe experimentally, detection of potential binding pockets for allosteric drugs would greatly benefit from accurate computational prediction of conformational coupling. Even more than for ligand binding in orthosteric pockets, detailed mechanistic understanding of the underlying molecular mechanisms will be indispensable to understand the subtleties of allosteric effects and evolutional divergence of function between homologous variants of protein drug targets. Numerous computational approaches have been presented which could contribute to this effort, including our own; what is needed going forward, will be to evaluate these methods to understand their strengths and flaws, to consolidate and refine them, and to investigate combined theoretical and experimental approaches to deepen the application interface between in silico and in vitro domains.

# 6. ACKNOWLEDGEMENTS

This thesis marks the end of a very long journey. I could not have done this without the unwavering support of my loving family, to whom this work is dedicated.

A special acknowledgement to my supervisor Prof. Dr. Iris Antes, who inspired my interest in this topic and gave me the opportunity to pursue it,

Prof. Dr. Dmitrij Frishman and Prof. Dr. Martin Zacharias, who helped me to push through the last difficult stretch after Iris' untimely passing,

My fellow colleagues Antoine, Ilke, Manuel, Okke, Max, Martin, Helmut, Lukas, and Chen, of and with whom I learned a lot,

my colleagues at MSAID, who encouraged me to finally get this out of the door

and last but not least my friends Fieri, Tobi, Michi and Alina, who with great patience listened to my ramblings and tolerated my flakiness.

To all of you, my heartfelt appreciation.

# LIST OF FIGURES

# REFERENCES

1. Korendovych IV. Rational and Semirational Protein Design. Methods Mol Biol. 2018;1685:15-23.

2. Schmidt TGM, Eichinger A, Schneider M, Bonet L, Carl U, Karthaus D, et al. The Role of Changing Loop Conformations in Streptavidin Versions Engineered for High-affinity Binding of the Strep-tag II Peptide. J Mol Biol. 2021;433(9):166893.

3. Rothlisberger D, Khersonsky O, Wollacott AM, Jiang L, DeChancie J, Betker J, et al. Kemp elimination catalysts by computational enzyme design. Nature. 2008;453(7192):190-5.

4. Korendovych IV, Kulp DW, Wu Y, Cheng H, Roder H, DeGrado WF. Design of a switchable eliminase. Proc Natl Acad Sci U S A. 2011;108(17):6823-7.

5. Morin A, Meiler J, Mizoue LS. Computational design of protein-ligand interfaces: potential in therapeutic development. Trends Biotechnol. 2011;29(4):159-66.

6. Sliwoski G, Kothiwale S, Meiler J, Lowe EW, Jr. Computational methods in drug discovery. Pharmacol Rev. 2014;66(1):334-95.

7. De Vivo M, Masetti M, Bottegoni G, Cavalli A. Role of Molecular Dynamics and Related Methods in Drug Discovery. Journal of medicinal chemistry. 2016:4035-61.

8. Pinzi L, Rastelli G. Molecular Docking: Shifting Paradigms in Drug Discovery. Int J Mol Sci. 2019;20(18).

9. Nussinov R, Tsai CJ. Allostery in disease and in drug discovery. Cell. 2013;153(2):293-305.

10. Lu S, Li S, Zhang J. Harnessing allostery: a novel approach to drug discovery. Med Res Rev. 2014;34(6):1242-85.

11. 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.

12. Yang JS, Seo SW, Jang S, Jung GY, Kim S. Rational engineering of enzyme allosteric regulation through sequence evolution analysis. PLoS Comput Biol. 2012;8(7):e1002612.

13. Sheik Amamuddy O, Veldman W, Manyumwa C, Khairallah A, Agajanian S, Oluyemi O, et al. Integrated Computational Approaches and Tools forAllosteric Drug Discovery. Int J Mol Sci. 2020;21(3).

14. Wang Q, Zheng M, Huang Z, Liu X, Zhou H, Chen Y, et al. Toward understanding the molecular basis for chemical allosteric modulator design. Journal of Molecular Graphics and Modelling. 2012;38:324-33.

15. Tsai C-J, Nussinov R. A Unified View of “How Allostery Works”. PLoS Computational Biology. 2014;10:e1003394.

16. Liu J, Nussinov R. Allostery: An Overview of Its History, Concepts, Methods, and Applications. PLoS Comput Biol. 2016;12(6):e1004966.

17. Motlagh HN, Wrabl JO, Li J, Hilser VJ. The ensemble nature of allostery. Nature. 2014;508(7496):331-9.

18. Gunasekaran K, Ma B, Nussinov R. Is allostery an intrinsic property of all dynamic proteins? Proteins: Structure, Function, and Bioinformatics. 2004;57(3):433-43.

19. Monod J, Wyman J, Changeux J-P. On the nature of allosteric transitions: A plausible model. Journal of Molecular Biology. 1965;12(1):88-118.

20. Koshland DE, Jr., Némethy G, Filmer D. Comparison of experimental binding data and theoretical models in proteins containing subunits. Biochemistry. 1966;5(1):365-85.

21. Perutz MF, Rossmann MG, Cullis AF, Muirhead H, Will G, North ACT. Structure of Hæmoglobin: A Three-Dimensional Fourier Synthesis at 5.5-Å. Resolution, Obtained by X-Ray Analysis. Nature. 1960;185(4711):416-22.

22. Perutz MF. Stereochemistry of Cooperative Effects in Haemoglobin: Haem–Haem Interaction and the Problem of Allostery. Nature. 1970;228(5273):726-34.

23. Perutz MF, Wilkinson AJ, Paoli M, Dodson GG. THE STEREOCHEMICAL MECHANISM OF THE COOPERATIVE EFFECTS IN HEMOGLOBIN REVISITED. Annual Review of Biophysics and Biomolecular Structure. 1998;27(1):1-34.

24. Changeux JP. 50 years of allosteric interactions: the twists and turns of the models. Nat Rev Mol Cell Biol. 2013;14(12):819-29.

25. Popovych N, Sun S, Ebright RH, Kalodimos CG. Dynamically driven protein allostery. Nat Struct Mol Biol. 2006;13(9):831-8.

26. Daily MD, Gray JJ. Local motions in a benchmark of allosteric proteins. Proteins. 2007;67(2):385-99.

27. Cooper A, Dryden DTF. Allostery without conformational change. European Biophysics Journal. 1984;11(2):103-9.

28. Tsai CJ, del Sol A, Nussinov R. Allostery: absence of a change in shape does not imply that allostery is not at play. J Mol Biol. 2008;378(1):1-11.

29. Tzeng SR, Kalodimos CG. Allosteric inhibition through suppression of transient conformational states. Nat Chem Biol. 2013;9(7):462-5.

30. Sekhar A, Kay LE. NMR paves the way for atomic level descriptions of sparsely populated, transiently formed biomolecular conformers. Proc Natl Acad Sci U S A. 2013;110(32):12867-74.

31. Manley G, Rivalta I, Loria JP. Solution NMR and computational methods for understanding protein allostery. J Phys Chem B. 2013;117(11):3063-73.

32. Kumar S, Ma B, Tsai C-J, Wolfson H, Nussinov R. Folding funnels and conformational transitions via hinge-bending motions. Cell Biochemistry and Biophysics. 1999;31(2):141-64.

33. Greener JG, Sternberg MJ. Structure-based prediction of protein allostery. Curr Opin Struct Biol. 2018;50:1-8.

34. Doman TN, McGovern SL, Witherbee BJ, Kasten TP, Kurumbail R, Stallings WC, et al. Molecular docking and high-throughput screening for novel inhibitors of protein tyrosine phosphatase-1B. J Med Chem. 2002;45(11):2213-21.

35. Salmaso V, Moro S. Bridging Molecular Docking to Molecular Dynamics in Exploring Ligand-Protein Recognition Process: An Overview. Front Pharmacol. 2018;9:923.

36. Tuccinardi T. What is the current value of MM/PBSA and MM/GBSA methods in drug discovery? Expert Opin Drug Discov. 2021;16(11):1233-7.

37. Genheden S, Ryde U. The MM/PBSA and MM/GBSA methods to estimate ligand-binding affinities. Expert Opinion on Drug Discovery. 2015;10(5):449-61.

38. Bowman GR, Bolin ER, Hart KM, Maguire BC, Marqusee S. Discovery of multiple hidden allosteric sites by combining Markov state models and experiments. Proc Natl Acad Sci U S A. 2015;112(9):2734-9.

39. Oleinikovas V, Saladino G, Cossins BP, Gervasio FL. Understanding Cryptic Pocket Formation in Protein Targets by Enhanced Sampling Simulations. J Am Chem Soc. 2016;138(43):14257-63.

40. Huang W, Lu S, Huang Z, Liu X, Mou L, Luo Y, et al. Allosite: a method for predicting allosteric sites. Bioinformatics. 2013;29(18):2357-9.

41. Greener JG, Sternberg MJ. AlloPred: prediction of allosteric pockets on proteins using normal mode perturbation analysis. BMC Bioinformatics. 2015;16:335.

42. Panjkovich A, Daura X. PARS: a web server for the prediction of Protein Allosteric and Regulatory Sites. Bioinformatics. 2014;30(9):1314-5.

43. Panjkovich A, Daura X. Exploiting protein flexibility to predict the location of allosteric sites. BMC Bioinformatics. 2012;13(1):273.

44. Guarnera E, Tan ZW, Zheng Z, Berezovsky IN. AlloSigMA: allosteric signaling and mutation analysis server. Bioinformatics. 2017;33(24):3996-8.

45. Xu Y, Wang S, Hu Q, Gao S, Ma X, Zhang W, et al. CavityPlus: a web server for protein cavity detection with pharmacophore modelling, allosteric site identification and covalent ligand binding ability prediction. Nucleic Acids Res. 2018;46(W1):W374-w9.

46. Li H, Chang YY, Lee JY, Bahar I, Yang LW. DynOmics: dynamics of structural proteome and beyond. Nucleic Acids Res. 2017;45(W1):W374-w80.

47. Goncearenco A, Mitternacht S, Yong T, Eisenhaber B, Eisenhaber F, Berezovsky IN. SPACER: Server for predicting allosteric communication and effects of regulation. Nucleic Acids Res. 2013;41(Web Server issue):W266-72.

48. Clarke D, Sethi A, Li S, Kumar S, Chang RWF, Chen J, et al. Identifying Allosteric Hotspots with Dynamics: Application to Inter- and Intra-species Conservation. Structure. 2016;24(5):826-37.

49. Guo J, Zhou HX. Protein Allostery and Conformational Dynamics. Chem Rev. 2016;116(11):6503-15.

50. Grossfield A, Zuckerman DM. Quantifying uncertainty and sampling quality in biomolecular simulations. Annu Rep Comput Chem. 2009;5:23-48.

51. Grossfield A, Patrone PN, Roe DR, Schultz AJ, Siderius DW, Zuckerman DM. Best Practices for Quantification of Uncertainty and Sampling Quality in Molecular Simulations [Article v1.0]. Living J Comput Mol Sci. 2018;1(1).

52. Feher VA, Durrant JD, Van Wart AT, Amaro RE. Computational approaches to mapping allosteric pathways. Curr Opin Struct Biol. 2014;25:98-103.

53. Liang Z, Verkhivker GM, Hu G. Integration of network models and evolutionary analysis into high-throughput modeling of protein dynamics and allosteric regulation: theory, tools and applications. Brief Bioinform. 2020;21(3):815-35.

54. Bowerman S, Wereszczynski J. Detecting Allosteric Networks Using Molecular Dynamics Simulation. Methods Enzymol. 2016;578:429-47.

55. Fuglestad B, Gasper PM, McCammon JA, Markwick PR, Komives EA. Correlated motions and residual frustration in thrombin. J Phys Chem B. 2013;117(42):12857-63.

56. del Sol A, Fujihashi H, Amoros D, Nussinov R. Residues crucial for maintaining short paths in network communication mediate signaling in proteins. Molecular systems biology. 2006;2:2006.0019.

57. 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.

58. Di Paola L, Giuliani A. Protein contact network topology: a natural language for allostery. Curr Opin Struc Biol. 2015;31:43-8.

59. van den Bedem H, Bhabha G, Yang K, Wright PE, Fraser JS. Automated identification of functional dynamic contact networks from X-ray crystallography. Nat Methods. 2013;10(9):896-902.

60. Ferreiro Diego U, Hegler Joseph A, Komives Elizabeth A, Wolynes Peter G. Localizing frustration in native proteins and protein assemblies. Proceedings of the National Academy of Sciences. 2007;104(50):19819-24.

61. Daily MD, Upadhyaya TJ, Gray JJ. Contact rearrangements form coupled networks from local motions in allosteric proteins. Proteins. 2008;71(1):455-66.

62. Mao W, Kaya C, Dutta A, Horovitz A, Bahar I. Comparative study of the effectiveness and limitations of current methods for detecting sequence coevolution. Bioinformatics. 2015;31(12):1929-37.

63. Lockless SW, Ranganathan R. Evolutionarily conserved pathways of energetic connectivity in protein families. Science. 1999;286(5438):295-9.

64. Suel GM, Lockless SW, Wall MA, Ranganathan R. Evolutionarily conserved networks of residues mediate allosteric communication in proteins. Nat Struct Biol. 2003;10(1):59-69.

65. Dunn SD, Wahl LM, Gloor GB. Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction. Bioinformatics. 2008;24(3):333-40.

66. Kass I, Horovitz A. Mapping pathways of allosteric communication in GroEL by analysis of correlated mutations. Proteins. 2002;48(4):611-7.

67. Morcos F, Pagnani A, Lunt B, Bertolino A, Marks DS, Sander C, et al. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. Proceedings of the National Academy of Sciences. 2011;108(49):E1293-E301.

68. Jones DT, Buchan DW, Cozzetto D, Pontil M. PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. Bioinformatics. 2012;28(2):184-90.

69. Socolich M, Lockless SW, Russ WP, Lee H, Gardner KH, Ranganathan R. Evolutionary information for specifying a protein fold. Nature. 2005;437(7058):512-8.

70. Liu Y, Gierasch LM, Bahar I. Role of Hsp70 ATPase Domain Intrinsic Dynamics and Sequence Evolution in Enabling its Functional Interactions with NEFs. PLOS Computational Biology. 2010;6(9):e1000931.

71. Wang N, Lodge JM, Fierke CA, Mapp AK. Dissecting allosteric effects of activator-coactivator complexes using a covalent small molecule ligand. Proc Natl Acad Sci U S A. 2014;111(33):12061-6.

72. Lakhani B, Thayer KM, Hingorani MM, Beveridge DL. Evolutionary Covariance Combined with Molecular Dynamics Predicts a Framework for Allostery in the MutS DNA Mismatch Repair Protein. The Journal of Physical Chemistry B. 2017;121(9):2049-61.

73. Bahar I, Lezon TR, Yang LW, Eyal E. Global dynamics of proteins: bridging between structure and function. Annu Rev Biophys. 2010;39:23-42.

74. Mitternacht S, Berezovsky IN. Binding leverage as a molecular basis for allosteric regulation. PLoS Comput Biol. 2011;7(9):e1002148.

75. Lange OF, Grubmuller H. Generalized correlation for biomolecular dynamics. Proteins. 2006;62(4):1053-61.

76. Lange OF, Grubmuller H. Full correlation analysis of conformational protein dynamics. Proteins. 2008;70(4):1294-312.

77. Sethi A, Eargle J, Black AA, Luthey-Schulten Z. Dynamical networks in tRNA:protein complexes. Proc Natl Acad Sci U S A. 2009;106(16):6620-5.

78. Vanwart AT, Eargle J, Luthey-Schulten Z, Amaro RE. Exploring residue component contributions to dynamical network models of allostery. J Chem Theory Comput. 2012;8(8):2949-61.

79. Kmiecik S, Gront D, Kolinski M, Wieteska L, Dawid AE, Kolinski A. Coarse-Grained Protein Models and Their Applications. Chem Rev. 2016;116(14):7898-936.

80. Saunders MG, Voth GA. Coarse-graining methods for computational biology. Annu Rev Biophys. 2013;42:73-93.

81. Riniker S, Allison JR, van Gunsteren WF. On developing coarse-grained models for biomolecular simulation: a review. Phys Chem Chem Phys. 2012;14(36):12423-30.

82. Zuckerman DM. Equilibrium Sampling in Biomolecular Simulation. Annual review of biophysics. 2011;40:41-62.

83. Yang YI, Shao Q, Zhang J, Yang L, Gao YQ. Enhanced sampling in molecular dynamics. J Chem Phys. 2019;151(7):070902.

84. Maximova T, Moffatt R, Ma B, Nussinov R, Shehu A. Principles and Overview of Sampling Methods for Modeling Macromolecular Structure and Dynamics. PLoS Computational Biology. 2016;12(4):e1004619.

85. Sharp K, Skinner JJ. Pump-probe molecular dynamics as a tool for studying protein motion and long range coupling. Proteins. 2006;65(2):347-61.

86. Atilgan C, Atilgan AR. Perturbation-response scanning reveals ligand entry-exit mechanisms of ferric binding protein. PLoS Comput Biol. 2009;5(10):e1000544.

87. Greene LH. Protein structure networks. Briefings in Functional Genomics. 2012;11:469-78.

88. 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.

89. Böde C, Kovács Ia, Szalay MS, Palotai R, Korcsmáros T, Csermely P, et al. Network analysis of protein dynamics. FEBS Letters. 2007;581:2776-82.

90. Vendruscolo M, Dokholyan NV, Paci E, Karplus M. Small-world view of the amino acids that play a key role in protein folding. Phys Rev E Stat Nonlin Soft Matter Phys. 2002;65(6 Pt 1):061910.

91. Greene LH, Higman Va. Uncovering network systems within protein structures. Journal of Molecular Biology. 2003;334:781-91.

92. Brinda KV, Vishveshwara S. A Network Representation of Protein Structures: Implications for Protein Stability. Biophysical Journal. 2005;89:4159-70.

93. Estrada E. Universality in protein residue networks. Biophys J. 2010;98(5):890-900.

94. Atilgan AR, Akan P, Baysal C. Small-world communication of residues and significance for protein dynamics. Biophysical journal. 2004;86:85-91.

95. Pande VS, Beauchamp K, Bowman GR. Everything you wanted to know about Markov State Models but were afraid to ask. Methods (San Diego, Calif). 2010;52(1):99-105.

96. Sengupta U, Strodel B. Markov models for the elucidation of allosteric regulation. Philosophical Transactions of the Royal Society B: Biological Sciences. 2018;373(1749):20170178.

97. Pande VS. Understanding Protein Folding Using Markov State Models. In: Bowman GR, Pande VS, Noé F, editors. An Introduction to Markov State Models and Their Application to Long Timescale Molecular Simulation. Dordrecht: Springer Netherlands; 2014. p. 101-6.

98. Bowman GR, Ensign DL, Pande VS. Enhanced modeling via network theory: Adaptive sampling of Markov state models. Journal of chemical theory and computation. 2010;6(3):787-94.

99. Konovalov KA, Unarta IC, Cao S, Goonetilleke EC, Huang X. Markov State Models to Study the Functional Dynamics of Proteins in the Wake of Machine Learning. JACS Au. 2021;1(9):1330-41.

100. McGibbon RT, Schwantes CR, Pande VS. Statistical Model Selection for Markov Models of Biomolecular Dynamics. The Journal of Physical Chemistry B. 2014;118(24):6475-81.

101. Suarez E, Wiewiora RP, Wehmeyer C, Noe F, Chodera JD, Zuckerman DM. What Markov State Models Can and Cannot Do: Correlation versus Path-Based Observables in Protein-Folding Models. J Chem Theory Comput. 2021;17(5):3119-33.

102. Kumar S, Nussinov R. Close-Range Electrostatic Interactions in Proteins. ChemBioChem. 2002;3(7):604-17.

103. Del Sol A, Araúzo-Bravo MJ, Amoros D, Nussinov R. Modular architecture of protein structures and allosteric communications: potential implications for signaling proteins and regulatory linkages. Genome biology. 2007;8:R92.

104. Amitai G, Shemesh A, Sitbon E, Shklar M, Netanely D, Venger I, et al. Network Analysis of Protein Structures Identifies Functional Residues. Journal of Molecular Biology. 2004;344:1135-46.

105. Gasper PM, Fuglestad B, Komives EA, Markwick PRL, McCammon JA. Allosteric networks in thrombin distinguish procoagulant vs. anticoagulant activities. Proceedings of the National Academy of Sciences. 2012;109(52):21216.

106. Rivalta I, Sultan MM, Lee NS, Manley GA, Loria JP, Batista VS. Allosteric pathways in imidazole glycerol phosphate synthase. Proc Natl Acad Sci U S A. 2012;109(22):E1428-36.

107. Doshi U, Holliday MJ, Eisenmesser EZ, Hamelberg D. Dynamical network of residue-residue contacts reveals coupled allosteric effects in recognition, catalysis, and mutation. Proc Natl Acad Sci U S A. 2016;113(17):4735-40.

108. Tse A, Verkhivker GM. Molecular Dynamics Simulations and Structural Network Analysis of c-Abl and c-Src Kinase Core Proteins: Capturing Allosteric Mechanisms and Communication Pathways from Residue Centrality. J Chem Inf Model. 2015;55(8):1645-62.

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

110. Opsahl T, Agneessens F, Skvoretz J. Node centrality in weighted networks: Generalizing degree and shortest paths. Social Networks. 2010;32(3):245-51.

111. Ichiye T, Karplus M. Collective motions in proteins: a covariance analysis of atomic fluctuations in molecular dynamics and normal mode simulations. Proteins. 1991;11(3):205-17.

112. Kraskov A, Stögbauer H, Grassberger P. Estimating mutual information. Phys Rev E. 2004;69(6):066138.

113. Holmes CM, Nemenman I. Estimation of mutual information for real-valued data with error bars and controlled bias. Phys Rev E. 2019;100(2):022404.

114. 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. 2017;33(17):2768-71.

115. Seeber M, Felline A, Raimondi F, Muff S, Friedman R, Rao F, et al. Wordom: A user-friendly program for the analysis of molecular structures, trajectories, and free energy surfaces. Journal of Computational Chemistry. 2011;32(6):1183-94.

116. Chakrabarty B, Parekh N. NAPS: Network Analysis of Protein Structures. Nucleic Acids Research. 2016;44(W1):W375-W82.

117. Doncheva NT, Klein K, Domingues FS, Albrecht M. Analyzing and visualizing residue networks of protein structures. Trends in Biochemical Sciences. 2011;36(4):179-82.

118. Contreras-Riquelme S, Garate J-A, Perez-Acle T, Martin AJM. RIP-MD: a tool to study residue interaction networks in protein molecular dynamics. PeerJ. 2018;6:e5998.

119. Ribeiro AA, Ortiz V. MDN: A Web Portal for Network Analysis of Molecular Dynamics Simulations. Biophys J. 2015;109(6):1110-6.

120. Serçinoğlu O, Ozbek P. gRINN: a tool for calculation of residue interaction energies and protein energy network analysis of molecular dynamics simulations. Nucleic Acids Research. 2018;46(W1):W554-W62.

121. Eargle J, Luthey-Schulten Z. NetworkView: 3D display and analysis of protein·RNA interaction networks. Bioinformatics. 2012;28(22):3000-1.

122. 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.

123. Karami Y, Laine E, Carbone A. Dissecting protein architecture with communication blocks and communicating segment pairs. BMC Bioinformatics. 2016;17(S2).

124. Karami Y, Bitard-Feildel T, Laine E, Carbone A. "Infostery" analysis of short molecular dynamics simulations identifies highly sensitive residues and predicts deleterious mutations. Sci Rep. 2018;8(1):16126.

125. Harris BZ, Lim WA. Mechanism and role of PDZ domains in signaling complex assembly. Journal of Cell Science. 2001;114(18):3219.

126. Fan JS, Zhang M. Signaling complex organization by PDZ domain proteins. Neurosignals. 2002;11(6):315-21.

127. Hung AY, Sheng M. PDZ domains: structural modules for protein complex assembly. J Biol Chem. 2002;277(8):5699-702.

128. 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.

129. 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.

130. 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.

131. 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.

132. 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.

133. Gautier C, Laursen L, Jemth P, Gianni S. Seeking allosteric networks in PDZ domains. Protein Eng Des Sel. 2018;31(10):367-73.

134. Bostick M, Kim JK, Esteve P-O, Clark A, Pradhan S, Jacobsen SE. UHRF1 Plays a Role in Maintaining DNA Methylation in Mammalian Cells. Science. 2007;317:1760-4.

135. Hashimoto H, Horton JR, Zhang X, Bostick M, Jacobsen SE, Cheng X. The SRA domain of UHRF1 flips 5-methylcytosine out of the DNA helix. Nature. 2008;455:826-9.

136. Sharif J, Muto M, Takebayashi S, Suetake I, Iwamatsu A, Endo TA, et al. The SRA protein Np95 mediates epigenetic inheritance by recruiting Dnmt1 to methylated DNA. Nature. 2007;450(7171):908-12.

137. Jenkins Y, Markovtsov V, Lang W, Sharma P, Pearsall D, Warner J, et al. Critical role of the ubiquitin ligase activity of UHRF1, a nuclear RING finger protein, in tumor cell growth. Mol Biol Cell. 2005;16(12):5621-9.

138. Mancini M, Magnani E, Macchi F, Bonapace IM. The multi-functionality of UHRF1: epigenome maintenance and preservation of genome integrity. Nucleic Acids Res. 2021;49(11):6053-68.

139. Xie S, Qian C. The Growing Complexity of UHRF1-Mediated Maintenance DNA Methylation. Genes (Basel). 2018;9(12).

140. Nabel Christopher S, Kohli Rahul M. Demystifying DNA Demethylation. Science. 2011;333(6047):1229-30.

141. Lu X, Han D, Boxuan Simen Z, Song C-X, Zhang L-S, Doré LC, et al. Base-resolution maps of 5-formylcytosine and 5-carboxylcytosine reveal genome-wide DNA demethylation dynamics. Cell Research. 2015;25:386-9.

142. Neri F, Incarnato D, Krepelova A, Rapelli S, Anselmi F, Parlato C, et al. Single-Base resolution analysis of 5-formyl and 5-carboxyl cytosine reveals promoter DNA Methylation Dynamics. Cell Reports. 2015;10:674-83.

143. Shen L, Wu H, Diep D, Yamaguchi S, D'Alessio AC, Fung HL, et al. Genome-wide analysis reveals TET- and TDG-dependent 5-methylcytosine oxidation dynamics. Cell. 2013;153:692-706.

144. Globisch D, Münzel M, Müller M, Michalakis S, Wagner M, Koch S, et al. Tissue distribution of 5-hydroxymethylcytosine and search for active demethylation intermediates. PLoS ONE. 2010;5:1-9.

145. Eleftheriou M, Pascual AJ, Wheldon LM, Perry C, Abakir A, Arora A, et al. 5-Carboxylcytosine levels are elevated in human breast cancers and gliomas. Clinical epigenetics. 2015;7.

146. Arita K, Ariyoshi M, Tochio H, Nakamura Y, Shirakawa M. Recognition of hemi-methylated DNA by the SRA protein UHRF1 by a base-flipping mechanism. Nature. 2008;455(7214):818-21.

147. Avvakumov GV, Walker JR, Xue S, Li Y, Duan S, Bronner C, et al. Structural basis for recognition of hemi-methylated DNA by the SRA domain of human UHRF1. Nature. 2008;455(7214):822-5.

148. Fang J, Cheng J, Wang J, Zhang Q, Liu M, Gong R, et al. Hemi-methylated DNA opens a closed conformation of UHRF1 to facilitate its histone recognition. Nat Commun. 2016;7:11197.

149. Gelato KA, Tauber M, Ong MS, Winter S, Hiragami-Hamada K, Sindlinger J, et al. Accessibility of different histone H3-binding domains of UHRF1 is allosterically regulated by phosphatidylinositol 5-phosphate. Mol Cell. 2014;54(6):905-19.

150. Harrison JS, Cornett EM, Goldfarb D, DaRosa PA, Li ZM, Yan F, et al. Hemi-methylated DNA regulates DNA methylation inheritance through allosteric activation of H3 ubiquitylation by UHRF1. Elife. 2016;5.

151. Schneider M. Modelling of Ligand Induced Protein Movements Based on Experimental Data: Technische Universität München; 2016.

152. Clerico EM, Meng W, Pozhidaeva A, Bhasne K, Petridis C, Gierasch LM. Hsp70 molecular chaperones: multifunctional allosteric holding and unfolding machines. Biochem J. 2019;476(11):1653-77.

153. Rosenzweig R, Nillegoda NB, Mayer MP, Bukau B. The Hsp70 chaperone network. Nat Rev Mol Cell Biol. 2019.

154. Kohler V, Andreasson C. Hsp70-mediated quality control: should I stay or should I go? Biol Chem. 2020;401(11):1233-48.

155. Mayer MP. Intra-molecular pathways of allosteric control in Hsp70s. Philos Trans R Soc Lond B Biol Sci. 2018;373(1749).

156. Mayer MP, Gierasch LM. Recent advances in the structural and mechanistic aspects of Hsp70 molecular chaperones. J Biol Chem. 2019;294(6):2085-97.

157. Mayer MP. The Hsp70-Chaperone Machines in Bacteria. Front Mol Biosci. 2021;8:694012.

158. Zuiderweg ER, Hightower LE, Gestwicki JE. The remarkable multivalency of the Hsp70 chaperones. Cell Stress Chaperones. 2017;22(2):173-89.

159. Radons J. The human HSP70 family of chaperones: where do we stand? Cell Stress Chaperones. 2016;21(3):379-404.

160. Balchin D, Hayer-Hartl M, Hartl FU. In vivo aspects of protein folding and quality control. Science. 2016;353(6294):aac4354.

161. Patury S, Miyata Y, Gestwicki JE. Pharmacological targeting of the Hsp70 chaperone. Curr Top Med Chem. 2009;9(15):1337-51.

162. Ferraro M, D'Annessa I, Moroni E, Morra G, Paladino A, Rinaldi S, et al. Allosteric Modulators of HSP90 and HSP70: Dynamics Meets Function through Structure-Based Drug Design. J Med Chem. 2019;62(1):60-87.

163. Gestwicki JE, Shao H. Inhibitors and chemical probes for molecular chaperone networks. J Biol Chem. 2019;294(6):2151-61.

164. Mayer MP, Kityk R. Insights into the molecular mechanism of allostery in Hsp70s. Front Mol Biosci. 2015;2:58.

165. Wang W, Liu Q, Liu Q, Hendrickson WA. Conformational equilibria in allosteric control of Hsp70 chaperones. Mol Cell. 2021;81(19):3919-33 e7.

166. Karzai AW, McMacken R. A Bipartite Signaling Mechanism Involved in DnaJ-mediated Activation of the Escherichia coli DnaK Protein. Journal of Biological Chemistry. 1996;271(19):11236-46.

167. Laufen T, Mayer MP, Beisel C, Klostermeier D, Mogk A, Reinstein J, et al. Mechanism of regulation of Hsp70 chaperones by DnaJ cochaperones. P Natl Acad Sci USA. 1999;96(10):5452-7.

168. Marcinowski M, Höller M, Feige MJ, Baerend D, Lamb DC, Buchner J. Substrate discrimination of the chaperone BiP by autonomous and cochaperone-regulated conformational transitions. Nature Structural &Amp; Molecular Biology. 2011;18:150.

169. Pobre KFR, Poet GJ, Hendershot LM. The endoplasmic reticulum (ER) chaperone BiP is a master regulator of ER functions: Getting by with a little help from ERdj friends. J Biol Chem. 2019;294(6):2098-108.

170. Wang J, Lee J, Liem D, Ping P. HSPA5 Gene encoding Hsp70 chaperone BiP in the endoplasmic reticulum. Gene. 2017;618:14-23.

171. Lee AS. Glucose-regulated proteins in cancer: molecular mechanisms and therapeutic potential. Nat Rev Cancer. 2014;14(4):263-76.

172. Ni M, Zhang Y, Lee AS. Beyond the endoplasmic reticulum: atypical GRP78 in cell viability, signalling and therapeutic targeting. Biochem J. 2011;434(2):181-8.

173. Shin BK, Wang H, Yim AM, Le Naour F, Brichory F, Jang JH, et al. Global profiling of the cell surface proteome of cancer cells uncovers an abundance of proteins with chaperone function. J Biol Chem. 2003;278(9):7607-16.

174. Arap MA, Lahdenranta J, Mintz PJ, Hajitou A, Sarkis AS, Arap W, et al. Cell surface expression of the stress response chaperone GRP78 enables tumor targeting by circulating ligands. Cancer Cell. 2004;6(3):275-84.

175. Kim Y, Lillo AM, Steiniger SCJ, Liu Y, Ballatore C, Anichini A, et al. Targeting Heat Shock Proteins on Cancer Cells:  Selection, Characterization, and Cell-Penetrating Properties of a Peptidic GRP78 Ligand. Biochemistry. 2006;45(31):9434-44.

176. Liu Y, Steiniger SC, Kim Y, Kaufmann GF, Felding-Habermann B, Janda KD. Mechanistic studies of a peptidic GRP78 ligand for cancer cell-specific drug delivery. Mol Pharm. 2007;4(3):435-47.

177. Zhang Y, Liu R, Ni M, Gill P, Lee AS. Cell surface relocalization of the endoplasmic reticulum chaperone and unfolded protein response regulator GRP78/BiP. J Biol Chem. 2010;285(20):15065-75.

178. Gopal U, Pizzo SV. Cell surface GRP78 signaling: An emerging role as a transcriptional modulator in cancer. J Cell Physiol. 2021;236(4):2352-63.

179. Carlos AJ, Ha DP, Yeh DW, Van Krieken R, Tseng CC, Zhang P, et al. The chaperone GRP78 is a host auxiliary factor for SARS-CoV-2 and GRP78 depleting antibody blocks viral entry and infection. J Biol Chem. 2021;296:100759.

180. Katopodis P, Randeva HS, Spandidos DA, Saravi S, Kyrou I, Karteris E. Host cell entry mediators implicated in the cellular tropism of SARS-CoV-2, the pathophysiology of COVID-19 and the identification of microRNAs that can modulate the expression of these mediators (Review). Int J Mol Med. 2022;49(2):20.

181. Chu H, Chan CM, Zhang X, Wang Y, Yuan S, Zhou J, et al. Middle East respiratory syndrome coronavirus and bat coronavirus HKU9 both can utilize GRP78 for attachment onto host cells. J Biol Chem. 2018;293(30):11709-26.

182. Das JK, Roy S, Guzzi PH. Analyzing host-viral interactome of SARS-CoV-2 for identifying vulnerable host proteins during COVID-19 pathogenesis. Infect Genet Evol. 2021;93:104921.

183. Marcinowski M, Rosam M, Seitz C, Elferich J, Behnke J, Bello C, et al. Conformational selection in substrate recognition by Hsp70 chaperones. J Mol Biol. 2013;425(3):466-74.

184. Bonomo J, Welsh JP, Manthiram K, Swartz JR. Comparing the functional properties of the Hsp70 chaperones, DnaK and BiP. Biophys Chem. 2010;149(1-2):58-66.

185. Karlin S, Brocchieri L. Heat Shock Protein 70 Family: Multiple Sequence Comparisons, Function, and Evolution. Journal of Molecular Evolution. 1998;47(5):565-77.

186. Gething MJ, Blond-Elguindi S, Buchner J, Fourie A, Knarr G, Modrow S, et al. Binding Sites for Hsp70 Molecular Chaperones in Natural Proteins. Cold Spring Harbor Symposia on Quantitative Biology. 1995;60:417-28.

187. Flynn GC, Pohl J, Flocco MT, Rothman JE. Peptide-binding specificity of the molecular chaperone BiP. Nature. 1991;353(6346):726-30.

188. Rüdiger S, Buchberger A, Bukau B. Interaction of Hsp70 chaperones with substrates. Nature Structural Biology. 1997;4(5):342-9.

189. Gragerov A, Gottesman ME. Different Peptide Binding Specificities of hsp70 Family Members. Journal of Molecular Biology. 1994;241(2):133-5.

190. Hageman J, van Waarde Maria AWH, Zylicz A, Walerych D, Kampinga Harm H. The diverse members of the mammalian HSP70 machine show distinct chaperone-like activities. Biochemical Journal. 2011;435(1):127-42.

191. Knarr G, Modrow S, Todd A, Gething M-J, Buchner J. BiP-binding Sequences in HIV gp160. Journal of Biological Chemistry. 1999;274(42):29850-7.

192. Schneider M, Rosam M, Glaser M, Patronov A, Shah H, Back KC, et al. BiPPred: Combined sequence- and structure-based prediction of peptide binding to the Hsp70 chaperone BiP. Proteins. 2016;84(10):1390-407.

193. 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.

194. Schneider M, Antes I. SenseNet, a tool for analysis of protein structure networks obtained from molecular dynamics simulations. PLoS One. 2022;17(3):e0265194.

195. Schneider M, Antes I. Comparison of allosteric signaling in DnaK and BiP using mutual information between simulated residue conformations. Proteins. 2022.

196. 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. Journal of Chemical Theory and Computation. 2015;11(8):3696-713.

197. Tian C, Kasavajhala K, Belfon KAA, Raguette L, Huang H, Migues AN, et al. ff19SB: Amino-Acid-Specific Protein Backbone Parameters Trained against Quantum Mechanics Energy Surfaces in Solution. Journal of Chemical Theory and Computation. 2020;16(1):528-52.

198. Huang J, MacKerell Jr AD. CHARMM36 all-atom additive protein force field: Validation based on comparison to NMR data. Journal of Computational Chemistry. 2013;34(25):2135-45.

199. Jorgensen WL, Maxwell DS, Tirado-Rives J. Development and Testing of the OPLS All-Atom Force Field on Conformational Energetics and Properties of Organic Liquids. Journal of the American Chemical Society. 1996;118(45):11225-36.

200. Kaminski GA, Friesner RA, Tirado-Rives J, Jorgensen WL. Evaluation and Reparametrization of the OPLS-AA Force Field for Proteins via Comparison with Accurate Quantum Chemical Calculations on Peptides. The Journal of Physical Chemistry B. 2001;105(28):6474-87.

201. Robertson MJ, Tirado-Rives J, Jorgensen WL. Improved Peptide and Protein Torsional Energetics with the OPLSAA Force Field. J Chem Theory Comput. 2015;11(7):3499-509.

202. González MA. Force fields and molecular dynamics simulations. École thématique de la Société Française de la Neutronique. 2011;12:169-200.

203. Leimkuhler BJ, Reich S, Skeel RD. Integration Methods for Molecular Dynamics. In: Mesirov JP, Schulten K, Sumners DW, editors. Mathematical Approaches to Biomolecular Structure and Dynamics. New York, NY: Springer New York; 1996. p. 161-85.

204. and CS, Darden TA. MOLECULAR DYNAMICS SIMULATIONS OF BIOMOLECULES: Long-Range Electrostatic Effects. Annual Review of Biophysics and Biomolecular Structure. 1999;28(1):155-79.

205. Xiongwu W, Bernard RB. Molecular Simulation with Discrete Fast Fourier Transform. In: Salih Mohammed S, editor. Fourier Transform. Rijeka: IntechOpen; 2012. p. Ch. 7.

206. Harvey SC, Tan RK-Z, Cheatham III TE. The flying ice cube: Velocity rescaling in molecular dynamics leads to violation of energy equipartition. Journal of Computational Chemistry. 1998;19(7):726-40.

207. Braun E, Moosavi SM, Smit B. Anomalous Effects of Velocity Rescaling Algorithms: The Flying Ice Cube Effect Revisited. Journal of Chemical Theory and Computation. 2018;14(10):5262-72.

208. Van Durme J, Maurer-Stroh S, Gallardo R, Wilkinson H, Rousseau F, Schymkowitz J. Accurate prediction of DnaK-peptide binding via homology modelling and experimental data. PLoS Comput Biol. 2009;5(8):e1000475.

209. Eswar N, Webb B, Marti-Renom MA, Madhusudhan MS, Eramian D, Shen M-y, et al. Comparative Protein Structure Modeling Using Modeller. Current Protocols in Bioinformatics. 2006;15(1):5.6.1-5.6.30.

210. Hartmann C, Antes I, Lengauer T. IRECS: A new algorithm for the selection of most probable ensembles of side-chain conformations in protein models. Protein Science. 2007;16(7):1294-307.

211. Antes I. DynaDock: A new molecular dynamics-based algorithm for protein–peptide docking including receptor flexibility. Proteins: Structure, Function, and Bioinformatics. 2010;78(5):1084-104.

212. Knarr G, Gething MJ, Modrow S, Buchner J. BiP binding sequences in antibodies. J Biol Chem. 1995;270(46):27589-94.

213. Zahn M, Berthold N, Kieslich B, Knappe D, Hoffmann R, Sträter N. Structural Studies on the Forward and Reverse Binding Modes of Peptides to the Chaperone DnaK. Journal of Molecular Biology. 2013;425(14):2463-79.

214. Fan R-E, Chang K-W, Hsieh C-J, Wang X-R, Lin C-J. LIBLINEAR: A Library for Large Linear Classification. J Mach Learn Res. 2008;9:1871–4.

215. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. J Mach Learn Res. 2011;12(null):2825–30.

216. Marino Perez L, Ielasi FS, Bessa LM, Maurin D, Kragelj J, Blackledge M, et al. Visualizing protein breathing motions associated with aromatic ring flipping. Nature. 2022;602(7898):695-700.

217. 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.

218. 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.

219. Dijkstra EW. A note on two problems in connexion with graphs. Numerische Mathematik. 1959;1(1):269-71.

220. Brandes U. On Variants of Shortest-Path Betweenness Centrality and their Generic Computation. Social Networks. 2008;30(2):136-45.

221. Luo P, Baldwin RL. Interaction between water and polar groups of the helix backbone: An important determinant of helix propensities. Proceedings of the National Academy of Sciences. 1999;96(9):4930-5.

222. Rüdiger S, Germeroth L, Schneider-Mergener J, Bukau B. Substrate specificity of the DnaK chaperone determined by screening cellulose-bound peptide libraries. Embo j. 1997;16(7):1501-7.

223. Roomp K, Antes I, Lengauer T. Predicting MHC class I epitopes in large datasets. BMC Bioinformatics. 2010;11(1):90.

224. Antes I, Siu SWI, Lengauer T. DynaPred: A structure and sequence based method for the prediction of MHC class I binding peptide sequences and conformations. Bioinformatics. 2006;22(14):e16-e24.

225. Johnson OT, Gestwicki JE. Multivalent protein-protein interactions are pivotal regulators of eukaryotic Hsp70 complexes. Cell Stress Chaperones. 2022;27(4):397-415.

226. Nordquist EB, English CA, Clerico EM, Sherman W, Gierasch LM, Chen J. Physics-based modeling provides predictive understanding of selectively promiscuous substrate binding by Hsp70 chaperones. PLoS Comput Biol. 2021;17(11):e1009567.

227. Gutierres MBB, Bonorino CBC, Rigo MM. ChaperISM: improved chaperone binding prediction using position-independent scoring matrices. Bioinformatics. 2020;36(3):735-41.

228. Zhou T, Xiong J, Wang M, Yang N, Wong J, Zhu B, et al. Structural Basis for Hydroxymethylcytosine Recognition by the SRA Domain of UHRF2. Molecular Cell. 2014;54(5):879-86.

229. Qian C, Li S, Jakoncic J, Zeng L, Walsh MJ, Zhou M-M. Structure and Hemimethylated CpG Binding of the SRA Domain from Human UHRF1 *<sup></sup>. Journal of Biological Chemistry. 2008;283(50):34490-4.

230. Greiner VJ, Kovalenko L, Humbert N, Richert L, Birck C, Ruff M, et al. Site-Selective Monitoring of the Interaction of the SRA Domain of UHRF1 with Target DNA Sequences Labeled with 2-Aminopurine. Biochemistry. 2015;54(39):6012-20.

231. Bianchi C, Zangi R. UHRF1 discriminates against binding to fully-methylated CpG-Sites by steric repulsion. Biophysical Chemistry. 2013;171:38-45.

232. Spruijt CG, Gnerlich F, Smits AH, Pfaffeneder T, Jansen PWTC, Bauer C, et al. Dynamic readers for 5-(Hydroxy)methylcytosine and its oxidized derivatives. Cell. 2013;152:1146-59.

233. Pfaffeneder T, Spada F, Wagner M, Brandmayr C, Laube SK, Eisen D, et al. Tet oxidizes thymine to 5-hydroxymethyluracil in mouse embryonic stem cell DNA. Nature Chemical Biology. 2014;10(7):574-81.

234. Wang L, Zhou Y, Xu L, Xiao R, Lu X, Chen L, et al. Molecular basis for 5-carboxycytosine recognition by RNA polymerase II elongation complex. Nature. 2015;523(7562):621-5.

235. Hashimoto H, Olanrewaju YO, Zheng Y, Wilson GG, Zhang X, Cheng X. Wilms tumor protein recognizes 5-carboxylcytosine within a specific DNA sequence. Genes & Development. 2014;28(20):2304-13.

236. Jin S-G, Zhang Z-M, Dunwell Thomas L, Harter Matthew R, Wu X, Johnson J, et al. Tet3 Reads 5-Carboxylcytosine through Its CXXC Domain and Is a Potential Guardian against Neurodegeneration. Cell Reports. 2016;14(3):493-505.

237. Grin I, Ishchenko AA. An interplay of the base excision repair and mismatch repair pathways in active DNA demethylation. Nucleic Acids Research. 2016;44(8):3713-27.

238. Kohli RM, Zhang Y. TET enzymes, TDG and the dynamics of DNA demethylation. Nature. 2013;502(7472):472-9.

239. Bochtler M, Kolano A, Xu G-L. DNA demethylation pathways: Additional players and regulators. BioEssays. 2017;39(1):e201600178.

240. Mistry H, Tamblyn L, Butt H, Sisgoreo D, Gracias A, Larin M, et al. UHRF1 is a genome caretaker that facilitates the DNA damage response to γ-irradiation. Genome Integrity. 2010;1(1):7.

241. Tian Y, Paramasivam M, Ghosal G, Chen D, Shen X, Huang Y, et al. UHRF1 Contributes to DNA Damage Repair as a Lesion Recognition Factor and Nuclease Scaffold. Cell Reports. 2015;10(12):1957-66.

242. 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.

243. 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.

244. Taylor NR. Small world network strategies for studying protein structures and binding. Computational and structural biotechnology journal. 2013;5:e201302006.

245. Dubay KH, Bothma JP, Geissler PL. Long-range intra-protein communication can be transmitted by correlated side-chain fluctuations alone. PLoS Comput Biol. 2011;7(9):e1002168.

246. Kappel K, Wereszczynski J, Clubb RT, McCammon JA. The binding mechanism, multiple binding modes, and allosteric regulation of Staphylococcus aureus Sortase A probed by molecular dynamics simulations. Protein Sci. 2012;21(12):1858-71.

247. LeVine MV, Weinstein H. NbIT - A New Information Theory-Based Analysis of Allosteric Mechanisms Reveals Residues that Underlie Function in the Leucine Transporter LeuT. PLoS Computational Biology. 2014;10.

248. Lenaerts T, Ferkinghoff-Borg J, Stricher F, Serrano L, Schymkowitz JW, Rousseau F. Quantifying information transfer by protein domains: analysis of the Fyn SH2 domain structure. BMC Struct Biol. 2008;8:43.

249. Noe F, Horenko I, Schutte C, Smith JC. Hierarchical analysis of conformational dynamics in biomolecules: transition networks of metastable states. J Chem Phys. 2007;126(15):155102.

250. McClendon CL, Friedland G, Mobley DL, Amirkhani H, Jacobson MP. Quantifying Correlations Between Allosteric Sites in Thermodynamic Ensembles. Journal of Chemical Theory and Computation. 2009;5(9):2486-502.

251. 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.

252. Zhou H, Tao P. REDAN: Relative Entropy-Based Dynamical Allosteric Network Model. Mol Phys. 2019;117(9-12):1334-43.

253. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021;596(7873):583-9.

254. Stein RA, McHaourab HS. SPEACH_AF: Sampling protein ensembles and conformational heterogeneity with Alphafold2. PLoS Comput Biol. 2022;18(8):e1010483.

255. Kalescky R, Liu J, Tao P. Identifying key residues for protein allostery through rigid residue scan. J Phys Chem A. 2015;119(9):1689-700.

256. Huang Z, Zhu L, Cao Y, Wu G, Liu X, Chen Y, et al. ASD: a comprehensive database of allosteric proteins and modulators. Nucleic Acids Res. 2011;39(Database issue):D663-9.

257. Zha J, Li M, Kong R, Lu S, Zhang J. Explaining and Predicting Allostery with Allosteric Database and Modern Analytical Techniques. Journal of Molecular Biology. 2022;434(17):167481.

258. Kryshtafovych A, Schwede T, Topf M, Fidelis K, Moult J. Critical assessment of methods of protein structure prediction (CASP)-Round XIII. Proteins. 2019;87(12):1011-20.

259. Kityk R, Vogel M, Schlecht R, Bukau B, Mayer MP. Pathways of allosteric regulation in Hsp70 chaperones. Nat Commun. 2015;6:8308.

260. Amaral M, Kokh DB, Bomke J, Wegener A, Buchstaller HP, Eggenweiler HM, et al. Protein conformational flexibility modulates kinetics and thermodynamics of drug binding. Nat Commun. 2017;8(1):2276.

261. Yang J, Zong Y, Su J, Li H, Zhu H, Columbus L, et al. Conformation transitions of the polypeptide-binding pocket support an active substrate release from Hsp70s. Nature Communications. 2017;8(1):1201.

262. Bracher A, Verghese J. The nucleotide exchange factors of Hsp70 molecular chaperones. Front Mol Biosci. 2015;2:10.

263. Ota N, Agard DA. Intramolecular signaling pathways revealed by modeling anisotropic thermal diffusion. Journal of Molecular Biology. 2005;351:345-54.

264. 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.

265. 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.

266. Kong Y, Karplus M. The signaling pathway of rhodopsin. Structure. 2007;15(5):611-23.

267. 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. J Chem Theory Comput. 2014;10(12):5677-89.

268. 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.

269. Vijayabaskar MS, Vishveshwara S. Interaction energy based protein structure networks. Biophys J. 2010;99(11):3704-15.

270. Flyvbjerg H, Petersen HG. Error estimates on averages of correlated data. The Journal of Chemical Physics. 1989;91(1):461-6.

271. Lin J. Divergence measures based on the Shannon entropy. IEEE Transactions on Information Theory. 1991;37(1):145-51.

272. Kullback S, Leibler RA. On Information and Sufficiency. The Annals of Mathematical Statistics. 1951;22(1):79-86.

273. Luitz M, Bomblies R, Ostermeir K, Zacharias M. Exploring biomolecular dynamics and interactions using advanced sampling methods. Journal of Physics: Condensed Matter. 2015;27:323101.

274. Bernardi RC, Melo MC, Schulten K. Enhanced sampling techniques in molecular dynamics simulations of biological systems. Biochim Biophys Acta. 2015;1850(5):872-7.

275. Bonati L, Piccini G, Parrinello M. Deep learning the slow modes for rare events sampling. Proc Natl Acad Sci U S A. 2021;118(44).

# APPENDIX

PROTEINS
STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

# BiPPred: Combined sequence- and structure-based prediction of peptide binding to the Hsp70 chaperone BiP

Markus Schneider,[1] Mathias Rosam,[2] Manuel Glaser,[1] Atanas Patronov,[1,4] Harpreet Shah,[1] Katrin Christiane Back,[2] Marina Angelika Daake,[2] Johannes Buchner,[2,3] and Iris Antes[1,4*]

[1] Department Biowissenschaftliche Grundlagen, Technische Universität München, Freising, Germany

[2] Department Chemie, Technische Universität München, Garching, Germany

[3] Center for Integrated Protein Science, Department of Chemistry, Technische Universität München, Munich, Germany

[4] Center for Integrated Protein Science, Departments of Bioscience, Technische Universität München, Munich, Germany

## ABSTRACT

**Substrate binding to Hsp70 chaperones is involved in many biological processes, and the identification of potential substrates is important for a comprehensive understanding of these events. We present a multi-scale pipeline for an accurate, yet efficient prediction of peptides binding to the Hsp70 chaperone BiP by combining sequence-based prediction with molecular docking and MMPBSA calculations. First, we measured the binding of 15mer peptides from known substrate proteins of BiP by peptide array (PA) experiments and performed an accuracy assessment of the PA data by fluorescence anisotropy studies. Several sequence-based prediction models were fitted using this and other peptide binding data. A structure-based position-specific scoring matrix (SB-PSSM) derived solely from structural modeling data forms the core of all models. The matrix elements are based on a combination of binding energy estimations, molecular dynamics simulations, and analysis of the BiP binding site, which led to new insights into the peptide binding specificities of the chaperone. Using this SB-PSSM, peptide binders could be predicted with high selectivity even without training of the model on experimental data. Additional training further increased the prediction accuracies. Subsequent molecular docking (DynaDock) and MMGBSA/MMPBSA-based binding affinity estimations for predicted binders allowed the identification of the correct binding mode of the peptides as well as the calculation of nearly quantitative binding affinities. The general concept behind the developed multi-scale pipeline can readily be applied to other protein-peptide complexes with linearly bound peptides, for which sufficient experimental binding data for the training of classical sequence-based prediction models is not available.**

Proteins 2016; 84:1390–1407.

**Key words: peptide binding prediction; protein–peptide docking; MMPBSA calculations; MMGBSA calculations; Hsp70-substrate specificity.**

## INTRODUCTION

The heat shock protein 70 kDa (Hsp70) family is a major chaperone class, which can be found throughout all kingdoms of life.[1] Hsp70s bind to their protein substrates through extended peptide stretches, thus suppressing protein aggregation and assisting folding.[2] Hsp70s consist of an N-terminal nucleotide binding domain (NBD) and a C-terminal substrate binding domain (SBD) (Fig. 1). Structurally, the SBD is composed of a β-sheet "sandwich" harboring a cleft for substrate binding and an α-helical domain, the so-called lid.[2,4] Substrate affinity in the SBD is regulated by ATP hydrolysis in the NBD: The ATP bound state shows low substrate affinity, while the hydrolysis of ATP to ADP leads to efficient substrate binding.[5–7] Upon substrate binding and ATP hydrolysis, the chaperone undergoes large conformational changes.[8] In the ADP state, both domains are weakly coupled and the substrate-binding site is closed

Additional Supporting information may be found in the online version of this article.
Grant sponsor: Deutsche Forschungsgemeinschaft; Grant number: SFB1035; Grant sponsors: CIPSM Cluster of Excellence, TUM-Laura Bassi Award.
Markus Schneider and Mathias Rosam contributed equally to this work.
*Correspondence to: Iris Antes, Department Biowissenschaftliche Grundlagen, Technische Universität München, Freising, Germany. E-mail: antes@tum.de
Received 10 February 2016; Revised 8 May 2016; Accepted 19 May 2016
Published online 10 June 2016 in Wiley Online Library (wileyonlinelibrary.com).
DOI: 10.1002/prot.25084

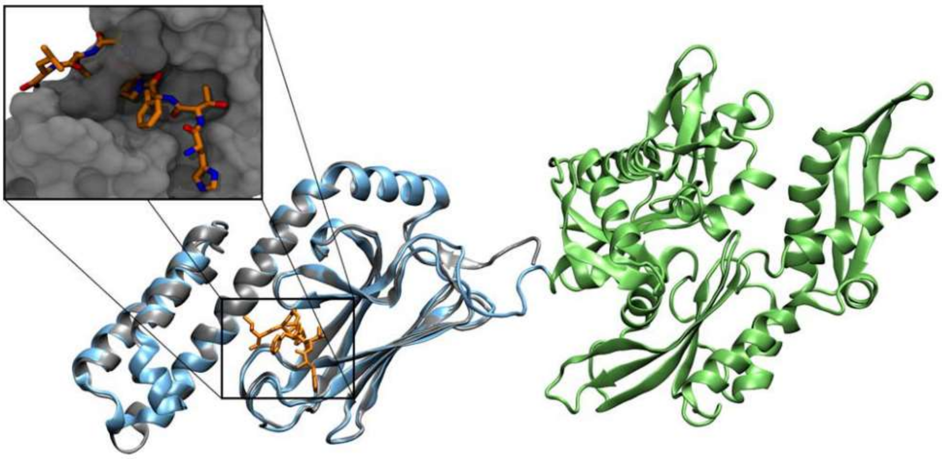


**Figure 1**

Structure of DnaK in its ADP-bound state (NMR, PDB-ID: 2KHO[3]) (green, nucleotide binding domain (NBD); iceblue, substrate binding domain (SBD)) with the superimposed homology model of the BiP-SBD (gray cartoon) with bound substrate peptide (HTFPAVL, orange licorice). The inset shows BiP's substrate binding cavity in surface representation with bound HTFPAVL.

(Fig. 1). Upon ATP binding, the domain interaction increases, leading to a respositioning of the lid, which in return allows the opening of the substrate binding site and substrate release.

The substrate specificity of Hsp70 chaperones has been studied extensively and it was observed that Hsp70s originating from different organisms and compartments differ in their substrate recognition.[9–11] It is commonly accepted that extended stretches of five to seven, preferably hydrophobic, residues are recognized.[1,2,12,13] The conserved binding site consists of an elongated groove with a deep, mainly hydrophobic binding pocket at the center (Fig. 1, inset). Structural and electrostatic complementarity of the central peptide residue with this central pocket is crucial for stable peptide binding.[14] Once stable side-chain interactions with this pocket are established, a hydrophobic bridge consisting of two binding site loops closes over the peptide backbone thus further stabilizing the peptide in its bound position.[14–16] Differences in the binding site in the various Hsp70s lead to variations in their substrate binding affinity and peptide kinetics and specificity.[14,17] For example, BiP, the endoplasmic Hsp70, accepts tryptophan residues in the binding site, whereas DnaK, the Hsp70 homolog in *Escherichia coli*, does not.[13,17,18] Both share the preference for leucine-enriched peptides.

In the present study, we focus on the peptide binding properties of BiP.[1,19] BiP is involved in many biological processes such as protein folding, quality control, and translocation into the ER.[20–23]

To identify binding sites of Hsp70 chaperones in proteins, computational prediction algorithms are needed as they allow a high-throughput, proteome-wide approach. However, despite long lasting efforts, the prediction of potential Hsp70 binding sites in proteins still remains challenging.[17,18,24,25] Several prediction models based on the sequences of known peptide substrates and nonbinders were developed for the identification of peptide stretches binding to Hsp70 binding sites. On the basis of the data from a bacteriophage-based study of the binding of random octa- to decapeptides to BiP, Blond-Elguindi *et al.* constructed a sequence-based, position-specific scoring matrix (PSSM) using the experimentally observed amino acid probabilities at each specific sequence position of the bound peptide as matrix terms. Evaluation of the prediction method on independent data sets[17,26] showed that the prediction accuracy varies between 60 and 80% depending on the tested protein substrate. The score was further improved by the inclusion of sequence information from known substrate proteins.[17] Rüdiger *et al.*[13,18] performed a systematic study of the substrate specificity of DnaK, the bacterial Hsp70, using peptide array data of 4260 tridecapeptides. This prediction approach is also based on the amino acid probability at each specific sequence position of the bound peptide. The method clusters them into a core and two flanking regions, and assigns "region-specific" scores for each amino acid. The algorithm can correctly predict about 80% of the strong binders and nonbinders in the benchmark set.

More recently, van Durme *et al.*[25] performed another study for DnaK, testing 172 decapeptides from seven known DnaK substrate proteins using the same experimental setup. The experimental data was used to first create a classical sequence-based PSSM based on the amino acid probability at each specific sequence position. Afterward, a combined sequence- and structure-based PSSM was established by performing an *in silico* structural position scan. For this the FoldX force field was

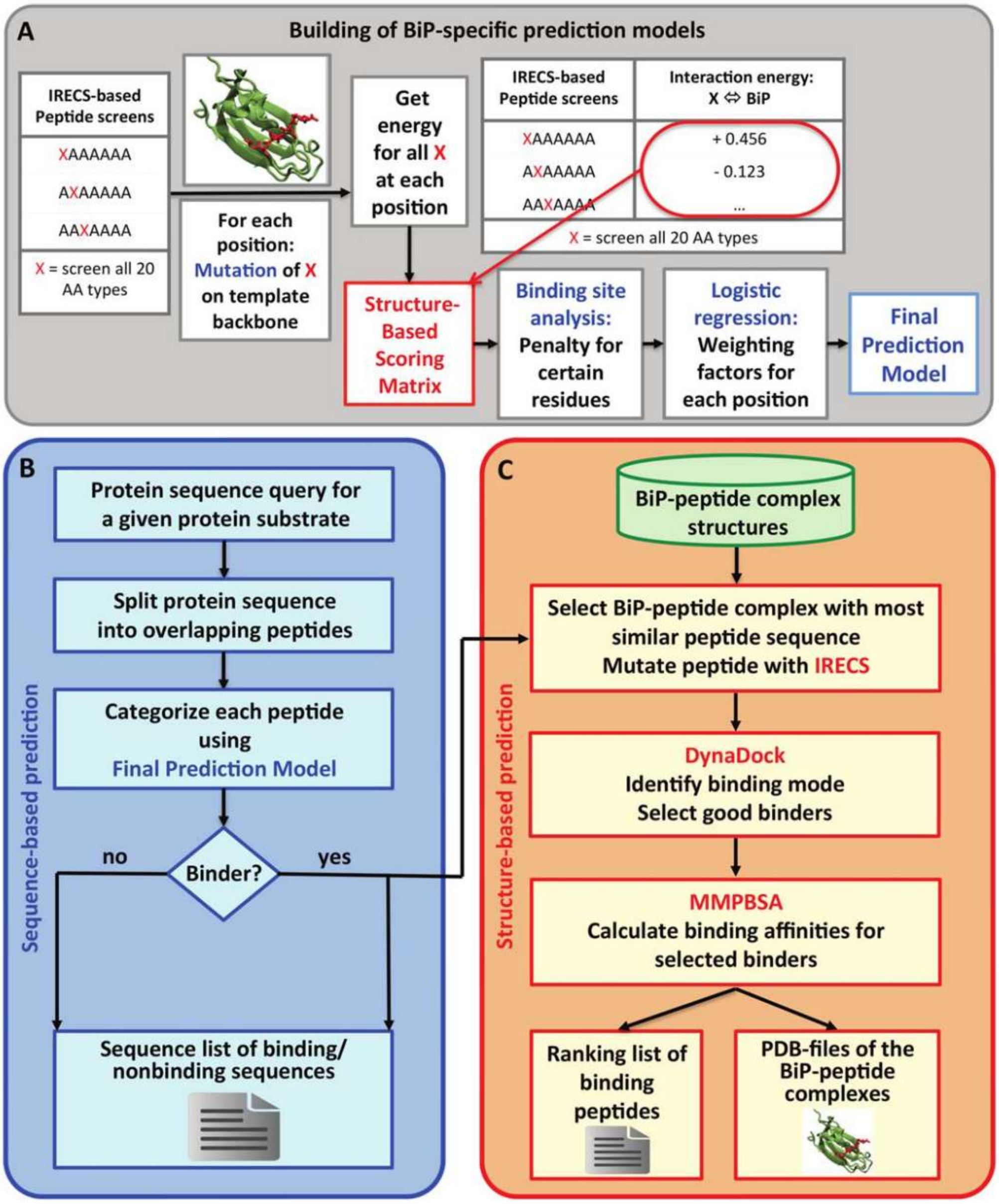

A
Building of BiP-specific prediction models
IRECS-based Peptide screens
XAAAAAA
AXAAAAA
AAXAAAA
X = screen all 20 AA types
For each position: Mutation of X on template backbone
Get energy for all X at each position
IRECS-based Peptide screens
Interaction energy: X ⇔ BiP
XAAAAAA + 0.456
AXAAAAA - 0.123
AAXAAAA ...
X = screen all 20 AA types
Structure-Based Scoring Matrix
Binding site analysis: Penalty for certain residues
Logistic regression: Weighting factors for each position
Final Prediction Model
B
Sequence-based prediction
Protein sequence query for a given protein substrate
Split protein sequence into overlapping peptides
Categorize each peptide using Final Prediction Model
Binder?
no
yes
Sequence list of binding/ nonbinding sequences
C
Structure-based prediction
BiP-peptide complex structures
Select BiP-peptide complex with most similar peptide sequence Mutate peptide with IRECS
DynaDock Identify binding mode Select good binders
MMPBSA Calculate binding affinities for selected binders
Ranking list of binding peptides
PDB-files of the BiP-peptide complexes

**Figure 2**

Pipeline for the development of the prediction models (**A**). Flowchart of the sequence-based (**B**) and of the structure-based (**C**) prediction steps of the final prediction pipeline.

used to thread each amino acid onto each position of the peptide using a polyalanine heptapeptide backbone, and to calculate its interaction energy with the DnaK binding site.[25] These interaction energies were added to the PSSM. The additional use of structural data improved the prediction accuracy considerably for the validation set, indicating that the use of structure-based information leads to more robust predictions than models based on peptide sequence data only. However, the differences in the performances of the individual models were strikingly larger for the validation set than for the benchmark set, which might be due to the small size of the validation set.

Regarding the overall accuracies of the prediction methods for Hsp70-peptide interactions, these are not as high as for other protein–peptide systems like for example major histocompatibility complex (MHC) peptide complexes. Extensive studies in the context of MHC-peptide binding led to three crucial prerequisites for the development of accurate sequence-based prediction models of protein-peptide binding.[27] First, the amount and comparability of the available experimental data is

**Table I**
Pentadecapeptides Tested for BiP Binding in Solution

| ID | Sequence[a] (N- to C-terminus) | Parent protein | Solubility[b] | DE[c] (%) | Pred. binders[d] | $\mathrm{Score_{max}}$[e] |
|---|---|---|---|---|---|---|
| 0 | HTFPAVL | $\mathrm{C_H1}$ | + | 101.9 | 1 | 1.00 |
| 75 | SSLGTQTYICNVNHK | $\mathrm{C_H1}$ | − − − | – | - | - |
| 81 | TYICNVNHKPSNTKV | $\mathrm{C_H1}$ | − − | – | - | - |
| 83 | ICNVNHKPSNTKVDK | $\mathrm{C_H1}$ | + | 35.1 | 7 | 0.84 |
| 213 | QHNKCECRPKKDRAR | VEGF | − | – | - | - |
| 225 | RARQENPCGPCSERR | VEGF | + | 11.2 | 3 | 0.69 |
| 238 | RRKHLFVQDPQTCKC | VEGF | + | 44.0 | 9 | 0.97 |
| 251 | KCSCKNTDSRCKARQ | VEGF | + | n.d. | 1 | 0.80 |
| 260 | RCKARQLELNERTCR | VEGF | + | 38.4 | 8 | 0.95 |

The displacement efficiency for Lucifer Yellow-labeled HTFPAVL (HTFPAVLGSC) was determined in fluorescence anisotropy measurements and represents the ratio of the change in anisotropy ($\Delta r$) for dissociation and $\Delta r$ for association. Dashes indicate that the parameter could not be determined due to artifacts during the measurement caused by peptides with poor solubility.

[a]To increase peptide stability, the N- and C-termini were synthesized in acetylated and amidated form, respectively.

[b]Subjective assessment of solubility upon addition of HKM buffer, that is, observation of visible aggregates in solution.

[c]DE: Displacement efficiency; n.d. not detectable.

[d]Number of heptamers within the 15mer peptide, predicted as binders with the CD-fitted IE/BA model corresponding to "Model 6" in Table II.

[e]$\mathrm{Score_{max}}$: highest score obtained within the group of predicted binders from column 5.

crucial. For a decent prediction accuracy, sequence-based prediction models should be based on at least 200 binding and 200 nonbinding sequences. Second, the data sets need to be well-balanced (equal amount of binding and nonbinding sequences). Third, a drop of 10–20% in accuracy can be observed if the exact binding register is not known and has to be predicted computationally.

In the case of Hsp70-peptide prediction models, the size of the experimental data sets used so far is either below or at the lower limit of 200 peptides in most studies. Furthermore, as there exist only a few binding sites in each protein substrate, a large imbalance between the number of binding and nonbinding sequences can lead to a bias in the prediction model. Third, in all studies, the length of the measured peptides ranges from 8 to 13 residues, thus the exact binding heptamer stretches (i.e., the exact binding registers) are experimentally not known. Therefore, the actual binding heptamer stretches were selected using various computational procedures, which might be error-prone. Next to these general accuracy-limiting factors, it was observed recently that peptides can bind to DnaK with their backbone placed alternatively in a so called "forward" or "reverse" direction[15] in the symmetrical binding site (Supporting Information Fig. S1); that is, the peptide is flipped by 180° with respect to its backbone direction. As the binding direction of the peptides in the binding assays is not known, the same (forward) direction (same as observed in the first experimental structures) is assumed for all peptides during the development of sequence-based prediction models. Therefore, the potential existence of a "reverse" binding mode introduces an additional error into the sequence-based models. Nevertheless, regarding the experimental structures available in the PDB at this time, 15 different peptides are bound in forward direction, while only 7 peptides were observed to bind in reverse orientation. As, in addition, the latter were all obtained by the same group and are thus structurally similar,[15] the introduced error should be tolerable.

Because of these challenges for the prediction of Hsp70-peptide binding, the major goal of this work was to design a prediction approach, which is based predominantly on structural "*ab initio*" modeling, using this data for the design of a structure-based position-specific scoring matrix (SB-PSSM). This allows predicting peptide binding to BiP reliably with high accuracy without the need of extensive high quality experimental binding data and knowledge of the exact binding mode of each peptide binder. We present a hierarchical approach which combines such a SB-PSSM-based prediction model with subsequent molecular docking and MMPBSA calculations allowing for identification, structural characterization, and binding affinity estimation of peptide sequences binding to BiP (Fig. 2). In addition, peptide array (PA) data and florescence anisotropy measurements were performed for the optimization and verification of the prediction results.

## MATERIALS AND METHODS

### Peptide array design and experiments

CelluSpot™ peptide array chips were purchased from Intavis (Cologne, Germany) and comprised 384 peptides spotted in duplicates onto the chip. The array was designed to contain peptides with a length of 15 amino acids overlapping by 14 residues, that is, 1 amino acid offset from spot to spot. The specific sequences of the parent proteins are summarized in Supporting Information Table SV. The peptide array was prepared according to the manufacturer's instructions. 14 μ*M* BiP was incubated with the peptides on the chip surface in HKM buffer (50 mM HEPES/KOH pH 7.5, 150 m*M* KCl, 10 m*M* $\mathrm{MgCl_2}$) containing 1 m*M* ADP at 37°C for 2 h.

**Table II**
Performance of the Different Prediction Models

| | | Data sets[a] | | Performance | |
|---|---|---|---|---|---|
| Model | SB-PSSM | Training set | Evaluation set | $AUC_{train}$[b] | $AUC_{eval}$[b] |
| 1 | IE[c] | — | CD | — | 0.48 |
| 2 | IE/4 | — | CD | — | 0.61 |
| 3 | IE/BA | — | CD | — | 0.83 |
| 4 | IE | CD | — | 0.71 | — |
| 5 | IE/4 | CD | — | 0.72 | — |
| 6 | IE/BA | CD | — | 0.85 | — |
| 7 | IE | $PA_{train}$ | $PA_{eval}$ | 0.68 | 0.74 |
| 7 | IE | $PA_{train}$ | CD | 0.68 | 0.59 |
| 8 | IE/4 | $PA_{train}$ | $PA_{eval}$ | 0.70 | 0.65 |
| 8 | IE/4 | $PA_{train}$ | CD | 0.70 | 0.51 |
| 9 | IE/BA | $PA_{train}$ | $PA_{eval}$ | 0.65 | 0.57 |
| 9 | IE/BA | $PA_{train}$ | CD | 0.65 | 0.58 |

[a] For the definition of the data sets see Material and Methods section.
[b] Area Under the Curve (AUC) values of the corresponding training sets ($AUC_{train}$) and the independent evaluation sets ($AUC_{eval}$).
[c] For the definition of the SB-PSSM models see Results section.

Bound BiP was detected by a primary anti-BiP antibody (1:5000) kindly provided by Linda Henderhot (St. Jude Children's Research Hospital, Memphis, TN) and a secondary anti-rabbit IgG antibody coupled to horseradish peroxidase (1:10,000; Sigma–Aldrich, St. Louis, MO).

## Peptide preparation and fluorescence anisotropy spectroscopy measurements

### Synthesized peptides

All peptides from Tables I–III were ordered from Biomatik (Cambridge, Canada) at a purity grade of 95% or higher. As published, HTFPAVL and SVFPLAP were synthesized without modification at their termini.[8,14] Because the termini of the peptides on the array were not charged, the individually ordered peptides were acetylated at the N terminus and amidated at the C terminus to increase stability.

### Peptide labeling and preparation of peptide stocks

HTFPAVLGSC was labeled with Lucifer Yellow as described before[14] and the synthesized peptides were dissolved in HKM buffer to a final concentration of 10 mM. Because of the lack of aromatic side chains, no extinction coefficients could be determined for the peptides. Therefore, the exact synthesized quantity (masses around 5 mg per peptide) and the molecular mass described in the quality control report of the manufacturer were used to calculate the amount of buffer needed to achieve a final concentration of 10 m*M* in the stock solution.

**Table III**
Predicted Heptapeptides Tested for BiP Binding in Solution

| ID | Sequence[a] (N- to C-terminus) | Parent protein | Solubility[b] | M3[c] | DE[d] (%) |
|---|---|---|---|---|---|
| 0 | HTFPAVL | $C_H1$ | + | B | 113.1 |
| HP1 | PGHPPRF | VpreB | + | NB | n.d. |
| HP2 | GPCSERR | VEGF | + | NB | – |
| HP3 | PQVPPRF | VpreB | + | *B* | n.d. |
| HP**4** | KDVARNR | VpreB | + | **B** | **39.4** |
| HP5 | QPEDEAM | VpreB | − | NB | n.d. |
| HP6 | MGARSSE | VpreB | + | NB | n.d. |
| HP7 | HPIETLV | VEGF | − − | B | 48.6 |
| HP**8** | PMAEGGG | VEGF | + | **B** | **75.9** |
| HP9 | FMDVYQR | VEGF | − − − | B | 49.3 |
| HP**10** | PPRFLLR | VpreB | + | **B** | **107.0** |

The displacement efficiency for Lucifer Yellow-labeled HTFPAVL (HTFPAVLGSC) was determined in fluorescence anisotropy measurements and represents the ratio of the change in anisotropy ($\Delta r$) for dissociation and $\Delta r$ for association. Dashes indicate that the parameter could not be determined due to artifacts during the measurement caused by peptides with poor solubility.
[a] To increase peptide stability, the N- and C-termini were synthesized in acetylated and amidated form, respectively.
[b] Subjective assessment of solubility upon addition of HKM buffer, that is, observation of visible aggregates in solution.
[c] M3: corresponds to "Model 3" in Table II; wrong predictions are highlighted as *italic* typeface. Peptides highlighted in **bold** typeface were found to bind to BiP and were further used for titration experiments. B, binder; NB, nonbinder.
[d] DE: Displacement efficiency; n.d., not detectable.

### *Fluorescence anisotropy spectroscopy*

For the detection of peptide association and dissociation kinetics to BiP, a Jasco FP-8500 spectrofluorimeter equipped with polarizers was thermostated at 37°C. Samples containing 1 μ*M* Lucifer Yellow (LY)-labeled HTFPAVLGSC and 1 m*M* ADP were equilibrated and measured in a 1-cm quartz cuvette at 37°C for ~15 min before 15 μM BiP was added. LY was excited at 428 nm and BiP-peptide association was followed at 525 nm with bandwidths of 5 and 10 nm for excitation and emission, respectively. Sensitivity was set to high and the time interval was 1 s. After reaching steady-state, a 150-fold molar excess of unlabeled peptide was added to the cuvette and dissociation was recorded. The kinetic parameters were derived from a single-exponential model similar to Ref. 14. Titration experiments were performed with increasing concentrations of the competing unlabeled peptide and fitted using Eq. (1)

$$y = F_{\mathrm{p}} - \left( (P_{\mathrm{t}}+x+K_{\mathrm{d}}) - \sqrt{(P_{\mathrm{t}}+x+K_{\mathrm{d}})^2 - (4P_{\mathrm{t}}x)} \right) \times \left( \frac{F_{\mathrm{p}} - F_{\mathrm{pl}}}{2P_{\mathrm{t}}} \right) \tag{1}$$

with the fluorescence signal of the peptide $F_{\mathrm{p}}$, the total concentration of the labeled peptide $P_{\mathrm{t}}$, the total concentration of the ligand (BiP) $x$, the affinity between peptide and ligand $K_{\mathrm{d}}$, and the fluorescence signal of the peptide-ligand complex $F_{\mathrm{pl}}$.

### BiP homology model

For the construction of the SB-PSSMs a previously described homology model of the BiP substrate binding domain (SBD) with a bound peptide (HTFPAVL)[14] was used (Fig. 1). It was created using the DnaK structure 1DKX.[4] The alignment of the protein sequences was performed with the align2d module of MODELLER.[28] Afterwards 400 structural models were created using MODELLER and the model with the best DOPE score was used for the further studies after manual inspection for soundness. The final model was energy minimized and slowly heated up to 300 K and equilibrated. A more detailed description of the modeling procedure is provided in Ref. 14

### Sequence-based prediction model

For the sequence-based prediction, a model was developed which uses an interaction energy-based structure-based position-specific scoring matrix (SB-PSSM), and is independent from the available experimental binding data [Fig. 2(A)]. Experimental data-based statistical learning was only used to adjust additional position-specific weights, which scale the contribution of each position of the bound peptide in the final score. All structural calculations use a previously built homology model of BiP.[14]

### Structure-based position-specific scoring matrix

For the elements of this matrix, first, a peptide library was created by introducing point mutations at each position of the bound peptide in the BiP homology model, while restraining its backbone conformation [Fig. 2(A)]. For this SB-PSSM the base peptide AAAPAAA was used instead of HTFPAVL to avoid any bias from the neighboring residues. A proline residue was placed in the central binding pocket instead of an alanine, as it was observed that this increases the overall stability of the peptide's position in the binding site. The mutations were performed with our in-house tool for protein side-chain prediction, IRECS,[29,30] by mutating all 20 proteinogenic amino acids onto all seven peptide sequence positions. This way a structural library of 7 × 20 BiP-peptide complexes was assembled and used for the derivation of the SB-PSSM. Afterwards, all 140 complexes were energy minimized using DynaCell[31] with the OPLS all-atom parameter set.[32] For the energy optimization, a step size of 0.002 nm and an energy convergence criterion of 1 kJ mol$^{-1}$ was used. Afterwards the pepscore-weighed Coulomb and Lennard–Jones interaction terms between the mutated residues and BiP were calculated according to Ref. 31 (for details about the pepscore scoring function see section "molecular docking simulations") for all energy minimized structures. These energy values were normalized over the whole matrix and the resulting normalized values form the basis for the SB-PSSM.

To further increase the prediction performance of the SB-PSSM, the binding site of the equilibrated BiP-HTFPAVL complex was comprehensively analyzed and for all peptides, which were mutated at position 4, molecular dynamics (MD) simulations were performed. The reason for the latter was that the central binding site was previously found to be very flexible, requiring a dynamic treatment to be able to accurately judge the binding properties of the different amino acids in that pocket.[14] For the molecular dynamics simulation, the same conditions as described in the MMPBSA section were used, but with the OPLS all-atom force field[32] instead of ff99SBildn, to be consistent with the IRECS mutation calculations. The SB-PSSM was further modified according to the results of these two analyses. A more detailed discussion of the modifications can be found in the Results section and a detailed description of the analyses is given in the Supporting Information.

In addition, in the final matrix, the values of the first and seventh position were set to zero as the variation in these amino acids was very small in the available binding data set (CD data set, see below). The reason for this simply is that they are considered unimportant for binding

**Table IV**
Experimental ($K_d$) and Computational Binding Data from DynaDock Docking, MMPBSA, and MMGBSA Calculations

| ID | Sequence[a] (N- to C-term) | $K_d$ (μM) | FF-score[b] (kJ mol$^{-1}$) | Pepscore[c] (kJ mol$^{-1}$) | Δ$G$ MMPBSA[d] (kcal mol$^{-1}$) | Δ$G$ MMGBSA[d] (kcal mol$^{-1}$) |
|---|---|---|---|---|---|---|
| HP4 | KDVARNR | 12.0 ± 3.1 | −2806.33 | −375.30 | −9.78 (0.50) | −63.44 (0.38) |
| | KDVARNR_r | – | −3113.55 | −402.35 | −15.46 (0.61) | −82.49 (0.43) |
| HP8 | PMAEGGG | 17.9 ± 6.0 | −2313.11 | −127.53 | 5.15 (0.39) | −39.09 (0.27) |
| | PMAEGGG_r | – | −2239.15 | −139.90 | 14.29 (0.50) | −34.92 (0.29) |
| HP10 | PPRFLLR | 9.7 ± 4.1 | −2290.83 | −439.79 | −23.23 (0.40) | −72.87 (0.31) |
| | PPRFLLR_r | – | −2051.27 | −436.52 | −5.71 (0.64) | −58.80 (0.37) |

All complexes were docked in forward and reverse binding mode ("_r") and the two binding modes were analyzed separately.
[a] To increase peptide stability, the N- and C-termini were synthesized in acetylated and amidated form, respectively.
[b] Docking pose with the best FF-score among the cluster representatives of the five biggest structural clusters.
[c] Docking pose with the best pepscore among the cluster representatives of the five biggest structural clusters (chosen for MMPBSA/MMGBSA analysis).
[d] The standard error of the mean is provided in parentheses.

and thus are (in contrary to the central three residues) not mutated in the corresponding experimental studies. This leads to an artificially high "sequence conservation," which biases the statistical learning step leading to unnaturally high weights for these positions and thus no gain in the prediction performance of the final model could be observed by the inclusion of these positions.

### Data sets for statistical learning

Experimental peptide binding data was collected from different resources: First, we included the data from the peptide array (PA) assay. As we assumed that each binding core should be defined as a heptamer, there are multiple possibilities for binding stretches in the tested 15mers and it is unknown which one of the possible heptapeptides in the corresponding 15mer peptide is responsible for the signal. Additionally, the signal of known heptameric binders can fluctuate depending on their position in the 15mer sequence. Therefore, we identified all 15mers in which the same heptamer is present and used the average over the PA signal intensities of these 15mers as a score to represent the heptamer's binding affinity. To create a robust data set, only the sequences with an average intensity >20% (binders) and those with an intensity <2% (nonbinders) were included in the PA final data set. Second, binding peptide sequences were taken from Knarr *et al.*[17,26] In this case we included all peptide sequences as binders in our training set, for which an ATPase stimulation factor of >1.5 was detected. Third, we included DnaK binding sequences from Zahn *et al.*[15] Data from both chaperones can be used in the statistical learning step as this step is only used to adjust the weights of the individual sequence positions. These weights are predominantly defined by the overall geometry of the binding site, which is the same in both chaperones. Fourth, we included several peptide sequences, which were identified in previous studies in our groups.[14]

On the basis of this data we defined three final data sets for the training of the SB-PSSM. The full PA data set was divided into two data sets, a training set ($\mathrm{PA_{train}}$) and an independent evaluation set ($\mathrm{PA_{eval}}$, 20% of the entire PA data set). The $\mathrm{PA_{train}}$ data set consists of 117 peptides (41 binding, 76 nonbinding sequences), while 30 peptides (19 binding, 11 nonbinding sequences) are included in the $\mathrm{PA_{eval}}$ set, respectively. The third data set (CD, Collected Data) consists of the binding peptides from resources 2–4[14,17,26] to which we added the definite nonbinding sequences from all sources (Supporting Information Table SIV).

### Statistical learning protocol

Training of the models was conducted using logistic regression realized by a C++ implementation of LIBLINEAR version 1.96 or Python 2.7 with the additional libraries skikit-learn 0.16.1, SciPy 0.16 and NumPy 1.9.2. L2-regularized logistic regression was used and the regularization strength parameter was optimized by threefold cross-validation for the best area under the curve (AUC) score in a response operator characteristic (ROC) curve. The samples were weighted inversely to the frequency of their respective class in the data set. The performance of the different prediction models was evaluated using ROC analysis. In the shown ROC plots the True Positive Rate (TPR) is plotted against the false positive rate (FPR) for different minimum scores required for a data point to be predicted as a binder. Good models have a fast-climbing TPR and a slow-climbing corresponding FPR, which leads to a high AUC in the plot. Random prediction algorithms are expected to yield an AUC of 0.5 (diagonal from (0,0) to (1,1)), while a perfect algorithm would have an AUC of 1.0 (step from (0,0) to (1,0)).

### Molecular docking simulations

As there exist no experimental structures of BiP-peptide complexes, the performance evaluation of the DynaDock method for Hsp70-peptide complexes was performed using seven DnaK-peptide complexes with known X-ray structures (PDB-ID: 1DKX,[4] 3DPO,[33] 3QNJ, 4E81, 4EZT, 4EZY, and 4EZZ[15]). For the three predicted binding peptides of BiP (Tables III and IV), the equilibrated homology model of the BiP-SBD with the HTFPAVL peptide (see above) was used and the amino acid side chains of the predicted peptides were mutated using the HTFPAVL peptide backbone with the side-chain placement tool IRECS.[29,30] In accordance to the experiment, the termini of the BiP peptide ligands were capped with N-terminal acetyl or C-terminal amide groups, respectively. To model the reverse binding mode of peptide ligands, the template peptide ligand was flipped along its backbone direction by 180°. The molecular docking simulations were performed using the DynaDock module of our in-house modeling program DynaCell.[31] A detailed description of the docking conditions is provided in the Supporting Information Text S1.

### MMPBSA and MMGBSA calculations

For the final, energetically best scoring docking poses from the DynaDock calculations 20 ns of molecular dynamics simulations were performed using the AmberTools14 package[34] and standard, established simulation conditions, as detailed in the Supporting Information Text S1. To estimate the free energy of binding of the bound peptides Molecular Mechanics-Poisson Boltzmann Surface Area (MMPBSA)[35–37] and Molecular Mechanics-Generalized Born Surface Area (MMGBSA)[36] calculations were performed, based on snapshot ensembles collected in 20 ps intervals from the last 5 ns of the 20 ns MD trajectories. MMPBSA and MMGBSA calculations were conducted with the MMPBSA.py script[38] from the AmberTools14 software package.[34] Further details (GB method, SA models, etc) are provided in the Supporting Information Text S1.

## RESULTS

In the following we present a new hierarchical approach for the prediction of peptides binding to the Hsp70 chaperone BiP. First, we will present the experimental data determined in this study, followed by the optimization and performance evaluation of the sequence-based prediction methods. Finally, we will discuss the results of the structural docking and binding free energy estimations. In Figure 2 a schematic overview over the approach is provided.

### Experimental peptide array and anisotropy studies

To obtain sufficient data for the parameterization of a sequence-based prediction model, a peptide array containing 384 overlapping peptides was designed from three secreted human proteins, the IgG1 antibody $C_H1$ domain, the vascular endothelial growth factor A (VEGF), and the surrogate light chain component (VpreB) (Fig. 3 and Supporting Information Table SV). As a control, the formerly identified binders SVFPLAP and HTFPAVL from $C_H1$ were included[14] together with the previously predicted nonbinder LRAEDMA (lowest BiP score) and the binder FTFSDYY (highest BiP score) from $V_H$.[26] All peptides had a length of 15 amino acids with an offset of one residue and were linked covalently with their C-terminus to cellulose spots on a glass chip.

Peptide spots associated with BiP were identified by immunodetection (Fig. 3). Both internal array replicates indicate a homogenous incubation with BiP and the absence of spots on the negative control shows the specificity of the detection antibodies. The array results were reproducible qualitatively but some minor differences existed in the quantitative evaluation, albeit the overall trend was not changed. The strongest BiP-peptide interactions were detected in the lower half of both arrays, indicating a stronger interaction with VEGF and VpreB than with $C_H1$. Quantification of the spot intensities allowed ranking the peptides and grouping them into a selection of protein- or array-wide top 10 binders (Supporting Information Table SI). In the array-wide ranking, the top 10 positions are predominantly occupied by VpreB peptides (seven peptides) showing the consensus sequence PRFSGSKDVARNR, followed by three peptides of VEGF occupying the positions 6, 8, and 10 in the array-wide ranking, respectively. The earliest entry for $C_H1$ appears at rank 59 (TYICNVNHKPSNTKV, peptide 81), featuring an array-wide relative signal intensity of only ~17%. This categorizes $C_H1$ in general as the weakest binding partner for BiP compared to VpreB and VEGF.

Surprisingly, from the known $C_H1$ binding sites SVFPLAP and HTFPAVL, only SVFPLAP containing 15mers were found in the $C_H1$-specific top 10 (3 appearances), but no HTFPAVL containing peptides (Supporting Information Table SI). Regarding the $C_H1$-wide comparison of the known binders, SVFPLAP-containing peptides yielded a maximum signal intensity of ~80% (peptide 4), while HTFPAVL peptides only reached a maximum intensity of ~40% (peptide 52; not present in top 10) (Supporting Information Table SI).

In summary, the peptide array data indicate that BiP can recognize peptides immobilized in a cellulose matrix with differing efficiency. Based on the array-wide ranking VpreB was identified as the most potent BiP binder, followed by VEGF and $C_H1$. However, the known binders SVFPLAP and HTFPAVL, which also served as positive control and share similar affinities for BiP in solution (12.5 and 11.1 μM, respectively[14]), do not show comparable signal intensities in the peptide array. This observation indicates that the affinities in solution do not necessarily correlate with the signal intensity on the chip.

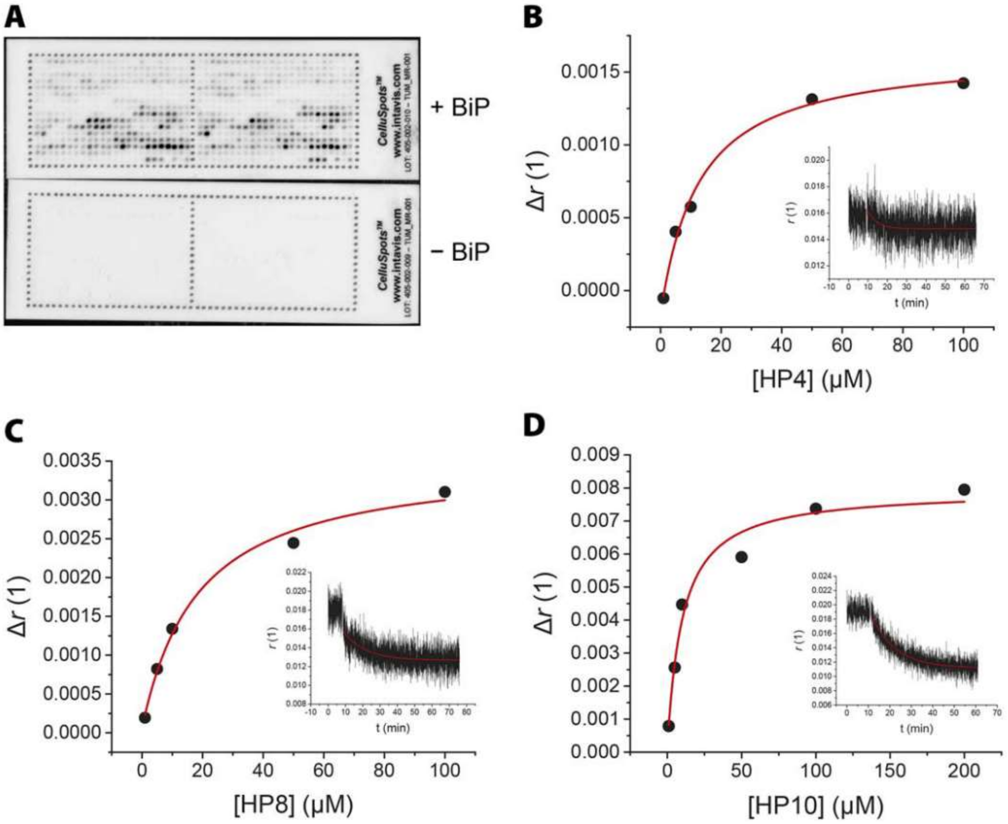


**Figure 3**

Identification of novel BiP binding sites in $C_H1$, VEGF, and VpreB and determination of binding affinities for selected true binders. (**A**) BiP was incubated with the peptide chip at 14 μM in HKM buffer containing 1 mM ADP for 2 h at 37°C. Bound BiP was detected with α-BiP (1:5,000) as primary and α-rabbit (1:10,000) as secondary antibody, respectively. Exposure time was 1 min and post-processing (Auto Contrast, overlay of luminescence and visible light images) was performed in Photoshop on the whole images. Each chip contained two identical arrays and the shown images are representative of at least three independent experiments. The addition of BiP is indicated on the right. (**B–D**) The difference in fluorescence anisotropy signal ($\Delta r$) between 1 μ*M* of fully bound and fully displaced HTFPAVLGSC by/from BiP was determined at different concentrations of the peptides HP4 (**B**), HP8 (**C**), and HP10 (**D**). $\Delta r$ was then plotted against the competitor concentration and fitted according to the formula in the material and methods section (red) to derive the $K_d$ values. The inset depicts an exemplary trace of HTFPAVLGSC displacement at 100 μ*M* competitor with a single-exponential fit. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

Because of the different BiP association behavior on the chip and in solution, a selection of the array-derived strong and medium binders was synthesized and their binding to BiP was tested *in vitro* by fluorescence anisotropy spectroscopy (Table I and Supporting Information Table SII). For these experiments HTFPAVL coupled to a C-terminal GSC linker (HTFPAVLGSC) was labeled with Lucifer Yellow (LY) and used to monitor the association with BiP. Once in steady-state, the BiP-HTPFPAVLGSC complex was dissociated with an excess of unlabeled HTFPAVL, thus proving that the peptide competes with its labeled counterpart for binding to BiP.

As visible from the changes in the anisotropy signal, all peptides with poor solubility (peptides 75, 81, and 213) produced artifacts during the measurements leading to a sudden signal increase upon their addition to the BiP-HTFPAVLGSC complex (Supporting Information Fig. S2, sharp signal spike between 30 and 40 min). These peptides could not be used for further experiments due to their undefined behavior and the light scattering properties of the formed aggregates.

The peptides 83, 225, 238, and 260 showed weak displacement of the labeled peptide as the anisotropy signal decreased only slightly over time. Even after 80 min, the strongest competitor, peptide 238, could displace only ~44% of the initially bound HTFPAVLGSC whereas unlabeled HTFPAVL displaced ~100% (Table I, Supporting Information Fig. S2). With 0.134 $\text{min}^{-1}$ the rate of displacement was higher than for HTFPAVL (Supporting Information Table SII). In a similar range, peptides 260 and 83 displaced ~38 and ~35% of the BiP-HTFPAVLGSC complex, respectively, but at different rates of 0.107 and 0.280 $\text{min}^{-1}$ (Table I, Supporting Information Fig. S2, Table SII). Peptide 225 only displayed a limited capability of binding to BiP with a displacement efficiency of 11%. These relatively low efficiencies and hence BiP binding capabilities were surprising, as the respective peptides showed strong signals on the peptide chip. This discrepancy indicates that immobilized peptides behave differently once they can move freely in solution and underlines the importance of a verification of binders in solution.

In summary, the fluorescence anisotropy measurements revealed that the binding intensity on the chip does not necessarily correlate with the peptide's ability to displace bound peptides from BiP in solution. For example, the best binder in solution was peptide 238, which did not even appear in the ranking of the top 10 binders on the chip.

### Design of the sequence-based prediction model

The study by van Durme *et al.* and several MHC-peptide binding studies showed that using binding energy data from structure-based calculations for the construction of the PSSM generally improves the performance of the prediction models.[25,27,39] However, all these approaches still rely strongly on the available experimental data, as either amino acid propensity-based data is additionally included in the PSSM or complex experimental data-dependent prediction models are trained on the basis of the calculated interaction energies using statistical learning procedures. Thus in both approaches the model's performance is still strongly correlated to the size and quality of the chosen experimental data, that is, the type of data used, the consistency of their measurement, and the overall experimental setup. As already discussed in the Introduction, in the case of Hsp70-peptide binding there exist serious limitations for such approaches due to the limited available data as well as the specific binding properties of the peptides (e.g., forward/reverse binding direction). Thus the aim of this study was to develop a computational procedure, which, in contrast to all previous studies, allows for the derivation of a highly accurate sequence-based prediction model without the use of experimental data. Therefore the derivation strategy presented here is based on the idea of developing an "*ab initio*" (i.e., without the use of experimental data) position-specific scoring matrix, solely obtained by structure-based modeling (SB-PSSM).

Specifically, all previous prediction approaches for Hsp70-peptide binding use either experimental data-based amino acid propensity scores alone or a combination of these with interaction energies from structural calculations as matrix elements in the PSSM. In contrary, we combine force field based interaction energies with results from extensive binding site analyses and molecular dynamics simulations for these terms. This leads to a PSSM, which is *a priori* independent of any experimental binding data (amino acid propensities). Nevertheless, as existing high-quality experimental data can also be used to further improve the performance of such a prediction model, we evaluated this possibility and additionally assigned one overall weighting factor to each sequence position of the peptide, that is, resulting in seven position-dependent weighting parameters (instead of 140 (7 × 20) amino acid and position-dependent propensity-based scores in the previous models). These weights were fitted on the basis of the available experimental data.

### Optimization and evaluation of the SB-PSSM-based prediction model

For the evaluation and optimization of the prediction model we focused on two aspects, namely the influence of the binding site analysis and molecular dynamics simulation-based modifications of the PSSM on the performance of the prediction model, and the influence of additional fitting of the position-based weights using experimental data.

Thus we built three different SB-PSSM matrices: The first matrix only contained the force field-based interaction energies (further referred to as IE, interaction energy), the second also contained MD-based modifications of the scores at peptide position 4 (binding into the central binding pocket) (IE/4, interaction energy/modified position 4), and the third matrix contained also all modifications at the other peptide positions based on the binding site analyses (IE/BA, interaction energy/binding site analyses).

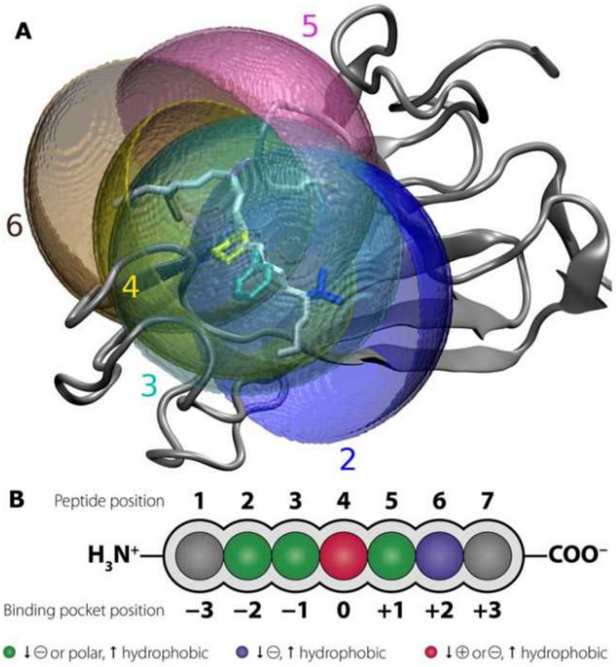
For these three matrices we investigated different settings for the training of the position-dependent weights and the model evaluation using three different experimental data sets: $PA_{train}$, $PA_{eval}$, and CD. For the first two data sets, the peptide array data was randomly divided into a training ($PA_{train}$) and an evaluation ($PA_{eval}$) set, representing 80 and 20% of the data, respectively. In addition, an independent third data set (CD, collected data) was built by a collection of experimentally verified binding sequences from literature and additional nonbinding sequences from the peptide array, as nearly no verified nonbinding data was available elsewhere. The CD set was collected in response to the discrepancies between experimental anisotropy and peptide array results observed in this study.

The performance and details of the 9 resulting prediction models are listed in Table II. The evaluation of the original IE matrix on the CD data set showed that the matrix had no distinct predictive value ($AUC = 0.48$) (see Model 1 in Table II). Additional training on the CD and peptide array ($PA_{train}$) data sets improved the results by about 20% leading to models with AUC values around 70% (Table II, Models 4 and 7), that is, showing a decent performance, which is in the same range as the previous Hsp70-peptide prediction models.

However, the performance of the unfitted IE model is about 10% lower than the accuracies of corresponding interaction energy-based matrices for other protein-peptide systems (e.g., for MHC-peptide binding the performance is around 60%). There are two main reasons for this relatively low performance, which can be attributed to the special geometry of the Hsp70 binding site.

First, in our previous work about the BiP-peptide binding features,[14] we observed that the central binding pocket of the binding site, interacting with the 4th residue of the heptamer peptides, is not only the most crucial for peptide specificity, but does also have a distinct conformational flexibility and is predominantly hydrophobic. This allows the pocket to accommodate hydrophobic and aromatic side chains of very different size, whereas charged side chains are normally rejected. In our mutation protocol this incompatibility led the IRECS algorithm to place the charged residues outside the binding pocket. Thus, no meaningful repulsive energies could be obtained, as outside the binding site, the side chains were placed on the protein surface in a more or less “interaction energy neutral position.” Therefore, more advanced sampling is necessary to properly describe the interaction between residue four and its binding pocket. As the pocket size and shape can change significantly, depending on the size and type of amino acid bound, we performed MD simulations for all BiP-peptide complexes obtained in our original IRECS-based mutation procedure, in which the peptide was mutated at position 4, that is, each AAAXAAA sequence with X located in or around the central binding pocket. These simulations were expected to provide a realistic picture of the binding site’s propensity to adapt to or to reject the corresponding amino acid.

The second accuracy-limiting feature is the open surface-like character of the regions of the binding site surrounding the central binding pocket. As a consequence, different backbone and side-chain positions are possible for the peptide residues in these regions and the exact conformation of the individual residue might depend on its neighboring residues. As the latter interdependence is not included in our alanine-based mutation protocol, this might lead to another drop in accuracy. A straightforward solution to this problem would be to sample all combinations of all 20 amino acids in each of the seven positions. This, however, would lead to 8 × $10^{15}$ combinations and is therefore not practically feasible. To overcome this issue, we designed a different, feasible strategy, which is based on a combined structural and interaction analysis of the bound peptide side-chain as obtained by the IRECS-based mutation protocol and the characteristic binding features of the binding site region responsible for binding the respective residue (see Fig. 4). For this purpose, sub-regions of the binding site were defined, including all residues located within a 10 Å sphere around the Cα-atoms of the residue of interest, which will be referred to as “residue sub-pockets.” A detailed description of the results of both analyses (MD and binding site analysis) is provided in the Supporting Information (Text S2).



C

| | | Peptide residue | | |
|---|---|---|---|---|
| 2 | 3 | 4 | 5 | 6 |
| T429 | T429 | F452 | E428 | A455 |
| T435 | V430 | S453 | T429 | S456 |
| K436 | T454 | T454 | V430 | Q459 |
| Q450 | A455 | A455 | G431 | P460 |
| I451 | R493 | Q459 | G432 | T461 |
| F452 | G494 | V462 | T463 | T463 |
| | V495 | T463 | K465 | K465 |
| | | I498 | K475 | D484 |
| | | V500 | D484 | |

**Figure 4**

Binding site analysis: (**A**) BiP binding site showing the residue centered spheres used for analysis of the individual residue surroundings. (**B**) Schematic drawing of the general properties that each peptide residue should have for good BiP binding properties. (**C**) Residues of BiP, which can interact with the corresponding peptide residue (i.e., which are located within the corresponding sphere).

On the basis of the molecular dynamics results (Fig. 5), we first modified the scores at position 4 in the SB-PSSM. A score of 1 was attributed to the side chains that were stable or that moved deeper into the cavity during simulation. To the side chains showing a semi-stable character, low cavity penetration, or missing polar contacts a score of 0.5 was assigned. Finally, amino acids with side chains that could not be placed near or in the binding pocket received a score of 0. The resulting

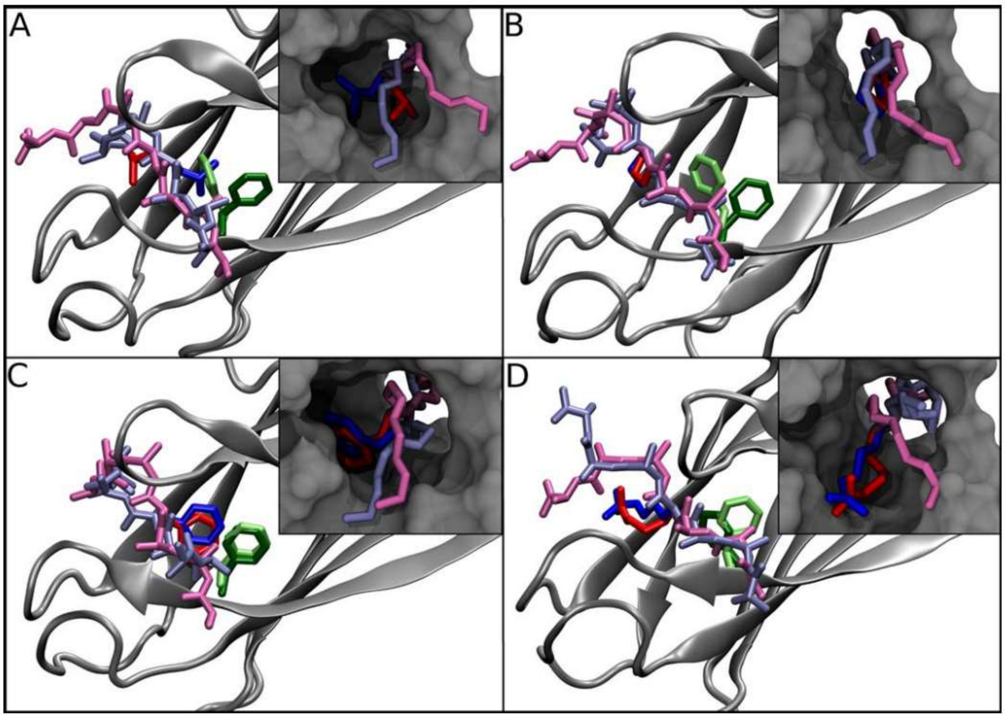


**Figure 5**

MD results for the AAAXAAA mutants, (**A**) X = L, (**B**) X = P, (**C**) X = F, (**D**) X = R. The starting conformation of the peptides is shown in red (the central amino acid residue in dark red, the rest of the peptide in light red) and the final conformation in blue (the central amino acid residue in dark blue, the rest of the peptide in light blue). The corresponding phenylalanine in the BiP binding site (F452) is shown in light (starting conformation) and dark green (final conformation). The insets show the central binding pocket in surface representation.

matrix is further referred to as IE/4. In a second step, we modified the scores for all other positions based on the binding site analyses. In that case the interaction energy-based scores were retained for "well" placed side chains, which were able to form stable interactions with the binding site. For side chains with medium binding characteristics a score of 0.5 was assigned, and amino acids with side chains unable to form any interaction received a score of 0. This led to the final matrix named IE/BA (Interaction Energy/Binding site Analysis), which contains all modifications (for a more detailed discussion, see Supporting Information).

To train and evaluate the new IE/4 and IE/BA matrices, we applied the same protocol as described above for the IE matrix. All results are listed in Table II, the ROC-curves derived for the Models 1–6 are given in Figure 6 and those for the Models 7–9 are shown in Supporting Information Figure S4. It can be observed that the performance of the unfitted Models (Model 1–3) increases consistently with the introduction of system-specific modifications using the CD data set for evaluation. The MD-based changes for position 4 increase the AUC value from 0.48 (Table II, Model 1) to 0.61 (Table II, Model 2). The final IE/BA matrix (Table II, Model 3), which includes all system-specific changes, features a remarkably high AUC value of 0.83, thus featuring already a very high selectivity even before any experimental data-based fitting of the position-based weights.

Nevertheless, we tried to further improve the performance by fitting weights for each residue position in the peptide, using the CD (Table II, Models 4–6) and $PA_{train}$ (Table II, Models 7–9) data sets. For the IE and IE/4 matrices significant performance improvements could be obtained by fitting these position weights by either of the both training data sets, leading to AUC values for the training sets between 0.68 and 0.72 (Table II, Models 4–5 and 7–8). However, evaluation of the models trained on the $PA_{train}$ data set (Models 7–8) led to inconsistent performance on the PA-based evaluation set ($PA_{eval}$) (AUCs between 0.57 and 0.74) and only to a very moderate performance of 0.51-0.59 (AUC) on the independent CD data set.

For the final IE/BA model, additional improvement by position-dependent weight fitting was only obtained by using the CD data set (Table II, Model 6), which led to a further increase in performance by 2% (Table II, Model 6 vs. Model 3). However, fitting of the position weights using the $PA_{train}$ data set (Table II, Model 9) led to a

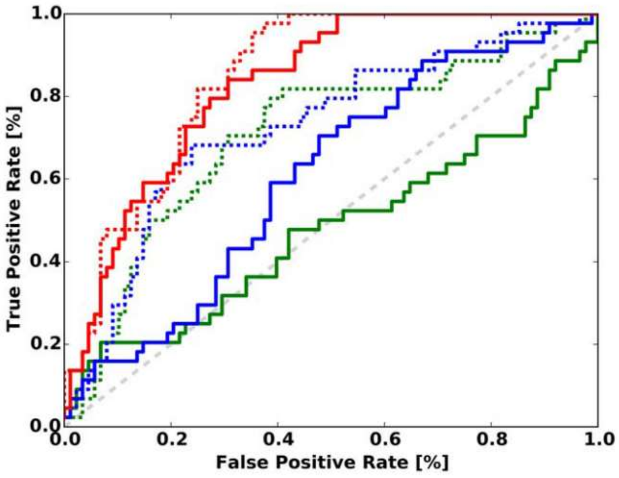


**Figure 6**

ROC plots for the prediction models from Table II: SB-PSSM: *IE = green, IE/4 = blue, IE/BA = red, original SB-PSSMs = straight line, CD-fitted = dotted line.*

decrease in accuracy for both, on the training set (AUC = 0.65) as well as on the evaluation on the two independent evaluation sets (AUCs of 0.57 and 0.58, respectively).

Overall, all AUC values obtained with models featuring $PA_{train}$ data set-fitted position weights show a lower accuracy than the CD-trained models and vary considerably in their performance. These differences in the performance depend predominantly on the training data set (between 0.51 and 0.74) and only slightly on the SB-PSSM matrix used. This indicates that the predictive capability of the peptide array data is around 65–70%, which is in agreement with the accuracy of the previous prediction models trained on PA data as discussed in the Introduction. It also agrees with the experimental discrepancies between the PA results and the fluorescence anisotropy measurements observed in the present study.

In summary, these results demonstrate the importance of a very careful system-specific design of the SB-PSSM as well as the necessity of a very conservative choice of the experimental data set used for fitting purposes.

### Identification of new binding peptide sequences using the final prediction model

Our next goal was to predict new binding heptamers in the PA data set and validate the prediction by measuring their displacement efficiency in fluorescence anisotropy experiments.

In the first study, we predicted the actual binding heptamer stretches within the measured 15mer binding peptides from Table I using the final CD-fitted IE/BA-based prediction model (Model 6 in Table II). As it was experimentally observed that peptides can bind in both sequence directions (Supporting Information Fig. S1),[15] we tested both, the forward and the reverse sequences. The number of resulting binders per 15mer together with the achieved maximum score are provided in Table I. A clear correlation can be found between the maximum score, the number of predicted binders, and the displacement efficiency. The correlation between the number of predicted binders and the displacement efficiency suggests that there might be multiple alternative binding registers, that is, more than one binding heptamer, within one 15mer. This would increase the overall stochastic probability of a binding event, potentially leading to more efficient displacement of HTFPAVLGSC. In general, all 15mers with a high displacement efficiency include at least seven predicted binding heptamers with a score of 0.84 or higher.

In a second study, we performed binding predictions on the whole PA data set to evaluate the prediction capacities of the IE/4 and the IE/BA matrices comprehensively. The predictions on the whole data set were performed using a PA-trained IE/4 model, as this was the best performing model available at that time. From the prediction results we selected the five highest ranking binding heptamer stretches (HP1-HP4, HP10) and the five lowest ranking nonbinding heptamers of the whole PA data set (HP5-HP9) (Supporting Information Table SII). Comparison of the experimental PA data and the prediction results showed a strong correlation, as all predicted strong binding heptamers were part of the top-10 VpreB/VEGF 15mer sequences with the highest signal intensities in the PA assays (Table SI) and all nonbinding heptamers could be located in 15mer peptides with very low intensities, respectively. This can be attributed to the PA-based training of the model. Afterward, the displacement efficiency was measured for all selected peptides by fluorescence anisotropy spectroscopy. The corresponding results are provided in Table III, Supporting Information Table SII, and Figure S3. The peptides HP4, HP7, HP8, HP9, and HP10 were identified as binders, displacing HTFPAVLGSC with different efficiencies. In contrast, the remaining heptapeptides HP1, HP2, HP3, HP5, and HP6 did not show any change in the anisotropy signal indicating that they could not be recognized by BiP. One peptide, HP2, repeatedly gave artifactual signals, perhaps caused by its instability in solution as calculated by ProtParam (http://web.expasy.org/protparam). Comparing these anisotropy data with the PA signal intensities, the correlation is again rather low (40%), in the same range as the correlation for the measured 15mers (Table I). Strikingly, the overall displacement efficiency for the predicted heptapeptide substrates was much higher than for the array-derived pentadecapeptides. Especially, peptide HP10 proved to be a potent BiP binder as it exerted an efficiency of 107%, characteristic for a true substrate of BiP (Table III, Supporting Information Fig. S3). The known binder

HTFPAVL yielded an efficiency of ~113% which is in the same range as HP10. The dissociation kinetics of the former and later peptides were also found to be very similar with 0.096 and 0.100 $\text{min}^{-1}$ for HTFPAVL and HP10, respectively (Table III, Supporting Information Fig. S3, Table SII). HP8 was also able to strongly compete with HTFPAVLGSC for BiP binding, as reflected by a displacement efficiency of ~76% (Table III). Here, the competition occurred at a slightly more elevated rate of 0.113 $\text{min}^{-1}$, a value still comparable to the control (Table III, Supporting Information Table SII). The highest displacement rate was achieved by HP7 ($\lambda_{\text{off}} = 0.184\ \text{min}^{-1}$) but this candidate could decrease the anisotropy signal only by about 49% (Table III, Supporting Information Table SII).

Finally, we performed additional predictions for all heptamers using the IE/BA matrix. Because the CD data set contains nonbinding sequences from the PA data set, the prediction on the PA data set might be biased using the CD-fitted IE/BA model. Therefore we used the non-fitted IE/BA matrix (Model 3 in Table II) for these calculations. Comparing these results to the anisotropy and the PA-based experimental binding data revealed a very strong correlation with the anisotropy data, as all peptides except one could be predicted correctly, thus leading to an accuracy of 90% with our non-fitted IE/BA model (Model 3 in Table II). At the same time the correlation to the PA-based data is in the same low range as the correlation between the two experimental data sets. These results demonstrate that the prediction model fitted to the PA data (PA-fitted IE/4 model) predicts this data very well, but not the peptide binding properties in solution, whereas the unfitted "*ab initio*" model shows a very high predictive power for peptide binding in solution. This confirms our original hypothesis that efficient, sequence-based peptide binding prediction should be possible on the basis of structural–biophysical interaction data and properties.

Thus, both independent evaluation studies on the 15mer and 7mer peptide sets demonstrate the high predictive power of the final unfitted as well as the CD-fitted IE/BA-based prediction models.

### Molecular docking and MMPBSA/MMGBSA binding affinity estimations

The results described above show that a reliable and selective identification of potential binding sequences of BiP is possible with the IE/BA prediction model. However, due to the special binding site features discussed above, an additional structure-based refinement of the prediction results is still necessary to reliably predict the binding direction (forward or reverse binding mode) and to provide an accurate estimation of the binding affinity for the ranking of the peptides.

Thus, in a second step, the DynaDock approach was evaluated for the prediction of Hsp70-peptide complex structures. The method was already successfully applied in a former study to identify important structural features of BiP-peptide binding.[14] Here we additionally evaluate its capability to predict the correct forward/backward binding mode of the peptide based on a new set of recent experimental structures of DnaK-peptide complexes.[15] As DnaK is the bacterial homolog of BiP, it features the same overall Hsp70-binding site geometry and is thus a suitable evaluation system. A detailed discussion of the evaluation results is provided in the Supporting Information (Text S1). In the performed molecular docking simulations, all peptides were docked in their experimental and in the alternative reverse orientation. These redocking experiments showed that in all cases the best-scoring peptide conformations feature RMSD values for the peptides smaller than 2 Å, demonstrating that with the applied DynaDock protocol it is not only possible to obtain accurately placed peptide conformations but also to identify them by the Hsp70-peptide interaction energy (Supporting Information Table SVI). In addition, in all cases except one, the binding orientation could be predicted correctly.

These results indicate that by combining fast IE/BA-based prediction of potential BiP-binding sequences and successive DynaDock simulations for the identified binders, highly accurate prediction results can be expected.

To specifically evaluate the hierarchical pipeline as shown in Figure 2 for BiP-peptide binding predictions, we performed molecular docking simulations for three of the predicted heptamers from Table III (HP4, HP8, and HP10, highlighted in bold typeface) using the same conditions as for the DynaDock evaluation studies. For the best-scoring docked complexes 20 ns of molecular dynamics simulations and successive MMPBSA/MMGBSA calculations were performed based on the last 5 ns of the simulations (Table IV). For the latter calculations the AMBER14 software package was used together with the AMBER ff99SBildn force field instead of the OPLS-all-atom force field used in the DynaDock studies, as in other studies performed in our group we obtained the best MMPBSA results by using the AMBER 14 suite of programs.[40] The results are provided in Table IV. For this analysis, also $K_d$ values were measured experimentally for these peptides using titration experiments.

For all three true binders, reasonable $K_d$ values for BiP recognition could be obtained with 12.0 μM for peptide HP4, 17.9 μM for peptide HP8, and 9.7 μM for peptide HP10 (Fig. 3, Table IV). Compared to HTFPAVL ($K_d = 11.1$ μM[14]), the affinities for BiP binding of all three true binders were in the same order of magnitude as expected for heptapeptides. For HP10, the affinity also correlated with the peptide's displacement efficiency since this peptide with the highest BiP affinity also represented the most efficient competitor for HTFPAVLGSC.

Regarding the molecular docking results (Table IV, FF-score), HP8 and HP10 should bind in a forward direction and HP4 in reverse. These binding modes are confirmed by follow-up MMPBSA/MMGBSA calculations, proving the robustness of the DynaDock-based prediction of the peptide orientation, especially considering that not only different methods, but also different force fields were used. Further analyzing the MMGBSA and MMPBSA results of the free energy of binding, qualitative agreement with the measured $K_d$ values can be observed for both methods. The MMPBSA approach provides the same binding affinity-based ranking for all peptides as the experiment (if HP4 is considered to bind in reverse mode), whereas the MMGBSA method leads to the ranking HP4 < HP10 < HP8, which are both very good results considering the rather small differences in the measured $K_d$ values. However, only the MMPBSA values are quantitatively in the same range as the measured $K_d$ values (i.e., to the corresponding $\Delta G$ values, which vary around $-6$ kcal/mol), whereas the MMGBSA data are ten times larger. The overestimation of binding affinities by the MMGBSA method agrees with all previous MMGBSA studies on protein–ligand binding affinity estimation and is most likely caused by use of the simple SASA (solvent accessible surface area) approximation of the non-polar term.[41] For the MMPBSA implementation a more advanced non-polar treatment is available and thus used in this study (see Materials and Methods), which allowed a quasi-quantitative estimation of experimental values by the MMPBSA method.

## DISCUSSION

In the first part of this work, we evaluated the overall accuracy and usefulness of peptide array measurements for the identification of binding peptides to the Hsp70 chaperone BiP. The inclusion of known binding sequences into the measured data set showed that the BiP-peptide affinity does not correlate well with the signal intensity on the chip. The well-studied $C_H1$ peptide HTFPAVL with demonstrated high affinity for BiP[14] was not present in the 10 $C_H1$ peptides with the highest intensity, while SVFPLAP with its lower BiP affinity is contained in three high-ranking peptides. To further evaluate this observation, several 15mer and heptamer peptides, which showed high intensities in the peptide array, were reinvestigated by fluorescence anisotropy spectroscopy experiments in solution (Tables I and III and Supporting Information Table SII). The results did not show a clear correlation between the peptide array intensities and the displacement efficiencies of the corresponding peptides in solution. This discrepancy indicates that the environment in solution and on the chip varies greatly, possibly due to the immobilization of the peptides and the accessible peptide concentration on each spot. Another potential reason might be the different lengths of the tested peptides, as 15mer peptides do have the potential to form short secondary structure elements, whereas heptamers do not.[42] This might alter the accessibility and thus the binding properties of a 7mer binding stretch in a 15mer peptide compared to individually tested 7mers, which could also be a potential explanation for the differences in the displacement efficiency values in Tables I and III, as the predicted heptapeptides show a tendency toward higher displacement values but at slower kinetics compared to the evaluated 15mers (Supporting Information Table SII). However, the number of peptides analyzed here is small and a comprehensive, systematic study would be necessary to draw final, unbiased conclusions. Overall, these results show that the peptide array data allows only a qualitative binary distinction between potential binding and nonbinding sequences. A subsequent BiP interaction analysis of the binding heptapeptides either experimentally in solution or computationally via docking calculations is still necessary.

As previous sequence-based prediction models for Hsp70-peptide binding are mainly using peptide array data, this might explain the limited accuracy of these models. Interestingly, in our case the use of the PA-based training and evaluation sets for position weight fitting of the different SB-PSSM matrices led to models which feature a similar accuracy as the previously published models (AUC values of 0.65 to 0.70 for the training and 0.51–0.74 for the validation sets) independent of the SB-PSSM used. $PA_{train}$-based position weight fitting to the IE/BA SB-PSSM even led to a decrease in accuracy. All $PA_{train}$-based models show an especially low AUC of around 0.50–0.60 if evaluated on the PA-independent CD data set, which predominantly contains peptide binding data determined by solution-based studies. In addition, in the second application study, the evaluation of the 7mer peptide binding properties (Table III), the PA-trained IE/4 model agreed perfectly with the experimental PA data, but showed only a 40–60% correlation to the fluorescence anisotropy displacement efficiency measurements and the predictions by the non-fitted IE/BA model.

All these observations reflect the discussed limited correlation between the experimental peptide array data on one side and the anisotropy results on the other, which can also explain why sequence-based prediction models trained on the peptide array data show a limited accuracy of about 70%. In addition, these models might be biased toward the PA data, thus explaining their even lower capability for predicting peptide binding in solution. Similar results were obtained by van Durme *et al.* who showed that although the performance of their sequence-based prediction model could be increased by the inclusion of structure-based data for the benchmark set (Matthews Correlation Coefficient (MCC) = 0.756), a strong decrease was observed for the validation set (MCC = 0.375) compared to the performance of the structure-based model alone (MCC = 0.593).

In addition, next to the intrinsic inaccuracies in the peptide array data, which are not Hsp70 specific, there are two more major reasons, why predictions based solely on sequence data from peptide (non)binders might not perform as well for Hsp70-peptide binding as for other protein–peptide systems. The predominant problem regarding the experimental peptide data is the selection of the exact binding heptamer stretches in the longer peptides measured. This computational selection was found to impair the prediction accuracy considerably also for other protein-peptide systems with experimentally undefined peptide binding core sequences (see MHC class II-peptide binding predictions). A previous study on DnaK binding could indeed show that the performance of the bacteriophage-based prediction model could be improved by the inclusion of specific heptamer binding data.[17,24] Van Durme *et al.*[25] relied on threading experiments to identify the correct binding heptamer within the 15mer peptides. As in our study successive peptides were tested, "average" heptamer-specific intensities could be calculated directly from the measured data, which should be more consistent with the measured data (see Material and Methods). Nevertheless, both approaches introduce an additional error, which is difficult to estimate. The second crucial issue is based on the recent observation that peptides can bind in both backbone directions, as the Hsp70 binding site is nearly symmetrical. For the training of a prediction model on peptide sequence data, a consistent binding direction is normally assumed, as experimental binding assays do not provide any information about the peptide's binding direction. This could lead to an additional error.

To minimize the influence of the issues discussed above, a structure-based position-specific scoring (SB-PSSM) was developed based on structural modeling and analysis, using the experimental data only to fit general weights for the individual residue positions in the peptide. As these weights are predominantly determined by the properties and shape of the binding site, they are less critically dependent on the correct orientation of all peptides in the data set. However, we still need to assume that the majority of the peptides bind in the forward direction as suggested by all fitted models.[15]

The evaluation of our first prediction model (IE) demonstrated that, due to the special geometry of the Hsp70 binding site, the system-independent straightforward mutation-based strategy needed to be adapted manually. Thus, we developed an improved SB-PSSM (IE/BA), which is not only based on interaction energy data, but also on MD simulations and on a static analysis of the binding site properties of BiP. Our final prediction matrix shows a very high selectivity and accuracy with an AUC value of 0.83 even without any parameterization on experimental data (Model 3, Table II). Position-weight fitting using the CD data set led to a further increase in accuracy of about 2% (Table II, Model 6 vs. Model 3). This small increase is presumably due to the limited size of the CD data set. It demonstrates, however, that the accuracy of the final model can still be improved by the use of accurate experimental binding data, indicating that with a large data set of certified binding and nonbinding heptamer sequences, even higher performances are possible.

Similar observations were made in several previous studies, in which it was demonstrated that the use of structure or interaction energy-based data can improve the performance and the robustness of a prediction model considerably, and lessen its dependence on the experimental data used.[25,27,39]

Comparing our approach to the previously published prediction models, the predominant difference is that all previous models are exclusively based on the sequence and binding data of the peptide substrates, but do not explicitly include binding site features. In contrast, we did not use the sequence information from peptide binding experiments for the PSSM, but instead included information of the structural features of the protein's binding site obtained by structural analysis and calculations together with interaction energies of the individual peptide residues. Thus our approach is *a priori* independent of any experimental binding information. Nevertheless, we showed that its performance can be further improved by an additional position-weight fitting to such data. With that strategy we were able to develop a prediction model, which shows a higher performance on an independent data set than the previous approaches based on peptide sequence data. However, the procedure cannot be automated yet and must be performed individually for each new protein-peptide system. Thus its development is much more time-consuming than the training of a standard prediction approach solely based on peptide sequence data. However, to our knowledge, no other comparable model, which performs equally well, exists at the moment. Therefore, if no peptide binding data is available, the current standard procedure is to perform molecular docking calculations. Although such calculations can lead to good results, they are too time-consuming for the screening of whole protein sequences and can only be performed for a preselected set of peptides, which needs to be obtained either experimentally or by non-system specific sequence analysis studies.

The practical evaluation of Model 3 (i.e., non-fitted IE/BA-based model) on the 7mer peptides from Table III showed an excellent correlation between the measured anisotropy data and the prediction results. Using Model 3, the binding properties of 9 out of 10 peptides could be predicted correctly. This independent evaluation study demonstrates impressively the power of "*ab initio*" structure-based prediction models. Therefore the new model can be used for a robust and solid identification of potential peptide binding sequences of BiP. As the selectivity of the model is solely based on structural and energetic data from the binding site analyses, it does not contain any bias with respect to the actual binding orientation of the peptides.

Using subsequent molecular docking calculations, very accurate bound peptide conformations could be obtained. In addition, it was possible to identify the correct binding direction of the peptides by their interaction energies. MMPBSA binding affinity estimations based on the energetically best docking solutions allowed to rank potential binding peptides with good accuracy and to obtain values that are quantitatively in the same range as the experimental values.

## CONCLUSIONS

Peptide arrays are a valuable tool for a fast, first screening of protein sequences and provide a general idea whether a peptide should be considered as binder or nonbinder. However, the BiP-binding properties of these candidates can differ considerably if measured in solution and thus the peptide array results needs to be verified in this case.

For the development of a SB-PSSM-based prediction model for BiP-peptide binding, the peptide array data was of limited value. This is in agreement with the limited performance of previous sequence- and structure-based prediction models of Hsp70-peptide binding, which rely mainly on experimental peptide binding data. However, very good prediction performances could be obtained with our final "*ab initio*" structure-based IE/BA prediction model. The corresponding SB-PSSM, optimized by careful analysis of the binding site properties of BiP, already showed high selectivity (AUC = 0.83) without any fitting to experimental data. By additional parameterization of position weights on the basis of solution-based, verified heptamer binding data, the accuracy of the model could be further improved. This is an encouraging result, as it demonstrates that it is possible to obtain highly predictive models for protein-peptide binding by system-specific structural modeling and analysis studies alone. This general concept allows the development of sequence-based prediction models for protein–peptide systems for which no or only few experimental data sets is available and for which no sequence-based prediction models exist currently.

## REFERENCES

1. Karlin S, Brocchieri L. Heat shock protein 70 family: multiple sequence comparisons, function, and evolution. J Mol Evol 1998;47: 565–577.
2. Gething M-J, Blond-Elguindi S, Buchner J, Fourie A, Knarr G, Modrow S, Nanu L, Segal M, Sambrook J. Binding sites for Hsp70 molecular chaperones in natural proteins. Cold Spring Harbor Symp Quantitative Biol 1995;60:417–428.
3. Bertelsen EB, Chang L, Gestwicki JE, Zuiderweg ER. Solution conformation of wild-type *E. coli* Hsp70 (DnaK) chaperone complexed with ADP and substrate. Proc Natl Acad Sci USA 2009;106:8471–8476.
4. Zhu X, Zhao X, Burkholder WF, Gragerov A, Ogata CM, Gottesman ME, Hendrickson WA. Structural analysis of substrate binding by the molecular chaperone DnaK. Science 1996;272:1606–1614.
5. Goloubinoff P, De Los Rios P. The mechanism of Hsp70 chaperones: (entropic) pulling the models together. Trends Biochem Sci 2007;32:372–380.
6. Swain JF, Dinler G, Sivendran R, Montgomery DL, Stotz M, Gierasch LM. Hsp70 chaperone ligands control domain association via an allosteric mechanism mediated by the interdomain linker. Mol Cell 2007;26:27–39.
7. Takeda S, McKay DB. Kinetics of peptide binding to the bovine 70 kDa heat shock cognate protein, a molecular chaperone. Biochemistry 1996;35:4636–4644.
8. Marcinowski M, Höller M, Feige MJ, Baerend D, Lamb DC, Buchner J. Substrate discrimination of the chaperone BiP by autonomous and cochaperone-regulated conformational transitions. Nat Struct Mol Biol 2011;18:150–158.
9. Gragerov A, Gottesman ME. Different peptide binding specificities of hsp70 family members. J Mol Biol 1994;241:133–135.
10. Hageman J, van Waarde MA, Zylicz A, Walerych D, Kampinga HH. The diverse members of the mammalian HSP70 machine show distinct chaperone-like activities. Biochem J 2011;435:127–142.
11. Wiech H, Buchner J, Zimmermann M, Zimmermann R, Jakob U. Hsc70, immunoglobulin heavy chain binding protein, and Hsp90 differ in their ability to stimulate transport of precursor proteins into mammalian microsomes. J Biol Chem 1993;268:7414–7421.
12. Flynn GC, Pohl J, Flocco MT, Rothman JE. Peptide-binding specificity of the molecular chaperone BiP. Nature 1991;353:726–730.
13. Rüdiger S, Buchberger A, Bukau B. Interaction of Hsp70 chaperones with substrates. Nat Struct Mol Biol 1997;4:342–349.
14. Marcinowski M, Rosam M, Seitz C, Elferich J, Behnke J, Bello C, Feige MJ, Becker CFW, Antes I, Buchner J. Conformational selection in substrate recognition by Hsp70 chaperones. J Mol Biol 2013;425: 466–474.
15. Zahn M, Berthold N, Kieslich B, Knappe D, Hoffmann R, Strater N. Structural studies on the forward and reverse binding modes of peptides to the chaperone DnaK. J Mol Biol 2013;425:2463–2479.
16. Rudiger S, Mayer MP, Schneider-Mergener J, Bukau B. Modulation of substrate specificity of the DnaK chaperone by alteration of a hydrophobic arch. J Mol Biol 2000;304:245–251.
17. Knarr G, Modrow S, Todd A, Gething MJ, Buchner J. BiP-binding sequences in HIV gp160. Implications for the binding specificity of BiP. J Biol Chem 1999;274:29850–29857.
18. Rüdiger S, Germeroth L, Schneider-Mergener J, Bukau B. Substrate specificity of the DnaK chaperone determined by screening cellulose-bound peptide libraries. EMBO J 1997;16:1501–1507.
19. Munro S, Pelham HR. An Hsp70-like protein in the ER: identity with the 78 kd glucose-regulated protein and immunoglobulin heavy chain binding protein. Cell 1986;46:291–300.
20. Dudek J, Greiner M, Muller A, Hendershot LM, Kopsch K, Nastainczyk W, Zimmermann R. ERj1p has a basic role in protein biogenesis at the endoplasmic reticulum. Nat Struct Mol Biol 2005; 12:1008–1014.
21. Kabani M, Kelley SS, Morrow MW, Montgomery DL, Sivendran R, Rose MD, Gierasch LM, Brodsky JL. Dependence of endoplasmic reticulum-associated degradation on the peptide binding domain and concentration of BiP. Mol Biol Cell 2003;14:3437–3448.
22. Kassenbrock CK, Garcia PD, Walter P, Kelly RB. Heavy-chain binding protein recognizes aberrant polypeptides translocated in vitro. Nature 1988;333:90–93.

23. Alder NN, Shen Y, Brodsky JL, Hendershot LM, Johnson AE. The molecular mechanisms underlying BiP-mediated gating of the Sec61 translocon of the endoplasmic reticulum. J Cell Biol 2005;168: 389–399.
24. Blond-Elguindi S, Cwirla SE, Dower WJ, Lipshutz RJ, Sprang SR, Sambrook JF, Gething M-JH. Affinity panning of a library of peptides displayed on bacteriophages reveals the binding specificity of BiP. Cell 1993;75:717–728.
25. Van Durme J, Maurer-Stroh S, Gallardo R, Wilkinson H, Rousseau F, Schymkowitz J. Accurate prediction of DnaK-peptide binding via homology modelling and experimental data. PLoS Computat Biol 2009;5:e1000475.
26. Knarr G, Gething M-J, Modrow S, Buchner J. BiP binding sequences in antibodies. J Biol Chem 1995;270:27589–27594.
27. Roomp K, Antes I, Lengauer T. Predicting MHC class I epitopes in large datasets. BMC Bioinform 2010;11:90.
28. John B, Sali A. Comparative protein structure modeling by iterative alignment, model building and model assessment. Nucleic Acids Res 2003;31:3982–3992.
29. Hartmann C, Antes I, Lengauer T. IRECS: a new algorithm for the selection of most probable ensembles of side-chain conformations in protein models. Protein Sci Publ Protein Soc 2007;16:1294–1307.
30. Hartmann C, Antes I, Lengauer T. Docking and scoring with alternative side-chain conformations. Proteins 2009;74:712–726.
31. Antes I. DynaDock: a new molecular dynamics-based algorithm for protein-peptide docking including receptor flexibility. Proteins 2010; 78:1084–1104.
32. Jorgensen WL, Maxwell DS, TiradoRives J. Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. J Am Chem Soc 1996;118:11225–11236.
33. Liebscher M, Roujeinikova A. Allosteric coupling between the lid and interdomain linker in DnaK revealed by inhibitor binding studies. J Bacteriol 2009;191:1456–1462.
34. Case DA, Babin V, Berryman JT, Betz RM, Cai Q, Cerutti DS, Cheatham TE, III, Darden TA, Duke RE, Gohlke H, Goetz AW, Gusarov S, Homeyer N, Janowski P, Kaus J, Kolossváry I, Kovalenko A, Lee TS, LeGrand S, Luchko T, Luo R, Madej B, Merz KM, Paesani F, Roe DR, Roitberg A, Sagui C, Salomon-Ferrer R, Seabra G, Simmerling CL, Smith W, Swails J, Walker RC, Wang J, Wolf RM, Wu X, Kollman PA. AMBER 14. University of California, San Francisco; 2014.
35. Kollman PA, Massova I, Reyes C, Kuhn B, Huo S, Chong L, Lee M, Lee T, Duan Y, Wang W, Donini O, Cieplak P, Srinivasan J, Case DA, Cheatham TE, III. Calculating structures and free energies of complex molecules: combining molecular mechanics and continuum models. Acc Chem Res 2000;33:889–897.
36. Srinivasan J, Cheatham TE, III, Cieplak P, Kollman PA, Case DA. Continuum solvent studies of the stability of DNA, RNA, and phosphoramidate−DNA helices. J Am Chem Soc 1998;120:9401–9409.
37. Chong LT, Duan Y, Wang L, Massova I, Kollman PA. Molecular dynamics and free-energy calculations applied to affinity maturation in antibody 48G7. Proc Natl Acad Sci US A 1999;96:14330–14335.
38. Bill R, Miller BR, III, McGee TD, Jr, Swails JM, Homeyer N, Gohlke H, Roitberg AE. MMPBSA.py: an efficient program for end-state free energy calculations. J Chem Theory Comput 2012;8:3314–3321.
39. Antes I, Siu SW, Lengauer T. DynaPred: a structure and sequence based method for the prediction of MHC class I binding peptide sequences and conformations. Bioinformatics 2006;22:e16–e24.
40. Salomon-Ferrer R, Götz AW, Poole D, Grand SL, Walker RC. Routine microsecond molecular dynamics simulations with AMBER on GPUs. 2. Explicit solvent particle mesh Ewald. J Chem Theory Comput 2013;9:3878–3888.
41. Sun H, Li Y, Tian S, Xu L, Hou T. Assessing the performance of MM/PBSA and MM/GBSA methods. 4. Accuracies of MM/PBSA and MM/GBSA methodologies evaluated by various simulation protocols using PDBbind data set. Phys Chem Chem Phys 2014;16: 16719–16729.
42. Luo P, Baldwin RL. Interaction between water and polar groups of the helix backbone: An important determinant of helix propensities. Proc Natl Acad Sci USA 1999;96:4930–4935.

RESEARCH ARTICLE

# Systematic analysis of the binding behaviour of UHRF1 towards different methyl- and carboxylcytosine modification patterns at CpG dyads

**Markus Schneider[1☯], Carina Trummer[2☯], Andreas Stengl[2], Peng Zhang[2,3], Aleksandra Szwagierczak[2], M. Cristina Cardoso[3], Heinrich Leonhardt[2], Christina Bauer[2¤], Iris Antes[1]***

**1** Center for Integrated Protein Science Munich at the TUM School of Life Sciences, Technische Universität München, Freising, Germany, **2** Center for Integrated Protein Science Munich at the Department of Biology II, Ludwig Maximilians University Munich, Planegg-Martinsried, Germany, **3** Cell Biology and Epigenetics at the Department of Biology, Technische Universität Darmstadt, Darmstadt, Germany

☯ These authors contributed equally to this work.
¤ Current address: Department of Biomedicine, University of Basel, Basel, Switzerland
* antes@tum.de

**Citation:** Schneider M, Trummer C, Stengl A, Zhang P, Szwagierczak A, Cardoso MC, et al. (2020) Systematic analysis of the binding behaviour of UHRF1 towards different methyl- and carboxylcytosine modification patterns at CpG dyads. PLoS ONE 15(2): e0229144. https://doi.org/10.1371/journal.pone.0229144

**Editor:** Albert Jeltsch, Universität Stuttgart, GERMANY

**Received:** September 4, 2019

**Accepted:** January 30, 2020

**Published:** February 21, 2020

**Data Availability Statement:** All relevant data are within the manuscript and its Supporting Information files.

**Funding:** This work was supported by the Deutsche Forschungsgemeinschaft (SFB 1035/A10 and SFB749/C08 to IA, SFB 1064/A17 and SFB 1243/A01 to HL, CIPSM to IA and HL, CA198/10-1 to MCC). MS was supported by the TUM International Graduate School of Science and Engineering (IGSSE). CB and CT were supported

## Abstract

The multi-domain protein UHRF1 is essential for DNA methylation maintenance and binds DNA via a base-flipping mechanism with a preference for hemi-methylated CpG sites. We investigated its binding to hemi- and symmetrically modified DNA containing either 5-methylcytosine (mC), 5-hydroxymethylcytosine (hmC), 5-formylcytosine (fC), or 5-carboxylcytosine (caC). Our experimental results indicate that UHRF1 binds symmetrically carboxylated and hybrid methylated/carboxylated CpG dyads in addition to its previously reported substrates. Complementary molecular dynamics simulations provide a possible mechanistic explanation of how the protein could differentiate between modification patterns. First, we observe different local binding modes in the nucleotide binding pocket as well as the protein's NKR finger. Second, both DNA modification sites are coupled through key residues within the NKR finger, suggesting a communication pathway affecting protein-DNA binding for carboxylcytosine modifications. Our results suggest a possible additional function of the hemi-methylation reader UHRF1 through binding of carboxylated CpG sites. This opens the possibility of new biological roles of UHRF1 beyond DNA methylation maintenance and of oxidised methylcytosine derivates in epigenetic regulation.

## Introduction

UHRF1 (also referred to as Np95) is an essential protein for DNA methylation maintenance in mammals. It consists of 5 domains: A ubiquitin-like domain, a Tandem-Tudor domain, a PHD domain, a DNA-binding SRA domain, and a RING domain with E3 ubiquitin ligase activity (Fig 1a) [1–3]. UHRF1 was originally reported to preferentially bind to hemi-

by the International Max Planck Research School for Molecular Life Sciences (IMPRS-LS). AS was trained and supported by the graduate school GRK1721 of the Deutsche Forschungsgemeinschaft. PZ received support from the “Vereinigung von Freunden der Technischen Universität Darmstadt”. 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.

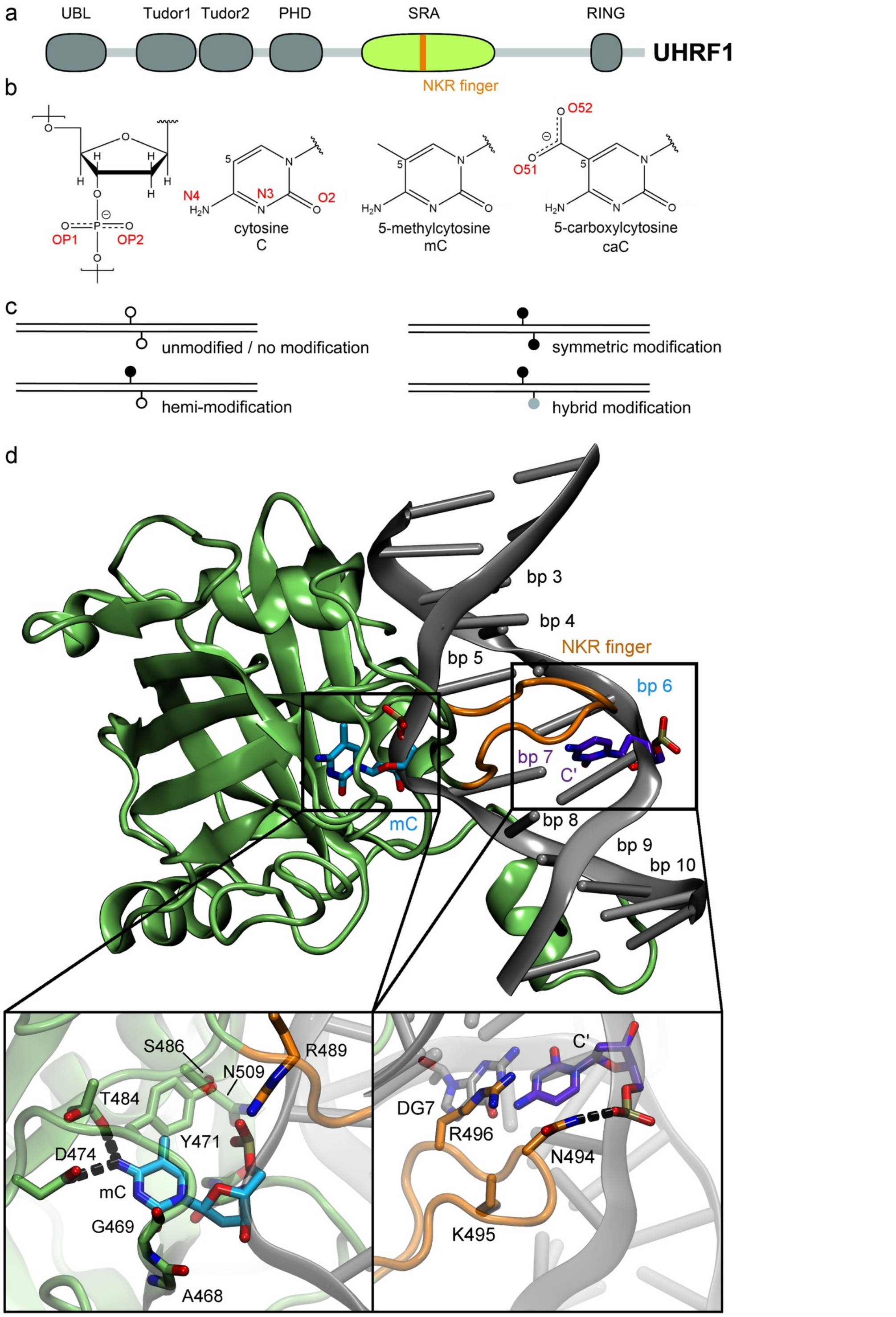


**Fig 1. Structure of the UHRF1—DNA complex.** (a) Schematic structure of UHRF1. The Tudor-like domains and the PHD-type zinc finger recognize the histone marks H3K9me2/3 and H3R2me0, respectively, while the SRA domain (in green, also referred to as YDG domain) is important for DNA binding. (b) Chemical structure and atom names of the modified DNA bases methylcytosine (mC) and carboxylcytosine (caC). (c) Schematic illustration of possible cytosine modification configurations on CpG dyads. (d) Representative molecular dynamics structure of the SRA domain of UHRF1 bound to hemi-methylated DNA. Insets show a magnification of the nucleotide binding pocket and NKR finger regions. DNA base pairs (bp) are numbered based on the strand binding the flipped-out base.

https://doi.org/10.1371/journal.pone.0229144.g001

methylated DNA, i.e. DNA harbouring 5-methylcytosine (mC) only on one strand. Upon binding of the methylated strand, UHRF1 recruits DNA methyltransferase 1 (DNMT1) for additional methylation of the second strand, yielding a symmetrically methylated CpG site [1–3]. This recruitment depends on specific histone ubiquitination, set by the RING domain of UHRF1 and recognized by a ubiquitin interaction motif of DNMT1 [4–6].

Besides mC, three other cytosine (C) modifications exist in mammalian cells, i.e. 5-hydroxymethylcytosine (hmC), 5-formylcytosine (fC), and 5-carboxylcytosine (caC) [7–9]. These variants are generated by the family of TET proteins through step-wise oxidation of mC and are discussed to be either intermediates in active DNA demethylation or independent epigenetic marks [10]. Their overall abundance in vivo is normally magnitudes lower than that of methylated sites [11], but the ratio increases under certain conditions. Higher hmC concentrations were observed in neuronal cells [12], while a study investigating breast and glioma tumour tissues found that a substantial portion of the samples exhibited increased caC levels [13]. Efforts to map mC, hmC, fC, and caC modifications in the genome showed that they accumulate at functionally distinct regions of transcription regulation [14–16]. One common conclusion of these studies was that methylation/demethylation of CpG sites is a highly dynamic and genome-wide process. In this light, low concentrations of some DNA modifications could represent a transient state in a high turnover process, while the accumulation at functionally diverse sites suggests that some variants might have a biological role beyond being demethylation intermediates. It has been demonstrated that several proteins recognize some oxidised variants with similar or even greater affinity than mC. The UHRF family member UHRF2, which features a highly similar domain architecture to UHRF1 [17, 18], is a reader with increased affinity for hmC in neuronal progenitor cells [19]. Other examples include SUVH5, which binds both mC and hmC with similar strength [20], while POL II, WT1 and TET3 specifically recognize caC [21–23]. It is currently unclear how frequent certain CpG modification patterns occur in vivo. DNA replication during S-phase will generally result in hemi-modified CpG sites. In case of mC, the subsequent restoration of the DNA modification to symmetry is well studied and described [24]. Nevertheless, the degree of persistent hemimethylation varies between cell types and genomic elements [25]. For hmC, fC, and caC, no maintenance pathways have been described so far. In vitro, TET proteins predominantly generate symmetric fC sites [26], whereas genomic mapping approaches suggest the existence of hmC and fC/caC in hemi-modified form [15, 27]. The occurrence of hybrid modifications with mC on one and an oxidised cytosine derivative on the other strand is also likely (Fig 1c).

Structural analysis revealed that the SRA domain of UHRF1 flips the methylated cytosine out of the DNA strand and envelopes it within its binding pocket. In addition, the protein binds to the DNA by inserting its thumb region into the minor groove and its NKR finger region into the major groove [2, 28, 29]. In a previous work, our groups showed by a combination of in vitro experiments and molecular dynamics (MD) simulations that UHRF1 binds hemi-modified hmC with similar affinity as hemi-mC [30]. Although subsequent studies revealed that UHRF1 binds hmC with lower affinity than mC, it still binds hmC with 1.3 to 3-fold higher affinity than unmodified C [19, 31, 32]. These results are in line with an unbiased mass spectrometry screen for epigenetic readers in embryonic stem cells, which demonstrated UHRF1 binds to all modified cytosines, but in particular to mC and hmC [19]. Experiments with UHRF1 and symmetrically modified mC sites, i.e. CpG sites in which both DNA strands feature methylcytosine, consistently show reduced binding affinity [1, 2, 28, 29]. This selectivity is commonly explained by a hydrogen bond between N494 at the tip of the NKR finger and the C' cytosine, i.e. the base that potentially carries the symmetric modification (Fig 1d) [29]. Throughout the manuscript we use a terminal apostrophe to mark bases on the distal DNA strand (e.g. C'). Bianchi et al. observed in a computational study that the presence of mC on both strands sterically impairs binding of the NKR finger of UHRF1 to the major groove [33]. In contrast to mC and hmC, the structural effects of fC and caC variants on UHRF1-DNA binding are still not well elucidated. Investigations of several SRA domains by Rajakumara et al. suggest a reduced affinity of UHRF1 towards hemi-hmC, -fC and–caC containing DNA [20]. Crystal structures of POL II and TDG, which exhibit specific activity towards caC, show that the caC carboxyl group participates in specific hydrogen bond networks, which are crucial for binding key recognition residues in the protein [21, 34].

It was recently shown that UHRF1 allosterically regulates its activity and binding properties through intramolecular conformational changes [35–38]. The formation of these extensive inter-domain interactions illustrates an inherent flexibility of UHRF1 and allows the protein to adapt to different substrates. As we already observed solid binding of UHRF1 to hemi-hmC, we sought to systematically analyse the binding behaviour of UHRF1 towards CpG sites containing C, mC, hmC, fC, and caC either in a hemi-, hybrid or symmetrically modified state. The highest binding affinities are observed for hemi-mC, symmetric caC, and the caC-mC' hybrid. To understand the differences in recognition of these modifications, we performed molecular dynamics simulations of mC- and caC-modified DNA in complex with the SRA domain of UHRF1 (see Fig 1d).

## Materials & methods

### Electrophoretic mobility shift assays (EMSAs)

Expression constructs for GFP-mUHRF1 and mUHRF2-GFP have been described previously [18, 39]. In general, protein purification and EMSAs were performed as reported in Spruijt et al. [19]. Briefly, a 2-fold serial dilution of protein (300 nM to 4.69 nM) in binding buffer (including 100 ng/μl BSA final concentration) was incubated with a 1:1 mixture of two fluorescently labelled 42 bp oligonucleotides (Eurofins Genomics) at a stable concentration of 250 nM each. After 30 min of incubation on ice, reactions were run over a 6% native PAGE in 0.5x TBE buffer (45 mM Tris-borate, 1 mM EDTA). ATTO647N-labelled DNA (“C$^{647}$") served as internal control and reference whereas ATTO550-labelled DNA carried one of the following cytosine variants at the central CG site: canonical C, mC, hmC, fC, or caC (“xC$^{550}$”). Fluorescent signal was detected with a Typhoon Trio+ scanner (GE Healthcare Life Sciences). Signal of bound and unbound fractions were quantified with ImageJ by plotting the mean grey values per lane and measuring the area under the selected peaks. Before quantitation, gel pictures were assigned random names to blind the experimenter during analysis. Box plots show $\frac{\text{ATTO550 bound fraction}}{\text{ATTO647 bound fraction}} \times \frac{\text{ATTO647 total signal}}{\text{ATTO 550 total signal}}$ with the C$^{550}$/C$^{647}$ experiment as control. All raw gel image scans with annotations are provided as S1 Fig.

### Microscale Thermophoresis (MST)

For MST, the SRA domain of mouse UHRF1 (residues 419–628) was cloned into a hexahistidine-tagged construct and protein was expressed in Escherichia coli BL21(DE3)-Gold cells (Stratagene). The purified SRA domain was labelled with a NT-647 dye using the Monolith NT™ His-Tag Labelling Kit RED-tris-NTA (NanoTemper Technologies) according to the manufacturer's instructions and 50 nM of the labelled protein was incubated for 20 min at room temperature with increasing concentrations of the corresponding DNA oligonucleotide (C-C', mC-C', caC-C', caC-caC', mC-caC') in PBS-T (0.05% Tween-20). The solutions were then aspirated into NT.115 Standard Treated Capillaries (NanoTemper Technologies) and placed into the Monolith NT.115 instrument (NanoTemper Technologies). Experiments were conducted with 60% LED power and 80% MST power. Obtained fluorescence signals were normalized ($F_{norm}$) and the change in $F_{norm}$ was plotted as a function of the concentration of the titrated binding partner using the MO. Affinity Analysis software version 2.3 (NanoTemper Technologies). For fluorescence normalization ($F_{norm} = F_{Hot}/F_{cold}$), the manual analysis mode was selected and cursors were set as follows: $F_{cold}$ = -1 to 0, $F_{hot}$ = 9 to 10 (see S2 Fig). Data of four to five independent measurements were analysed and means were fitted to obtain the respective $K_D$ values. More detailed information and additional experimental procedures can be found in S1 Text.

## Force field parameterization of modified cytosine bases

We generated parameters for the parmbsc1 force field [40] for both deoxy-5-methylcytosine (mC) and deoxy-5-carboxylcytosine (caC) using the mC structure and bonded parameters template from Lankas et al. [41], which was originally derived for parmbsc0 [42]. The atom type of the C3' atom was changed from CT to CE to adjust the template to parmbsc1. Fixed point atom charges were derived for both mC and caC following the procedure in ref. [43] using the R.E.D Dev webserver [44–48]. Atom types were assigned and final parameter files prepared using the programs antechamber and prepgen of the AmberTools17 package [49]. The final parameter files are provided in S1 File.

## Molecular dynamics simulations

Molecular dynamics simulations were performed with the Amber16/AmberTools17 software suite [49] using the Amber14SB force field for protein and parmbsc1 for nucleic acid parameters [40, 50]. All systems were based on the crystal structure of a mouse UHRF1 SRA domain bound to DNA featuring a single mC (PDB-ID: 3FDE). The same structure had been used in our previous work analysing the binding of 5-hydroxymethylcytosine [30] and featured the best resolution (1.41 Å) of published UHRF1 structures at the time of this study. Cytosine modifications were modelled and topologies prepared using leap (AmberTools). Each system was solvated in a box of TIP3P water [51] with a minimum face distance of 15 Å and 150 mM NaCl. A direct space cutoff of 12 Å was used for nonbonded potentials and PME summation was applied for electrostatic interactions. Energy minimization was performed until convergence to 0.01 kcal * mol$^{-1}$ * Å$^{-1}$ using the XMIN minimizer. Then, the volume of the solvent box was modified such that the density increased in 0.02 kg * m$^{3}$ steps and energy minimization was repeated for each step until a target density of 1.00 kg * m$^{3}$ was reached. For all molecular dynamics simulations hereafter, a time step of 1 fs and SHAKE [52] for bonds connected to hydrogens were used. The system was gradually heated from 0 to 300 K over 1.7 ns, applying a variation of the step-wise heatup protocol established within our group [53]. Within these steps, restraints of 2.39 kcal * mol$^{-1}$ * Å$^{-2}$ were applied to all heavy atoms until 20 K and on protein/DNA backbone atoms until 200 K. For 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. During the following simulations at 300 K, a slow coupling Berendsen thermostat with a coupling time of 10 ps was used in combination with a Berendsen barostat and a respective relaxation time of 5 ps. Backbone phosphates and oxygens of terminal DNA residues were harmonically restrained with a constant of 2.39 kcal * mol$^{-1}$ * Å$^{-2}$ while resetting target coordinates in 500 ps intervals. For all replicas, different initial velocities and random seeds for the Langevin thermostat were generated at the beginning of each step of the heatup protocol (i.e. for each temperature simulated). Each replicon was simulated for 200 ns, yielding a total simulation time of 1 μs per system (5 replicas). In two out of thirty simulations (caC-caC'_r2 and mC-caC'_r2), the DNA structure diverged notably from the others (RMSD > 4 Å; see S3 and S4 Figs). In the case of caC-caC'_r2, the distortion correlates with an interaction between the protein's free C-terminal helix and the DNA strand, bending it out of position, which is clearly an artefact due to the use of the isolated SRA domain. Therefore, and as it is in general difficult to determine whether such diverging trajectories show a rare but physically relevant conformational change or a simulation artefact, we excluded these two replicas from our analysis. The remaining simulations showed stable RMSD curves after about 20 ns. To allow for proper equilibration and to minimize any bias towards the initial structure, we extracted only the last 100 ns of each trajectory and afterwards merged the trajectories of all five replicas into a single system-specific trajectory that was used for all computational analyses.

Trajectory post-processing was performed with CPPTRAJ [54] version 17.00 unless otherwise indicated. Salt bridges were calculated using the "nativecontacts" command and a cutoff of 5 Å, saving both native and non-native time series and selecting interactions with opposite formal charges involving Arg, Lys, Glu, Asp and nucleotide residues. Hydrogen bonds were extracted using the "hbond" command, a cutoff distance of 4 Å and an angle cutoff of 120°. CPPTRAJ outputs were merged and converted into networks using our analysis tools AIFGen and CONAN (manuscript in preparation). Root mean square deviation (RMSD) and root mean square fluctuation (RMSF) calculations were performed for non-hydrogen atoms using the CPPTRAJ "rmsd" and "atomicfluct" commands after aligning each simulation frame to the protein's Cα atoms without the terminal regions (residues 432 to 586). For RMSD, the reference frame was the simulation's initial structure, while for RMSF the protein was aligned to its simulation average. DNA major and minor groove widths were calculated using the method of El Hassan and Calladine [55] as implemented in the "nastruct" command in CPPTRAJ (version 18.01). Figures of protein and DNA structures were prepared using VMD 1.9.3 [56]. Plots and supporting calculations (e.g. gaussian kernel estimates) were generated with matplotlib 2.0.0 [57].

# Results

## Experimental investigation of the binding behaviour of UHRF1 towards different cytosine variants

For systematic analysis of the binding specificities of UHRF1 towards the five known cytosine variants, we performed EMSA experiments with full-length UHRF1 in complex with 42 bp oligonucleotides harbouring C, mC, hmC, fC, or caC at a central CpG site (Fig 2a). To correct for general DNA binding affinity, two DNA fragments were used in direct competition in each EMSA experiment: A 647-labeled unmodified oligonucleotide and a 550-labeled oligonucleotide carrying the modification of interest in either hemi-modified (xC-C') or symmetric (xC-xC') state. 647-labeled unmodified DNA is used as internal control and reference for quantification. This allows direct comparison of UHRF1 binding affinity to all modifications without the need for pair-wise competition assays. Generally, EMSAs showed binding of UHRF1 to all studied DNA variants (example gel pictures are shown in Fig 2b). However, quantitation of the shifted fractions reveals a 1.5-fold preference for hemi-mC and a statistically significant 2-fold preference for symmetric caC (Fig 2c). All other modification variants, including hemi-caC, were bound with comparable strength to unmodified DNA. Similarly, we observed a 2-fold preference of UHRF2 for symmetric caC (S5 Fig).

Upon UHRF1 binding, the melting temperature of CpG-containing DNA is slightly reduced compared to its unbound state or a non-CpG-control, indicating a destabilization of the DNA duplex (S6a Fig). Complementary to our EMSA results, the SRA domain of UHRF1 substantially shifted the melting temperature of symmetrically carboxylated DNA to lower temperatures, whereas a weaker shift was observed for unmodified and hemi-methylated DNA (S6 Fig). To rule out that the thermal shift observed for symmetrically carboxylated DNA is due to different binding stoichiometries, we examined DNA-protein complex formation by

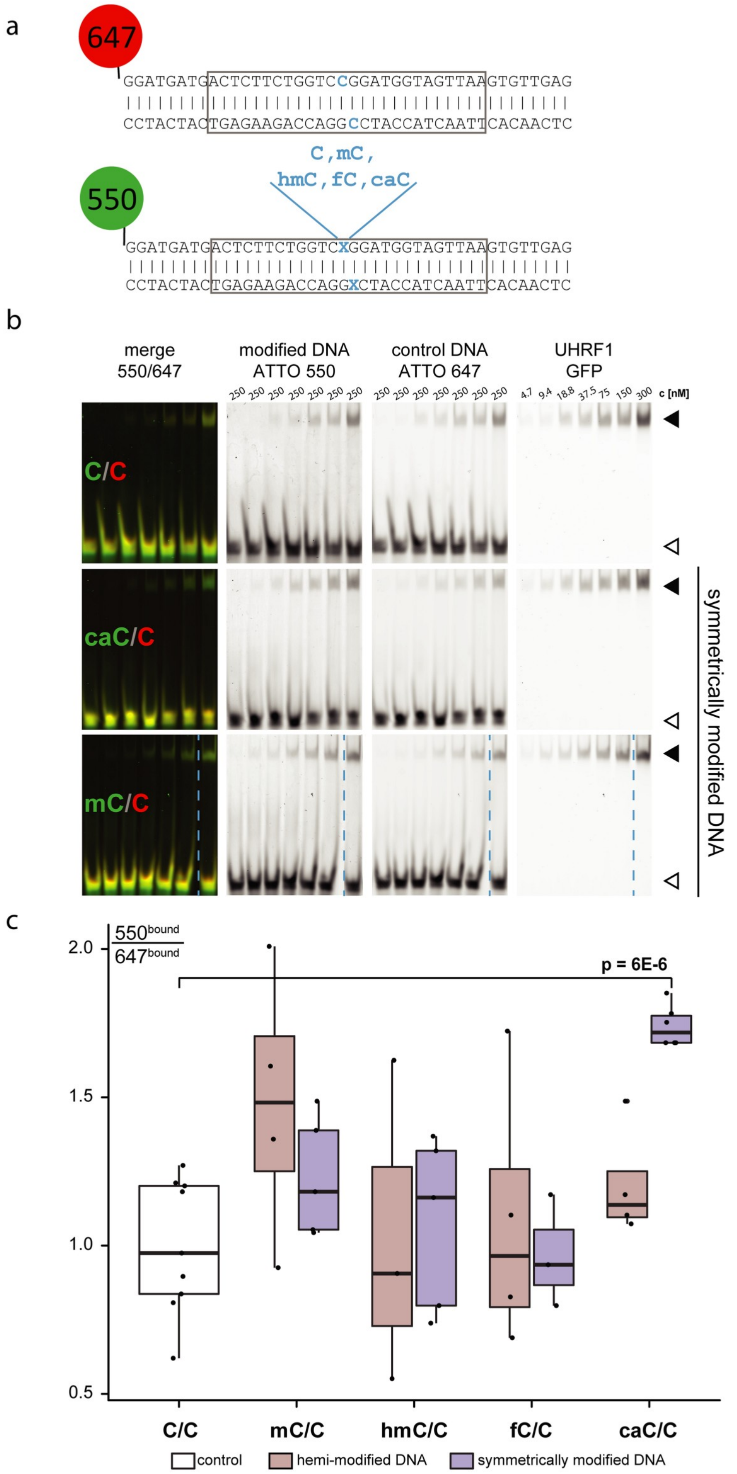


**Fig 2. Binding of UHRF1 to differentially modified CpG sites.** (a) DNA used in EMSA experiments. The 550-labelled DNA contains a central CG site harbouring different cytosine modifications: Unmodified C, mC, hmC, fC, or caC. The modification resides either on one strand (hemi-modification) or on both strands (symmetric modification). The 647-labelled oligonucleotide is always unmodified and serves as an internal control and reference. Grey boxes indicate sequences of the shorter DNA fragments used in Fig 3. (b) Representative images of EMSAs. Fluorescently labelled DNA oligonucleotides of 42 bp are incubated with GFP-UHRF1 at increasing protein concentrations. Black arrowheads indicate the DNA-protein complex (bound fraction); white arrowheads show free DNA. Dashed blue lines indicate empty gel lanes that have been removed for presentation purposes. (c) Quantitation of the bound fraction of symmetric and hemi-modified DNA incubated with wild type UHRF1, p value of two-tailed student's t-test.

https://doi.org/10.1371/journal.pone.0229144.g002

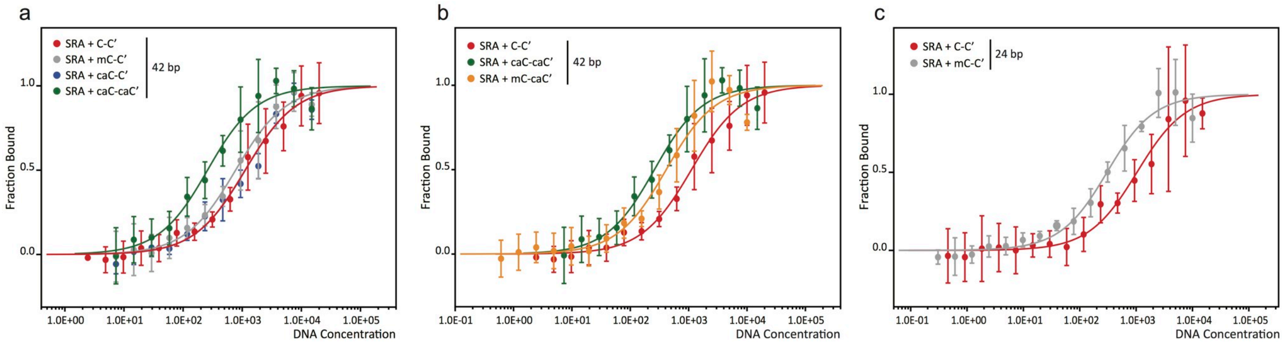


**Fig 3. Microscale Thermophoresis experiments of UHRF1-SRA bound to DNA with modified CpG sites.** (a,b) Dissociation constants of UHRF1 bound to a 42 bp DNA oligonucleotide: 1.10±0.15 μM for C-C', 0.75±0.11 μM for mC-C', 1.10±0.29 μM for caC-C', 0.23±0.05 μM for caC-caC', and 0.39±0.11 μM for mC-caC'. (c) Dissociation constants of UHRF1 bound to a 24 bp oligonucleotide; 1.01±0.20 μM for C-C' and 0.28±0.06 μM for mC-C'. Curves show the fitted average values of 4–5 independent experiments.

https://doi.org/10.1371/journal.pone.0229144.g003

size-exclusion chromatography. Binding of the SRA domain to the modified DNA oligonucleotides led to a comparable shift in retention time for all modifications tested (S7 Fig), indicating a uniform binding stoichiometry for UHRF1 independent of the DNA's modification state.

To better characterize the binding of UHRF1 to hemi-mC, hemi-caC and symmetric caC, we determined the respective dissociation constants ($K_D$) with Microscale Thermophoresis [58] (MST) experiments (Fig 3a). We observed slightly stronger binding of hemi-mC ($K_D$ = 0.75±0.11 μM vs. 1.10±0.15 μM for unmodified DNA) and considerably enhanced binding of symmetric caC ($K_D$ = 0.23±0.05 μM). In agreement with the EMSA results, hemi-carboxylated DNA ($K_D$ = 1.10±0.29 μM) is bound with similar affinity as unmodified DNA. Taken together, we performed three independent experimental assays, i.e. EMSAs, melting temperature analysis and MST, which consistently confirm a binding preference of UHRF1 towards symmetric caC.

Additionally, as the enzymatic reactions involved in generation of mC and caC modifications suggest the potential existence of hybrid mC-caC' sites, we determined the $K_D$ of the SRA domain of UHRF1 and a mC-caC' oligonucleotide and observed binding comparable to symmetric caC ($K_D$ = 0.39±0.11 μM vs. 0.23±0.05 μM). In summary, UHRF1 exhibits a binding preference for caC modifications opposite of mC or caC, but not C.

Since the difference in $K_D$ between unmodified and hemi-methylated DNA was smaller than expected from the literature [1, 32, 36, 59, 60], we repeated the MST experiments with shorter DNA oligonucleotides of 24 bp to reduce the number of unspecific binding sites (Fig 3c). With this new setup we observed a 3.6-fold preference of the SRA domain of UHRF1 towards hemi-methylated CpG sites ($K_D$ = 0.28±0.06 μM for mC-C' vs. 1.01±0.20 μM for C-C'). This ratio is in very good agreement with data by Greiner et al. [60] and Zhou et al. [32] (Table 1), who reported a 3.5 or 3.4-fold smaller $K_D$ for hemi-methylated CpGs for a 12 bp oligonucleotide, respectively, compared to unmodified DNA. Generally, caution is advised when published $K_D$ values of UHRF1 and differentially modified DNA are compared, since applied methods, DNA substrates and protein constructs used vary greatly among studies, resulting in a broad range of $K_D$ values from 1.8 nM to 9.23 μM (Table 1). Nonetheless, previous studies and our results not only demonstrate the sensitivity of UHRF1 to different types of cytosine modification, but also the dependency of measured binding affinities on modification density, i.e. the number of DNA modifications compared to unmodified DNA stretches.

**Table 1. Published $K_D$ values for UHRF1 and DNA with differentially modified CpG sites.**

| Citation | Method | Affinity | DNA substrate | protein construct |
|---|---|---|---|---|
| Bostick, M. et al., 2007, 10.1126/science.1147939 | EMSA | $K_D$(mC-C’) = 1.8 nM | 39mer, 13 modification sites | murine SRA |
| | | $K_D$(mC-mC’) = 12.1 nM | | |
| Fang, J., 2016, 10.1038/ncomms11197 | Fluorescence Polarization | $K_D$(UHRF1) = 0.35 μM | 12mer, 1 modification site | human UHRF1, different constructs with mC-C’ |
| | | $K_D$(SRA) = 9.23 μM | | |
| | | $K_D$(SRA+Spacer[a]) = 0.49 μM | | |
| Greiner, V. J., 2015, 10.1021/acs.biochem.5b00419 | FRET | $K_D$(mC-C’) = 0.08 μM | 12mer, 1 modification site | human SRA |
| | | $K_D$(mC-mC’) = 0.25 μM | | |
| | | $K_D$(C-C‘) = 0.28 μM | | |
| | | $K_D$(T-C‘) = 0.55 μM | | |
| Qian, C., 2008, 10.1074/jbc.C800169200 | Fluorescence Polarization | $K_D$(mC-C’) = 0.2 μM | 13mer, 1 modification site | human SRA |
| Zhou, T., 2014, 10.1016/j.molcel.2014.04.003 | Fluorescence Polarization | $K_D$(C-C’) = 8.61 μM | 12mer, 1 modification site | human SRA |
| | | $K_D$(mC-C’) = 2.56 μM | | |
| | | $K_D$(hmC-hmC’) = 7.97 μM | | |
| Schneider, Trummer et al., 2019 | MST | $K_D$(C-C’) = 1.01 μM | 24mer, 1 modification site | murine SRA |
| | | $K_D$(mC-C’) = 0.28 μM | | |
| Schneider, Trummer et al., 2019 | MST | $K_D$(C-C’) = 1.10 μM | 42mer, 1 modification site | murine SRA |
| | | $K_D$(mC-C’) = 0.75 μM | | |
| | | $K_D$(caC-C’) = 1.10 μM | | |
| | | $K_D$(caC-caC’) = 0.23 μM | | |
| | | $K_D$(mC-caC’) = 0.39 μM | | |

[a] Spacer: amino acid stretch C-terminal of SRA domain

https://doi.org/10.1371/journal.pone.0229144.t001

## Molecular dynamics simulations of the UHRF1-SRA domain bound to CpG sites with mC and caC modifications

For methylated CpG sites, UHRF1 binds stronger to mC-C’ modified DNA than to the symmetric modification variant mC-mC’ (Table 1) [1, 60]. As discussed above, in our experiments the opposite was observed for caC modifications, as caC-caC’ DNA was preferred over caC-C’. To understand this behaviour, we performed MD simulations of UHRF1-DNA complexes with different nucleotide modifications, i.e. hemi-modified and symmetrically modified mC and caC as well as the hybrid modification variants mC-caC’ and caC-mC’. As simulation of the full binding process for all variants was not feasible due to the high complexity and computational cost of such simulations, we focused on studying the complex with the flipped-out modified base bound in the protein’s binding pocket, based on the experimental structure of mC-C’ bound to UHRF1 (PDB-ID: 3FDE). Various experimental data indicate that this is the most relevant state for recognition: Fluorescence kinetics experiments [61] showed that the stability of the DNA flipped state is correlated to the lifetime of the flipped state bound to protein. Regarding flipping propensity, previous simulation studies showed no substantial intrinsic difference between mC and caC [62] and furthermore, NMR experiments of Dickerson–Drew dodecamers showed that both mC and caC bases were slightly less likely to flip compared to unmodified cytosines [63]. Finally, in a study of another base-flipping protein, bacterial cytosine-5-methyltransferase, it was found that specific protein-base interactions were responsible for facilitating and stabilizing the flipped out state [64]. We chose to simulate the second potentially modified base on the distal strand in the flipped-in state, motivated by the following observations: First, stable flipping of the distal base has only been observed for proteins which can bind in a 2:1 protein-DNA ratio to the same CpG site, like UHRF2 or SUVH5, but not UHRF1 [2, 28, 29, 32, 65]. Second, the NKR finger can recognize modifications on the distal strand directly, as demonstrated by the crystal structure contacts of N494 [2, 29, 66] and third, it was observed that a single mutation of this residue abolishes the selectivity of UHRF1 between mC-C' and mC-mC' [29]. Finally, computational studies reported that the first stable intermediate in the flipping process requires a flip angle of at least 50° [62, 67]. It is difficult to imagine how direct interactions of the NKR finger could be sustained with the modified base in this position. For these reasons, we consider the complex conformation with a flipped-out pocket bound base and a flipped-in base on the distal DNA strand as the most relevant for explaining the selectivity of UHRF1.

Therefore, we did not aim at the simulation and analysis of the binding process itself and its related binding affinities, but rather at identifying similarities and differences in the binding modes of the different DNA modifications, i.e. which regions of the protein are likely to sense the chemical differences of these modification types and how this influences their interaction patterns. In contrast to mC, the caC modification contains an additional carboxyl group, which can form additional salt bridges and hydrogen bonds. Thus, we analysed whether this difference in interaction capacity could affect the polar interaction network and the local conformations of the binding pocket and NKR finger regions, which are in direct contact with the two modification sites.

Analysis of mC and caC recognition in the UHRF1-SRA nucleotide binding pocket. In Fig 4 we provide the interaction networks of the flipped base in the nucleotide binding pocket as derived from our MD simulations. Nodes represent residues of the protein and atoms of the modified DNA bases (see naming conventions in Fig 1b), while edges show the average number of hydrogen bonds (black lines) and salt bridges (red lines) between two nodes during the simulation. The canonical binding mode of mC-C' (Fig 4a) is characterized by strong hydrogen bonds between the mC atom N4 to T484 and D474 (1.84 and 1.04 hydrogen bonds on average per analysed simulation frame, respectively) and between the pyrimidine oxygen O2 and G470 and A468 (1.0 and 0.98 hydrogen bonds on average). Thus, the base is effectively locked at these two positions with the N4 and O2 atoms acting as handles. In addition, the mC backbone atom OP1 (phosphate oxygen 1) forms one stable hydrogen bond with G453 and the adjacent OP2 forms approximately two (1.86) salt bridges with R489, the latter being located at the beginning of the NKR finger. Overall, the binding pocket of the mC-C' simulation shows a regular and stable polar interaction pattern. This pattern is nearly identical to the one observed in the mC-mC' and mC-caC' simulations (Fig 4c and 4e), indicating that modifications on the distal strand have little effect on the conformation and interactions of the nucleotide binding pocket containing flipped mC.

Analysis of the binding mode of the hemi-modified caC-C' system (Fig 4b) shows that this modification leads to a very different interaction pattern: The previously observed hydrogen bonds of the nucleotide N4 atom are substantially weakened (-1.87 hydrogen bonds), while interactions of O2 are dispersed from two to three amino acids (-0.2 hydrogen bonds total). Although several hydrogen bond donors such as S486, N509, and the backbone atoms of I454 and G453 are available in the binding pocket, the carboxyl atoms O51 and O52 of caC predominantly interact with R489, forming very strong interactions (1.92 salt bridges on average) with this residue. This interaction pattern is unexpected, since the caC modification is located within the binding site, whereas R489 is located at its edge, usually interacting only with the DNA backbone. This may cause a force pulling the base out of position and could explain the weaker hydrogen bonds formed by the base's N4 nitrogen. The NKR finger region consisting of residues 488 to 502 is a flexible loop important for DNA binding with residues N494, K495,

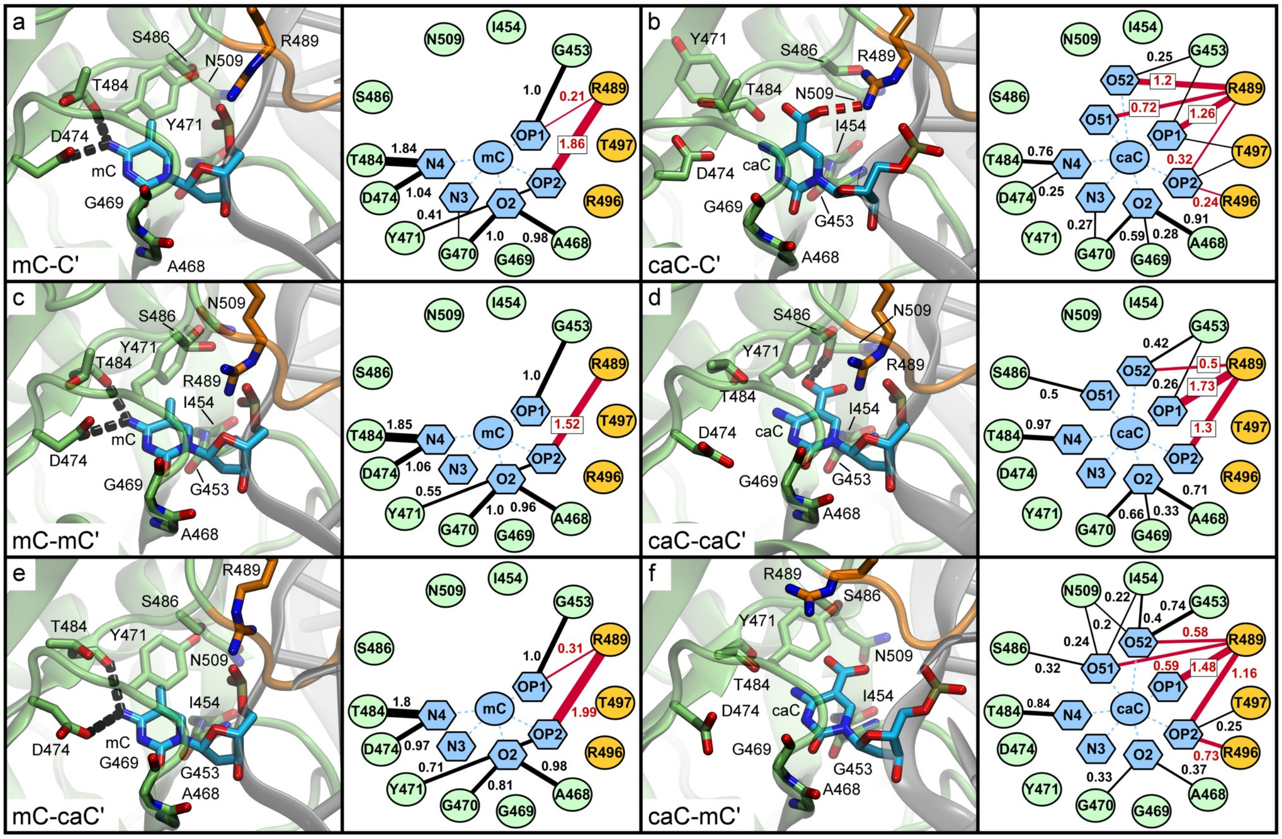


**Fig 4. Interaction networks of the nucleotide binding pocket based on molecular dynamics simulations of UHRF1-SRA.** Structures show representative conformations of the flipped-out modified DNA base within the binding pocket as observed during MD simulations. To the right of each structure a corresponding network of hydrogen bonds (black lines) and salt bridges (red lines) averaged over the course of the simulation is shown. Numbers next to edges show the average number of interactions per time frame. Edges representing interactions occurring in $\leq$ 15% of simulation time are omitted for clarity. For node pairs featuring both hydrogen bonds and salt bridges, only salt bridges are displayed.

https://doi.org/10.1371/journal.pone.0229144.g004

and R496 at its tip. Observing that R489 is involved directly in interactions with the carboxyl oxygens establishes a direct link between the flipped-out base and the NKR finger, which predominantly interacts with the distal DNA strand. The interaction pattern of the caC-caC' system (Fig 4d) is consistent with this observation. In this system, the caC N4 and O2 atoms show an overall similar interaction pattern to the hemi-modified variant. However, distinct differences are seen in the interaction with R489: The salt bridges between the carboxyl oxygens and R489 are much weaker (only 0.5), whereas the residue forms very strong interactions (3.03) with the backbone atoms OP1 and OP2 (+ 0.96 compared to mC-C'). To compensate for the weaker R489 interactions, O51 and O52 form fluctuating weak ($\leq$ 0.5) hydrogen bonds with S486 and G453 in the binding pocket. The caC-mC' system (Fig 4f) shows a mixture between these patterns, as R489 establishes 1.17 salt bridges to O51 and O52 of caC and 2.64 salt bridges to the caC backbone. The hydrogen bonds of the carboxyl oxygens are more dispersed compared to the caC-caC' system, interacting weakly ($<$ 0.5) with S486, N509, and I454 and moderately strong (0.74) with G453. In turn, O2 establishes only 0.7 hydrogen bonds to G470, G469, and A468, which is 1 less than in caC-caC'. The differences we observed in the binding modes of caC-C', caC-caC' and caC-mC' indicate that the caC carboxyl oxygens have several possible interaction partners in the nucleotide binding pocket and the interaction networks are more heterogenous compared to bound mC. In addition to interactions within the binding pocket (S486, N509, I454, G453), caC oxygens O51/O52 can establish alternative interactions outside of the main pocket, particularly with the NKR finger residue R489. In combination with our observation that the overall interaction pattern of R489 is strongly dependent on the xC' modification on the distal strand, this suggests that the binding mode is influenced by the NKR finger, which senses that modification.

Another notable difference between the interaction networks is the hydrogen bond of the Y471 hydroxyl atom to the OP2 atom of the modified base, which is absent in the carboxylated variants (Fig 4). As Y471 has been described previously to form a hydrophobic cage, closing like a lid over the modified base [2], we analysed whether the distances between the tyrosine and pyrimidine rings were influenced by the nucleotide modification. S8 Fig shows that for both mC-C' and mC-mC' the distances cluster in two close narrow peaks with tyrosine being stabilized in its position, while for the carboxylated variants the distances fluctuate between multiple distinct conformations due to changes in the nucleotide binding mode. The distance histograms tend to differ more between replicas than during a single simulation, indicating that Y471 flips between distinct conformations with characteristic transition times roughly in the ~ 10–100 ns range or longer. Interestingly, the distribution of mC-caC' shows a similar pattern to the other methylated variants, but an additional small peak at 8–9 Å, indicating a partial destabilization of the Y471 lid. In summary, carboxylation of the flipped base leads to a different local conformation of the binding pocket compared to methylation. While during the simulations of complexes featuring a flipped mC base very similar binding modes were observed, strong differences were found in the binding modes of complexes containing a flipped caC depending on the xC' modification on the distal strand. These differences suggest potential conformational long-range correlations between the binding pocket and the NKR finger, in particular R489, which can interact directly with the carboxyl modification of the flipped-out base.

**Analysis of mC and caC recognition on the distal DNA strand by the UHRF1-SRA NKR finger.** Our observations so far indicated that the NKR finger could play an important role for UHRF1 to differentiate between carboxylated and methylated CpG sites. As for the binding pocket, we analysed the interaction networks between the finger residues and the second modification site on the distal DNA strand (Fig 5). In the native binding conformation represented by the mC-C' simulation (Fig 5a), N494 forms 0.76 hydrogen bonds with the OP2 atom of the unmodified DNA base backbone. This interaction has been described previously as one of the key features for differentiating between hemi-methylated and symmetrically methylated DNA [29, 33]. This is in line with our simulation of mC-mC' in which this interaction is not observed (Fig 5c), as N494 is pushed away from its native position by steric repulsion of the additional methyl group. Interestingly, a similar trend is observed for caC-C' (Fig 5b), for which the N494-OP2 hydrogen bond is also much weaker (0.13) compared to mC-C' despite the lack of any modification on the distal DNA strand. This indicates a shift in the conformation of the NKR finger similar to the mC-mC' system, only that in this case the cause is not the modified base on the distal strand, but it appears that the shift might be mediated by the conformations of R489 as described above. Investigating the interaction pattern of the caC-caC' system (Fig 5d), we observed additional strong salt bridges (3.32) between R496 and the caC' O51/O52 atoms. No interactions are formed between the modified base and N494, likely related to steric repulsion similar to the methyl group as in mC-mC'. The interaction pattern of mC-caC' (Fig 5e) is similar to caC-caC', but with slightly weaker individual interactions as R496 forms only 1.74 salt bridges to the carboxyl oxygens (- 1.58), albeit with support from spurious interactions of K495 (0.61). In contrast, the interaction pattern of caC-mC' (Fig 5f) resembles mC-mC' with an additional loss of 0.51 hydrogen bonds between N494 and the N4 base atom of mC', with nearly no polar interactions remaining between the NKR finger and the modified base.

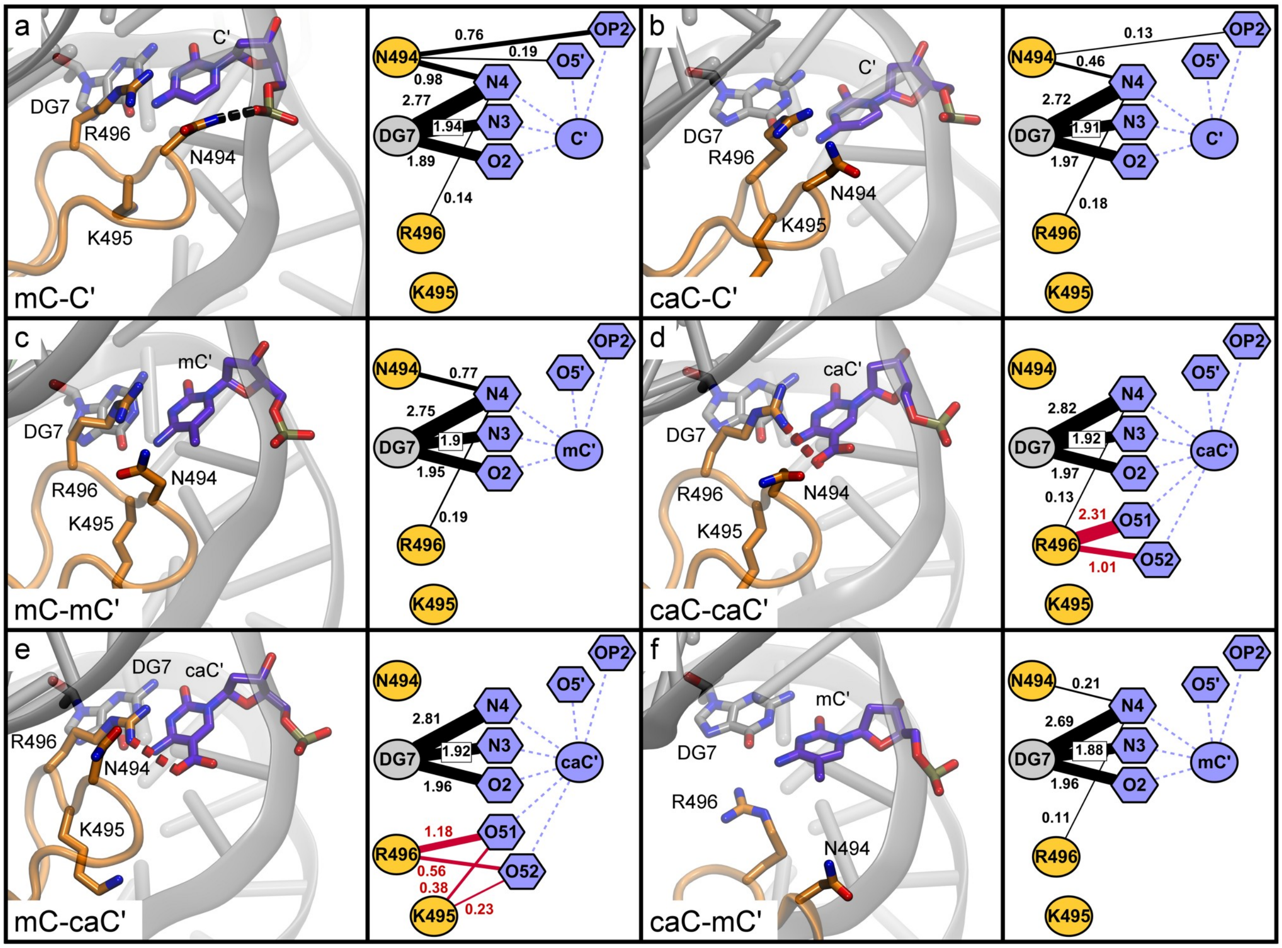


**Fig 5. Interaction networks of the NKR finger based on molecular dynamics simulations of UHRF1-SRA.** Structures show representative conformations of the NKR finger close to the distal (symmetrical) DNA modification site as observed during the MD simulations. To the right of each structure a corresponding network of hydrogen bonds (black lines) and salt bridges (red lines) over the course of the simulation is shown. Numbers next to edges show the average number of interactions per time frame. Edges representing interactions occurring in ≤ 10% of simulation time are omitted for clarity. For node pairs featuring both hydrogen bonds and salt bridges, only salt bridges are displayed.

https://doi.org/10.1371/journal.pone.0229144.g005

R496 is generally a strong interaction partner for the DNA in all simulated systems, partaking in hydrogen bonds with adjacent bases and stacking interactions with the modified base. The interactions of the carboxyl group seem to modulate this role, either directly through salt bridges or by influencing stacking, although stacking effects are not quantifiable using classical force fields. As our analyses showed that only mC-C' retained the native interaction pattern of the NKR finger, we were interested in whether there was any effect on the flexibility of the finger. To quantify this, we compared the Root Mean Square Fluctuation (RMSF) for all protein residues (Fig 6a). Overall, very similar residue flexibility is observed for most regions of the protein independent of DNA modifications. Only two regions show substantial differences: The first is located in the region between residues 468 and 475, which corresponds to the conformational flexibility of Y471 discussed above. The second region featuring pronounced differences is located between residues 488 and 502 forming the NKR finger (Fig 6b). Although the NKR finger shows a different conformation in the mC-mC' simulation, the flexibility of the finger is comparable to the mC-C' reference system. In contrast, for the caC-C', caC-caC', and mC-caC' systems, the finger shows increased flexibility with a slightly different pattern:

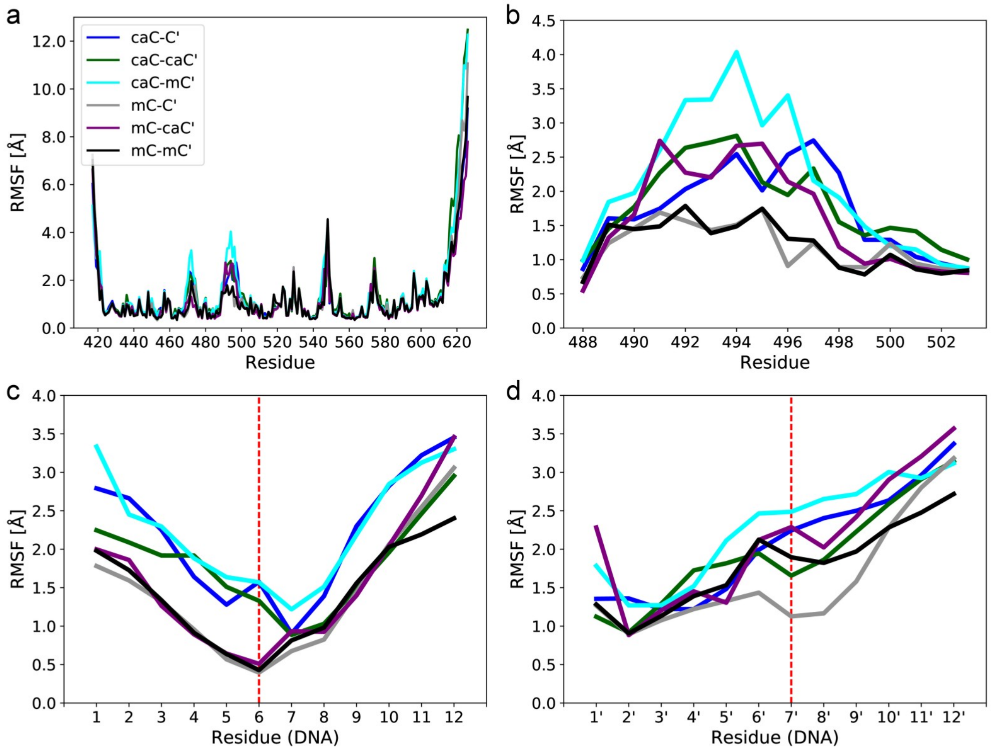


**Fig 6. Root Mean Square Fluctuation (RMSF) of protein and DNA regions in molecular dynamics trajectories of UHRF1-SRA.** (a) Full protein. (b) NKR finger. (c) DNA strand containing the flipped xC base bound by the protein. (d) Distal DNA strand containing the modified xC' base. Red dashed lines show the xC/xC' modification sites.

https://doi.org/10.1371/journal.pone.0229144.g006

The hemi-modified variant being more flexible in the 495–499 region and both the caC-caC' and mC-caC' variants more flexible between residues 490 and 494. Finally, the largest finger flexibility of all systems is observed for caC-mC', in line with the previously observed loss of interactions of the NKR finger.

UHRF1 encloses the flipped base by inserting a thumb into the minor groove and the NKR finger into the major groove of the DNA strands. Having observed differences in interaction pattern and flexibility of the NKR finger depending on the CpG modification pattern, we asked how the DNA structure around the modified sites was affected. Fig 6c shows that overall flexibility of the bound strand increases if caC is in the binding pocket, including particularly strong differences at the flipped xC base in position 6. For the distal strand, flexibility compared to mC-C' increases in all systems around the modified base 7' (Fig 6d), likely reflecting the loss of the stabilizing hydrogen bond between N494 and the DNA backbone. For a more detailed analysis, we examined how the modified bases affected the minor and major grooves, as they are strongly influenced by shifts in the DNA backbone. A small but consistent increase of minor groove width by about 1–2 Å is observed between base pairs 3 to 5 in all simulations containing caC in the binding pocket, while widths decrease by roughly the same amount between base pairs 7 and 9 (S9 Fig; locations of base pairs are shown in Fig 1d). The major groove follows a similar but weaker trend due to the large variances within replicas (S10 Fig). Although individual effects are small, their consistency and anti-symmetry with respect to the modified bases 6 and 7' is notable. Therefore, the flipped base appears to be important for the local flexibility of the DNA backbone, which is more rigid for mC and more flexible for caC. This could potentially contribute to the increased flexibility of NKR finger residues, particularly R489, which is in a prime position to sense distortions due to its strong salt bridges with the phosphate backbone of the flipped base. These observations agree with our interaction network analyses, showing that binding of a flipped caC base leads to conformational rearrangements including the DNA strands in locations close to the modification sites.

In summary, our simulations reveal that all DNA modifications investigated lead to differences in the conformation and binding pattern of the nucleotide binding pocket and NKR finger compared to the native conformation of the mC-C' system. Interestingly, in the hemi-carboxylated variant caC-C', local conformational changes in the binding pocket are transmitted to the NKR finger via R489, which in turn becomes more flexible and thus compromises the essential N494 hydrogen bond to the C' backbone on the distal strand [29]. The symmetrically carboxylated variant caC-caC' also shows increased NKR finger flexibility, but different interaction patterns, particularly for R489 and R496. The latter forms strong salt bridges with the caC' modified base, possibly compensating for the loss of the N494 hydrogen bond. This is in strong contrast to the recognition of hemi- and symmetrically methylated CpG sites, which show much smaller differences. Our additional analysis of the hybrid modification variant mC-caC' suggests that the NKR finger can recognize and interact with the caC' modification without large changes in the binding pocket containing a flipped mC. In the opposite case of caC-mC', a heterogeneous binding pocket conformation is met with an almost complete loss of NKR finger interactions with the mC' base. Based on this simulation data, we formulate the hypothesis that UHRF1 binding of a flipped-out caC base leads to conformational changes in the protein, which can propagate to and induce shifts in the protein's NKR finger and the DNA backbone. In turn, modification of the distal DNA strand can influence the overall binding mode via steric repulsion or attractive interactions with the NKR finger, coupling recognition of both modification sites.

## Discussion

The role of UHRF1 as a specific hemi-mC reader is well established [1, 3]. Reported dissociation constants range from 1.8 nM to 9.23 μM depending on the protein construct and DNA substrate [1, 32, 36, 59, 60] (Table 1). Here, we use a relatively long DNA fragment (42 bp) with a single modified CpG site, whereas other studies have used either oligonucleotides with multiple methylated sites [1] or shorter DNA fragments with one modification site [29, 32]. We observe a relatively low preference of hemi-methylated over unmodified DNA compared to published data [1, 19, 29, 32], which we explain by the lower density of methylated sites in our experiments. To verify this relation, we also measured binding of a shorter DNA fragment which increased the affinity of UHRF1 for hemi-mC to the order of what has been reported in literature [32, 60]. A possible explanation can be given by the proposed "sliding" mechanism of UHRF1 [60, 61, 68, 69]: In this model, fast unspecific binding occurs between the protein and DNA, followed by a sliding "scan" for a modified base. Thus, the relative differences in apparent binding affinities would decrease with the length of the DNA fragments, which corresponds to our observations. In three independent assays, we observe that UHRF1 prefers binding symmetrically carboxylated CpG sites over the hemi-carboxylated variant, which is the opposite behaviour as observed for methylcytosine. Interestingly, we also measure increased affinity of UHRF1 towards hybrid mC-caC' sites. To find a possible explanation for the underlying molecular mechanisms of these differences, we performed MD simulations of the UHRF1-SRA domain in complex with hemi-, hybrid, and symmetrically modified DNA based on the crystal structure of mC-C', which features the flipped-out base in the protein's binding pocket and the second potentially modified base on the distal strand in the flipped-in state. As discussed in the results section, we preferred this approach over simulating the entire flipping process.

Our simulations revealed substantial differences in the conformations and binding patterns of the nucleotide binding pocket and the NKR finger between caC and mC modifications. If caC is bound in the binding pocket, these two regions appear to be coupled and able to influence each other in a more pronounced manner than for mC. In the caC-C' system, this coupling leads to reduced hydrogen bonding between N494 and the DNA backbone, which is an essential interaction for binding [29]. The same interaction is interrupted by steric repulsion when mC' and caC' modifications are present on the distal strand, sterically pushing the NKR finger out of its native binding position. The simulations provide no indication that the mC' modification could be beneficial to overall binding, but the caC' modification forms stable salt bridges to the NKR finger, which might compensate for the loss of the N494-DNA hydrogen bond. Thus, the caC' oxygens push the NKR finger away from its hydrogen bond with the DNA backbone and at the same time offer salt bridges to bind the finger in its new position. In this light, we propose that the carboxyl group of both, the caC and caC' bases, has a strong influence on their local interaction network partners in UHRF1, leading to conformational changes in which R489, N494, and R496 play key roles in differentiating DNA modifications. Other proteins are already known to recognize caC' modifications using finger regions: TET3, one of the three dioxygenases that generate hmC, fC, and caC, was also shown to specifically bind symmetrically carboxylated CpG sites with a finger-like structure containing a NRRT sequence [23]. Comparing the NKRT sequence of UHRF1 to the NRRT sequence of TET3, it is intriguing to speculate that such a flexible stretch of basic amino acids facilitates the binding of distant carboxyl groups.

The biological role of UHRF1 binding to symmetrically carboxylated DNA remains to be determined, considering the low abundance of this modification in cells. For this reason, it is likely that the majority of UHRF1 in a proliferating cell population interacts with hemi-methylated CpG sites, but a certain fraction may encounter and bind mC-caC' and caC-caC' depending on the cell type and cell cycle phase. Carboxylcytosine has been suggested to be an intermediate of active DNA demethylation and is detected at gene regulatory elements and promoters of actively transcribed genes, indicating dynamic DNA methylation turnover [14–16]. Several DNA repair mechanisms have been associated with this demethylation [70–72], most prominently removal of fC and caC by TDG and the base excision repair pathway [8, 73–75]. Interestingly, both UHRF1 and UHRF2 have been shown to play a role in DNA damage response [76–78]. Additionally, the bona fide UHRF1 interaction partner DNMT1 has been described to change its genomic localization upon oxidative stress [79, 80]. Furthermore, besides being demethylation intermediates, fC and caC are thought to influence DNA replication and genome stability [81, 82]. By transiently pausing RNA polymerases, fC and caC may lead to precise fine-tuning of gene expression [21]. Accordingly, the binding of UHRF1 to caC as demonstrated in our study could also represent a way of locus-specific gene expression regulation in addition to its well-established role in recognizing hemi-mC sites and initiating DNA maintenance methylation. Last but not least, UHRF1 has recently been described as a regulator of bivalent promoters and an interactor of SETD1A [83]. Interestingly, both functions have been attributed to TET proteins as well [84, 85]. This raises the intriguing possibility that UHRF1 integrates several epigenetic marks at bivalent domains and that caC, generated by TET proteins, is one of these marks involved in maintenance of the bivalent state. However, further work is needed to determine whether and where exactly UHRF1 binds caC sites in vivo and what implications this might have on epigenetic gene regulation.

## Supporting information

**S1 Fig. Raw gel images of EMSA experiments.** All raw gel scans that have been used to generate the EMSA results presented in Fig 2b/2c and S5 Fig. An overview of all individual

quantitative values and the corresponding statistics is provided on page 1.
(PDF)

**S2 Fig. Normalized MST traces of UHRF1 bound to C-C', mC-C', mC-caC' and caC-caC'.** Fluorescence traces that have been used to generate the binding curves in Fig 3. Traces are shown individually for all modifications and are coloured by experimental replicate. Blue and red bars indicate the time points that were used for the analysis; blue: $t_{cold}$ (pre infra-red laser), red: $t_{hot}$ (post infra-red laser).
(TIF)

**S3 Fig. Root Mean Squared Deviation (RMSD) of DNA atoms in molecular dynamics trajectories of UHRF1-SRA.** Coordinates were fitted to the initial crystal structure using the Cα atoms of protein residues 432 to 586. Only the last 1000 frames of each trajectory were used for analysis (vertical lines). Horizontal lines were added at 4 Å to highlight trajectories with strong structural distortions.
(TIF)

**S4 Fig. Root Mean Squared Deviation (RMSD) of protein atoms in molecular dynamics trajectories of UHRF1-SRA.** Coordinates were fitted to the initial crystal structure using the Cα atoms of protein residues 432 to 586. Only the last 1000 frames of each trajectory were used for analysis (vertical lines). Horizontal lines were added at 4 Å to highlight trajectories with strong structural distortions.
(TIF)

**S5 Fig. EMSAs of UHRF2 with differentially modified DNA.** Quantitation of the bound fraction of EMSAs of wild type UHRF2-GFP with 42 bp DNA oligonucleotides carrying different cytosine modifications. Experiments and analyses have been performed as in Fig 2.
(TIF)

**S6 Fig. Melting temperatures of modified DNA in presence of UHRF1-SRA.** (a) The melting temperature of double-stranded DNA containing C-C' in a CpG context (red) or no CpG site (black) with (solid lines) or without (dotted lines) a 5-fold excess of the SRA domain of UHRF1, measured using high resolution melting temperature (HRM) analysis. As control, proteins were digested by proteinase K before HRM analysis (right panel). Experiments were performed independently three times; one representative experiment is depicted as average of three technical replicates. (b) Melting temperatures as in (a) with DNA harbouring symmetric caC (green) or hemi-mC (gray) at the central CpG site.
(TIF)

**S7 Fig. Size exclusion chromatograms of differentially modified DNA in the presence or absence of UHRF1-SRA.** To test for different binding stoichiometries of the SRA domain towards differentially modified DNA, ATTO550-labeled DNA oligonucleotides were incubated with a 10-fold excess of SRA. Size exclusion chromatograms of analyzed DNA oligonucleotides at an absorbance of 554 nm (a) and 260nm/280nm (b) show a clear and comparable shift in retention time for the SRA-bound DNA (left peaks) compared to free DNA (right peaks).
(TIF)

**S8 Fig. Histograms of distances between Y471 and the flipped-out DNA base in molecular dynamics trajectories of UHRF1-SRA.** Individual replicas are shown as separate bars stacked on top of each other. Distances were measured between the geometric centres of the phenyl and pyrimidine rings. Red lines show a gaussian kernel estimate of the probability density

function (pdf). The estimated pdf of the mC-C' system is shown as black dashed lines.
(TIF)

**S9 Fig. Distribution of DNA minor groove widths in molecular dynamics trajectories of UHRF1-SRA.** Blue faces represent gaussian kernel estimates of the underlying values. Black bars show distribution means and standard deviations.
(TIF)

**S10 Fig. Distribution of DNA major groove widths in molecular dynamics trajectories of UHRF1-SRA.** Blue faces represent gaussian kernel estimates of the underlying values. Black bars show distribution means and standard deviations.
(TIF)

**S1 Text. Additional experimental procedures.**
(DOCX)

**S1 File. Parameter files for mC/caC used during molecular dynamics simulations.**
(ZIP)

## Author Contributions

**Conceptualization:** Markus Schneider, Carina Trummer, Heinrich Leonhardt, Christina Bauer, Iris Antes.

**Data curation:** Markus Schneider, Carina Trummer, Christina Bauer.

**Formal analysis:** Markus Schneider, Carina Trummer, Andreas Stengl, Peng Zhang, Christina Bauer.

**Funding acquisition:** M. Cristina Cardoso, Heinrich Leonhardt, Iris Antes.

**Investigation:** Carina Trummer, Andreas Stengl, Peng Zhang, Aleksandra Szwagierczak, Christina Bauer.

**Methodology:** Markus Schneider, Carina Trummer, Heinrich Leonhardt, Christina Bauer, Iris Antes.

**Project administration:** Heinrich Leonhardt, Iris Antes.

**Resources:** Heinrich Leonhardt, Iris Antes.

**Software:** Markus Schneider.

**Supervision:** Heinrich Leonhardt, Christina Bauer, Iris Antes.

**Validation:** Markus Schneider, Carina Trummer, Christina Bauer.

**Visualization:** Markus Schneider, Carina Trummer, Christina Bauer.

**Writing – original draft:** Markus Schneider, Christina Bauer, Iris Antes.

**Writing – review & editing:** Markus Schneider, Carina Trummer, Christina Bauer, Iris Antes.

## References

1. Bostick M, Kim JK, Esteve P-O, Clark A, Pradhan S, Jacobsen SE. UHRF1 Plays a Role in Maintaining DNA Methylation in Mammalian Cells. Science. 2007; 317:1760–4. https://doi.org/10.1126/science.1147939 PMID: 17673620.

2. Hashimoto H, Horton JR, Zhang X, Bostick M, Jacobsen SE, Cheng X. The SRA domain of UHRF1 flips 5-methylcytosine out of the DNA helix. Nature. 2008; 455:826–9. https://doi.org/10.1038/nature07280 PMID: 18772888.

3. Sharif J, Muto M, Takebayashi S, Suetake I, Iwamatsu A, Endo TA, et al. The SRA protein Np95 mediates epigenetic inheritance by recruiting Dnmt1 to methylated DNA. Nature. 2007; 450(7171):908–12. https://doi.org/10.1038/nature06397 PMID: 17994007.

4. Nishiyama A, Yamaguchi L, Sharif J, Johmura Y, Kawamura T, Nakanishi K, et al. Uhrf1-dependent H3K23 ubiquitylation couples maintenance DNA methylation and replication. Nature. 2013; 502 (7470):249–53. https://doi.org/10.1038/nature12488 PMID: 24013172.

5. Qin W, Wolf P, Liu N, Link S, Smets M, La Mastra F, et al. DNA methylation requires a DNMT1 ubiquitin interacting motif (UIM) and histone ubiquitination. Cell Res. 2015; 25(8):911–29. https://doi.org/10.1038/cr.2015.72 PMID: 26065575.

6. Ishiyama S, Nishiyama A, Saeki Y, Moritsugu K, Morimoto D, Yamaguchi L, et al. Structure of the Dnmt1 Reader Module Complexed with a Unique Two-Mono-Ubiquitin Mark on Histone H3 Reveals the Basis for DNA Methylation Maintenance. Mol Cell. 2017; 68(2):350–60 e7. https://doi.org/10.1016/j.molcel.2017.09.037 PMID: 29053958.

7. Tahiliani M, Koh KP, Shen Y, Pastor WA, Bandukwala H, Brudno Y, et al. Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1. Science. 2009; 324 (5929):930–5. https://doi.org/10.1126/science.1170116 PMID: 19372391.

8. He YF, Li BZ, Li Z, Liu P, Wang Y, Tang Q, et al. Tet-mediated formation of 5-carboxylcytosine and its excision by TDG in mammalian DNA. Science. 2011; 333(6047):1303–7. https://doi.org/10.1126/science.1210944 PMID: 21817016.

9. Pfaffeneder T, Hackner B, Truss M, Munzel M, Muller M, Deiml CA, et al. The discovery of 5-formylcytosine in embryonic stem cell DNA. Angew Chem Int Ed Engl. 2011; 50(31):7008–12. https://doi.org/10.1002/anie.201103899 PMID: 21721093.

10. Nabel CS, Kohli RM. Molecular biology. Demystifying DNA demethylation. Science. 2011; 333 (6047):1229–30. https://doi.org/10.1126/science.1211917 PMID: 21885763.

11. Pfaffeneder T, Spada F, Wagner M, Brandmayr C, Laube SK, Eisen D, et al. Tet oxidizes thymine to 5-hydroxymethyluracil in mouse embryonic stem cell DNA. Nat Chem Biol. 2014; 10(7):574–81. https://doi.org/10.1038/nchembio.1532 PMID: 24838012.

12. Globisch D, Münzel M, Müller M, Michalakis S, Wagner M, Koch S, et al. Tissue distribution of 5-hydroxymethylcytosine and search for active demethylation intermediates. PLoS ONE. 2010; 5:1–9. https://doi.org/10.1371/journal.pone.0015367 PMID: 21203455.

13. Eleftheriou M, Pascual AJ, Wheldon LM, Perry C, Abakir A, Arora A, et al. 5-Carboxylcytosine levels are elevated in human breast cancers and gliomas. Clinical epigenetics. 2015; 7. https://doi.org/10.1186/s13148-015-0117-x PMID: 26300993.

14. Lu X, Han D, Boxuan Simen Z, Song C-X, Zhang L-S, Doré LC, et al. Base-resolution maps of 5-formylcytosine and 5-carboxylcytosine reveal genome-wide DNA demethylation dynamics. Cell Research. 2015; 25:386–9. https://doi.org/10.1038/cr.2015.5 PMID: 25591929.

15. Neri F, Incarnato D, Krepelova A, Rapelli S, Anselmi F, Parlato C, et al. Single-Base resolution analysis of 5-formyl and 5-carboxyl cytosine reveals promoter DNA Methylation Dynamics. Cell Reports. 2015; 10:674–83. https://doi.org/10.1016/j.celrep.2015.01.008 PMID: 25660018.

16. Shen L, Wu H, Diep D, Yamaguchi S, D'Alessio AC, Fung HL, et al. Genome-wide analysis reveals TET- and TDG-dependent 5-methylcytosine oxidation dynamics. Cell. 2013; 153:692–706. https://doi.org/10.1016/j.cell.2013.04.002 PMID: 23602152.

17. Bronner C, Achour M, Arima Y, Chataigneau T, Saya H, Schini-Kerth VB. The UHRF family: oncogenes that are drugable targets for cancer therapy in the near future? Pharmacol Ther. 2007; 115(3):419–34. https://doi.org/10.1016/j.pharmthera.2007.06.003 PMID: 17658611.

18. Pichler G, Wolf P, Schmidt CS, Meilinger D, Schneider K, Frauer C, et al. Cooperative DNA and histone binding by Uhrf2 links the two major repressive epigenetic pathways. J Cell Biochem. 2011; 112 (9):2585–93. https://doi.org/10.1002/jcb.23185 PMID: 21598301.

19. Spruijt CG, Gnerlich F, Smits AH, Pfaffeneder T, Jansen PWTC, Bauer C, et al. Dynamic readers for 5-(Hydroxy)methylcytosine and its oxidized derivatives. Cell. 2013; 152:1146–59. https://doi.org/10.1016/j.cell.2013.02.004 PMID: 23434322.

20. Rajakumara E, Nakarakanti NK, Nivya MA, Satish M. Mechanistic insights into the recognition of 5-methylcytosine oxidation derivatives by the SUVH5 SRA domain. Scientific Reports. 2016; 6:20161. https://doi.org/10.1038/srep20161 PMID: 26841909

21. Wang L, Zhou Y, Xu L, Xiao R, Lu X, Chen L, et al. Molecular basis for 5-carboxycytosine recognition by RNA polymerase II elongation complex. Nature. 2015; 523:621–5. https://doi.org/10.1038/nature14482 PMID: 26123024.

22. Hashimoto H, Olanrewaju YO, Zheng Y, Wilson GG, Zhang X, Cheng X. Wilms tumor protein recognizes 5-carboxylcytosine within a specific DNA sequence. Genes Dev. 2014; 28(20):2304–13. https://doi.org/10.1101/gad.250746.114 PMID: 25258363.

23. Jin S-G, Zhang Z-M, Dunwell TL, Harter MR, Wu X, Johnson J, et al. Tet3 reads 5-carboxylcytosine through its CXXC domain and is a potential guardian against neurodegeneration. Cell Rep. 2016; 14:493–505. https://doi.org/10.1016/j.celrep.2015.12.044 PMID: 26774490.

24. Gowher H, Jeltsch A. Mammalian DNA methyltransferases: new discoveries and open questions. Biochemical Society Transactions. 2018; 46(5):1191–202. https://doi.org/10.1042/BST20170574 PMID: 30154093

25. Arand J, Spieler D, Karius T, Branco MR, Meilinger D, Meissner A, et al. In vivo control of CpG and non-CpG DNA methylation by DNA methyltransferases. PLoS Genet. 2012; 8(6):e1002750. https://doi.org/10.1371/journal.pgen.1002750 PMID: 22761581.

26. Xu L, Chen YC, Chong J, Fin A, McCoy LS, Xu J, et al. Pyrene-based quantitative detection of the 5-formylcytosine loci symmetry in the CpG duplex content during TET-dependent demethylation. Angew Chem Int Ed Engl. 2014; 53(42):11223–7. https://doi.org/10.1002/anie.201406220 PMID: 25159856.

27. Yu M, Hon GC, Szulwach KE, Song CX, Zhang L, Kim A, et al. Base-resolution analysis of 5-hydroxymethylcytosine in the mammalian genome. Cell. 2012; 149(6):1368–80. https://doi.org/10.1016/j.cell.2012.04.027 PMID: 22608086.

28. Arita K, Ariyoshi M, Tochio H, Nakamura Y, Shirakawa M. Recognition of hemi-methylated DNA by the SRA protein UHRF1 by a base-flipping mechanism. Nature. 2008; 455(7214):818–21. https://doi.org/10.1038/nature07249 PMID: 18772891.

29. Avvakumov GV, Walker JR, Xue S, Li Y, Duan S, Bronner C, et al. Structural basis for recognition of hemi-methylated DNA by the SRA domain of human UHRF1. Nature. 2008; 455(7214):822–5. https://doi.org/10.1038/nature07273 PMID: 18772889.

30. Frauer C, Hoffmann T, Bultmann S, Casa V, Cardoso MC, Antes I, et al. Recognition of 5-hydroxymethylcytosine by the Uhrf1 SRA domain. PLoS ONE. 2011; 6:1–8. https://doi.org/10.1371/journal.pone.0021306 PMID: 21731699.

31. Hashimoto H, Liu Y, Upadhyay AK, Chang Y, Howerton SB, Vertino PM, et al. Recognition and potential mechanisms for replication and erasure of cytosine hydroxymethylation. Nucleic Acids Research. 2012; 40(11):4841–9. https://doi.org/10.1093/nar/gks155 PMID: 22362737

32. Zhou T, Xiong J, Wang M, Yang N, Wong J, Zhu B, et al. Structural Basis for Hydroxymethylcytosine Recognition by the SRA Domain of UHRF2. Molecular Cell. 2014; 54:879–86. https://doi.org/10.1016/j.molcel.2014.04.003 PMID: 24813944.

33. Bianchi C, Zangi R. UHRF1 discriminates against binding to fully-methylated CpG-Sites by steric repulsion. Biophysical Chemistry. 2013; 171:38–45. https://doi.org/10.1016/j.bpc.2012.10.002 PMID: 23245651.

34. Hashimoto H, Zhang X, Cheng X. Activity and crystal structure of human thymine DNA glycosylase mutant N140A with 5-carboxylcytosine DNA at low pH. DNA Repair. 2013; 12:535–40. https://doi.org/10.1016/j.dnarep.2013.04.003 PMID: 23680598.

35. Gelato KA, Tauber M, Ong MS, Winter S, Hiragami-Hamada K, Sindlinger J, et al. Accessibility of different histone H3-binding domains of UHRF1 is allosterically regulated by phosphatidylinositol 5-phosphate. Mol Cell. 2014; 54(6):905–19. https://doi.org/10.1016/j.molcel.2014.04.004 PMID: 24813945.

36. Fang J, Cheng J, Wang J, Zhang Q, Liu M, Gong R, et al. Hemi-methylated DNA opens a closed conformation of UHRF1 to facilitate its histone recognition. Nat Commun. 2016; 7:11197. https://doi.org/10.1038/ncomms11197 PMID: 27045799.

37. Harrison JS, Cornett EM, Goldfarb D, DaRosa PA, Li ZM, Yan F, et al. Hemi-methylated DNA regulates DNA methylation inheritance through allosteric activation of H3 ubiquitylation by UHRF1. Elife. 2016; 5. https://doi.org/10.7554/eLife.17101 PMID: 27595565.

38. Vaughan RM, Dickson BM, Whelihan MF, Johnstone AL, Cornett EM, Cheek MA, et al. Chromatin structure and its chemical modifications regulate the ubiquitin ligase substrate selectivity of UHRF1. Proc Natl Acad Sci U S A. 2018; 115(35):8775–80. https://doi.org/10.1073/pnas.1806373115 PMID: 30104358.

39. Rottach A, Frauer C, Pichler G, Bonapace IM, Spada F, Leonhardt H. The multi-domain protein Np95 connects DNA methylation and histone modification. Nucleic Acids Res. 2010; 38(6):1796–804. https://doi.org/10.1093/nar/gkp1152 PMID: 20026581.

40. Ivani I, Dans PD, Noy A, Perez A, Faustino I, Hospital A, et al. Parmbsc1: a refined force field for DNA simulations. Nat Methods. 2016; 13(1):55–8. https://doi.org/10.1038/nmeth.3658 PMID: 26569599.

41. Lankaš F, Cheatham TE, Špačáková Na, Hobza P, Langowski J, Šponer J. Critical Effect of the N2 Amino Group on Structure, Dynamics, and Elasticity of DNA Polypurine Tracts. Biophysical Journal. 2002; 82(5):2592–609. https://doi.org/10.1016/s0006-3495(02)75601-4 PMID: 11964246

42. Perez A, Marchan I, Svozil D, Sponer J, Cheatham TE 3rd, Laughton CA, et al. Refinement of the AMBER force field for nucleic acids: improving the description of alpha/gamma conformers. Biophys J. 2007; 92(11):3817–29. https://doi.org/10.1529/biophysj.106.097782 PMID: 17351000.

43. Cieplak P, Cornell Wendy D, Bayly C, Kollman Peter A. Application of the multimolecule and multiconformational RESP methodology to biopolymers: Charge derivation for DNA, RNA, and proteins. Journal of Computational Chemistry. 1995; 16(11):1357–77. https://doi.org/10.1002/jcc.540161106

44. Vanquelef E, Simon S, Marquant G, Garcia E, Klimerak G, Delepine JC, et al. R.E.D. Server: a web service for deriving RESP and ESP charges and building force field libraries for new molecules and molecular fragments. Nucleic Acids Res. 2011; 39(Web Server issue):W511–7. https://doi.org/10.1093/nar/gkr288 PMID: 21609950.

45. Wang F, Becker J-P, Cieplak P, Dupradeau F-Y. R.E.D. Python: Object oriented programming for Amber force fields. Université de Picardie—Jules Verne, Sanford Burnham Prebys Medical Discovery Institute. 2013.

46. Dupradeau FY, Pigache A, Zaffran T, Savineau C, Lelong R, Grivel N, et al. The R.E.D. tools: advances in RESP and ESP charge derivation and force field library building. Phys Chem Chem Phys. 2010; 12 (28):7821–39. https://doi.org/10.1039/c0cp00111b PMID: 20574571.

47. Bayly CI, Cieplak P, Cornell W, Kollman PA. A well-behaved electrostatic potential based method using charge restraints for deriving atomic charges: the RESP model. The Journal of Physical Chemistry. 1993; 97(40):10269–80. https://doi.org/10.1021/j100142a004

48. Frisch MJ, Trucks GW, Schlegel HB, Scuseria GE, Robb MA, Cheeseman JR, et al. Gaussian 09 Revision A.2. 2009.

49. Case DA, Cerutti DS, T.E. Cheatham I, Darden TA, Duke RE, Giese TJ, et al. AMBER 2017. University of California, San Francisco. 2017.

50. 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.

51. 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. https://doi.org/10.1063/1.445869

52. 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.

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

54. 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.

55. El Hassan MA, Calladine CR. Two distinct modes of protein-induced bending in DNA. J Mol Biol. 1998; 282(2):331–43. https://doi.org/10.1006/jmbi.1998.1994 PMID: 9735291

56. 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.

57. Hunter JD. Matplotlib: A 2D Graphics Environment. Computing in Science & Engineering. 2007; 9 (3):90–5. https://doi.org/10.1109/MCSE.2007.55

58. Seidel SA, Dijkman PM, Lea WA, van den Bogaart G, Jerabek-Willemsen M, Lazic A, et al. Microscale thermophoresis quantifies biomolecular interactions under previously challenging conditions. Methods. 2013; 59(3):301–15. https://doi.org/10.1016/j.ymeth.2012.12.005 PMID: 23270813.

59. Qian C, Li S, Jakoncic J, Zeng L, Walsh MJ, Zhou MM. Structure and hemimethylated CpG binding of the SRA domain from human UHRF1. J Biol Chem. 2008; 283(50):34490–4. https://doi.org/10.1074/jbc.C800169200 PMID: 18945682.

60. Greiner VJ, Kovalenko L, Humbert N, Richert L, Birck C, Ruff M, et al. Site-Selective Monitoring of the Interaction of the SRA Domain of UHRF1 with Target DNA Sequences Labeled with 2-Aminopurine. Biochemistry. 2015; 54(39):6012–20. https://doi.org/10.1021/acs.biochem.5b00419 PMID: 26368281.

61. Kilin V, Gavvala K, Barthes NPF, Michel BY, Shin D, Boudier C, et al. Dynamics of Methylated Cytosine Flipping by UHRF1. Journal of the American Chemical Society. 2017; 139(6):2520–8. https://doi.org/10.1021/jacs.7b00154 PMID: 28112929

62. Helabad MB, Kanaan N, Imhof P. Base Flip in DNA Studied by Molecular Dynamics Simulations of Differently-Oxidized Forms of Methyl-Cytosine. International Journal of Molecular Sciences. 2014; 15 (7):11799–816. https://doi.org/10.3390/ijms150711799 PMID: 24995694

63. Szulik MW, Pallan PS, Nocek B, Voehler M, Banerjee S, Brooks S, et al. Differential Stabilities and Sequence-Dependent Base Pair Opening Dynamics of Watson–Crick Base Pairs with 5-Hydroxymethylcytosine, 5-Formylcytosine, or 5-Carboxylcytosine. Biochemistry. 2015; 54(5):1294–305. https://doi.org/10.1021/bi501534x PMID: 25632825

64. Huang N, Banavali NK, MacKerell AD. Protein-facilitated base flipping in DNA by cytosine-5-methyltransferase. Proceedings of the National Academy of Sciences. 2003; 100(1):68.

65. Rajakumara E, Law JA, Simanshu DK, Voigt P, Johnson LM, Reinberg D, et al. A dual flip-out mechanism for 5mC recognition by the Arabidopsis SUVH5 SRA domain and its impact on DNA methylation and H3K9 dimethylation in vivo. Genes Dev. 2011; 25(2):137–52. https://doi.org/10.1101/gad.1980311 PMID: 21245167.

66. Arita K, Isogai S, Oda T, Unoki M, Sugita K, Sekiyama N, et al. Recognition of modification status on a histone H3 tail by linked histone reader modules of the epigenetic regulator UHRF1. Proc Natl Acad Sci U S A. 2012; 109(32):12950–5. https://doi.org/10.1073/pnas.1203701109 PMID: 22837395.

67. Bianchi C, Zangi R. Dual base-flipping of cytosines in a CpG dinucleotide sequence. Biophysical Chemistry. 2014; 187–188:14–22. https://doi.org/10.1016/j.bpc.2013.12.005 PMID: 24469333.

68. Hashimoto H, Horton JR, Zhang X, Cheng X. UHRF1, a modular multi-domain protein, regulates replication-coupled crosstalk between DNA methylation and histone modifications. Epigenetics. 2009; 4:8–14. https://doi.org/10.4161/epi.4.1.7370 PMID: 19077538.

69. Bronner C, Fuhrmann G, Chédin FL, Macaluso M, Dhe-Paganon S. UHRF1 Links the Histone code and DNA Methylation to ensure Faithful Epigenetic Memory Inheritance. Genetics & epigenetics. 2010; 2009(2):29–36.

70. Grin I, Ishchenko AA. An interplay of the base excision repair and mismatch repair pathways in active DNA demethylation. Nucleic Acids Res. 2016; 44(8):3713–27. https://doi.org/10.1093/nar/gkw059 PMID: 26843430.

71. Kohli RM, Zhang Y. TET enzymes, TDG and the dynamics of DNA demethylation. Nature. 2013; 502:472–9. https://doi.org/10.1038/nature12750 PMID: 24153300.

72. Bochtler M, Kolano A, Xu GL. DNA demethylation pathways: Additional players and regulators. Bioessays. 2017; 39(1):1–13. https://doi.org/10.1002/bies.201600178 PMID: 27859411.

73. Maiti A, Drohat AC. Thymine DNA glycosylase can rapidly excise 5-formylcytosine and 5-carboxylcytosine: potential implications for active demethylation of CpG sites. J Biol Chem. 2011; 286(41):35334–8. https://doi.org/10.1074/jbc.C111.284620 PMID: 21862836.

74. Muller U, Bauer C, Siegl M, Rottach A, Leonhardt H. TET-mediated oxidation of methylcytosine causes TDG or NEIL glycosylase dependent gene reactivation. Nucleic Acids Res. 2014; 42(13):8592–604. https://doi.org/10.1093/nar/gku552 PMID: 24948610.

75. Weber AR, Krawczyk C, Robertson AB, Kusnierczyk A, Vagbo CB, Schuermann D, et al. Biochemical reconstitution of TET1-TDG-BER-dependent active DNA demethylation reveals a highly coordinated mechanism. Nat Commun. 2016; 7:10806. https://doi.org/10.1038/ncomms10806 PMID: 26932196.

76. Luo T, Cui S, Bian C, Yu X. Uhrf2 is important for DNA damage response in vascular smooth muscle cells. Biochem Biophys Res Commun. 2013; 441(1):65–70. https://doi.org/10.1016/j.bbrc.2013.10.018 PMID: 24134842.

77. Mistry H, Tamblyn L, Butt H, Sisgoreo D, Gracias A, Larin M, et al. UHRF1 is a genome caretaker that facilitates the DNA damage response to gamma-irradiation. Genome Integr. 2010; 1(1):7. https://doi.org/10.1186/2041-9414-1-7 PMID: 20678257.

78. Tian Y, Paramasivam M, Ghosal G, Chen D, Shen X, Huang Y, et al. UHRF1 contributes to DNA damage repair as a lesion recognition factor and nuclease scaffold. Cell Reports. 2015; 10:1957–66. https://doi.org/10.1016/j.celrep.2015.03.038 PMID: 25818288.

79. Laget S, Miotto B, Chin HG, Esteve PO, Roberts RJ, Pradhan S, et al. MBD4 cooperates with DNMT1 to mediate methyl-DNA repression and protects mammalian cells from oxidative stress. Epigenetics. 2014; 9(4):546–56. https://doi.org/10.4161/epi.27695 PMID: 24434851.

80. O'Hagan HM, Wang W, Sen S, Destefano Shields C, Lee SS, Zhang YW, et al. Oxidative damage targets complexes containing DNA methyltransferases, SIRT1, and polycomb members to promoter CpG Islands. Cancer Cell. 2011; 20(5):606–19. https://doi.org/10.1016/j.ccr.2011.09.012 PMID: 22094255.

81. Kamiya H, Tsuchiya H, Karino N, Ueno Y, Matsuda A, Harashima H. Mutagenicity of 5-Formylcytosine, an Oxidation Product of 5-Methylcytosine, in DNA in Mammalian Cells1. The Journal of Biochemistry. 2002; 132(4):551–5. https://doi.org/10.1093/oxfordjournals.jbchem.a003256 PMID: 12359069

82. Shibutani T, Ito S, Toda M, Kanao R, Collins LB, Shibata M, et al. Guanine- 5-carboxylcytosine base pairs mimic mismatches during DNA replication. Sci Rep. 2014; 4:5220. https://doi.org/10.1038/srep05220 PMID: 24910358.

**83.** Kim KY, Tanaka Y, Su J, Cakir B, Xiang Y, Patterson B, et al. Uhrf1 regulates active transcriptional marks at bivalent domains in pluripotent stem cells through Setd1a. Nat Commun. 2018; 9(1):2583. https://doi.org/10.1038/s41467-018-04818-0 PMID: 29968706.

**84.** Verma N, Pan H, Dore LC, Shukla A, Li QV, Pelham-Webb B, et al. TET proteins safeguard bivalent promoters from de novo methylation in human embryonic stem cells. Nat Genet. 2018; 50(1):83–95. https://doi.org/10.1038/s41588-017-0002-y PMID: 29203910.

**85.** Deplus R, Delatte B, Schwinn MK, Defrance M, Mendez J, Murphy N, et al. TET2 and TET3 regulate GlcNAcylation and H3K4 methylation through OGT and SET1/COMPASS. EMBO J. 2013; 32(5):645–55. https://doi.org/10.1038/emboj.2012.357 PMID: 23353889.

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 $\alpha, \beta$ 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, $\chi$ 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 $\delta$ is the Kronecker delta and the $\delta_{kl}$ 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

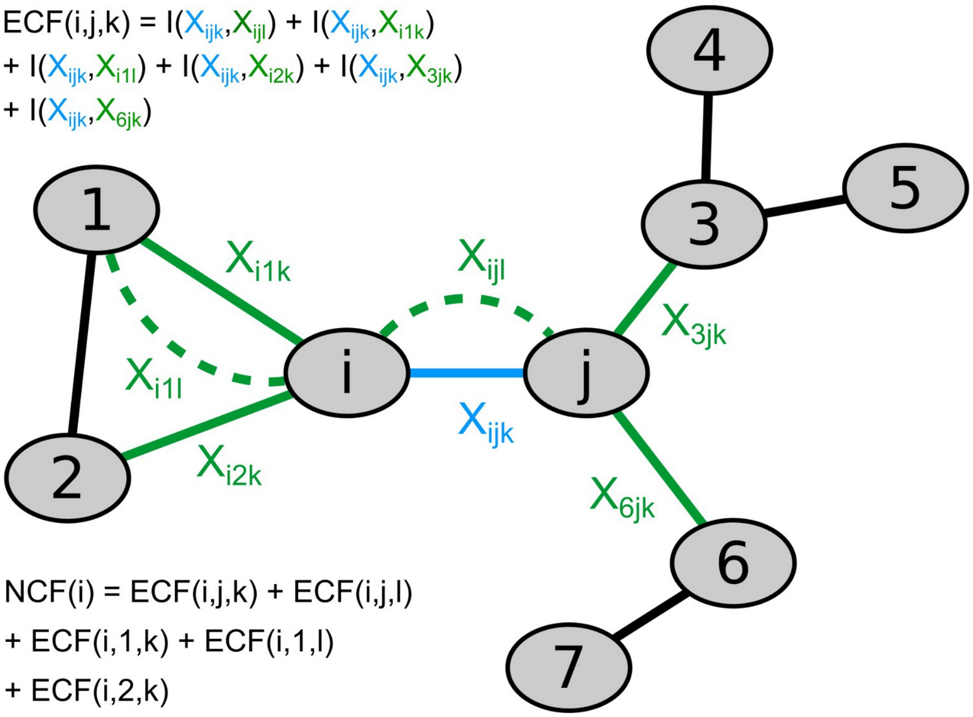

ECF(i,j,k) = I(X_ijk,X_ijl) + I(X_ijk,X_i1k)
+ I(X_ijk,X_i1l) + I(X_ijk,X_i2k) + I(X_ijk,X_3jk)
+ I(X_ijk,X_6jk)
4
5
1
3
X_i1k
X_ijl
X_3jk
X_i1l
i
j
X_ijk
2
X_i2k
X_6jk
6
NCF(i) = ECF(i,j,k) + ECF(i,j,l)
+ ECF(i,1,k) + ECF(i,1,l)
+ ECF(i,2,k)
7

**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

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

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 $N$, $\sigma_{jk}$ 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].

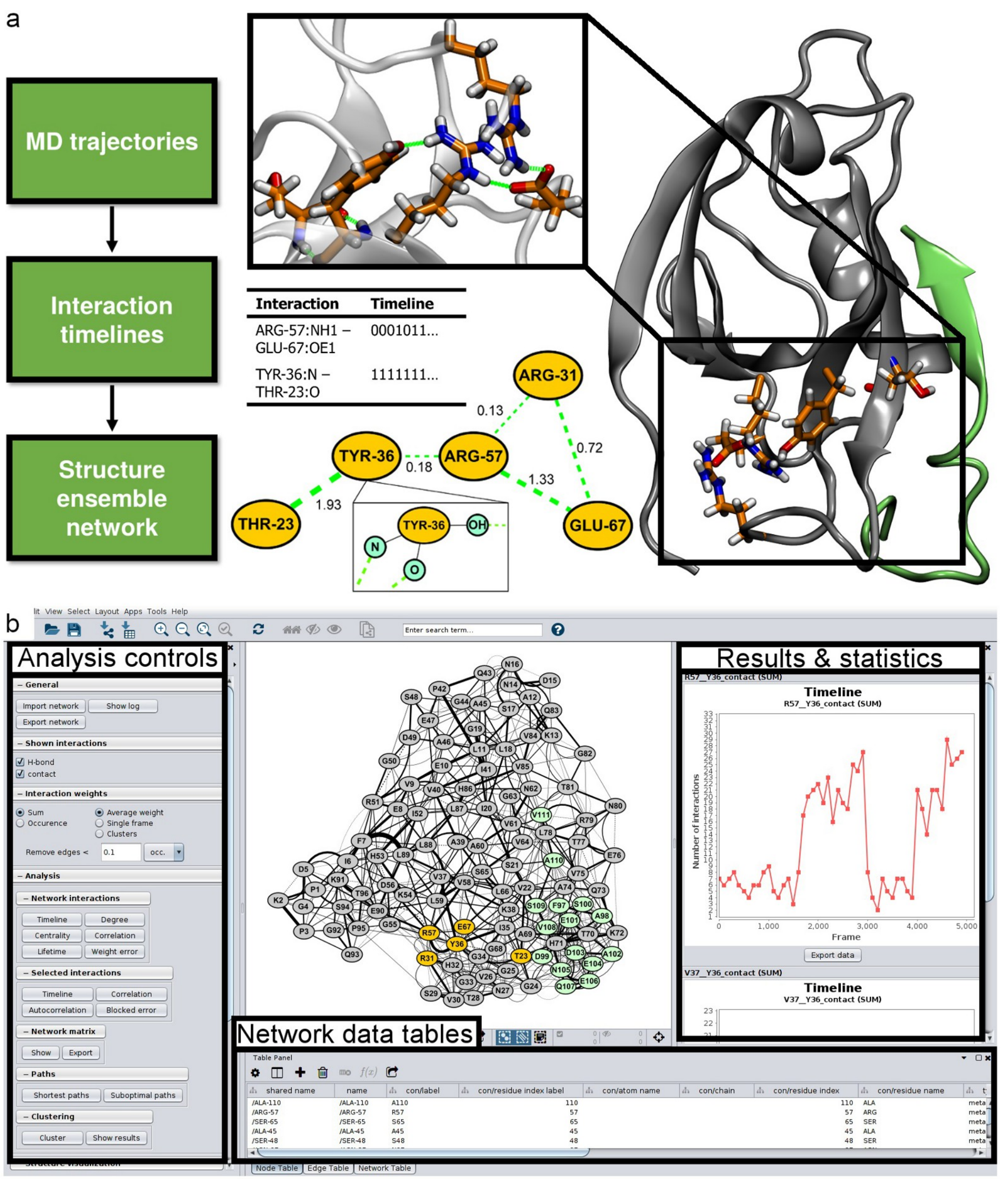
### Prediction of allosteric residues

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



| Interaction | Timeline |
|---|---|
| ARG-57:NH1 – GLU-67:OE1 | 0001011... |
| TYR-36:N – THR-23:O | 1111111... |

General: Import network, Show log, Export network

Shown interactions: ☑ H-bond ☑ contact

Interaction weights: ● Sum ○ Occurence ● Average weight ○ Single frame ○ Clusters

Remove edges < 0.1 occ.

Analysis

Network interactions: Timeline, Degree, Centrality, Correlation, Lifetime, Weight error

Selected interactions: Timeline, Correlation, Autocorrelation, Blocked error

Network matrix: Show, Export

Paths: Shortest paths, Suboptimal paths

Clustering: Cluster, Show results

R57_Y36_contact (SUM)

Timeline
R57_Y36_contact (SUM)

V37_Y36_contact (SUM)

Timeline
V37_Y36_contact (SUM)

Table Panel

| shared name | name | con/label | con/residue index label | con/atom name | con/chain | con/residue index | con/residue name | t |
|---|---|---|---|---|---|---|---|---|
| /ALA-110 | /ALA-110 | A110 | 110 | | | 110 | ALA | meta |
| /ARG-57 | /ARG-57 | R57 | 57 | | | 57 | ARG | meta |
| /SER-65 | /SER-65 | S65 | 65 | | | 65 | SER | meta |
| /ALA-45 | /ALA-45 | A45 | 45 | | | 45 | ALA | meta |
| /SER-48 | /SER-48 | S48 | 48 | | | 48 | SER | meta |

**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

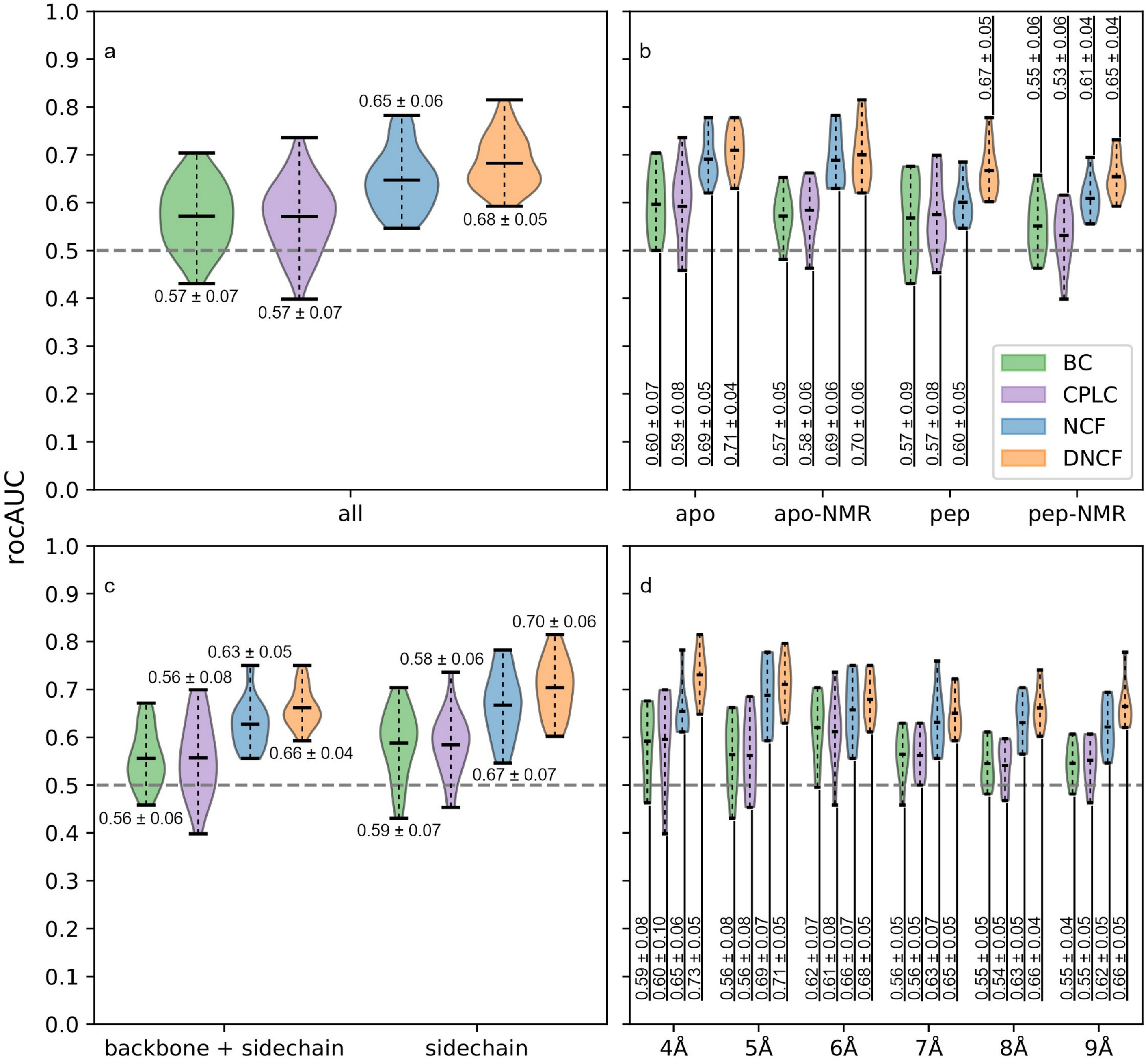


**Fig 3. Influence of network parameters on prediction model performance based on the NMR-Ala reference set.** Shaded areas show distribution estimates based on a gaussian kernel with added labels for mean and standard deviation. (a) Distributions including all parameter combinations. (b) Source of analyzed 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

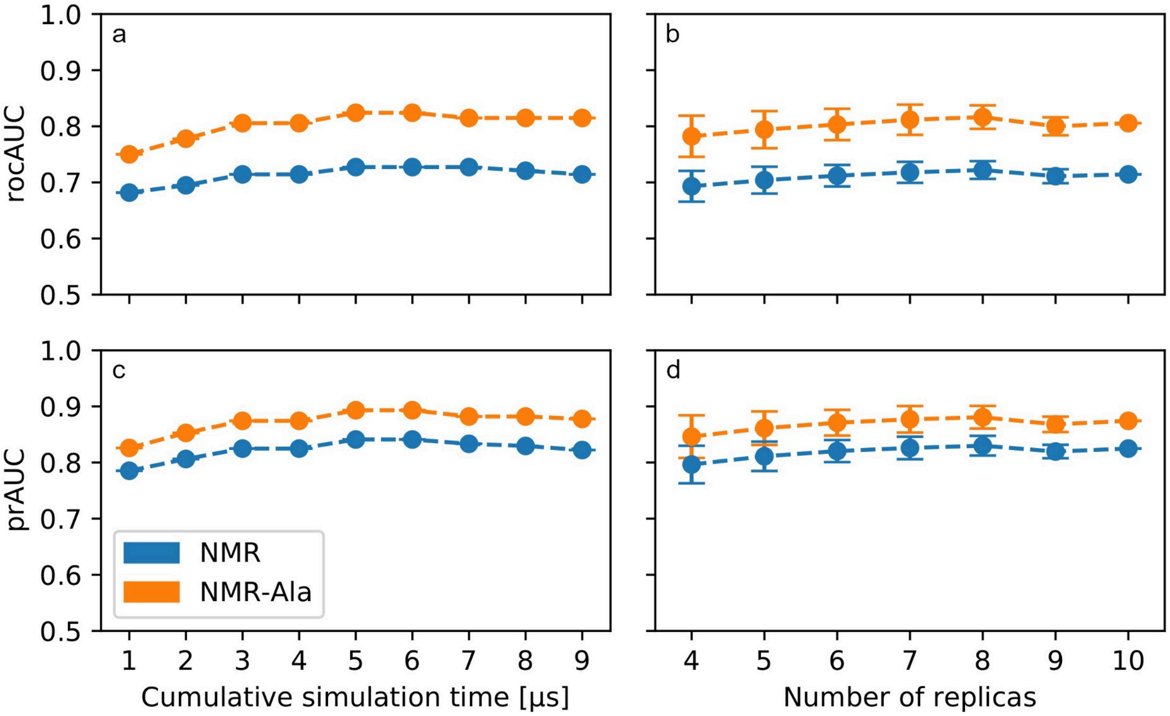


**Fig 4. Effect of simulation time and number of replicas on prediction performance of the final DNCF model.** (a,c) Timelines of all ten replicas were truncated, merged to the specified cumulative simulation time and analyzed successively. 1 μs of cumulative simulation time corresponds to a simulation 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

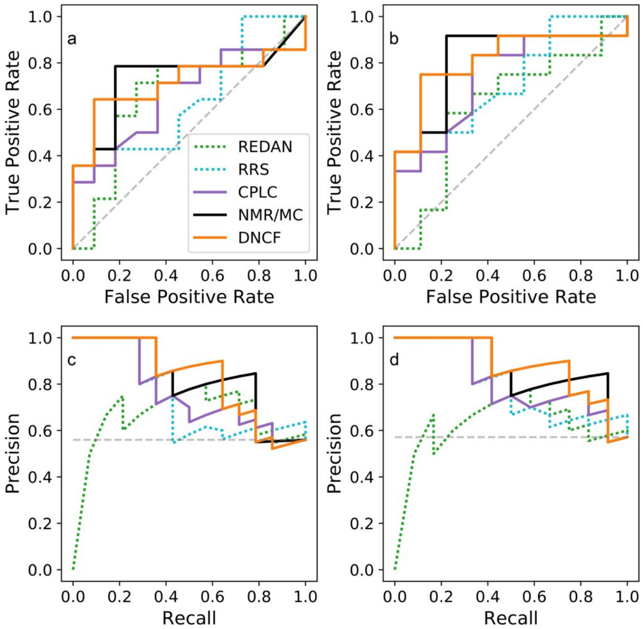


**Fig 5. ROC and PR curves of selected prediction models.** (a) ROC curve based on the NMR reference set. (b) ROC curve based on the NMR-Ala reference set. (c) PR curve based on the NMR reference set. (d) PR curve based on the NMR-Ala reference set.

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

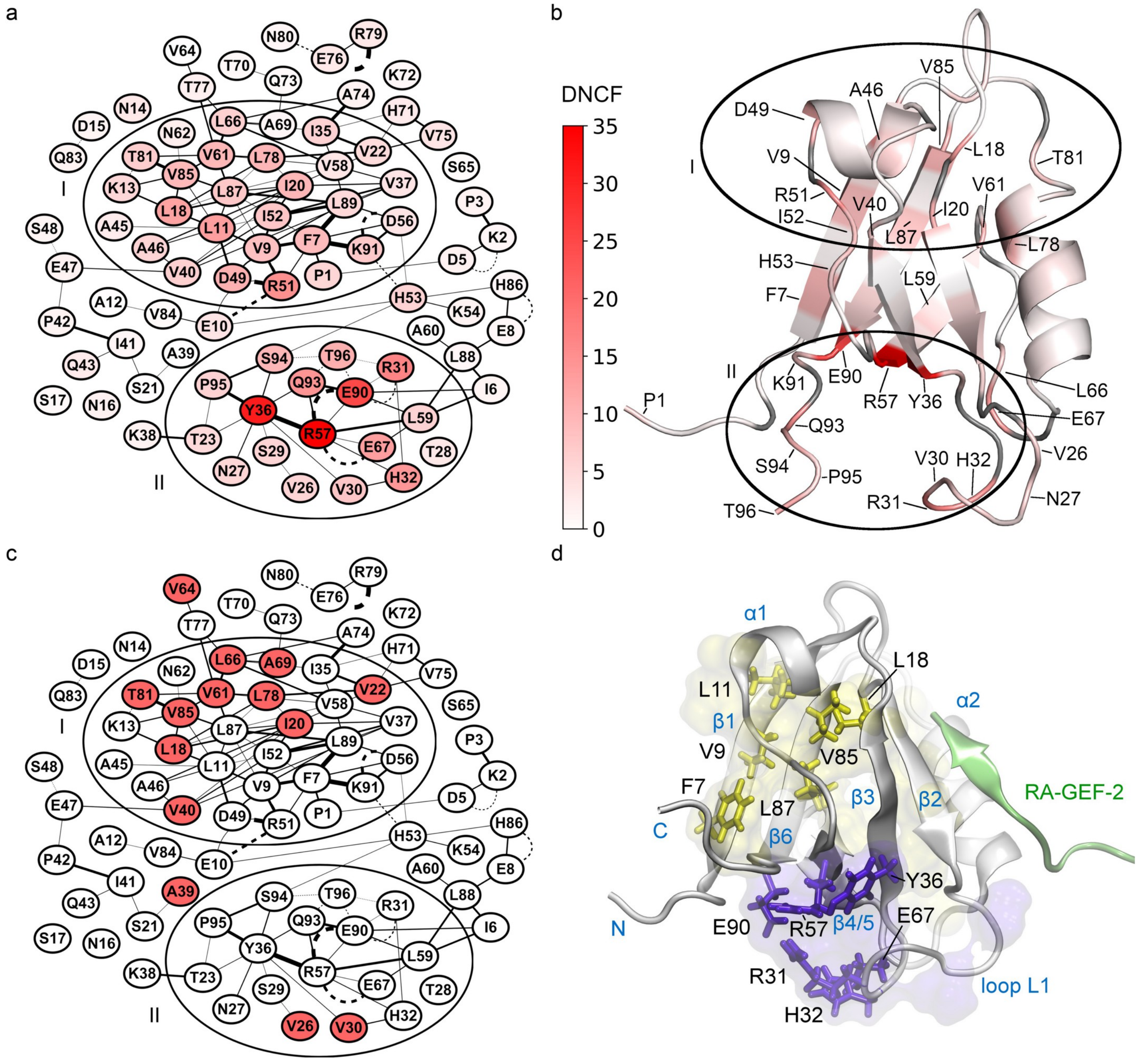


**Fig 6. Allosteric predictions of the final DNCF model mapped to PDZ2 structures.** For visual clarity, only edges occurring in ≥0.1% of simulation time are shown. (a) Network representation of DNCF predictions. Nodes are colored from low (white) to high (red) DNCF scores. (b) DNCF scores mapped to 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

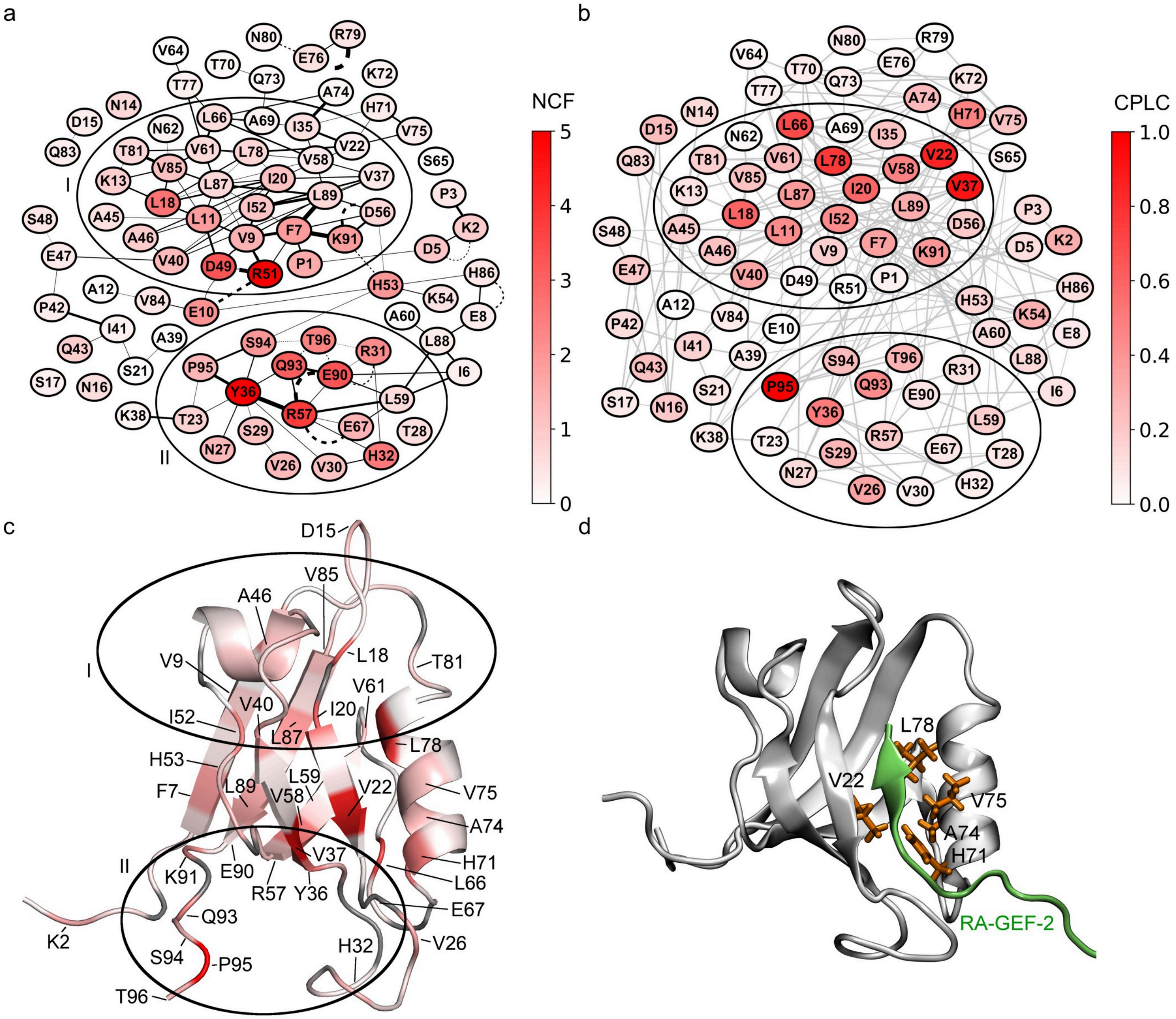


**Fig 7. Allosteric predictions of the final NCF and CPLC models mapped to PDZ2 structures.** Nodes colored from low (white) to high (red) scores. (a) Network representation of NCF predictions. For visual clarity, only edges occurring in ≥0.1% of simulation time are shown. (b) Network representation 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

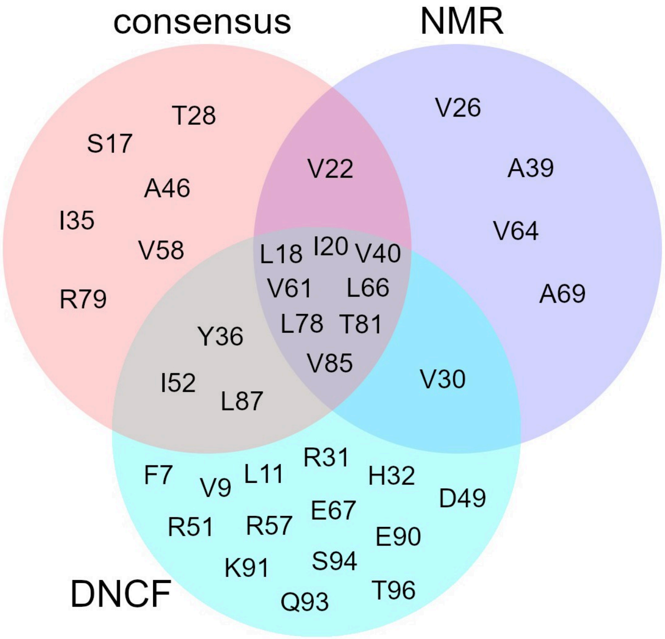


**Fig 8. Intersection of the DNCF allosteric core set, NMR reference set, and the computational prediction consensus set.**

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

Received: 30 March 2022 | Revised: 6 August 2022 | Accepted: 13 September 2022

DOI: 10.1002/prot.26425

RESEARCH ARTICLE

# Comparison of allosteric signaling in DnaK and BiP using mutual information between simulated residue conformations

Markus Schneider | Iris Antes[†]

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

**Correspondence**
Markus Schneider, TUM Center for Functional Protein Assemblies and TUM School of Life Sciences, Technische Universität München, Klebelstrasse 5, 85356 Freising, Bavaria, Germany.
Email: markusg.schneider@tum.de

**Funding information**
Deutsche Forschungsgemeinschaft, Grant/Award Numbers: SFB749/C08, SFB 1035/A10

## Abstract

The heat shock protein 70 kDa (Hsp70) chaperone system serves as a critical component of protein quality control across a wide range of prokaryotic and eukaryotic organisms. Divergent evolution and specialization to particular organelles have produced numerous Hsp70 variants which share similarities in structure and general function, but differ substantially in regulatory aspects, including conformational dynamics and activity modulation by cochaperones. The human Hsp70 variant BiP (also known as GRP78 or HSPA5) is of therapeutic interest in the context of cancer, neurodegenerative diseases, and viral infection, including for treatment of the pandemic virus SARS-CoV-2. Due to the complex conformational rearrangements and high sequential variance within the Hsp70 protein family, it is in many cases poorly understood which amino acid mutations are responsible for biochemical differences between protein variants. In this study, we predicted residues associated with conformational regulation of human BiP and *Escherichia coli* DnaK. Based on protein structure networks obtained from molecular dynamics simulations, we analyzed the shared information between interaction timelines to highlight residue positions with strong conformational coupling to their environment. Our predictions, which focus on the binding processes of the chaperone's substrate and cochaperones, indicate residues filling potential signaling roles specific to either DnaK or BiP. By combining predictions of individual residues into conformationally coupled chains connecting ligand binding sites, we predict a BiP specific secondary signaling pathway associated with substrate binding. Our study sheds light on mechanistic differences in signaling and regulation between Hsp70 variants, which provide insights relevant to therapeutic applications of these proteins.

KEYWORDS

allostery, HSP70 heat-shock proteins, molecular dynamics simulation, protein structure networks, signal transduction

[†] Died August 4, 2021.

## 1 | INTRODUCTION

The heat shock protein 70 kDa (Hsp70) molecular chaperone is a class of proteins found across a wide range of prokaryotic and eukaryotic organisms, with no fewer than 13 isoforms in humans alone.[1–3] Their functional roles include protein (re-)folding, membrane translocation, regulation of apoptosis, and disaggregation of denatured proteins in cooperation with other chaperone systems.[1–7] Being both ubiquitous and critical to cell damage mitigation, they are also of therapeutic interest for a variety of conditions such as cancer and neurodegenerative disorders.[2,3,7–12] The ATP-driven conformational cycle allows Hsp70s to transiently bind exposed hydrophobic stretches of substrate proteins and selectively modulate their folding process. Structurally, a Hsp70 protein is divided into a number of distinct modular domains: The nucleotide-binding domain (NBD) is an actin-like ATPase with two rotatable lobes (NBD-I and NBD-II), connected to the substrate binding domain (SBD) via a flexible linker region (NBD–SBD linker). The SBD is further divided into a β-sandwich core forming a cleft for binding of target peptides or proteins (SBDβ) and an α-helical lid which can dynamically open and close over the binding cleft (SBDα), followed by an unstructured C-terminal tail.[2,6,7,11,13,14] In the ATP-bound state of the conformational cycle, the NBD and SBDβ are predominantly docked onto each other, with the SBDα lid in the "open" conformation and stabilized by contacts with the NBD (Figure 1). In this docked conformation, the NBD–SBD linker is nestled in a cleft formed on the surface of NBD-II. Upon binding of a substrate polypeptide in the SBDβ binding cleft, the NBD–SBD interface partially undocks, allowing the NBD lobes to rotate into a position activating ATP hydrolysis. This leads to complete undocking of the NBD–SBD domains, freeing the NBD–SBD linker and stabilizing Hsp70 in a conformation with two separate domains, the NBD binding ADP and the SBD binding the substrate. The cycle is finally

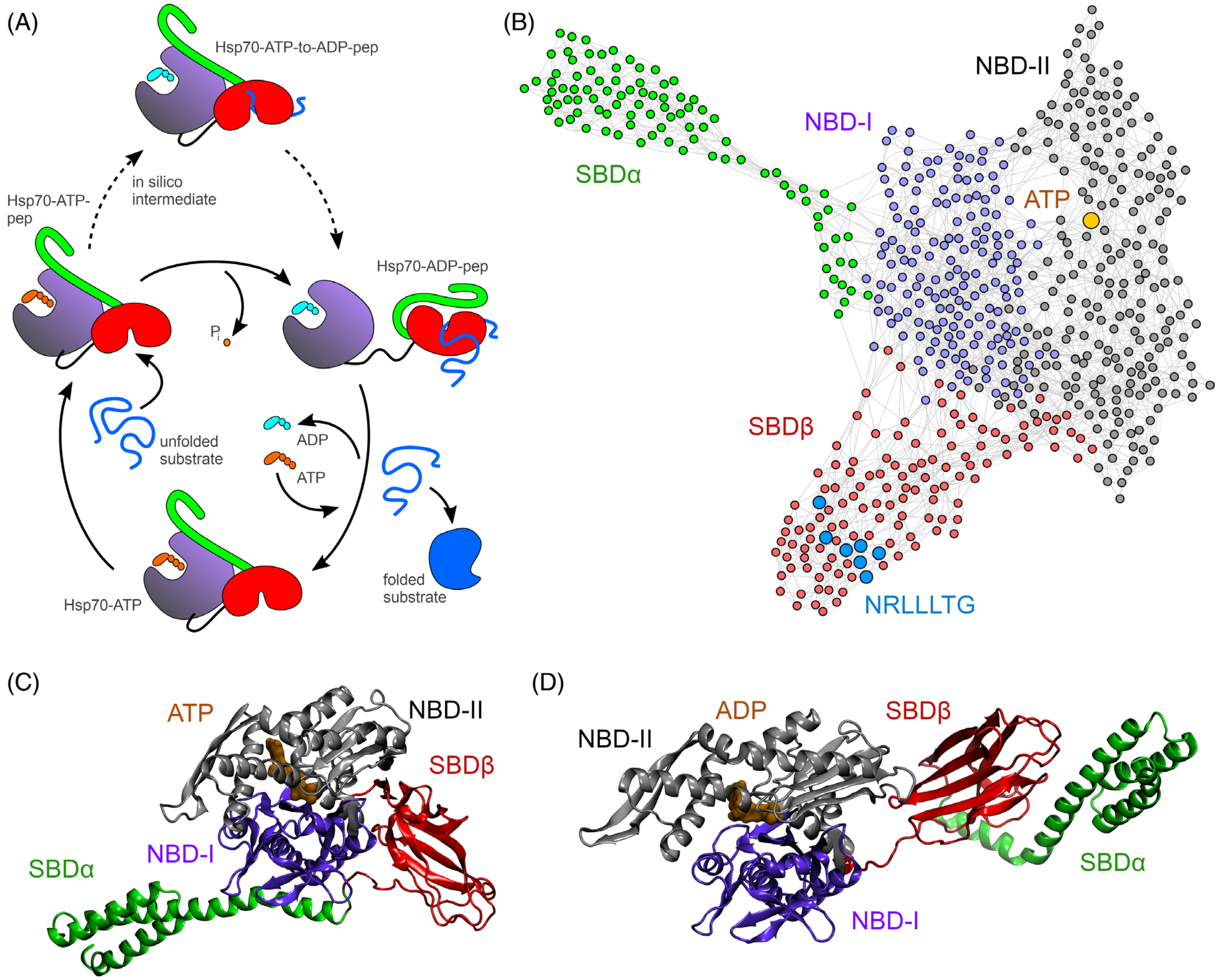


**FIGURE 1** Structural organization and conformations of heat shock protein 70 kDa (Hsp70) chaperones. (A) Simplified representation of the Hsp70 conformational cycle. (B) Protein structure network of DnaK-ATP bound to the NRLLLTG peptide. (C,D) Representative structures of DnaK-ATP (C) and DnaK-ADP (D) extracted from molecular dynamics simulations with subdomain coloring. ATP/ADP nucleotides are shown in brown. NBD, nucleotide-binding domain; SBD, substrate binding domain.

completed by nucleotide exchange of ADP to ATP and subsequent re-docking of the NBD and SBD domains.[2,6,7,11,14–17] This complex orchestration is achieved through multiple pathways transmitting conformational changes and fluctuations throughout the protein, with the nucleotide and protein substrate ligands acting in concert to advance the conformational cycle.[2,6,7,11,13,14,16,18,19] In addition, co-chaperones like J-domain proteins (JDPs) or nucleotide exchange factors (NEF) accelerate these processes by forming transient complexes with Hsp70 and modulating its functional and conformational cycle.[2,6,7,11,15,20–23]

Of the diverse family of Hsp70 proteins, the *Escherichia coli* variant DnaK is by far the most extensively studied. A comprehensive series of biochemical experiments has investigated effects of point mutations on protein activity,[13] gradually assembling a model for understanding the underlying allosteric mechanisms advancing the conformational cycle. While Hsp70s from other organisms generally share the same structural architecture, it has been noted that they can differ substantially in substrate recognition, allosteric signaling, and co-chaperone interactions.[1,2,7,15,16,24–29] The human Hsp70 isoform binding immunoglobulin protein (BiP; also known as GRP78 or HSPA5) is usually found in the endoplasmatic reticulum, where it folds membrane and secretes proteins.[30,31] However, under specific conditions BiP or its isoforms can appear in certain cell types within the cytosol, nucleus, mitochondria, and on the cell surface (csBiP).[32–34] Tumors and cells under stress show elevated levels of csBiP compared to normal cells, which has prompted investigations into applications for cancer detection and therapy.[35–39] Moreover, csBiP is implicated as a coreceptor promoting host cell entry for several coronaviruses, including the pandemic virus SARS-CoV-2,[40–44] and has been suggested as a therapeutic factor for severe COVID-19 cases.[41,44] Elucidating the evolutionary differences distinguishing Hsp70 variants could deepen our understanding of this important protein class and help to tailor drugs to the specific properties of the targeted protein. The sequence homology between BiP and DnaK is below 50%, while several functional differences have been reported between the two variants, such as propensity of NBD–SBD docking, post-translational modifications or interactions with cochaperones modulating allosteric signaling.[7,16,24,31,45–47] It is mostly unknown which mutations are responsible for the observed functional and regulatory differences,[48] which are difficult to pinpoint due to the large size of Hsp70s (>600 amino acids), high sequential variance within the family and the complexity of the conformational cycle. In this work, we investigated residues associated with conformational control in the Hsp70 proteins DnaK and BiP using our recently developed difference node correlation factor (DNCF) method, which is based on estimating the shared information between interaction timelines obtained from molecular dynamics (MD) simulations and evaluating how this shared information is modulated by, for example, binding of a ligand.[49] This method implements a new variant within the class of graph-based allosteric prediction frameworks, which have been applied with success to a diverse range of proteins such as CFTR,[50] GPCRs,[51] myoglobins,[51] Hsp90,[52] and others.

Although there have been a number of studies reporting computational predictions of allostery in Hsp70s,[53–60] there is limited information on how evolutionary differences affect transferability of allosteric models between DnaK and BiP. Our predictions complement the set of experimentally determined individual residues by predicting pathways of conformationally coupled residues, which we presume to be involved in the initiation of conformational changes and allostery following a ligand binding event. On this basis, we suggest a number of residue positions which might explain the functional differences between DnaK and BiP.

## 2 | MATERIALS AND METHODS

### 2.1 | Protein structures

Structures for full length DnaK-ATP (PDB-ID: 4B9Q), BiP-ATP (PDB-ID: 5E84), and DnaK-ADP (PDB-ID: 2KHO; first model in file was used) were acquired from the RCSB PDB web site. Protein models were adjusted to reflect the sequences in Text S1 using IRECS[61] to mutate side chains in 4B9Q and MODELLER (v 9.18)[62] to mutate and add missing residues in 5E84, selecting the model with the best DOPE score out of 100 candidates. Structures containing only the NBD were derived from the full-length structures by cutting at the N-terminal side of the NBD–SBD linker. For DnaK-NBD-ADP, the structure of full-length DnaK-ADP was cut at the linker and ADP/$Mg^{2+}$ was added based on the 4B9Q structure. The structure for BiP–NBD–ADP was created as a homology model using MOD based on a template of yeast BiP (PDB-ID: 3QFU), using the same procedure as described above. For all systems, ATP-to-ADP variants were obtained by cutting the terminal ATP phosphate from the corresponding ATP-bound structures. Ions present in crystal structures were removed with the exception of magnesium located in the nucleotide binding pocket. The NRLLLTG peptide was added to relevant systems based on the conformation found in the DnaK–peptide complex (PDB-ID: 1DKX). To allow for easier comparison between systems, indices describing residue positions were adjusted in all systems to reflect the DnaK sequence (UniProt-ID: P0A6Y8), adding PDB residue insertion codes as needed (see full sequence alignment in Text S2). Furthermore, a complete mapping of the DnaK residue indices to the UniProt sequence numbering of BiP (UniProt-ID: P11021) is provided in Table S1. Protein regions were defined by the following residue index ranges, based on DnaK: NBD-I from 1 to 177, NBD-II from 178 to 383, SBDβ from 384 to 506, and SBDα from 507 to 603.

### 2.2 | MD simulations

MD simulations were performed using the Amber16-AmberTools16/17 software suite[63] with the Amber14SB force field,[64] and TIP3P water[65] using ATP/ADP parameters from Meagher et al.[66] The system was solvated in a cubic water box using a minimum solute-face distance of 12 Å and neutralized with NaCl. For the nonbonded interactions a 12 Å direct space cutoff and particle mesh Ewald (PME) summation for long-ranged 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 density of 1.00 kg $m^3$. Systems were gradually heated from 0 to 300 K over 1.5 ns using a variant of our published heatup protocol,[67] restraining all heavy atoms with a force constant of 3.00 kcal $mol^{-1}$ $Å^{-2}$ until 20 K and all protein backbone atoms until 200 K. SHAKE[68] was applied to all bonds involving hydrogen and an integration time step of 1 fs was used during heatup, increasing to 2 fs for subsequent production runs. For heating and temperature control, a Langevin thermostat was used with a collision frequency of 4 $ps^{-1}$, and beginning from the final 0.5 ns of the heatup, a Berendsen barostat was employed with a relaxation time of 1 ps. For each system, three independent replica runs were simulated for 400 ns each, starting from separate heatup runs and with randomized Langevin seeds. The initial 100 ns of each run were removed before analysis to reduce bias toward initial structures. Atom interactions were extracted from MD trajectories with CPPTRAJ,[69] using the "nativecontacts" command for contact timelines (distance cutoff 5 Å; saving both native and nonnative time series), and the "hbond" command for hydrogen bonds (distance cutoff 3.5 Å; angle cutoff 135°).

## 2.3 | Protein structure networks

For analyses of protein structure networks and related quantities we used our network analysis tool SenseNet (version 1.1.0),[49] a plugin for Cytocape 3 (version 3.6.1).[70] CPPTRAJ outputs of contact and hydrogen bond timelines were processed using AIFgen[49] and loaded into SenseNet. Edges representing interactions occurring in less than 10% of the total simulation time were removed from the networks to minimize the influence of spurious interactions. DNCF scores were calculated in SenseNet as described before,[49] using the "Correlation" function set to the "Neighbor" and "Mutual information difference" modes. 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 set to "Match Location"). As reference for DNCF calculations, we selected the corresponding networks from Hsp70-ATP (for analyses of the full-length protein) or Hsp70-NBD-ATP (for analyses of the isolated NBD domain). The DNCF method evaluates the changes in conformational coupling in neighboring interactions between a target and a reference simulation, for example, between Hsp70-ATP and Hsp70-ATP-to-ADP. Contacts and hydrogen bonds between residues are described as a timeline encoding the number of interactions in each time frame of the MD trajectory. The DNCF score is calculated as follows: Each carbon–carbon contact and each hydrogen bond in the network is represented by a separate edge $X$ in the network. Another edge $Y$ is said to be neighboring if it shares at least one node with $X$. In other words, the neighboring interactions represented by $X$ and $Y$ share at least one common residue (e.g., two different hydrogen bonds formed by one residue to different interaction partners). For each pair of neighboring interactions $X$ and $Y$ in the target simulation, the equivalent interactions $\widehat{X}$ and $\widehat{Y}$ are obtained from the reference simulation. Then, the change in shared information of the selected interaction pair between timelines from the target and reference simulation is evaluated using the difference in pointwise mutual information as

$$I(X;Y) = \sum_{x \in \cup\left(X,\widehat{X}\right)} \sum_{y \in \cup\left(Y,\widehat{Y}\right)} \left| p(x,y) \cdot \log_2\left(\frac{p(x,y)}{p(x)p(y)}\right) - \widehat{p}(x,y) \cdot \log_2\left(\frac{\widehat{p}(x,y)}{\widehat{p}(x)\widehat{p}(y)}\right) \right|, \qquad (1)$$

with $\widehat{X}, \widehat{Y}$ denoting the timelines from the reference simulation matching the locations of $X$ and $Y$ of the target simulation, and $p, \widehat{p}$ representing the probabilities of interaction states within the target and reference timelines. Finally, the DNCF score for each residue is obtained by summing the contributions of Equation (1) for all interactions that residue is participating in. More methodological details, including an extensive discussion on network parameters and simulation setups, can be found in our previous work.[49]

Random walks weighted by DNCF scores ("DNCF-RW") were performed using the "Random Walk" function of SenseNet in "Targeted Symmetric" mode, starting from the node representing the central leucine of the NRLLLTG peptide substrate and stopping the search when the ATP node was reached (or vice versa). Given a starting ("current") node for the random walk, the next node to be visited is selected from the list of connected neighbor nodes with the probability distribution

$$p(i) = \frac{\mathrm{DNCF}(i)}{\sum_{n \in N} \mathrm{DNCF}(n)}, \qquad (2)$$

where the candidate node $i$ is part of the set of neighbors $N$, that is, nodes connected to the current node. Revisiting nodes was permitted, but their contribution was only counted once. The search was restarted if the target node was not found after 1000 steps, and in addition with a probability of 0.1 at each step, ensuring that the search rejected pathways substantially longer than the shortest possible path (see Section 3). Shortest paths between two nodes were calculated using Dijkstra's algorithm, as implemented by the "Shortest path" function of SenseNet. Plots were generated using matplotlib (version 3.0.3)[71] with pictures of molecular structures by VMD (version 1.9.3)[72] and open-source PyMOL (Schrodinger, LLC. 2010. The PyMOL Molecular Graphics System, Version 1.8.4.0).

## 2.4 | Analysis of NBD lobe rotation

All clustering analyses were performed with the "cluster" command of CPPTRAJ. First, trajectories were aligned to the Cα atoms of NBD-I.

Following this alignment, root mean square deviations (RMSD) calculations of the NBD lobe rotation state were performed by calculating the RMSD values of Cα atoms in NBD-II. Next, trajectory frames were hierarchically clustered by their pairwise RMSD (using complete linkage) until two clusters remained. The centroid structures of each cluster were chosen as representatives. Rotational (screw) axes describing the relative lobe motion was calculated with CPPTRAJ in a two-step process: First, the representative cluster structures obtained before were aligned to the Cα atoms of NBD-I of a reference structure, that is, clusters of Hsp70-NBD-ATP-to-ADP trajectories were aligned to the crystal structures to Hsp70-NBD-ATP. Next, structures were aligned to NBD-II of the reference, extracting the corresponding rotational and translational matrices of that motion, from which the axes and angles of lobe rotation were subsequently calculated. An analogous procedure was applied to calculate the screw axes for all individual trajectory frames in order to yield the distribution of rotation angles during simulation.

## 3 | RESULTS AND DISCUSSION

We set out to predict protein residues responding to different ligand configurations in the Hsp70 proteins DnaK and BiP using the DNCF analysis,[49] which is based on evaluating the mutual information between the timelines of residue interactions in a protein structure network. First, we performed MD simulations of DnaK and BiP in different configurations, that is, bound to ADP, ATP, and the peptide substrate NRLLLTG (Table S2). The set of simulated systems includes Hsp70 in the ATP bound conformation ("Hsp70-ATP") and an in-silico modeled conformation bound to ATP and the NRLLLTG peptide ("Hsp70-ATP-pep"), which was chosen to approximate the substrate binding phase of the conformational cycle (Figure 1A). The process of substrate binding leading up to ATP hydrolysis involves an intermediate structure characterized by partial undocking of the NBD–SBD interface, which is not easily accessible to experimental methods of structure determination. A structure of DnaK-ATP-pep was reported recently, though only after the production of our simulations and our analyses had concluded, and no corresponding structure is currently available for BiP.[73] In the absence of this intermediate structure at the time of this work, we created variants based on the Hsp70-ATP structures, replacing ATP with ADP in silico ("Hsp70-ATP-to-ADP" and "Hsp70-ATP-to-ADP-pep"), which represents an artificial conformation for investigating the ability of protein residues to sense the ATP terminal phosphate. Each system was simulated for a total length of 1.2 μs, distributed over three independent runs of 400 ns each. The trajectories appear stable within expected variance as observed from the evolution of RMSD and force field energy terms (Figures S1–S6). A detailed discussion of these analyses can be found in Text S3.

From these trajectories, interaction timelines of carbon contacts and hydrogen bonds were extracted, transformed into networks and subsequently analyzed with SenseNet as described in our previous publication.[49] In these networks, residues are represented as nodes, which are connected by edges corresponding to residue–residue interactions. A residue pair can be connected by either a carbon contact interaction, a hydrogen bond interaction or both (using two separate edges). Furthermore, each interaction is associated with a timeline encoding the interactions state, that is, number of interactions, between two residues at different snapshots of the simulation. For example, a timeline of "1023" associated with a hydrogen bond between residues A and B indicates the presence of 1, 0, 2, and 3 hydrogen bonds at different timeslots of the simulation. Conformational correlation can then be measured by analyzing the correlation between timelines of different interactions. In the DNCF analysis,[49] this correlation is modeled by evaluating the mutual information between the timelines of neighboring interactions in the network (see Section 2). The DNCF score can be intuitively understood as the answer to the following question: Provided that we observe 0/1/2/3 or any larger number of hydrogen bonds (or carbon contacts) between residues A and B at a particular time frame of the simulation; does this influence the likelihood of observing a specific number of hydrogen bonds (or carbon contacts) in its close environment? If there is a correlation, we quantify the amount of shared information between interaction timelines. In the final step, the DNCF score calculates by how much this shared information between timelines (of its interactions) changes between two simulations. For example, the DNCF score of DnaK-ATP-pep (with DnaK-ATP as reference) indicates whether the shared information between specific interaction timelines changes due to the introduction of the peptide ligand. The DNCF score of residue A thus corresponds to the difference of shared information (in bits) between two simulations summed over all interactions (hydrogen bonds and carbon contacts) involving residue A.

By calculating the mutual information between interaction timelines of neighboring nodes in the network, residues with strong conformational coupling to their local environment can be predicted. In this context, the term "information" should not be understood as flow of bits through the protein but as contact transfer indicating routes of correlated conformations between adjacent residues. This information is summarized into a DNCF score for each residue, which corresponds to the change in shared information encoded in the interactions between two different system configurations. It is important to note that the mutual information between residues depends not only on the residues itself but on the rigidity or packing of the environment. For example, a binding event may rigidify a protein region and in turn alter the mutual information transfer and coupling of contacts between residues. Hence, it can open new routes or pathways of mutual interaction transfer between neighboring residues. Our DNCF analysis aims at predicting residues contributing to the Hsp70 allosteric network by assessing differences in the conformational coupling of residue interactions between different Hsp70 configurations, that is, either when bound to ATP, ADP, or the NRLLLTG peptide substrate.

After performing MD simulations for the selected Hsp70 configurations, hydrophobic contacts and hydrogen bonds were extracted from the trajectories and transformed into structure networks (Figure 1B). The layout of nodes in these networks was chosen to approximate the structural organization of the protein, which allows to inspect the interfaces between subdomains and trace pathways

**TABLE 1** Mann–Whitney-*U* tests for association of DNCF scores with a dataset of experimentally verified allosteric residues

| System | MWU *p*-value | rocAUC |
|---|---|---|
| DnaK-ATP-to-ADP | $1.26 \times 10^{-11}$ | 0.84 |
| DnaK-ATP-pep | $1.28 \times 10^{-11}$ | 0.84 |
| DnaK-ATP-to-ADP-pep | $4.09 \times 10^{-13}$ | 0.86 |
| BiP-ATP-to-ADP | $1.50 \times 10^{-7}$ | 0.76 |
| BiP-ATP-pep | $1.82 \times 10^{-10}$ | 0.82 |
| BiP-ATP-to-ADP-pep | $7.37 \times 10^{-12}$ | 0.84 |

Abbreviation: rocAUC, receiver operating characteristic.

between key regions. In the ATP bound form of DnaK, both NBD lobes form an interface to the SBD, but only NBD-I has direct contact to the SBDα lid region (Figure 1B–D). The same structural organization is observed in the network of BiP. Using Dijkstra's algorithm, we determined that depending on the system, a minimum of six to seven edges need to be traversed to reach the central leucine of the NRLLLTG peptide starting from the ATP/ADP node (Table S3). This indicates a substantial distance over which a conformational signal has to be transmitted between the nucleotide and the substrate peptide binding site located within the SBD. In order to predict which residues might respond to this process in DnaK, we first calculated DNCF scores for networks based on simulations of three different configurations, that is, DnaK-ATP-to-ADP, DnaK-ATP-pep and DnaK-ATP-to-ADP-pep. As reference configuration for the DNCF calculations, we chose the network generated from the trajectory of DnaK bound to ATP ("DnaK-ATP"). Using this setup, the DNCF scores of the DnaK-ATP-to-ADP system, using DnaK-ATP as a reference, are elevated by the conformational differences induced by the in-silico exchange of ATP to ADP. In order to evaluate the agreement of our predictions with experimental data, we compared the resulting DNCF scores (Table S4) to a set of experimentally verified residues associated with allosteric effects, as found predominantly in DnaK. This dataset is composed of residue positions for which mutations affected the coupling between binding of the polypeptide substrate, nucleotide binding at the NBD, ATP hydrolysis, and NBD–SBD docking[8,13,19,22,74–84] (Table S5). As there is no comparable dataset available for specifically "non-allosteric" residues in this system, for the purpose of evaluation we categorized all residues not present in the experimental dataset as "non-allosteric"; assuming that the majority of relevant allosteric residues are already known (in DnaK), the error induced by misallocating a presumably low number of unknown allosteric residues is expected to be limited. The vast majority of experimentally verified residues were determined in DnaK as the most frequently investigated Hsp70 representative, whereas available data for other Hsp70 variants was too limited to allow for quantitative validation. In addition, we were careful to exclude functional mutants with no clear relation to an allosteric effect, that is, a mutation that was more likely to influence ligand binding affinities than communication. We began our evaluation by observing the distribution of DNCF scores within the networks, finding that all systems diverged substantially from the hypothetical normal distribution, with a notable tendency toward a log-normal shape (Figure S7). Therefore, we used the nonparametric Mann–Whitney-*U* (MWU) test to evaluate whether known allosteric residues exhibited higher DNCF scores, and found a significant ($p < .01$) increase in all tested systems (Table 1). Next, the difference between these two groups was quantified using the area under the receiver operating characteristic curve (rocAUC). The DNCF scores of all DnaK systems achieved rocAUC values of ≥ 0.84, with DnaK-ATP-to-ADP-pep yielding the top rocAUC of 0.86 (Table 1, Figure S8). Substantial association of DNCF scores with the experimental set of allosteric residues is also observed for the corresponding BiP simulations, although rocAUC scores are reduced by 0.02–0.08. Intuitively, the rocAUC indicates the probability of a randomly selected allosteric residue having a higher DNCF score than a randomly selected non-allosteric residue; the observed rocAUC decrease in BiP systems thus corresponds to a lower probability of correctly ranked residue pairs by 2%–8%. This decrease might be caused by subtle differences between the allosteric networks of BiP and DnaK, as the latter was the primary source for the experimental dataset. Thus, the rocAUC rankings do not necessarily indicate a difference in prediction quality between the system, but rather reflect the biases of the experimental dataset. Nevertheless, DNCF scores of BiP systems are still strongly correlated with known allosteric residues in DnaK, as prediction performance remains much higher than for a random model (rocAUC = 0.5). Overall, due to the consistently strong agreement of DNCF scores with experimental data in all systems, we conclude that our analysis is able to detect known allosteric residues in DnaK/BiP, which are important for the conformational coupling between the nucleotide binding region and the substrate binding region. From this basis we proceeded to predict additional candidates with potential coupling function, particularly those which may fulfill specific roles in either protein. As DnaK/BiP-ATP-to-ADP-pep consistently showed the best agreement with experimental data, we chose to focus on these configurations for further in-depth analyses.

We next investigated the structural distribution of DNCF scores within Hsp70-ATP-to-ADP-pep networks (Figure 2). The highest scoring residues, that is, within the top 10% of the network distribution, were extracted (Table 2) and mapped to the protein structures (Figure 3). Beginning with the DnaK network, Figure 2A,C shows that high-scoring residues are organized into localized clusters. The majority of residues with high DNCF scores are located in proximity to the shortest network path between the ATP/ADP nucleotide and NRLLLTG peptide (Figure 3A,B). This aligns with the experimental dataset of allosteric residues (Table S5), which were primarily determined by investigating the coupling between ATP hydrolysis and peptide binding. An additional cluster extends from the direct NBD–SBD pathway into a separate region, close to the NBD–SBD linker region and the binding site of the J-Protein cochaperone DnaJ[85] (Figures 2A, 3B). The linker itself (residues 388–394) exhibits slightly higher than average DNCF scores, less than expected considering the linker's well-established importance for controlling the NBD–SBD docking dynamics.[18,56] However, several of the adjacent high scoring residues are involved in the interface between DnaK and DnaJ (Figures 2A and

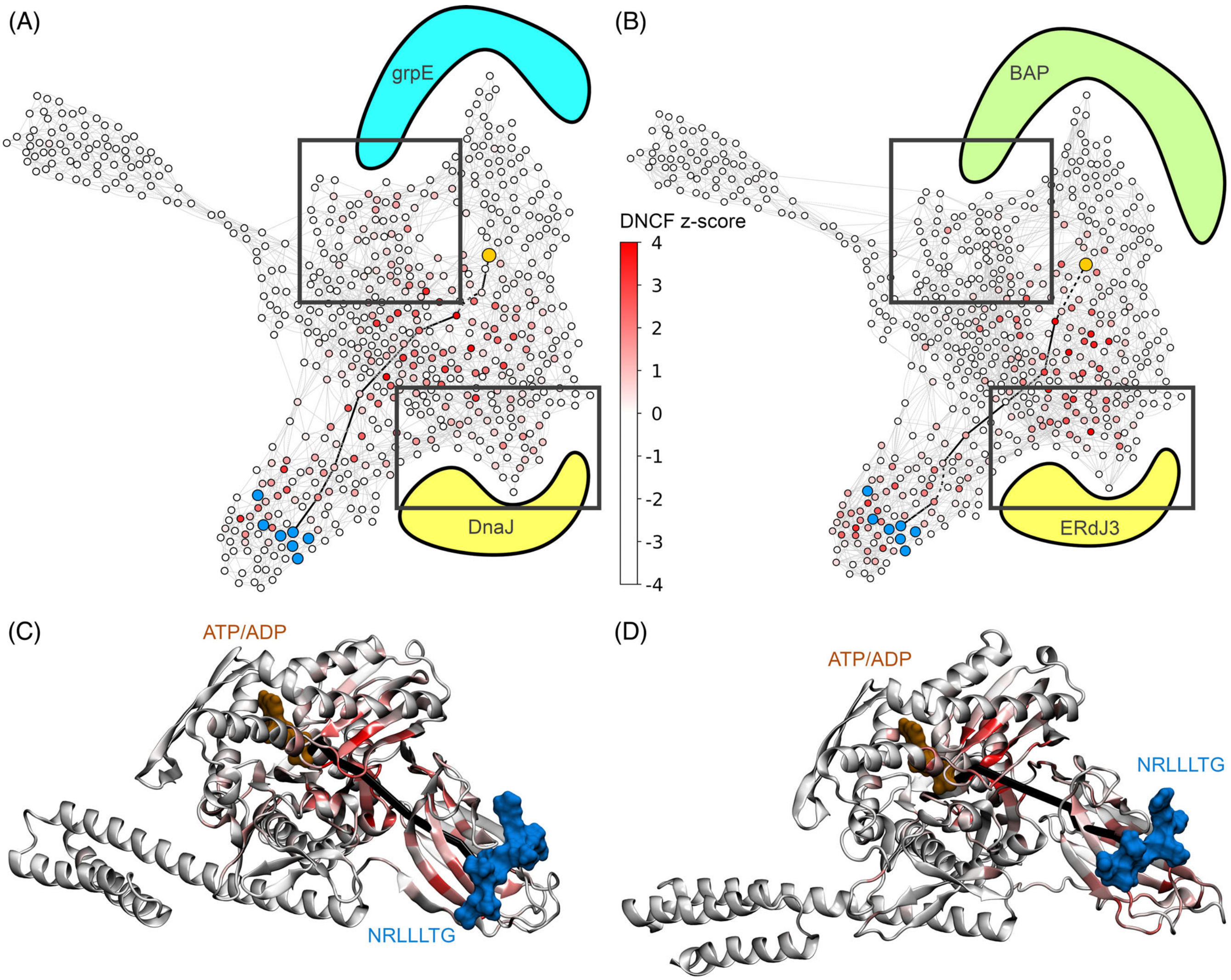


**FIGURE 2** Structures and residue interaction network of heat shock protein 70 kDa proteins. (A,B) Protein structure networks obtained from molecular dynamics simulations of DnaK-ATP-to-ADP-pep (A) and BiP-ATP-to-ADP-pep (B). Nodes are colored according to the *z*-score normalized DNCF scores of their associated residues. Cochaperones DnaJ and grpE are indicated as colored shapes to visualize the location of their DnaK binding sites as observed from PDB structures (PDB-IDs: 5NRO, 1DKG), but were not present during simulations. Corresponding locations for the BiP cochaperones are estimated by homology: BAP from yeast Sil1 (PDB-ID: 3QML) and ERdJ3 from *Escherichia coli* DnaJ (PDB-ID: 5NRO). (C,D) Representative structures extracted from molecular dynamics simulations of DnaK-ATP-to-ADP-pep (C) and BiP-ATP-to-ADP-pep (D) with residues colored according to their *z*-score normalized DNCF scores

3B), which plays a substantial role in initiating the undocking of the NBD–SBD domain preceding ATP hydrolysis.[16,85] The observation that residues surrounding the linker have higher scores than the linker itself suggests that DnaJ binding may trigger a cascade of conformational changes involving residues such as R167, I168, I169, I207, K214, T395, and D481 (Figure 3B), leading to subsequent unbinding of the actual linker residues. Residues R167, I168, I169, and D481 are known to affect ATP hydrolysis and/or its stimulation by DnaJ,[19] while I207 was found to co-evolve strongly with SBD residues.[86] During the preparation of this manuscript, a structure of DnaK-ATP-pep was published in the suggested allosterically active conformation,[73] which is characterized by partial undocking of NBD and SBD domains. The conformational differences, compared to previous crystal structures of full length DnaK, were found to be concentrated in the protein region between residues 220 and 231.[73] This corroborates with a large cluster in our predictions, namely T221, N222, T225, H226, L227, and D231 (Figure 3A,B). We expect that it should be interesting to include MD simulations based on this conformation in future analyses, provided the corresponding structure can be obtained for BiP. In summary, we were able to find several localized clusters of predicted allosteric residues in DnaK, of which a substantial number are supported by previously established experimental evidence.

In addition to clusters characterized by distinct structural regions, DNCF scores also show a tendency to cluster within the protein sequence (Figure 4A). Both the SBDβ and NBD domains (including the NBD-I and NBD-II lobes) contribute high DNCF scores (Figure S9), and only the SBDα domain appears to lack any pronounced residues. Analyzing the localization of high scoring residues in more detail, we

TABLE 2 Prediction of residues which contribute to the coupling between substrate and nucleotide binding in DnaK/BiP according to DNCF scores

| System | Residues |
|---|---|
| DnaK-ATP-to-ADP-pep | L9 A58 F67 **K70** R71 R75 E81 R84 T141 V142 **P143** A144 **Y145 F146 N147 D148 R151** I168 I169 **N170 E171** P172 Y193 L195 **T199** F200 **D201 I202** I207 K214 **T221** N222 T225 H226 **L227** D231 L320 V340 T395 L397 L399 I401 T437 I438 Q442 L454 Q456 F457 N458 L459 I472 F476 **D481** D490 Q497 |
| BiP-ATP-to-ADP-pep | L9 F42 **K70** R71 R75 K100 V142 A144 **Y145** I169 **N170 E171 T173** I177 Y179 G180 L181 R184 F193 **D194** L195 **T199** F200 **D201 V202** L205 I207 F216 **T221** N222 T225 **L227** E230 F232 V340 Q384B T386 **L389 L391 D393** L399 I401 E402 M408 L411 T428 T435 V436 I438 E442 F457 L459 T460 I462 P470 Q471 I472 F476 D490 T493 N497 K498 |

*Note*: Residue positions corroborated by the set of experimentally verified residues are marked in bold font.

find particular enrichment at the subdomain interfaces between NBD-I, NBD-II, and SBDβ, while residues in the protein core and the NBD–SBDα interface trend toward lower scores (Figure S10). These observations suggest that the extensive subdomain interfaces formed in DnaK play a key role for conformational control, potentially by modulating residue packing and flexibility (e.g., NBD lobes) and changes in the equilibrium of interdomain binding (e.g., NBD–SBD docking). Comparing the DNCF score distributions between DnaK and BiP, we find that the trends for the NBD–SBD interface and its linker region are highly similar between these related proteins (Figures 2B,D, 3C,D, 4B, S11, and S12). However, while the rough structural locations of allosteric regions seemed well preserved in general, the sets of predicted residues in the top 10% percentile diverge substantially (Table 2), and in addition include residues which are unique to either protein variant, such as K214 in DnaK and Q384b in BiP. Therefore, we next focused our efforts on investigating the differences between the predicted sets of allosteric residues for DnaK and BiP.

Having observed similar DNCF score distributions between DnaK and BiP simulations, we set out to determine which residues were shared between both proteins or specific to either protein variant. The correlation between DNCF scores of DnaK-ATP-to-ADP-pep and BiP-ATP-to-ADP-pep (Figure 4C) is lower than the average between different configurations of the same protein (Spearman's $r$: 0.74 vs. average of simulations involving BiP: 0.87 ± 0.02 or DnaK: 0.9 ± 0.02) (Table 3), which suggests systematic differences between DnaK and BiP. Based on these differences, we created residue sets of predictions specific to each Hsp70 variant, that is, likely to contribute to allosteric signaling in one system but not in the other. For this, we selected residues which were specific to DnaK or BiP, that is, residues which were (i) within the top 15% of the DNCF scores in DnaK-ATP-to-ADP-pep as well as (ii) concurrently in the lower 15% of the log-normal DNCF score distribution estimated for the experimentally determined allosteric set in BiP-ATP-to-ADP-pep, and vice versa. In addition, we selected the residues which occurred in the top 10% of both systems as the "common" set of conserved allosteric residues (Table 4). The regions containing conserved allosteric residues (Figure 5) resemble the clusters of top scoring residues detected before (Figure 3). Out of the 30 allosteric candidates predicted specifically in either DnaK or BiP, 13 are related to amino acid mutations or insertions (Table 4). Residue positions with specific differences between DnaK and BiP are found in several regions: The first cluster, which is specific to DnaK, (Figure 5A,B) contains residues which contribute to the NBD–SBDβ interface (N147, D148, Q150, D481). Mutational studies have shown that these residues are important for stabilizing the NBD–SBDβ interface as well as allosteric signaling in DnaK.[19] The fact that these residue positions do not feature as prominently in BiP in our predictions suggests diminished dynamics at these locations compared to DnaK. This interpretation is backed by experimental data: Introduction of the D481N point mutation in DnaK, which is the wild type residue variant for BiP, is capable of disturbing the equilibrium of docked–undocked conformations at the NBD–SBDβ interface.[16] A similar trend toward rigidification of the same interface has been reported in multiple instances for BiP compared to DnaK.[24,31,87] In DnaK, D148 contacts the SBD via Q442 and is an essential residue for communication of the peptide binding signal from the SBD to the NBD in DnaK.[19] However, in BiP the corresponding position on the SBD side harbors a negatively charged residue (Q442E), creating electrostatic repulsion to D148. In combination, these data point toward substantial changes in the interaction pattern of the N147N-D148D-Q150Q-D481N-Q442E cluster between DnaK and BiP, which may explain differences in the dynamics of the NBD–SBD interface and allosteric communication. The second cluster is composed of residues specific to either DnaK or BiP (Figure 5B,D) and is found in the vicinity of the NBD–SBD linker (G180G, G184R, K214–, –[384b]Q, V386T, T395C), indicating another potential key region for differential regulation of NBD–SBDβ docking in these two protein variants. As it is this region that binds the J-domain in DnaK,[85] the characteristic domain shared between BiP's ERdJ1–ERdJ7 cochaperone families,[31] it appears likely that differences in the linker environment reflect evolutionary specialization to different sets of cochaperones. Another cluster, which is specific to BiP, is formed by residues located in the SBD loops (T428T, E430S, A435T, P470P, K491K, S493T, G494G, K498K, I501I). Residues 428 and 430 are part of the SBDβ's L1,2 loop and Residues 491–501 are part of the β8 sheet, two structural elements which have been shown to assume multiple distinct conformations in BiP.[88] In our data, these residues showed high DNCF scores exclusively for BiP, which might indicate increased conformational flexibility of the SBD in BiP compared to DnaK. Finally, positions 61, 62, and 65 are located close to the binding interface of NEF grpE (Figures 2A and S13) and one can speculate that these residues may be utilized to facilitate opening of the NBD loops. It is not surprising that this

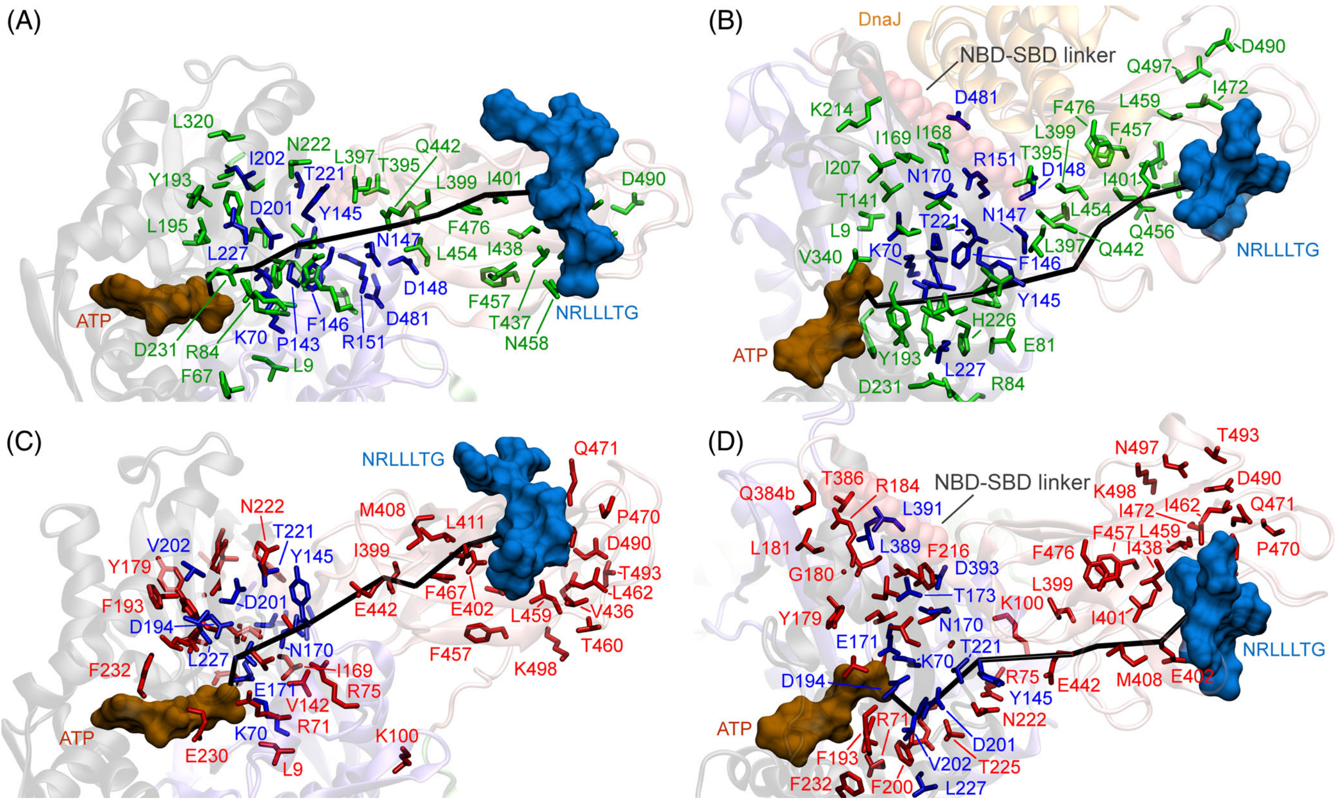


FIGURE 3 Residues showing strong conformational coupling with the nucleotide and peptide substrate according to their DNCF scores. Residues marked in blue are part of both the predicted set and the set of experimentally verified residues. The shortest inter-residue path between ADP and the central leucine of the NRLLLTG peptide is shown in black. The nucleotide-binding domain–substrate binding domain linker is indicated as red spheres. (A,B) Residues within the top 10% of DNCF scores mapped onto DnaK (blue, green). The location of the J-domain of the DnaJ cochaperone in the complex (PDB-ID: 5NRO) is indicated as an orange cartoon, though it was not present during simulation. (C,D) Residues within the top 10% of DNCF scores mapped onto BiP (blue, red)

region lacks high-scoring residues in our BiP simulations (Figure 2B), as NEFs of eukaryotic organisms including BiP (e.g., BAP[89,90]) are thought to have evolved independently from grpE and might thus utilize a different mechanism.[91]

Communication of the peptide binding signal through the protein to induce ATP hydrolysis is an essential step in the conformational cycle of Hsp70, which aligns with our observations of predicted allosteric residues clustering along the shortest path between the nucleotide binding site in the NBD and peptide binding pocket in SBDβ. However, the DNCF predictions, in isolation, account only for the conformational influence of individual residues, particularly those close to the nucleotide and substrate binding sites as these adapt fastest to the different configurations probed in our simulations, that is, in-silico exchange of ATP to ADP and the NRLLLTG peptide. Thus, conformationally coupled residues located in the intermediate region between the NBD and SBD may be overlooked within the limited ns–μs timescale of our simulations, which is much shorter than the estimated ms–s timescale characteristic for processes within the Hsp70 conformational cycle.[15,84] To address this problem, we set out to combine our predictions of individual residues into a chain of conformationally coupled residues, focusing specifically on the process of protein activation triggered by binding of the peptide substrate. We chose to perform our analyses on the Hsp70-ATP-pep systems as the configuration representing the closest approximation to that step of the conformational cycle. Starting from the node representing the central leucine of the NRLLLTG peptide in the network, we performed a weighted random walk traversing edges until the ATP node was reached, while keeping track of the visited nodes. The probability of jumping from one node to a neighboring node was chosen to be proportional to their relative DNCF scores, such that a node with twice the score than its alternative was two times as likely to be chosen for the next step (see Section 2). The procedure was then repeated after interchanging source and target nodes, that is, starting from ATP and finishing at the central leucine of the peptide. These runs, both in the forward and backward direction, were performed 10 000 times each and summed to yield the final result. This approach combines the advantages of two strategies: First, the DNCF method provides information about the conformational coupling of individual residues to their environment, and how this coupling is affected by different ligand binding states. Then, this information is supplemented with a search for the shortest paths connecting two regions, that is, nucleotide and peptide binding pockets, a technique that serves as the foundation for the class of centrality-based methods to predict functional residues in

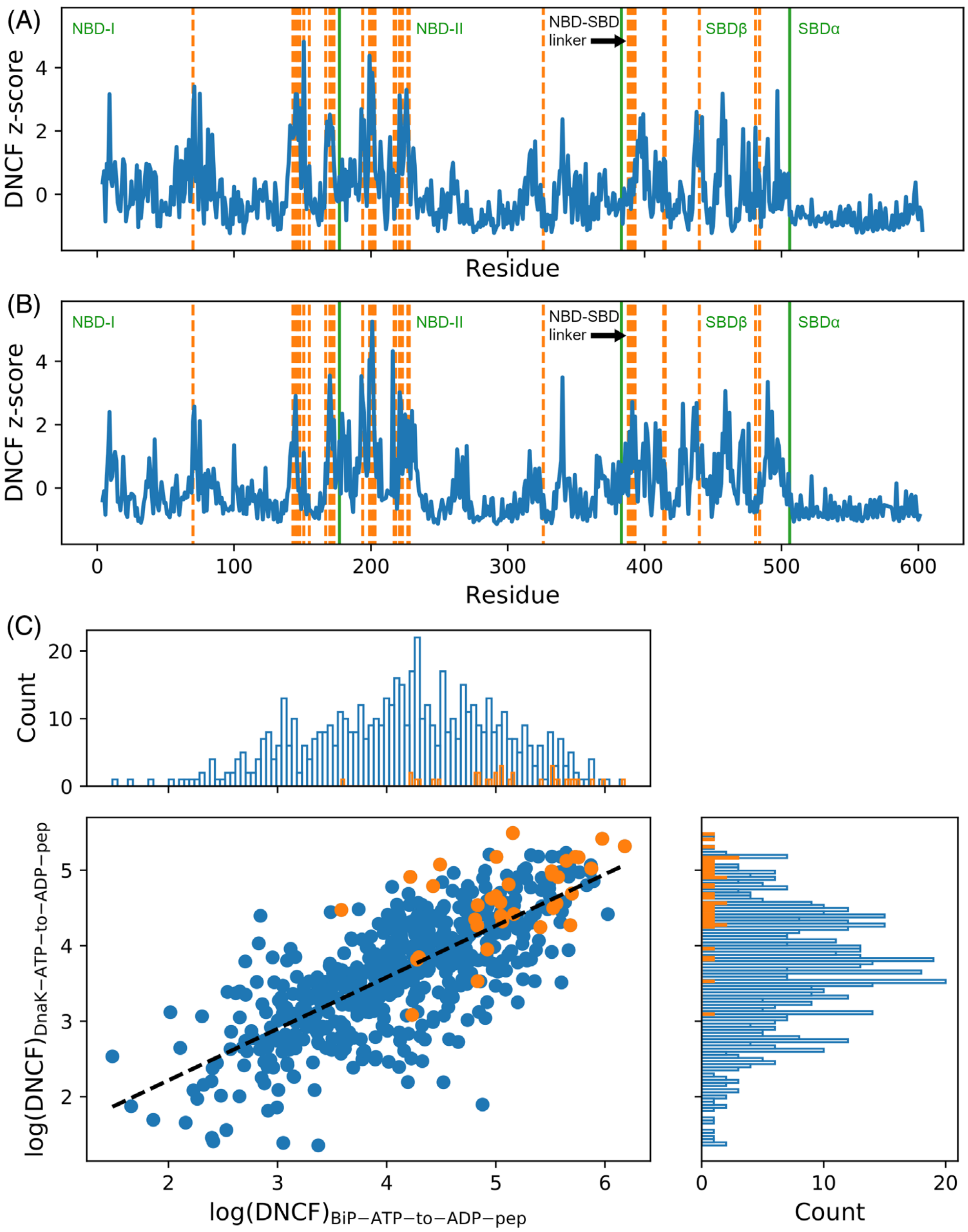


**FIGURE 4** Correlation of DNCF scores obtained from molecular dynamics simulations of DnaK and BiP. Residues with experimentally verified allosteric roles are shown in orange. (A,B) Normalized DNCF scores of (A) DnaK-ATP-to-ADP-pep and (B) BiP-ATP-to-ADP-pep plotted over the protein sequence. (C) Scatterplot showing the correlation between DNCF scores of DnaK-ATP-to-ADP-pep and BiP-ATP-to-ADP-pep. NBD, nucleotide-binding domain; SBD, substrate binding domain.

**TABLE 3** Spearman's correlation coefficients for DNCF scores obtained from network analysis of molecular dynamics trajectories

| | DnaK-ATP-to-ADP-pep | DnaK-ATP-pep | DnaK-ATP-to-ADP | BiP-ATP-to-ADP-pep | BiP-ATP-pep |
|---|---|---|---|---|---|
| DnaK-ATP-to-ADP-pep | | | | | |
| DnaK-ATP-pep | 0.93 | | | | |
| DnaK-ATP-to-ADP | 0.89 | 0.88 | | | |
| BiP-ATP-to-ADP-pep | 0.74 | 0.72 | 0.61 | | |
| BiP-ATP-pep | 0.76 | 0.74 | 0.65 | 0.89 | |
| BiP-ATP-to-ADP | 0.68 | 0.62 | 0.59 | 0.85 | 0.87 |

*Note*: System combinations featuring the same protein are highlighted in green.

proteins.[92–94] Our combined approach yields a score that accounts both for the conformational coupling of individual residues and their interconnectivity, that is, their closeness to the regions of interest and other residues with high DNCF scores. Figure 6 shows the signaling pathways predicted by this score, denoted as DNCF-RW ("DNCF Random Walk"; raw scores reported in Table S6).

**TABLE 4** Prediction of residues with specific signaling properties to either DnaK or BiP according to DNCF and DNCF-RW scores

| System | Method | Residues |
|---|---|---|
| DnaK/BiP common[a] | DNCF | L9L, K70K, R71R, R75R, V142V, A144A, Y145Y, I169I, N170N, E171E, Y193F, L195L, T199T, F200F, D201D, I202V, I207I, T221T, N222N, T225T, L227L, V340V, L399L, I401I, I438I, **Q442E,** F457F, L459L, I472I, F476F, D490D, **Q497N** |
| BiP[b] | DNCF | **-(106a)I**, G180G, **G184R**, E267E, K270K, **V365I**, –**(384b)Q**, **V386T**, T428T, **E430S**, **A435T**, N451N, **D460T**, P470P, K491K, **S493T**, G494G, K498K, I501I |
| DnaK[b] | DNCF | N61N, P62P, T65T, N147N, D148D, Q150Q, **T185E**, **K214-**, F357F, **T395C**, **D481N** |
| DnaK/BiP common[a] | DNCF-RW | K70K, R71R, I73I, R75R, P143P, A144A, Y145Y, F146F, N147N, D148D, Q150Q, R151R, N170N, L195L, G196G, G198G, T199T, F200F, D201D, T225T, L227L, E230E, L397L, **S398T,** L399L, I401I, E402E, M408M, L411L, F426F, V436V, T437T, I438I, V440V, L441Y, **Q442E,** L454L, **Q456T,** F457F, **N458D,** L459L, I472I, V474V, F476F, L484L |
| BiP[b] | DNCF-RW | P37P, **P113E**, D156D, E430S, –**(506a)R**, L507L |
| DnaK[b] | DNCF-RW | **R84Q** |

*Note*: Residue codes at the beginning/end mark the DnaK/BiP sequence variants, respectively. Missing residues are indicated by a dash and insertion codes by lower case letters and parentheses. Residue positions differing between DnaK and BiP are highlighted in bold.

[a]Residues with increased scores in both systems.

[b]Residues with increased scores only in the denoted system.

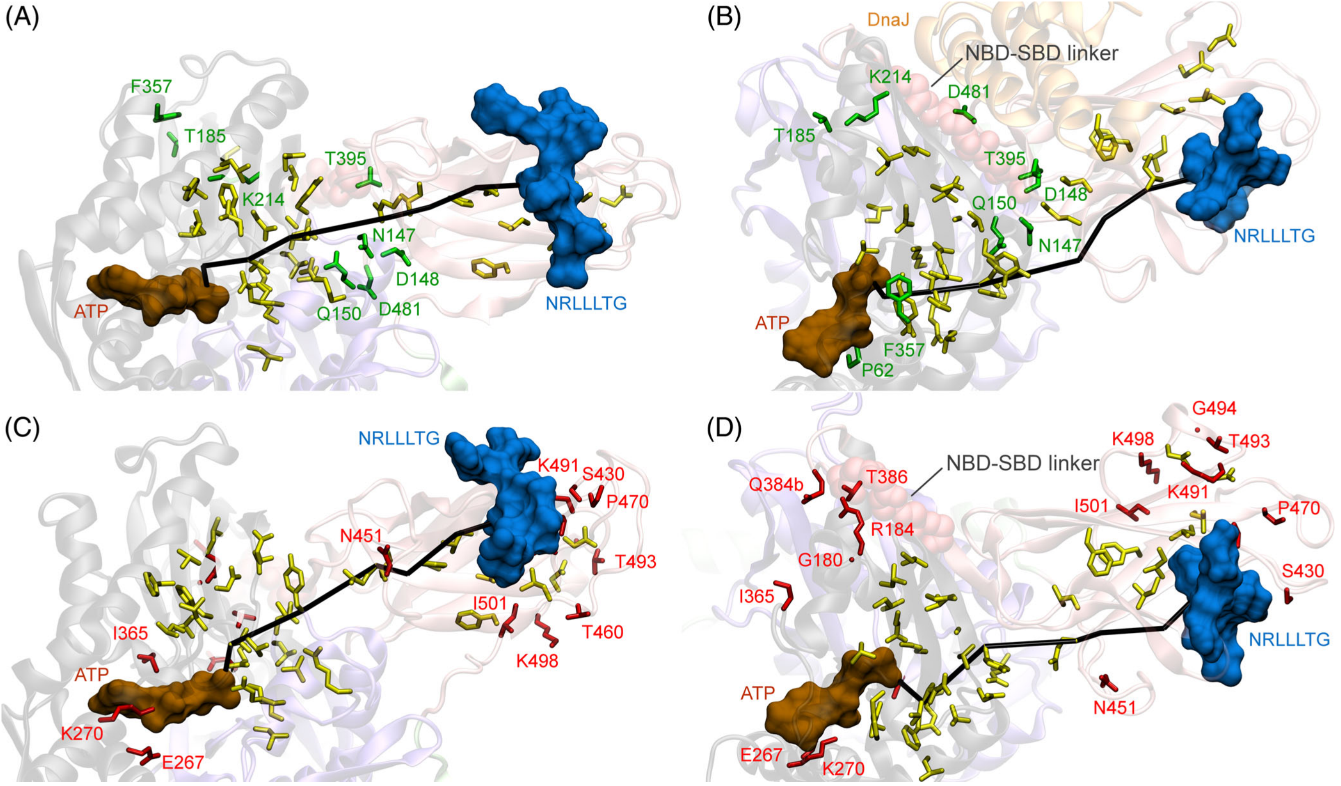


**FIGURE 5** Residues showing strong conformational coupling with the nucleotide and peptide substrate specific to either DnaK or BiP as predicted by their DNCF scores. Yellow residues mark residues with increased scores in both proteins. The shortest inter-residue path between ADP and the central leucine of the NRLLLTG peptide is shown in black. The nucleotide-binding domain–substrate binding domain(NBD–SBD) linker is indicated as red spheres. (A,B) Residues with specifically increased DNCF scores in DnaK-ATP-to-ADP-pep compared to BiP (green). The location of the DnaJ cochaperone in the complex (PDB-ID: 5NRO) is indicated as an orange cartoon, though it was not present during simulation. (C,D) Residues with specifically increased DNCF scores in BiP-ATP-to-ADP-pep compared to DnaK (red)

As described above, we selected sets of shared and specific residues defined by the highest and lowest score percentiles, that is, residues which are both within the top 15% of the evaluated score in one system and within the lower 15% percentile of distribution obtained from the set of experimentally verified allosteric residues in the other system. The set of residues common to both systems consists of

**TABLE 6** Predicted residues with specific roles in nucleotide-binding domain (NBD) lobe rotation compared to the full-length protein

| System | Method | Residues |
|---|---|---|
| BiP-NBD | DNCF | V4, N64, A69 |
| DnaK-NBD | DNCF | G229, D233, E267, K270 |
| BiP-NBD/full length common | DNCF | L9, F42, K70, R71, R75, E171, Y179, L181, F193, D194, L195, T199, F200, D201, V202, L205, I207, F216, T225, L227, E230, F232, V340 |
| DnaK-NBD/full length common | DNCF | L9, F67, K70, R71, R75, E81, R84, T141, Y145, Y193, L195, T199, F200, D201, I207, H226, L227, V340 |

45 residues, describing a contiguous surface of conformationally coupled residues between the nucleotide and substrate binding pockets (Table 4, Figure 6). R84 is the single residue found to contribute specifically to DnaK (Figure 6A,B), with the difference in scores arising from the R84Q mutation present in BiP. In contrast, the network of BiP shows a specific cluster of residues with increased scores close to the R(−506a) insertion. This residue insertion does not occur in DnaK, but is highly conserved in eukaryotes[25] and is a prominent interaction partner forming hydrogen bonds with D148, Q152, and D156 at the interface between the SBDβ core, SBDα, and NBD domains (Figure 6C,D). The R(-506a) residue has also been found to adapt to the binding of peptide substrates[28] and plays an important role in stabilizing the docking of NBD–SBD domains.[25] The cluster furthermore consists of residues P37P, P113E, D156D, and L507L,

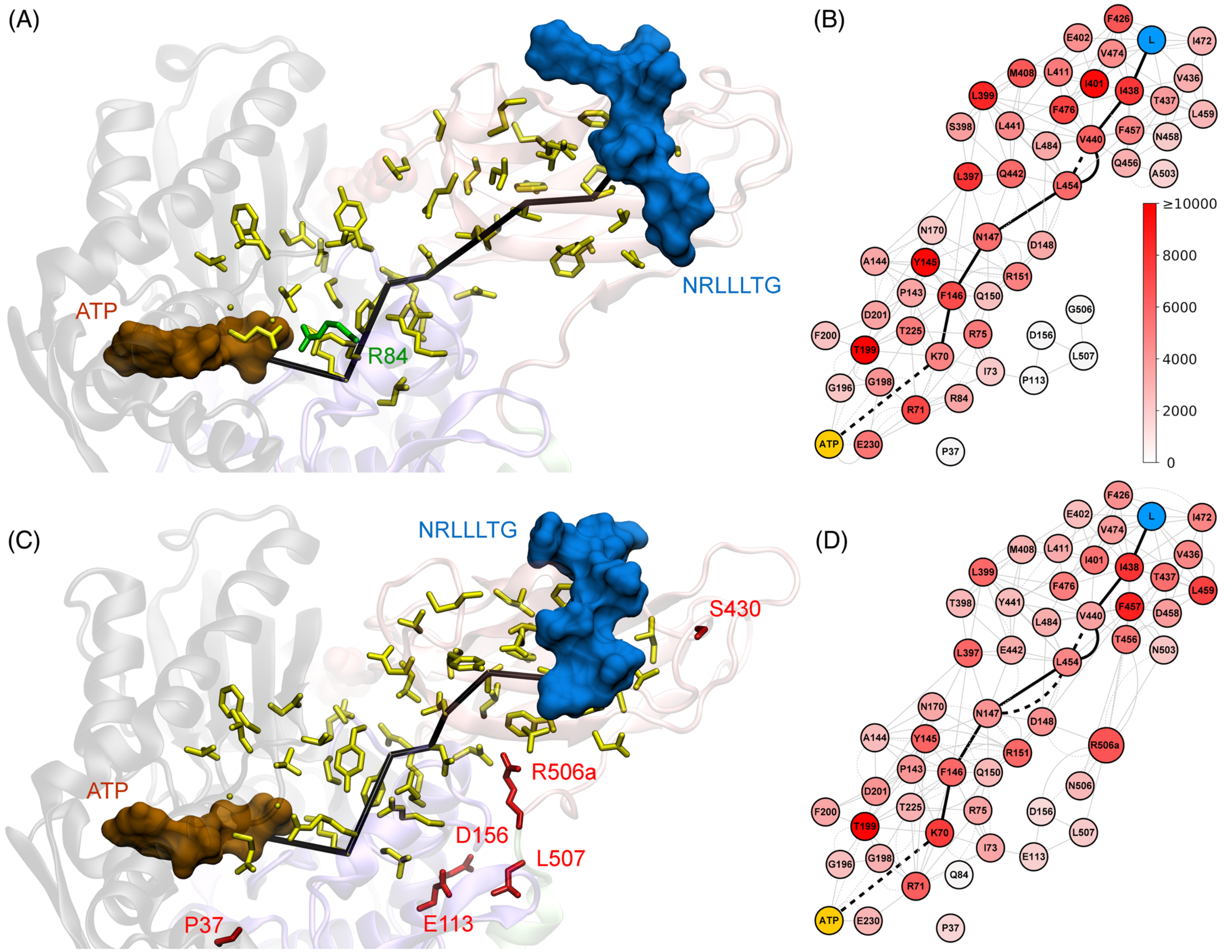


**FIGURE 6** Cluster of conformationally coupled residues between the substrate and nucleotide binding sites in heat shock protein 70 kDa predicted by a targeted random walk weighted by DNCF scores (DNCF-RW). Node colors in the networks indicate the number of times each node was visited during the DNCF-RW random walk. Stick representations mark residues predicted as specific to DnaK (green), specific to BiP (red) or shared between DnaK and BiP (yellow). The shortest pathway between ADP and the NRLLLTG peptide is indicated in black. Within networks, solid edges denote carbon contacts and dashed edges indicate hydrogen bonds. (A,B) Structure and network of DnaK-ATP-pep. (C,D) Structure and network of BiP-ATP-pep

which together form an alternative pathway between NBD and SBD in addition to the main pathway present in both DnaK and BiP. In summary, our observations using the DNCF and the DNCF-RW methods suggest a high degree of conservation between DnaK and BiP with respect to the major conformationally coupled regions located between the binding pockets of the nucleotide and peptide substrate. However, differences were found in the marginal regions of the clusters, which could be the result of adaptation to their specific organism or organelle contexts. In particular, we detected three regions showing differences, which might be the result of evolutionary adaptation in DnaK and BiP: First, the NBD–SBD linker and its surrounding residues; second, the peptide binding pocket; and third, the linker region connecting the SBDβ with the SBDα Lid domain via the -(506a)R residue insert.

After investigating substrate dependent signaling in Hsp70s, we were interested in addressing the conformational effects of ATP hydrolysis on NBD lobe dynamics, particularly whether this process was controlled by the same set of residues as determined above. The partial undocking of NBD and SBD after binding a substrate peptide is the prerequisite step to prime the NBD for ATP hydrolysis. Following cleavage of the nucleotide's terminal phosphate, the NBD lobe subdomains rotate by ~17.5°, as observed by comparing ATP-bound crystal structure and ADP- bound crystal structure of DnaK (only NBD domain; Figure S14).[84,95] In order to study the rotation of the NBD lobes, we performed simulations of the isolated DnaK/BiP NBD as an approximation to the undocked SBD state, in which both NBD and SBD are separated and only connected by the NBD–SBD linker. Three nucleotide configurations were simulated, namely Hsp70-NBD-ATP, Hsp70-NBD-ADP, and Hsp70-NBD-ATP-to-ADP. Using the RMSD of NBD-II to track the rotation state of simulations, we observed that the NBD lobes of DnaK-ATP-to-ADP rotated toward the ADP conformer in 2 out of 3 simulation replicas, while they did not rotate substantially in DnaK-ATP (Figure 7A,B). To determine whether the axis of rotation observed in the simulation matched what was expected from the crystal structures, we conducted a hierarchical clustering of trajectory frames on NBD-II until only two clusters remained. The first cluster corresponded to trajectory frames close to the initial structure, while structures showing substantial rotation to come closer to the

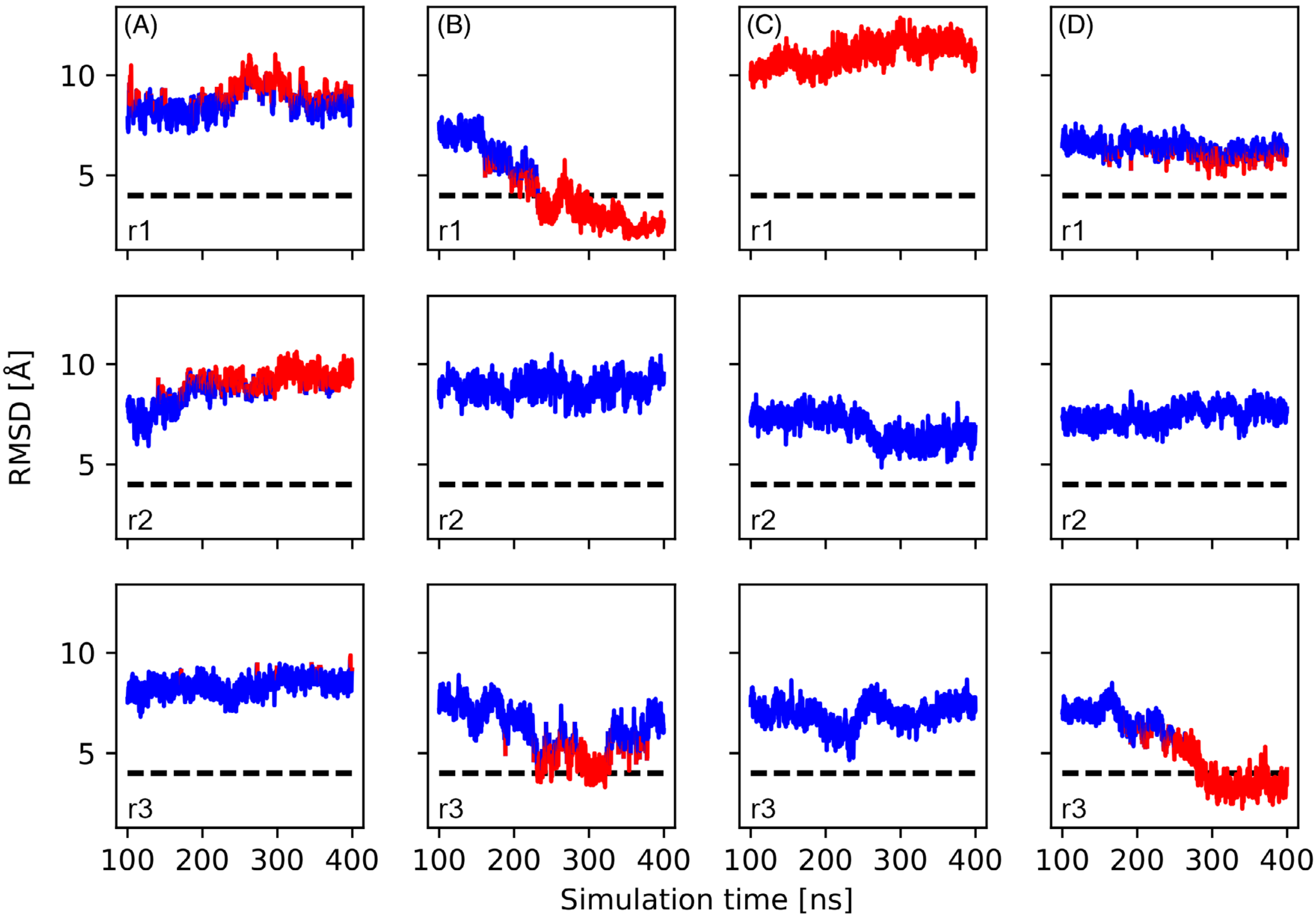


**FIGURE 7** Root mean square deviations (RMSD) of nucleotide-binding domain-II (NBD-II) during molecular dynamics simulations of the isolated NBD domain of DnaK/BiP. The lobe conformation associated with bound ADP was used as reference for RMSD calculation. Trajectory frames shown in blue and red indicate membership to the two top clusters remaining after hierarchical clustering. Each column shows values obtained from replicas r1–r3 for the different systems: (A) DnaK-NBD-ATP; (B) DnaK-NBD-ATP-to-ADP; (C) BiP-NBD-ATP; (D) BiP-NBD-ATP-to-ADP

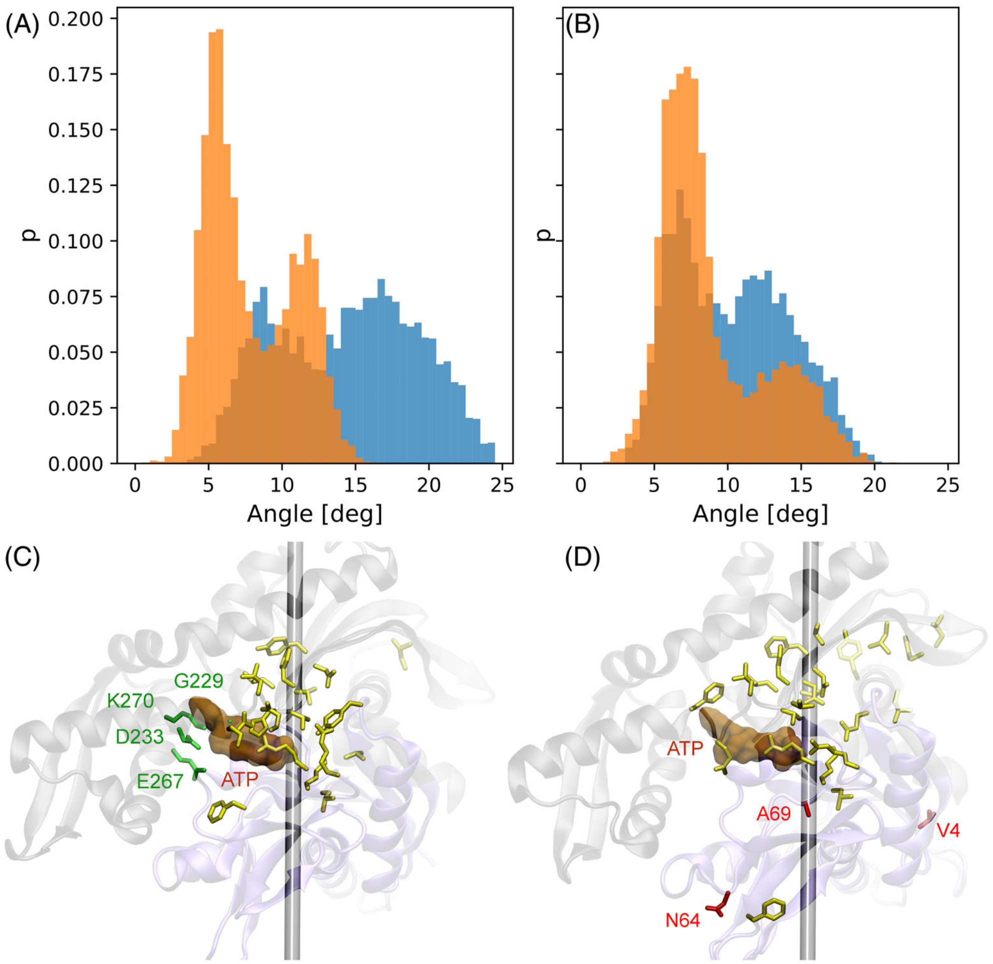


**FIGURE 8** Simulation of nucleotide-binding domain (NBD) lobe rotation using molecular dynamic simulations of DnaK and BiP. (A, B) Histograms of NBD lobe rotation angles obtained from molecular dynamics trajectories of (A) BiP-NBD and (B) DnaK-NBD. The orange histograms show Hsp70-NBD-ATP while the blue histograms show the Hsp70-NBD-ATP-to-ADP variant. (C,D) Residues with specifically increased DNCF scores in the trajectories of NBD versus full length proteins in (C) DnaK (green) and (D) BiP (red)

**TABLE 5** Rotation of nucleotide-binding domain (NBD) lobes during molecular dynamics simulations

| System | Angle mean[a] (degree) | Angle standard deviation[a] (degree) | Angle to reference axis[b] (degree) |
|---|---|---|---|
| DnaK-NBD-ATP | 4.34 | 2.05 | 78.40 |
| DnaK-NBD-ATP-to-ADP | 14.4 | 4.83 | 18.40 |
| BiP-NBD-ATP | 7.97 | 2.36 | 83.33 |
| BiP-NBD-ATP-to-ADP | 10.4 | 3.79 | 26.12 |

[a] Mean and standard deviation of the NBD lobe rotation angle compared to the starting conformation during molecular dynamics.

[b] Angle between the principal rotation axis of the simulation and the reference rotation axis obtained from the ATP-bound crystal structure and ADP-bound crystal structure.

DnaK-NBD-ADP conformation formed the second cluster (Figure 7A,B). The average NBD rotation during simulations, compared to the starting structure, was 14.4 ± 4.83° for DnaK-NBD-ATP-to-ADP and only 4.34 ± 2.05° for DnaK-NBD-ATP, indicating a clear increase in conformational dynamics (Figure 8A). We determined the median representative of the second cluster and calculated the rotation axis of NBD lobes compared to the initial structure as the "principal rotation" of the simulation. As reference, we calculated the rotation axis from the crystal structures of DnaK-NBD-ATP and DnaK-NBD-ADP. The principal rotation axis of DnaK-NBD-ATP-to-ADP was shifted compared to the reference axis by 18.40°, showing a much closer alignment to the reference axis than the shift of 78.4° of the DnaK-NBD-ATP system (Table 5). We performed the same analysis for BiP-NBD-ATP-to-ADP and BiP-NBD-ATP, revealing similar trends: The NBD lobes of BiP-NBD-ATP-to-ADP rotated by 10.4 ± 3.79°, while in BiP-NBD-ATP they rotated by 7.97 ± 2.36° (Figure 8B). Again, the principal rotation axis of BiP-NBD-ATP-to-ADP was much closer to the reference axis, with a shift of only 26.12°, compared to BiP-NBD-ATP with a shift of 83.33° (Table 5). This reduction in the average rotation angle of BiP compared to DnaK is explained by observing that only one out of three BiP-NBD-ATP-to-ADP replicas showed a substantial rotation toward the ADP state (Figure 7C,D). In summary, the in-silico exchange of ATP to ADP was sufficient to trigger rotation of NBD lobes toward the expected conformers in some of our simulations. Note that we do not use this data to make a quantitative prediction of the propensity of this rotation in DnaK and BiP proteins, for which the analyzed number of replicas is too small. Instead, we utilize these simulations qualitatively to gain insight into the rough timescale on which this rotation can occur and detect candidates for residues which are specifically associated with NBD lobe rotation.

Next, we extracted interaction networks from the simulations and applied the DNCF method to determine changes in residue dynamics. Comparing the DNCF scores of DnaK-NBD-ATP-to-ADP to its corresponding full-length protein, we observe solid but not perfect correlation (Spearman's *r*: 0.82). We repeated the analyses detailed above to determine residues with specifically increased DNCF scores comparing the NBD and full-length protein simulations. Total 18 residues are found to be shared in the top 10% DNCF percentiles of DnaK-NBD-ATP-to-ADP and its full-length variant, whereas 23 residues are shared for the corresponding BiP systems (Table 6). Four residues close to the nucleotide were detected with high scores specifically in the NBD simulations, with three of them acting as direct interaction partners (G229, E267, and K270) and one located in the close vicinity (D233) (Figure 8C). This location puts them in a prime position to sense the nucleotide and provide flexibility to the NBD lobes depending on DnaK's ATP/ADP state.[58,96] In addition, G229 is located adjacent to G228, another glycine for which mutations exhibit defective chaperone function.[97] In BiP-NBD, residues V4, N64, and A69 were detected as specific for NBD lobe rotation (Figure 8D). A69 is adjacent to K70, a residue essential to ATP hydrolysis in Hsp70s,[75] indicating an association of NBD lobe rotation with conformational changes close to residues regulating catalysis. In contrast, the other two residues V4 and N64 are located far away from the axis of lobe rotation. It is possible that these observations are influenced by statistical noise, as only one out of three simulations of BiP-NBD showed NBD lobe rotation (Figure S14). Overall, the signaling properties of the isolated NBD domains appear to be very similar to the full-length protein, however with a number of residues arising with potentially specific functions for the rotational motion.

## 4 | CONCLUSION

In this study, we performed MD simulations of the Hsp70 chaperones DnaK and BiP, extracted networks of hydrophobic and hydrogen bond interactions and performed DNCF and DNCF-RW analyses to predict residues exchanging information about their conformational states with their environment, prompted by different ligand configurations. These residues are presumed to be associated with allosteric pathways of the Hsp70 system, that is, residues for which mutation has a notable effect on the coupling between the processes of peptide substrate binding, cochaperone-mediated activation and ATP hydrolysis. Our predictions based on the DNCF method were found to be in quantitative agreement with a set of experimentally verified allosteric residues. As the experimental dataset is limited by the number of tested mutants and reliance on DnaK as the predominant model, our predictions can aid by potentially filling gaps in our understanding of Hsp70 allostery and by pinpointing signaling differences between related protein variants, such as between DnaK and BiP. The strong agreement with experimental data further indicates that the artificial Hsp70 conformations constructed for our analysis are able to provide useful insights, despite reflecting only a part of the complete biological picture. As more and more structures of different states within the Hsp70 conformational cycle become available, like the recent publication of the partially undocked DnaK-ATP-pep conformation,[73] further MD simulations based on these novel structures will be useful for further refinement of analyses. The structures we investigated in this work—and thus the pathways we predict—correspond to one specific phase of the conformational cycle, namely the substrate mediated activation of ATP hydrolysis and subsequent undocking of the NBD–SBD interface. All simulated systems are conformationally related to the Hsp70-ATP conformation with relatively limited structural differences, that is, binding of a peptide or exchange of the nucleotide to ADP. The DNCF analysis is thus primed toward the propagation of conformational changes arising from these signal triggers. However, comparing too divergent conformations using the DNCF method, for example, the domain-docked Hsp70-ATP and the fully undocked Hsp70-ADP conformation would not yield as much useful information, as the DNCF method would simply pick up these dramatic but self-evident conformational differences. Investigating for example, the re-docking of the Hsp70-ADP conformation will require a different set of simulations, where the fully undocked Hsp70-ADP conformation is simulated alongside conformationally related variants that are more likely to initiate re-docking. Clusters formed by our predictions aligned with regions already known to be important for mediating Hsp70 conformational changes and function: The interfaces between NBD and SBD subdomains and the binding site of the JDP DnaJ. Investigated in more detail, we detected a number of residues which were predicted to be specific to either DnaK and BiP. About 40% of these differences arise directly from mutations, while others point to inherent differences between the dynamics of DnaK and BiP, such as the stability of (sub)-domain interfaces and substrate binding pocket conformational plasticity, which have been described previously on a biochemical level. By combining DNCF scores with a targeted random walk, we were able to integrate predictions of individual residues into a proposed pathway responsible for communicating binding of a substrate in the SBD to the NBD. This pathway corresponds to a series of residues whose neighboring interactions are substantially coupled and modulated by substrate and/or nucleotide binding. In this context, communication within this pathway would manifest through the possibility of subtle conformational changes or correlated fluctuations that can occur along the proposed chains of interactions. Our data revealed an alternative pathway existing in BiP but not DnaK, centered around the –(506a)R residue, which is a highly conserved position in eukaryotic Hsp70 variants.[25] Finally, we investigated the conformational control exerted by ATP/ADP over the NBD lobes by simulating the isolated NBD domain in different configurations. We observe that the in-silico transformation of ATP to ADP is sufficient to trigger spontaneous lobe rotation during simulation toward the conformations expected from crystal structures, indicating that the terminal ATP phosphate acts as a strong mechanical wedge locking the lobes in place. Given the relatively short simulation times necessary to observe these rotations in some replicas, there appears to be a relatively low kinetic barrier to rotation after ATP hydrolysis, suggesting that this specific process does not require external assistance by cochaperones, provided that the SBDβ domain is completely undocked from the NBD. Furthermore, we identified a number of residues in DnaK which are in direct contact with the ATP/ADP nucleotide and can thus act as sensors for the nucleotide hydrolysis state, acting as focal points for initiating NBD lobe rotation. In total, our findings shed light on the pathways of allosteric communication in Hsp70s, suggesting the involvement of additional residues beyond what has been experimentally verified. We found that while many signaling residues are conserved between DnaK and BiP, there are also specific differences reflecting the divergent evolution of the two proteins. These specific residues may contribute to an explanation of the differences in biochemical behavior between Hsp70s found in different organisms and organelles. Studies elucidating differential mechanisms within a protein family provide important insights into the regulatory fine-tuning of the system, which are essential for development of targeted orthosteric or allosteric inhibitors. A possible avenue for application is indicated by a study series creating specific allosteric inhibitors for Hsp90, targeting the TRAP1 mitochondrial paralog but with no effect on cytoscolic Hsp90.[98–101] Our observations deepen our understanding of allosteric communication in the Hsp70 system and how a ubiquitous but diverse protein class has adapted to different cellular environments and cochaperone interaction partners. As Hsp70 have also been suggested to be promising therapeutic factors in a range of contexts, among them neurodegenerative diseases[2,3,7–12] and csBiP as a coreceptor of the pandemic SARS-CoV-2 virus,[40–44] investigating such evolutionary differences in further detail may become a key step in developing medical applications.

## ACKNOWLEDGMENTS

We thank Martin Zacharias for constructive discussions and valuable feedback for improving the manuscript. This work is dedicated to the memory of Iris Antes, who passed away unexpectedly on the 4th of August 2021. Open Access funding enabled and organized by Projekt DEAL.

## FUNDING INFORMATION

This work was supported by the Deutsche Forschungsgemeinschaft (https://www.dfg.de/; SFB 1035/A10 and SFB749/C08 to Iris Antes). Markus Schneider 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.

## CONFLICT OF INTEREST

The authors have declared that no competing interests exist.

## PEER REVIEW

The peer review history for this article is available at https://publons.com/publon/10.1002/prot.26425.

## DATA AVAILABILITY STATEMENT

The data that support the findings of this study are openly available in the Dryad database (https://doi.org/10.5061/dryad.1g1jwstzs). 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 distributed over the Cytoscape App Store. Cytoscape 3 can be downloaded for free at https://cytoscape.org/. The source code for SenseNet and AIFgen is included in the java archive (.jar) files used to run these programs.

## ORCID

*Markus Schneider* https://orcid.org/0000-0002-8169-6577

## REFERENCES

1. Zuiderweg ER, Hightower LE, Gestwicki JE. The remarkable multivalency of the Hsp70 chaperones. *Cell Stress Chaperones*. 2017;22(2):173-189.
2. Rosenzweig R, Nillegoda NB, Mayer MP, Bukau B. The Hsp70 chaperone network. *Nat Rev Mol Cell Biol*. 2019;20:665-680.
3. Radons J. The human HSP70 family of chaperones: where do we stand? *Cell Stress Chaperones*. 2016;21(3):379-404.
4. Balchin D, Hayer-Hartl M, Hartl FU. In vivo aspects of protein folding and quality control. *Science*. 2016;353(6294):aac4354.
5. Clerico EM, Tilitsky JM, Meng W, Gierasch LM. How hsp70 molecular machines interact with their substrates to mediate diverse physiological functions. *J Mol Biol*. 2015;427(7):1575-1588.
6. Kohler V, Andreasson C. Hsp70-mediated quality control: should I stay or should I go? *Biol Chem*. 2020;401(11):1233-1248.
7. Clerico EM, Meng W, Pozhidaeva A, Bhasne K, Petridis C, Gierasch LM. Hsp70 molecular chaperones: multifunctional allosteric holding and unfolding machines. *Biochem J*. 2019;476(11):1653-1677.
8. Zuiderweg ER, Bertelsen EB, Rousaki A, Mayer MP, Gestwicki JE, Ahmad A. Allostery in the Hsp70 chaperone proteins. *Top Curr Chem*. 2013;328:99-153.
9. Gestwicki JE, Shao H. Inhibitors and chemical probes for molecular chaperone networks. *J Biol Chem*. 2019;294(6):2151-2161.
10. Patury S, Miyata Y, Gestwicki JE. Pharmacological targeting of the Hsp70 chaperone. *Curr Top Med Chem*. 2009;9(15):1337-1351.
11. Mayer MP. The Hsp70-chaperone Machines in Bacteria. *Front Mol Biosci*. 2021;8:694012.
12. Ferraro M, D'Annessa I, Moroni E, et al. Allosteric modulators of HSP90 and HSP70: dynamics meets function through structure-based drug design. *J Med Chem*. 2019;62(1):60-87.
13. Mayer MP. Intra-molecular pathways of allosteric control in Hsp70s. *Philos Trans R Soc Lond B Biol Sci*. 2018;373(1749):20170183.
14. Mayer MP, Kityk R. Insights into the molecular mechanism of allostery in Hsp70s. *Front Mol Biosci*. 2015;2:58.
15. Mayer MP, Gierasch LM. Recent advances in the structural and mechanistic aspects of Hsp70 molecular chaperones. *J Biol Chem*. 2019;294(6):2085-2097.

16. Zhuravleva A, Clerico EM, Gierasch LM. An interdomain energetic tug-of-war creates the allosterically active state in Hsp70 molecular chaperones. *Cell*. 2012;151(6):1296-1307.
17. Lai AL, Clerico EM, Blackburn ME, et al. Key features of an Hsp70 chaperone allosteric landscape revealed by ion-mobility native mass spectrometry and double electron-electron resonance. *J Biol Chem*. 2017;292(21):8773-8785.
18. Zhuravleva A, Gierasch LM. Allosteric signal transmission in the nucleotide-binding domain of 70-kDa heat shock protein (Hsp70) molecular chaperones. *Proc Natl Acad Sci USA*. 2011;108(17):6987-6992.
19. Kityk R, Vogel M, Schlecht R, Bukau B, Mayer MP. Pathways of allosteric regulation in Hsp70 chaperones. *Nat Commun*. 2015;6:8308.
20. Rosam M, Krader D, Nickels C, et al. Bap (Sil1) regulates the molecular chaperone BiP by coupling release of nucleotide and substrate. *Nat Struct Mol Biol*. 2018;25(1):90-100.
21. Karzai AW, McMacken R. A bipartite signaling mechanism involved in DnaJ-mediated activation of the *Escherichia coli* DnaK protein. *J Biol Chem*. 1996;271(19):11236-11246.
22. Laufen T, Mayer MP, Beisel C, et al. Mechanism of regulation of Hsp70 chaperones by DnaJ cochaperones. *Proc Natl Acad Sci USA*. 1999;96(10):5452-5457.
23. Marcinowski M, Höller M, Feige MJ, Baerend D, Lamb DC, Buchner J. Substrate discrimination of the chaperone BiP by autonomous and cochaperone-regulated conformational transitions. *Nat Struct & Mol Biol*. 2011;18:150-158.
24. Wieteska L, Shahidi S, Zhuravleva A. Allosteric fine-tuning of the conformational equilibrium poises the chaperone BiP for post-translational regulation. *Elife*. 2017;6:e.2943.
25. Meng W, Clerico EM, McArthur N, Gierasch LM. Allosteric landscapes of eukaryotic cytoplasmic Hsp70s are shaped by evolutionary tuning of key interfaces. *Proc Natl Acad Sci USA*. 2018;115(47):11970-11975.
26. Schneider M, Rosam M, Glaser M, et al. BiPPred: combined sequence- and structure-based prediction of peptide binding to the Hsp70 chaperone BiP. *Proteins*. 2016;84(10):1390-1407.
27. Marcinowski M, Rosam M, Seitz C, et al. Conformational selection in substrate recognition by Hsp70 chaperones. *J Mol Biol*. 2013;425(3):466-474.
28. Umehara K, Hoshikawa M, Tochio N, Tate SI. Substrate binding switches the conformation at the lynchpin site in the substrate-binding domain of human Hsp70 to enable allosteric interdomain communication. *Molecules*. 2018;23(3):528.
29. Voith von Voithenberg L, Barth A, Trauschke V, et al. Comparative analysis of the coordinated motion of Hsp70s from different organelles observed by single-molecule three-color FRET. *Proc Natl Acad Sci USA*. 2021;118(33):e2025578118.
30. Wang J, Lee J, Liem D, Ping P. HSPA5 gene encoding Hsp70 chaperone BiP in the endoplasmic reticulum. *Gene*. 2017;618:14-23.
31. Pobre KFR, Poet GJ, Hendershot LM. The endoplasmic reticulum (ER) chaperone BiP is a master regulator of ER functions: getting by with a little help from ERdj friends. *J Biol Chem*. 2019;294(6):2098-2108.
32. Lee AS. Glucose-regulated proteins in cancer: molecular mechanisms and therapeutic potential. *Nat Rev Cancer*. 2014;14(4):263-276.
33. Shin BK, Wang H, Yim AM, et al. Global profiling of the cell surface proteome of cancer cells uncovers an abundance of proteins with chaperone function. *J Biol Chem*. 2003;278(9):7607-7616.
34. Ni M, Zhang Y, Lee AS. Beyond the endoplasmic reticulum: atypical GRP78 in cell viability, signalling and therapeutic targeting. *Biochem J*. 2011;434(2):181-188.
35. Arap MA, Lahdenranta J, Mintz PJ, et al. Cell surface expression of the stress response chaperone GRP78 enables tumor targeting by circulating ligands. *Cancer Cell*. 2004;6(3):275-284.
36. Kim Y, Lillo AM, Steiniger SCJ, et al. Targeting heat shock proteins on cancer cells: selection, characterization, and cell-penetrating properties of a peptidic GRP78 ligand. *Biochemistry*. 2006;45(31):9434-9444.
37. Liu Y, Steiniger SC, Kim Y, Kaufmann GF, Felding-Habermann B, Janda KD. Mechanistic studies of a peptidic GRP78 ligand for cancer cell-specific drug delivery. *Mol Pharm*. 2007;4(3):435-447.
38. Zhang Y, Liu R, Ni M, Gill P, Lee AS. Cell surface relocalization of the endoplasmic reticulum chaperone and unfolded protein response regulator GRP78/BiP. *J Biol Chem*. 2010;285(20):15065-15075.
39. Gopal U, Pizzo SV. Cell surface GRP78 signaling: an emerging role as a transcriptional modulator in cancer. *J Cell Physiol*. 2021;236(4):2352-2363.
40. Carlos AJ, Ha DP, Yeh DW, et al. The chaperone GRP78 is a host auxiliary factor for SARS-CoV-2 and GRP78 depleting antibody blocks viral entry and infection. *J Biol Chem*. 2021;296:100759.
41. Katopodis P, Randeva HS, Spandidos DA, Saravi S, Kyrou I, Karteris E. Host cell entry mediators implicated in the cellular tropism of SARS-CoV-2, the pathophysiology of COVID-19 and the identification of microRNAs that can modulate the expression of these mediators (review). *Int J Mol Med*. 2022;49(2):20.
42. Das JK, Roy S, Guzzi PH. Analyzing host-viral interactome of SARS-CoV-2 for identifying vulnerable host proteins during COVID-19 pathogenesis. *Infect Genet Evol*. 2021;93:104921.
43. Chu H, Chan CM, Zhang X, et al. Middle East respiratory syndrome coronavirus and bat coronavirus HKU9 both can utilize GRP78 for attachment onto host cells. *J Biol Chem*. 2018;293(30):11709-11726.
44. Shahriari-Felordi M, Alikhani HK, Hashemian SR, Hassan M, Vosough M. Mini review ATF4 and GRP78 as novel molecular targets in ER-stress modulation for critical COVID-19 patients. *Mol Biol Rep*. 2022;49:1545-1549.
45. Behnke J, Feige MJ, Hendershot LM. BiP and its nucleotide exchange factors Grp170 and Sil1: mechanisms of action and biological functions. *J Mol Biol*. 2015;427(7):1589-1608.
46. Serlidaki D, van Waarde M, Rohland L, et al. Functional diversity between HSP70 paralogs caused by variable interactions with specific co-chaperones. *J Biol Chem*. 2020;295(21):7301-7316.
47. Li H, Musayev FN, Yang J, et al. A novel and unique ATP hydrolysis to AMP by a human Hsp70 binding immunoglobin protein (BiP). *Protein Sci*. 2021;31(4):797-810.
48. Bonomo J, Welsh JP, Manthiram K, Swartz JR. Comparing the functional properties of the Hsp70 chaperones, DnaK and BiP. *Biophys Chem*. 2010;149(1–2):58-66.
49. Schneider M, Antes I. SenseNet, a tool for analysis of protein structure networks obtained from molecular dynamics simulations. *PLoS One*. 2022;17(3):e0265194.
50. Proctor EA, Kota P, Aleksandrov AA, He L, Riordan JR, Dokholyan NV. Rational coupled dynamics network manipulation rescues disease-relevant mutant cystic fibrosis transmembrane conductance regulator. *Chem Sci*. 2015;6(2):1237-1246.
51. 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;2006(2):2006.0019. doi:10.1038/msb4100063
52. Blacklock K, Verkhivker GM. Computational modeling of allosteric regulation in the hsp90 chaperones: a statistical ensemble analysis of protein structure networks and allosteric communications. *PLoS Comput Biol*. 2014;10(6):e1003679.
53. Penkler D, Sensoy O, Atilgan C, Tastan BO. Perturbation-response scanning reveals key residues for allosteric control in Hsp70. *J Chem Inf Model*. 2017;57(6):1359-1374.
54. Stetz G, Verkhivker GM. Computational analysis of residue interaction networks and coevolutionary relationships in the Hsp70

chaperones: a community-hopping model of allosteric regulation and communication. *PLoS Comput Biol*. 2017;13(1):e1005299.
55. Stetz G, Verkhivker GM. Dancing through life: molecular dynamics simulations and network-centric modeling of allosteric mechanisms in Hsp70 and Hsp110 chaperone proteins. *PLoS One*. 2015;10(11):e0143752.
56. English CA, Sherman W, Meng W, Gierasch LM. The Hsp70 interdomain linker is a dynamic switch that enables allosteric communication between two structured domains. *J Biol Chem*. 2017;292(36):14765-14774.
57. Chiappori F, Merelli I, Milanesi L, Colombo G, Morra G. An atomistic view of Hsp70 allosteric crosstalk: from the nucleotide to the substrate binding domain and back. *Sci Rep*. 2016;6:23474.
58. Ung PM-U, Thompson AD, Chang L, Gestwicki JE, Carlson HA. Identification of key hinge residues important for nucleotide-dependent allostery in *E. coli* Hsp70/DnaK. *PLoS Comput Biol*. 2013;9(11):e1003279.
59. Nicolaï A, Delarue P, Senet P. Decipher the mechanisms of protein conformational changes induced by nucleotide binding through free-energy landscape analysis: ATP binding to Hsp70. *PLoS Comput Biol*. 2013;9(12):e1003379.
60. Chiappori F, Merelli I, Colombo G, Milanesi L, Morra G. Molecular mechanism of allosteric communication in Hsp70 revealed by molecular dynamics simulations. *PLoS Comput Biol*. 2012;8(12):e1002844.
61. Hartmann C, Antes I, Lengauer T. IRECS: a new algorithm for the selection of most probable ensembles of side-chain conformations in protein models. *Protein Sci: Publ Protein Soc*. 2007;16(7):1294-1307.
62. Eswar N, Webb B, Marti-Renom MA, et al. Comparative protein structure modeling using Modeller. *Curr Protoc Bioinform* 2006; Chapter 5:Unit-5 6. doi:10.1002/0471250953.bi0506s15
63. Case DA, Cerutti DS, Cheatham TE, et al. *AMBER 2017*. University of California; 2017.
64. 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):3696-3713.
65. Jorgensen WL, Chandrasekhar J, Madura JD, Impey RW, Klein ML. Comparison of simple potential functions for simulating liquid water. *J Chem Phys*. 1983;79(2):926-935.
66. Meagher KL, Redman LT, Carlson HA. Development of polyphosphate parameters for use with the AMBER force field. *J Comput Chem*. 2003;24(9):1016-1025.
67. 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):1082-1089.
68. Miyamoto S, Kollman PA. Settle: an analytical version of the SHAKE and RATTLE algorithm for rigid water models. *J Comput Chem*. 1992;13(8):952-962.
69. 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):3084-3095.
70. Shannon P, Markiel A, Ozier O, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res*. 2003;13(11):2498-2504.
71. Hunter JD. Matplotlib: a 2D graphics environment. *Comput Sci Eng*. 2007;9(3):90-95.
72. Humphrey W, Dalke A, Schulten K. VMD: visual molecular dynamics. *J Mol Graph*. 1996;14(1):33-38.
73. Wang W, Liu Q, Liu Q, Hendrickson WA. Conformational equilibria in allosteric control of Hsp70 chaperones. *Mol Cell*. 2021;81(19):3919-3933.e7.
74. Vogel M, Bukau B, Mayer MP. Allosteric regulation of Hsp70 chaperones by a proline switch. *Mol Cell*. 2006;21(3):359-367.
75. Barthel TK, Zhang J, Walker GC. ATPase-defective derivatives of *Escherichia coli* DnaK that behave differently with respect to ATP-induced conformational change and peptide release. *J Bacteriol*. 2001;183(19):5482-5490.
76. Burkholder WF, Panagiotidis CA, Silverstein SJ, Cegielska A, Gottesman ME, Gaitanaris GA. Isolation and characterization of an *Escherichia coli* DnaK mutant with impaired ATPase activity. *J Mol Biol*. 1994;242(4):364-377.
77. Gässler CS, Buchberger A, Laufen T, et al. Mutations in the DnaK chaperone affecting interaction with the DnaJ cochaperone. *Proc Natl Acad Sci USA*. 1998;95(26):15229-15234.
78. Vogel M, Mayer MP, Bukau B. Allosteric regulation of Hsp70 chaperones involves a conserved interdomain linker. *J Biol Chem*. 2006;281(50):38705-38711.
79. Suh W-C, Burkholder WF, Lu CZ, Zhao X, Gottesman ME, Gross CA. Interaction of the Hsp70 molecular chaperone, DnaK, with its cochaperone DnaJ. *Proc Natl Acad Sci*. 1998;95(26):15223-15228.
80. Kamath-Loeb AS, Lu CZ, Suh W-C, Lonetto MA, Gross CA. Analysis of three DnaK mutant proteins suggests that progression through the ATPase cycle requires conformational changes. *J Biol Chem*. 1995;270(50):30051-30059.
81. Smock RG, Rivoire O, Russ WP, et al. An interdomain sector mediating allostery in Hsp70 molecular chaperones. *Mol Syst Biol*. 2010;6:414.
82. Mayer MP, Laufen T, Paal K, McCarty JS, Bukau B. Investigation of the interaction between DnaK and DnaJ by surface Plasmon resonance spectroscopy. *J Mol Biol*. 1999;289(4):1131-1144.
83. Montgomery DL, Morimoto RI, Gierasch LM. Mutations in the substrate binding domain of the Escherichia coli 70 kda molecular chaperone, DnaK, which alter substrate affinity or interdomain. *J Mol Biol*. 1999;286(3):915-932.
84. Kityk R, Kopp J, Sinning I, Mayer MP. Structure and dynamics of the ATP-bound open conformation of Hsp70 chaperones. *Mol Cell*. 2012;48(6):863-874.
85. Kityk R, Kopp J, Mayer MP. Molecular mechanism of J-domain-triggered ATP hydrolysis by Hsp70 chaperones. *Mol Cell*. 2018;69(2):227-237.e4.
86. General IJ, Liu Y, Blackburn ME, Mao W, Gierasch LM, Bahar I. ATPase subdomain IA is a mediator of interdomain allostery in Hsp70 molecular chaperones. *PLoS Comput Biol*. 2014;10(5):e1003624.
87. Yang J, Nune M, Zong Y, Zhou L, Liu Q. Close and allosteric opening of the polypeptide-binding site in a human Hsp70 chaperone BiP. *Structure*. 2015;23(12):2191-2203.
88. Yang J, Zong Y, Su J, et al. Conformation transitions of the polypeptide-binding pocket support an active substrate release from Hsp70s. *Nat Commun*. 2017;8(1):1201.
89. Yan M, Li J, Sha B. Structural analysis of the Sil1–Bip complex reveals the mechanism for Sil1 to function as a nucleotide-exchange factor. *Biochem J*. 2011;438(3):447-455.
90. Shomura Y, Dragovic Z, Chang HC, et al. Regulation of Hsp70 function by HspBP1: structural analysis reveals an alternate mechanism for Hsp70 nucleotide exchange. *Mol Cell*. 2005;17(3):367-379.
91. Bracher A, Verghese J. The nucleotide exchange factors of Hsp70 molecular chaperones. *Front Mol Biosci*. 2015;2:10.
92. O'Rourke KF, Gorman SD, Boehr DD. Biophysical and computational methods to analyze amino acid interaction networks in proteins. *Comput Struct Biotechnol J*. 2016;14:245-251.
93. Greene LH. Protein structure networks. *Brief Funct Genom*. 2012;11:469-478.
94. Di Paola L, Giuliani A. Protein contact network topology: a natural language for allostery. *Curr Opin Struc Biol*. 2015;31:43-48.
95. Bertelsen EB, Chang L, Gestwicki JE, Zuiderweg ER. Solution conformation of wild-type E. coli Hsp70 (DnaK) chaperone complexed with ADP and substrate. *Proc Natl Acad Sci USA*. 2009;106(21):8471-8476.

96. Liu Y, Gierasch LM, Bahar I. Role of Hsp70 ATPase domain intrinsic dynamics and sequence evolution in enabling its functional interactions with NEFs. *PLoS Comput Biol*. 2010;6(9):e1000931.

97. Chang L, Thompson AD, Ung P, Carlson HA, Gestwicki JE. Mutagenesis reveals the complex relationships between ATPase rate and the chaperone activities of Escherichia coli heat shock protein 70 (Hsp70/DnaK). *J Biol Chem*. 2010;285(28):21282-21291.

98. Moroni E, Agard DA, Colombo G. The structural asymmetry of mitochondrial Hsp90 (Trap1) determines fine tuning of functional dynamics. *J Chem Theory Comput*. 2018;14(2):1033-1044.

99. Serapian SA, Moroni E, Ferraro M, Colombo G. Atomistic simulations of the mechanisms of the poorly catalytic mitochondrial chaperone Trap1: insights into the effects of structural asymmetry on reactivity. *ACS Catalysis*. 2021;11(14):8605-8620.

100. Sanchez-Martin C, Moroni E, Ferraro M, et al. Rational Design of Allosteric and Selective Inhibitors of the molecular chaperone TRAP1. *Cell Rep*. 2020;31(3):107531.

101. Sanchez-Martin C, Menon D, Moroni E, et al. Honokiol Bis-Dichloroacetate is a selective allosteric inhibitor of the mitochondrial chaperone TRAP1. *Antioxid Redox Signal*. 2021;34(7):505-516.

## SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

**How to cite this article:** Schneider M, Antes I. Comparison of allosteric signaling in DnaK and BiP using mutual information between simulated residue conformations. *Proteins*. 2023; 91(2):237-255. doi:10.1002/prot.26425

# SenseNet Manual

Version 1.1.0

# Contents

# 1 Introduction

SenseNet (“Structure ENSEmble NETworks”) maps structure ensembles of biomolecules to atom interaction networks and provides functions for their analysis and visualization. It is available as a plugin for the free network visualization software *Cytoscape* [1].

Protein structures are frequently analysed to gain insights into the effects of ligand binding, residue mutations or conformational changes. In contrast to isolated structures generated by crystallography or other experimental sources, ensembles obtained from Molecular Dynamics (MD) provide additional information such as interaction lifetimes or correlation between conformations, allowing to investigate dynamic properties of biomolecules. In a structure ensemble network, each node represents one or a group of atoms while edges correspond to the interactions between these atoms (e.g. hydrophobic contacts or hydrogen bonds). Each edge is associated with a ’timeline’ which indicates the presence of an interaction for each structure in the ensemble. Analysis functions are provided to extract information from these timelines and map results to network nodes and edges. Finally, SenseNet offers comprehensive visualization functions for side-by-side analyses of networks and 3D structures.

# 2 Installation

The recommended method for obtaining SenseNet is the Cytoscape App Store or alternatively from our website at `https://bioinformatics.wzw.tum.de`. The latter option requires you to install the plugin manually. To do this, place the ’SenseNet’ .jar file into the ’CytoscapeConfiguration/3/apps/installed’ folder. Make sure to remove any old version of SenseNet from this folder before starting Cytoscape.

# 3 Data model

SenseNet maps protein or other macromolecular structures to a network of nodes, which correspond to individual atoms or groups of atoms, and edges, representing interactions between atoms. Atoms may be grouped into a ’metanode’; the individual atom nodes that are contained in this group are hence called its ’subnodes’. In a network, either a metanode or its subnodes may be ’active’ (i.e. present) at a given time. This is achieved by expanding (replacing a metanode by its subnodes) or collapsing (replacing all subnodes by their metanode). The metanodes of a network can be expanded or collapsed by a double click, and all analysis methods which act on the ’active’ set of nodes and edges take the current state of the network into account. For example, an analysis can be performed while some selected residues are expanded into individual atom nodes, while other residues are represented by a single node. When collapsing subnodes, all edges connecting to these nodes are replaced by metaedges which represent the cumulative interactions of the replaced edges. Separate metaedges are created for each interaction type (e.g. contacts or hydrogen bonds). A detailed description of edge sets and how they change can be found in section 5.3.

In order to model a structure ensemble, each edge is associated with a timeline represented as either a vector of integer values (e.g. presence or absence of contacts in each time frame) or a vector of real numbers (e.g. interaction energies). The timeline of a metaedge is called a ’metatimeline’ and is calculated from its subedge timelines. The ’Sum’ and ’Occurrence’ frame weight methods yield two different metatimeline variants

$$\begin{aligned} X_{ijk,sum} &= \sum_{\alpha \in i} \sum_{\beta \in j} X_{\alpha\beta k} \\ X_{ijk,occ} &= \text{min}(1, X_{ijk,sum}) \end{aligned} \tag{1}$$

in which $X$ corresponds to a timeline, $i, j$ are metanodes, $k$ is an interaction type and $\alpha, \beta$ are subnodes of $i, j$.

The weight of an edge describes the strength of an interaction. It is determined by the ’Timeline weight method’, which is by default the average over all time frames. Alternatively, users can set the network to any single time frame or use averages of time blocks, e.g. as obtained from clustering.

Difference networks can be used to compare networks of two similar, but different ensembles (e.g. structures with one or more point mutations). A difference network is created by mapping interactions between equivalent atoms of two networks onto each other. Two atoms are considered equivalent if they have the same chain name, residue index, residue inset, residue alternative location and atom name (see PDB specification [2]). Notably, the residue name is not compared in order to allow comparisons for residue point mutations. All interactions between two equivalent atoms which have the same interaction type are considered equivalent. Interactions for which no equivalent can be found are compared to an empty timeline of all zeroes. Once all equivalent interactions are mapped, the timeline vectors are subtracted element-wise

$$X_{\alpha\beta k,diff} = X_{\alpha\beta k} - X_{\alpha\beta k,ref} \tag{2}$$

where $X_{\alpha\beta k}$ is the timeline of the compared network, and $X_{\alpha\beta k,ref}$ is the timeline of the reference network. Metatimelines are calculated analogously to eq. 1.

# 4 User Guide

All plugin functions can be accessed via the Cytoscape GUI. The controlling elements can be found either in the 'SenseNet' tab on in the control panel or in the top menu within 'Apps - SenseNet'.

## 4.1 General

**Import network** *Parameters*

- **Import networks** Choose one or more input sources to import. Use the plus and minus buttons to add/remove fields. See also section 5.4 for more details on some of the file formats.
    - **Input type**
        - **AIF file** Import network from an AIF (Atom Interaction Format) .aif/.zaif file.
            - **.aif/.zaif file** Input file: Either in AIF or compressed ZAIF format.
            - **Frame sieve** Read only every nth frame. Useful to save memory.
            - **Skip timelines** Skip timeline depending on a threshold. For example the 'Skip timelines $<$ 0.05 avg.' option does not import a timeline if its average is below 0.05.
        - **CPPTRAJ H-bonds** Import network from CPPTRAJ [3] hbond command output.
            - **H-bond file** Generated by 'avgout'.
            - **Timeline file** Generated by 'uuseries'.
            - **Interaction type** Interaction type name.
            - **Ignore backbone** Choose to ignore contacts involving backbone atoms (atom names C,O,N,CA).
            - **Frame sieve** Same as for 'AIF file'.
            - **Skip timelines** Same as for 'AIF file'.
        - **CPPTRAJ nativecontacts** Import network from CPPTRAJ nativecontacts output.
            - **Contacts file** Generated by 'writecontacts'.
            - **Native timeline file** Generated by 'seriesout'.
            - **Nonnative timeline file** Generated by 'seriesnnout'.
            - **Contacts .pdb file** Generated by 'contactpdb'.
            - **Interaction type** Interaction type name.

        - **Ignore backbone** Choose to ignore contacts involving backbone atoms (atom names C,O,N,CA).
        - **Ignore intra-residue** Choose to ignore contacts within the same residue.
        - **Frame sieve** Same as for 'AIF file'.
        - **Skip timelines** Same as for 'AIF file'.
    - **PDB structure H-bonds** Import network from H-bonds found in a PDB file. Hydrogens are required to find H-bonds. All atoms in the PDB file whose names start with 'H' are considered hydrogens.
        - **PDB file** Select .pdb file to load.
        - **Distance cut-off** Maximum donor-acceptor distance.
        - **Angle cut-off** Minimum donor-hydrogen-acceptor angle.
        - **Donor mask** Atom mask for donor atoms (see mask reference in section 5.2).
        - **Acceptor mask** Atom mask for acceptor atoms (see mask reference in section 5.2).
        - **Interaction type** Interaction type name.
    - **PDB structure contacts** Import network from contacts found in a PDB file.
        - **PDB file** Select .pdb file to load.
        - **Distance cut-off** Maximum contact distance.
        - **Atom mask** Atom name mask to calculate contacts for (see mask reference in section 5.2).
        - **Interaction type** Interaction type name.
        - **Ignore backbone** Choose to ignore contacts involving backbone atoms (atom names C,O,N,CA).
        - **Ignore intra-residue** Choose to ignore contacts within the same residue.
    - **DSSP secondary structure** Import network of secondary structure elements. All residues belonging to the same secondary structure element (helix/sheet) are connected sequentially.
    - **DSSP file** Select .dssp file to load.
    - **Interaction type** Interaction type name.
- **Difference network** Check box to select input files for creating a difference network. When checked, the 'Import reference networks' panel will appear. The differences are calculated by subtracting the timelines of equivalent interactions of the reference network from the network loaded in the top import panel (**Import networks**).

- **Metanode definition** Choose how atom nodes are grouped together.
  - **Group definition** Grouping settings.
    - **Amino acids** Group atoms by their respective amino acids
    - **Backbone/Sidechain** Group atoms by their respective amino acids, but create seperate nodes for the backbone/sidechain portions. The **Backbone atom names** text field contains all atom names that will be categorized as backbone, separated by commas.
- **Network options**
  - **Create visual style** Check to automatically create a new visual style for the network.
  - **Network name** Displayed name of the network. Automatically filled when imported files are chosen. Can also be set manually.
  - **Remove edges** Setting to remove (deactivate) edges below a certain threshold. Equivalent to the setting with the same name in the 'Interaction weights' tab. Can be changed at any time.

**Show log** Show task logs.

*Parameters*

- **Log category** Select log category to display.
  - **Global** Continuous list of task logs since session start.
  - **Task** Last log of task type selected in the **Log type** box.

**Export network** Export current network in AIF format. The exported file can e.g. be used for importing a multisource network as a single file containing the combined information of all original import files.

## 4.2 Shown interactions

Only interaction types with checked boxes are shown in the network. The interaction type of each edge is read from the **shared interaction** column.

## 4.3 Interaction weights

This panel controls how individual atom timelines are combined into metatimelines. The radio buttons on the left chooses how each frame of the metatimeline is determined ('Frame weight method'; see eq. 1). The buttons on the right select how to determine the total weight of the metatimeline ('Timeline weight method'):

- **Average weight** Average over all frames in the timeline

- **Single frame** Select and show the network state at an individual time frame
- **Clusters** Show average of frames within a cluster. Requires previous clustering analysis

Weighting is performed on all imported edges (see section 5.3 for an explanation of edge sets). For metaedges, all subedges of the same interaction type are considered. Whenever weighting is performed, the results are written into the 'weight' and 'standard deviation' edge columns.

## 4.4 Analysis

### 4.4.1 Network interactions

The analysis functions in this panel act on the whole network. Results are usually presented as tables/plots in the result panel on the right.

**Timeline** Shows the timelines for all imported edges. For metaedges, the metatimeline is calculated according to the selected weight method.

*Parameters*

- **Frame weight** Method for determining metatimelines. See section 4.3

**Degree** Calculates weighted degree for active edges. Results are written into the 'degree' node table column.

*Parameters*

- **Degree weight** Method for calculating weights for adjacent edges.
  - **Edge weight sum** Sum values of edge columns.
- **Weight column** Source column for edge weights.
- **Negative weights** Method to treat negative edge weights.
  - **abs(x)** Use absolute value of x.

**Centrality** Calculate weighted centrality measures for active nodes, based on shortest paths. The algorithms are implemented as described in refs. [4, 5]. If two nodes are connected by multiple edges, they are treated as one, using one of several merging methods. Results are written into the 'centrality' node table column.

*Parameters*

- **Centrality type** Centrality measures to compute. A transformation function maps edge weights to determine the distance between nodes (see algorithm 10 in ref. [4])

- **Multiple edges weights** Method for merging parallel edges between two nodes.
  - **Sum/Min/Max** Use the weight sum/average/min/max of parallel edges as total weight.
  - **Edge count** Use the number of edges between two node pairs as total weight.
  - **Uniform** All edges are assigned an uniform weight. Parallel edges are ignored. This option effectively gives the centralities of an unweighted network.
- **Weight column** Column of edge weights.
- **Distance transformation** Function to transform edge weights to distances. The function is commonly chosen such that a high weight corresponds to a low distance.
- **Negative weights** Method to treat negative edge weights.
- **Normalization** Type of normalization to apply to centrality values.
  - **None** Do not perform normalization.
  - **Min-max range** Subtract the minimum centrality of the network from each value and divide by the range of values. The normalized value is limited between 0 (lowest centrality) and 1 (highest centrality).
  - **Max node pairs** Divide each node's centrality by the theoretical maximum number of node pairs, excluding that node, in an undirected network: $\frac{(N-1)(N-2)}{2}$, where $N$ is the total number of active nodes.

**Correlation** Determine correlation between **active** edges. The edge neighbour correlation factor $\mathrm{ECF}$ is calculated as

$$\mathrm{ECF(i)} = \sum_{j \in A} |c(i,j)| \tag{3}$$

where $i, j$ are network edges, $c(i,j)$ is a correlation function of edges $i$ and $j$, and $A$ is the set of edges that $i$ is compared to. Correlation factor types may use different edge sets for $A$.

*Parameters*

- **Correlation factor type**
  - **Neighbour** Edge neighbour correlation factor. Here, $A$ is the set of edges adjacent to $i$ (i.e. separated by at most one node)
- **Correlation method** Correlation measure to use (see text below).
- **Frame weight** Method for determining metatimelines. See section 4.3.
- **Reference network** For methods calculating correlation differences.

- **Edge mapping** Determine how edges are matched between active and reference networks.
  - **Shared name** Match edges with identical ’shared name’ entries in the edge table.
  - **Match location** Require that the ’shared name’ entries of source and target nodes approximately match between networks. The approximation is that only the ’residue name’ is allowed to differ. The ’altloc’ and ’residue insert’ identifiers are not considered part of the residue name and thus still have to be identical. This option is intended for comparing two networks differing e.g. in point mutations of single residues in order to match edges representing conserved interactions.

The ’Mutual information’ correlation method determines correlation as

$$I(X;Y) = \sum_{x \in X} \sum_{y \in Y} p(x,y) \log_2 \left( \frac{p(x,y)}{p(x)p(y)} \right) \tag{4}$$

in which, $X$ and $Y$ are **integer** timelines, $p(x,y)$ is the joint probability function of values $x, y$, and $p(x), p(y)$ are marginal probability functions of $x, y$. This method should be used when the timeline contains discrete values, such as the count of interactions. The unit for mutual information results is ’bits’.

’Mutual information difference’ calculates the sum of absolute changes in expected pointwise mutual information for each event

$$\begin{aligned} I(X;Y) = \sum_{x \in (X \bigcup \hat{X})} \sum_{y \in (Y \bigcup \hat{Y})} \Bigg| & \, p(x,y) \log_2 \left( \frac{p(x,y)}{p(x)p(y)} \right) \\ & - \hat{p}(x,y) \log_2 \left( \frac{\hat{p}(x,y)}{\hat{p}(x)\hat{p}(y)} \right) \Bigg| \end{aligned} \tag{5}$$

where the $\hat{X}, \hat{Y}$ denote timelines in the reference network corresponding to $X, Y$. Edges are considered equivalent if their ’shared name’ columns match. If no match can be found for an edge, the reference timeline is replaced by a vector of zeroes. This measure is useful for determining differences in dynamic behaviours between simulations, f.e. a protein with and without a ligand.

Alternatively, choosing the ’Pearson’ correlation method will calculate the Pearson correlation coefficient between interaction timelines

$$r = \frac{\sum_i (X(i) - \mu_x)(Y(i) - \mu_y)}{\sigma_x \sigma_y} \tag{6}$$

where $i$ is a discrete time frame, $X(i), Y(i)$ are functions yielding the corresponding timeline value at position $i$, with associated sample means $\mu$ and standard deviations $\sigma$. This method is recommended when the timeline contains continuous values like interaction energies.

The resulting correlation factors for each edge are written into the 'correlation factor' column.

**Lifetime** Calculates estimates for the interaction lifetimes. These are calculated from the intermittent autocorrelation function

$$C_L(k) = \frac{1}{N} \sum_{i=0}^{N-k} \frac{X_{occ}(i) X_{occ}(i+k)}{X_{occ}(i)^2} \tag{7}$$

in which $k$ is the discrete lag step, $N$ is the total number of time frames and $X_{occ}(i)$ is the occurrence weighted metatimeline function. The calculated lifetime is intermittent, i.e. the interaction may break and reform between the compared time frames. The average lifetime is estimated from the autocorrelations following the same protocol as for the autocorrelation time during error estimation.

If the 'Replicas' setting is set to $n > 1$, the timeline is divided into $n$ equal sized blocks, which are analyzed separately. The average of block lifetimes is written to the 'lifetime' edge column.

**Weight error** Calculates weight error estimates for all imported edges. These functions aim at approximating the standard error of the timeline weight.

*Parameters*

- **Error method** Method for calculating the error. Currently, only 'Autocorrelation' is available.
- **Frame weight** Method for determining metatimelines. See section 4.3.
- **Replica weight** Method to merge errors from multiple replicas.
  - **Max/Avg/Min** Use maximum, average or minimum of replicas as final error value.

The autocorrelation method is based on the approach outlined in ref. [6]. The timeline autocorrelation for different discrete lag steps is calculated as

$$C(k) = \frac{1}{N\sigma^2} \sum_{i=0}^{N-k} (X(i) - \mu)(X(i+k) - \mu) \tag{8}$$

with $k$ as the discrete lag step, $N$ as the total number of time frames, $X(i)$ as the timeline function at each discrete time frame, $\mu$ as the mean of $X$, and $\sigma^2$ as the variance of $X$. A single exponential of the form $A * e^{Bx}$ is fitted by weighted linear regression in log space with weights $\frac{1}{k+1}$. To reduce the influence of noisy function tails, only autocorrelation values above 0.1 are considered for fitting. The integral of the fitted exponential is calculated analytically and serves as estimate for the autocorrelation time $\tau$. The estimate for the independent sample size is then determined as

$$N_{ind} = \frac{N}{\tau} \tag{9}$$

which is used to estimate the standard error of independent samples

$$\sigma_{e,ind} = \frac{\sigma}{\sqrt{N_{ind}}} \tag{10}$$

If the ’Replicas’ setting is set to $n > 1$, the timeline is divided into $n$ equal sized blocks, which are analyzed separately. The final error is determined as by the ’Replica weight method’.

The results of this analysis are written into the ’error estimate’, ’autocorrelation sample size’, and ’autocorrelation time’ edge columns.

**Entropy** For each network edge, determine Shannon’s information entropy

$$H(X) = -\sum_{x \in X} p(x) \log_2(p(x)) \tag{11}$$

where $X$ is the edge’s integer interaction timeline. Results are written into the ’entropy’ edge column. The unit of reported results is ’bits’.

*Parameters*

- **Frame weight** Method for determining metatimelines. See section 4.3.

**Random Walk** Perform a random walk through the network. At each step, the next visited node is selected randomly from the list of nodes connected to the current node via one or more edges (neighbors). Multiple edges between node pairs do not affect the selection probability.

*Parameters*

- **Walk mode**
  - **Default** Regular random walk starting from the specified node and finishing after the specified number of steps, tracking visited nodes along the way. Restarts do not reset the list of visited nodes.
  - **Targeted** Random walk starting and finishing at the specified nodes. If the target node is not reached after the specified number of steps, the run is restarted. Restarts always reset the list of visited nodes.
  - **Targeted-Symmetric** Like ’Targeted’, but each run is performed an additional second time after exchanging start and target nodes. The total number of runs is doubled.
- **Weighting mode**
  - **Unweighted** The next node is selected from the list of neighbors using a uniform probability distribution.

- **Weighted** The next node is selected using a weighted probability factor obtained from a node table value specified by 'Weight Column'. The probability of a node neighbor to be selected for the next step is then

$$p(i) = \frac{w(i)}{\sum_{n \in N} w(n)} \tag{12}$$

  where $i$ is the neighbor candidate from the node neighbor list $N$ and $w(i)$ is the weight obtained from the 'Weight Column' for node $i$.
- **Weight column** Column used for obtaining weights and calculate weighted probabilities.
- **Max steps** Maximum number of walk steps in a single run.
- **Restart probability** Probability of restarting the run at each step. The list of visited nodes may be retained or cleared, depending on the 'Walk mode'.
- **Num runs** Number of independent runs to be performed. Nodes visited multipled times during one individual run are counted only once. Individual runs are added up to yield the final result.
- **Random seed** Seed value for random generator. Choosing the same seed guruantees reproduction of a specific analysis outcome, provided that the network and all other parameters are the same.

The number of random walk visits for each node are written into the 'visited' edge column.

### 4.4.2 Selected interactions

Show detailed analyses for one or more selected edges.

**Timeline** Equivalent to the corresponding function in section 4.4.1, but shows a plot of the timeline using the currently active frame weight method.

**Correlation** Calculates correlation measures between a single selected edge and all other imported edges (see section 4.4.1).

**Autocorrelation** Equivalent to the corresponding function in section 4.4.1. Uses the currently active frame weight method and plots autocorrelation functions for each selected edge.

**Blocked error** Plots the blocked standard errors following a procedure from ref. [6]. Uses the currently active frame weight method.

In essence, the standard error (see eq. 10) is calculated multiple times between blocks of time frames. For the minimum block size of 1, the result is the conventional standard error. For a block size of 2, each time frame is averaged with its successor yielding $\frac{n}{2}$ non-overlapping 'block averages', where $n$ is the total number of time frames. The standard error is then calculated between these block averages as if each was an independent data point. This is repeated for increasing block sizes. The maximum block size is set as $\frac{n}{4r}$, where $n$ is the total number of time frames and $r$ is the number of replicas. The total standard error can be estimated from the plot as the value to which the blocked standard errors converge.

### 4.4.3 Network matrix

These functions allow plotting and exporting networks in matrix form.

**Show** Display matrix of active network edges as a dotplot.

*Parameters*

- **Weight column** Select column to use for weighting.
- **Node index column** Select column to use as indices for the X and Y axes of the plot. Only columns of integers are allowed.
- **Min/Max value** Set min and max value for weight color scale.

**Export** Export matrix of active network edges.

*Parameters*

- **Weight column** Select column to use for weighting.
- **Node name column** Select column to use as names for the X and Y axes of the plot.
- **Output file** File to write matrix to.

### 4.4.4 Paths

The following functions provide functions for the identification of pathways between two selected nodes.

**Shortest paths** Find all shortest paths by traversing active edges between two selected nodes, starting from the node that was selected first. Edges are considered to represent equal distances, and parallel edges are ignored. The paths are presented in a table in the result panel. In addition, two measures of interaction strength are shown: The 'timeline sum' is the sum of average interactions of edges contributing to the pathway. In contrast, 'timeline occurrence' gives the average occurrence weights along that path. If multiple edges are present between two nodes, they are treated as if their timelines were merged.

**Suboptimal paths** Find all paths of a fixed length range between two nodes. Only active edges are considered and all edges are assumed to represent the same distance. Minimum and maximum path lengths can be set as parameters. Otherwise, output is equivalent to the 'Shortest paths' function.

### 4.4.5 Clustering

From this panel, functions for clustering of time frames can be accessed.

**Cluster** Start a new clustering run, grouping time frames with similar network states until a limit is reached. For this purpose, each time frame in the network is represented by an interaction matrix of dimension $N \times N$, where $N$ is the number of nodes. Each entry in the matrix corresponds to the selected weight. If multiple edges are present between two nodes, their weights are summed. The distance between two time frames is calculated as the Frobenius norm of their matrix differences. This set of distances is then used for clustering.

*Parameters*

- **Clustering method**
  - **Agglomerative** Hierarchical agglomerative clustering.
  - **Linkage** Select linkage mode for agglomerative clustering.
- **Target cluster count/Epsilon** Select to stop clustering either at N clusters or when the minimum intercluster distance drops below a certain limit.
- **Sieve** Select to use only every Nth frame for clustering.
- **Frame weight** Select weight mode for the frame interaction matrix. See section 4.3.

## 4.5 Structure visualization

This panel contains functions to connect Cytoscape to a 3D structure viewer session. Networks can be linked to structures or trajectories loaded in the viewer. Node and edge selections in a linked network are highlighted in the structure.

**Connect viewer** Start a structure viewer and connect to it.

*Parameters*

- **Viewer** Select one of the available viewers to start (VMD, PyMOL or UCSF Chimera). See section 2 for viewer installation requirements.
- **Load session** Select a session file for the viewer to load after starting.

**Model link** Shows whether the currently focused network is linked to a structure.

**Link network** Link current network to a structure in the connected viewer.

*Parameters*

- **Single structure** Load a single structure into the viewer and link the network to it.
  - **Structure file** Select file to load structure from.
  - **Format** Automatically determined by structure file extension. Can be set manually. Available options depend on the connected viewer.
  - **Model name** Model name to use for the structure. Must be unique.
- **Trajectory** Load a trajectory of structures into the viewer and link the network to it.
  - **Structure file** Select file to load structure or topology from.
  - **Format** same as in 'Single structure'. Depending on the connected viewer, topology formats are accepted as well.
  - **Model name** same as in 'Single structure'.
  - **Trajectory file** Select file to load trajectory from.
  - **Trajectory format** same as 'Format', but for trajectories.
- **Preloaded** Link network to a structure already present in the viewer
  - **Model name** Select model name in structure viewer to link network to.

**Unlink network** Remove structure link of the current network. An option is given to remove the structure from the viewer as well.

**Pause link** Temporarily disable structure link of the current network. Until the link is unpaused, selection changes are not updated between network and structure viewer.

**Transfer colors** Color linked structure according to node colors in the network. For VMD and UCSF Chimera, the visually closest color of the defined color set is used. For PyMOL, the colors are transferred exactly as shown in the network.

## 4.6 Style

### 4.6.1 Node style

**Auto style** Map continuous node attributes to visual style. Creates a copy of the current style with a '_auto' suffix. If a style with that name already exists, it will be overridden.

*Parameters*

- **Style property** Select visual property to map values to. Different style settings are available for each property.
  - **Color** Map node attribute to fill color.
    - **Min/Mid/Max value** Minimum/Middle/Maximum value to map. Default values are determined automatically according the range of values in the network for the selected column.
    - **Low/Mid/High color** Color gradient from low to high.
  - **Size** Map node attribute to node size.
  - **Min/Max value** see 'Color'.
  - **Min/Max size** Minimum/Maximum node size.
- **Column** Column to map values from. All columns containing 'Double' values can be chosen.

**Label format** Select node label style. See section 5.1 for an explanation of naming conventions.

**Renumber** Renumber residue indices in labels.

*Parameters*

- **Chain** Chain(s) to renumber. Select one character ('A','B', etc.) or '*' to select all chains.
- **First residue index** First residue index to renumber.
- **Last residue index** Last residue index to renumber or '-1' to select up until and including the last residue index of selected chain(s).
- **Offset** Offset to add to the selected residues. Can be positive or negative.

Note that renumbering always acts on the **original residue numbering**. Therefore, if you renumbered residue 1 to become residue 100, you would have to select residue index 1 again to renumber it a second time. In order to return all numberings to the original imported state, you can use the 'Reset numbering' button in the dialog.

Labeling and renumbering never changes the 'shared name' or 'residue index' columns. Instead, results are written into the 'label' and 'residue index label' columns in the node table. This is done to avoid accidental ambiguity and loss of data. See section 5.1 for an explanation of the underlying concepts.

### 4.6.2 Edge style

**Auto style** Analogous to the corresponding function in section 4.6.1.

## 4.7 Settings

**Structure viewer** Configuration for structure viewers

- **PyMOL/VMD/UCSF Chimera location** Location to search for respective executables. If no location is given, the plugin will attempt to run the displayed command from the operating system's PATH variable
- **Max shown residues** Maximum number of residues to show as sticks before an error is thrown
- **Max shown interactions** Maximum number of interactions highlighted before an error is thrown
- **Selected interaction color** Highlight color for selected interactions
- **Selected residue color** Highlight color for selected residues
- **Zoom to selection** Check to enable auto-zoom to selected residues

# 5 Concepts

## 5.1 Labels and identifiers

Within the plugin, certain naming conventions are used to map nodes and edges to their structural counterparts. The standard naming style for nodes is

<**Chain**>/<**Altloc**><**Residue name**><**Residue insert**>
**-**<**Residue index**>**(-**<**Mutated residue name**>**)(:**<**Atom name**>**)**
**(#**<**Group tag**>**)**

The standard naming style for edges is

<**Node name 1**>_<**Bridge name**>_<**Node name 2**>
_<**Interaction type**>

The data fields correspond to the RCSB PDB [2] standard[1]. 'Group tag' is used when a non-amino acid metanode definition is used (such as backbone / side-chain). 'Bridge name' is used for interactions that involve more than two atoms, for example the name of the hydrogen in a hydrogen bond interaction. For amino acids, three-letter codes are used for residue names, but longer names

[1] http://www.wwpdb.org/documentation/file-format-content/format33/v3.3.html

are possible. Empty fields are allowed. Field separators enclosed in parentheses will only appear if the corresponding field is filled. For edges with a bridge atom, node names 1 and 2 are the source and target nodes, respectively. If no bridge atom is present (e.g. in metaedges), the node names appear in alphabetical order. This is done to ensure that edge names are predictable for symmetric interactions. the An example for a standard node label is 'A/TYR-290', and for a standard edge label it is 'A/TYR-290__A/VAL-80_H-bond'.

A node or edge 'name' (as found in the 'shared name' or 'name' columns) always follows the standard naming style, is assigned at import and is never changed. 'Names' are meant to identify edges as uniquely as possible, but **there is no guarantee that a name is unique in the network**. Node label customization functions create 'labels', which are allowed to omit information (such as chain names). Edge labels are automatically updated accordingly to the node labels. Labeling functions always write into separate 'label' columns.

## 5.2 Atom masks

When importing networks using the plugin, sometimes a selection of atom names is needed. Whenever an 'atom mask' is required, a Java regular expression[2] needs to be provided. For most purposes, a tiny subset of the regular expression language is sufficient for a proper selection. Examples of often used patterns are

- **.*** All atom names (. = any character, * = zero or more repetitions of the preceding character)
- **C.*** Atom names starting with C
- **F.***|**O.***|**N.*** Atom names starting with F,O, or N (| = 'or')

## 5.3 Network operations

The number of nodes and edges currently present can change due to filtering or expanding/collapsing of metanodes. Hence, it is important for each function that deals with edges to define on which set it operates. The sets are defined as follows:

- **Active nodes/edges** All edges that are currently present in the network. This includes nodes that are present, but hidden from view (f.e. using Cytoscape's 'hide nodes' feature). A node or edge that becomes inactive is temporarily deleted from the network. When an inactive node/edge becomes active, it is restored.
- **Imported nodes/edges** All nodes/edges from the original import, regardless of whether they are active or not.

The following operations can render a node/edge active or inactive:

[2] https://docs.oracle.com/javase/7/docs/api/java/util/regex/Pattern.html

- **Expanding/Collapsing a metanode** The metanode and associated metaedges are deactivated. Subnodes and associated edges are activated.
- **Edge weight filtering** Edges which are filtered out due to low weights are deactivated.
- **Subnetwork creation** When subnetworks are created from selected nodes, the selected nodes become the active node set for that network.
- **Manual deletion** Manually deleting nodes/edges f.e. using the 'DEL' key are deactivated, but can be reactivated f.e. when weight filtering is updated or a metanode is expanded/collapsed.

Generally, **the set of active nodes and edges is equivalent to the currently visible network**. Nodes and edges that are only hidden due to visualization (such as using Cytoscape's 'hide nodes/edges' feature) are still considered 'active' and part of the network. Inactive nodes are temporarily deleted from the network and are hence invisible both in the presentation and for network analysis algorithms. Hence, tools like NetworkAnalyzer or other analysis plugins can be used normally.

## 5.4 AIF file format

The AIF (Atom Interaction Format) was created as a convenient way to define interaction networks based on timeline data. All lines start with a record indicator, followed by one or more comma-separated data fields. Interaction data may be given as TIMELINE and DIFFERENCE_TIMELINE records, which define the following fields

- interaction type (string)
- source atomname (string)
- target atomname (string)
- source residue index (int)
- target residue index (int)
- source residue name (string)
- target residue name (string)
- source residue insert (string)
- target residue insert (string)
- source altloc (string)
- target altloc (string)
- source chain (string)
- target chain (string)
- bridge names (whitespace delimited list of strings)
- timeline (whitespace delimited list of ints/floats)

An AIF file must have the following properties:

- Lines are separated by Unix style newline ('Linefeed') characters.
- A line consists of one or more fields, which are separated by a comma. Leading and trailing whitespace are ignored for each field. Fields may be empty.
- The first field of each line denotes the record type. The record type defines how many fields follow in the same line and what their field data types are. The record type is case insensitive.
- Lines starting with a '#' character indicate comment lines and should be ignored by parsers.
- Empty lines or lines that contain only whitespace should be ignored by parsers.
- Tabs count as regular characters (not whitespace) and should be avoided entirely.

In order to save disk space, AIF files may be zipped (.zaif).

# 6 External tools guide

The SenseNet plugin can interact with a number of programs. This section gives details on how to set up and use those programs together with the plugin.

## 6.1 Command line interface

In order to allow automatized workflows, SenseNet allows some of its functions to be called either via the Cytoscape automation console, script files or the CyREST interface. They fulfill the same purpose as their equally named GUI counterparts and are called by preceding them with the "sensenet" namespace tag (e.g. "senseset importAif"). In addition, functions provide documentation by using the "help" command (e.g. "help sensenet importAif"). Validity of input is checked after parsing the command and corresponding error messages will appear, spelling out problems and allowed input options.

## 6.2 Structure viewers

The plugin can interact with several structure viewers (PyMOL, VMD and UCSF Chimera) in order to map the network onto a molecule structure. In order to use the plugin together with one of these viewers, it is only necessary that the plugin can start the viewer. By default, it will attempt to look up the installation location from the operating system's PATH variable (equivalent to typing 'pymol','vmd' or 'chimera' on the command line). Alternatively, the installation location can be set manually in the 'Settings' menu. The viewer needs to be started using the 'Connect viewer' button in order to link a network.

The plugin was tested in combination with PyMOL 2.1.0, VMD 1.9.2 and UCSF Chimera 1.12.

## 6.3 CPPTRAJ

CPPTRAJ [3] is part of the AmberTools program suite and can be used to process and analyze molecular dynamics trajectories. The following sections describe how to use CPPTRAJ to write interaction timeline data, which can be used to create interaction networks. The scripts were tested with the CPPTRAJ version contained in AmberTools17.

### 6.3.1 nativecontacts

The following output files of the CPPTRAJ nativecontacts command are necessary for network import:

- **contacts.out** Contact table
- **contacts.series** Timeline series for native contacts
- **contacts.nonnative.series** Timeline series for non-native contacts
- **contacts.pdb** PDB file as output by nativecontacts

These files can be created using the following CPPTRAJ commands (adjust **bold** arguments as necessary)

```
parm md.prmtop
trajin md.nc
nativecontacts @C* distance 5.0 \
      writecontacts contacts.out contactpdb contacts.pdb \
      series seriesout contacts.series \
      savenonnative seriesnnout contacts.nonnative.series
run
```

### 6.3.2 hbond

The following output files of the CPPTRAJ hbond command are needed for importing a network:

- **hbonds.out** Interaction table
- **hbonds.series** Interaction timeline series

These files can be generated by the following CPPTRAJ commands (adjust **bold** arguments as necessary)

```
parm md.prmtop
trajin md.nc
hbond !(:WAT,Na+,Cl-) dist 3.5 angle 135 \
      avgout hbonds.out \
      series uuseries hbonds.series
run
```

# 7 Troubleshooting

## 7.1 Installation

- *Could not find plugin for installation* Make sure that your current working directory contains the .jar file found in the zip archive
- *Could not find Cytoscape app directory* Check whether Cytoscape is installed on your system. If you never started Cytoscape on your machine, try starting it once before installing the plugin. The CytoscapeConfiguration directory will be created on its first start.

## 7.2 Session files

Cytoscape offers to save the current working state in a .cys session file. The plugin is fully compatible with this function. However, reading session files generated with older versions of the plugin may fail. In general, you can expect session to work if the first two digits of the plugin version are identical. For example, a session created with version 1.2.0 is guaranteed to be readable by plugins of version 1.2.X, but not by version 1.1.0.

## 7.3 Slow analyses and 'out of memory' errors

Keeping the full timelines of long trajectories consumes a lot of memory, especially for large systems. Reducing the amount of analysed time frames, e.g. by using the 'sieve' option in CPPTRAJ or during import, can accelerate analyses substantially. For a workstation with 8 GB RAM and a protein of about 200 amino acids, we found a number of  5000 frames to work well.

## 7.4 Subnetworks and changes in node/edge tables

It is often useful to create one or more subnetworks to analyze specific regions in the protein. It is important to remember that the node and edge data tables are shared between the network and all subnetworks. This means that an analysis performed in a subnetwork also changes the data tables in all other networks belonging to the same group.

# References

[1] P Shannon, A Markiel, O Ozier, N S Baliga, J T Wang, D Ramage, N Amin, B Schwikowski, and T Ideker. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res*, 13:2498–2504, 2003.

[2] Peter W. Rose, Bojan Beran, Chunxiao Bi, Wolfgang F. Bluhm, Dimitris Dimitropoulos, David S. Goodsell, Andreas Prlić, Martha Quesada, Gregory B. Quinn, John D. Westbrook, Jasmine Young, Benjamin Yukich, Christine Zardecki, Helen M. Berman, and Philip E. Bourne. The rcsb protein data bank: redesigned web site and web services. *Nucleic Acids Research*, 39(suppl_1):D392–D401, 2011.

[3] D. R. Roe and 3rd Cheatham, T. E. Ptraj and cpptraj: Software for processing and analysis of molecular dynamics trajectory data. *J Chem Theory Comput*, 9(7):3084–95, 2013.

[4] Ulrik Brandes. On variants of shortest-path betweenness centrality and their generic computation. *Social Networks*, 30(2):136 – 145, 2008.

[5] Antonio del Sol, Hirotomo Fujihashi, Dolors Amoros, and Ruth Nussinov. Residues crucial for maintaining short paths in network communication mediate signaling in proteins. *Molecular Systems Biology*, 2(1):2006.0019, 2006.

[6] A. Grossfield and D. M. Zuckerman. Quantifying uncertainty and sampling quality in biomolecular simulations. *Annu Rep Comput Chem*, 5:23–48, 2009.