



OPEN

## Sampling of the conformational landscape of small proteins with Monte Carlo methods

Nana Heilmann, Moritz Wolf, Mariana Kozłowska, Elaheh Sedghamiz, Julia Setzler, Martin Brieg & Wolfgang Wenzel 

Computer simulation provides an increasingly realistic picture of large-scale conformational change of proteins, but investigations remain fundamentally constrained by the femtosecond timestep of molecular dynamics simulations. For this reason, many biologically interesting questions cannot be addressed using accessible state-of-the-art computational resources. Here, we report the development of an all-atom Monte Carlo approach that permits the modelling of the large-scale conformational change of proteins using standard off-the-shelf computational hardware and standard all-atom force fields. We demonstrate extensive thermodynamic characterization of the folding process of the  $\alpha$ -helical Trp-cage, the Villin headpiece and the  $\beta$ -sheet WW-domain. We fully characterize the free energy landscape, transition states, energy barriers between different states, and the per-residue stability of individual amino acids over a wide temperature range. We demonstrate that a state-of-the-art intramolecular force field can be combined with an implicit solvent model to obtain a high quality of the folded structures and also discuss limitations that still remain.

Conformational change is one of the most prominent mechanisms of protein function and regulation<sup>1–4</sup>. The stability of the protein as a whole, as well as the stability of its particular conformational sub-ensembles is essential to understand and regulate protein function. Due to its dynamic nature, it is difficult to observe conformational changes directly at the single molecule level in experiment<sup>5,6</sup>, while computational methods, generally, struggle with the timescales involved<sup>5,7–10</sup>. However, the benefits of the simulation are prominent, as it provides detailed information on the mechanism of the protein folding, and identifies key intermediates and barriers to folding. In the last 20 years, there have been an enormous number of studies focused on the folding of small peptides using specialized force fields and simulation methods<sup>11–18</sup>, reflecting a small subsection of an active field that has evolved over decades. The universal workhorse of all these methods is the molecular dynamics (MD) method that remains constrained by the admissible timestep, which is limited by the fastest frequency of the system, i.e. typically bond-stretch vibrations in the  $10^{14}$  Hz range<sup>19</sup>. At the same time, the vast majority of natural processes have a much larger inherent timescale, ranging from microseconds to seconds<sup>20</sup>.

To capture a single folding event, long MD simulations are required, which incurs either extreme computational cost or the need for a specialized supercomputer architecture<sup>21</sup>. The development of the supercomputer Anton permitted observations of many folding transitions for a range of small fast-folding proteins<sup>22</sup>, and even larger proteins, e.g. G-proteins<sup>23</sup>, using transferable biophysical force fields such as CHARMM and AMBER and explicit solvent models<sup>24,25</sup>. Where such hardware is not available, many strategies have been developed to circumvent the time/length scale problem by either simplifying or coarse graining the force field to accelerate the simulation protocol<sup>26–28</sup>, or to subdivide the simulations in many small non-equilibrium simulations<sup>29</sup>. MD simulations with implicit solvent also enable faster conformational sampling<sup>30</sup>. Indeed, the accuracy of the folding mechanism and the size of the simulated protein are limited, i.e.  $\sim 100$  residues<sup>27,31</sup>, which depend on the protein complexity and the quality of the solvent model<sup>32,33</sup>. Some other methods, such as enhanced sampling techniques<sup>34</sup>, make it possible to reach longer timescales, but none of the above-mentioned approaches offer the same straightforward analysis provided by the “virtual” experiment of simulating the process as it occurs in nature<sup>35</sup>.

As an alternative, Monte Carlo (MC) simulation, which has no inherent timescale, has been explored as a simulation approach. MC simulations yield all thermodynamic data that can be extracted from MD simulations, and permit reconstruction of the kinetic information on long-time scales, such as folding, association or function, but do not provide direct insights into kinetics<sup>36</sup>. Many biological processes can be described as transitions

Institute of Nanotechnology (INT), Karlsruhe Institute of Technology (KIT), Hermann-von-Helmholtz-Platz 1, 76344 Eggenstein-Leopoldshafen, Germany. ✉ email: wolfgang.wenzel@kit.edu

between a few distinct conformational sub-ensembles. The barriers between these ensembles, which are directly sampled in MD, can be computed from thermodynamic averages in MC methods, which permits reconstruction of large scale kinetics even on the basis of a thermodynamic simulation<sup>37</sup>. In MC, special moves may be designed, which do not follow the local force, hence the change of the conformation per energy evaluation in each step may be larger, and the simulation may concentrate on a few particularly important degrees of freedom, such as dihedral angles for peptides. These advantages have the potential to accelerate molecular simulations for peptides and proteins, provided that a suitable force field can be found. Typically, in MC simulation, only a small part of the system in a single move is modulated. Efficient algorithms for multi-particles moves become more and more expensive with the number of molecules, therefore, MC with explicit solvent is particularly problematic<sup>38,39</sup>. Instead, implicit solvation models are well suited for MC simulations and can frequently speed up simulations by orders of magnitude. However, this is often connected to limited hydrogen bonds representation, over-stabilized salt bridges, incorrect ion distribution and neglect of the temperature dependence of the solvation free energy<sup>31,40,41</sup>.

Since the 90 s, the MC folding algorithms have been applied to peptides and proteins using different MC program packages developed to model aggregation and folding/unfolding behavior of peptides. Both, all-atom and coarse grained representations of proteins are used, e.g. as is known for the Rosetta model<sup>42–44</sup>. It implements a knowledge-guided Metropolis Monte Carlo sampling approach using a phenomenological energy function and relies heavily on the data derived from the experimental structure. Therefore, users must incorporate other biochemical information to obtain native-like models, especially for large and complex proteins. For instance, it is difficult to model and design a topology or structure that has never, or only very rarely, been observed in the Protein Data Bank<sup>44</sup>. PROFASI<sup>15</sup>, an all atom MC based C++ code, and SMMP<sup>45</sup>, a FORTRAN based MC code, are also known for simulation of small proteins. They are computationally fast methods and able to capture structural and thermodynamic properties of a diverse set of sequences. At the same time, coarse grained models for protein folding<sup>46,47</sup>, e.g. CABS (C-alpha, beta, and side chain)<sup>48</sup>, which uses various MC schemes, were reported with successful performance in binding studies of intrinsically disordered proteins (IDPs)<sup>49</sup>. By an efficient treatment of large time scale dynamics, they provide significant extension of the structural transitions and better conformational sampling while maintaining sufficient accuracy<sup>47,50</sup>. All atom/coarse grained multi-scale modeling techniques, such as reported by Zacharias et al.<sup>51</sup> and Feig et al.<sup>52</sup> (MMTSB model), applied for the scoring of the protein conformation, peptide folding and prediction of the missing protein fragments, have been also developed<sup>52,53</sup>. All reported MC simulations are limited to employ specifically designed force fields and algorithms, which may impact their common usage.

The improvement and the development of force fields has been an instrumental step in the advances made in protein simulation. It is therefore important to investigate whether state-of-the-art force fields, that were originally designed for explicit solvent simulations, can also be employed in accurate and predictive MC. In this context, two important questions naturally arise in the use of Monte Carlo methods: (1) will the combination of an accurate intramolecular force field, developed for all-atom MD simulations, together with the implicit solvent models yield to quantitative results, and (2) can the free-energy landscape be sampled sufficiently well, relying on simplified moves defined in MC protocols independently on the forces on the atoms?

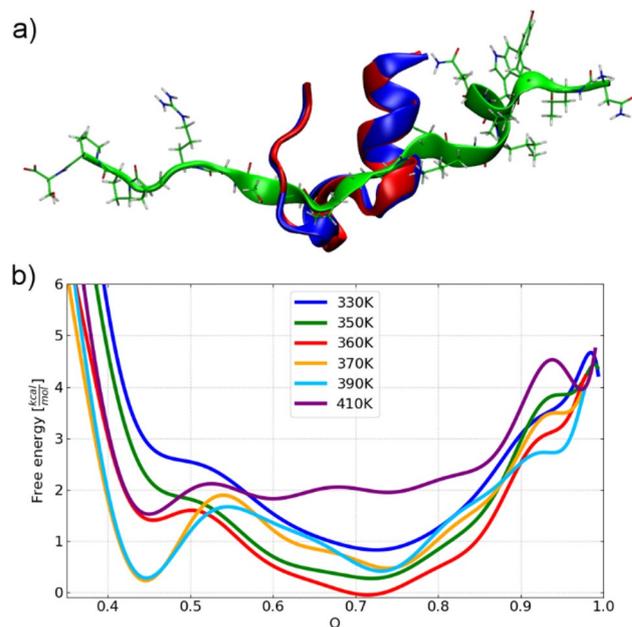
In the following, we aim to answer these two questions by employing a Monte Carlo based protocol<sup>54</sup>, using an accurate implicit solvent model and a transferable all-atom intramolecular AMBER99SB\*-ILDN force field<sup>55</sup>. This force field, in the most cases in the combination with explicit water, has been shown to perform well in mimicking experimental data using MD simulations of different peptides<sup>24,55</sup>. Here, we show the sampling of the conformational landscape of three conventional peptides: the 20 amino acid Trp-cage miniprotein (PDB code 1L2Y)<sup>56</sup>, the Villin headpiece (PDB code 1vii) comprising of 36 residues (including N-terminal methionine)<sup>57</sup>, and the 35-residue WW-domain (PDB code 2f21)<sup>58</sup>. All of the proteins belong to widely studied all-helical and  $\beta$ -stranded mini-proteins<sup>42,59–62</sup>. We focus on the reproducing of their folding free energy landscapes, barriers, and transition states in order to demonstrate thermodynamic characterization of small proteins using Monte Carlo simulations with an all-atom force field.

## Results and discussion

We investigated the conformational landscape of three, well studied, proteins of different size and tertiary structure: Trp-cage<sup>56</sup>, Villin headpiece<sup>57</sup> and WW-domain<sup>58</sup>, using Monte Carlo simulations starting both from folded and unfolded structures. For each protein we sampled a wide temperature range to characterize the folding and unfolding equilibria. The computational methodology used is described in detail in section “Methods”. We start with the Trp-cage protein, being the smallest peptide investigated, and then focus on the Villin headpiece and WW-domain to show transferability and efficiency of the all-atom force field in Monte Carlo approach.

**Trp-cage protein.** The Trp-cage protein (PDB 1L2Y)<sup>56</sup> has been of high interest for both experimentalists and theoreticians as this short, 20 amino acid, protein is a fast folding protein (ca. 4.1  $\mu$ s), enabling the introduction of different protein mutations to understand the ways of enhancing protein stability or improving drug binding efficiency (e.g., in treatment of type II diabetes mellitus)<sup>63,64</sup>. Moreover, this miniprotein has been used for the last two decades to benchmark force fields and modeling techniques against detailed structural, thermodynamic and kinetic data<sup>65–74</sup>. MC simulations of Trp-cage in the current study were performed in the temperature range of 330 K–410 K and sampling of 200 Million MC steps on the AMD EPYC 7551P node using 15 or 30 cores required for 181 h and 108 h of CPU time, respectively.

Trp-cage consists of a short  $\alpha$ -helix between residues 2–9, a single turn of  $3_{10}$ -helix (residues 11–14), and a hydrophobic core made of proline residues (Pro-12, Pro-18, Pro-19) and Tyr-3, Trp-6 (see the structure in blue in Fig. 1a). Its folding is known to be modulated by cooperative interactions between water molecules and polar



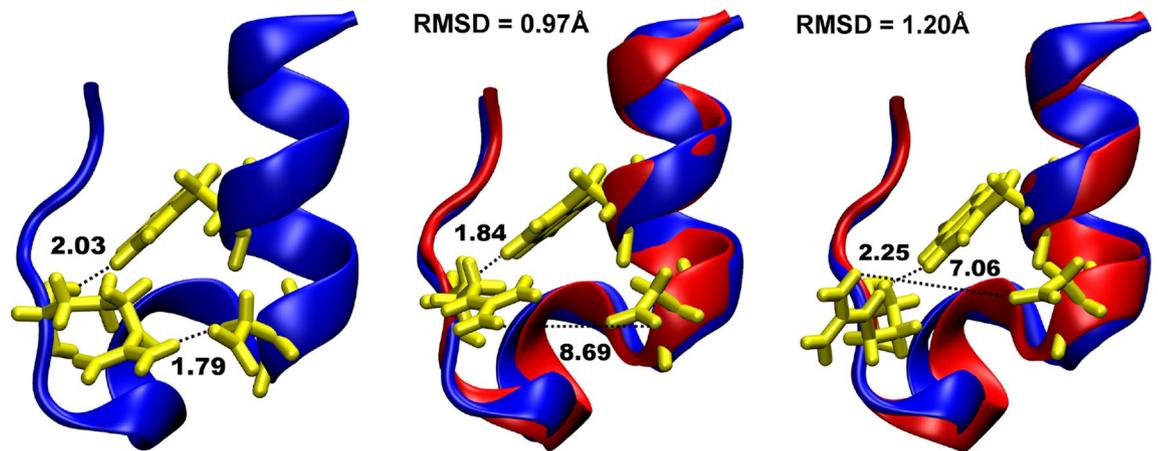
**Figure 1.** Sampling of the conformational landscape of the Trp-cage miniprotein. **(a)** Overlay of the native structure of 1L2Y (in blue) compared to the refolded structure (in red) obtained in MC simulation run at 370 K (full MC trajectory is given in Fig. S1a) started from the unfolded protein structure marked in green. **(b)** Free energy profiles as a function of the reaction coordinate  $Q$  (fraction of native contacts) at different temperatures computed from the Monte Carlo simulations. Refolded and intermediate ensembles were found at  $Q \sim 0.73$  and  $0.45$ , respectively. Visualization was done in VMD (version 1.9.2beta1) <https://www.ks.uiuc.edu/Research/vmd/>.

groups of a protein, therefore, proper treatment of the solvation environment is essential to get the correct folding behavior<sup>75–77</sup>. Indeed, several implicit solvent models, used in all-atom MD simulations, have been reported to yield correct refolded structures of Trp-cage<sup>78–80</sup>. Here, we use the AMBER99SB\*-ILDN force field with the generalized Born based implicit solvent model with a solvent accessible surface area term for nonpolar solvation effects (see “Methods”). Figure 1a depicts the overlay of the native (in blue) and the refolded (in red) structure of 1L2 after its full unfolding (in green). The respective MC simulation run was performed at 370 K (see Fig. S1a) starting from the unfolded Trp-cage with the fraction of native contacts of 0.07. The refolded structures match good with the native Trp-cage (Ca-RMSD of the refolded protein, depicted in red in Fig. 1, is of 0.86 Å) with small deviations around the helix turn. This shows high quality of the force field and the sampling efficiency of the MC algorithm with the accumulated acceptance ratio of 60%.

To establish the folding temperature of the protein, the free energy profile was calculated using the potential of mean force (PMF) projected on the fraction of native contacts,  $Q$ . This measure is a widely used reaction coordinate for the folding process<sup>77,81</sup> (see Fig. 1b). It reflects the similarity between the native and the predicted structure of the protein, i.e.  $Q \sim 0.9–1.0$  is the closest to the native structure obtained in NMR. Two main states of the conformational ensemble describing its folded ( $Q \sim 0.73$ ) and unfolded states were observed. We also find a partially unfolded state ( $Q \sim 0.45$ ) with an energy barrier of 1.20 and 0.80 kcal·mol<sup>-1</sup> at 370 K and 390 K, respectively. Similar states of Trp-cage with the free energy barrier of 0.80 kcal·mol<sup>-1</sup> were reported by Zhou<sup>82</sup> using highly parallel replica exchange MD with explicit solvation and in experiment<sup>83</sup>. We note that the primary minimum at  $Q \sim 0.73$  is rather broad, which results from the relatively weak stabilization of the native structure in this small protein because of the implicit solvation. This was overcome in MD with explicit water, where  $Q \sim 0.9–1.0$  of the folded Trp-cage was reported<sup>75</sup>. Nevertheless, the absolute value of the  $Q$  at the minimum depends on the details of the definition of the native contacts taken from the NMR ensemble.

We also observe the broad range of the folding temperature of the protein, starting from 370 K. At low temperature, i.e. 330 K, we observe essentially only a single minimum, which gives way to a free energy surface with two minima around 350 K. The estimated folding temperature, i.e. the temperature when both minima are equally probable, is significantly higher than was experimentally observed, i.e. 315–317 K<sup>64,69</sup>, or calculated using all-atom force fields with explicit solvation, i.e. 321–326 K<sup>76,84</sup>, but it is in line with the folding temperatures obtained using implicit solvent models (375–400 K)<sup>85,86</sup>. This may result from the lack of the temperature dependence of the implicit solvent model<sup>31</sup>. The generalized Born solvent-accessible surface area (GBSA)-type implicit solvation models, used here, are known to over-stabilize the folded states of proteins, especially those stabilized by solvent-exposed salt bridges<sup>85,87,88</sup>. The breaking/formation of such H-bonded salt bridge, formed between Asp-9 and Arg-16 (N-H ··· O bond of 1.79 Å in Fig. 2) outside of the central hydrophobic core region exposed to the solvent, regulates Trp-cage folding and refolding, inducing the observed increase in the folding temperature.

A detailed analysis of the refolded Trp-cage demonstrates the accuracy of the force field and MC approach with respect to structure. Several refolded structures are depicted in Figs. 2 and S2. The Ca-RMSD of the refolded



**Figure 2.** Local minima representing refolded conformers of the Trp-cage in MC simulation started from the unfolded protein structure with  $Q=0.07$  at 370 K (see Fig. S1a). The native structure is in blue, while refolded representatives are in red. Asp-9 and Arg-16 residues forming hydrogen-bonded salt bridge (with the length in the native state of 1.79 Å) and Trp-6 forming hydrogen-bond with Arg-16 (with the length in the native state of 2.03 Å) are colored in yellow. Visualization was done in VMD (version 1.9.2beta1) <https://www.ks.uiuc.edu/Research/vmd/>.

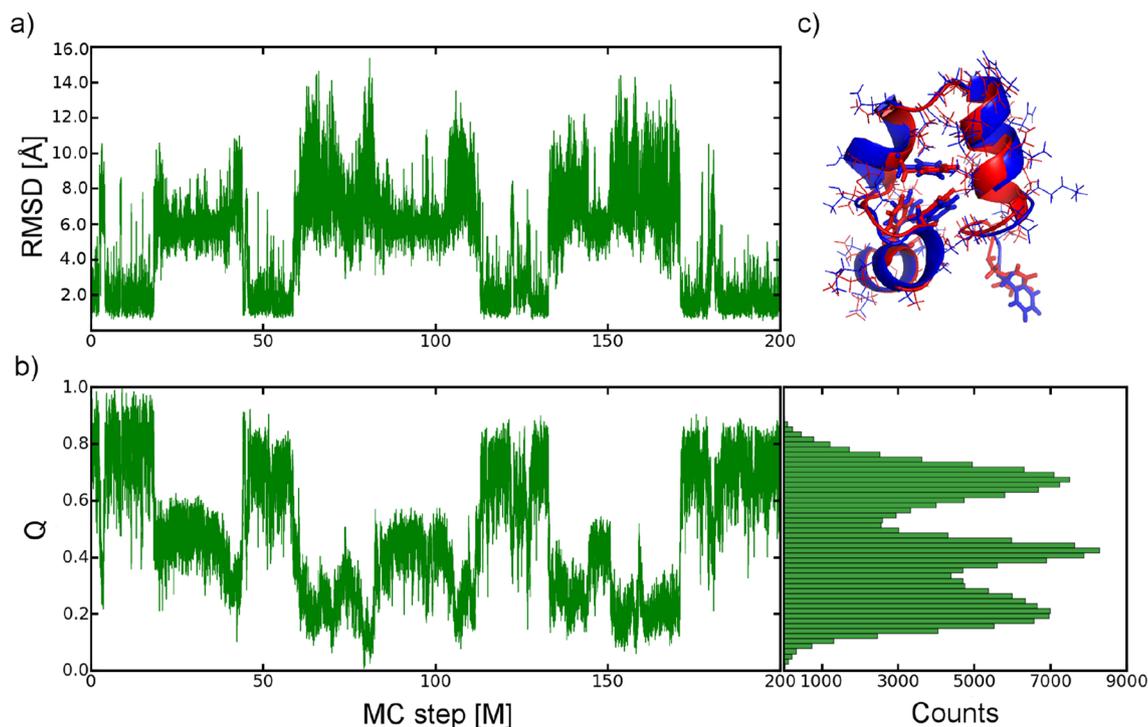
protein is 0.97 Å with the average  $C\alpha$ -RMSD deviation during the MC run, started from the unfolded structure (shown in green in Fig. 1a), is 1.96 (0.44) Å. Secondary structure of both,  $\alpha$ -helix and proline-end, are reproduced correctly in the refolded protein (see Fig. 2) with backbone RMSD of 0.73 Å and 0.47 Å, respectively. This is connected to the helical propensity of the helical domain of Trp-cage, which was shown to be stable even at 400 K in MD simulations with explicit water, which results from the force field parameterization that is largely helix-based<sup>89</sup>. Among the most flexible residues in the refolded protein, marked in Fig. S2, is Arg16, which participates in the salt bridge formation of the Trp-cage. This results in the lack of salt bridge formation in the refolded structures, as shown in Fig. 2, where the  $H\cdots O$  distance between Asp-9 and Arg-16 was found to be far larger than the native one. This result probably originates from the drawbacks of the implicit solvation model, as mentioned above. Even though the salt bridge is unstable in our MC simulations, the refolded structures of Trp-cage conserve the two main secondary elements, with the most notable difference in a  $3_{10}$ -helix-like turn, and form the tertiary structure due to the stability of other H-bonding interactions present in Trp-cage. Among them is a  $N-H\cdots O$  hydrogen bond between Trp-6 (H-bond donor) and Arg-16 (H-bond acceptor) with the length of 2.03 Å in the NMR-structure (marked in Fig. 2), which is stable in the refolded Trp-cage (see Fig. S2). Together with the salt bridge between Asp-9 and Arg-16, this H-bond regulates fast folding of Trp-cage<sup>78,89</sup>.

**Villin headpiece.** Next, we generated multiple long MC trajectories of the Villin headpiece, known as HP-36<sup>57</sup>: the smallest autonomously folded protein without disulfide bonds, oligomerization or stabilizing ligands, at a wide range of temperatures (360–460 K). The Villin headpiece consists of three helices: (i) between residues 4–8, (ii) 15–18 and (iii) 23–30, as depicted in Fig. 3c. Loop, turn and a closely packed hydrophobic core held the helices together in a compact structure. Most MD simulations observe one and/or two, folding pathways of HP-36, where helix2 serves as the structural starter of the Villin folding through intermediate and transition states to the folded structure<sup>8</sup>. Recently, Wang et al. have shown the third folding pathway, which starts from the hydrophobic core or/and helix3<sup>90</sup>.

The main aim of our analysis was to predict the near-native refolded structure of Villin using efficient sampling of its conformational space performing MC simulations with all-atom force field. Our MC algorithm results in the multiple transitions between different states of the protein (see Fig. 3a,b), enabling quantitative prediction of the protein thermodynamics. The quality of the refolded structures is good, we find an all-atom/backbone RMSD of 1.49 Å/0.76 Å after refolding to the NMR structure<sup>57</sup>, and unfolded conformations with RMSD > 12 Å (all-atom/backbone RMSD) are observed (see Fig. 3a). Excluding the first three unstructured residues, the refolded conformations agree to an all-atom/backbone RMSD of 1.12 Å/0.46 Å. In the high RMSD ensemble, all elements of the tertiary structure and significant fractions of the secondary structure are lost. Refolded conformations, see Fig. 3c, completely recover the native secondary and tertiary structure.

Similarly to the longest, presently available, molecular dynamics simulations<sup>21,22,90,91</sup>, in all simulations we observe multiple folding and unfolding events, which we use to extract thermodynamic information via the PMF projected on the fraction of the native contacts,  $Q$  (Fig. 3b). The free energy profile obtained, depicted in Fig. 4a, features three distinct minima corresponding to the native ensemble N ( $Q \sim 0.8$ ), the denatured ensemble D ( $Q \sim 0.2$ ) and a folding intermediate I ( $Q \sim 0.45$ ). The existence of intermediate conformations is difficult to observe both in experiment and theory, but they have been clearly identified through kinetic analysis<sup>91</sup>.

To demonstrate the shift of the folding temperature as the result of the implicit solvation, as we have shown also for Trp-cage, we computed the specific heat capacity of the HP-36 as a function of temperature (see Fig. S4) and performed temperature calibration of the MC simulations by 83 K. In such a way, we established the folding



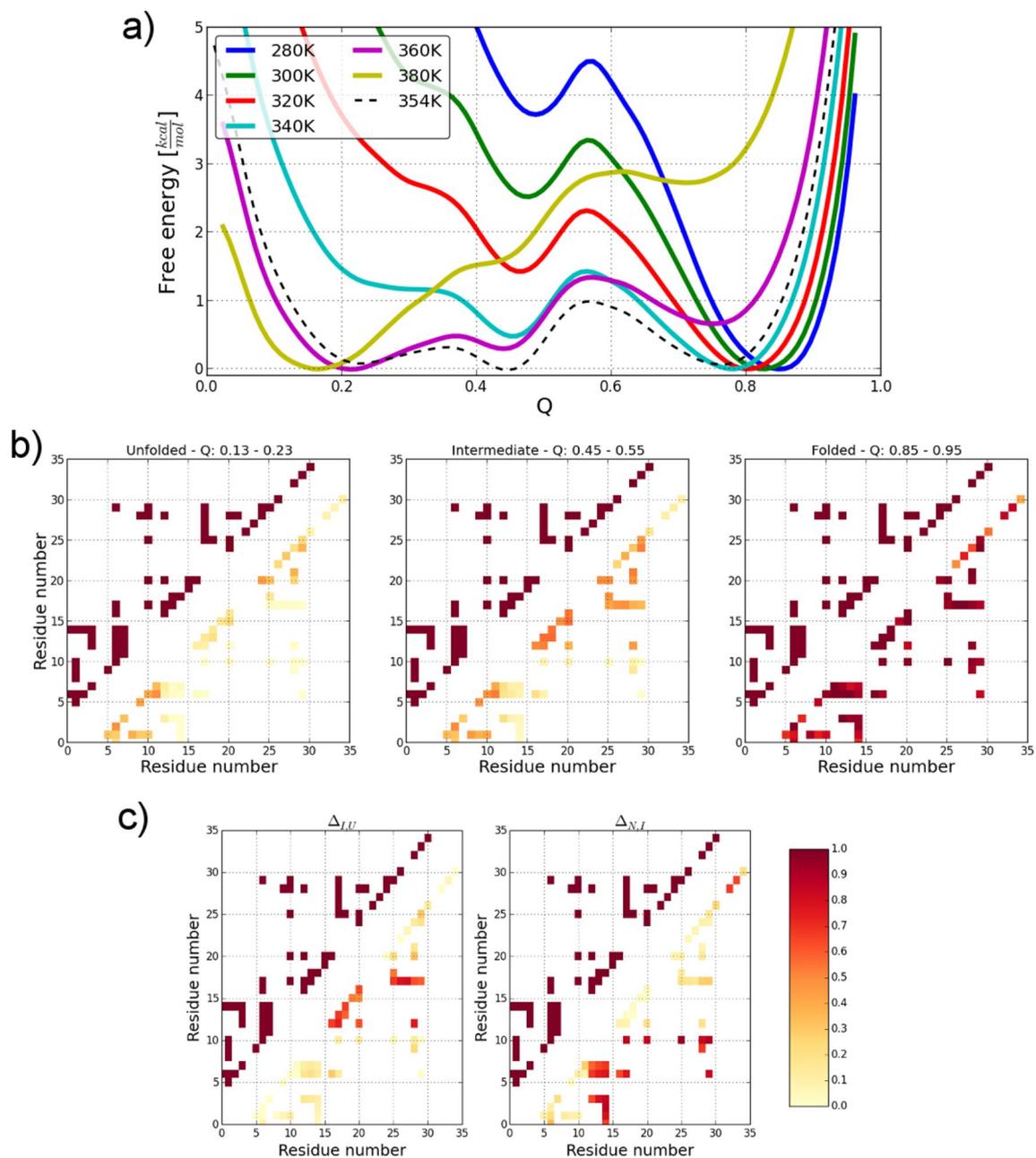
**Figure 3.** Section of a trajectory of the Villin headpiece simulations started from the native 1VII at 360 K (including temperature shift shown in Fig. S4): **(a,b)** RMSD and fraction of the native contacts,  $Q$ , as a function of step-size showing many transitions between well-defined native and non-native ensembles. At the right side in **(b)**, the occupancy of the sub-ensembles, corresponding to the folded, intermediate and unfolded conformations are shown. **(c)** Overlay of the refolded (in red,  $Q=0.8$ ) and the experimental (in blue) conformation of the protein. Visualization was done in VMD (version 1.9.2beta1) <https://www.ks.uiuc.edu/Research/vmd/>.

temperature of the Villin headpiece as 354 K (see Fig. 4a), which is only 12 K shifted in comparison to the experimental value (342 K)<sup>57</sup>.

At this temperature, the native state is stabilized by the energy barrier of  $\Delta G_{N,I} = 1$  kcal·mol<sup>-1</sup>, what is in agreement with the experimental data and all-atom simulations using the same intramolecular force field and explicit water<sup>8,20,91,92</sup>. We also observe smaller energy barrier between the intermediate and the denatured state, i.e.  $\Delta G_{I,N} = 0.4$  kcal·mol<sup>-1</sup> (as also known from experiment)<sup>91</sup>.

To better understand the most significant structural changes and residues responsible for the protein folding, we performed the detailed analysis of each of the three ensembles extracted from the trajectories. In Fig. 4b, the contact maps of the unfolded, intermediate, and folded ensemble are given, while in panel in Fig. 4c, the differences between the unfolded and intermediate, as well as, the intermediate and folded ensembles are demonstrated. In the folded ensemble (last panel in Fig. 4b, refolded structure with the lowest C $\alpha$ -RMSD), we see a good agreement between the contacts of the folded ensemble in the simulation and the native contacts derived from the NMR structure. The unfolded ensemble shows residual secondary structure with significantly decreased probability, in comparison to the other ensembles, and essentially no tertiary contacts. As the fraction of native contacts increases, helix2 forms, but few tertiary contacts are present. This is better seen in the difference maps between the unfolded and the intermediate (left panel in Fig. 4c) and the intermediate and folded ensemble (right panel in Fig. 4c). There, large changes in the stability of tertiary contacts finally lead to the stabilization of the native conformation.

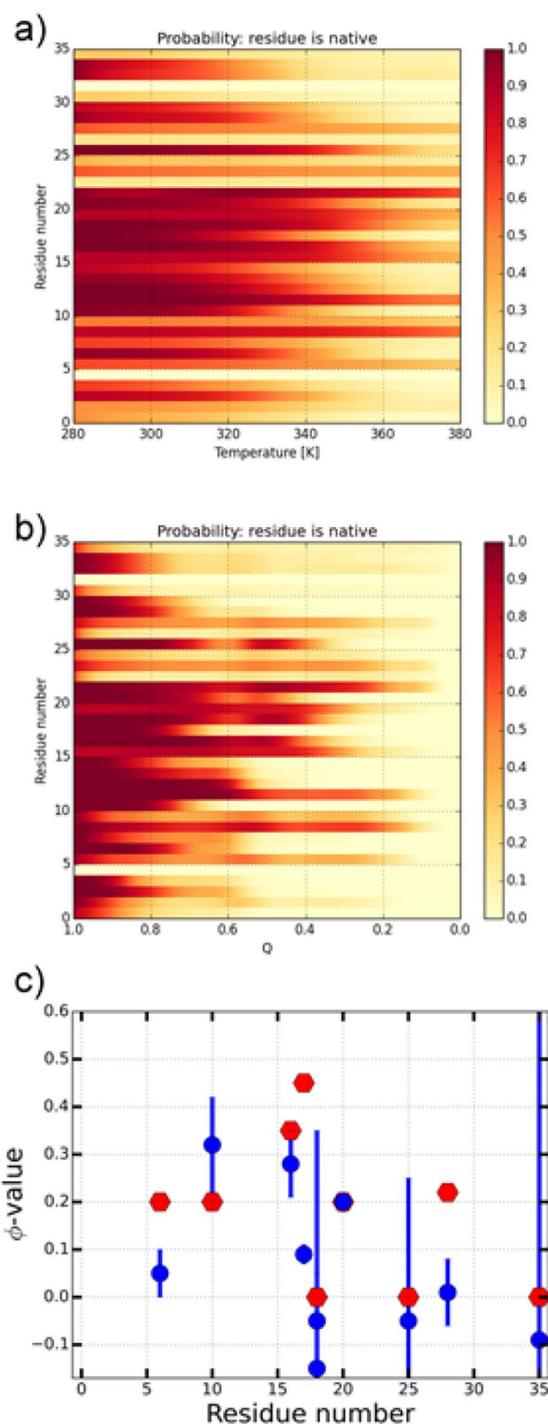
The stability of the individual residues, measured by the presence of local contacts, as a function of temperature (Fig. 5a), offers a view on the parts of the protein relevant for protein folding and their thermal stability with detailed structural information on protein folding transition states. The complementary per-residue stability of individual residues at the transition temperature of 354 K, as a function of the reaction coordinate  $Q$ , is shown in Fig. 5b. There, the transition state ensemble near  $Q \sim 0.6$  is clearly visible as a light vertical band, where native contacts are diminished partially reforming in the intermediate state. Identification of the transition state ensemble between the intermediate and the native conformation permits a computation of the  $\phi$ -values, which are the measure of the presence of native contacts in the transition state.  $\phi$ -values are descriptors widely used as experimental perturbation (mutation) to probe the free energy landscape. They are also used to check on simulation accuracy<sup>93</sup>. As shown in Fig. 5c, the computed  $\phi$ -values using our approach (marked in red), show good agreement with experimental observations<sup>91</sup>. Combining the data obtained, the contribution of individual amino acids to protein stability is analyzed in detail: Residue Ala-18, for example, is found to be thermally highly stable (averaged over all conformations, Fig. 5a), but diminishes significantly in its native environment in the



**Figure 4.** Analysis of conformational landscape of the Villin headpiece. **(a)** Free energy profiles as a function of the reaction coordinate  $Q$  at different temperatures and at the interpolated folding temperature (marked with dashed line). Native, intermediate, and unfolded ensembles were found at  $Q \sim 0.8, 0.45, 0.2$ , respectively. **(b)** Contact maps of the unfolded, intermediate, and folded ensemble (left to right) weighted by the occurrence in the ensemble (bottom right half of the figures) in comparison to the native contacts derived from the NMR structure (top left parts of the figures). **(c)** Contact difference maps showing the changes in secondary and tertiary contacts in the unfolded-intermediate transition and intermediate-native transition.

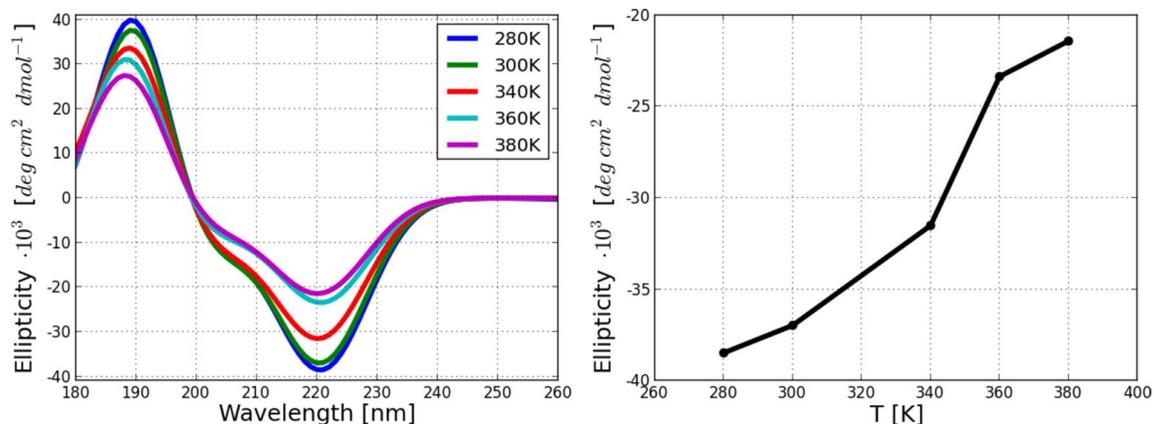
transition state (light band at  $Q \sim 0.6$  in Fig. 5b), in comparison to both, the native and intermediate ensemble, where its  $\phi$ -value is correspondingly low.

The folding equilibrium at a wide range of temperatures can be also characterized by investigating the helical content of a protein via the circular dichroism (CD) spectroscopy<sup>94</sup>. In order to directly compare our data with experiment, we have computed the ellipticity of Villin headpiece, shown in Fig. 6 (see section “Methods” for details). Strong peaks in all three  $\alpha$ -helical bands at 190 nm, 208 nm and 220 nm have been found. The signals weaken as a function of temperature, confirming denaturation of the protein, being in a good agreement with experimental measurements<sup>57,91</sup> and MD simulations<sup>92</sup>. The temperature dependence of ellipticity as a function of temperature (panel on the right in Fig. 6) demonstrates the ellipticity decrease with temperature increase since all three helices lose their stability.

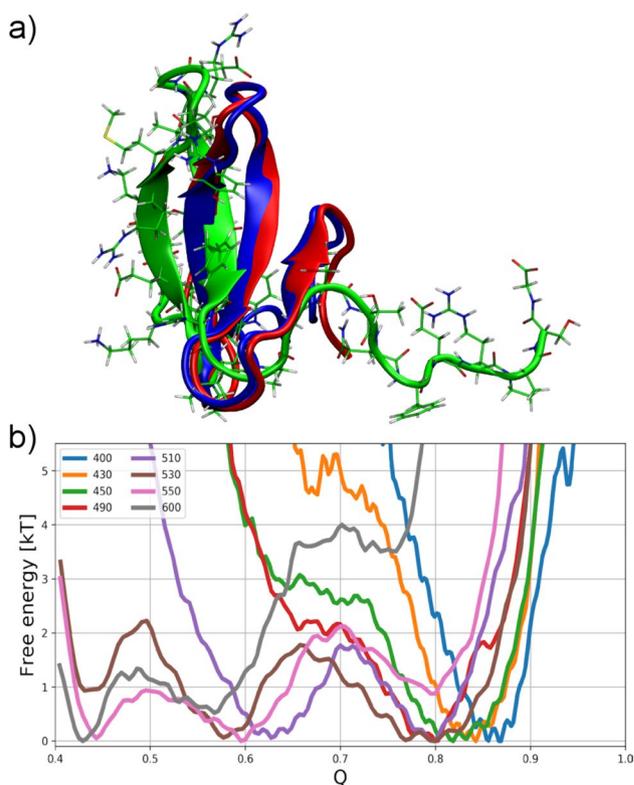


**Figure 5.** Thermal stability of Villin headpiece from MC simulations. **(a)** Probability the residue is in its native environment as a function of temperature. **(b)** Probability the residue is in its native environment as a function of established native contacts  $Q$ . White horizontal stripes in the data occur for residues that have few native contacts. **(c)**  $\phi$ -values at the folding temperature, obtained from the Boltzmann-weighted fraction of native contacts for each residue for all ensembles at and close to the transition state barrier. Values in blue correspond to experimental values at 310 K<sup>70</sup>, red symbols result from the simulation.

**WW domain.** We have observed that the folding of both Trp-cage and Villin headpiece, which are alpha-helical proteins, was in good agreement with the native structures. As a third example, we investigated the folding of a  $\beta$ -sheet-containing protein, i.e. the WW-domain, shown in blue in Fig. 7a. The WW domain is often used as a model protein in the investigation of  $\beta$ -sheet folding<sup>58,95,96</sup>. We performed folding simulations of the



**Figure 6.** Calculated circular dichroism (CD) spectra of Villin headpiece. Left panel: Ellipticity as a function of wavelength at different temperatures of MC simulation. Right panel: ellipticity at 220 nm as a function of temperature.



**Figure 7.** Conformational space of the WW domain in the MC simulations started from the native structure (see Fig. S6). (a) Representative structures of the intermediate energy minimum at  $Q=0.44$  (in green), refolded protein at  $Q=0.80$  (in red) compared to the native structure (in blue). Other refolded local minima are depicted in Fig. S5. (b) The free energy landscape of the WW domain mutant for different temperatures as a function of the fraction of established native contacts  $Q$  computed from the Monte Carlo simulations. Visualization was done in VMD (version 1.9.2beta1) <https://www.ks.uiuc.edu/Research/vmd/>.

GTT mutant of the WW-domain<sup>97</sup> in the regime  $1 > Q > 0.4$  at the temperature range of 400–600 K (every 20 K), which allows enhanced sampling of the refolding events between the native and the intermediate configurations. In most cases, the outer strands of the WW-domain unfold, while the core stays intact. We restrained the  $Q$  range for this protein because we observed that simulations that reached  $Q < 0.4$  did not return to  $Q > 0.4$  within the allocated computational time. This may be related to the fact that for  $Q < 0.4$  most of the  $\beta$ -sheet character is lost and the nucleation of  $\beta$ -sheets from random coil structures is a rare event. In the following, we therefore confine the investigation to the refolding of the native configuration from partially unfolded configurations which retain some  $\beta$ -sheets (see below). The issue of nucleation of  $\beta$ -sheets from random coils has been addressed elsewhere<sup>98</sup>. It is known that the  $\beta$ -sheet nucleus is stabilized by a solvent exposed N–H $\cdots$ O hydrogen

bond between Ser-13 to Arg-17 (see Fig. S7) which may be difficult to stabilize on its own in an implicit solvation model<sup>99</sup>. It has also been reported that implicit solvation models struggle with a correct description of the transition state of WW domain<sup>100</sup>. To overcome this limitation, we confined the simulation to the window in  $Q$  reported above, corresponding to the intermediate ensemble through implemented a repulsive potential in  $Q$  that drives the simulations back to  $Q > 0.4$ .

We performed MC simulations for a set of different temperatures to estimate the equilibrium folding temperature. The eight resulting free energy landscapes are shown in Fig. 7b. The minimum of the free energy surface with the highest fraction of established native contacts ( $Q \sim 0.80$ – $0.85$ ) corresponds to the global free energy minimum at low temperatures. By increasing the temperature up to 510 K, a second minimum is observed at  $Q \sim 0.6$ , where the outer two strands of the WW domain detach from the protein core.

A further minimum in the free energy landscape of the protein, i.e. at  $Q \sim 0.45$ , is observed at around 530 K. At higher temperatures, the protein is more prone to stay in the unfolded state. The simulations at 530–550 K result in multiple partial unfolding and refolding of the WW domain even during Monte Carlo simulations with 64 million MC steps (see Fig. S6). The refolded structures of the protein, for example at 530 K, with the fraction of the native contacts of  $Q \sim 0.8$  (see Fig. 7b), are in a good agreement with the reported crystal structure of WW domain (PDB code 2f21)<sup>80</sup> with the C $\alpha$ -RMSD  $< 2.0$  Å, as shown in Fig. 7a and S5. The equilibrium folding temperature was found at 530 K, again higher than known from experiment (345 K) and MD simulations (395 K)<sup>98,101</sup>. At the same time, we observe the free energy barriers of folding up to 2kBT (see Fig. 7b), i.e. up to 2.1 kcal·mol<sup>-1</sup> at 530 K, similarly as was reported by Shaw et al.<sup>22</sup>.

## Discussion

In summary, we have found that the combination of a state-of-the-art intramolecular forcefield with an accurate implementation of an implicit, physics-based solvent model leads to reproducible refolding of the Trp-cage, Villin headpiece and WW domain proteins. While the refolded structures are in good agreement with the experimental data, in comparison to other MC predictions, and can be recovered from unfolded configurations for helical proteins and partially unfolded  $\beta$ -stranded structures, the folding temperatures are systematically overestimated. Our data show that using off-the-shelf computational hardware and the combination of the intramolecular part of the all-atom AMBER99ILDN\* force field with an implicit solvent model can characterize the relevant states of helical proteins with sufficient accuracy. The reasons for the deviation in the folding temperatures are presently unclear. While the implicit solvent model has no temperature dependence, it should perform well near room temperature where most of these proteins fold. An overestimation of the folding temperature in simulation<sup>85–88</sup> with an intramolecular force field that performs better in explicit solvent, means that the entropy of the unfolded configurations is underestimated. Implementation of other force fields, e.g. CHARMM with CMAP correction may improve dynamical and structural properties of proteins in their unfolded state, thus, increase the quality of MC sampled configurations<sup>102</sup>. Further studies will consider also the folding behavior of large proteins, which may differ significantly from those studied here.

One remaining problem is therefore the accuracy of the implicit solvent model that lacks temperature dependence and the proper description of the solvent exposed hydrogen bonding, which may lead to folding at the elevated temperatures<sup>103</sup>. Moreover, the differences may arise from the imbalance of the intramolecular energies of the force field calibrated for explicit water simulations and the implicit solvent model. The use of the recently reported implicit solvent models, e.g. ff14SBonlysc + GB-Neck2<sup>31,33,104</sup>, where advanced fitting of GB solvation energies and the relative solvation energies to Poisson-Boltzmann method for a set of proteins and peptides has been made, should be tested. Improvement in the accuracy of the solvation energies and effective radii may result in better agreement of conformational sampling in comparison to MD with explicit solvation than in GBSA. Moreover, the computational cost of more advanced models like the three dimensional reference interaction site implicit solvent model (3D-RISM)<sup>105</sup>, especially with the closure relation proposed by Kovalenko and Hirata (3D-RISM-KH)<sup>105–108</sup>, need to be considered. The latest version of this approach operates with the solvent representation by the spatial distributions of the solvent molecules around a solute macromolecule, therefore results in better solvation structure of a protein and its thermodynamics. To our knowledge, the accuracy of the 3D-RISM-KH was demonstrated for the folding of the miniprotein 1L2Y and protein G<sup>106</sup>. Further investigations regarding the efficiency and accuracy of implicit solvent models are needed, which further improve the MC approach using the standard off-the-shelf computational hardware and standard all-atom force fields as demonstrated in the present report.

## Conclusions

We have demonstrated that Monte Carlo simulations make it possible to efficiently sample the conformational landscape of the folding of small proteins using standard hardware without the need of extreme high-performance computing. There are two contributing factors that make the reported simulations fast. First: the efficient implementation of an implicit solvent model that significantly reduces the number of degrees of freedom (presently MC simulations are not feasible in explicit water, because there are no efficient collective moves for all the water coordinates). Secondly: acceleration by the usage of the Monte Carlo algorithm with its large effective time step, i.e. near the folding equilibrium temperature, the trajectories show a transition approximately every  $5 \times 10^7$  energy evaluations for the Villin headpiece. Correlating this with the experimental folding time, a single MC step covers the same distance in conformational space as an MD simulation of  $5 \times 10^{-13}$  s. The “time step” in MC is, thus, about two orders of magnitude larger than the typical MD timestep. More improvement of the current model is needed, including implementation of the algorithm for GPU acceleration.

## Methods

**Force field.** The simulations were performed with the AMBER99SB\*-ILDN force field<sup>40</sup> and an implicit solvent model consisting of a generalized Born (GB)<sup>80,109</sup> term that models polar solvation effects and a solvent accessible surface area (SASA)<sup>110</sup> term that models nonpolar solvation effects. The Born radii were computed with the PowerBorn method<sup>111</sup> and the SASA with the PowerSASA method<sup>97</sup>. The dielectric constant of the protein was taken to be  $\epsilon_p = 1$  and that of water  $\epsilon_w = 80$ . The surface tension of the nonpolar solvation term was  $\gamma = 5.42 \text{ cal/mol}\cdot\text{\AA}^2$ . No long-range cutoffs or approximate methods were used in the evaluation of the force field or implicit solvent model. In Monte Carlo simulations sometimes, large moves on atoms or groups of atoms are proposed that lead to near-zero atomic distances, which leads to infinite potential energies. To enhance the numerical stability of the simulation, the distance computation between atoms was modified, and a constant offset of 0.001 Å was added to each interatomic distance for the computation of the Coulomb, Lennard–Jones, and GB terms. This small offset has negligible numerical effects in the low energy regions, but modifies the force field in the unphysical cases when atoms are clashing.

**Simulation protocol.** All calculations were performed with the SIMONA<sup>54</sup> code, which is available under <https://www.int.kit.edu/nanosim/simona>. The Trp-cage, Villin Headpiece and WW domain simulations were based on the pdb-files with codes 1L2Y<sup>56</sup>, 1VII<sup>57</sup> and 2F21<sup>58</sup>, respectively. The force field parameters were assigned with the pdb2gmx program of Gromacs<sup>112</sup>. The structure was minimized with Gromacs<sup>19</sup> and relaxed with backbone or sidechain moves using SIMONA at 50 K. The structure with the lowest energy was used as the reference for the native structure. We performed five to ten simulations each comprising 200 million MC-steps at different simulation temperatures, depending on the protein, i.e. 330–450 K, 360–460 K and 400–600 K for Trp-cage, Villin Headpiece and WW domain, respectively. An individual MC step comprises either a randomly selected backbone and sidechain dihedral rotation or a concerted move with equal probability. The angle change in the dihedral moves was drawn from a Gaussian distribution with a width of 18.3° for Villin Headpiece and 20° for Trp-cage and WW domain. In a “concerted move” a segment of 4 amino acids modified, changing all dihedral angles under the constraint that the endpoints of the segment do not change. In addition, rigid body rotations were applied by rotating the molecule around a random axis through its geometric center with a uniformly distributed rotation angle of up to 5°. The Metropolis acceptance criterion with Markov chain model was used to construct collective moves with the acceptance probability of 0.6 (60%), preserving detailed balance. The simulations operated at an effective time step of 260 fs/MC step, accelerating the sampling of the conformational space by about two orders of magnitude over all-atom explicit-solvent MD simulations. The first 10% of the steps of each simulation was discarded to permit equilibration. The ellipticity and CD spectra were computed with the analyses program CdPro<sup>113</sup>. Visualization of proteins was done using VMD<sup>114</sup>.

Received: 15 June 2020; Accepted: 12 October 2020

Published online: 23 October 2020

## References

- Henzler-Wildman, K. & Kern, D. Dynamic personalities of proteins. *Nature* **450**, 964–972 (2007).
- Eisenmesser, E. Z. *et al.* Intrinsic dynamics of an enzyme underlies catalysis. *Nature* **438**, 117–121 (2005).
- Bock, L. V. *et al.* Energy barriers and driving forces in tRNA translocation through the ribosome. *Nat. Struct. Mol. Biol.* **20**, 1390–1396 (2013).
- Lange, O. F. *et al.* Recognition dynamics up to microseconds revealed from an RDC-derived ubiquitin ensemble in solution. *Science* **320**, 1471–1475 (2008).
- Snow, C. D., Nguyen, H., Pande, V. S. & Gruebele, M. Absolute comparison of simulated and experimental protein-folding dynamics. *Nature* **420**, 102–106 (2002).
- Cecconi, C. Direct observation of the three-state folding of a single protein molecule. *Science* **309**, 2057–2060 (2005).
- Duan, Y. Pathways to a protein folding intermediate observed in a 1-microsecond simulation in aqueous solution. *Science* **282**, 740–744 (1998).
- Lei, H., Wu, C., Liu, H. & Duan, Y. Folding free-energy landscape of villin headpiece subdomain from molecular dynamics simulations. *Proc. Natl. Acad. Sci. USA* **104**, 4925–4930 (2007).
- Herges, T. & Wenzel, W. Free-energy landscape of the Villin headpiece in an all-atom force field. *Structure* **13**, 661–668 (2005).
- Schug, A., Herges, T. & Wenzel, W. Reproducible protein folding with the stochastic tunneling method. *Phys. Rev. Lett.* **91**, 158102 (2003).
- Vila, J. A., Ripoll, D. R. & Scheraga, H. A. Atomically detailed folding simulation of the B domain of staphylococcal protein A from random structures. *Proc. Natl. Acad. Sci. USA* **100**, 14812–14816 (2003).
- Hills, R. D., Lu, L. & Voth, G. A. Multiscale coarse-graining of the protein energy landscape. *PLoS Comput. Biol.* **6**, e1000827 (2010).
- Suárez, E., Adelman, J. L. & Zuckerman, D. M. Accurate estimation of protein folding and unfolding times: beyond Markov state models. *J. Chem. Theory Comput.* **12**, 3473–3481 (2016).
- Irbäck, A., Mitternacht, S. & Mohanty, S. An effective all-atom potential for proteins. *PMC Biophys.* **2**, 2 (2009).
- Irbäck, A. & Mohanty, S. PROFASI: A Monte Carlo simulation package for protein folding and aggregation. *J. Comput. Chem.* **27**, 1548–1555 (2006).
- Ding, F., Tsao, D., Nie, H. & Dokholyan, N. V. Ab initio folding of proteins with all-atom discrete molecular dynamics. *Structure* **16**, 1010–1018 (2008).
- Mohanty, S., Meinke, J. H., Zimmermann, O. & Hansmann, U. H. E. Simulation of Top7-CFR: A transient helix extension guides folding. *Proc. Natl. Acad. Sci. USA* **105**, 8004–8007 (2008).
- Meinke, J. H. & Hansmann, U. H. E. Free-energy-driven folding and thermodynamics of the 67-residue protein GS- $\alpha_3$  W-A large-scale Monte Carlo study. *J. Comput. Chem.* **30**, 1642–1648 (2009).
- Hess, B., Kutzner, C., van der Spoel, D. & Lindahl, E. GROMACS 4: algorithms for highly efficient, load-balanced, and scalable molecular simulation. *J. Chem. Theory Comput.* **4**, 435–447 (2008).

20. Veitshans, T., Klimov, D. & Thirumalai, D. Protein folding kinetics: timescales, pathways and energy landscapes in terms of sequence-dependent properties. *Fold. Des.* **2**, 1–22 (1997).
21. Lindorff-Larsen, K., Piana, S., Dror, R. O. & Shaw, D. E. How fast-folding proteins fold. *Science* **334**, 517–520 (2011).
22. Shaw, D. E. *et al.* Atomic-level characterization of the structural dynamics of proteins. *Science* **330**, 341–346 (2010).
23. Dror, R. O. *et al.* Structural basis for nucleotide exchange in heterotrimeric G proteins. *Science* **348**, 1361–1365 (2015).
24. Lindorff-Larsen, K. *et al.* Systematic validation of protein force fields against experimental data. *PLoS ONE* **7**, e32131 (2012).
25. Lee, H. Structures, dynamics, and hydrogen-bond interactions of antifreeze proteins in TIP4P/Ice water and their dependence on force fields. *PLoS ONE* **13**, e0198887 (2018).
26. de Jong, D. H. *et al.* Improved parameters for the martini coarse-grained protein force field. *J. Chem. Theory Comput.* **9**, 687–697 (2013).
27. Nguyen, H., Maier, J., Huang, H., Perrone, V. & Simmerling, C. Folding simulations for proteins with diverse topologies are accessible in days with a physics-based force field and implicit solvent. *J. Am. Chem. Soc.* **136**, 13959–13962 (2014).
28. Garcia, A. E. & Onuchic, J. N. Folding a protein in a computer: An atomic description of the folding/unfolding of protein A. *Proc. Natl. Acad. Sci. USA* **100**, 13898–13903 (2003).
29. Noé, F., Schütte, C., Vanden-Eijnden, E., Reich, L. & Weikl, T. R. Constructing the equilibrium ensemble of folding pathways from short off-equilibrium simulations. *Proc. Natl. Acad. Sci. USA* **106**, 19011–19016 (2009).
30. Ferrara, P., Apostolakis, J. & Caflisch, A. Evaluation of a fast implicit solvent model for molecular dynamics simulations. *Proteins Struct. Funct. Genet.* **46**, 24–33 (2002).
31. Shao, Q. & Zhu, W. How well can implicit solvent simulations explore folding pathways? A quantitative analysis of  $\alpha$ -helix bundle proteins. *J. Chem. Theory Comput.* **13**, 6177–6190 (2017).
32. Shao, Q. & Zhu, W. Assessing AMBER force fields for protein folding in an implicit solvent. *Phys. Chem. Chem. Phys.* **20**, 7206–7216 (2018).
33. Tao, P. & Xiao, Y. Using the generalized Born surface area model to fold proteins yields more effective sampling while qualitatively preserving the folding landscape. *Phys. Rev. E* **101**, 062417 (2020).
34. Harada, R. & Shigeta, Y. Temperature-shuffled structural dissimilarity sampling based on a root-mean-square deviation. *J. Chem. Inf. Model.* **58**, 1397–1405 (2018).
35. McCammon, J. A. & Karplus, M. Simulation of protein dynamics. *Annu. Rev. Phys. Chem.* **31**, 29–45 (1980).
36. Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. & Teller, E. Equation of state calculations by fast computing machines. *J. Chem. Phys.* **21**, 1087–1092 (1953).
37. Liang, F. & Wong, W. H. Evolutionary Monte Carlo for protein folding simulations. *J. Chem. Phys.* **115**, 3374–3380 (2001).
38. Nerenberg, P. S. & Head-Gordon, T. New developments in force fields for biomolecular simulations. *Curr. Opin. Struct. Biol.* **49**, 129–138 (2018).
39. Perez, A., Morrone, J. A., Simmerling, C. & Dill, K. A. Advances in free-energy-based simulations of protein folding and ligand binding. *Curr. Opin. Struct. Biol.* **36**, 25–31 (2016).
40. Kleinjung, J. & Fraternali, F. Design and application of implicit solvent models in biomolecular simulations. *Curr. Opin. Struct. Biol.* **25**, 126–134 (2014).
41. Anandakrishnan, R., Drozdetski, A., Walker, R. C. & Onufriev, A. V. Speed of conformational change: comparing explicit and implicit solvent molecular dynamics simulations. *Biophys. J.* **108**, 1153–1164 (2015).
42. Shimada, J. & Shakhnovich, E. I. The ensemble folding kinetics of protein G from an all-atom Monte Carlo simulation. *Proc. Natl. Acad. Sci. USA* **99**, 11175–11180 (2002).
43. Cragnell, C., Durand, D., Cabane, B. & Skeep, M. Coarse-grained modeling of the intrinsically disordered protein Histatin 5 in solution: Monte Carlo simulations in combination with SAXS: Monte Carlo simulations in combination with SAXS. *Proteins Struct. Funct. Bioinform.* **84**, 777–791 (2016).
44. Rohl, C. A., Strauss, C. E. M., Misura, K. M. S. & Baker, D. Protein structure prediction using rosetta. *Methods Enzymol.* **383**, 66–93 (2004).
45. Eisenmenger, F., Hansmann, U. H. E., Hayryan, S. & Hu, C.-K. [SMMP] A modern package for simulation of proteins. *Comput. Phys. Commun.* **138**, 192–212 (2001).
46. Khalili, M., Liwo, A. & Scheraga, H. A. Kinetic studies of folding of the b-domain of staphylococcal protein A with molecular dynamics and a united-residue (UNRES) model of polypeptide chains. *J. Mol. Biol.* **355**, 536–547 (2006).
47. Liwo, A. *et al.* A unified coarse-grained model of biological macromolecules based on mean-field multipole–multipole interactions. *J. Mol. Model.* **20**, 2306 (2014).
48. Pulawski, W., Jamroz, M., Kolinski, M., Kolinski, A. & Kmiecik, S. Coarse-grained simulations of membrane insertion and folding of small helical proteins using the CABS model. *J. Chem. Inf. Model.* **56**, 2207–2215 (2016).
49. Kurcinski, M., Kolinski, A. & Kmiecik, S. Mechanism of folding and binding of an intrinsically disordered protein as revealed by ab initio simulations. *J. Chem. Theory Comput.* **10**, 2224–2231 (2014).
50. Kmiecik, S. *et al.* Coarse-grained protein models and their applications. *Chem. Rev.* **116**, 7898–7936 (2016).
51. Fiorucci, S. & Zacharias, M. Binding site prediction and improved scoring during flexible protein-protein docking with ATTRACT. *Proteins Struct. Funct. Bioinform.* **78**, 3131–3139 (2010).
52. Feig, M., Karanicolas, J. & Brooks, C. L. MMTSB Tool Set: enhanced sampling and multiscale modeling methods for applications in structural biology. *J. Mol. Graph. Model.* **22**, 377–395 (2004).
53. Nastica-Labouze, J., Meli, M., Derreumaux, P., Colombo, G. & Mousseau, N. A Multiscale approach to characterize the early aggregation steps of the amyloid-forming peptide GNNQQNY from the yeast prion Sup-35. *PLoS Comput. Biol.* **7**, e1002051 (2011).
54. Strunk, T. *et al.* SIMONA 1.0: An efficient and versatile framework for stochastic simulations of molecular and nanoscale systems. *J. Comput. Chem.* **33**, 2602–2613 (2012).
55. Lindorff-Larsen, K. *et al.* Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins* **78**, 1950–1958 (2010).
56. Neidigh, J. W., Fesinmeyer, R. M. & Andersen, N. H. Designing a 20-residue protein. *Nat. Struct. Biol.* **9**, 425–430 (2002).
57. McKnight, C. J., Matsudaira, P. T. & Kim, P. S. NMR structure of the 35-residue villin headpiece subdomain. *Nat. Struct. Biol.* **4**, 180–184 (1997).
58. Jager, M. *et al.* Structure-function-folding relationship in a WW domain. *Proc. Natl. Acad. Sci. USA* **103**, 10648–10653 (2006).
59. Xu, J., Huang, L. & Shakhnovich, E. I. The ensemble folding kinetics of the FBP28 WW domain revealed by an all-atom Monte Carlo simulation in a knowledge-based potential: Folding Kinetics of FBP28 WW Domain. *Proteins* **79**, 1704–1714 (2011).
60. Hansmann, U. H. E. Protein-folding simulations in generalized ensembles. *Int. J. Quantum Chem.* **90**, 1515–1521 (2002).
61. Mohanty, S., Irbäck, A., Mitternacht, S., Favrin, G. & Hansmann, U. H. E. Protein folding, aggregation and unfolding in Monte Carlo simulations. *Phys. Procedia* **7**, 68–71 (2010).
62. Irbäck, A. & Mohanty, S. Folding thermodynamics of peptides. *Biophys. J.* **88**, 1560–1569 (2005).
63. Drucker, D. J. *et al.* Exenatide once weekly versus twice daily for the treatment of type 2 diabetes: a randomised, open-label, non-inferiority study. *The Lancet* **372**, 1240–1250 (2008).
64. Barua, B. *et al.* The Trp-cage: optimizing the stability of a globular miniprotein. *Protein Eng. Des. Sel.* **21**, 171–185 (2008).

65. Simmerling, C., Strockbine, B. & Roitberg, A. E. All-atom structure prediction and folding simulations of a stable protein. *J. Am. Chem. Soc.* **124**, 11258–11259 (2002).
66. Snow, C. D., Zagrovic, B. & Pande, V. S. The Trp cage: folding kinetics and unfolded state topology via molecular dynamics simulations. *J. Am. Chem. Soc.* **124**, 14548–14549 (2002).
67. Chowdhury, S., Lee, M. C., Xiong, G. & Duan, Y. Ab initio folding simulation of the Trp-cage mini-protein approaches NMR resolution. *J. Mol. Biol.* **327**, 711–717 (2003).
68. Ding, F., Buldyrev, S. V. & Dokholyan, N. V. Folding Trp-cage to NMR resolution native structure using a coarse-grained protein model. *Biophys. J.* **88**, 147–155 (2005).
69. Streicher, W. W. & Makhatazde, G. I. Unfolding thermodynamics of Trp-Cage, a 20 residue miniprotein, studied by differential scanning calorimetry and circular dichroism spectroscopy<sup>†</sup>. *Biochemistry* **46**, 2876–2880 (2007).
70. Ulmschneider, J. P., Ulmschneider, M. B. & Di Nola, A. Monte Carlo vs molecular dynamics for all-atom polypeptide folding simulations. *J. Phys. Chem. B* **110**, 16733–16742 (2006).
71. Kentsis, A., Gindin, T., Mezei, M. & Osman, R. Calculation of the free energy and cooperativity of protein folding. *PLoS ONE* **2**, e446 (2007).
72. Fossat, M. J., Kitazawa, S., McCallum, S., Garcia, A. & Royer, C. A Novel Trp cage conformer revealed by combining high pressure NMR and MD simulations. *Biophys. J.* **112**, 61a (2017).
73. Meshkin, H. & Zhu, F. Atomic simulations of Trp-cage folding by umbrella sampling using Q function as reaction coordinate. *Biophys. J.* **116**, 46a–47a (2019).
74. Gupta, M., Nayar, D., Chakravarty, C. & Bandyopadhyay, S. Comparison of hydration behavior and conformational preferences of the Trp-cage mini-protein in different rigid-body water models. *Phys. Chem. Chem. Phys.* **18**, 32796–32813 (2016).
75. Best, R. B. & Mittal, J. Balance between  $\alpha$  and  $\beta$  structures in Ab initio protein folding. *J. Phys. Chem. B* **114**, 8790–8798 (2010).
76. Paschek, D., Day, R. & Garcia, A. E. Influence of water–protein hydrogen bonding on the stability of Trp-cage miniprotein. A comparison between the TIP3P and TIP4P-Ew water models. *Phys. Chem. Chem. Phys.* **13**, 19840 (2011).
77. Cho, S. S., Levy, Y. & Wolynes, P. G. P versus Q: Structural reaction coordinates capture protein folding on smooth landscapes. *Proc. Natl. Acad. Sci. USA* **103**, 586–591 (2006).
78. Hu, Z., Tang, Y., Wang, H., Zhang, X. & Lei, M. Dynamics and cooperativity of Trp-cage folding. *Arch. Biochem. Biophys.* **475**, 140–147 (2008).
79. Chen, J., Im, W. & Brooks, C. L. Balancing solvation and intramolecular interactions: toward a consistent generalized born force field. *J. Am. Chem. Soc.* **128**, 3728–3736 (2006).
80. Im, W., Lee, M. S. & Brooks, C. L. Generalized born model with a simple smoothing function. *J. Comput. Chem.* **24**, 1691–1702 (2003).
81. Šali, A., Shakhnovich, E. & Karplus, M. How does a protein fold?. *Nature* **369**, 248–251 (1994).
82. Zhou, R. Trp-cage: folding free energy landscape in explicit water. *Proc. Natl. Acad. Sci. USA* **100**, 13280–13285 (2003).
83. Rovó, P. *et al.* Structural insights into the Trp-Cage folding intermediate formation. *Chem. Eur. J.* **19**, 2628–2640 (2013).
84. Day, R., Paschek, D. & Garcia, A. E. Microsecond simulations of the folding/unfolding thermodynamics of the Trp-cage mini-protein. *Proteins* **78**, 1889–1899 (2010).
85. Pitera, J. W. & Swope, W. Understanding folding and design: Replica-exchange simulations of “Trp-cage” miniproteins. *Proc. Natl. Acad. Sci. USA* **100**, 7587–7592 (2003).
86. Duan, L. *et al.* Simulation of the thermodynamics of folding and unfolding of the Trp-cage mini-protein TC5b using different combinations of force fields and solvation models. *Sci. China Chem.* **53**, 196–201 (2010).
87. Bursulaya, B. D. & Brooks, C. L. Comparative study of the folding free energy landscape of a three-stranded  $\beta$ -sheet protein with explicit and implicit solvent models. *J. Phys. Chem. B* **104**, 12378–12383 (2000).
88. Zhou, R. & Berne, B. J. Can a continuum solvent model reproduce the free energy landscape of a-hairpin folding in water?. *Proc. Natl. Acad. Sci. USA* **99**, 12777–12782 (2002).
89. Paschek, D., Nymeyer, H. & Garcia, A. E. Replica exchange simulation of reversible folding/unfolding of the Trp-cage miniprotein in explicit solvent: on the structure and possible role of internal water. *J. Struct. Biol.* **157**, 524–533 (2007).
90. Wang, E., Tao, P., Wang, J. & Xiao, Y. A novel folding pathway of the villin headpiece subdomain HP35. *Phys. Chem. Chem. Phys.* **21**, 18219–18226 (2019).
91. Kubelka, J., Henry, E. R., Cellmer, T., Hofrichter, J. & Eaton, W. A. Chemical, physical, and theoretical kinetics of an ultrafast folding protein. *Proc. Natl. Acad. Sci.* **105**, 18655–18662 (2008).
92. Piana, S., Lindorff-Larsen, K. & Shaw, D. E. Protein folding kinetics and thermodynamics from atomistic simulation. *Proc. Natl. Acad. Sci. USA* **109**, 17845–17850 (2012).
93. Harada, R. & Kitao, A. The fast-folding mechanism of Villin headpiece subdomain studied by multiscale distributed computing. *J. Chem. Theory Comput.* **8**, 290–299 (2012).
94. Sreerama, N., Venyaminov, S. Y. U. & Woody, R. W. Estimation of the number of  $\alpha$ -helical and  $\beta$ -strand segments in proteins using circular dichroism spectroscopy. *Protein Sci.* **8**, 370–380 (2008).
95. Davis, C. M. & Dyer, R. B. The role of electrostatic interactions in folding of  $\beta$ -proteins. *J. Am. Chem. Soc.* **138**, 1456–1464 (2016).
96. Wirth, A. J., Liu, Y., Prigozhin, M. B., Schulten, K. & Gruebele, M. Comparing fast pressure jump and temperature jump protein folding experiments and simulations. *J. Am. Chem. Soc.* **137**, 7152–7159 (2015).
97. Klenin, K. V., Tristram, F., Strunk, T. & Wenzel, W. Derivatives of molecular surface area and volume: simple and exact analytical formulas. *J. Comput. Chem.* **32**, 2647–2653 (2011).
98. Piana, S. *et al.* Computational design and experimental testing of the fastest-folding  $\beta$ -sheet protein. *J. Mol. Biol.* **405**, 43–48 (2011).
99. Deechongkit, S. *et al.*  $\beta$ -Sheet folding mechanisms from perturbation energetics. *Curr. Opin. Struct. Biol.* **16**, 94–101 (2006).
100. Fuller, A. A. *et al.* Evaluating  $\gamma$ -turn mimics as  $\beta$ -sheet folding nucleators. *Proc. Natl. Acad. Sci. USA* **106**, 11067–11072 (2009).
101. Robustelli, P., Piana, S. & Shaw, D. E. Developing a molecular dynamics force field for both folded and disordered protein states. *Proc. Natl. Acad. Sci. USA* **115**, E4758–E4766 (2018).
102. Buck, M., Bouguet-Bonnet, S., Pastor, R. W. & MacKerell, A. D. Importance of the CMAP correction to the CHARMM22 protein force field: dynamics of Hen lysozyme. *Biophys. J.* **90**, L36–L38 (2006).
103. Anandakrishnan, R., Izadi, S. & Onufriev, A. V. Why computed protein folding landscapes are sensitive to the water model. *J. Chem. Theory Comput.* **15**, 625–636 (2019).
104. Nguyen, H., Roe, D. R. & Simmerling, C. Improved generalized born solvent model parameters for protein simulations. *J. Chem. Theory Comput.* **9**, 2020–2034 (2013).
105. Hirata, F. *Molecular Theory of Solvation* (Springer, Netherlands, 2004).
106. Omelyan, I. & Kovalenko, A. MTS-MD of biomolecules steered with 3D-RISM-KH mean solvation forces accelerated with generalized solvation force extrapolation. *J. Chem. Theory Comput.* **11**, 1875–1895 (2015).
107. Kovalenko, A. & Hirata, F. Self-consistent description of a metal–water interface by the Kohn-Sham density functional theory and the three-dimensional reference interaction site model. *J. Chem. Phys.* **110**, 10095–10112 (1999).
108. Kovalenko, A. Multiscale modeling of solvation in chemical and biological nanosystems and in nanoporous materials. *Pure Appl. Chem.* **85**, 159–199 (2013).

109. Feig, M. & Brooks, C. L. Recent advances in the development and application of implicit solvent models in biomolecule simulations. *Curr. Opin. Struct. Biol.* **14**, 217–224 (2004).
110. Lee, B. & Richards, F. M. The interpretation of protein structures: Estimation of static accessibility. *J. Mol. Biol.* **55**, 379 (1971).
111. Brieg, M. & Wenzel, W. PowerBorn: a Barnes–Hut tree implementation for accurate and efficient born radii computation. *J. Chem. Theory Comput.* **9**, 1489–1498 (2013).
112. Pronk, S. *et al.* GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit. *Bioinformatics* **29**, 845–854 (2013).
113. Sreerama, N., Venyaminov, S.Yu. & Woody, R. W. Estimation of protein secondary structure from circular dichroism spectra: inclusion of denatured proteins with native proteins in the analysis. *Anal. Biochem.* **287**, 243–251 (2000).
114. Humphrey, W., Dalke, A. & Schulten, K. VMD: visual molecular dynamics. *J. Mol. Graphics* **14**, 33–38 (1996).

## Acknowledgements

This research has been funded by the BMBF within projects “Rational Design of Peptide-Surface Interactions” and “Molecular Interaction Engineering (IM-1)” and the Deutsche Forschungsgemeinschaft (DFG) priority programme DiSPBiotech (SPP1934) within the project A9. We acknowledge the financial support by the DFG under Germany Excellence Strategy via the Excellence Cluster 3D Matter Made to Order (Grant No. EXC-2082/1–390761711). M.K. acknowledges funding by the Ministry of Science, Research and Art of Baden-Württemberg (Germany) under Brigitte-Schlieben-Lange-Programm. Authors are grateful to Dr. Timo Strunk (Nanomatch GmbH) for fruitful discussions and SIMONA technical support. This work was performed on the supercomputer ForHLR funded by the Ministry of Science, Research and the Arts Baden-Württemberg and by the Federal Ministry of Education and Research and supercomputer resources of HERMIT based at HLRS Stuttgart within PRACE. We acknowledge support by the KIT-Publication Fund of the Karlsruhe Institute of Technology.

## Author contributions

W.W. developed conceptualization of the article; N.H., M.W., M.B. developed methodology and made code implementations; M.K., J.S., N.H. and M.W. performed simulations; N.H., M.W., M.K., E.S., J.S., M.B. and W.W. analyzed the data; N.H., M.W., M.K., J.S. visualized the results obtained; N.H., M.K., E.S., J.S. and W.W. wrote the manuscript; M.K., E.S., W.W. reviewed and edited the manuscript. Supervision, project administration and funding acquisition were done by W.W. All authors discussed the results and contributed to the final manuscript.

## Funding

Open Access funding enabled and organized by Projekt DEAL.

## Competing interests

The authors declare no competing interests.

## Additional information

**Supplementary information** is available for this paper at <https://doi.org/10.1038/s41598-020-75239-7>.

**Correspondence** and requests for materials should be addressed to W.W.

**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).

**Publisher’s note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2020