Rapid Simulation of Unprocessed DEER Decay Data for Protein Fold Prediction

Diego del Alamo,[1,2] Maxx H. Tessmer,[3] Richard A. Stein,[2] Jimmy B. Feix,[4] Hassane S. Mchaourab,[1,2] and Jens Meiler[1,5,*]

[1]Department of Chemistry and Center for Structural Biology; [2]Department of Molecular Physiology and Biophysics, Vanderbilt University, Nashville, Tennessee; [3]Department of Microbiology and Immunology; [4]Department of Biophysics, Medical College of Wisconsin, Milwaukee, Wisconsin; and [5]Institut for Drug Discovery, Leipzig University, Leipzig, Germany

ABSTRACT   Despite advances in sampling and scoring strategies, Monte Carlo modeling methods still struggle to accurately predict de novo the structures of large proteins, membrane proteins, or proteins of complex topologies. Previous approaches have addressed these shortcomings by leveraging sparse distance data gathered using site-directed spin labeling and electron paramagnetic resonance spectroscopy to improve protein structure prediction and refinement outcomes. However, existing computational implementations entail compromises between coarse-grained models of the spin label that lower the resolution and explicit models that lead to resource-intense simulations. These methods are further limited by their reliance on distance distributions, which are calculated from a primary refocused echo decay signal and contain uncertainties that may require manual refinement. Here, we addressed these challenges by developing RosettaDEER, a scoring method within the Rosetta software suite capable of simulating double electron-electron resonance spectroscopy decay traces and distance distributions between spin labels fast enough to fold proteins de novo. We demonstrate that the accuracy of resulting distance distributions match or exceed those generated by more computationally intensive methods. Moreover, decay traces generated from these distributions recapitulate intermolecular background coupling parameters even when the time window of data collection is truncated. As a result, RosettaDEER can discriminate between poorly folded and native-like models by using decay traces that cannot be accurately converted into distance distributions using regularized fitting approaches. Finally, using two challenging test cases, we demonstrate that RosettaDEER leverages these experimental data for protein fold prediction more effectively than previous methods. These benchmarking results confirm that RosettaDEER can effectively leverage sparse experimental data for a wide array of modeling applications built into the Rosetta software suite.

SIGNIFICANCE   Computational methods struggle to generate protein structural models using data obtained with double electron-electron resonance spectroscopy (DEER), which measures nanometer-scale distances between probes bound to the protein backbone. In addition to these data being sparse, their precision is dependent on the quality of the primary spectroscopic readout from which they are derived. We developed RosettaDEER to directly interrogate the raw spectroscopic data for protein structural modeling and found that even low-quality DEER data enable the identification of native-like protein structural models. Moreover, by predicting the folds of two example proteins de novo, we demonstrate that this approach leverages experimental data more effectively than existing methods. These results highlight the utility of DEER decay data for protein structural modeling.

INTRODUCTION

Structural biology increasingly relies on integrated methods to model the structure and dynamics of proteins and protein assemblies (1,2). Multiple complementary experimental methodologies can describe the structure and dynamics of proteins that elude determination from a single technique, such as integral membrane proteins, conformationally flexible proteins, and those that fall outside the size limitations of solution-state nuclear magnetic resonance and cryogenic electron microscopy. By integrating experimental data from multiple approaches, computational modeling can build accurate models in regions with sparse experimental data. One promising source of high-resolution experimental data for integrated structural biology combines site-directed spin labeling and electron paramagnetic resonance spectroscopy

(SDSL-EPR) (3,4). Previous studies have employed SDSL-EPR and computation in tandem to predict protein structures de novo (5–12), model conformational changes (13–16), and dock rigid bodies (17–19).

Existing modeling methods largely focus on data gathered using four-pulse double electron-electron resonance spectroscopy (20) (DEER, also called PELDOR), which can report on distances of up to 60–80 Å between stable unpaired electrons conjugated to the protein backbone by SDSL (21,22). However, incorporation of these distances as interatomic restraints for modeling purposes is confounded by the conformational freedom of these paramagnetic probes. The central challenge is to convert interspin distance information into structural restraints that report on the protein backbone (23–25). Additionally, the need to incorporate two spin labels into the protein sequence per restraint results in sparse coverage of the experimental data that can introduce ambiguities into computational modeling (8). As a result, only a few experimental restraints are generally available to describe the protein fold.

These sparse data sets have nonetheless been leveraged for protein structure prediction and refinement by a range of computational modeling approaches that represent the spin labels either implicitly or explicitly. Implicit models, such as the motion-on-a-cone (CONE) model (5), use knowledge-based potentials to translate interspin distance values into backbone restraints, typically between C_β atoms (26). Introducing these restraints led to measurable improvements in de novo structure prediction benchmarks by programs employing Monte Carlo sampling strategies (5,7–11), gradient minimization (6,12), and molecular dynamics (27). However, because these potentials fail to account for the environment or the relative orientations of the spin labels, they tend to be ambiguous (26). Explicit methods, by contrast, model spin labels as individual side chains (16,28–31), ensembles of side chains (17,32–34), or ensembles of dummy atoms (13,14,35). The added detail improves accuracy of modeling but makes implementations too computationally intensive for de novo protein structure prediction and limits the utility of these methods to modeling small-scale conformational changes (13,14,16).

Despite their diversity, these methods largely share a common limitation in their reliance on distance distributions rather than the primary spectroscopic readout. Other computational methodologies directly incorporate primary experimental data, such as two-dimensional NMR spectra (36) and cryogenic electron microscopy electron density maps (37) to fold and refine proteins. The feasibility of using DEER dipolar coupling decay traces as modeling restraints has only recently been explored (16). Whereas processing spectroscopic decay traces into distance distributions risks introducing ambiguities and artifacts (38–43), simulating a decay trace from a distance distribution is well described and mathematically straightforward (16,22,40,44).

Here, we introduce RosettaDEER, a method in the macromolecular modeling suite Rosetta that is capable of rapidly simulating distance distributions and DEER decay traces

between spin labels, as well as evaluating a model's agreement with experimental data. RosettaDEER's computational efficiency enables prediction of protein structures de novo with greater accuracy than the default energy function or the CONE model. Owing to Rosetta's Monte Carlo sampling strategy (45), the experimental data can be used directly without analysis or background correction. Thus, as with other forward modeling approaches (46), the quality of the primary spectroscopic data can be significantly poorer than what would ordinarily be required for rigorous transformation into distance distributions using common fitting strategies. This method reinforces the utility of DEER in conjunction with computational modeling to accurately model protein structures.

MATERIALS AND METHODS

Assembly of diverse experimental data sets

RosettaDEER was implemented in the Rosetta software suite (45), trained on distance data gathered in T4 lysozyme obtained from the laboratory of Hassane S. Mchaourab, and tested and cross-validated using both raw spectroscopic and analyzed distance data gathered in five laboratories (Table S1). Data for the ExoU C-terminus (11), Bax (47), and Mhp1 (14) were obtained from and analyzed by the laboratories of Dr. Jimmy Feix, Dr. Enrica Bordignon, and Dr. Hassane S. Mchaourab, respectively. New ExoU double-cysteine mutants were purified, spin labeled, measured, and analyzed as previously described (Fig. S1; (11)). Raw data for CDB3 (48) and bovine rhodopsin (49) were obtained from the laboratories of Dr. Albert Beth and Dr. Wayne Hubbell, respectively, and were analyzed using DEERAnalysis2016 (38); the last 200 and 500 ns were removed from experimental decay traces shorter and longer than 1.5 μs, respectively.

Generation of DEER distance distributions

The accuracy of various methods that simulate distance distributions between spin labels were compared using Bax (Protein Data Bank, PDB: 1F16, NMR state 8), ExoU (PDB: 3TU3), CDB3 (PDB: 1HYN chains R/S), rhodopsin (PDB: 1GZM chain A), and Mhp1 (PDB: 2JLN). The methods compared were MMM (32), MDDS (35), MtsslWizard (33), Pronox (34), and TagDock (18) (Fig. 1, A and B; Table S2). MMM2017 was run locally on both cryogenic mode (175 K) and ambient mode (298 K) with default settings. MDDS was run using the CHARMM-GUI web server (50) with default settings. MtsslWizard was run locally from Pymol 1.7.2.1 using tight fitting unless no rotamers could be placed, in which case loose fitting was used (because Mhp1 residue 324 could not be labeled using loose fitting, distances between it were omitted). Pronox was run from the University of Southern California web server using a bias of 0.9 and a van der Waals radius scaling factor of 0.75, the latter of which was reduced to 0.4 if rotamers could not be placed. TagDock was run locally with SCWRL4 (51) and a bump radius of 0.85. Measurements using the CONE model (5,7) were determined by adding 1.79 Å to the C_β-C_β distance.

RosettaDEER method description

The Rosetta rotamer library for the paramagnetic probe methanethiosulfonate spin label (MTSSL) (30) served as the basis for the coarse-grained rotameric ensemble used in this study. For each of 54 possible rotameric configurations, the unpaired electron was assumed to occupy the nitroxide bond midpoint; it was from these coordinates that distances would be measured. These coordinates were consolidated into a common frame defined by the C_α atom at the origin, the backbone carbonyl carbon along the z axis, and the backbone nitrogen in the y-z plane (Fig. 1 C). The
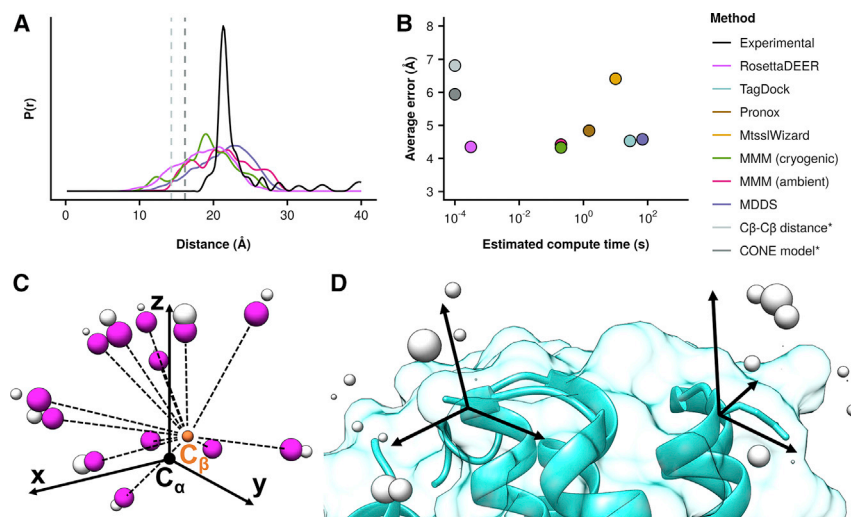
FIGURE 1 Simulations of distance distributions between nitroxide probes using RosettaDEER. (*A*) An example of an experimentally observed distance distribution in apo Mhp1 51/278, shown in black. Distance distributions were simulated using RosettaDEER, MMM, and MDDS from the occluded Mhp1 structure (PDB: 2JLN). The average distance between C_β atoms and the average distance calculated using the CONE model shown in light gray and dark gray, respectively. (*B*) The estimated average time required to simulate distance distributions (*the lower limit of quantitation exceeded the C_β-C_β distance compute time). (*C*) Coarse-grained rotameric ensemble representation of the methanethiosulfonate spin label. Centers of mass, shown in purple, are used for clash evaluation, whereas electron coordinates, shown in gray, serve as measurement coordinates. (*D*) Distance distributions between residues are simulated by superimposing coordinates, evaluating clashes and measuring all resulting pairwise distances.

remainder of each rotamer was represented by a single pseudoatom with a radius of 2.4 Å that was placed at 87.5% of the distance between each nitroxide bond midpoint coordinate and an idealized C_β coordinate; if this pseudoatom clashed with the protein model, its corresponding electron coordinate was not used for distance measurements. The placement of this pseudoatom coincides with the center of mass of the nitroxide ring of MTSSL (Fig. S2; Table S3). For cases in which the steric environment prevented the placement of any rotamers, the van der Waals radii of the pseudoatoms were gradually lowered until at least one rotamer could be accommodated. Distance distributions between two residues reflect all pairwise distance measurements between their respective coordinates after evaluating clashes; we smoothed each of these distance values into Gaussian distributions with a 0.5 Å standard deviation. The resulting distance distributions were then binned to 0.5 Å.

The resulting coordinate frame, which consisted of 54 unweighted coordinates and their positions with respect to protein backbone, did not account for the dynamics of the spin label (e.g., the configurations and positions it preferentially occupies) and was highly redundant, with coordinates often being placed <1 Å apart (Fig. S3). We addressed both issues using a scheme outlined in Fig. S3. Neighboring coordinates were merged using *k*-means clustering to generate a series of coordinate sets ranging from three positions to 53 total positions. The weights of these resulting positions were then optimized using 49 previously published experimental distance distributions between 37 residues gathered in T4 lysozyme (35). During each of half a million iterations, a Monte Carlo Metropolis algorithm randomly modified the weight of a coordinate and either accepted or rejected the change based on the improved agreement with the experimental T4 lysozyme distance data. This algorithm was carried out on each set of clustered coordinates 1000 times. The resulting set of weights with the best agreement consisted of 17 coordinates, four of which were fit to be zero. This set was introduced as the default set of coordinates for RosettaDEER and was used for all subsequent experiments described here.

Simulation of DEER dipolar coupling decay traces and comparison to experimental values

Because the simulation of DEER decay traces has been extensively described (16,22,38,40), here we limit our discussion to their generation from distance distributions for the purpose of evaluating protein structural models. The traces simulated by RosettaDEER (V_{sim}) reflect coupling between spin labels attached to the same macromolecule (V_{intra}), as well as an intermolecular "background" component reflecting coupling between spin labels across different macromolecules:

$$V_{sim}(t_i, \lambda, k, \boldsymbol{r}, \boldsymbol{w}) = e^{kt}\left(1 - \lambda\left(1 - V_{intra}(t_i, \boldsymbol{r}, \boldsymbol{w})\right)\right) \quad (1)$$

This background is assumed to be homogeneous across three dimensions and is modeled using a slope k and a modulation depth λ. The simulated distribution consists of a vector of distances $\boldsymbol{r}$ (in nanometers) and their corresponding amplitudes $\boldsymbol{w}$. Simulated traces obtained this way are converted into scores (S_{DEER}) by comparing them to the corresponding experimental spectra (V_{exp}) using the following cost function:

$$S_{DEER} = \frac{1}{n}\sum_{i=1}^{n}\left(V_{exp}(t_i) - V_{sim}(t_i, \lambda, k, \boldsymbol{r}, \boldsymbol{w})\right)^2, \quad (2)$$

where n is the number of time points in the data.

To convert a distance distribution into a spectroscopic signal that can be compared to experimental data, RosettaDEER first simulates V_{intra} for each 0.5 Å bin j between 15 and 100 Å:

$$V_{intra}(t_i, \boldsymbol{r}, \boldsymbol{w}) = \left(\sum_{j=1}^{m} \boldsymbol{w}_j \int_0^{\frac{\pi}{2}} \sin\theta\cos\left(\frac{(1 - 3\cos^2\theta)\mu_0\mu_B^2 g_A g_B t_i}{4\pi\hbar r_j^3}\right)d\theta\right)\left(\sum_{j=1}^{m}\boldsymbol{w}_j\right)^{-1}, \quad (3)$$

where t is the time point of a trace in μs, μ_B is the Bohr magneton, μ_0 is the vacuum permeability constant, g_X is the g-factor of electron X, θ is the angle between the interelectron vector and the bulk magnetic field, and m is the number of distance bins.

Background parameters k and λ are then determined and optimized in two stages. Initial values for both parameters were first determined by incrementing λ with step size 0.01 and log-transforming Eq. 1 to determine k using linear regression:

$$\widehat{k} = \left(\sum_{i=1}^{n} t_i \ln\left(\frac{V_{\exp}(t_i)}{1 - \lambda(1 - V_{intra}(t_i, \boldsymbol{r}, \boldsymbol{w}))} \right) \right) \left(\sum_{i=1}^{n} t_i^2 \right)^{-1} \tag{4}$$

Subsequent attempts to fit simulated intramolecular decay traces were achieved using gradient minimization to solve for λ and linear regression to solve for k. Convergence was reached when $|\Delta\lambda| < 0.0025$. The iterative strategy used to obtain the initial guess was repeated for cases in which λ exceeded reasonable values, the lower and upper bounds of which are defined by default as 0.02 and 0.50. This range corresponds to modulation depth values that would ordinarily be obtained from Q-band DEER on well-labeled double-cysteine mutants without using an arbitrary waveform generator. Deviations from experimentally observed values for these two parameters were found to frequently occur during the initial stages of extended chain de novo folding, in which simulated distance distributions deviated drastically from experimental values and led to erroneous background parameter results.

Rosetta model generation and evaluation

Rosetta models were generated with two approaches to not only sample a large conformational space but also ensure native-like models at a high density. The native-like models were generated with RosettaCM (52) using either full-length or truncated native models as inputs. Coverage of a large conformational space was accomplished by de novo protein folding without experimental restraints. Bax, ExoU, and CDB3 were scored using the ref2015 energy function (53), and rhodopsin and Mhp1 were scored using RosettaMembrane (54). The transmembrane regions for rhodopsin and Mhp1 were predicted using OCTOPUS (55). These models were evaluated using $RMSD_{100SSE}$, which measures the size-normalized root mean-square deviation (RMSD) over residues in secondary structures (56). Enrichment of these models was evaluated as $log\left((TP_{P,Score} \times N_{total})/(P \times N_{total}) \right)$, where N_{total} refers to the total number of models being considered, P refers to the proportion of models considered native-like by C_α $RMSD_{100SSE}$, and $TP_{P,Score}$ refers to the number of true positives identified in the top $P \times N_{total}$ models by score (9). We treated the top 10% of models as native-like ($P =$ 0.1), thus scaling the metric from -1 (none of the top 10% of models by $RMSD_{100SSE}$ were in the top 10% by score) to 1 (all of the top 10% of models by $RMSD_{100SSE}$ were also in the top 10% by score), with a value of 0 indicating that the number of native-like models found in the top 10% by score was equal to what is expected by chance.

Oscillation frequencies of decay traces for distributions with an average distance r_{avg} (in angstroms) were calculated as $r_{avg}^3/(5.2 \times 10^4)$ μs (40). Decay traces with fewer than three oscillations were not used to evaluate enrichment as a function of decay trace duration.

De novo protein structure prediction benchmark

The protein structure prediction protocol we used largely follows a previously published template (57) and consists of three stages. In the first stage, 10,000 models were generated using extended chain AbInitio with either RosettaDEER restraints, CONE model restraints (5), or no restraints. This protocol relies on the insertion of fragments obtained from a July 2011 copy of the Protein Data Bank and was obtained from the Robetta online server (58); homologous protein structures were excluded from these fragment libraries. The contribution of the RosettaDEER score term was adjusted so that its dynamic range was similar to that of the Rosetta energy function (57). Because the proportion of DEER restraints relative to the protein length was comparable for Bax and ExoU, the impact of the number of restraints on the weight of the score term was not considered (59).

Models generated this way were then clustered to a radius of 7.5 Å C_α $RMSD_{100}$ using Durandal (60). Each cluster was evaluated by scoring its models using both RosettaDEER and the full-atom Rosetta energy function (53), obtaining the cluster averages for both values, and adding their Z-scores with respect to those of other clusters. After discarding sparsely populated clusters ($<5\%$ of the size of the largest cluster), the top 10 best-scoring models by combined Z-score were selected from the five best-scoring clusters for subsequent modeling.

An additional 1200 models were generated from these 50 models using RosettaCM (52), which also relies on fragment insertion but ensures that the input model's topology is retained throughout the modeling process. The scripts were obtained from a recently published refinement protocol (61), and no experimental restraints were used. Models generated during this stage were again clustered to 7.5 Å C_α $RMSD_{100}$ and scored, except only the RosettaDEER score was used to evaluate the quality of these models.

During the third and final stage, models in the best-scoring cluster were minimized using FastRelax (62), which introduces and repacks side chains while performing gradient descent on a full-atom depiction of the entire model. Models generated at this stage were scored exclusively using the native Rosetta energy function, with the lowest-scoring model selected as the output model.

RESULTS

Modeling nitroxide spin labels using RosettaDEER

A strategy to model proteins using DEER data must reliably simulate distance distributions between spin-labeled residues. To quantify the computational cost and efficiency of this task, we considered a panel of five proteins in which both atomic-detail structures and experimental DEER data were available (Table S1; (11,14,47–49)). Distance distributions between residue pairs that have been previously measured experimentally were simulated using a number of methods, and the resulting error was quantified as the difference between the average values of the simulated and experimental distance distributions (example shown in Fig. 1 A). In addition, we measured how rapidly each program calculated these distance distributions. (Fig. 1 B). Consistent with previous results (10,26,33), the average values of experimental distance distributions gathered in monomeric proteins, but not the homodimer CDB3, agree more closely with those of simulated distributions than their corresponding C_β-C_β distances, from which restraints such as the CONE model are derived (Table S2; (5)). By contrast, none of the methods examined here reliably reproduced the width of the distance distributions. This is likely attributable to oversampling of available conformational space of the spin label, which results from the exclusive use of van der Waals repulsive energies to limit possible rotameric configurations. Finally, the data revealed how simulation times varied substantially between these methods.

These results further illustrate that increasing computational complexity did not lead to more accurate distance distributions. We hypothesized that, for the same reason, decreasing the computational complexity would not lead to less accurate distance distributions. Therefore, RosettaDEER's design prioritized computational efficiency (see Materials and Methods). Rather than measure distances from full-atom rotamers or mobile dummy atoms, RosettaDEER uses a probability density function to capture high-occupancy electron positions that would be explored by MTSSL and map them onto the protein structure (Fig. 1, C and D). For each of these coordinates, an evaluation of a potential van der Waals overlap was performed between a pseudoatom representing the nitroxide ring's center of mass and the rest of the protein. Placing this pseudoatom at an idealized location, consistent with spin-labeled protein structures in the PDB (Fig. S2; Table S3), reduced the number of atoms for this evaluation to one per rotamer, thus maximizing computational efficiency. Fig. 1, A and B demonstrate that RosettaDEER's simplified representation of the spin label allows the generation of distance distributions three to five orders of magnitude faster than other approaches but with comparable accuracy.

Comparison of simulations with experimental DEER decay traces

Most existing methods that leverage DEER experimental data for structural modeling require that the primary spectroscopic readout first be processed into a distance distribution. A conventional approach, such as the Rosetta CONE model, is outlined in Fig. 2 A. This involves 1) manually identifying and removing the "background" signal, which corresponds to coupling between spin labels across macromolecules; 2) using Tikhonov regularization to convert the remaining intramolecular signal into a distance distribution; and 3) selecting a single distance value from this distribution to restrain the modeling process (3). An additional bias is often required to convert these distance data into backbone restraints (5,6,30,63).

We reasoned that these preprocessing steps could be avoided by simulating a spectroscopic signal from candidate models for direct comparison to the experimental data. As with other forward approaches to fitting DEER data (40,46), the steps are as follows: 1) the model is used to generate a distance distribution, 2) this distance distribution is converted into a spectroscopic signal consisting solely of the effect of coupling between spin labels attached to the same macromolecule, and 3) the slope of the "background" coupling and depth of modulation needed to optimally fit the simulated and experimental decay traces are determined.

This final step represents the outstanding challenge in the proposed pipeline because most modeling programs, including Rosetta, focus on isolated protein structural
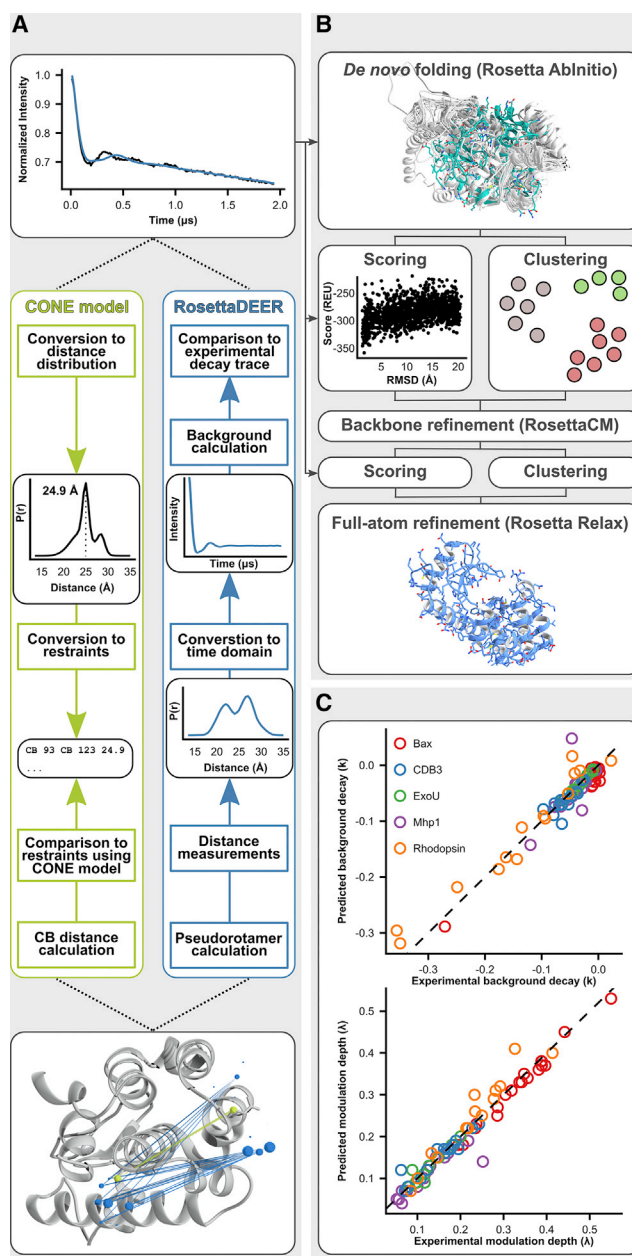
FIGURE 2 RosettaDEER simulations of distance distributions and decay traces. The forward approach taken by RosettaDEER contrasts with the preprocessing required by the CONE model. (A) A flowchart illustrating how both the CONE model and RosettaDEER use experimental DEER data to model proteins (example shown is T4 lysozyme residues 93 and 123). (B) Incorporation of DEER experimental restraints into Rosetta structure prediction pipeline. (C) Recovery of experimental background coupling and modulation depth parameter values.

models. We instead used a two-parameter exponential function to simulate the background coupling (k) and modulation depth (λ) (see Materials and Methods). The values of these parameters were determined by minimizing the sum of the squared residuals. The optimum values obtained strongly correlated with those obtained using DeerAnalysis (38), with r^2 values exceeding 0.90 for both parameters

(Fig. 2 C), despite the fact that the inaccuracies in the distance distributions affected the fit (Fig. S4). In fact, we found that this correspondence correlated less strongly with the goodness of fit in the distance domain than it did with the quality of the experimental data in the time domain (Fig. S5; Table S2).

Enrichment of native-like models using experimental decay traces

Being able to simulate DEER traces from candidate structural models without any preprocessing offers the possibility to reframe the problem currently faced by translating the DEER traces into distance distributions. Whereas methods such as Tikhonov regularization convert individual DEER traces into distance distributions, RosettaDEER, in conjunction with Monte Carlo modeling, would instead seek to determine the structural model most consistent with both an energy function and the experimental data. To investigate whether unprocessed DEER traces can be used to discriminate native-like models from incorrectly folded models, we generated a series of 1000–2000 misfolded models for each of the five proteins in our test set and scored their agreement with experimental DEER data. In addition, we generated 1000 docked models of the homodimer CDB3, which retained the native fold for the protomer but not the oligomeric interface. Similarity to the native model was measured by C_α $RMSD_{100SSE}$ (56), which is the size-normalized RMSD across secondary structural elements (Fig. 3). RosettaDEER's effectiveness at this task was measured by the enrichment parameter, which is defined in the Materials and Methods and quantifies a scoring function's ability to discriminate native-like models from incorrectly folded models.

RosettaDEER consistently scored native-like models of the monomeric proteins more favorably than poorly folded models (Fig. 3). This was also observed with correctly docked models of CDB3. Moreover, it generally outperformed the CONE model in enriching native-like models (Fig. S6). Perhaps unsurprisingly, the simultaneous use of Rosetta's energy function often improved enrichment because it overwhelmingly considers short-range interactions and is therefore expected to complement the evaluation of longer-range, fold-level information provided by DEER restraints (Fig. S6; (53)). We note that RosettaDEER could not effectively identify misfolded models of CDB3, which we attribute to the fact that DEER restraints reflect distances across the center of symmetry rather than within the protomer. Nevertheless, these results suggest that RosettaDEER's inability to perfectly recreate the experimental DEER data did not impede its ability to identify correctly folded models, suggesting that it could be effectively used for structure prediction.

The fact that structural models are scored based on their consistency with the primary spectroscopic data led us to hypoth-
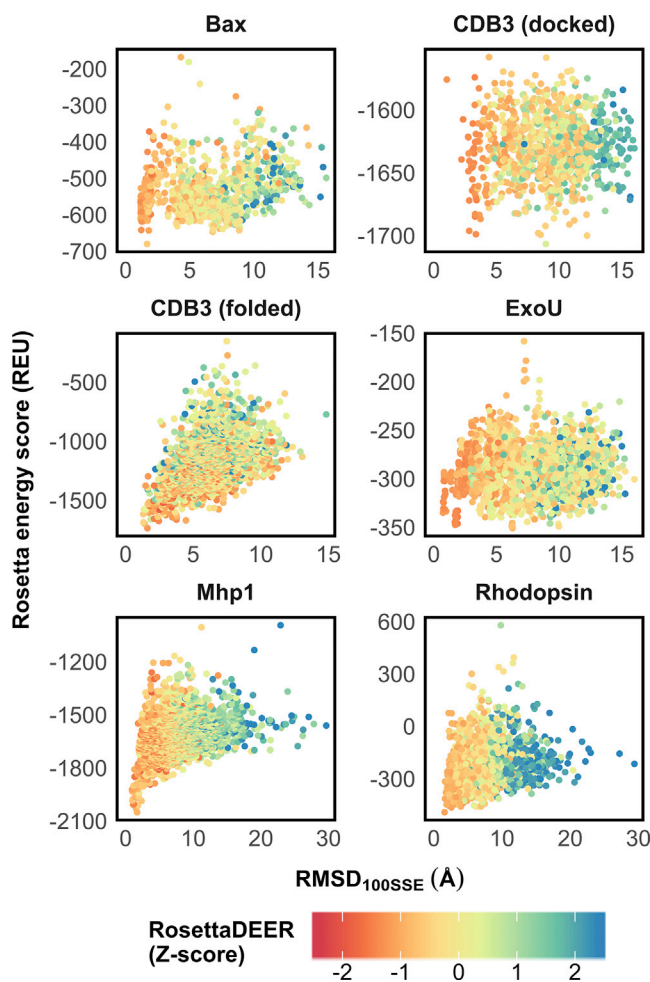

FIGURE 3 Evaluation of models using DEER decay traces. Models with C_α $RMSD_{100SSE}$ ranging from 0.5 Å to 20.0–30.0 Å were scored using both the Rosetta energy function and RosettaDEER.

esize that they could be evaluated using lower-quality data than what would be necessary for conversion into precise distance distributions. We were specifically interested in evaluating the importance of the experimental data's time window, which must undergo roughly 0.8 and 1.6 oscillations for Tikhonov regularization to accurately identify a distance distribution's average and standard deviation, respectively (22). This hypothesis was tested by artificially truncating the experimental data in the time domain and measuring enrichment as a function of how many oscillations were included (see Materials and Methods, Fig. S6). Strikingly, RosettaDEER could enrich native-like models of Bax, ExoU, rhodopsin, and Mhp1 with highly truncated data (<0.8 oscillations), albeit to a reduced degree. We found that the addition of data in the time domain beyond one oscillation failed to lead to any measurable improvements in enrichment, despite its importance in allowing RosettaDEER to identify the correct background coupling parameters (Fig. S5). These results suggest that RosettaDEER is more permissive than Tikhonov regularization with respect to the effect of data quality on protein structural modeling.

De novo folding of Bax and ExoU

To further illustrate RosettaDEER's capability to identify native-like models, we folded Bax and ExoU de novo using experimental DEER decay data. These two proteins were chosen because native-like models cannot be identified using the default Rosetta energy function alone (Figs. 3 and S6). The structure prediction protocol we used is similar to one used to model proteins using other types of sparse data (57,58) and is illustrated in Fig. 2 B and described in detail in Materials and Methods. We first generated an initial set of 10,000 models using Rosetta AbInitio folding supplemented by experimental restraints through RosettaDEER, experimental restraints through the CONE model (5), or no restraints. These models were then clustered, and models from the best-scoring clusters were refined and recombined into 1200 new models without using experimental data. After a second round of clustering, models from the cluster with the best agreement to the experimental data were refined and minimized, and the model with the best Rosetta energy score was returned as the predicted model.

In the absence of experimental restraints, few of the models generated by AbInitio folding resembled the native fold (Fig. 4 A). Perhaps strikingly, providing DEER restraints with the CONE model had no effect on the proportion of native-like models of ExoU generated this way (a measurable improvement was observed when folding Bax). This contrasts with the proportion of native-like models generated using RosettaDEER, which was substantially higher in the case of both proteins.

Although agreement between models and experimental structures loosely correlated with both RosettaDEER score and Rosetta energy score for both proteins, an abundance of incorrectly folded models obscured this trend (Fig. 4 B; RosettaDEER and Rosetta energy scores were jointly considered by adding the Z-scores of each). As a result, we were unable to identify native-like models for either Bax of ExoU from score values alone. The 10 best-scoring models by these metrics were generally incorrectly folded (5–10 Å C_α RMSD$_{100SSE}$) and buried amphipathic features found on the surface of the native model. This shortcoming is typically addressed by clustering because native-like models are more likely to be found near the centers of large clusters with favorable average scores (64). We therefore clustered Bax and ExoU models with a radius of 7.5 Å and evaluated these clusters by taking the Z-scores of both the average Rosetta energy and RosettaDEER score and adding them together (Fig. 4 B). In the case of both proteins, this step placed native-like models in the best-scoring clusters. Focusing our attention on the five best-scoring clusters allowed us to discard 85.3% of the Bax models and 61.3% of the ExoU models while retaining a majority of the native-like models in each case.

Each cluster at this stage represented a broad population of models that satisfied the DEER data. To test whether
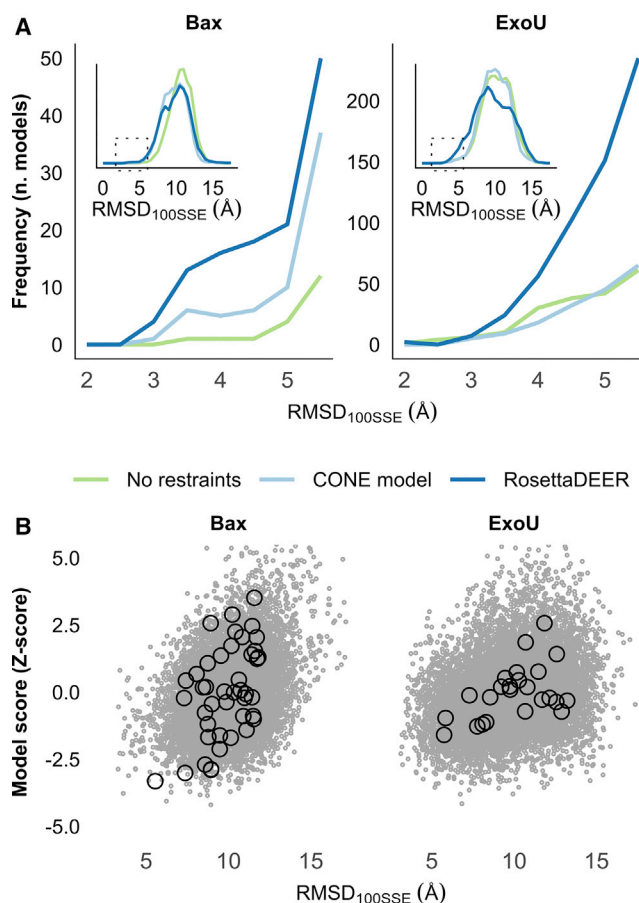

FIGURE 4 Structure prediction of Bax and ExoU using DEER decay data. (A) De novo protein folding of native-like models using DEER decay restraints with RosettaDEER, C_β-C_β distance restraints with the CONE model, or no restraints. Inset: spread of all models generated using these three methods. (B) Accuracy of de novo folded models (gray dots) and clusters (black circles) as a function of combined DEER and Rosetta Z-score.

refining models without experimental restraints would reveal the native fold, 10 models from each of the top five clusters were refined and recombined using RosettaCM (52). This step retained the topology of the input models but permitted minor backbone rearrangements that allowed misfolded models to optimize away from conformations consistent with the experimental data. As a result, the cluster with the most native-like models after this resampling stage scored the most favorably by RosettaDEER. After minimization of models in this cluster (62), the best-scoring model by Rosetta score for both Bax and ExoU had near-native folds (<3.5 Å C_α RMSD$_{100SSE}$; Figs. 5 and S8).

DISCUSSION

RosettaDEER predicts and refines protein structures by integrating DEER spectroscopy data and Rosetta computational modeling protocols. To our knowledge, the novel aspects of
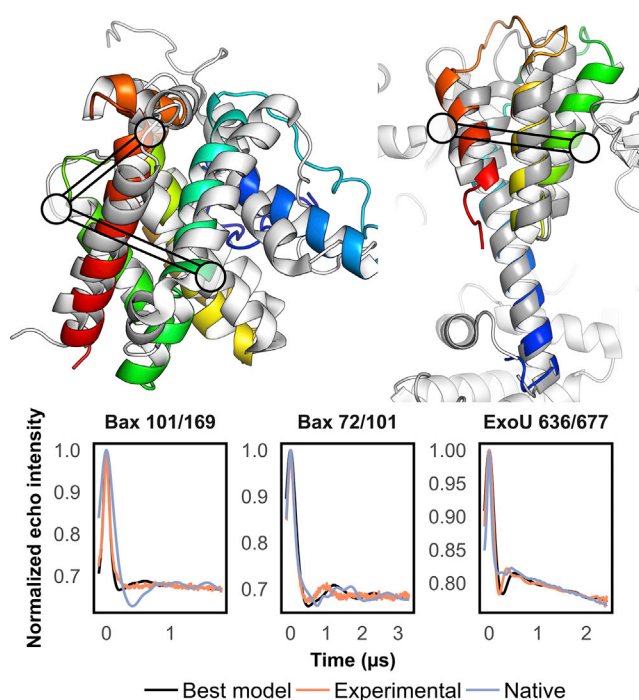
FIGURE 5 Predicted models of Bax and ExoU at 3.2 and 2.1 Å C_α RMSD$_{100SSE}$. (*Top*) Models were obtained from 10,000 de novo folded models, the best-scoring of which were refined into 1200 additional models. Native models shown in white. (*Bottom*) Example DEER traces in which the best model outperformed the native. Corresponding residues indicated as circles in (*A*) and (*B*).

this method are a simplified representation of the commonly used spin label MTSSL and a strategy to rapidly simulate DEER decay traces for comparison to uncorrected experimental traces. The robustness of the method was demonstrated by benchmarking every step on five sparse data sets. Despite the simplified spin label representation, the distance distributions simulated by RosettaDEER are comparable to those generated using more computationally complex rotamer library approaches. Moreover, even though simulated spectra fail to perfectly fit experimental DEER traces, this integrated approach efficiently identifies conformations that simultaneously satisfy the data and the Rosetta energy function. Our findings illustrate how RosettaDEER can complement similar methods that are more computationally intensive but able to use DEER decay data to perform high-resolution refinement of protein structures (16).

The de novo folding benchmark with the small soluble proteins ExoU and Bax highlights the success of this strategy. Both proteins possess surface-exposed amphipathic regions that insert into the membrane. Bax transitions from a soluble monomer into a membrane-bound oligomer using its C-terminal helix (47), whereas ExoU is hypothesized to move into the membrane using a flexible loop between its two C-terminal helices (65). Consistent with previous results (10,11), the Rosetta energy function favored models that packed these substructures in the protein core, leading

to incorrectly folded models and lack of correlation between the Rosetta score and model accuracy. As a result, orthogonal experimental data that define the structure are critical to de novo folding. Our folding benchmark suggests that RosettaDEER more effectively leverages the experimental data than the C_β-based CONE model. Moreover, even low-quality data can be used to discriminate native-like from incorrectly folded models. We appreciate that, for larger proteins, structure determination from DEER experiments alone would require extensive experimental data. Integrating RosettaDEER with other types of sparse experimental data could therefore reduce the number of DEER restraints required for accurate modeling.

The strategy of RosettaDEER to predict the structures of these two proteins leverages the experimental data by folding and optimizing protein structures with and without restraints, respectively. The first step leads to a substantial reduction in the search space and a concomitant increase in the number of models that satisfy the restraints, although not all of these models are correctly folded. After clustering the models to remove those that correspond to narrow energy minima, the second step, optimization without restraints, allows clusters with incorrectly folded models to reach energy minima inconsistent with the data. This filtering procedure restores the experimental data's ability to identify native-like models because the most native-like models of Bax and ExoU at this stage were not identifiable using the Rosetta energy function. Overall, this protocol decreases both the number of incorrectly folded structures that fit the data and the conformational search space inherent to the protein folding problem.

Despite its success illustrated here, the current implementation of RosettaDEER assumes that a single protein conformation describes the data. For example, the distance distributions of Mhp1, the most conformationally flexible protein examined in this data set, were generally more poorly simulated using available methods than those collected in other proteins. Experimental applications of the DEER technique often focus on monitoring ensembles of protein conformations and require computational methods that interpret this data with the capability to generate multiple models and examine their consistency with sparse experimental data. This is the next step for RosettaDEER. Furthermore, a Rosetta de novo folding protocol for membrane-associated proteins that includes a model membrane would be desirable for proteins such as Bax and ExoU.

SUPPORTING MATERIAL

AUTHOR CONTRIBUTIONS

D.d.A., H.S.M., and J.M. conceived the study and wrote the manuscript. D.d.A. developed and implemented RosettaDEER and performed

computational experiments with input from R.A.S; M.H.T. and J.B.F. designed and performed DEER experiments on ExoU.

ACKNOWLEDGMENTS

REFERENCES

1. Steven, A. C., and W. Baumeister. 2008. The future is hybrid. *J. Struct. Biol.* 163:186–195.

2. Xia, Y., A. W. Fischer, …, J. Meiler. 2018. Integrated structural biology for α-Helical membrane protein structure determination. *Structure.* 26:657–666.e2.

3. Jeschke, G. 2018. The contribution of modern EPR to structural biology. *Emerg. Top. Life Sci.* 2:9–18.

4. Sahu, I. D., and G. A. Lorigan. 2018. Site-directed spin labeling EPR for studying membrane proteins. *BioMed Res. Int.* 2018:3248289.

5. Alexander, N., M. Bortolus, …, J. Meiler. 2008. De novo high-resolution protein structure determination from sparse spin-labeling EPR data. *Structure.* 16:181–195.

6. Yang, Y., T. A. Ramelot, …, M. A. Kennedy. 2010. Combining NMR and EPR methods for homodimer protein structure determination. *J. Am. Chem. Soc.* 132:11910–11913.

7. Hirst, S. J., N. Alexander, …, J. Meiler. 2011. RosettaEPR: an integrated tool for protein structure determination from sparse EPR data. *J. Struct. Biol.* 173:506–514.

8. Kazmier, K., N. S. Alexander, …, H. S. Mchaourab. 2011. Algorithm for selection of optimized EPR distance restraints for de novo protein structure determination. *J. Struct. Biol.* 173:549–557.

9. Fischer, A. W., N. S. Alexander, …, J. Meiler. 2015. BCL:MP-fold: membrane protein structure prediction guided by EPR restraints. *Proteins.* 83:1947–1962.

10. Fischer, A. W., E. Bordignon, …, J. Meiler. 2016. Pushing the size limit of de novo structure ensemble prediction guided by sparse SDSL-EPR restraints to 200 residues: the monomeric and homodimeric forms of BAX. *J. Struct. Biol.* 195:62–71.

11. Fischer, A. W., D. M. Anderson, …, J. Meiler. 2017. Structure and dynamics of type III secretion effector protein ExoU as determined by SDSL-EPR spectroscopy in conjunction with de novo protein folding. *ACS Omega.* 2:2977–2984.

12. Ling, S., W. Wang, …, C. Tian. 2016. Structure of an *E. coli* integral membrane sulfurtransferase and its structural transition upon SCN(-) binding defined by EPR-based hybrid method. *Sci. Rep.* 6:20025.

13. Kazmier, K., S. Sharma, …, H. S. Mchaourab. 2014. Conformational dynamics of ligand-dependent alternating access in LeuT. *Nat. Struct. Mol. Biol.* 21:472–479.

14. Kazmier, K., S. Sharma, …, H. S. Mchaourab. 2014. Conformational cycle and ion-coupling mechanism of the Na+/hydantoin transporter Mhp1. *Proc. Natl. Acad. Sci. USA.* 111:14752–14757.

15. Raghuraman, H., S. M. Islam, …, E. Perozo. 2014. Dynamics transitions at the outer vestibule of the KcsA potassium channel during gating. *Proc. Natl. Acad. Sci. USA.* 111:1831–1836.

16. Marinelli, F., and G. Fiorin. 2019. Structural characterization of biomolecules through atomistic simulations guided by DEER measurements. *Structure.* 27:359–370.e12.

17. Hilger, D., Y. Polyhach, …, G. Jeschke. 2007. High-resolution structure of a Na+/H+ antiporter dimer obtained by pulsed electron paramagnetic resonance distance measurements. *Biophys. J.* 93:3675–3683.

18. Edwards, S. J., C. W. Moth, …, T. P. Lybrand. 2014. Automated structure refinement for a protein heterodimer complex using limited EPR spectroscopic data and a rigid-body docking algorithm: a three-dimensional model for an ankyrin-CDB3 complex. *J. Phys. Chem. B.* 118:4717–4726.

19. Bhatnagar, J., J. H. Freed, and B. R. Crane. 2007. Rigid body refinement of protein complexes with long-range distance restraints from pulsed dipolar ESR. *Methods Enzymol.* 423:117–133.

20. Pannier, M., S. Veit, …, H. W. Spiess. 2000. Dead-time free measurement of dipole-dipole interactions between electron spins. *J. Magn. Reson.* 142:331–340.

21. McHaourab, H. S., P. R. Steed, and K. Kazmier. 2011. Toward the fourth dimension of membrane protein structure: insight into dynamics from spin-labeling EPR spectroscopy. *Structure.* 19:1549–1561.

22. Jeschke, G. 2012. DEER distance measurements on proteins. *Annu. Rev. Phys. Chem.* 63:419–446.

23. Abdullin, D., G. Hagelueken, and O. Schiemann. 2016. Determination of nitroxide spin label conformations via PELDOR and X-ray crystallography. *Phys. Chem. Chem. Phys.* 18:10428–10437.

24. Iwahara, J., C. D. Schwieters, and G. M. Clore. 2004. Ensemble approach for NMR structure refinement against (1)H paramagnetic relaxation enhancement data arising from a flexible paramagnetic group attached to a macromolecule. *J. Am. Chem. Soc.* 126:5879–5896.

25. Alexander, N. S., A. M. Preininger, …, J. Meiler. 2014. Energetic analysis of the rhodopsin-G-protein complex links the α5 helix to GDP release. *Nat. Struct. Mol. Biol.* 21:56–63.

26. Sale, K., L. Song, …, P. Fajer. 2005. Explicit treatment of spin labels in modeling of distance constraints from dipolar EPR and DEER. *J. Am. Chem. Soc.* 127:9334–9335.

27. MacCallum, J. L., A. Perez, and K. A. Dill. 2015. Determining protein structures by combining semireliable data with atomistic physical models by Bayesian inference. *Proc. Natl. Acad. Sci. USA.* 112:6985–6990.

28. Krug, U., N. S. Alexander, …, J. Meiler. 2016. Characterization of the domain orientations of *E. coli* 5′-nucleotidase by fitting an ensemble of conformers to DEER distance distributions. *Structure.* 24:43–56.

29. Dastvan, R., E. M. Brouwer, …, T. F. Prisner. 2016. Relative orientation of POTRA domains from cyanobacterial Omp85 studied by pulsed EPR spectroscopy. *Biophys. J.* 110:2195–2206.

30. Alexander, N. S., R. A. Stein, …, J. Meiler. 2013. RosettaEPR: rotamer library for spin label structure and dynamics. *PLoS One.* 8:e72851.

31. Marinelli, F., and J. D. Faraldo-Gómez. 2015. Ensemble-biased metadynamics: a molecular simulation method to sample experimental distributions. *Biophys. J.* 108:2779–2782.

32. Polyhach, Y., E. Bordignon, and G. Jeschke. 2011. Rotamer libraries of spin labelled cysteines for protein studies. *Phys. Chem. Chem. Phys.* 13:2356–2366.

33. Hagelueken, G., R. Ward, …, O. Schiemann. 2012. MtsslWizard: in silico spin-labeling and generation of distance distributions in PyMOL. *Appl. Magn. Reson.* 42:377–391.

34. Hatmal, M. M., Y. Li, …, I. S. Haworth. 2012. Computer modeling of nitroxide spin labels on proteins. *Biopolymers.* 97:35–44.

35. Islam, S. M., R. A. Stein, …, B. Roux. 2013. Structural refinement from restrained-ensemble simulations based on EPR/DEER data: application to T4 lysozyme. *J. Phys. Chem. B.* 117:4740–4754.

36. Meiler, J., and D. Baker. 2003. Rapid protein fold determination using unassigned NMR data. *Proc. Natl. Acad. Sci. USA.* 100:15404–15409.

37. Wang, R. Y., Y. Song, …, F. DiMaio. 2016. Automated structure refinement of macromolecular assemblies from cryo-EM maps using Rosetta. *eLife.* 5:1–22.

38. Jeschke, G., V. Chechik, …, H. Jung. 2006. DeerAnalysis2006—a comprehensive software package for analyzing pulsed ELDOR data. *Appl. Magn. Reson.* 30:473–498.

39. Brandon, S., A. H. Beth, and E. J. Hustedt. 2012. The global analysis of DEER data. *J. Magn. Reson.* 218:93–104.

40. Hustedt, E. J., F. Marinelli, …, H. S. Mchaourab. 2018. Confidence analysis of DEER data and its structural interpretation with ensemble-biased metadynamics. *Biophys. J.* 115:1200–1216.

41. Sen, K. I., T. M. Logan, and P. G. Fajer. 2007. Protein dynamics and monomer-monomer interactions in AntR activation by electron paramagnetic resonance and double electron-electron resonance. *Biochemistry.* 46:11639–11649.

42. Worswick, S. G., J. A. Spencer, …, I. Kuprov. 2018. Deep neural network processing of DEER data. *Sci. Adv.* 4:eaat5218.

43. Srivastava, M., and J. H. Freed. 2017. Singular value decomposition method to determine distance distributions in pulsed dipolar electron spin resonance. *J. Phys. Chem. Lett.* 8:5648–5655.

44. Hogben, H. J., M. Krzystyniak, …, I. Kuprov. 2011. Spinach–a software library for simulation of spin dynamics in large spin systems. *J. Magn. Reson.* 208:179–194.

45. Leaver-fay, A., M. Tyka, …, P. Bradley. 2011. ROSETTA 3: an object-oriented software suite for the simulation and design of macromolecules. *Methods Enzymol.* 487:545–574.

46. Stein, R. A., A. H. Beth, and E. J. Hustedt. 2015. A straightforward approach to the analysis of double electron-electron resonance data. *Methods Enzymol.* 563:531–567.

47. Bleicken, S., G. Jeschke, …, E. Bordignon. 2014. Structural model of active Bax at the membrane. *Mol. Cell.* 56:496–505.

48. Zhou, Z., S. C. DeSensi, …, A. H. Beth. 2005. Solution structure of the cytoplasmic domain of erythrocyte membrane band 3 determined by site-directed spin labeling. *Biochemistry.* 44:15115–15128.

49. Altenbach, C., A. K. Kusnetzow, …, W. L. Hubbell. 2008. High-resolution distance mapping in rhodopsin reveals the pattern of helix movement due to activation. *Proc. Natl. Acad. Sci. USA.* 105:7439–7444.

50. Jo, S., X. Cheng, …, W. Im. 2014. CHARMM-GUI PDB manipulator for advanced modeling and simulations of proteins containing nonstandard residues. *Adv. Protein Chem. Struct. Biol.* 96:235–265.

51. Krivov, G. G., M. V. Shapovalov, and R. L. Dunbrack, Jr. 2009. Improved prediction of protein side-chain conformations with SCWRL4. *Proteins.* 77:778–795.

52. Song, Y., F. DiMaio, …, D. Baker. 2013. High-resolution comparative modeling with RosettaCM. *Structure.* 21:1735–1742.

53. Alford, R. F., A. Leaver-Fay, …, J. J. Gray. 2017. The Rosetta all-atom energy function for macromolecular modeling and design. *J. Chem. Theory Comput.* 13:3031–3048.

54. Yarov-Yarovoy, V., J. Schonbrun, and D. Baker. 2006. Multipass membrane protein structure prediction using Rosetta. *Proteins.* 62:1010–1025.

55. Viklund, H., and A. Elofsson. 2008. OCTOPUS: improving topology prediction by two-track ANN-based preference scores and an extended topological grammar. *Bioinformatics.* 24:1662–1668.

56. Carugo, O., and S. Pongor. 2001. A normalized root-mean-square distance for comparing protein three-dimensional structures. *Protein Sci.* 10:1470–1473.

57. Ovchinnikov, S., L. Kinch, …, D. Baker. 2015. Large-scale determination of previously unsolved protein structures using evolutionary information. *eLife.* 4:e09248.

58. Kim, D. E., D. Chivian, and D. Baker. 2004. Protein structure prediction and analysis using the Robetta server. *Nucleic Acids Res.* 32:W526–W531.

59. Weiner, B. E., N. Alexander, …, J. Meiler. 2014. BCL:Fold–protein topology determination from limited NMR restraints. *Proteins.* 82:587–595.

60. Berenger, F., R. Shrestha, …, K. Y. J. Zhang. 2012. Durandal: fast exact clustering of protein decoys. *J. Comput. Chem.* 33:471–474.

61. Park, H., S. Ovchinnikov, …, D. Baker. 2018. Protein homology model refinement by large-scale energy optimization. *Proc. Natl. Acad. Sci. USA.* 115:3054–3059.

62. Conway, P., M. D. Tyka, …, D. Baker. 2014. Relaxation of backbone bond geometry improves protein energy landscape modeling. *Protein Sci.* 23:47–55.

63. Sale, K., J. L. Faulon, …, M. M. Young. 2004. Optimal bundling of transmembrane helices using sparse distance constraints. *Protein Sci.* 13:2613–2627.

64. Shortle, D., K. T. Simons, and D. Baker. 1998. Clustering of low-energy conformations near the native structures of small proteins. *Proc. Natl. Acad. Sci. USA.* 95:11158–11162.

65. Tessmer, M. H., D. M. Anderson, …, J. B. Feix. 2017. Cooperative substrate-cofactor interactions and membrane localization of the bacterial phospholipase A2(PLA2) enzyme. *J. Biol. Chem.* 292:3411–3419.