Cooperative protein allosteric transition mediated by a fluctuating transmission network

Matthias Post,[1] Benjamin Lickert,[1] Georg Diez,[1] Steffen Wolf,[1, a)] and Gerhard Stock[1, b)]
Biomolecular Dynamics, Institute of Physics, Albert Ludwigs University, 79104 Freiburg, Germany.

Allosteric communication between distant protein sites represents a key mechanism of biomolecular regulation and signal transduction. Compared to other processes such as protein folding, however, the dynamical evolution of allosteric transitions is still not well understood. As example of allosteric coupling between distant protein regions, we consider the global open-closed motion of the two domains of T4 lysozyme, which is triggered by local motions in the hinge region. Combining extensive molecular dynamics simulations with machine learning of contact features, we identify a network of interresidue distances that move in a concerted manner. The cooperative process originates from a cogwheel-like motion of the hydrophobic core in the hinge region, which constitutes a flexible transmission network. Through rigid contacts and the protein backbone, the small local changes of the hydrophobic core are passed on to the distant terminal domains and lead to the emergence of a rare global conformational transition. As in an Ising-type model, the cooperativity of the allosteric transition can be explained via the interaction of local fluctuations.

INTRODUCTION

Proteins may coexist in various conformational states of different function, e.g., in an active and an inactive state. Transitions between these states can occur spontaneously or may be triggered by an external stimulus such as ligand binding. A prime example is allosteric communication, where a binding event at one site of a protein may affect the structure and dynamics of another distant site[1–7]. Even without the presence of a ligand, many allosteric systems thermally populate both protein states, thus representing a bistable system. Common to all cases is that the associated structural rearrangement requires the interaction of distant regions of the system. Although these long-range or allosteric couplings appear to be ubiquitous in protein systems[1], surprisingly little is known about the dynamical mechanism, its microscopical details, and the time evolution of the associated conformational transition[5]. We note that the situation is quite different for the protein folding problem, where several decades of theoretical and experimental work have resulted in a quite well-established picture how folding and unfolding proceed. This includes general scenarios such as cooperative two-state and multistate downhill folding[8], as well as dynamical mechanisms such as zipping or diffusion limited processes[9]. The dynamical process underlying an allosteric transition, on the other hand, appears much more elusive due to the typical smallness of the various local structural changes, which are quite challenging to observe in experiments[10–12]. While allosteric transitions have been rarely witnessed in unbiased molecular dynamics (MD) simulations[6,13–17], numerous residue-based network models have been suggested, which aim to predict intramolecular signaling[18]. Furthermore Markov state models[19], which describe the conformational dynamics via memoryless jumps between metastable states, have been applied to allostery[14,20].

In this work we consider T4 lysozyme (T4L) as a well-studied example of a bistable two-domain protein[21–26]. Aiming to destroy bacterial cell walls by catalyzing the cleavage of glycosidic bonds, T4L performs an open-closed transition of its two domains that resembles a Pac-Man, see Fig. 1a. This motion of the "mouth" region was recently shown to be triggered by a locking mechanism in the "hinge" region of T4L, revealing that these two distant regions are allosterically coupled[26]. (Note that we use the term 'allostery' for general long-range communication, independent from ligand binding.) With lifetimes of the open and closed state in the order of a few microseconds, T4L is one of the few protein systems, whose long-range transitions can be sufficiently sampled by un-

a) email: steffen.wolf@physik.uni-freiburg.de
b) email: stock@physik.uni-freiburg.de

biased simulations. Adopting the 61 μs-long all-atom MD simulation of Ernst et al.[26], here we elucidate the microscopic mechanism of this coupling. Previous work showed that the identification of the underlying reaction coordinates poses a challenge to various dimensionality reduction approaches[25–28]. Combining a correlation analysis of contact features[29] with a careful visual inspection of the conformational changes, we show that the open-closed transition of T4L amounts to a cooperative process, where a large number of interatomic distances from the hinge to the mouth region act in a simultaneous manner. In particular, we find that the hydrophobic core of the hinge region constitutes a flexible transmission network, which represents the molecular machinery at the heart of the open-closed transition.

RESULTS AND DISCUSSION

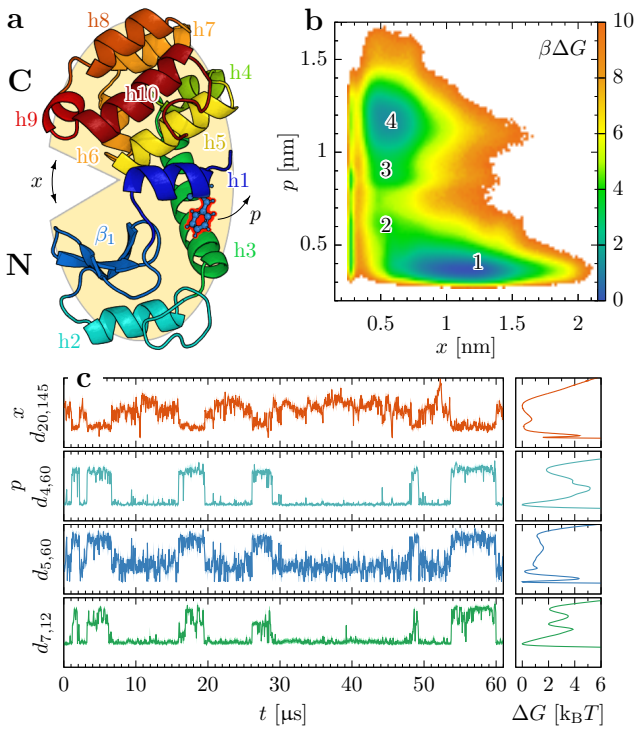

FIG. 1. (a) Structure of the N- and C-domain of T4L, indicating the opening coordinate x of the "mouth" region and the locking coordinate p that describes the motion of key residue Phe4 (red) in the "hinge" region. (b) Two-dimensional free energy landscape $\Delta G(x,p)$ (in units of k_BT), indicating four conformational states. (c) MD time traces of various distance coordinates describing the open-closed transition and their corresponding free energy profiles.

Two-dimensional model: the locking mechanism. The open-closed motion of T4L is usually described by the mouth opening width x (Fig. 1a), which can be represented, e.g., by the distance $d_{20,145}$ between residues 20 and 145 of the N- and C-domain, respectively. Employing the MD trajectory by Ernst et al.[26], Fig. 1c reveals that the open-closed transitions of $x(t)$ occur on a microsecond timescale, while the transition path time lasts only nanoseconds. From the data we estimate equilibrium populations of 70 % (open) and 30 % (closed) and mean waiting times of $\tau_{o\to c} \sim 4\,\mu s$ and $\tau_{c\to o} \sim 2\,\mu s$, which is in excellent agreement with recent experimental results[24]. The corresponding free energy profile $\Delta G(x)$, however, exhibits a rather low energy barrier ($\Delta G^{\#} \sim 1 k_BT$). Adopting the standard expression for the transition rate, $k = 1/\tau = k_0\, e^{-\Delta G^{\#}/k_BT}$, it is obvious that the rate is almost solely caused by the prefactor k_0[30]. This indicates that the opening width x does not reveal the underlying mechanism of the process and is therefore not a suitable one-dimensional reaction coordinate.

Indeed, when we employed targeted MD simulations to enforce the open-close motion by pulling along x, the two-state behavior of T4L could not be recovered[26]. Rather, the transition was found to be triggered by a locking mechanism, by which the side chain of Phe4 changes from a solvent-exposed to a hydrophobically-buried state. Describing the motion of Phe4 via the distance between Phe4 and Lys60, $d_{4,60} \equiv p$, this coordinate discriminates between locked ($p \lesssim 0.7$, hydrophobically buried) and free ($p \gtrsim 0.7$, solvent exposed) states. Figures 1b,c show the free energy landscape $\Delta G(x,p)$ and the time traces $x(t)$, $p(t)$ of the resulting two-dimensional model of T4L. The landscape exhibits four metastable conformational states labeled by numbers **1** - **4**, including the prominent open locked state **1** and closed free state **4**, as well as two sparsely populated intermediate closed states **2** and **3**.

The locking mechanism is found to stabilize the open and closed states and can therefore be considered as the origin of the bistable behavior of T4L. Demonstrating the coupling of opening coordinate x and locking coordinate p, the free energy landscape $\Delta G(x,p)$ also provides first mechanistic insights of the conformational transition[26]. In particular $\Delta G(x,p)$ indicates that most transitions proceed along the route **1**→**2**→**3**→**4** and back along

4→3→2→1. Since the corresponding highest free energy barrier is still quite small ($\sim 5\,k_{\mathrm{B}}T$), however, the model cannot provide an microscopic explanation for the microsecond transition times. This is related to the fact that the two-dimensional model does not account for the propagation of the conformational change from Phe4 in the hinge region to the mouth of T4L.

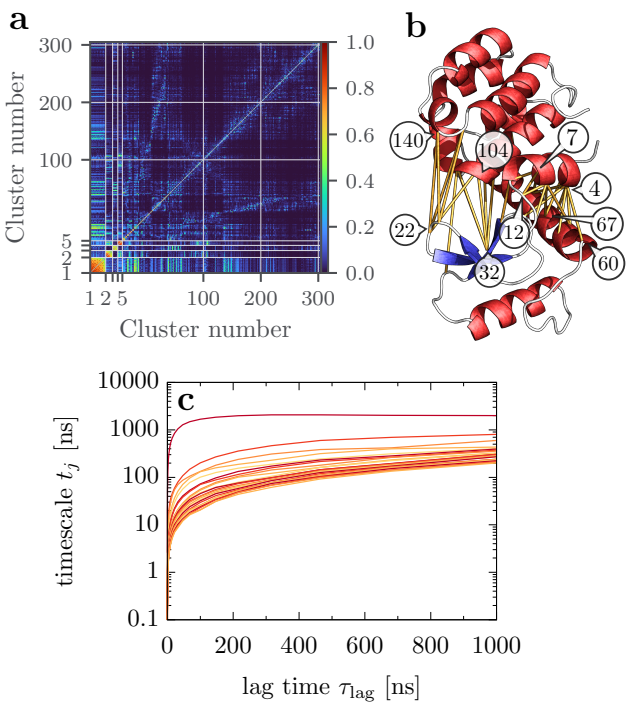

FIG. 2. (a) Block-diagonal form of the correlation matrix built from 402 contact distances of T4L. The first block (cluster 1) contains 27 highly correlated coordinates that account for the open-closed transition. (b) Illustration of the 27 relevant contact distances (yellow lines), indicating residues Phe4, Phe67, Leu7 and Phe104 of the hydrophobic core, as well as important distances $d_{4,60}$ (locking coordinate p), $d_{22,140}$ describing the opening of the mouth, and $d_{7,12}$ accounting for the (de)stabilization of helix 1. (c) Timescales t_j of the corresponding TICA components as a function of the employed lag time. The well-separated slowest timescale t_1 (top line) accounts for the open-close motion, while the other quite similar timescales reflect fluctuations within the open and the closed state.

Correlation analysis: Cooperative dynamics. To describe the mechanistic coupling between the mouth and hinge regions in T4L, we want to identify internal coordinates that change significantly during the open-closed transition. As discussed in detail by Ernst et al.[26], the contact distances of T4L represent a suitable set of coordinates for this purpose. Assuming that a contact is formed if the distance d_{ij} between the closest non-hydrogen atoms of residues i and j is shorter than 4.5 Å[31], 402 interresidue contacts were identified for T4L[26]. To study which of these contact distances are interrelated to each other, we calculated the linear correlation matrix

$$\sigma_{ij,kl} = \frac{\langle \delta d_{ij}\, \delta d_{kl} \rangle}{\langle \delta d_{ij}^2 \rangle^{1/2}\, \langle \delta d_{kl}^2 \rangle^{1/2}} \tag{1}$$

where $\delta d_{ij} = d_{ij} - \langle \delta d_{ij} \rangle$ and $\langle \ldots \rangle$ denotes the time average over the MD data. To rearrange this matrix in an approximately block-diagonal form, we employed a machine learning technique called Leiden algorithm[32], which is a community detection approach (see Methods). The resulting block-diagonalized correlation matrix in Fig. 2a shows that only 27 of the 402 contact distances correlate strongly (comprised in cluster 1, i.e., the first block), while the remaining coordinates, distributed over the remaining $\sim$ 300 clusters, exhibit only minor correlation. The latter finding is associated with the fact that in T4L intraprotein contacts are quite stable and only fluctuate around their mean distance, while contacts on the protein surface frequently form and break and hence fluctuate randomly.

Displaying the 27 highly correlated distances in the molecular structure of T4L, Fig. 2b shows that these coordinates span the space connecting the mouth and hinge regions. They include 13 mouth-opening coordinates and 6 coordinates involving Phe4, while the remaining coordinates are mostly located in the hinge region. As demonstrated by the time evolution of some important examples (Fig. 1c), these coordinates follow the microsecond open-closed transitions of T4L. Hence, the above correlation analysis indicates that the open-closed transition has to be mediated via 27 highly correlated contact distances that connect the hinge and mouth regions of T4L.

While the overall time evolutions of the coordinates in Fig. 1c are clearly correlated, the various contact distances are seen to differ in their feasible fluctuations in the open and closed state. E.g., opening coordinate x fluctuates considerably in the open state and only little in the closed state, while it is the other way round for the locking coordinate p. This is because amino acid side-chains that are in contact generally fluctuate significantly less than in the case of a broken contact. It suggests that only if the fluctuations of the various contacts by chance cause the simultaneous switching of all relevant interactions, the open-close transition may take place. In

other words, the transition seems to require a concerted or cooperative action of all involved coordinates of the intramolecular contact network of T4L.

Timescale analysis: simultaneous motions. To further characterize this presumed cooperative motion, we employed time-lagged independent component analysis[33]. TICA represents a linear transformation, $\boldsymbol{y} = A\boldsymbol{x}$, which aims to identify the 'slowest coordinates' y_j of the system. That is, TICA maximizes the autocorrelation functions $C_j(\tau_{\rm lag}) = \langle y_j(\tau_{\rm lag}) y_j(0)\rangle \sim e^{-\tau_{\rm lag}/t_j}$ for a chosen lag time $\tau_{\rm lag}$, and orders the decay times t_j decreasingly ($t_1 > t_2 > t_3 > \ldots$). Using the 27 distance coordinates x_j as input data, Fig. 2c shows the resulting timescales t_j as a function of the employed lag time. We find that the timescale t_1 of the first component is one order of magnitude slower than the other timescales, which all are quite similar. While t_1 represents the open-close motion, the other timescales t_j reflect the above discussed fluctuations of the various coordinates in the open and the closed state. This significant separation between t_1 and all other timescales clearly suggest a trivial two-state Markov state model based on a single coordinate reflecting the open-close motion. To gain a microscopic understanding of the mechanism of the open-close transition and test our hypothesis of a cooperative action, however, we need to include additional coordinates into the model, which according to Fig. 2c do not show a timescale separation.

Combining our correlation analysis with the structural characterization given in the next section, we identified the following reaction coordinates. Besides the above discussed opening coordinate $x \equiv d_{20,145}$ and the locking coordinate $p \equiv d_{4,60}$, we included the contact distances $d_{5,60}$ and $d_{22,137}$, which also report on the dynamics of the hinge and mouth regions, respectively. Moreover, it turns out advantageous to represent the re-orientation of Phe4 by its second side-chain dihedral angle, referred to as χ_4 (instead of several related distance coordinates). Finally the distance $d_{7,12}$ will be of importance, as it is found to account for the (de)stabilization of helix 1. To illustrate to what extend these coordinates from the hinge to the mouth region indeed move simultaneously, we zoom into the time evolution of the system during a representative open→closed transition. Assigning the start of the transition to time $t = 0$, Fig. 3a reveals that distance $d_{5,60}$ changes first, before right afterwards coordinates $d_{4,60}$, $d_{22,137}$ and $d_{20,145}$ change virtually simultaneously. A few nanosecond later χ_4 responds, followed by $d_{7,12}$ about 10 ns later. Of particular interest is the simultaneous response of distances $d_{4,60}$ (hinge) and $d_{22,137}$ (mouth), which are far apart (about 2.5 nm) from each other (see Fig. 2b). This synchronism indeed seems to indicate a direct mechanical coupling of the two protein sections.

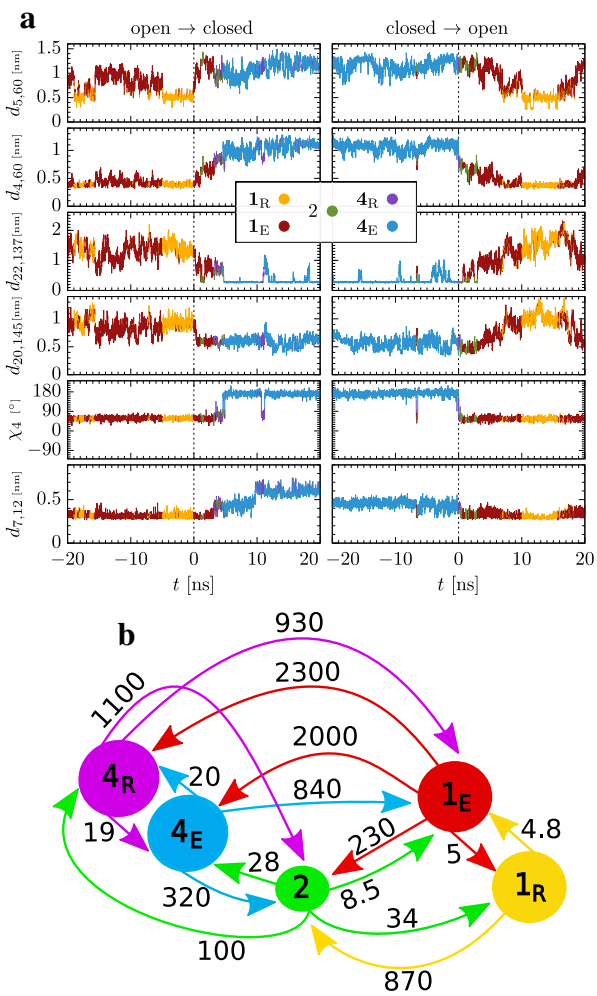

FIG. 3. (a) Time evolution of some key coordinates of T4L during a representative open→closed (left, occurring at 3.2 µs in Fig. 1) and closed→open (right, at 16.6 µs) transition, respectively. The color encodes the instantaneous conformational state of the system revealed by density-based clustering. (b) Scheme of the resulting Markov state model indicating the main transition times (in units of ns).

Structural characterization: transmission network. To explain the above findings, Fig. 4a depicts the overall conformational rearrangement of T4L during the open-closed transition. While the N- and C-terminal pro-

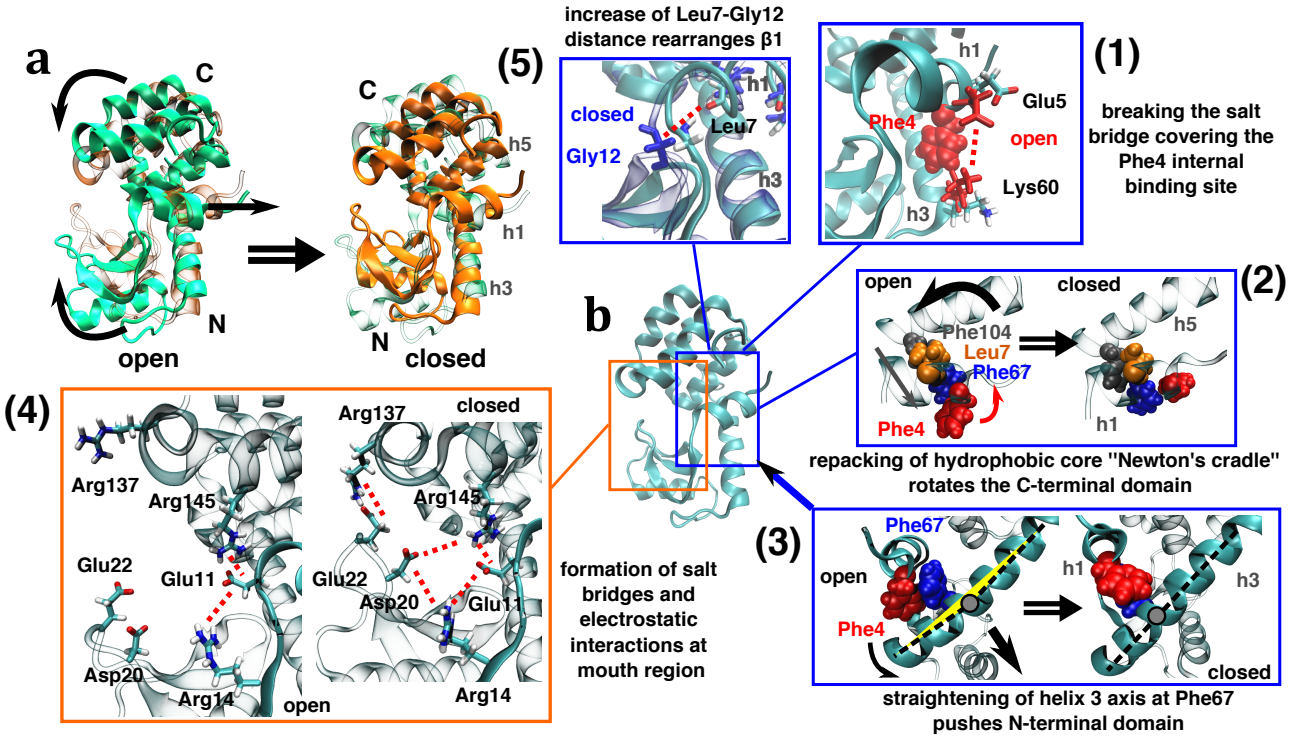

FIG. 4. Structural details of the open-closed mechanism in T4L, with black arrows indicating the direction of backbone motions. (a) illustrates the overall picture, (b) accounts for the individual steps of the conformational rearrangement, see the text.

tein domains perform a mirrored rotational motion, helix 1 executes a piston-like move into the opposite direction. This structural coupling between N- and C-terminal domains as well as with helix 1 is facilitated by a network of persistent salt bridges indicated in Figs. 4b and S1. That is, helix 1 is connected to the C-terminal domain via several salt bridges between the two negatively charged residues Asp10 and Glu11 in helix 1 and the two positively charged residues Arg145 and Arg148 in helix 10 of the C-terminal domain. Along helix 3, we find a salt bridge between Asp72 and Arg76. Furthermore, helix 3 is connected to the N-terminal domain via Arg52-Glu62, and to the C-terminal domain via Arg80-Glu108. All these electrostatic connections serve as structural tethers that provide a balance between stability and flexibility, which we consider as a prerequisite for a bistable system.

In Fig. 4b, panels (1) to (5) comprise the individual local sites mediating the conformational rearrangement from the protein open to the closed state. Moreover, the analysis suggests the intramolecular coordinates shown in Fig. 3a, which are indicative of the local structural changes. The structural evolution is illustrated in real time in the Supplementary Movie S1.

(1) To allow the exit of Phe4 from its hydrophobic pocket, first a salt bridge between Glu5 and Lys60 needs to open. This motion is naturally described by the contact distance $d_{5,60}$. The subsequent outward move of Phe4 can be monitored by the locking coordinate $d_{4,60}$.

(2) The outward motion of Phe4 is transferred via Phe67, Leu7 and Phe104 to its adjacent helix 5. That is, Phe4, Phe67, Leu7 and Phe104 form a hydrophobic core that causes a rolling motion of helix 5 over helix 1 similar to two cogwheels in contact. As helix 5 represents an integral part of the C-terminal domain, this motion of helix 5 rotates the C-terminal domain as shown in Fig. 4a.

(3) At the same time, the conformational change of the Phe67 side chain is transferred farther to the backbone of helix 3, which causes a straightening of the helix. This straightening acts on the N-terminal domain via the backbone and side chain packings, causing the forward movement of the N-terminal domain. In combination, the above described motions of the C- and the N-terminal domains lead to the closing of the mouth of T4L.

(4) The closed conformation is stabilized via a salt bridge at the tip of the mouth between Glu22 and Arg137, which can be followed via distance $d_{22,137}$, and an electrostatic interaction complex involving Glu11, Arg14, Asp20 and Arg145 that can be monitored by $d_{20,145}$ in the middle of the mouth. Both salt bridge centers impose structural near-order on the mouth region. Since these coordinates respond simultaneously with the locking coordinate $d_{4,60}$ (which indicates that the Phe4 side chain passes over the side chain of Phe67), the complex conformational rearrangements described in steps (2) to (4) occur in a concerted manner. Subsequently, the Phe4 side chain rotates into a new conformation and becomes completely exposed to the solvent. This can be monitored by the side-chain dihedral angle χ_4.

(5) Finally, a destabilization of the last winding of helix 1 leads to a rupture of a backbone hydrogen bond between Leu7 and Gly12, resulting in an increase in the distance $d_{7,12}$. As a consequence, the motion of Gly12 causes a change in the position of the β_1 strand such that Arg14 is rotated away from the electrostatic complex, thus weakening its coupling with Glu11 and Asp20.

To go back to the open state, the above steps simply occur in reverse order. Here the crucial issue is the locking of Phe4, i.e., its move from the solvent-exposed to a hydrophobically-buried state, which is indicated by coordinates $d_{7,12}$, χ_4 and p.

In summary, we have shown that the hydrophobic core (Phe4, Phe67, Leu7 and Phe104) mediating a cogwheel-like motion of helices 1 and 5 and the straightening of helix 3 constitutes a flexible transmission network, which represents the molecular machinery underlying the open-closed transition. While common motions of Phe4, Phe67 and Phe104 seemed plausible from the comparison of various wild-type and mutant crystal structures of T4L[21,22], here we show that these residues together with Leu7 in fact perform a concerted motion. Since helices 3 and 5 are rigidly connected to the N- and C-terminal domains, respectively, the motion of the transmission network in the hinge region is directly and instantaneously coupled to the mouth region of T4L. This long-range or allosteric coupling is the origin of the concerted action of the 27 distances connecting the mouth and the hinge region of T4L (Fig. 2). Finally, we note that the overall process requires the initial unlocking of Phe4 to start the transition (Fig. 4b1), its reorientation into the bulk water, and a backbone hydrogen bond (Fig. 4b5) to finalize the transition.

State model description: heterogeneity of pathways. The finding of several steps in the cooperative mechanism raises the question to what extent these steps are associated with metastable intermediate states of the process. To this end, we employed the six coordinates (x, p, $d_{5,60}$, $d_{22,137}$, χ_4, $d_{7,12}$) discussed in Fig. 3a and performed robust density-based clustering[34] in this space (see Methods and Fig. S2 for details). In line with the nomenclature of the previous four-state model (Fig. 1c), the resulting five conformational states are referred to as states $\mathbf{1}_\mathrm{R}$ (with a population of 33.5 %), $\mathbf{1}_\mathrm{E}$ (35.5 %), $\mathbf{2}$ (1.3 %), $\mathbf{4}_\mathrm{E}$ (15.1 %) and $\mathbf{4}_\mathrm{R}$ (14.5 %), see Fig. 3b. Compared to the previous model that relied on coordinates (x, p) only, notably the two main states **1** (open) and **4** (closed) each split up into two similarly populated substates, indexed by "R" (for relaxed) and "E" (for excited). In both cases, the sub-state E represents the "pole position" to start the open↔closed transition, while the sub-state R accounts for the fully relaxed equilibrium state. Displaying the probability distributions of all six coordinates in each state, Fig. S3 reveals that the splitting of state **1** is caused by coordinate $d_{5,60}$, while the splitting of state **4** is caused by coordinate $d_{7,12}$. While the intermediate state **2** is weakly populated and short-lived as in the previous model, it turns out to be mechanistically relevant (see below). In contrast, the weakly populated previous intermediate state **3** is now a part of state $\mathbf{4}_\mathrm{E}$. The spatial extension and the connectivity of the five states are well characterized by the two-dimensional free energy surfaces along the above defined six coordinates shown in Fig. S4.

To relate these conformational states to steps (1) - (5) in Fig. 4b, we consider the resulting state trajectory during the representative open→closed transition shown in Fig. 3a. As suggested above, we find that transition $\mathbf{1}_\mathrm{R} \rightarrow \mathbf{1}_\mathrm{E}$ accounts for the initial unlocking of Phe4 (Fig. 4b1). The subsequent main cooperative process (Figs. 4b2 - 4), on the other hand, is associated with two transitions, $\mathbf{1}_\mathrm{E} \rightarrow \mathbf{2}$ and $\mathbf{2} \rightarrow \mathbf{4}_\mathrm{E}$. State **2** represents the first

stabilized closed structure, that is, it succeeded to stabilize one of the frequent attempts to close the mouth along opening coordinate x. Moreover, we find that helix 3 is already straightened in this state, reflecting step (3). Only by reaching state $\mathbf{4}_{\mathrm{E}}$, though, the cogwheel-like motion of the hydrophobic core and helices 3 and 5 (steps (2) and (3)) can be completed, with Phe4 moving into the solvent. Finally, transition $\mathbf{4}_{\mathrm{E}} \to \mathbf{4}_{\mathrm{R}}$ describes the stabilization of the closed state (step (5)), which typically takes longer than the nanosecond timescale shown in Fig. 3a.

To illustrate the heterogeneity of the open↔closed transitions and discuss the possible pathways of the reaction, Fig. S5 shows additional transition events apart from the one shown in Fig. 3a. Moreover, we calculated the resulting transition matrix $T(\tau_{\mathrm{lag}})$ containing the probabilities T_{ij} that the system jumps from state i to j within lag time τ_{lag}. To distinguish completed transitions from transient fluctuations, dynamical coring[35] was applied which requires that the trajectory spends a minimum time τ_{cor} in the new state for the transition to be counted (Fig. S6). Using $\tau_{\mathrm{cor}} = \tau_{\mathrm{lag}} = 0.7\,\mathrm{ns}$, the resulting transition times $t_{ij} = \tau_{\mathrm{lag}}/T_{ij}$ of the state model are shown in Fig. 3b. We note that the state doublets ($\mathbf{1}_{\mathrm{R}}$, $\mathbf{1}_{\mathrm{E}}$) and ($\mathbf{4}_{\mathrm{R}}$, $\mathbf{4}_{\mathrm{E}}$) interconvert frequently with 5 and 20 ns transition time, respectively. Apart from the main pathway $\mathbf{1}_{\mathrm{R}} \to \mathbf{1}_{\mathrm{E}} \to \mathbf{2} \to \mathbf{4}_{\mathrm{E}} \to \mathbf{4}_{\mathrm{R}}$ discussed above, we find various shortcuts of the reaction that, e.g., omit state $\mathbf{1}_{\mathrm{E}}$ (by going directly from $\mathbf{1}_{\mathrm{R}}$ to $\mathbf{2}$) or omit state $\mathbf{2}$ (by going directly from $\mathbf{1}_{\mathrm{E}}$ to $\mathbf{4}_{\mathrm{E}}$), or omit state $\mathbf{4}_{\mathrm{E}}$ (by going directly from $\mathbf{2}$ to $\mathbf{4}_{\mathrm{R}}$).

Constructing a Markov state model from the transition matrix, we find that its predictions of the state-to-state waiting times are only in qualitative agreement with the corresponding MD results, i.e., they deviate from MD results by a factor of $\sim$2 (Fig. S6). This reflects the fact that during a successful cooperative open-close reaction the various state-to-state transitions occur almost simultaneously (Fig. 3a), i.e., they are not independent from each other[36]. This is opposed to the timescale separation assumed by a Markov state model, where the lag time (representing the randomization time of the fast intrastate dynamics) should be significantly shorter than the transition times of the model.

CONCLUSIONS

To investigate the microscopic mechanism underlying the open-closed transition of T4L, we have combined a correlation analysis of contact features (Fig. 2) with a study of their time evolution (Fig. 3) and a careful visualization of the conformational changes (Fig. 4 and Supplementary Movie S1). The allosteric mechanism in T4L is found to consist of three ingredients.

(i) To start the open-closed transition, a salt bridge between Glu5 and Lys60 needs to open, in order to allow the exit of Phe4 from its hydrophobic pocket (Fig. 4b1). The finalization of the overall transition requires a rearrangement of salt bridges in the mouth region (Fig. 4b4) and a hydrogen bond of helix 1 (Fig. 4b5), which both stabilize the closed state. In the case of the converse reaction, first the latter three interactions are reversed, while the final formation of the Glu5-Lys60 salt bridge stabilizes the closed state. That is, both open and closed states are stabilized by specific interactions, which initially need to be overcome to get the system in a pole position for the desired transition.

(ii) Following this initiation, the hydrophobic core (Phe4, Phe67, Leu7 and Phe104) of the hinge region mediates a cogwheel-like motion of helices 1 and 5 (Fig. 4b2) as well as the straightening of helix 3 (Fig. 4b3). In this way, the hydrophobic core constitute a flexible transmission network, which represents the molecular machine driving the open-closed transition.

(iii) Since helices 3 and 5 are tightly connected to the N- and C-terminal domain, respectively, the motion of the transmission network in the hinge region is directly and instantaneously coupled to the mouth region of T4L. Hence the small local changes of the hydrophobic core are passed on via rigid contacts and the protein backbone to distant sites.

We anticipate that these three components of the process are ubiquitous for allosteric transitions of bistable proteins. For example, a repacking of a hydrophobic core as well as switching of polar interaction partners was reported for the nitrogen regulartory protein C[14].

Moreover, the above described lever-arm effect is reminiscent of structural changes in G protein-coupled receptors, where small changes in the ligand binding site are structurally amplified to large-scale protein surface changes at the G protein binding site[37,38]. A similar effect was also reported for the microbial rhodopsin bacteriorhodopsin, where the retinal cofactor pushing against Trp182 causes a large-scale outward motion of helix F[39], and for the connection between ATP hydrolysis and protein conformational changes in heat shock protein 90[17]. The interplay between rigid secondary elements (such as α-helices and β-sheets) and flexible protein sections (such as loops or linkers) to mediate structural rearrangements was also discussed by Nussinov and Thirumalai[6,40].

The simultaneous involvement of residues of the N- and C-terminal domains as well as of helix 1 demonstrates a high cooperativity of the open-closed transition. This finding is in line with recent unfolding experiments on T4L[23], which revealed that helix 1 is responsible for the high degree of cooperativity within the protein. By coupling the two domains, helix 1 gives rise to an all-or-none, two-state unfolding behavior. It is interesting to note that the cooperativity of the (un)folding of T4L seems to transfer to the cooperativity of its functional dynamics. This insight could allow us to employ –at least to some extent– the well-established theories and concepts of protein folding to the less understood theoretical description of allosteric transitions. In this context we note that recent time-resolved experimental and computational studies of a photoswitchable PDZ2 domain indicated that allosteric communication shares some properties with multistate downhill folding[5,16]. Similar to the various degrees of cooperativity in protein folding[41], allosteric transitions may as well show a variety of mechanisms.

According to the timescale analysis in Fig. 2c, the significant separation between the first and all other timescales clearly suggested a trivial two-state Markov state model based on a single coordinate reflecting the open-close motion. Including five additional coordinates into the model to gain a microscopic understanding of the process, we obtained five conformational states (Fig. 3b), which showed a splitting of the two main open and closed protein states reflecting their stabilization described in (i). Apart from this and a weakly populated optional intermediate state, however, the resulting Markov state model provided only little insight in the cooperative open-close transition. This is because in a cooperative process the various motions occur almost simultaneously (Fig. 3a) and are therefore not independent from each other, which is opposed to the timescale separation assumed by a Markov model.

Cooperative processes such as two-state protein folding, on the other hand, have been successfully described by Ising-type models[42]. Here we consider an Ising model below the critical temperature, where local couplings between two-state entities lead to the emergence of well-defined global orientations of the system. Applied to the functional dynamics of a bistable protein, we associate the local two-state systems with the side-chain contacts, their next-neighbor couplings with interactions with residues in side-chain or backbone proximity, and the resulting global orientations with the open and closed protein conformations, see Fig. 5. In direct analogy to the system's stabilization in a single global orientation by an external field, in allostery a protein state can be stabilized via the (un)binding of a ligand. The intensity of the external field corresponds to the strength of protein-ligand interaction, which may range from weak (causing merely a a population shift of the bound and unbound states) to strong (giving rise to a discrete on-off behavior).

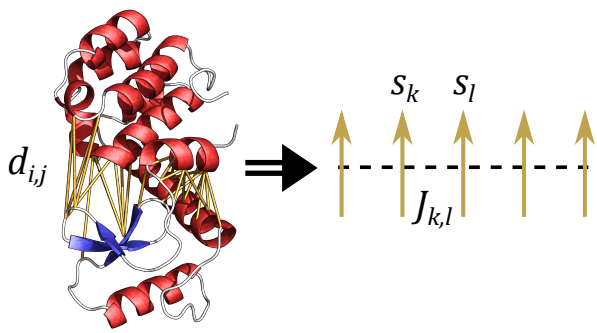

FIG. 5. Scheme of an Ising model to interpret the cooperative transition in T4L. The contact distances $d_{i,j}$ are represented by spin-type variables s_k, which interact with their next neighbors via couplings $J_{k,l}$. Here $s_k = \uparrow$ and $s_k = \downarrow$ means that contact $k = (i, j)$ favors the open and closed protein state, respectively.

The Ising model approach explains the cooperativity of the global open-closed transition via the interaction of local fluctuations[43]. While all local contacts fluctuate on their own, a successful conformational transition

requires the simultaneous flipping of all contacts at the same time. In other words, the system needs to pass through a narrow region in phase space, which constitutes a rare event[44]. In particular, the need of a simultaneous change of numerous contacts explains the long waiting times despite small barriers along individual coordinates (Fig. 1). A similar combination of small barriers and transitions requiring maximal cooperativity between amino acids was found for the folding of small proteins[45]. The fluctuation-driven character of the open-closed transition is also evident in nonequilibrium pulling simulations of Phe4 using dissipation-corrected targeted MD[46]. While the method qualitatively reproduces the free energy profile along the locking coordinate p (Fig. S7), we find exceptionally strong fluctuating forces and (related via the fluctuation-dissipation theorem) friction factors that are two orders of magnitude larger than factors observed in protein-host complexes[47]. Using these friction factors, subsequent Langevin equation simulations[36] recovered the open-closed transition times within a factor ∼2–4, thus confirming that their magnitude is correct.

In conclusion, we have shown for a case study of T4L that the combination of extensive MD simulations, machine-learning assisted dimensionality reduction, visual inspection of the structural changes, and models from protein folding theory may provide a microscopic understanding of the allosteric mechanism in unprecedented detail. As in an Ising model, global conformational rearrangements and allosteric communication over long distances are found to emerge from local fluctuations and short-distance couplings.

METHODS

MD simulations This work is based on the 61 μs-long all-atom MD trajectory of Ernst et al.[26], which used the GROMACS v2016 software package[48] and the Amber99*ILDN force field[49–51]. Moreover, dissipation-corrected targeted MD simulations were performed[46], which employed the Gromacs PULL code[48], and subsequent Langevin simulations[47], see Fig. S7 for details.

Definition of coordinates The Leiden algorithm[32] is a community detection algorithm which identifies subparts of strongly interacting nodes in graphs. For the identification of collective motion in T4L we constructed a dissimilarity-graph in which the nodes represent the coordinates and the edges indicate how uncorrelated they are. For this graph, the Leiden algorithm maximizes an objective function, indicating to what extent the graph exhibits a clusters-like topology. Here we employed the constant Potts model as the objective function $\mathcal{H}_{\mathrm{CPM}} = \sum_c \left[e_c - \gamma \binom{n_c}{2} \right]$, where e_c denotes the sum of edge weights and n_c the number of nodes in cluster c. γ is a resolution parameter that governs the degree to which coordinates within a cluster must be correlated. Setting $\gamma = 0.515$, four larger clusters were found, while the majority of the coordinates are mostly uncorrelated to the rest of the system. Cluster 1 corresponds to 27 contact distances, listed in Table S1, that correlate highly with the open-closed transition.

Characterization of metastable states Robust density-based clustering[34] identifies states based on a local free energy estimate which is calculated for each trajectory point by counting the number of neighbors found inside a d-dimensional hypersphere of fixed radius R centered at the point. We applied this method to the coordinates $x = d_{20,145}$, $p = d_{4,60}$, $d_{5,60}$, $d_{22,137}$, χ_4 and $d_{7,12}$. To ensure that all six coordinates span similar value ranges the angle χ_4 was divided by 180. By using a clustering radius of $R = 0.034$ and by varying the required minimal population $P_{\min}$ of a state, the resulting number of states shows a plateau of 24 states for $100 \leq P_{\min} \leq 200$ (Fig. S2). Of those 24 states, only the first 9 states have populations > 1 %. By inspecting the location of these states in the six-dimensional space, we were able to associate them with one of the 5 coarse-grained states $\mathbf{1}_{\mathrm{R}}$ (microstates 2 and 5), $\mathbf{1}_{\mathrm{E}}$ (microstate 1), $\mathbf{2}$ (microstate 9), $\mathbf{4}_{\mathrm{R}}$ (microstates 4 and 8) and $\mathbf{4}_{\mathrm{E}}$ (microstates 3, 6 and 7). The remaining 15 microstates were assigned to the closest coarse-grained state.

Markov state model We first applied dynamical coring to improve the state definition[35]. Here, short-living fluctuation are removed from the dynamics by requiring for valid transitions that the trajectory spends at least a minimum time τ_{cor} in the new state. To determine τ_{cor}, we considered the probability $P_{\mathrm{stay},i}(t)$ to stay in state i which should decay like a single exponential function. Short-lived fluctuation induce a fast initial decay, which can be removed by using $\tau_{\mathrm{cor}} = 700$ ps. To test the relia-

bility of the MSM, we generated a Markov chain Monte Carlo trajectory and compared the average waiting times of MD and MSM, i.e., the average times $t_{\mathrm{wait,i\to j}}$ needed to reach state j after entering state i.

Data availability All data shown are available from the corresponding authors upon reasonable request.

CONFLICTS OF INTEREST

There are no conflicts of interest to declare.

ACKNOWLEDGEMENTS

The authors thank Peter Hamm, Thorsten Hugel, Peter Bolhuis and Daniel Nagel for fruitful discussions. This work has been supported by the Deutsche Forschungsgemeinschaft (DFG) via the Research Unit FOR 5099 "Reducing complexity of nonequilibrium" (project No. 431945604). The authors acknowledge support by the bwUniCluster computing initiative, the High Performance and Cloud Computing Group at the Zentrum für Datenverarbeitung of the University of Tübingen, the state of Baden-Württemberg through bwHPC and the DFG through grant No. INST 37/935-1 FUGG.

[1] K. Gunasekaran, B. Ma and R. Nussinov, *Proteins*, 2004, **57**, 433–443.
[2] Q. Cui and M. Karplus, *Protein Science*, 2008, **17**, 1295–1307.
[3] H. N. Motlagh, J. O. Wrabl, J. Li and V. J. Hilser, *Nature (London)*, 2014, **508**, 331 – 339.
[4] N. V. Dokholyan, *Chem. Rev.*, 2016, **116**, 6463–6487.
[5] G. Stock and P. Hamm, *Phil. Trans. B*, 2018, **373**, 20170187.
[6] D. Thirumalai, C. Hyeon, P. I. Zhuravlev and G. H. Lorimer, *Chem. Rev.*, 2019, **119**, 6788–6821.
[7] S. J. Wodak, E. Paci, N. Dokholyan, I. N. Berezovsky, A. Horovitz, J. Li, V. J. Hilser, I. Bahar, J. Karanicolas, G. Stock, P. Hamm, R. H. Stote, J. Eberhardt, Y. Chebaro, A. Dejaegere, M. Cecchini, J.-P. Changeux, D. Keri, P. Barth, P. G. Bolhuis, J. Vreede, P. Faccioli, S. Orioli, R. Ravasio, L. Yan, C. Brito, M. Wyart, P. Gkeka, I. Rivalta, G. Palermo, J. A. McCammon, J. Panecka-Hofman, R. C. Wade, A. D. Pizio, M. Y. Niv, R. Nussinov, C.-J. Tsai, H. Jang and T. McLeish, *Structure*, 2019, **27**, 566 – 578.
[8] M. Gruebele, *Protein Folding, Misfolding and Aggregation*, The Royal Society of Chemistry, 2008, pp. 106–138.
[9] K. A. Dill, S. Banu Ozkan, M. Scott Shell and T. R. Weikl, *Annu. Rev. Biophys.*, 2008, **37**, 289 – 316.
[10] S. Brüschweiler, P. Schanda, K. Kloiber, B. Brutscher, G. Kontaxis, R. Konrat and M. Tollinger, *J. Am. Chem. Soc.*, 2009, **131**, 3063–3068.
[11] O. Bozovic, C. Zanobini, A. Gulzar, B. Jankovic, D. Buhrke, M. Post, S. Wolf, G. Stock and P. Hamm, *Proc. Natl. Acad. Sci. USA*, 2020, **117**, 26031 – 26039.
[12] O. Bozovic, J. Ruf, C. Zanobini, B. Jankovic, D. Buhrke, P. J. M. Johnson and P. Hamm, *J. Phys. Chem. Lett.*, 2021, **12**, 4262–4267.
[13] M. D. Vesper and B. L. de Groot, *PLoS Computational Biology*, 2015, **9**, e1003232.
[14] F. Pontiggia, D. Pachov, M. Clarkson, J. Villali, M. Hagan, V. Pande and D. Kern, *Nat. Commun.*, 2015, **6**, 7284.
[15] C. A. Smith, D. Ban, S. Pratihar, K. Giller, M. Paulat, S. Becker, C. Griesinger, D. Lee and B. L. de Groot, *Proc. Natl. Acad. Sci. USA*, 2016, **113**, 3269–3274.
[16] S. Buchenberg, F. Sittel and G. Stock, *Proc. Natl. Acad. Sci. USA*, 2017, **114**, E6804–E6811.
[17] S. Wolf, B. Sohmen, B. Hellenkamp, J. Thurn, G. Stock and T. Hugel, *Chem. Sci.*, 2021, **12**, 3350 – 3359.
[18] M. Bhattacharyya, S. Ghosh and S. Vishveshwara, *Curr. Prot. Pept. Sci.*, 2016, **17**, 4 – 25.
[19] G. R. Bowman, V. S. Pande and F. Noé, *An Introduction to Markov State Models*, Springer, Heidelberg, 2013.
[20] U. Sengupta and B. Strodel, *Phil. Trans. R. Soc. B*, 2018, **373**, 20170178.
[21] M. Dixon, H. Nicholson, L. Shewchuk, W. Baase and B. Matthews, *J. Mol. Biol.*, 1992, **227**, 917 – 933.
[22] X. Zhang, J. A. Wozniak and B. W. Matthews, *J. Mol. Biol.*, 1995, **250**, 527 – 552.
[23] E. Shank, C. Cecconi, J. Dill, S. Marqusee and C. Bustamante, *Nature (London)*, 2010, **465**, 637 – 640.
[24] H. Sanabria, D. Rodnin, K. Hemmen, T.-O. Peulen, S. Felekyan, M. R. Fleissner, M. Dimura, F. Koberling, R. Kühnemuth, W. Hubbell, H. Gohlke and C. A. M. Seidel, *Nat. Commun*, 2020, **11**, 1231.
[25] J. S. Hub and B. L. de Groot, *PLoS Comput. Biol.*, 2009, **5**, e1000480.
[26] M. Ernst, S. Wolf and G. Stock, *J. Chem. Theory Comput.*, 2017, **13**, 5076 – 5088.
[27] S. Brandt, F. Sittel, M. Ernst and G. Stock, *J. Phys. Chem. Lett.*, 2018, **9**, 2144 – 2150.
[28] F. Sittel and G. Stock, *J. Chem. Phys.*, 2018, **149**, 150901.
[29] P. Ravindra, Z. Smith and P. Tiwary, *Mol. Syst. Des. Eng.*, 2020, **5**, 339–348.
[30] R. B. Best and G. Hummer, *Proc. Natl. Acad. Sci. USA*, 2010, **107**, 1088 – 1093.
[31] M. Ernst, F. Sittel and G. Stock, *J. Chem. Phys.*, 2015, **143**, 244114.
[32] V. Traag, L. Waltman and N. van Eck, *Sci. Rep.*, 2019, **9**, 5233.
[33] G. Perez-Hernandez, F. Paul, T. Giorgino, G. De Fabritiis and F. Noé, *J. Chem. Phys.*, 2013, **139**, 015102.
[34] F. Sittel and G. Stock, *J. Chem. Theory Comput.*, 2016, **12**, 2426–2435.
[35] D. Nagel, A. Weber, B. Lickert and G. Stock, *J. Chem. Phys.*, 2019, **150**, 094111.
[36] B. Lickert and G. Stock, *J. Chem. Phys.*, 2020, **153**, 244112.
[37] C.-M. Suomivuori, N. R. Latorraca, L. M. Wingler, S. Eismann, M. C. King, A. L. W. Kleinhenz, M. A. Skiba, D. P. Staus, A. C. Kruse, R. J. Lefkowitz and R. O. Dror, *Science*, 2020, **367**, 881–887.
[38] O. Fleetwood, J. Carlsson and L. Delemotte, *eLife*, 2021, **10**, e60715.
[39] T. Weinert, P. Skopintsev, D. James, F. Dworkowski, E. Panepucci, D. Kekilli, A. Furrer, S. Brünle, S. Mous, D. Ozerov, P. Nogly, M. Wang and J. Standfuss, *Science*, 2019, **365**, 61–65.
[40] E. Papaleo, G. Saladino, M. Lambrughi, K. Lindorff-Larsen, F. L. Gervasio and R. Nussinov, *Chem. Rev.*, 2016, **116**, 6391–6423.
[41] P. Malhotra and J. B. Udgaonkar, *Prot. Sci.*, 2016, **25**, 1924–1941.
[42] E. R. Henry, R. B. Best and W. A. Eaton, *Proc. Natl. Acad. Sci. USA*, 2013, **110**, 17880–17885.
[43] T. C. B. McLeish, T. L. Rodgers and M. R. Wilson, *Phys. Biol.*, 2013, **10**, 056004.
[44] C. Dellago and P. Bolhuis, *Adv. Polymer Sci.*, 2016, **221**,.

[45]S.-H. Chong and S. Ham, *Chem. Sci.*, 2021, **12**, 5944–5951.
[46]S. Wolf and G. Stock, *J. Chem. Theory Comput.*, 2018, **14**, 6175 – 6182.
[47]S. Wolf, B. Lickert, S. Bray and G. Stock, *Nat. Commun.*, 2020, **11**, 2918.
[48]M. J. Abraham, T. Murtola, R. Schulz, S. Pall, J. C. Smith, B. Hess and E. Lindahl, *SoftwareX*, 2015, **1**, 19 – 25.
[49]V. Hornak, R. Abel, A. Okur, B. Strockbine, A. Roitberg and C. Simmerling, *Proteins*, 2006, **65**, 712–725.
[50]R. B. Best and G. Hummer, *J. Phys. Chem. B*, 2009, **113**, 9004–9015.
[51]K. Lindorff-Larsen, S. Piana, K. Palmo, P. Maragakis, J. L. Klepeis, R. O. Dror and D. E. Shaw, *Proteins*, 2010, **78**, 1950–1958.