The Journal of Chemical Physics

Selection of representative structures from large biomolecular ensembles

Cite as: J. Chem. Phys. **156**, 144102 (2022); https://doi.org/10.1063/5.0082444
Submitted: 15 December 2021 • Accepted: 02 March 2022 • Accepted Manuscript Online: 04 March 2022 • Published Online: 12 April 2022

Arthur Voronin and Alexander Schug

The Journal of Chemical Physics ARTICLE scitation.org/journal/jcp

Selection of representative structures from large biomolecular ensembles

Cite as: J. Chem. Phys. 156, 144102 (2022); doi: 10.1063/5.0082444
Submitted: 15 December 2021 • Accepted: 2 March 2022 •
Published Online: 12 April 2022

Arthur Voronin[1,2] and Alexander Schug[3,4,a)]

AFFILIATIONS

[1] Steinbuch Centre for Computing, Karlsruhe Institute of Technology, Eggenstein-Leopoldshafen, Germany
[2] Department of Physics, Karlsruhe Institute of Technology, Karlsruhe, Germany
[3] Jülich Supercomputing Center, Institute for Advanced Simulation, Jülich, Germany
[4] Faculty of Biology, University of Duisburg-Essen, Duisburg, Germany

[a)] **Author to whom correspondence should be addressed:** al.schug@fz-juelich.de

ABSTRACT

Despite the incredible progress of experimental techniques, protein structure determination still remains a challenging task. Due to the rapid improvements of computer technology, simulations are often used to complement or interpret experimental data, particularly for sparse or low-resolution data. Many such *in silico* methods allow us to obtain highly accurate models of a protein structure either *de novo* or via refinement of a physical model with experimental restraints. One crucial question is how to select a representative member or ensemble out of the vast number of computationally generated structures. Here, we introduce such a method. As a representative task, we add co-evolutionary contact pairs as distance restraints to a physical force field and want to select a good characterization of the resulting native-like ensemble. To generate large ensembles, we run replica-exchange molecular dynamics (REMD) on five mid-sized test proteins and over a wide temperature range. High temperatures allow overcoming energetic barriers while low temperatures perform local searches of native-like conformations. The integrated bias is based on co-evolutionary contact pairs derived from a deep residual neural network to guide the simulation toward native-like conformations. We shortly compare and discuss the achieved model precision of contact-guided REMD for mid-sized proteins. Finally, we discuss four robust ensemble-selection algorithms in great detail, which are capable to extract the representative structure models with a high certainty. To assess the performance of the selection algorithms, we exemplarily mimic a "blind scenario," i.e., where the target structure is unknown, and select a representative structural ensemble of native-like folds.

I. INTRODUCTION

Proteins are the key biomolecular players in cells and organize life at the nanoscale. They are involved in virtually all significant biomolecular tasks be it the regulation of genes, conformational transitions, energy regulation in the cell, signaling, enzymatic function, structural stability, or protein synthesis. To gain any detailed understanding of protein function, the knowledge of the respective 3D structure is typically required. The majority of proteins have a classical structure–function relation, where one native fold is representative for its biological function. One interesting exception are so-called intrinsically disordered proteins (IDPs). Such proteins are more flexible in nature and have a set of different structure ensembles, separated by low-energy barriers, instead of one stable and characteristic native fold. This heterogeneity as well as fast transitions between structure ensembles during interactions makes studies of IDPs and their functional interpretation much more difficult.[1,2]

Historically, experimental techniques, such as x-ray diffraction and nuclear magnetic resonance (NMR), have been used for high-resolution structure determination, but recently, in particular, cryogenic electron microscopy (cryo-EM) has achieved spectacular successes.[3,4] Other experimental methods, such as Förster resonance energy transfer (FRET) or small-angle x-ray scattering (SAXS), do not directly provide high-resolution structures. In contrast, the measured data contain ambiguous structure information at lower spatial resolution and must be interpreted carefully to deduce a target structure. In such cases, the integration of computational techniques can complement experimental data sufficiently to also provide high-resolution structures.

In particular, molecular dynamics (MD) has had tremendous success complementing or refining experimental data,[5–8] which by itself would be insufficient for full biomolecular structural characterization. Furthermore, MD has provided valuable insight into biomolecular folding and function by itself.[9–11] Although computational costly, it was shown that it is even possible to fold proteins *de novo* on the ms timescale on specialized supercomputers, such as Anton.[12,13]

In short, MD relies on time-integration of a physics-based force field, thus offering time-resolved insight into biomolecular dynamics akin to a virtual microscope with atomic resolution. A key challenge of MD lies in overcoming local barriers, as proteins can get trapped in local minima during the simulation, which can be tackled by advanced sampling methods.[14–16] One popular solution to this problem is given by replica-exchange molecular dynamics (REMD), where one runs copies of the simulated system at a range of temperatures in parallel and allows each replica to switch places with other replicas at adjacent temperatures.[17–20] Such exchanges between temperature levels lead to trajectory jumps and disrupt simulating the correct system dynamics at a fixed temperature but still maintain a thermodynamically correct description of the system.

In this study, we want to address the following two main objectives:

(1) We aim to investigate the achievable model precision of REMD applied on medium-sized proteins between 39 and 92 residues. As our simulation protocol, we use contact-guided replica-exchange simulations and add an energetic bias in the form of distance restraints by a sigmoid potential.[21] In general, distance restraints can be taken from any source in order to guide the REMD simulation toward specific conformations. It is possible to use, e.g., (sparse) NMR data[22] or contact maps from co-evolutionary analysis methods, such as direct coupling analysis (DCA).[23–25] In our case, we use contact maps derived from the neural network ResTriplet.[26] In addition, we want to test the limits of a physical force fields acting on such large systems during the contact-guided REMD protocol. Our previously performed study focused on small proteins up to a size of up to 35 residues.[21] Furthermore, it was shown that folding of similar large protein targets is possible with REMD when using a specialized, residue-specific force field within 2 μs-long trajectories.[27]

(2) REMD as an enhanced sampling method can generate large amounts of structural ensembles. With this in mind, our second objective is to find a robust solution for ensemble selection. As we require a stable metric to validate our protocols, we aim at reproducing the native state from the selected ensemble. We test four different ensemble-selection algorithm chains and compare their performance in great detail. Ensemble-selection methods are very task-orientated and are often applied during protein refinement, such as in the CASP11 competition.[28] In this case, Feig *et al.* performed initial MD simulations and obtained a refined structure by averaging an ensemble of previously scored and filtered trajectory snapshots.

In Sec. II, we briefly summarize all relevant methods required for our work. We introduce the five simulated protein systems and give an overview of their characteristics. Furthermore, we explain the generation of our initial starting structures for our contact-guided REMD simulations. Finally, we state the setup conditions, such as REMD temperature distribution and bias contact enrichment. We also give an overview of the used software and hardware for our production runs.

In Sec. III, we give a brief overview of the achieved accuracy of the REMD simulations and continue with a detailed explanation of our applied method to select representative ensembles. Afterward, we have an in-depth discussion and evaluation of the observed performance and state the pros and cons of our four investigated algorithm chain pipelines.

Finally, in Sec. IV, we conclude our findings and recap the most important points of our study. Here, we focus primarily on the discussed ensemble selection algorithms and give a brief overview of their selection criteria, performance ratings, and use-case related aspects.

II. METHODS

A. Simulated systems

We selected five mid-sized monomeric proteins of varying complexity for our study, which span a variety of different folds. The proteins have lengths between 39 and 92 residues. The first test protein is the lambda repressor [protein data bank (PDB) id: 1LMB[29]] with a folding time in the order of 49 μs. We simulated only the second dimer chain whose structure is composed of six α-helices in different orientations and has a length of 92 residues. Our second test protein is the albumin-binding domain (PDB id: 1PRB[30]) with a length of 53 residues and an extremely short folding time of ~3.9 μs. The structure consists of three α-helices that are orientated as a helical bundle. We were also interested in simulations using only small parts of an entire protein. Furthermore, we wanted to investigate systems that are purely β-sheets and selected, therefore, the WW domain of human Pip1 Fip mutant (PDB id: 2F21[31]). This domain has a reported folding time of ~21 μs. It is also the shortest test structure with a length of 39 residues. Our test system with a mixed structure of α-helices and β-sheets is given by the N-terminal of L9 protein (PDB id: 2HBA[32]). It has a length of 52 residues with a folding time of 29 μs. Finally, we investigated simulations with the BBL protein (PDB id: 2WXC[33]) with a length of 47 residues consisting of α-helices with a folding time of 49 μs. Table I gives a brief overview of the discussed protein systems, whereas Fig. 1 shows their tertiary structures. Reported protein folding times were obtained as average lifetime in the unfolded state observed in MD simulations.[11]

B. Generation of starting structures

To minimize the correlation between the initial starting structure and the achievable models with contact-guided REMD,[21] we decided to use a wide ensemble of starting conformations. For this purpose, we generated 5000 decoys for each studied protein using `PyRosetta`.[35] We started the decoy creation from the protein sequence and used a folding algorithm, which applies fragment insertion of 3mer and 9mer fragments to speed up the process. Fragment files were generated using the Robetta fragment server[36] while excluding homologues. Afterward, we calculated C_α distance matrices of the 5000 decoys and clustered them using KMEANS with

TABLE I. Overview of the simulated systems. The left side of the table contains structure-related information. The right side lists the number of used bias contacts during contact-guided REMD and the corresponding true-positive rate (TPR).

Name/description	PDB id	Folding time (μs)	Length	Bias contacts	Bias TPR (%)
Lambda repressor	1LMB	49	92	70	87
Albumin-binding domain	1PRB	3.9	53	40	82
WW domain of human Pin1 Fip mutant	2F21	21	39	30	96
N-terminal of L9 protein (NTL9)	2HBA	29	52	40	95
BBL	2WXC	49	47	35	91

72 cluster centers. Finally, we selected one decoy from each cluster with the lowest Rosetta score ("ref2015" score[37]) as unique starting structures for each individual replica. The selected structures were assigned according to their Rosetta score ranking during the REMD system setups. Appendix C (Tables 1 and 2) of the supplementary material give an overview of the initial decoy accuracy prior to the REMD simulations for all replicas.

C. Setup of contact-guided REMD simulations

All MD simulations of this study were performed with `GROMACS` 2020.[38,39] We used the `AMBER99SB-ILDN` force field[40] and the `TIP3P` explicit-solvent model.[41] The system setups were achieved with `pyrexMD`,[42] a self-developed Python package that provides a Jupyter-Notebooks based environment to design, run, and analyze MD projects from the beginning to end. It automates many system-specific and arduous tasks during the REMD setup and eliminates possible application errors, such as mismatching system sizes across replicas and incorrect mapping of bias contacts. During the setup process, each replica was equilibrated shortly for 200 ps in NVT and NPT simulations at their respective temperature. Prior to the production run, we added ~3/4 L (L: protein sequence length) bias contacts to guide the REMD simulations toward native-like conformations by applying a sigmoid potential.[21] Used bias contacts were predicted with ResTriplet.[26] Table I gives an overview of the simulated systems, including the number of used bias contacts and the corresponding true-positive rates. A detailed list with the used bias contacts can be taken from Appendix C (Table 3) of the supplementary material. Appendix D (Figs. S1–S5) of the supplementary material shows the contact maps of the proteins, visualizing both native contacts and included bias contacts.

All REMD simulations were performed on the `JUWELS`[43] computer cluster. We used standard compute nodes each consisting of 2× Intel Xeon Platinum 8168 CPU (2× 24 cores, 2.7 GHz) and 96 GB DDR4 memory with 2666 MHz. The .mdp file options of the REMD simulations can be taken from Appendix A of the of the supplementary material. Each REMD simulation comprised a simulated time of 500 ns with a time step of 2 fs and 72 replicas. The used REMD temperature distribution[21] (see Appendix B of the supplementary material) assigned temperatures from 280 K at replica 1 to ~445 K at replica 72.

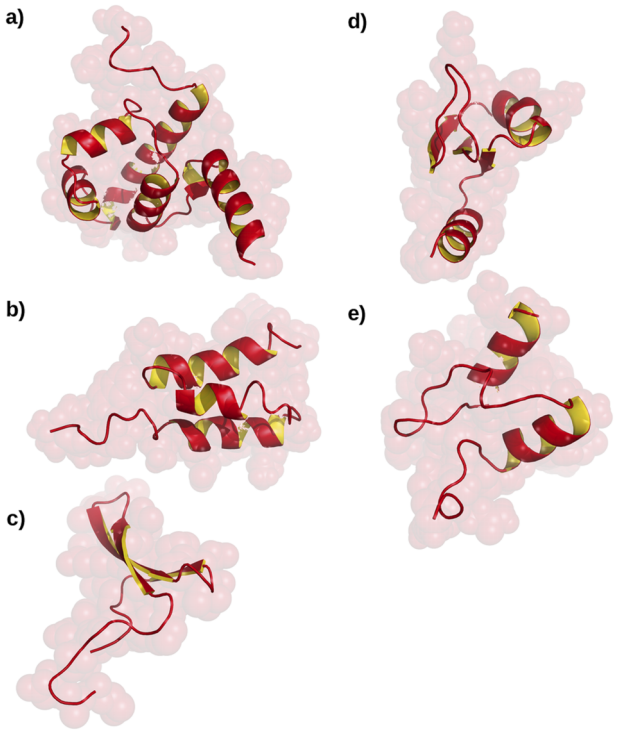

FIG. 1. Tertiary structures of simulated proteins. (a) Lambda repressor (1LMB). (b) Albumin-binding domain (1PRB). (c) WW domain of human Pin1 Fip mutant (2F21). (d) N-terminal of L9 protein (2HBA). (e) BBL (2WXC). Structures are visualized with PyMOL.[34]

III. RESULTS AND DISCUSSION

A. Achieved model accuracy

RMSD-based structure comparison strongly correlates with the largest displacement between two models where a few misplaced residues can result in disproportionally large RMSD values. For this reason, we focus on another metric called the global distance test (GDT[44–46]). After fitting the model to a target structure, pairwise C_α distances of the same residue are measured. These distances are used to calculate percentage values P_x based on different cutoff thresholds. Finally, a score to estimate the global structure similarity is calculated. The most common score is the total score (TS), which is defined as

TABLE II. Best achieved model accuracy during contact-guided REMD for the lowest-temperature replica. The protein's PDB id, the occurring secondary structure (ss) motifs, the system size (approx. atom count), the global distance test total score (GDT TS), and the backbone root-mean-square-deviation (RMSD) relative to the known protein structure are listed.

PDB id	ss motifs	System size	GDT TS	RMSD (Å)
1LMB	α	54×10^3	97	1.0
1PRB	α	47×10^3	78	1.9
2F21	β	36×10^3	89	1.8
2HBA	α, β	41×10^3	88	1.8
2WXC	α	34×10^3	81	1.9

$$GDT_{\mathrm{TS}} = \frac{1}{4}(P_1 + P_2 + P_4 + P_8) \in [0, 100], \quad (1)$$

where P_x denotes the percentage of residues with displacements below a distance cutoff of x Å.

The best results were achieved for the largest test system. The lowest-temperature replica of the 1LMB REMD simulation showed backbone RMSDs of ~1.0 Å after 250 ns simulated time, as shown in Fig. S6. Accordingly, GDT values higher than 90 were achieved, peaking with a value of 97 during the REMD trajectory. Table II summarizes the best achieved model accuracy, i.e., the backbone RMSD and global distance test total score (GDT TS) with regard to the known target structure, during the 500 ns long REMD simulations for all five test systems. The complete RMSD and GDT time evolution of the lowest-temperature replica can be looked up in Appendix D (Figs. S6–S10) of the supplementary material.

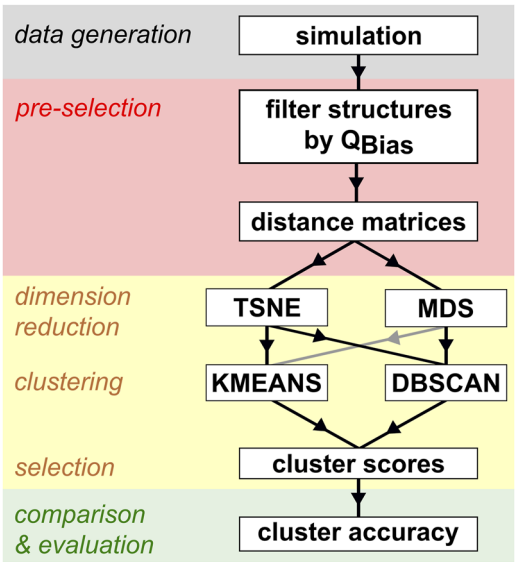

FIG. 2. Flowchart showing all steps of the investigated ensemble selection algorithm chains. Each stage is visualized in a different color: (I) data generation (gray), (II) pre-selection of structures (red), (III) ensemble selection algorithms (yellow), and (IV) comparison and evaluation (green) of the algorithm chains.

B. Ensemble selection algorithm chains

Our primary goal of this study was to find a method to reliably select a structure ensemble representing high GDT structures. We investigated four algorithm chains in great detail, which are able to achieve this goal with high certainty. Figure 2 shows an overview of all performed tasks in form of a flowchart. Starting with the entire REMD trajectory of the lowest-temperature replica, the first important step is to reduce the dataset by pre-selecting structures using a meaningful quantity. Our dataset was generated via application of contact-guided REMD. Therefore, we opted to use Q_{Bias}, the fraction of realized bias contacts in a structural model, as our metric to filter the generated structures. Q_{Bias} does not differentiate between true-positive or false-positive contacts; hence, $Q_{Bias} = 1$ might not always be structurally possible. As shown in Figs. 3(a) and S11(a)–S14(a), the fraction of realized bias contacts is positive correlated with the GDT scores. The only exception was observed for 2F21 [cf. Fig. S13(a)].

We decided to reduce the total frame count from 50 000 down to 2000 by pre-selecting structures with the highest Q_{Bias} values. These selections are visualized by blue dots in Figs. 3 and S12–S14. Next, we calculated the C_α distance matrices of the 2000 selected

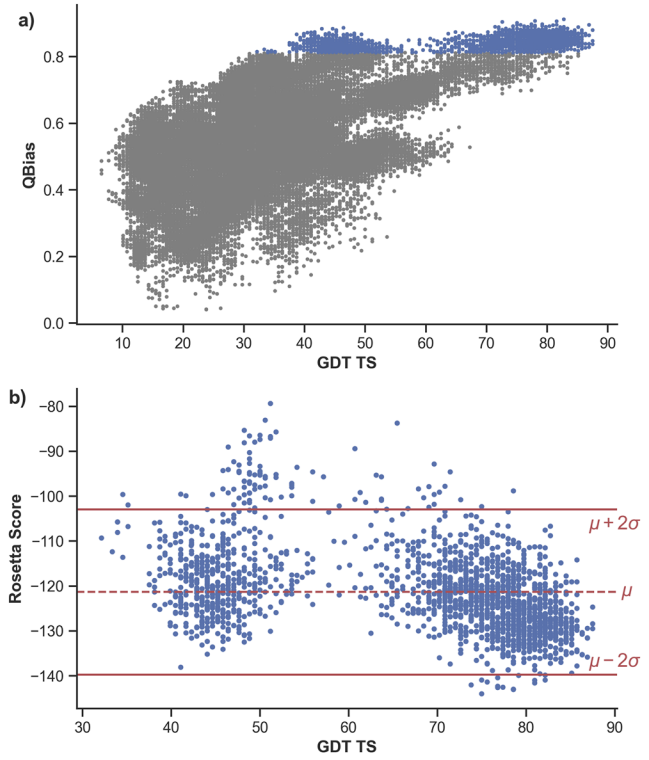

FIG. 3. (a) Scatter plot showing the relation between Q_{Bias} (fraction of realized bias contacts) and GDT TS (global distance test total score) for the 2HBA REMD simulation. Gray and blue colored dots represent the entire REMD trajectory composed of 50 000 structures. Blue dots highlight the 2000 structures with the highest Q_{Bias} values, which were pre-selected for the ensemble selection. (b) Scatter plot displaying the correlation between the Rosetta score and GDT TS of the 2000 pre-selected structures. The mean score μ (red dashed line) and $\mu \pm 2\sigma$ (red solid lines) are also shown, which were used as thresholds to filter outliers during the cluster score calculations.

structures, which we used as input for the dimension reduction of the four possible pipelines during the selection algorithm (yellow section in Fig. 2).

We investigated two different methods for dimension reduction: T-distributed stochastic neighbor embedding (TSNE[47]) and multidimensional scaling (MDS[48]). Both methods produce 2D representations (cf. Figs. 4 and 5) of the used distance matrices, where local adjacent points represent structures of high similarity. TSNE can visualize small structural differences better by creating many separated point clusters due to the t-distributed push–pull moves of samples during the algorithm. In other words, this algorithm aims to separate different structures from each other. MDS, on the other hand, visualizes structural differences better. That is because the distance between MDS points is always proportional to the difference of the corresponding distance matrices. The dimension reduction is of great importance for our selection method and fulfills two purposes. First, it enables the creation of readable 2D plots not only to compare structures but also to evaluate the performance of an entire ensemble selection algorithm chain. Second, the dimension reduction also eliminates a huge amount of randomness that could occur during the following clustering step. This makes the algorithm chains generally more robust and produces very similar results across independent executions.

We also investigated two different data clustering methods: KMEANS[49] and DBSCAN[50,51] (density-based spatial clustering of applications with noise). KMEANS is one of the most-basic clustering algorithms and has two important parameter specifications, namely, the number of clusters k and the number of runs with independant inizializations n_{init}. In each run, initial cluster centers are randomly chosen from the dataset and data points are assigned to the nearest cluster center. Next, cluster centers are shifted to the mean of all points belonging to a cluster and previous steps are repeated until convergence where no further changes occur. After n_{init} independent runs, KMEANS selects the best result based on the smallest sum of cluster variances. Due to the random selection of initial cluster centers, this clustering method has a certain degree of randomness attached. As previously mentioned, this random aspect can be reduced by previously performing a dimension reduction. The second clustering method, DBSCAN, is density-based and distinguishes between cluster points and noise. This clustering method also requires two important parameter specifications, i.e., ε and min_{pts}. ε describes the maximum distance between two samples, which are considered in the neighborhood. min_{pts} specifies how many data points within ε around sample X are required to consider X as a core sample and part of a cluster. If a core sample is identified, the cluster grows by including points within the ε neighborhood, which can also be core points or just simply reachable neighbors. Finally, all points that are not within the neighborhood of core points are considered noise. Due to this density-based procedure, each clustering result is identical for two individual runs using the same parameters. For comparison reasons, we decided to keep the total cluster count fixed for each clustering method, namely, 30 clusters for KMEANS and 21 clusters for DBSCAN.

After the clustering process is over, the next step is given by the cluster selection. In our case, we intend to select native-like structure ensembles. For this purpose, we calculated Rosetta scores[37] and mapped them to the individual TSNE or MDS representations. We

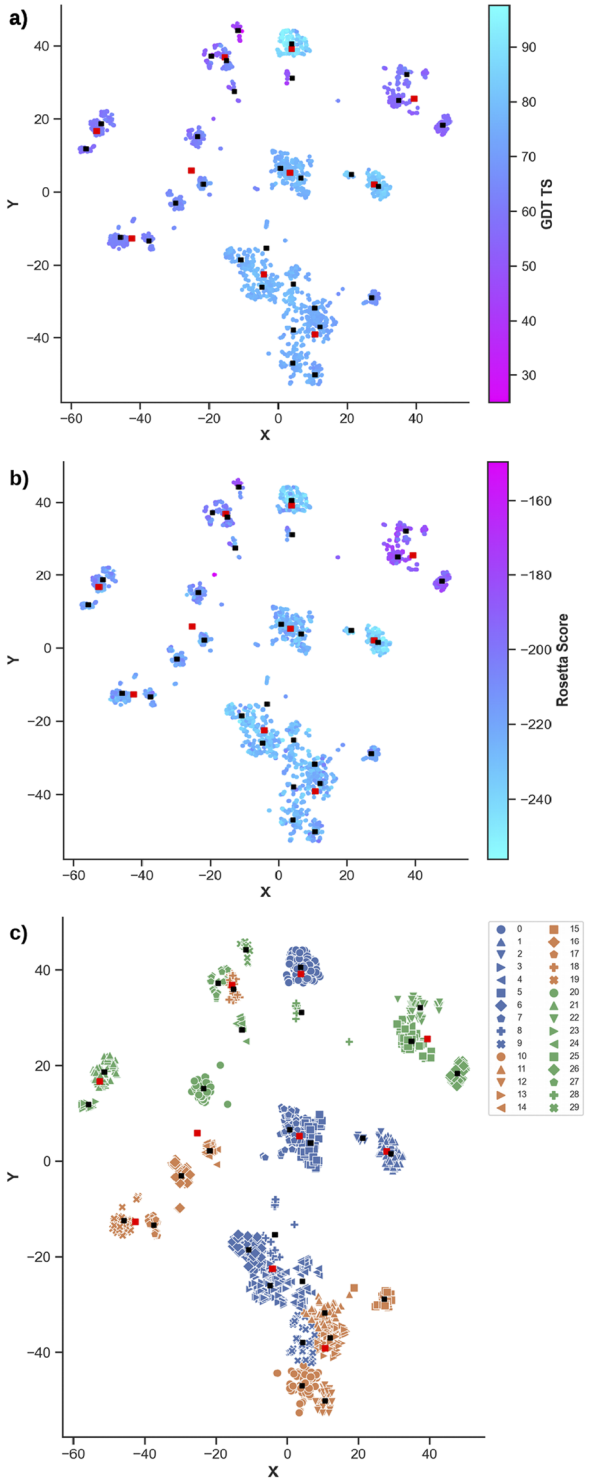

FIG. 4. TSNE representation of selected 1LMB structures with highlighted KMEANS cluster centers (k = 30: black squares, k = 10: red squares). (a) Relation with global distance test total scores (GDT TS). (b) Relation with Rosetta scores. (c) Relation with KMEANS cluster mapping. Cluster indices are ranked by average cluster accuracy (0: best, 29: worst).

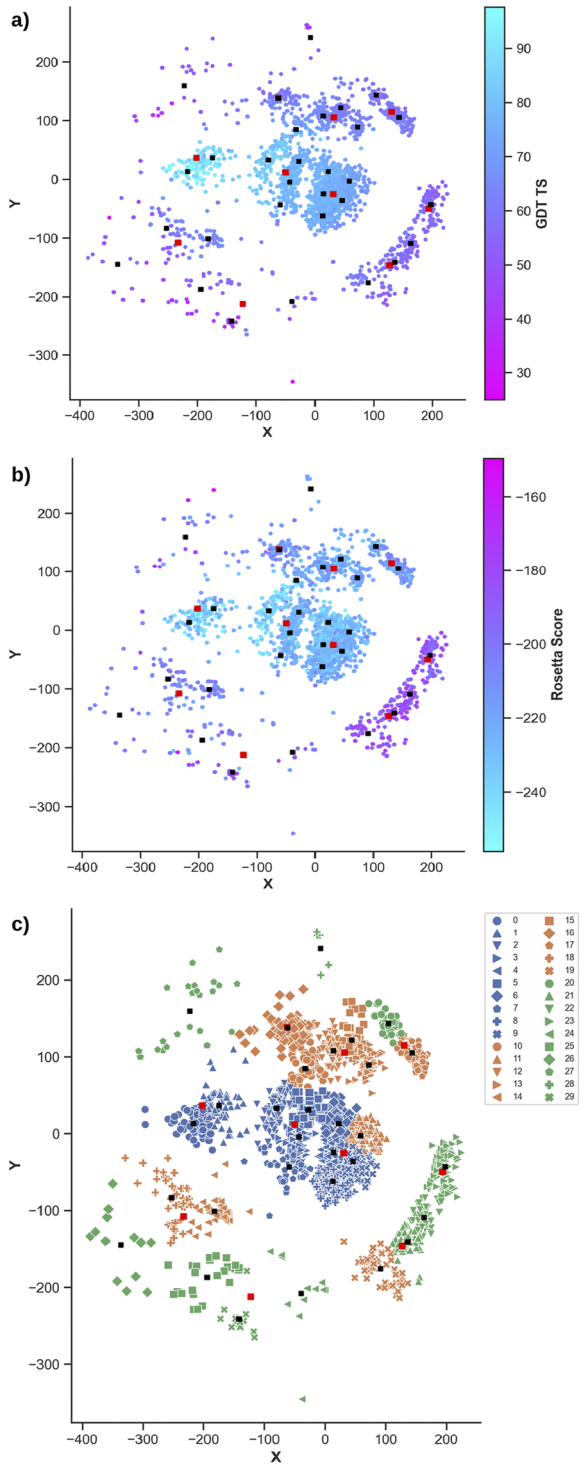

FIG. 5. MDS representation of selected 1LMB structures with highlighted KMEANS cluster centers (k = 30: black squares, k = 10: red squares). (a) Relation with global distance test total scores (GDT TS). (b) Relation with Rosetta scores. (c) Relation with KMEANS cluster mapping. Cluster indices are ranked by average cluster accuracy (0: best, 29: worst).

then identify four clusters with the lowest mean scores and pick them in increasing order. The general idea is that the 2D representation in addition to the Rosetta scores can be viewed as an energy landscape of protein structures. In order to select native-like structures, we, therefore, want to select the deepest valleys of the landscape and choose the structure ensembles by picking the low-scoring clusters. However, note that the Rosetta score by itself does not always discriminate native from native-like or even non-native folds sufficiently. As exemplified in Fig. 3(b), structures with low Rosetta scores are observed for structures with GDT scores primarily between 70 and 90 but also between 40 and 50. Nevertheless, the counts clearly indicate that low Rosetta scores have a higher probability to represent high GDT structures. Based on this observation, we can identify such structures out of the clusters only after taking their respective Rosetta score statistics into account, which favors our target ensembles with slightly lower mean scores. Finally, the comparison and evaluation regarding the performance of the four presented algorithm chains can be achieved by looking at the cluster accuracy statistics. Analogously to the cluster scoring with the mean Rosetta scores, we assigned each cluster their corresponding mean GDT value. We labeled all clusters based on their GDT ranking with indices 0–29 (0: best, 29: worst) for KMEANS or 0–20 (0: best, 19: worst, 20: noise) for DBSCAN. This allows an easier interpretation of the selected cluster during the discussion. Note that the calculation of GDT scores as well as the ranking of cluster labels is only possible because we used test systems where the native structure is already known. This information is only used for evaluation purposes and is not required to select the structure ensembles themselves. Furthermore, each time cluster labels are mentioned, they have been already sorted according to their accuracy. Tables III and IV give a performance overview by listing the selected clusters of each algorithm chain using KMEANS or DBSCAN, respectively. In addition, the tables state a performance rating, which indicates the importance of selected clusters. The rating is given by the weighted sum of selected clusters. However, meaningful weights are only assigned to clusters with labels 0–3 representing the highest GDT ensembles. Mathematically, this is provided by

$$rating = \sum_i w_i(cluster), \tag{2}$$

with the weights $w_0 = 4, w_1 = 3, w_2 = 2, w_3 = 1$, and $w_{i>3} = 0$.

TABLE III. Performance of algorithm chains with KMEANS clustering. Selection order of clusters is based on mean Rosetta scores, while cluster labels are ranked by accuracy (mean GDT scores, 0: best cluster). Rating is calculated according to Eq. (2) and indicates the importance of selected clusters by weighing only clusters 0–3.

	TSNE → KMEANS		MDS → KMEANS	
Protein	Selected clusters	Rating	Selected clusters	Rating
1LMB	1-0-6-2	9/10	2-0-1-6	9/10
1PRB	2-0-1-3	10/10	0-2-1-10	9/10
2F21	1-0-2-3	10/10	1-0-3-2	10/10
2HBA	3-2-0-1	10/10	0-1-7-8	7/10
2WXC	5-0-2-1	9/10	1-10-0-2	9/10

TABLE IV. Performance of algorithm chains with DBSCAN clustering. Selection order of clusters is based on mean Rosetta scores, while cluster labels are ranked by accuracy (mean GDT scores, 0: best cluster). Rating is calculated according to Eq. (2) and indicates the importance of selected clusters by weighing only clusters 0–3.

	TSNE → DBSCAN		MDS → DBSCAN	
Protein	Selected clusters	Rating	Selected clusters	Rating
1LMB	1-0-2-4	9/10	2-0-3-1	10/10
1PRB	1-0-2-14	9/10	2-0-1-12	9/10
2F21	0-1-6-2	9/10	0-1-8-2	9/10
2HBA	3-0-1-7	8/10	1-4-0-3	8/10
2WXC	1-0-3-9	8/10	2-3-0-1	10/10

In general, all compared algorithm chains yielded very good results regarding ensemble selections. Note that it was always possible to select the two highest GDT ensembles (labels 0 and 1) and in some cases even up to the four highest GDT ensembles. Algorithms using TSNE for dimension reduction were exceptional stable and produced ratings with 9/10 or higher for TNSE → KMEANS and 8/10 or higher for TNSE → DBSCAN. The direct rating comparison slightly favors the TNSE → KMEANS algorithm. In addition, this procedure is very straightforward and does not require any case-specific parameter tuning, as compared to the TSNE → DBSCAN pipeline. The selected ensembles resulting from algorithms using MDS are promising as well. MDS → KMEANS tends to produce ratings with ~9/10. However, the test case with 2HBA yielded a rating of only 7/10. MDS → DBSCAN pipelines produced ratings between 8/10 and 10/10. Additional information regarding selected clusters and the corresponding cluster accuracy can be looked up in Appendix C (Tables 4 and 5) of the supplementary material.

C. Discussion and evaluation

In this section, we want to focus on specific aspects, which should be considered when evaluating the performance of different ensemble-selection algorithms. One example is given by the mapping correlations between GDT scores, Rosetta scores, and selected cluster ensembles. These relations are illustrated in Fig. 4. Figure 4(a) displays color-coded GDT scores mapped to the TSNE representation of the 2000 pre-selected 1LMB structures. It contains all the important information regarding structure accuracy, which is only accessible if the target structure is already known. Therefore, it can be viewed as the "ground truth" in terms of the similarity of the individual structures compared to the native fold. Figure 4(b), on the other hand, shows the Rosetta score mapping instead. This information is always accessible and a high similarity between Figs. 4(a) and 4(b) means that the Rosetta score mapping is accurate and can be used to deduce good structure ensembles. Note that GDT scores and Rosetta scores have inverse scaling, i.e., good structures have high GDT scores and statistically low Rosetta scores. However, as previously mentioned, low Rosetta scores do not always indicate high GDT values. As exemplified in Figs. 3(b) and S12(b)–S14(b), low Rosetta scores can correspond to both high and low GDT structures. Finally, Fig. 3(c) shows the final cluster mapping. Depicted cluster labels are sorted by accuracy, where cluster 0 has the highest mean GDT scores and represents the best ensemble. By comparing Figs. 3(a)–3(c), it is possible to get an overview, where the four selected ensembles (cf. Tables III and IV) are located within the Rosetta score landscape and how good each selection algorithm performs.

When comparing similar figures for each individual test protein, we can clearly see that 1LMB, which is the largest test protein with a length of 92 residues, has a very accurate Rosetta score landscape. Figures 4(a) and 4(b) look almost identical. The same is obviously true for the MDS variant shown in Fig. 5(a) due to the fixed structure-to-score mappings, whereas the locations on the 2D representations differ. When comparing Fig. 4 with Fig. 5, we can observe the main difference between TSNE and MDS representations. TSNE plots tend to have more distinct sample groups, which result from the t-distributed push–pull projection. In both representations, highly similar structures are located very close to each other. However, in TSNE, the distance information is not conserved to the same degree as for MDS. This means that both GDT and Rosetta score landscapes are much easier to understand for MDS, as compared to TSNE. For example, Fig. 5(a) has exactly one distinct local minimum (left centered region), whereas Fig. 4(a) has multiple local minima spread around. This feature can be utilized to guess bad ensemble selections for algorithms using MDS. For example, if three out of four clusters are close to each other but one is far away during MDS → KMEANS, the one isolated cluster has a high probability to be a bad choice due to inaccurate Rosetta score mappings. Although TSNE does not have such a reliable way to tell false-positives, there exists a workaround. As shown in Fig. 4, KMEANS cluster centers for $k = 30$ and $k = 10$ are shown additionally as black and red squares, respectively. By clustering all samples with a high and low number of cluster centers, we can probe the associated cluster scores on different scales. In another step, we can check for the lowest scoring $k = 10$ cluster center and identify the three nearest $k = 30$ cluster centers. In our case, this information was always sufficient and could be used as some sort of confidence boost, when selecting the four final structure ensembles. Most of the time, at least two of the three nearest $k = 30$ clusters had cluster labels smaller than 3 and were part of the best selections based on their associated GDT accuracy.

Another important aspect of the evaluation is given by the robustness of each investigated algorithm chain. In general, using TSNE for dimension reduction makes the ensemble selection very robust. Individual executions of either KMEANS or DBSCAN on a TSNE representation tend to lead to very similar results. MDS, on the other hand, introduces a certain amount of stochasticity when combined with KMEANS clustering. The distance conservation typically leads to an accumulation of structures into few but dense groups without enough distance between each other to indicate a clear separation of structures. In some extreme cases, structures can even accumulate mainly into a single stack, e.g., as shown for 2F21 in Fig. S22 or 2WXC in Fig. S30. When applying KMEANS on such a 2D representation, independent runs can generate slightly varying results due to the initial random assignment of cluster centers. This results in minor movement of cluster borders and can lead to variations of the final ensemble selections. The variations can also be observed with different KMEANS initialization methods (e.g., random or Forgy[52]), since it primarily depends on the density of data points. Over many independent runs, we observed changes of ±1 for individual ratings according to Eq. (2). This means that the average rating, which is calculated across all test cases, stays approximately constant.

Comparing the last possible algorithm chain, one negative aspect stands out for the pipeline MDS → DBSCAN. The density-based clustering method is highly dependent on the 2D representation, which is applied beforehand. As previously mentioned, DBSCAN requires the specification of the two parameters ε and $\min_{pts}$. Note that both parameters have a correlation with distance, whereas ε is clearly dominating and, therefore, the most important parameter to choose. MDS representations, such as Fig. S22 for 2F21 or Fig. S30 for 2WXC, can produce unreliable results and will require case-specific parameter choices. However, because we forced a fixed cluster count of 21 for algorithms using DBSCAN, we were still able to achieve good results comparable to the other investigated alternatives. The biggest difference compared to other pipelines is that MDS → DBSCAN manages to select very small ensembles with extremely high structure accuracy. For example, the final selection for 1LMB contained an ensemble with only five structures and a maximum GDT of 94.64, or 2HBA resulted in an ensemble with eight structures and a maximum GDT of 85.71.

IV. CONCLUSION

One of our goals was to test the limitations of contact-guided REMD when applied on medium-sized proteins. Overall, we can deduce that this method is capable of achieving relatively good structure refinement for proteins up to the tested size of ~90 residues. We were able to observe GDT values above 80 in all of our 500 ns long simulations. For 2F21 and 2HBA, we achieved GDTs of nearly 90, whereas 1LMB reached an outstanding GDT score of 97 after only 250 ns. The observed performance and best-achieved model precision was more dependent on the true-positive rate of the bias contacts as compared to the secondary structure motifs or size of the protein. Proteins that mainly consist of α-helices are showing great results even after very short REMD simulations. Larger proteins and a high ratio of β-sheet motifs with respect to the total protein size generally require longer trajectories before showing good results. In such cases, longer REMD trajectories above 1 μs might be required to achieve better results due to additional replica turnarounds.

Our primary goal was to find a robust solution for ensemble selection. Here, we validated our selection method by trying to reproduce the native state mimicking a "blind scenario" with unknown target structure. In general, such a task is very challenging. There exist many different measurements or scoring formulas, which can be used to assess the quality of a protein structure. However, each on their own is not sufficient enough to guarantee outstanding structure selections. This was exemplarily shown by the correlations between observed Rosetta and GDT scores for our five test systems. Nevertheless, we showed that it is possible to reliably obtain our wanted protein ensemble by executing specific algorithm chains. We investigated a total of four different algorithm chains in great detail and objectively compared their performance.

Starting with the structures taken from the simulation, each chain requires a pre-selection of trajectory frames to reduce the frame count to a manageable amount. Here, structural data were generated using contact-guided REMD. We showed that, in this case, the fraction of realized bias contacts, Q_{Bias}, is a suitable quantity to reduce the frame count due to the primarily positive correlation with GDT scores. The next step of the algorithm chain performs a dimension reduction of the pre-selected structures and their C_α distance matrices. The main intentions are to improve the overall robustness of the algorithm chain by minimizing or even completely negating randomness from the following clustering step. In addition, it enables a readable 2D representation of structures, which can be extended via Rosetta and GDT score mappings to compare the algorithm performance. Here, we compared two variants of dimension-reduction, namely, t-distributed stochastic neighbor embedding (TSNE) and multidimensional scaling (MDS). The next step involved the clustering of the 2D representations into structure ensembles using either KMEANS or DBSCAN algorithms. The four possible pipelines were

(1) TSNE → KMEANS,
(2) MDS → KMEANS,
(3) TSNE → DBSCAN,
(4) MDS → DBSCAN.

TABLE V. Comparison overview of the four investigated algorithm chains. Clusters were selected by calculating mean Rosetta scores and picking the four lowest-scoring clusters. The total cluster count varies based on the used clustering method (KMEANS or DBSCAN). Average rating was calculated using Eq. (2) and normalized across all five test proteins.

Algorithm	Cluster selection	Average rating	Positive features	Negative features
TSNE → KMEANS	Top 4/30	9.6/10	Straightforward/no parameter tuning	Selection can include noise data
MDS → KMEANS	Top 4/30	8.8/10	Distance preservation allows to guess false-positives	Dense sample regions increase randomness of cluster borders
TSNE → DBSCAN	Top 4/21	8.6/10	Reduced noise	Slightly parameter dependent
MDS → DBSCAN	Top 4/21	9.2/10	Possible to identify small ensembles with extremely high structure accuracy	DBSCAN parameters correlate with distance, i.e., heavily depend on case-specific MDS representation

After calculating mean Rosetta scores for each cluster, we selected the four clusters with the lowest means as our final picks. Note that, while REMD leads to thermodynamically correct ensembles, our clustering does not maintain this property. Because we used proteins with known native structures, we were able to evaluate the performance of each algorithm chain by comparing the selected ensembles with the corresponding cluster accuracy. For this purpose, we introduced a numerical rating allowing us to easily compare the performance by weighing only the four ensembles corresponding to the most-refined structures. We also discussed the pros and cons of each algorithm chain in great detail and summarized them in Table V.

Overall, we showed that the presented algorithmic workflows performed very well in all test cases. Most notably, we were always able to obtain the two most native-like structure ensembles (i.e., clusters with label 0 and 1). However, it is still not possible to perfectly rank the selected ensembles based on accuracy if the target structure is truly unknown. The final ensemble selections primarily depend on the accuracy on the underlying energy function. As shown for four of our five test proteins, Rosetta scores are not accurate enough to reliably distinguish between low and high GDT conformations. Still, it is possible to apply small tricks to distinguish between particular good and bad picks. For instance, it is possible to use the distance preservation of MDS to eliminate bad picks, if, e.g., one of the selections is located far away from the others. Another example was given by performing two separate KMEANS clustering with $k = 10$ and $k = 30$. By changing the number of cluster centers k, comparing the individual distances of cluster centers and their energy rankings, one can deduce the real accuracy ranking in some cases.

Although we exemplarily aimed here for native-like ensemble selections, the key aspects of our methodology should be applicable to other ensemble-selection objectives as well. This would require an alteration of only two steps, namely, the pre-selection (here: filter structures by Q_{Bias}) and the scoring function during the final ensemble selection (here: mean Rosetta scores of clusters). By doing so, one should achieve similar results for other ensemble targets. For example, the application of an energy function that favors β-sheets could detect and select structure ensembles with high amounts of β-sheets.

SUPPLEMENTARY MATERIAL

See the supplementary material for Appendixes A–D for a sample .mdp file of the REMD simulations, the used REMD temperature distribution,[21] as well as various additional tables and figures related to this work.

ACKNOWLEDGMENTS

This work was supported by the Helmholtz Association Initiative and Networking Fund (Project No. ZT-I-0003). The authors gratefully acknowledge the Gauss Centre for Supercomputing e.V. (www.gauss-centre.eu) for funding this project by providing computing time through the John von Neumann Institute for Computing (NIC) on the GCS Supercomputer JUWELS at the Jülich Supercomputing Centre (JSC).[43] This work was also performed on the HoreKa supercomputer funded by the Ministry of Science, Research and the Arts Baden-Württemberg and by the Federal Ministry of Education and Research.[53]

AUTHOR DECLARATIONS

Conflict of Interest

The authors have no conflicts to disclose.

DATA AVAILABILITY

Raw data were generated at the GCS supercomputer JUWELS[43] and the supercomputer HoreKa[53] large scale facility. Derived data supporting the findings of this study are available from the corresponding author upon reasonable request.

REFERENCES

[1] M. Varadi and P. Tompa, "The protein ensemble database," in *Intrinsically Disordered Proteins Studied by NMR Spectroscopy* (Springer, 2015), pp. 335–349.
[2] T. Mittag and J. D. Forman-Kay, "Atomic-level characterization of disordered protein ensembles," Curr. Opin. Struct. Biol. **17**, 3–14 (2007).
[3] L. Gremer, D. Schölzel, C. Schenk, E. Reinartz, J. Labahn, R. B. G. Ravelli, M. Tusche, C. Lopez-Iglesias, W. Hoyer, H. Heise *et al.*, "Fibril structure of amyloid-β(1–42) by cryo–electron microscopy," Science **358**, 116–119 (2017).
[4] J. A. Geraets, K. R. Pothula, and G. F. Schröder, "Integrating cryo-EM and NMR data," Curr. Opin. Struct. Biol. **61**, 173–181 (2020).
[5] I. Reinartz, C. Sinner, D. Nettels, B. Stucki-Buchli, F. Stockmar, P. T. Panek, C. R. Jacob, G. U. Nienhaus, B. Schuler, and A. Schug, "Simulation of FRET dyes allows quantitative comparison against experimental data," J. Chem. Phys. **148**, 123321 (2018).
[6] M. Weiel, I. Reinartz, and A. Schug, "Rapid interpretation of small-angle X-ray scattering data," PLoS Comput. Biol. **15**, e1006900 (2019).
[7] M. R. Panman, E. Biasin, O. Berntsson, M. Hermann, S. Niebling, A. J. Hughes, J. Kübel, K. Atkovska, E. Gustavsson, A. Nimmrich *et al.*, "Observing the structural evolution in the photodissociation of diiodomethane with femtosecond solution X-ray scattering," Phys. Rev. Lett. **125**, 226001 (2020).
[8] M. Weiel, M. Götz, A. Klein, D. Coquelin, R. Floca, and A. Schug, "Dynamic particle swarm optimization of biomolecular simulation parameters with flexible objective functions," Nat. Mach. Intell. **3**, 727–734 (2021).
[9] M. Karplus and J. Kuriyan, "Molecular dynamics and protein function," Proc. Natl. Acad. Sci. U. S. A. **102**, 6679–6685 (2005).
[10] D. Paschek, S. Hempel, and A. E. García, "Computing the stability diagram of the Trp-cage miniprotein," Proc. Natl. Acad. Sci. U. S. A. **105**, 17754–17759 (2008).
[11] K. Lindorff-Larsen, S. Piana, R. O. Dror, and D. E. Shaw, "How fast-folding proteins fold," Science **334**, 517–520 (2011).
[12] D. E. Shaw, M. M. Deneroff, R. O. Dror, J. S. Kuskin, R. H. Larson, J. K. Salmon, C. Young, B. Batson, K. J. Bowers, J. C. Chao *et al.*, "Anton, a special-purpose machine for molecular dynamics simulation," Commun. ACM **51**, 91–97 (2008).
[13] D. E. Shaw, R. O. Dror, J. K. Salmon, J. Grossman, K. M. Mackenzie, J. A. Bank, C. Young, M. M. Deneroff, B. Batson, K. J. Bowers *et al.*, "Millisecond-scale molecular dynamics simulations on anton," in *Proceedings of the Conference on High Performance Computing Networking, Storage and Analysis* (ACM New York, NY, 2009), pp. 1–11.
[14] S. A. Adcock and J. A. McCammon, "Molecular dynamics: Survey of methods for simulating the activity of proteins," Chem. Rev. **106**, 1589–1615 (2006).
[15] G. A. Tribello, M. Bonomi, D. Branduardi, C. Camilloni, and G. Bussi, "PLUMED 2: New feathers for an old bird," Comput. Phys. Commun. **185**, 604–613 (2014).

[16]E. K. Peter, D. J. Manstein, J.-E. Shea, and A. Schug, "CORE-MD II: A fast, adaptive, and accurate enhanced sampling method," J. Chem. Phys. **155**, 104114 (2021).
[17]Y. Sugita and Y. Okamoto, "Replica-exchange molecular dynamics method for protein folding," Chem. Phys. Lett. **314**, 141–151 (1999).
[18]T. Okabe, M. Kawata, Y. Okamoto, and M. Mikami, "Replica-exchange Monte Carlo method for the isobaric–isothermal ensemble," Chem. Phys. Lett. **335**, 435–439 (2001).
[19]K. Y. Sanbonmatsu and A. E. García, "Structure of Met-enkephalin in explicit aqueous solution using replica exchange molecular dynamics," Proteins: Struct., Funct., Bioinf. **46**, 225–234 (2002).
[20]A. Schug, T. Herges, and W. Wenzel, "All-atom folding of the three-helix HIV accessory protein with an adaptive parallel tempering method," Proteins: Struct., Funct., Bioinf. **57**, 792–798 (2004).
[21]A. Voronin, M. Weiel, and A. Schug, "Including residual contact information into replica-exchange MD simulations significantly enriches native-like conformations," PLoS One **15**, e0242072 (2020).
[22]O. F. Lange, N.-A. Lakomek, C. Farès, G. F. Schröder, K. F. A. Walter, S. Becker, J. Meiler, H. Grubmüller, C. Griesinger, and B. L. De Groot, "Recognition dynamics up to microseconds revealed from an RDC-derived ubiquitin ensemble in solution," Science **320**, 1471–1475 (2008).
[23]A. Schug, M. Weigt, J. N. Onuchic, T. Hwa, and H. Szurmant, "High-resolution protein complexes from integrating genomic information with molecular simulation," Proc. Natl. Acad. Sci. U. S. A. **106**, 22124–22129 (2009).
[24]M. Weigt, R. A. White, H. Szurmant, J. A. Hoch, and T. Hwa, "Identification of direct residue contacts in protein–protein interaction by message passing," Proc. Natl. Acad. Sci. U. S. A. **106**, 67–72 (2009).
[25]F. Morcos, A. Pagnani, B. Lunt, A. Bertolino, D. S. Marks, C. Sander, R. Zecchina, J. N. Onuchic, T. Hwa, and M. Weigt, "Direct-coupling analysis of residue coevolution captures native contacts across many protein families," Proc. Natl. Acad. Sci. U. S. A. **108**, E1293–E1301 (2011).
[26]Y. Li, C. Zhang, E. W. Bell, D. J. Yu, and Y. Zhang, "Ensembling multiple raw coevolutionary features with deep residual neural networks for contact-map prediction in CASP13," Proteins: Struct., Funct., Bioinf. **87**, 1082–1091 (2019).
[27]F. Jiang and Y.-D. Wu, "Folding of fourteen small proteins with a residue-specific force field and replica-exchange molecular dynamics," J. Am. Chem. Soc. **136**, 9536–9539 (2014).
[28]M. Feig and V. Mirjalili, "Protein structure refinement via molecular-dynamics simulations: What works and what does not?," Proteins: Struct., Funct., Bioinf. **84**, 282–292 (2016).
[29]L. J. Beamer and C. O. Pabo, "Refined 1.8 Å crystal structure of the λ repressor-operator complex," J. Mol. Biol. **227**, 177–196 (1992).
[30]M. U. Johansson, M. de Château, M. Wikström, S. Forsén, T. Drakenberg, and L. Björck, "Solution structure of the albumin-binding GA module: A versatile bacterial protein domain," J. Mol. Biol. **266**, 859 (1997).
[31]M. Jäger, Y. Zhang, J. Bieschke, H. Nguyen, M. Dendle, M. E. Bowman, J. P. Noel, M. Gruebele, and J. W. Kelly, "Structure–function–folding relationship in a WW domain," Proc. Natl. Acad. Sci. U. S. A. **103**, 10648–10653 (2006).
[32]J.-H. Cho, W. Meng, S. Sato, E. Y. Kim, H. Schindelin, and D. P. Raleigh, "Energetically significant networks of coupled interactions within an unfolded protein," Proc. Natl. Acad. Sci. U. S. A. **111**, 12079–12084 (2014).
[33]H. Neuweiler, T. D. Sharpe, T. J. Rutherford, C. M. Johnson, M. D. Allen, N. Ferguson, and A. R. Fersht, "The folding mechanism of BBL: Plasticity of transition-state structure observed within an ultrafast folding protein family," J. Mol. Biol. **390**, 1060–1073 (2009).
[34]W. L. DeLano *et al.*, "PyMOL: An open-source molecular graphics tool," CCP4 Newsl. Protein Crystallogr. **40**, 82–92 (2002).
[35]S. Chaudhury, S. Lyskov, and J. J. Gray, "PyRosetta: A script-based interface for implementing molecular modeling algorithms using Rosetta," Bioinformatics **26**, 689–691 (2010).
[36]D. E. Kim, D. Chivian, and D. Baker, "Protein structure prediction and analysis using the Robetta server," Nucleic Acids Res. **32**, W526–W531 (2004).
[37]R. F. Alford, A. Leaver-Fay, J. R. Jeliazkov, M. J. O'Meara, F. P. DiMaio, H. Park, M. V. Shapovalov, P. D. Renfrew, V. K. Mulligan, K. Kappel *et al.*, "The Rosetta all-atom energy function for macromolecular modeling and design," J. Chem. Theory Comput. **13**, 3031–3048 (2017).
[38]D. Van Der Spoel, E. Lindahl, B. Hess, G. Groenhof, A. E. Mark, and H. J. C. Berendsen, "GROMACS: Fast, flexible, and free," J. Comput. Chem. **26**, 1701–1718 (2005).
[39]E. Lindahl, M. J. Abraham, B. Hess, and D. van der Spoel, GROMACS 2020 manual, 2020.
[40]K. Lindorff-Larsen, S. Piana, K. Palmo, P. Maragakis, J. L. Klepeis, R. O. Dror, and D. E. Shaw, "Improved side-chain torsion potentials for the Amber ff99SB protein force field," Proteins: Struct., Funct., Bioinf. **78**, 1950–1958 (2010).
[41]W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein, "Comparison of simple potential functions for simulating liquid water," J. Chem. Phys. **79**, 926–935 (1983).
[42]A. Voronin and A. Schug, "pyrexMD: Workflow-orientated Python package for replica exchange molecular dynamics," J. Open Source Software **6**, 3325 (2021).
[43]D. Krause, "JUWELS: Modular tier-0/1 supercomputer at the Jülich Supercomputing Centre," J. Large-Scale Res. Facil. **5**, A135 (2019).
[44]A. Zemla, "LGA: A method for finding 3D similarities in protein structures," Nucleic Acids Res. **31**, 3370–3374 (2003).
[45]A. Zemla, Č. Venclovas, J. Moult, and K. Fidelis, "Processing and analysis of CASP3 protein structure predictions," Proteins: Struct., Funct., Bioinf. **37**, 22–29 (1999).
[46]V. Modi and R. L. Dunbrack, Jr., "Assessment of refinement of template-based models in CASP11," Proteins: Struct., Funct., Bioinf. **84**, 260–281 (2016).
[47]L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," J. Mach. Learn. Res. **9**, 2579 (2008).
[48]I. Borg and P. J. Groenen, *Modern Multidimensional Scaling: Theory and Applications* (Springer Science & Business Media, 2005).
[49]D. Steinley, "K-means clustering: A half-century synthesis," Br. J. Math. Stat. Psychol. **59**, 1–34 (2006).
[50]M. Ester, H.-P. Kriegel, J. Sander, X. Xu *et al.*, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *KDD'96 Proceedings of the Second International Conference on Knowledge Discovery and Data Mining* (AAAI press, 1996), pp. 226–231.
[51]E. Schubert, J. Sander, M. Ester, H. P. Kriegel, and X. Xu, "DBSCAN revisited, revisited: Why and how you should (still) use DBSCAN," ACM Trans. Database Syst. **42**, 1–21 (2017).
[52]J. M. Peña, J. A. Lozano, and P. Larrañaga, "An empirical comparison of four initialization methods for the *K*-means algorithm," Pattern Recognit. Lett. **20**, 1027–1040 (1999).
[53]S. Raffeiner, Hochleistungsrechner Karlsruhe (HoreKa), URL: https://publikationen.bibliothek.kit.edu/1000136028, 2021.