Refinement of highly flexible protein structures using simulation-guided spectroscopy.

Jennifer M. Hays[1], Marissa K. Kieber[2], Jason Z. Li[2], Ji In Han[2], Kelley W. Moremen[3], Linda Columbus[2*], and Peter M. Kasson[1,4*].

[1]Departments of Biomedical Engineering and Molecular Physiology, University of Virginia, Charlottesville VA 22908, [2]Department of Chemistry, University of Virginia, Charlottesville VA 22908, [3]Complex Carbohydrate Research Center, University of Georgia, Athens GA 30602, [4]Science for Life Laboratory, Program in Molecular Biophysics, Uppsala University, Uppsala SE 75124.

* kasson@virginia.edu or columbus@virginia.edu.

Abstract

Highly flexible proteins present a special challenge for structure determination because they are multi-structured yet not disordered, and the resulting conformational ensembles are essential for understanding their function. Determining such ensembles is difficult because many measurements that capture multiple conformational populations provide only sparse data. A powerful opportunity exists to leverage molecular simulations to select which experiments will yield most efficient refinement. We have developed an information-theoretic approach to guide experiments by ranking which measurements will best refine the underlying conformational ensemble. We have prospectively tested this approach on the Neisserial virulence-associated protein Opa_{60}, identifying maximally informative residue pairs, measuring distance distributions using DEER spectroscopy, and assessing how well these data refine the Opa_{60} conformational ensemble. Our simulation-guided approach significantly outperforms standard residue-pair selection, providing a systematic method for iterative experiment selection and incorporation. It further yields structural insight into how Opa_{60} engages cellular receptors while maintaining overall conformational flexibility.

Introduction

Conformationally flexible biomolecules play a critical role in immunity and infection, particularly in problems of flexible molecular recognition [1-3]. This flexibility is precisely the reason their conformational ensembles are difficult to refine and, thus, a structure – function relationship remains elusive. Determining the structures of conformations in equilibrium and how these structures interact with other proteins and substrates to carry out functions such as cellular uptake and colonization is a particular challenge and will almost certainly require multimodal or hybrid approaches. Because experimental techniques tend to either capture a small number of highly resolved states or a large number of under-determined states, a better method is required to select optimal experimental restraints and integrate these restraints to model the conformational ensemble. Here we outline the specific challenges involved in the structure determination of flexible systems and present a general method that leverages molecular dynamics (MD) simulations and information theory to guide double electron-electron resonance (DEER, also known as PELDOR) experiments and, subsequently, refine the conformational ensemble of the highly flexible Neisserial virulence-associated protein Opa_{60}.

Refinement of conformational ensembles of flexible biomolecular systems is difficult because so many states contribute to the ensembles[4-7]. Both experimental and computational techniques often fail to fully capture flexible ensembles at high resolution. High-resolution experimental methods, such as X-ray crystallography, capture a few low-energy conformations and therefore lose information about the

conformational diversity of the system. Conversely, spectroscopic techniques such as single-molecule FRET and DEER provide quantitative data about relative populations of different conformations but report on only a very small subset of atomic distances. Finally, MD simulations can model the behavior of these systems but often require prohibitively large amounts of sampling in order to yield a converged approximation of the free energy landscape.

Since DEER experiments yield sparse data, choosing which measurements to perform is a critical challenge. Exhaustive sampling is not feasible: for a protein of 100 amino acids, a DEER spectroscopist would in the most extreme case have to measure 5,000 residue-residue pairs. Random selection of pairs may succeed for rigid to moderately flexible proteins, but these approaches are less likely to succeed as protein flexibility increases because a poorly selected pair distinguishes much less well among flexible conformations than a well selected pair. As a result, spectroscopists must often perform more experiments than are strictly necessary in order to refine the conformational ensemble. In order to address this challenge, we have developed an iterative methodology that leverages MD simulations to inform DEER experiments in order to refine the conformational ensembles of flexible proteins requiring as few DEER measurements as possible. MD simulations [8-12] or Monte Carlo methods [13-19] are often used to help incorporate DEER measurements into structural refinement, but using simulations to guide selection of DEER experiments is a challenging task for which general approaches have not yet been developed. Early efforts at probe selection from MD simulations have relied on pre-existing high-resolution structural and kinetic models and used retrospective testing [20]. Here, we develop a model-free formulation that is much more general and perform prospective prediction and experimental testing to validate it. The method is robust and applicable to any experimental restraint such as those obtained with FRET and NMR-derived restraints (e.g. NOEs and PREs).

We have used this methodology to model the conformational ensemble of the Neisserial virulence-associated protein Opa_{60}. Opacity associated (Opa) proteins bind to human carcinoembryonic antigen-related cellular adhesion molecules (CEACAMs) triggering cellular uptake and mediating cellular invasion [21]. Opa_{60} is a canonical eight-stranded β-barrel integral membrane protein with four extracellular loops (Fig. 2-Supplement 1) that are dynamic on the nanosecond time scale and predominantly disordered [22]. The long ligand-binding loops have high sequence diversity [23] and are hypothesized to be flexible in both the unbound state and during CEACAM engagement, which aids in immune evasion [22]. Despite high mobility and high sequence diversity among Opa variants, the loops bind with high affinity to CEACAM1 [24]. Fully understanding the nature of this molecular recognition event thus requires measuring the conformational ensembles of both unliganded and CEACAM-bound Opa_{60}. To help understand this process of binding and subsequent invasion, we have determined the conformational ensemble of Opa_{60} through DEER experiments directed by MD simulations. This model reveals previously uncharacterized loop-loop interaction modes and potential sites for CEACAM engagement. Our methodology is directly extensible to measuring conformational ensembles of flexible proteins such as those that mediate many host-pathogen interactions.

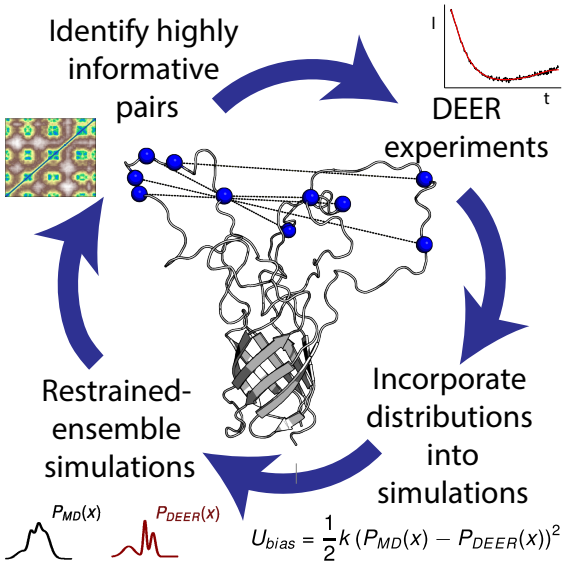

Figure 1: An iterative experimental and computational approach for refining the conformational ensembles of flexible proteins. Analyzed DEER data are incorporated into our scheme as distance distributions. These distributions are used to drive restrained-ensemble MD simulations in which the simulation distance distribution is biased toward the experimental distribution. Analysis of the MD simulations yields a set of maximally informative, minimally redundant pairs. Incorporating these pairs, as opposed to a set chosen based on current practice in the field, into a subsequent round of MD simulations permits efficient refinement of the conformational ensemble.

Theory

Refining flexible conformational ensembles with pair-wise distance information involves three steps (Fig. 1): choosing pairs to measure, performing those measurements, and building a multi-conformation model using the resulting data. The primary challenge we address here is the selection of the pairs. We therefore provide a detailed discussion of DEER pair selection and then only briefly summarize the way in which we use DEER restraints to bias simulations.

Selection of pairs: mRMR Algorithm. An ideal set of DEER pairs would have two properties: 1) each selected pair should resolve *as many other distances* in the system as possible and 2) each selected pair should resolve *distances that are distinct* from those determined by each of the other measured pairs. These criteria are optimally satisfied by selecting pairs using the information-theoretic criteria of maximum relevance and minimum redundancy (mRMR)[25]. We describe the algorithm as it applies to residue-residue pair selection below.

Typically, a conformation is represented as a 3N dimensional vector of atomic positions where N is the number of atoms. For selecting DEER pairs, however, a more natural choice of coordinate system is the set of distances between all possible residue-residue pairs. If the protein has *n* residues, there are $(n^2-n)/2$ possible pairs, and we can define the following conformation variable:

$$\vec{C} = (X_1 \quad X_2 \quad \cdots \quad X_i \quad \cdots \quad X_{(n^2-n)/2}),$$

where X_i is the distance between the ith pair of residues.

In order to determine the most informative pairs, we calculate the mutual information (MI) between a pair X_i and the conformation variable C:

$$I_i(X_i, C) = \sum_{\{x_i\},\{c\}} P(x_i, c) \log \frac{P(x_i, c)}{P(x_i)P(c)},$$

where $P(x_i, c)$ is the joint probability function of pair *i* and configuration *c* and $P(x_i)$ and P(c) are the marginal probability distribution functions of pair *i* and configuration *c*, respectively.

The second criterion for pair selection is statistical independence. Because pairs that are each highly informative but highly correlated provide redundant data about conformation, MI alone is not the best

metric for pair selection. Therefore, we both maximize the pair-conformation MI and minimize the pair-pair MI [25-26]:

$$\max_{i} \frac{1}{N} I_i(X_i, C)$$

$$\min_{i,j} \frac{1}{N^2} I_{i,j}(X_i, X_j)$$

where N is the number of pairs to be selected and

$$I_{i,j}(X_i, X_j) = \sum_{\{x_i\}} \sum_{\{x_j\}} P_{X_i, X_j}(x_i, x_j) \log \frac{P_{X_i,X_j}(x_i, x_j)}{P_{X_i}(x_i) P_{X_j}(x_j)}.$$

As we show below, this yields a means to guide DEER experiments by identifying an optimal set of pairs from MD simulations in a model-free fashion.

Incorporation of DEER data into a model: restrained-ensemble simulations. The incorporation of experimental data into MD simulations as Hamiltonian biases has been studied extensively [9, 27-31]. Here we have chosen to include DEER data into MD simulations using a modified version of Roux and Islam's restrained-ensemble methodology in order to incorporate the multi-modal distributions typical of flexible systems. The original method is designed to capture side-chain fluctuations in small, rigid systems. We therefore adapted it to accommodate the backbone motion of the Opa loops by having each ensemble member contain an independent representation of the protein backbone as opposed to a single common representation as in the original Roux method [9]. Simulation details, including a discussion of the Roux method and modifications, can be found in the Methods.

We have not treated spin label side-chain conformations explicitly because in flexible Opa proteins the distance differences resulting from different side-chain rotameric states are small compared to backbone rearrangements of the Opa loops. For systems where incorporation of the spin labels is desired, our method can be simply extended to include them explicitly at a greater, but still feasible, computational cost[32].

Results

Selection and measurement of mRMR pairs. We assessed the success of mRMR-based refinement by computing optimal residue pairs to refine the conformational ensemble of Opa_{60}, measuring those pairs experimentally, and incorporating the results into hybrid refinement. In general, it would be straightforward to calculate optimal pairs from a set of unrestrained MD simulations using mRMR. In this case, because a rough conformational ensemble of unliganded Opa_{60} was already available based on NMR-restrained MD simulations [22], we analyzed this dataset of approximately 2 μs aggregate simulation time to obtain a ranked list of highly informative Opa_{60} pairs using the mRMR algorithm (Fig 2-Supplement 2).

The five highest-ranking pairs are shown on the Opa_{60} structure in Fig 2c and differ substantially from pairs identified using MI alone (Fig 2b, Fig 2-Supplement 2), those independently selected by spectroscopists using a typical triangulation strategy with at least one of the sites in the pair on a structured region of the protein[33] (referred to as spectroscopist-selected pairs; Fig 2d), and pairs identified using ENM-based flexibility analysis (Fig 2-Supplement 3). The top five scoring mRMR pairs span multiple biologically significant regions of the extracellular loops, namely the hypervariable and semivariable regions (Fig 2-Supplement 1), denoted HV1, HV2, and SV, and the top ten pairs capture all possible combinations of these regions. By contrast, the top ten pairs identified using MI alone all span a single loop-loop pair, likely losing important information about the dynamics of the other loop. The pairs identified with an ENM span only a single loop-loop pair and likely perform more poorly than the MI pairs because in this case the ENM poorly captures loop flexibility measured via NMR relaxation timescales (Fig 2-Supplement 3). The spectroscopist-selected pairs are primarily short barrel-loop distances. Although a naïve approach might be simply to select residues at the ends of the loops, the length of the loops permit distances too long to be measured via DEER, so spectroscopic best practice is to se-

lect a somewhat more conservative set of pairs. The mRMR and MI-based selection, however, which utilizes MD simulations, could identify loop-loop pairs that remain within 55 Å and are thus good targets for DEER measurement. However, even had a spectroscopist decided to measure loop-loop distances, the chance of selecting a pair within the top 25% most informative pairs identified via mRMR would be only 7%, showing a strong advantage for the systematic selection methods developed here.

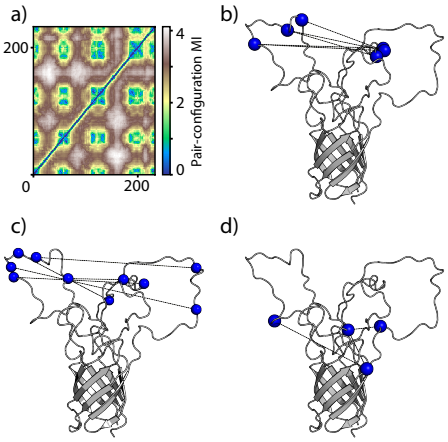

Figure 2: mRMR captures highly informative residue pairs that are minimally redundant. The mutual information between the residue pair distance and the configuration vector calculated over all Opa-Opa pairs is plotted in panel (a). Pairs selected via mutual information alone (b) are informative but redundant pairs. Pairs selected via mRMR (c) remove this redundancy. These pairs are better distributed across the Opa_{60} structure than the top-ranking pairs via MI alone or pairs selected by spectroscopists (d) according to current practice in the field. The Opa loop nomenclature is schematized in Figure 2 – Supplement 1. Residue rankings via mRMR are tabulated in Figure 2 – Supplement 2. An ENM-based scoring performs poorly for Opa loops and is shown and assessed in Figure 2 – Supplement 3. Further statistics on mRMR and MI values of measured pairs are given in Figure 2 – Supplement 4.

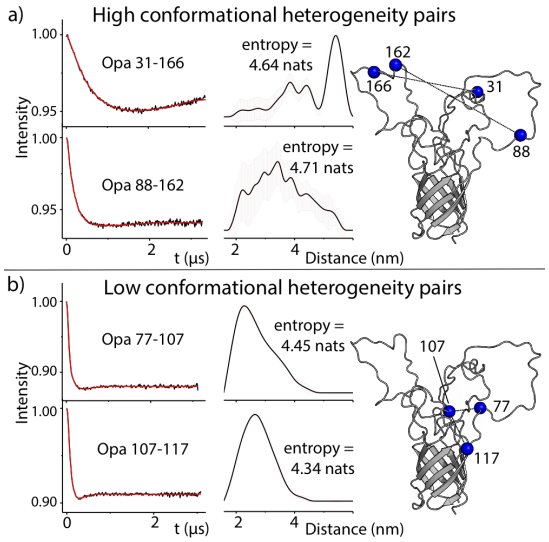

Figure 3: As expected, residue pairs identified by mRMR yield distributions with higher conformational heterogeneity. DEER-derived distributions of pairs identified by the mRMR algorithm, shown in (a), reveal a greater degree of conformational heterogeneity than pairs independently selected by DEER spectroscopists, shown in (b). Conformational heterogeneity is quantified as the entropy of each DEER-derived distribution. These high and low conformational heterogeneity pairs are represented on the Opa_{60} structure at right. Spin-echo decays and fits are given in Figure 3 – Supplement 1 and continuous-wave EPR spectra in Figure 3 – Supplement 2.

In order to quantitatively compare mRMR-selected pairs and spectroscopist-selected pairs, we utilized information-theoretic metrics first to assess the conformational heterogeneity of each pair and then to assess the utility of each set of pairs in refining the conformational ensemble. In all pairs tested, mRMR-guided selection identifies more loop-loop (rather than loop-barrel) distances that show greater conformational heterogeneity and lead to improved subsequent refinement of the conformational ensemble (Fig 3a). All measured DEER spin-echo decays and processed distributions are given in Figure 3-Supplement 1. Amongst the high-ranking mRMR pairs, the 88-162 distance distribution is extremely broad (entropy 4.71 nats), while the 31-166 distribution is more structured with two distinct modes (entropy 4.64 nats), indicating that the loops containing HV1 and HV2 may occupy distinct conformations with respect to each other. These differing distance distributions, although both high in entropy, immediately suggest structural hypotheses for loop-loop interaction as follows: the HV1 loop interacts non-specifically with the HV2 loop whereas the multiple distinct peaks observed between the SV loop and the HV2 loop indicate multiple distinct interactions. Hybrid refinement of these pairs using DEER measurements further refines these interaction modes (discussed below and in Fig. 4-Supplements 2-3). Of the spectroscopist-selected pairs, residues 77-107 and 107-117 have approximately Gaussian distributions with low entropy values of 4.45 and 4.34 nats, respectively. These distributions are also approximately centered on distances that were already observed with high probability in our NMR hybrid-refinement simulations or a microsecond-length extension of those simulations (Fig. 4-Supplement 1). These distributions therefore provide little additional information about the conformational ensemble beyond what was already known from previous work.

Incorporation and evaluation of mRMR vs. spectroscopist-selected pairs. We performed two sets of molecular dynamics simulations: one in which DEER data from two high-ranking mRMR pairs were incorporated and one in which DEER data from two spectroscopist-selected pairs were incorporated. Simulations were performed using the restrained-ensemble method with 20 ensemble members per set and 100 ns of simulation per ensemble member (see Methods). Over the first 50 ns of simulation, all restrained MD distances converged to closely approximate the experimentally measured distributions (Fig. 4) and remained there over the rest of the 100-ns simulation runs. Small fluctuations around the experimental target values are reflective of larger-scale conformational motions driven by the MD Hamiltonian. Both sets of simulations explore all of the experimentally observed distances, which is not the case for the unrestrained MD simulations (Fig. 4-Supplement 1), and thus indicate that the restraints drive sampling of multiple loop conformations supported by the DEER data but not well explored by unrestrained simulations. This demonstrates that the restrained-ensemble simulations successfully incorporate experimental data, however, what is novel and striking is that incorporating mRMR-selected data improves agreement with DEER distributions measured only subsequently and not used in the refinement (Fig 5).

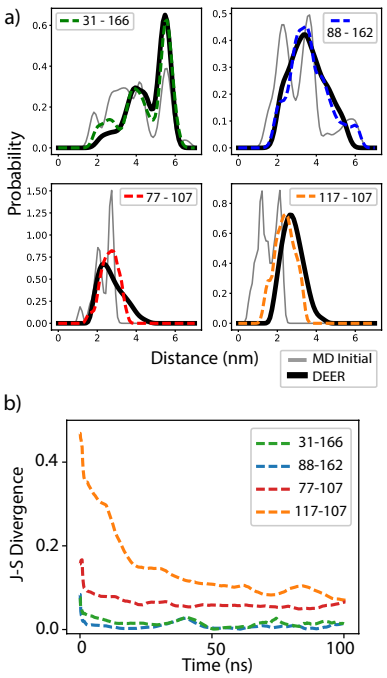

Figure 4: Restrained-ensemble simulations converge rapidly to experimental distributions. Convergence of restrained-ensemble simulations to DEER-derived distributions over 100 ns is plotted in (a) for both the high-scoring mRMR pairs and spectroscopist-selected pairs. Convergence of both ensembles is quantified in (b) using Jensen-Shannon divergence. Distance distributions are compared to the starting conformational ensembles in Figure 4 – Supplement 1. Comparison of loop-loop contact modes is given in Figure 4 – Supplement 2 for mRMR-selected pairs and Figure 4 – Supplement 3 for spectroscopist-selected pairs.

In order to further evaluate the mRMR method of pair selection, we analyzed the dimensionality of the conformational ensembles from hybrid refinement using different DEER datasets. Given sufficient sampling, a better-refined conformational ensemble will have lower dimensionality, approaching the "true" ensemble in the lower limit. We therefore developed a quantitative measure for the dimensionality of a conformational ensemble in residue-residue distance space (see Methods). Because residue-residue distances are not completely independent, we lumped together highly related distance variables at different thresholds of relatedness (ε) and at each threshold calculated the number of independent distance variables required to describe the ensemble. At every information-theoretic-based threshold tested, hybrid simulations incorporating DEER measurements from high-ranking mRMR pairs refine the conformational ensemble to lower dimensionality than simulations incorporating measurements from spectroscopist-selected pairs (Fig. 5). This indicates that DEER data from mRMR pairs more efficiently refine the conformational ensemble than data from pairs selected according to current state of the art.

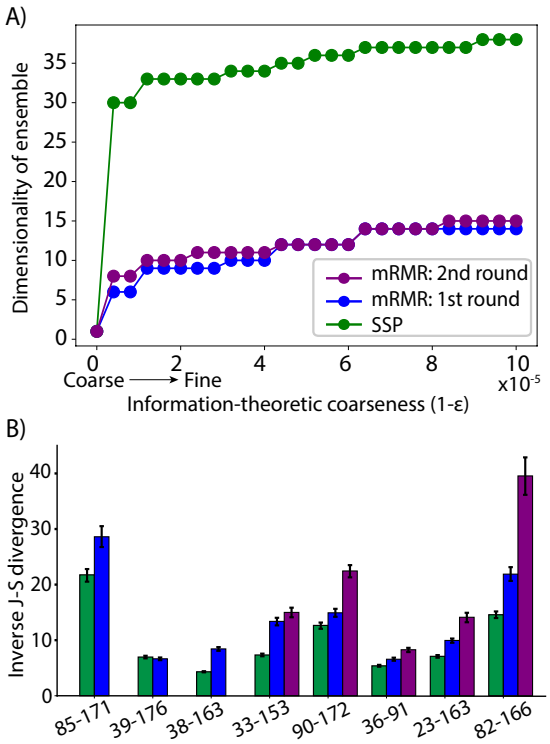

Figure 5: High mRMR pairs out-perform spectroscopist-selected pairs (SSP) in refining the Opa_{60} conformational ensemble. To assess convergence of the conformational ensemble, we calculated the dimensionality of the conformational ensemble (a) at any information theoretic resolution ε (see Methods for details). Ensembles refined using high-mRMR pairs are of lower dimensionality than ensembles refined using spectroscopist-selected pairs (labeled SSP) by 20-25. Each refined ensemble was also evaluated by its ability to predict 8 residue pairs identified via mRMR after the first round of refinement and measured using DEER (b). Conformational ensembles refined using mRMR-selected pairs predict these new DEER distributions significantly better than conformational ensembles refined using spectroscopist-selected pairs (SSP) in seven of eight cases, quantified as inverse J-S divergences. Three of these DEER pairs were used for a second round of mRMR refinement; the resulting conformational ensemble out-performs both 1st-round ensembles in predicting the five pairs not used for refinement. Error bars represent 90% confidence using 1000 bootstrap replicates.

mRMR-based refinement of the Opa_{60} conformational ensemble yields a preliminary hypothesis for CEACAM engagement. We obtained a preliminary sampling of the Opa_{60} conformational ensemble by incorporating five highly informative DEER-derived distributions of pairs identified by the mRMR algorithm into a second round of restrained-ensemble simulations. This model resolves previously unrefined loop-loop interaction modes and reveals potential sites for CEACAM engagement.

The refined conformational ensemble yields several sets of specific loop-loop interactions. To analyze these systematically, we identified the most abundant loop conformations in the structural ensemble by clustering PCA-transformed loop-loop contact maps (Fig 6-Supplement 1). Ten well-separated clusters were formed in the projected subspace. Strikingly, in the centroids of three of the four the most populated clusters (Fig 6), HV1 interacts with HV2 or SV1. In all conformations, HV2 does not interact with SV. HV1 and HV2 loops interact with approximately twice the likelihood than that of SV and HV1 loops. In each of these cases, the third loop is extended and does not interact with the other two loops. Analysis of contact maps (Fig 6-Supplement 1) supports these interaction patterns.

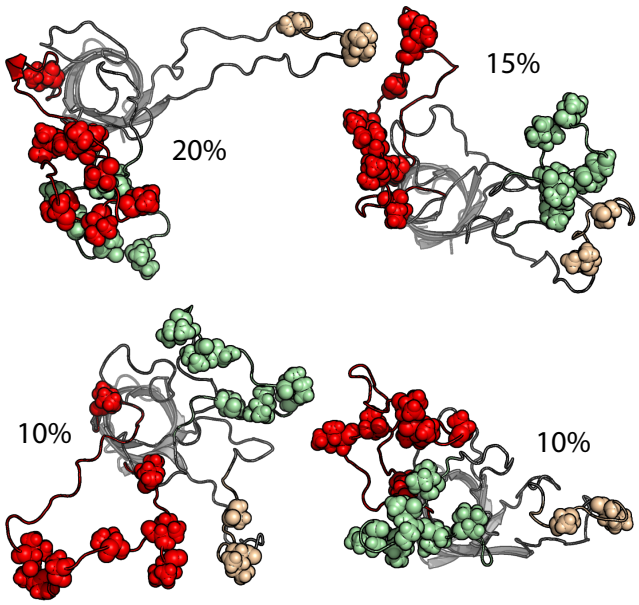

Figure 6. Preliminary refinement of Opa_{60} indicates that CEACAM engagement may occur via a single, laterally extended loop or through two contacting, vertically extended loops. Rendered are structures from four conformational clusters identified in restrained-ensemble simulations using mRMR-selected DEER restraints (see Methods for clustering details). Renderings show the barrel from above, as Opa_{60} would appear to its receptor CEACAM1, with hydrophobic residues from the HV1 loop in green, HV2 in red, and SV in tan. Clusters are labeled with the fraction of the overall conformational ensemble in that cluster. Three of the four structures have one loop laterally extended (further from the center of the barrel in the rendered view) and the two interacting loops more vertically positioned (above the barrel in the rendered views). Loop-loop contact modes are shown in Figure 6 – Supplement 1.

Opa_{60} conformations recognized by CEACAM. HV1/HV2 chimeric Opa proteins have previously shown that specific HV1/HV2 combinations are required for CEACAM engagement[34], leading to a model in which HV1 and HV2 together directly engage receptors. Our results on Opa conformations in the absence of CEACAM are compatible with this but also yield additional structural models for CEACAM binding. Two possibilities for CEACAM engagement exist: it could bind to one of the extended loops (SV or HV2), or it could bind to the combined surface formed by two interacting two loops (HV1/SV or HV1/HV2). Each of these possibilities are consistent with the structural data. The Opa extracellular loops have a surprisingly high number of hydrophobic residues for flexible sequences, which likely mediate loop-loop interactions and CEACAM engagement. Prior identification of two hydrophobic residues on CEACAM essential for Opa binding further supports engagement mediated by hydrophobic residues[34]. These new hypotheses generated by the experimentally-derived conformational ensemble can now be further tested with carefully designed binding experiments.

To determine whether CEACAM binds the SV-extended Opa conformation or the HV2-extended Opa conformation, we performed additional DEER experiments on labeled Opa proteins with and without bound CEACAM. The resulting data show substantial shifts in Opa loop-loop distance distributions upon CEACAM binding, consistent with conformational selection: a subset of the distances present in the *apo* protein increase, while others decrease (Fig 7a). We analyzed this quantitatively by fitting the CEACAM-bound ensemble as a linear combination of conformational states identified in the *apo* ensemble. The results were striking: HV2-extended conformations dramatically increased to account for 75% of the bound ensemble, while SV2 and splayed-loop conformations decreased to account for 0% and 25% of the bound ensemble, respectively. Visual analysis of the structures supports this finding because the long HV1-HV2 distances unambiguously exclude an HV1-HV2 interface and thus the SV-extended conformations. Indeed, the HV2-extended conformations show robust agreement with the

increase in probability density at the 45-Å and 51-Å peaks in the CEACAM-bound distributions (Fig 7b). By comparison, if we had measured the spectroscopist-selected pairs in the *apo* and CEACAM-bound forms, we would not have been able to differentiate SV-extended and HV2-extended conformations, since the distance distributions overlap in one pair (107-117) and the HV2-extended distances too short to measure via DEER in the other (77-107).

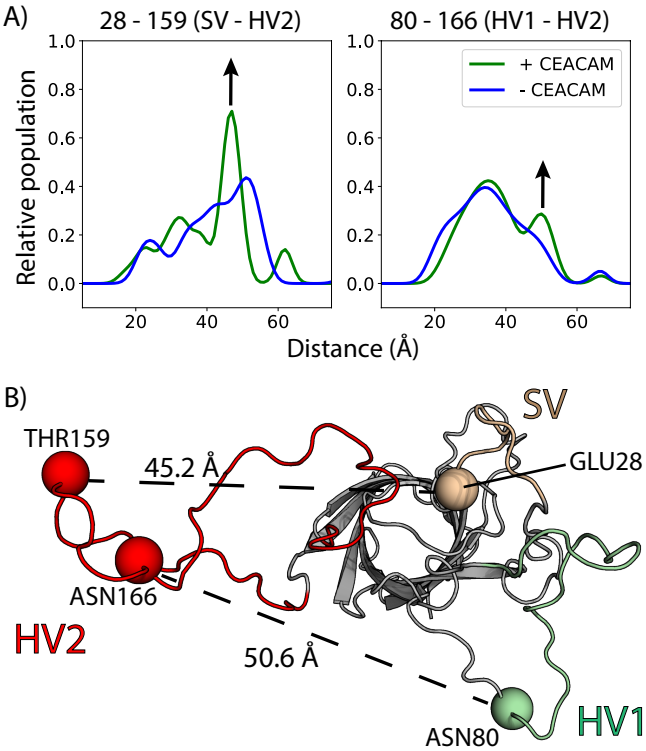

Figure 7. Binding of CEACAM selects for HV2-extended Opa conformations. DEER measurements of Opa-Opa residue pair distances show substantial shifts in the distance distributions upon binding to CEACAM (a), suggestive of conformational selection. Using these data to interpret the *apo* conformational ensemble, we find that HV2-extended conformations (Fig. 6) dramatically increased (accounting for 75% of the bound ensemble) while SV2 and splayed-loop populations decreased (0% and 25%, respectively). Rendering of an HV2-extended Opa conformation (b) makes the basis for this clear: the 45-Å distance between residues 28 and 159 and the 51-Å distance between 80 and 166 correspond to the major peaks in the DEER distributions collected for the Opa-CEACAM complex. Spin-echo decays and fits are given in Figure 7 – Supplement 1.

These results, obtained after two rounds of mRMR-guided pair selection, would likely not have been obtained using current state-of-the-art pair selection methods. Indeed, mRMR-guided pair selection produces strikingly more informative results than spectroscopist-guided pair selection after just a *single* round of refinement. Analysis of the loop-loop contacts in first round of mRMR-guided refinement yields four structures, all of which show one loop protruding laterally (Fig. 4-Supplement 2); the second round of mRMR-guided refinement better resolves the conformational heterogeneity among these extended-loop structural motifs. In contrast to the mRMR-guided refinement, refinement using comparator pairs predicted conformations with compact, closely interacting loops to have higher probability than those with splayed loops or a single laterally extended loop (Fig. 4-Supplement 3). Since these conformational ensembles poorly predict the additional DEER distance measurements (Fig. 5), they are incorrect and would have required further DEER measurements to yield similar hypotheses for the determinants of Opa-CEACAM binding.

Conclusions

We have demonstrated a method for selecting label locations to obtain distance distributions that optimally refine the conformational ensembles of flexible proteins. Our method was developed and tested using DEER spectroscopy but applies equally well to other methods that can provide distance distributions between pairs of labels. The method was prospectively validated for refinement of Opa_{60} Neisserial virulence-associated protein, and we are actively working to use it to refine bimolecular complexes in flexible molecular recognition such as occurs in the binding of CEACAM receptors by Opa_{60}.

In the case of Opa60 engagement of CEACAM, we identified a set of loop conformations that account for the *apo* conformational ensemble. Further DEER measurements indicated a conformational selection event upon CEACAM binding, and we have shown that HV2-extended Opa conformations are the only ones consistent with the CEACAM-engaged complex. Previous mutational data showed that specific HV1/HV2 loop sequence combinations were required for CEACAM engagement. Our data suggest a new interpretation of these findings: we speculate that since HV1 and HV2 contact each other in a much higher proportion of unbound conformations than bound conformations, certain HV1/HV2 sequence combinations could overstablize unproductive conformations and thus interfere with binding. Neither this structural hypothesis nor the underlying identification of the Opa conformations recognized by CEACAM would have been possible without the use of the mRMR method.

High-resolution refinement of flexible proteins is anticipated to require several rounds of the procedure we have described. Indeed, one of the advantages of this procedure is that it can be initiated using undersampled MD trajectories in early rounds of refinement rather than requiring a well-converged computational estimate to begin. At each iteration, the mRMR algorithm identifies under-determined regions of the free energy landscape and specifically selects pairs that improve the hybrid estimator of the ensemble. In later rounds of the procedure, the MD trajectories will converge to the experimentally-determined ensemble. Convergence is established when no additional refinement is required after mRMR prediction and measurement of a set of residue-residue pairs provides no new information compared to the current estimate of the conformational ensemble.

Our results demonstrate that current state-of-the art techniques for selecting spin-label sites for spectroscopic experiments are suboptimal and can be improved with our iterative methodology. The DEER-derived distributions of mRMR-selected pairs reveal critical information about conformational heterogeneity of flexible proteins and, when incorporated into simulations, are more efficiently matched by the MD ensemble. Finally, incorporation of the distributions of mRMR pairs leads to improved refinement of the conformational ensemble by reducing the effective dimensionality of the ensemble. Ultimately, the same information can be obtained about an ensemble with significantly fewer spectroscopic measurements.

Methods

Selection of residue pairs. Residue-residue pairs for DEER experiments were selected for the first round of refinement using mRMR on a set of 20x200ns Opa_{60} MD simulations we have previously reported [22] as an initial estimate of the conformational ensemble. The second-round pairs were selected using the mRMR-restrained ensemble generated in the first round of refinement. In both rounds, a matrix of all $C\alpha$- $C\alpha$ distances was calculated for each conformational snapshot at 500-ps intervals, and histogram distributions were computed using 1-Å bins. These histograms were normalized and used to calculate pair-configuration and pair-pair mutual information as given in the Theory section. An ordered list of highest-ranking mRMR Opa-Opa residue pairs was then generated using greedy mRMR selection [25]. Two sets of comparator pairs were also generated: a ranked list of Opa_{60} residue-residue pairs based on pair-configuration MI alone and a set of top pairs independently selected by a DEER spectroscopist in a blind fashion. Code for mRMR selection of residues for DEER experiments is available from https://github.com/kassonlab/mRMR-DEER.

Expression, purification, spin-labeling, and refolding of Opa_{60} proteins. Opa_{60} expression, purification, and refolding were performed as described previously [22, 24, 35-36]. The *opa60* gene was sub-cloned into a pET28b vector (EMD chemicals, Gibbstown, NJ) containing N and C terminal His_6 – tags. Cysteine residues were introduced at regions of interest on Opa using PIPE Mutagenesis, and gene sequencing

confirmed the mutations (Genewiz Inc., South Plainfield, NJ). The pET28b vectors containing a mutated *opa60* gene were transformed into BL21(DE3) *E. coli* cells, and cultures were grown in Luria-Burtani (LB) media. Opa protein expression to inclusion bodies was induced with 1 mM isopropyl-β-thio-D-galactoside (IPTG). Cells were harvested and resuspended in lysis buffer [50 mM Tris-HCl, pH 8.0, 150 mM NaCl, and 1 mM TCEP-HCl (tris(2-carboxyethyl)phosphine hydrochloride)]. Following cell lysis, insoluble fractions were pelleted and resuspended overnight with lysis buffer containing 8 M urea. Cell debris was removed via centrifugation and unfolded Opa proteins in the soluble fraction were purified using Co^{2+} immobilized metal affinity chromatography, eluting in 20 mM sodium phosphate, pH 7.0, 150 mM NaCl, 680 mM imidazole, 8 M urea, and 1 mM TCEP. Purified Opa proteins were loaded on a PD-10 column (GE Healthcare Biosciences, Pittsburg, PA) to remove TCEP. Opa proteins were eluted with buffer (20 mM sodium phosphate, pH 7.0, 150 mM NaCl, and 8M urea) directly into five molar excess MTSL/R1 spin label [S-(2, 2, 5, 5-tetramethyl-2,5-dihydro-1H-pyrrol-3-yl)methyl methanesulfonothioate, Toronto Research Chemicals Inc., Toronto, Canada, stored as 100 mM stock in acetonitrile] for proteins containing a single cysteine and ten molar excess MTSL for Opa proteins with two cysteine residues. The proteins were spin labeled overnight at room temperature. Excess spin label was removed using a second PD-10 column, and the eluted protein was concentrated to approximately 150 - 200 μM. The labeled proteins were rapidly diluted 20-fold into 20 mM Tris-HCl, pH 8.0, 500 mM NaCl, 3 M urea, and 4.6 mM n-dodecylphosphocholine (FC-12, Anatrace), upon which Opa proteins fold into the detergent micelles over the course of three days at room temperature [22, 36]. Folding efficiency was assessed with SDS-PAGE. Samples were dialyzed against 3 x 4L of 20 mM sodium phosphate, 150 mM NaCl for an hour each, removing any free spin. Opa proteins were concentrated to approximately 200 – 400 μM.

Continuous wave EPR to evaluate Opa_{60} spin labeling. Opa_{60} spin labeling was assessed using an X-band Bruker EMX continuous wave spectrometer with an ER4123D dielectric resonator (Bruker Biospin, Billerica, MA) at room temperature (Fig. 3-Supplement 2). Five μL of each sample (100 – 300 μM) was measured in Pyrex capillaries (0.6 mm id x 0.84 od, Vitrocom, Mountain Lakes, NJ). Spectra were baseline corrected and normalized using Lab-VIEW software generously provided by C. Altenbach (University of California at Los Angeles).

Double-electron electron resonance spectroscopy of Opa_{60} micelles. Double-labeled Opa_{60} proteins in detergent micelles were measured using pulsed EPR with a Q-band Bruker E580 Spectrometer fitted with an ER5106-QT Q-band Flexline Resonator (Bruker Biospin) at 80 K. All samples were prepared to a final protein concentration between approximately 100 and 200 μM with 10% deuterated glycerol. The samples were loaded into quartz capillaries with a 1.6 mm od x 1.1 mm id (Vitrocom) and flash frozen in liquid nitrogen. A four pulse DEER sequence was used with one 16 ns $\pi/2$, two 32 ns π observed pulses (at an observed frequency υ_1), and a π pump pulse (at a frequency υ_2) optimized at approximately 32 ns [37]. The pump frequency (υ_2) was set at the maximum of the nitroxide spectrum and the observed frequency (υ_1) is set to 75 MHz lower. Increasing inter-pulse delays at 16 ns increments were used with a 16-step phase cycle during data collection. Accumulation times were typically between 18 and 24 hours, with a dipolar evolution time between 2 and 3 μs. Dipolar evolution data were processed using DEERAnalysis2016 software using Tikhonov regularization to generate distance distributions [38]. Background subtraction of the distance distribution yielded error at each distance which was plotted as ranges representing fits that are within 15% root-mean-square-deviation of the best fit.

Molecular dynamics simulations. Simulations were performed using a modified version of Gromacs 5.2 available at https://github.com/kassonlab/reMD-gromacs-5.2 and the CHARMM36 forcefield[39-40]. Simulations were run under NPT conditions using the velocity-rescaling thermostat [41] at 310 K with a 2-ps time constant and pressure maintained at 1 bar using the Parrinello-Rahman barostat with a 10-ps time constant. Covalent bonds were constrained using LINCS, and long-range electrostatics were treated using Particle Mesh Ewald[42]. System setup and equilibration details are given below.

Incorporation of experimental distributions into MD simulations. In order to compare the performance of mRMR-guided and spectroscopist-guided refinement, two sets of ensemble simulations were run, each using the equilibrated 20 structures as initial states. The first set incorporated the DEER distance distributions of the high-ranking mRMR pairs 31-166 and 88-162—for operational reasons, the top mRMR pairs were shifted to these nearby pairs that have similar MI and mRMR values (Fig. 2-Supplement 4)—while the second set incorporated the spectroscopist-selected pairs 77-107 and 107-117. A subsequent set of ensemble simulations were run by resampling snapshots from the mRMR-guided ensemble and incorporating a second round of DEER-derived distributions. Restrained-ensemble biasing potentials were applied to match MD distance histograms to DEER-derived distance distributions as previously developed by Roux [9] with modifications detailed below. Simulated distance distributions converged to match the experimental distributions by 50 ns as measured by Jensen-Shannon divergence. Scripts for running ensemble simulations can be found at https://github.com/kassonlab/RestrainedEnsemble.

Setup and equilibration of MD simulations. The 20 lowest free-energy structures of Opa_{60} previously identified [22] (PDB ID 2MAF) were selected as initial states. Each Opa_{60} molecule was inserted into a membrane of 494 DMPC molecules as follows: the beta-barrel was aligned to previously embedded beta-barrel of a single structure from the Fox simulations [22]. The protein and membrane were energy-minimized using the steepest-descent integrator for either 5000 steps or until the largest force was less than 1000 kJ/mol/nm^2, whichever occurred first. Each system was solvated independently with approximately 300,000 TIP3P water molecules and ions were added to obtain a system with 150 mM NaCl and no net charge. The final systems were independently energy-minimized again using steepest-descent for 5000 steps or until the largest force was less than 1000 kJ/mol/nm^2. Finally, a brief 5000 step equilibration was run using the NPT conditions described in the main text Methods.

Initial states for the second iteration of mRMR were obtained by resampling the mRMR-restrained ensemble simulations according to the joint distribution of the underlying DEER distributions (the individual distributions were assumed to be independent). The solvation, energy minimization, and initial equilibration protocols were identical to those of the ensembles described above.

Restrained-ensemble biasing potentials. Both DEER-derived and MD-derived distance distributions were smoothed with a Gaussian filter. The smoothing parameter σ was chosen to reflect the experimental uncertainty in the fine modes of the DEER-derived distance distributions, 2 Å for the high-scoring mRMR pairs and 1 Å for the spectroscopist-selected pairs. Histograms were calculated using 1-Å bins.

Rather than updating the bias potential U_{bias} at every MD step, distance data were collected for all ensemble members for a period of 100 ps followed by a U_{bias} update. Additionally, a boxcar averaging filter was applied so that the simulation distance distributions were calculated using the last 10 ns of data for the first round of simulations and 25 ns for the second round of simulations. These modifications were implemented in order to obtain sufficient sampling for generating the MD distance distributions. Final distance distributions were calculated using the last 25 ns of data, while convergence monitoring using the Jensen-Shannon divergence was performed on a 10-ns window prior to the referenced time point. An initial spring constant K=10 kJ/mol/nm^2 was used for the first 40 ns in all three sets of simulations. After 40 ns, K was increased to 100 kJ/mol/nm^2 in the mRMR-guided simulations in order to reverse the increase in J-S divergence observed from approximately 30-40ns.

Information-theoretic clustering. The final trajectories of both the mRMR-restrained and control-restrained ensembles were sampled at 0.5 ns intervals, and all Cα- Cα distances were calculated using the Gromacs toolset. Histograms of each Cα- Cα pair were constructed using 1-Å bins, and all pairwise mutual information values were calculated as:

$$I\left(X_1^{C\alpha}, X_2^{C\alpha}\right) = \sum_{\{x_1^{C\alpha}\},\{x_2^{C\alpha}\}} P\left(x_1^{C\alpha}, x_2^{C\alpha}\right) \ln \frac{P\left(x_1^{C\alpha}, x_2^{C\alpha}\right)}{P\left(x_1^{C\alpha}\right) P\left(x_2^{C\alpha}\right)}.$$

Because closely related sets of pairs (high $I(X^{C_\alpha}{}_1, X^{C_\alpha}{}_2)$) contain redundant information, it is possible to obtain an approximation of the Opa_{60} ensemble by knowing the distributions of only a *subset* of all Cα-Cα distances; that is, by grouping together sets of highly related pairs, one can obtain an approximation of the dimensionality ensemble. The quality of the approximation depends on how much information is lost by grouping together more and more diverse Cα- Cα pairs.

In order to quantitatively evaluate the dimensionality of the ensemble after incorporation of the mRMR or spectroscopist-selected pairs, we clustered closely related sets of Cα- Cα pairs using complete-linkage hierarchical clustering with an MI-based distance metric $D(X^{C_\alpha}{}_1, X^{C_\alpha}{}_2) = 1 - I(X^{C_\alpha}{}_1, X^{C_\alpha}{}_2)/H(X^{C_\alpha}{}_1, X^{C_\alpha}{}_2)$, where $H(X^{C_\alpha}{}_1, X^{C_\alpha}{}_2)$ is the joint entropy of the pairwise Cα- Cα distance distributions. The maximum cluster diameter after each clustering step may be thought of as a measure of resolution, or quality of the approximation; as the cluster diameter increases, information about the ensemble is lost as increasingly more independent Cα- Cα pairs are grouped together and considered redundant.

The information-theoretic resolution is reported in Fig. 5 as δ, or 1 – max(cluster diameter).

Analysis of loop conformations. Contact matrices were calculated for all inter-loop contacts in snapshots taken at 500-ps intervals using a distance cutoff of 6 Å. Principal components analysis was performed to obtain a new orthogonal basis set for loop-loop contacts; for restrained-ensemble simulations performed using mRMR-guided DEER data, all snapshots formed compact and well-separated clusters in the subspace formed by the first three principal components. These clusters and their corresponding centroids thus reflect the major contact modes between loops. This contact-matrix-based analysis was chosen because the loops are highly flexible, making the rigid-body alignment that underlies RMSD-based clustering less accurate.

Expression and purification of glycosylated N-terminal domain CEACAM1 proteins. An expression and purification protocol for glycosylated N-CEACAM1 proteins was adapted from previously published work[43]. A 250 mL suspension culture of HEK293S cells was transfected with the NCEACAM1-pGEn2 plasmid using polyethyleneimine (linear 25 kDa, Polysciences Inc., Warrington PA) as described previously[43], where NCEACAM1 is the human *ceacam1* gene encoding the N-terminal domain of CEACAM1 (residues 34-141). Cysteine residues were introduced into *ceacam1* via site-directed mutagenesis and confirmed by sequencing. Glycosylated NCEACAM was produced over five days at 37°C, after which cell debris was removed via centrifugation (20 min, 150 x g, 4°C).

Glycosylated NCEACAM1 was purified from the supernatant via Co^{2+} immobilized metal affinity chromatography (IMAC), eluting in ten column volumes of elution buffer (25 mM HEPES, 300 mM NaCl, 680 mM imidazole, pH 7.0) at 4°C. The eluent was dialyzed into 4L of 25 mM HEPES, 300 mM NaCl, 10% glycerol, pH 7.0 containing approximately 3.5 μM tobacco etch virus (TEV) protease and endoglycosidase F1 (EndoF1). Excess GFP was removed using Co^{2+} IMAC, and the flow-through containing CEACAM was collected. NCEACAM1 was further purified from GFP, TEV, and EndoF1 using a HR Sephacryl S-200 Gel Filtration column (GE Healthcare) equilibrated with 20 mM HEPES, pH 7.0, 150 mM NaCl, and 10% glycerol. Opa and CEACAM samples were concentrated to approximately 200 μM and mixed at a 2:1 CEACAM:Opa molar ratio. Samples were incubated for 30 minutes with gentle nutation at room temperature prior to adding 10% deuterated glycerol and flash freezing.

Analysis of Opa conformations selected for by CEACAM. Opa residue-residue distance distributions measured in the presence of CEACAM were fitted as a linear combination of the SV-extended, HV2-extended, and splayed-loop ensembles resulting from final analysis of the *apo* Opa conformational ensemble. The distributions of residue-residue pairs 28-159 and 80-166 from restrained-ensemble simulations were calculated for the SV-extended, HV2-extended, and splay-loop ensembles by sampling at 0.5 ns intervals and smoothing the resulting distributions via a Gaussian filter with bin size 1 Å and σ 2 Å. The experimental distributions were similarly smoothed. The experimental distributions were fit as a linear combination of the three ensembles using a least-squares optimization procedure:

$$\min\left[\left(\alpha P_{SV}(x) + \beta P_{HV2}(x) + \gamma P_{splay}(x) - P_{DEER}(x)\right)^2 \ \forall\, x\right],$$

subject to the constraints

$$\alpha, \beta, \gamma \geq 0$$
$$\alpha + \beta + \gamma = 1.$$

The best approximation for the bound conformational ensemble is therefore the set of conformations defined by re-weighting the *apo* ensembles by the parameters α, β, γ.

Acknowledgements

The authors thank G. Cortina for many helpful discussions. Computational resources were provided by XSEDE Comet at San Diego Supercomputing Center through allocation MCB150128. Additional resources were provided by SNIC and the Center for Parallel Computation at KTH.

Competing Interests

The authors declare that they have no competing interests.

References

1. Boehr, D. D.; Nussinov, R.; Wright, P. E., The role of dynamic conformational ensembles in biomolecular recognition. *Nat Chem Biol* **2009,** *5* (11), 789-96.
2. Jimenez, R.; Salazar, G.; Yin, J.; Joo, T.; Romesberg, F. E., Protein dynamics and the immunological evolution of molecular recognition. *Proc Natl Acad Sci U S A* **2004,** *101* (11), 3803-8.
3. Willcox, B. E.; Gao, G. F.; Wyer, J. R.; Ladbury, J. E.; Bell, J. I.; Jakobsen, B. K.; van der Merwe, P. A., TCR binding to peptide-MHC stabilizes a flexible recognition interface. *Immunity* **1999,** *10* (3), 357-65.
4. Bernado, P.; Mylonas, E.; Petoukhov, M. V.; Blackledge, M.; Svergun, D. I., Structural characterization of flexible proteins using small-angle X-ray scattering. *J Am Chem Soc* **2007,** *129* (17), 5656-64.
5. Wei, G.; Xi, W.; Nussinov, R.; Ma, B., Protein Ensembles: How Does Nature Harness Thermodynamic Fluctuations for Life? The Diverse Functional Roles of Conformational Ensembles in the Cell. *Chem Rev* **2016,** *116* (11), 6516-51.
6. Levin, E. J.; Kondrashov, D. A.; Wesenberg, G. E.; Phillips, G. N., Jr., Ensemble refinement of protein crystal structures: validation and application. *Structure* **2007,** *15* (9), 1040-52.
7. Bonvin, A. M.; Brunger, A. T., Conformational variability of solution nuclear magnetic resonance structures. *J Mol Biol* **1995,** *250* (1), 80-93.
8. Islam, S. M.; Stein, R. A.; McHaourab, H. S.; Roux, B., Structural refinement from restrained-ensemble simulations based on EPR/DEER data: application to T4 lysozyme. *J Phys Chem B* **2013,** *117* (17), 4740-54.
9. Roux, B.; Islam, S. M., Restrained-ensemble molecular dynamics simulations based on distance histograms from double electron-electron resonance spectroscopy. *J Phys Chem B* **2013,** *117* (17), 4733-9.
10. Jao, C. C.; Hegde, B. G.; Chen, J.; Haworth, I. S.; Langen, R., Structure of membrane-bound alpha-synuclein from site-directed spin labeling and computational refinement. *Proc Natl Acad Sci U S A* **2008,** *105* (50), 19666-71.
11. Lindert, S.; Meiler, J.; McCammon, J. A., Iterative Molecular Dynamics-Rosetta Protein Structure Refinement Protocol to Improve Model Quality. *J Chem Theory Comput* **2013,** *9* (8), 3843-3847.
12. Deplazes, E.; Louhivuori, M.; Jayatilaka, D.; Marrink, S. J.; Corry, B., Structural investigation of MscL gating using experimental data and coarse grained MD simulations. *PLoS Comput Biol* **2012,** *8* (9), e1002683.
13. Boura, E.; Rozycki, B.; Herrick, D. Z.; Chung, H. S.; Vecer, J.; Eaton, W. A.; Cafiso, D. S.; Hummer, G.; Hurley, J. H., Solution structure of the ESCRT-I complex by small-angle X-ray scattering, EPR, and FRET spectroscopy. *Proc Natl Acad Sci U S A* **2011,** *108* (23), 9437-42.
14. Hirst, S. J.; Alexander, N.; McHaourab, H. S.; Meiler, J., RosettaEPR: an integrated tool for protein structure determination from sparse EPR data. *J Struct Biol* **2011,** *173* (3), 506-14.
15. Hanson, S. M.; Dawson, E. S.; Francis, D. J.; Van Eps, N.; Klug, C. S.; Hubbell, W. L.; Meiler, J.; Gurevich, V. V., A model for the solution structure of the rod arrestin tetramer. *Structure* **2008,** *16* (6), 924-34.
16. Rao, J. N.; Jao, C. C.; Hegde, B. G.; Langen, R.; Ulmer, T. S., A combinatorial NMR and EPR approach for evaluating the structural ensemble of partially folded proteins. *J Am Chem Soc* **2010,** *132* (25), 8657-68.
17. Ward, R.; Zoltner, M.; Beer, L.; El Mkami, H.; Henderson, I. R.; Palmer, T.; Norman, D. G., The orientation of a tandem POTRA domain pair, of the beta-barrel assembly protein BamA, determined by PELDOR spectroscopy. *Structure* **2009,** *17* (9), 1187-94.
18. Jeschke, G., MMM: A toolbox for integrative structure modeling. *Protein Sci* **2017**.

19. Jeschke, G., Ensemble models of proteins and protein domains based on distance distribution restraints. *Proteins* **2016,** *84* (4), 544-60.
20. Mittal, S.; Shukla, D., Predicting Optimal DEER Label Positions to Study Protein Conformational Heterogeneity. *J Phys Chem B* **2017**.
21. McCaw, S. E.; Liao, E. H.; Gray-Owen, S. D., Engulfment of Neisseria gonorrhoeae: revealing distinct processes of bacterial entry by individual carcinoembryonic antigen-related cellular adhesion molecule family receptors. *Infect Immun* **2004,** *72* (5), 2742-52.
22. Fox, D. A.; Larsson, P.; Lo, R. H.; Kroncke, B. M.; Kasson, P. M.; Columbus, L., Structure of the Neisserial outer membrane protein Opa(6)(0): loop flexibility essential to receptor recognition and bacterial engulfment. *J Am Chem Soc* **2014,** *136* (28), 9938-46.
23. Sadarangani, M.; Pollard, A. J.; Gray-Owen, S. D., Opa proteins and CEACAMs: pathways of immune engagement for pathogenic Neisseria. *FEMS Microbiol Rev* **2011,** *35* (3), 498-514.
24. Martin, J. N.; Ball, L. M.; Solomon, T. L.; Dewald, A. H.; Criss, A. K.; Columbus, L., Neisserial Opa Protein-CEACAM Interactions: Competition for Receptors as a Means of Bacterial Invasion and Pathogenesis. *Biochemistry* **2016,** *55* (31), 4286-94.
25. Peng, H.; Long, F.; Ding, C., Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans Pattern Anal Mach Intell* **2005,** *27* (8), 1226-38.
26. Ding, C.; Peng, H., Minimum redundancy feature selection from microarray gene expression data. *J Bioinform Comput Biol* **2005,** *3* (2), 185-205.
27. Roux, B.; Weare, J., On the statistical equivalence of restrained-ensemble simulations with the maximum entropy method. *J Chem Phys* **2013,** *138* (8), 084107.
28. White, A. D.; Voth, G. A., Efficient and Minimal Method to Bias Molecular Simulations with Experimental Data. *J Chem Theory Comput* **2014,** *10* (8), 3023-30.
29. Bruenger, A. T.; Karplus, M., Molecular dynamics simulations with experimental restraints. *Accounts of chemical research* **1991,** *24* (2), 54-61.
30. Boomsma, W.; Ferkinghoff-Borg, J.; Lindorff-Larsen, K., Combining experiments and simulations using the maximum entropy principle. *PLoS Comput Biol* **2014,** *10* (2), e1003406.
31. Pitera, J. W.; Chodera, J. D., On the Use of Experimental Observations to Bias Simulated Ensembles. *J Chem Theory Comput* **2012,** *8* (10), 3445-51.
32. Islam, S. M.; Roux, B., Simulating the distance distribution between spin-labels attached to proteins. *J Phys Chem B* **2015,** *119* (10), 3901-11.
33. Altenbach, C.; Kusnetzow, A. K.; Ernst, O. P.; Hofmann, K. P.; Hubbell, W. L., High-resolution distance mapping in rhodopsin reveals the pattern of helix movement due to activation. *Proc Natl Acad Sci U S A* **2008,** *105* (21), 7439-44.
34. Virji, M.; Evans, D.; Hadfield, A.; Grunert, F.; Teixeira, A. M.; Watt, S. M., Critical determinants of host receptor targeting by Neisseria meningitidis and Neisseria gonorrhoeae: identification of Opa adhesiotopes on the N-domain of CD66 molecules. *Mol Microbiol* **1999,** *34* (3), 538-51.
35. Dewald, A. H.; Hodges, J. C.; Columbus, L., Physical Determinants of beta-Barrel Membrane Protein Folding in Lipid Vesicles. *Biophys. J.* **2011,** *100* (9), 2131-2140.
36. Fox, D. A.; Columbus, L., Solution NMR resonance assignment strategies for -barrel membrane proteins. *Protein Sci.* **2013,** *22* (8), 1133-1140.
37. Pannier, M.; Veit, S.; Godt, A.; Jeschke, G.; Spiess, H. W., Dead-time free measurement of dipole-dipole interactions between electron spins. *J. Magn. Reson.* **2000,** *142* (2), 331-340.
38. Jeschke, G.; Chechik, V.; Ionita, P.; Godt, A.; Zimmermann, H.; Banham, J.; Timmel, C. R.; Hilger, D.; Jung, H., DeerAnalysis2006 - a comprehensive software package for analyzing pulsed ELDOR data. *Appl. Magn. Reson.* **2006,** *30* (3-4), 473-498.
39. Pronk, S.; Pall, S.; Schulz, R.; Larsson, P.; Bjelkmar, P.; Apostolov, R.; Shirts, M. R.; Smith, J. C.; Kasson, P. M.; van der Spoel, D.; Hess, B.; Lindahl, E., GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit. *Bioinformatics* **2013,** *29* (7), 845-54.
40. Huang, J.; MacKerell, A. D., Jr., CHARMM36 all-atom additive protein force field: validation based on comparison to NMR data. *J Comput Chem* **2013,** *34* (25), 2135-45.
41. Bussi, G.; Donadio, D.; Parrinello, M., Canonical sampling through velocity rescaling. *J Chem Phys* **2007,** *126* (1), 014101.
42. Darden, T.; York, D.; Pedersen, L., Particle Mesh Ewald - an N.Log(N) Method for Ewald Sums in Large Systems. *Journal of Chemical Physics* **1993,** *98* (12), 10089-10092.

43. Zhuo, Y.; Yang, J. Y.; Moremen, K. W.; Prestegard, J. H., Glycosylation Alters Dimerization Properties of a Cell-surface Signaling Protein, Carcinoembryonic Antigen-related Cell Adhesion Molecule 1 (CEACAM1). *J Biol Chem* **2016,** *291* (38), 20085-95.

Figure Supplement Legends.

Figure 2 – Supplement 1. Loop nomenclature and variable regions of Opa_{60}. Loops 1, 2, and 3 are shown in red, light green, and teal, respectively, on an Opa_{60} structure. Alpha carbons of the hypervariable regions (HV1 and HV2) as well as the semivariable region (SV1) are rendered as spheres.

Figure 2 – Supplement 2. Ranking of top residue-residue pairs via mRMR and mutual information alone.

Figure 2 – Supplement 3: ENM-based scoring of flexibility correlates poorly with NMR data and identifies less informative loop regions. Elastic network models provide a computationally efficient means of approximating some protein motions. To assess this approach for Opa loop prediction and DEER pair selection, a Gaussian Network Model was used to predict $C\alpha$ B-factors for Opa_{60}. The ten top-scoring residues are shown on the structure in (A). Many of the residues are located near the base of a single loop, while only two are located on a different loop in a more flexible region. Additionally, the ENM does not accurately reproduce the relative loop residue motion observed via NMR. The ENM-predicted B-factors correlate poorly with experimentally determined T_1 decays[1] (B); $r < 0.2$. The "high flexibility" residues identified by the ENM ends up closely resembling standard spectroscopist-guided pair-selection, with one residue in a region of high stability and one residue in a region of higher flexibility. Thus, mRMR-based pair selection on molecular dynamics trajectories, although computationally more expensive, yields more informative DEER pairs for Opa_{60}.

Figure 2 – Supplement 4. Top mRMR-predicted pairs and measured pairs have near-identical pair-configuration mutual information and mRMR values. For operational reasons, the residue-residue pairs measured via DEER were slightly different than the top mRMR-predicted pairs. As shown in the histogram in a) and mRMR table b), the predicted and measured pairs are closely linked, having near-identical pair-configuration MI and mRMR scores. The mRMR table shows values of the mRMR statistic for the second residue-residue pair selected over all combinations of predicted and measured pairs. These statistics vary by less than 5%.

Figure 3 – Supplement 1. Measured spin-echo decays and fitted distributions. Fits are superimposed in red on the decays. The red error bars in the distance distributions represent uncertainty due to the background subtraction form factor that produce fits within 15% RMSD of the best fit.

Figure 3 – Supplement 2. CW-EPR of the four MTSL spin-labeled Opa_{60}. Numbers indicate labeled amino acid positions.

Figure 4 – Supplement 1. Distance distributions of mRMR-selected pairs (top) and spectroscopist-selected pairs (bottom) from DEER and starting simulation datasets. Distance distributions are plotted from the hybrid NMR refinement simulations (without DEER restraints) used to calculate mRMR values; the distance distributions after DEER restrained-ensemble simulation are given in Fig. 4a and show an improved fit to the data. In addition to the published 20x100ns hybrid NMR refinement simulations, distributions are plotted for an additional hybrid NMR refinement simulation extended to 1.3 μs in length. Distance distributions calculated from DEER data are plotted in black.

Figure 4 – Supplement 2. Loop-loop contact modes identified by first-round refinement using mRMR-selected DEER pairs. Principal components analysis was performed on loop-loop contact matrices from restrained-ensemble simulations of mRMR-selected residue pairs. The first four principal components, which account for 25% of the total variance, are rendered in panel (**a**); the first three of

these separate restrained-ensemble snapshots into four non-overlapping clusters (**b**). The centroids of these clusters are rendered in panel (**c**), showing different loop-loop contact modes. Strikingly, the HV2 loop (red) protrudes laterally in all of these structures, while the SV and HV1 form multiple distinct sets of contacts. As above, hydrophobic residues are rendered as spheres and loops are colored with HV2 in red, HV1 in light green, and SV in tan, respectively.

Figure 4 – Supplement 3. Loop-loop contact modes identified by first-round refinement of spectroscopist-selected DEER pairs. Principal components analysis was performed on loop-loop contact matrices from restrained-ensemble simulations of spectroscopist-selected residue pairs. The first four principal components, which account for 25% of the total variance, are rendered in panel (**a**), and snapshots are plotted in a projection onto the first three principal components in panel (**b**). The centroids of these clusters are rendered in panel (**c**); in contrast to the mRMR-based refinement, these centroids primarily identify structures with all loops closely interacting and only two with the HV2 loop extended, thus requiring additional DEER pairs to yield a clear structural hypothesis regarding receptor recognition by Opa_{60}. As above, hydrophobic residues are rendered as spheres and loops are colored with HV2 in red, HV1 in light green, and SV in tan, respectively.

Figure 6 – Supplement 1. Loop-loop contact modes identified by second-round refinement using mRMR-selected DEER pairs. Principal components analysis was performed on loop-loop contact matrices from restrained-ensemble simulations of the second-round mRMR-selected residue pairs. The first three of principal components separate restrained-ensemble snapshots into ten well-separated clusters, rendered in different colors (**a**). Average contact maps for each of the three conformational states formed by these ten clusters are shown in (**b).** The centroids, rendered in (**c**), clearly show three types of loop-loop interactions: 40% of the conformational ensemble, represented by four leftmost cluster centroids, show HV1 (green) and HV2 (red) in contact, while the SV (tan) region is extended. 20% of the conformational ensemble shows the HV2 loop extended with contacts between HV1 and SV. The remainder of the ensemble is conformationally heterogeneous.

Figure 7 – Supplement 1. Measured spin-echo decays and fitted distributions in the presence and absence of CEACAM-1. Fits are superimposed in red on the decays. The red error bars in the distance distributions represent uncertainty due to the background subtraction form factor that produce fits within 15% RMSD of the best fit.