UNIVERSITY OF VIRGINIA

DOCTORAL THESIS

Selection and integration of optimal experiments for refinement of heterogeneous conformational ensembles

---

Author:
Jennifer M. HAYS

Supervisor:
Dr. Peter M. KASSON

A dissertation presented to the faculty of the
School of Engineering and Applied Science
in partial fulfillment of the requirements for the degree
Doctor of Philosophy

in the

Department of Biomedical Engineering

November 25, 2019

Approval Sheet

This dissertation is in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Biomedical Engineering

---

Jennifer M. Hays
Author
This dissertation has been read and approved by the examining committee:

Dr. Peter Kasson
Dissertation Advisor, Department of Biomedical Engineering

Dr. Craig Meyer
Committee Chair, Department of Biomedical Engineering

Dr. Michael Lawrence
Committee Member, Department of Biomedical Engineering

Dr. Linda Columbus
Committee Member, Department of Chemistry

Dr. Kateri DuBay
Committee Member, Department of Chemistry

Accepted for the School of Engineering and Applied Science

Craig Benson
Dean, School of Engineering and Applied Science

UNIVERSITY OF VIRGINIA

Abstract

Doctor of Philosophy

Selection and integration of optimal experiments for refinement of heterogeneous conformational ensembles

by Jennifer M. HAYS

Multistructured biomolecular systems play crucial roles in a wide variety of cellular processes but have resisted traditional methods of structure determination which are often optimized to resolve only a few low-energy states. Experimental measurements that do yield data on multiple conformational populations remain extremely challenging, largely because multiple measurements cannot be performed simultaneously. This leads to two major limitations: the data are often sparse over atomic degrees of freedom, making *experiment selection* a critical step in conformational refinement, and difficult to *integrate*, particularly since separate measurements cannot provide information on the joint distribution. This work addresses these two outstanding challenges in refining heterogeneous conformational ensembles.

In Chapter 2, we develop a molecular simulations and information-theory based approach to select which double electron-electron resonance (DEER) experiments best refine conformational ensembles. The approach is tested on three flexible proteins. For proteins where a clear mechanistic hypothesis exists, experiments that test this hypothesis are systematically identified. When available data do not yield such mechanistic hypotheses, experiments that significantly outperform structure-guided approaches in conformational refinement are identified. This approach offers a particular advantage when refining challenging, underdetermined protein conformational ensembles.

In Chapter 3, we develop a method to incorporate sparse, multimultimodal spectroscopic data into high-resolution estimates of conformational ensembles. We have tested our method by integrating DEER measurements on the SNARE protein syntaxin-1a into biased molecular dynamics simulations. We find that our method substantially outperforms existing state-of-the-art methods in capturing syntaxin's open–closed conformational equilibrium and further yields new conformational states that are consistent with experimental data and may help in understanding syntaxin's function.

In Chapter 4, we develop a method to estimate conformational ensembles from multiple, separately-acquired measurements by inferring their joint distribution. We have tested the method on a simplified model of an alternating-access transporter and find that the method correctly estimates both the joint distribution and the conformational

ensemble. Although the method is demonstrated on a toy system, it may be easily extended to more complex biological systems such as syntaxin.

Together, these three novel methods for refining heterogeneous conformational ensembles from spectroscopic data will greatly accelerate the structural understanding of such systems.

Acknowledgements

I would like to thank my advisor, Peter Kasson, for his support and mentorship over the five years of my PhD. He is an excellent researcher from whom I have learned a great deal, not only professionally, but personally.

I am especially grateful to Peter for connecting me with my current hematologist, Dr. Michael Douvas, whom I would also like to thank, along with Margie, and all the nurses and doctors at the UVA Emily Couric Cancer Center. I owe them a great deal more than just this dissertation.

I would like to thank our experimental collaborators who are responsible for the spectroscopic experiments in the following chapters. Linda Columbus, David Cafiso, and Marissa Keiber have provided critical data for my work and important feedback on how to think about their data like an experimentalist.

I appreciate the support and scientific insight provided by all of the members of the Kasson Research Group, but I am particularly indebted to M. Eric Irrgang, who wrote the software that enabled the research presented in the third chapter of this work. He is not only a brilliant software developer, but an excellent "statistical mechanic."

I would like to than my mother, Jean Hays, and my friends Carol Rowley, Jane Ryngaert, Rebecca Beiter, Rachel Ende, Anna Dusenberry, along with all other Catholic women who have pursued graduate work in the sciences. Their virtuous efforts have provided a constant reminder that science is, fundamentally, concerned with the pursuit of Truth.

I would also like to thank my father, Patrick Hays, for his tireless support of my now three-decade-long education. His love has taken three important forms: helping me with physics homework, helping me with math homework, and paying various institutions relatively large sums of money to assign me more physics and math homework. Thanks Dad.

Finally, I would like to thank my closest friend, Jessica Kidwell, and my husband, Michael Wagner, for loving me all these years. You have both been the source of so much laughter and joy - I will never be able to repay you both.

Financial support for this work was provided by a Commonwealth Fellowship from the University of Virginia, an NIH Biophysics Training Grant, a Blue Waters Graduate Fellowship, an ARCS Endowment Fellowship, a MolSSI Software Fellowship, and grants NSF OAC-1835780 and NIH R01GM115790.

Contents

List of Figures

List of Tables

List of Abbreviations

DEER	**D**ouble **E**lectron-**E**lectron **R**esonance
MD	**M**olecular **D**ynamics
NMR	**N**uclear **M**agnetic **R**esonance
smFRET	**s**ingle-**m**olecule **F**örster **R**esonance **E**nergy **T**ransfer
Opa	**Opa**city-associated
CEACAM	**C**arcino-**e**mbryonic **a**ntigen-related **c**ell **a**dhesin **m**olecule
mRMR	**m**inimum-**R**edundancy, **M**aximum-**R**elevancy
BRER	**B**ias-**R**esampling **E**nsemble **R**efinement
EESM	**E**nsemble **E**stimation from **S**eparate **M**easurements

This work is dedicated to the memory of Rachel Marie Quiñones who passed away from complications of Hodgkins Lymphoma in July of 2018.

Chapter 1

Introduction

1.1 An overview of the challenges in studying heterogeneous biomolecular systems

Multi-structured biomolecular systems are important in a wide variety of cellular processes and diseases, including flexible recognition events during infection and in signal transduction pathways.[1–7] For many of these systems, function depends both on large-scale conformational change and on small-scale fluctuations. It is therefore crucial to resolve their conformational heterogeneity at atomic resolution to understand function. Traditional methods of structure determination have not been optimized to do this; instead, high-resolution experimental techniques, such as X-ray crystallography and nuclear magnetic resonance (NMR), are optimized to yield information on a few low-energy states, while techniques that *do* report on heterogeneity, like double electron-electron resonance (DEER) spectroscopy and single-molecule Förster Resonance Energy Transfer (smFRET), provide data on only a few atomic degrees of freedom. Increasing interest in heterogeneous ensembles has driven significant effort to leverage the latter types of experimental data to guide high-resolution refinement. Despite this effort, a general, rigorous method for obtaining estimates of flexible ensembles using sparse, distributional experimental data has remained elusive.

In the following chapters, we develop a general, iterative methodology that utilizes distributional experimental data and molecular dynamics (MD) simulation to estimate the conformational ensembles of flexible proteins (Fig. 1.1). Our approach addresses three outstanding challenges in structural refinement which we detail in the remainder of the Introduction and in the subsequent three chapters: optimal selection of low-throughput spectroscopic experiments (Chapter 2), incorporation of distributional data into high-resolution estimates of conformational ensembles (Chapter 3), and finally, estimation of conformational ensembles from separately-acquired, yet correlated, experimental measurements (Chapter 4). The methods are designed to be generalizable to any experimental technique that yields non-parametric data on an ensemble, but they

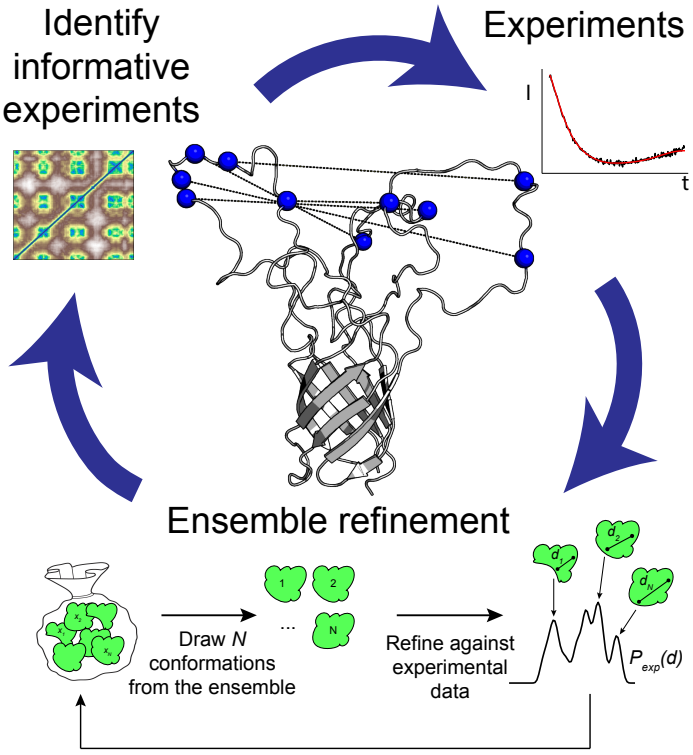

FIGURE 1.1: **Schema for refining heterogeneous conformational ensembles.** Data from MD simulations are used to select optimal experiments according to the methods described in Chapter 2. Data from these experiments are then incorporated into MD simulation using the methods described in Chapter 3. If it is impossible to obtain experimental information on the correlation structure of multiple measurements, the method developed in Chapter 4 is used to infer that correlation structure and improve the ensemble estimate.

were motivated by double electron-electron resonance (DEER) experiments performed on the two biological systems described in Section 1.2.

Broadly, there are two steps in refining flexible systems using sparse experimental data: selecting a set of experiments to perform, then integrating the resulting data to produce an estimate of the ensemble. Existing methods for experiment selection rely on having significant prior structural data on the ensemble. One recently developed method relies on extensive simulation to build kinetic models and has been tested only retrospectively, serving primarily to validate computational estimates rather than leverage computation to prospectively guide selection.[8] Other selection methods require choosing one or a small set of structures for estimating optimal spectroscopic label placement, and thus fall short as system complexity and flexibility increases.[9–11] These methods implicitly depend on having a predetermined set of structures at hand; often it is impossible to experimentally determine and prohibitively expensive to compute a

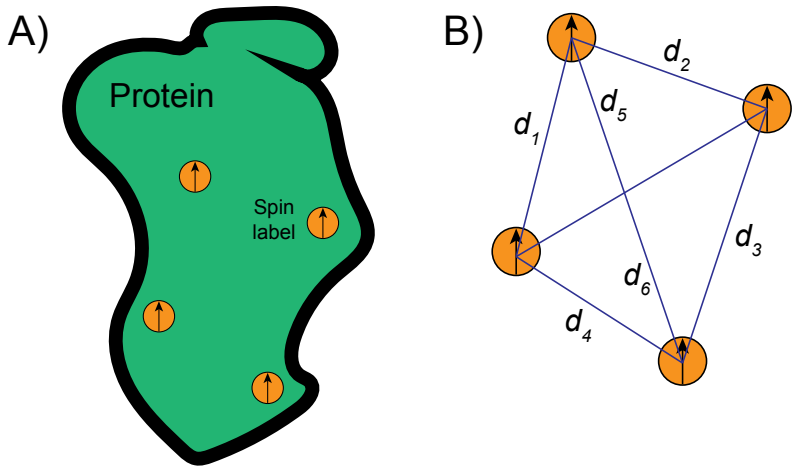

FIGURE 1.2: **Label-based distance measurements must be performed sequentially.** If four labels are introduced simultaneously as shown in A), then *six* distance variables are measured, shown in B), rather than two. The resulting distribution becomes extremely challenging to deconvolve into six pair-wise distributions.

well-sampled ensemble for highly flexible systems. We have addressed this problem by developing a method that utilizes *undersampled* estimates of the conformational ensemble to select a set of optimal experiments (Chapter 2).[12]

Once optimal experimental data have been acquired, they must be integrated into an estimate of the ensemble. Current methods for doing this integration have generally focused on experimental techniques that yield ensemble-average data,[13–22] rendering them unsuitable for systems where the details of the underlying distributions are critical for biological function. Methods that *do* integrate fully distributional data have been optimized to capture details of side-chain fluctuations rather than backbone heterogeneity,[23–25] and, even when re-optimized to accommodate backbone fluctuation, exhibit pathological behavior when driving sampling of systems with well-separated backbone conformations (Fig 2.6).[26] We have developed a method for integrating non-parametric, distributional data that is specifically designed to capture large-scale conformational change (Chapter 3).[26]

In all cases, the methods used to incorporate distributional data into simulation assume that the data are independent, which is often not appropriate. This assumption is typically made because it is extremely challenging to acquire experimental information on the correlation structure of the distributions. Label-based methods such as DEER and smFRET must be performed sequentially; because the labels are indistinguishable, introduction of more than two labels at a time leads to cross-talk between all labels, making individual pair-wise distributions very difficult to recover (Fig. 1.2). Rather than attempt to deconvolve multi-spin experimental data, which are prone to producing

extraneous peaks,[27] we have developed a method for estimating the joint distribution of separately-acquired experimental measurements (Chapter 4). Not only can this method estimate the correlation structure of separate experiments, but it directly enables estimation of a conformational ensemble that agrees with the joint distribution.

We developed these methods in response to specific challenges that are commonly encountered when studying two classic paradigms of heterogeneous systems: flexible molecular recognition and assembly of heterogeneous biomolecular complexes. The specific systems that motivated these methods are outlined below.

1.2 Common paradigms of heterogeneous ensembles

1.2.1 Molecular recognition: Opa_{60}-CEACAM engagement

The Neisserial outer membrane protein Opa_{60} is a critical component in attachment to host cells and subsequent cellular uptake of the bacterium, yet the mechanisms by which it engages its host are still not well understood.[28,29] Opa proteins consist of a β-barrel and four highly-mobile loops which are known to engage a small surface area of their receptor CEACAMs on host cells.[30] It is also known that these loops exhibit extreme sequence variability among Opas. Despite this combination of sequence and conformational flexibility, Opa engages CEACAM with nM affinity.[31] These findings beg a fundamental biophysical question: how does Opa, with no apparent binding motifs or obvious "bound state" conformation, engage its receptor? We hypothesize that the loops remain conformationally flexible in the Opa-CEACAM bound state, thus minimizing the entropic penalty of binding while contributing only a small enthalpic term to the free energy (Fig. 1.3). We have begun to test this hypothesis using our methodology for optimal selection of spectroscopic experiments.

In Chapter 2, we refine the *apo* Opa_{60} ensemble by rigorously selecting and performing a set of optimal DEER experiments. A quantitative methodology for experiment selection is essential for this system: with over 5,000 possible inter-loop pairs to choose from, it becomes extremely difficult to select pairs that yield important information on the ensemble without significant prior structural data or biochemical insight, neither of which are available for Opa_{60}. Our preliminary results reveal a clear structural hypothesis for how Opa_{60} might engage its receptor that would not have been apparent with other refinement techniques. These results will be used to guide additional spectroscopic experiments on the CEACAM-bound ensemble.

Not only will refinement of the Opa_{60}-CEACAM ensemble illuminate new fundamental properties of flexible molecular recognition, it will facilitate development of new

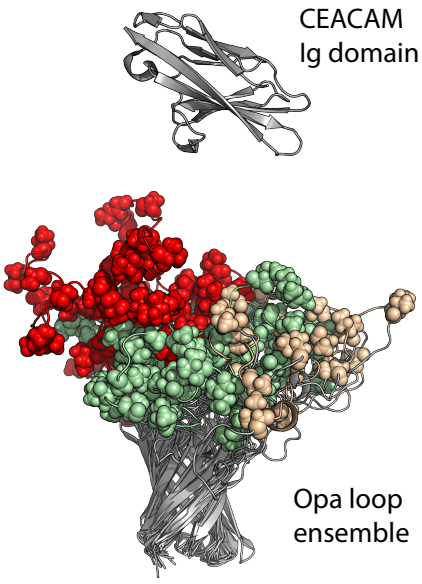

FIGURE 1.3: **Opa_{60} loop heterogeneity may minimize the entropic penalty of binding CEACAM** The twenty lowest-energy structures from NMR experiments on Opa_{60} are rendered along with the Ig domain of CEACAM1. The different variable regions of the loops are rendered in red, green, and tan. The Opa_{60} loop ensemble is strikingly heterogeneous. We hypothesize that this heterogeneity is maintained even when bound to CEACAM, minimizing the entropic penalty of binding.

models for targeted drug delivery. Because Opas evade immune surveillance and trigger phagocytic uptake of cellular contents, liposomes expressing modified Opas may be used to deliver therapeutics to cells that have been previously difficult to target.

1.2.2 Molecular assembly: SNARE-mediated exocytosis

Soluble N-ethylmaleimide-sensitive factor attachment receptor (SNARE) complexes are composed of a four-helical bundle of proteins which drive neuronal vesicle fusion and thus facilitate release of neurotransmitters into the synaptic cleft.[32–34] Syntaxin-1a, one of the proteins that participates in SNARE formation, exhibits a complex conformational equilibria (Fig 1.4). NMR and Fluorescence Interference-Contrast (FLIC) studies on the transmembrane and SNARE-binding domains of syntaxin suggest that the SNARE-binding H3 domain is α-helical when near a membrane,[35,36] yet DEER data taken on the soluble domain suggest that H3 is disordered.[37] We hypothesize that the conformational equilibrium of syntaxin is modulated by the presence of the membrane and have begun to test this using our methodology for integrating distributional experimental data.

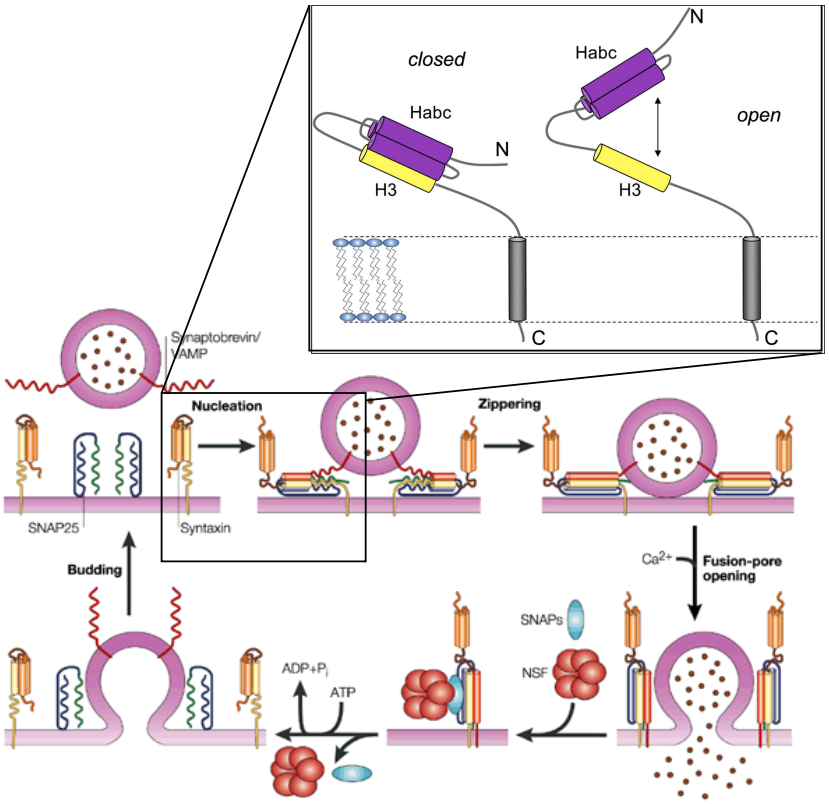

FIGURE 1.4: **Syntaxin-1a participates in formation of the SNARE complex and drives synaptic vesicle fusion.** Formation of the SNARE involves assembly of four proteins, SNAP25, synaptobrevin, VAMP, and syntaxin. A conformational change of the SNARE drives synaptic vesicle fusion (main panel). Prior to formation of the SNARE, syntaxin exhibits a dynamic equilibrium, alternating between closed and open states (inset). This equilibrium is not well understood, particularly how it changes upon membrane interaction. This figure is adapted from [37, 38]

We obtained a preliminary estimate of the conformational ensemble of the soluble domain of syntaxin-1a using experimental DEER data and found a new family of states that may be important for SNARE complex formation. It is generally thought that syntaxin must be in an open state in order to participate in formation of the SNARE, but the open state has yet to be fully characterized.[35,38,39] The common understanding of the open state, which is based on crystal structures of syntaxin in complex with the other SNARE proteins,[40,41] is that the H3 domain remains well-ordered while the linker region connecting Habc and H3 becomes flexible. Our experimentally-refined ensemble reveals new open conformations of syntaxin-1a that differ substantially from the commonly proposed models (Chapter 3).[26]

There are two ways to further refine the soluble conformational ensemble: because no data exist on the correlations between these measurements, we can improve the ensemble estimate by inferring the joint distribution and appropriately reweighting the initial ensemble (Chapter 4). Additional targeted DEER experiments could then resolve the presence of multiple open-state populations and test whether the conformations observed in simulation are indeed present. These studies would yield further insight into the syntaxin conformational ensemble and SNARE complex assembly in general.

The approach described above will be particularly powerful when used to study the syntaxin membrane interaction. In this case, multiple types of experimental data (DEER and FLIC) acquired under different laboratory conditions must be integrated to obtain an estimate of the conformational ensemble. Because the methods described in Chapters 3 and 4 obey the principle of maximum entropy, i.e., they minimally perturb the MD Hamiltonian, we can use them to leverage the solution ensemble of Chapter 3 to estimate the membrane-bound ensemble. The details of this refinement procedure are provided in Chapter 5.

1.3 Summary of MD simulation and DEER spectroscopy

1.3.1 MD Simulation

Molecular dynamics simulation is a computational technique that is often used to study the physics of many-body systems at high resolution; one defines a Hamiltonian that specifies the interactions between particles along with other thermodynamic properties of the system such as temperature and pressure, then uses that Hamiltonian to numerically integrate Newton's laws of motion. The result of an MD simulation is a rich dataset that describes the time-evolution of the system. However, these simulations are limited by the accuracy of the Hamiltonian and, most significantly, by the time-scales they can access. Typical simulations can reach microsecond time-scales, while even the longest simulation runs barely reach a millisecond.[42] Thus, standard MD simulations

produce incomplete estimates of ensembles, particularly in the case of highly flexible systems. We therefore incorporate DEER data into our simulations to improve sampling of experimentally-valid regions of phase space.

1.3.2 DEER Spectroscopy

Double electron-electron resonance spectroscopy is an experimental technique that measures distance distributions in the range of 1.6 nm to 6.0 nm between pairs of electron spins. In a manner similar to NMR experiments, interactions between spins are observed in the time domain in response to a series of applied magnetic pulses.[43] It is then possible to transform the dipolar coupling times of the electron spins into distance distribution functions.[44] For proteins that do not have naturally occuring paramagnetic centers, electron spins are introduced via site-directed spin labeling (SDSL).[45,46] This allows measurement of a wide variety of pairwise distance distributions for proteins. However, because each measurement must be made separately, DEER data are often sparse over atomic degrees of freedom. Thus, it becomes critical to select maximally informative pairwise distance variables for measurement; this is the subject of the following chapter.

Chapter 2

Simulation-guided spectroscopy

2.1 The mRMR algorithm: theory and applications

The contents of this section are published as a research article in:

Refinement of Highly Flexible Protein Structures using Simulation-Guided Spectroscopy. Jennifer M. Hays, Marissa K. Kieber, Jason Z. Li, Ji In Han, Linda Columbus, and Peter M. Kasson. *Angewandte Chemie International* 2018 (130) 17356 –17360.

Heterogeneous conformational ensembles play critical roles in molecular recognition and cellular regulation, yet high-resolution structure determination has typically required reducing these ensembles to only a few states.[1,2,4] Since the full equilibrium ensemble is often key to understanding biochemical function, other experimental techniques have been developed to probe the full ensemble distribution rather than either a few low-energy states or an equilibrium average.[7,13–15] However, these experiments measure only a small number of atomic degrees of freedom:[47–50] for instance, double electron-electron resonance (DEER) and single-molecule Förster Resonance Energy Transfer (smFRET) spectroscopy, which utilize pairs of labeled amino acids to obtain distance distributions, typically provide data for ≈ 10 measurements per system. Thus, experiment selection is currently the limiting factor in how much information can be obtained on an ensemble.

Prior quantitative approaches to experiment selection have relied on pre-existing high-resolution structural and kinetic models. Recent studies have shown, retrospectively, that leveraging either Markov State Models[8] or normal modes calculated from elastic network models[9] can select good labels for DEER experiments. But for systems where traditional structural or kinetic models are incomplete or fundamentally underdetermined due to conformational flexibility, it remains challenging to determine which pairs of residues should be chosen for labeling. We have therefore developed a general, information-theoretic formalism to select optimal spectroscopic experiments. We summarize the theory and show the application of this method to three conformationally heterogeneous bacterial proteins.

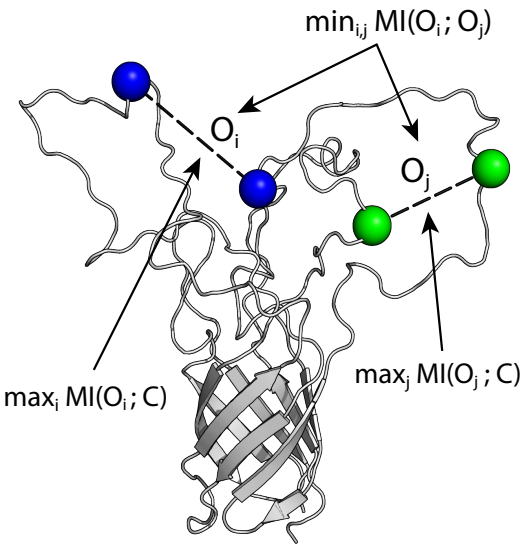

FIGURE 2.1: **The maximum-relevancy, minimum-redundancy (mRMR) method applied to spectroscopic observables O_i.** The optimal set of spectroscopic experiments that report on variables O_i are maximally informative of the conformation C and minimally redundant with each other. Informativeness and redundancy are quantified via mutual information (MI).

An optimal set of spectroscopic experiments has two properties: each experiment yields the maximum amount of information on the conformational ensemble and minimally redundant information with other experiments in the set to avoid wasting labeling and measurement effort (Fig 2.1). The maximum-relevance, minimum redundancy (mRMR) algorithm exactly satisfies these criteria.[51,52] To select N spectroscopic experiments, we maximize the mutual information (MI) between the set of spectroscopic observables O_i and the conformation C:

$$\max_i \frac{1}{N} I(O_i, C) \tag{2.1}$$

where C is the set of $n(n-1)/2$ pairwise distance variables. We simultaneously minimize the pairwise MI between spectroscopic variables O_i and O_j (Fig 2.1):

$$\min_{i,j} \frac{1}{N^2} I(O_i, O_j) \tag{2.2}$$

where $I(X, Y)$ is the mutual information between random variables X and Y:

$$I(X,Y) = \sum_{\{x\}} \sum_{\{y\}} P_{X,Y}(x,y) \log \frac{P_{X,Y}(x)}{P_X(x) P_Y(y)}$$

This method is particularly useful because it identifies, *by design*, those observables which are maximally underdetermined in a structural ensemble. These underdetermined observables are precisely those that would be especially challenging for traditional structure determination.

In our study, we make two approximations: we use deliberately undersampled estimates of the protein conformational ensemble to select labels for further refinement, and we approximate the spectroscopic variable O_i as the C_α -C_α pairwise distance distribution between labeled residues. We make the first approximation to demonstrate the strong advantage of our method for heterogenous ensembles: by identifying underdetermined degrees of freedom we can improve an incomplete estimate of the conformational ensemble rather than requiring a well-sampled starting model. The second approximation is an implementation rather than theoretical concern and we will discuss how it can be removed. The success of the mRMR method and these approximations is demonstrated below on a set of flexible bacterial outer membrane proteins.

β-barrel membrane proteins are excellent candidates for the mRMR approach because many contain flexible regions that are difficult to characterize experimentally yet have regions of secondary structure that make spectroscopic experiments tractable.[30,31,53–55] We have performed molecular dynamics (MD) simulations on three bacterial outer membrane proteins and applied the mRMR algorithm to select optimal DEER experiments. We have chosen FhuA, an E. coli iron transporter,[56] OprG, a pseudomonal small-molecule transporter,[57] and Opa_{60}, a Neisserial virulence-associated protein that binds cell-surface proteins but does not function as a transporter.[29] The FhuA conformational ensemble has been characterized via DEER experiments guided by pre-existing mechanistic hypotheses that relate conformational changes of the Ton box domain to ligand recognition;[58] it is thus a good test system for determining whether the mRMR algorithm identifies similar labels to those identified by spectroscopists. OprG, a more challenging system, has been studied using a combination of NMR and mutational experiments,[53] but the mechanisms by which transport is regulated remain unknown. Finally, Opa_{60} represents a particularly challenging system since it displays substantial, experimentally underdetermined conformational flexibility that controls its binding mechanism.[30] We have therefore studied this final system prospectively: choosing a set of residue-residue pairs using the mRMR algorithm, measuring them with DEER, incorporating the experimental data into MD simulation, and evaluating this ensemble versus one refined with spectroscopist-selected pairs (SSP). Simulations alone lack the fidelity to reliably predict structural ensembles of flexible proteins but can serve as a good platform for hybrid refinement combining physical information with experimental data.

For each protein, we generated initial estimates of the conformational ensembles using ensemble MD simulations that were deliberately undersampled at 2 µs per protein.

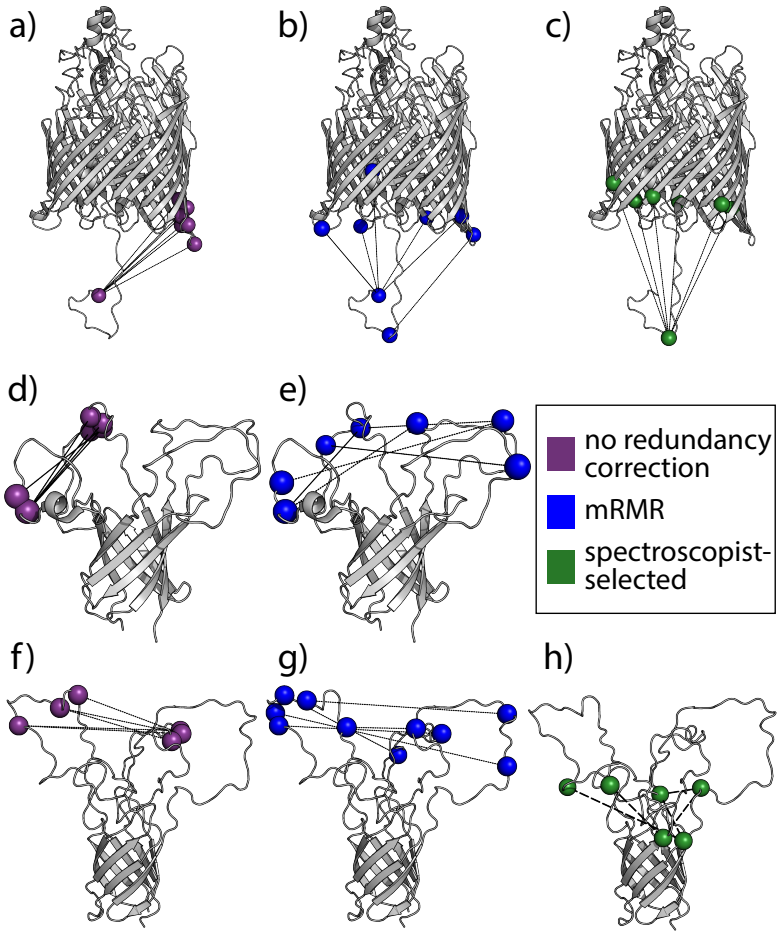

FIGURE 2.2: **Capture of highly informative, minimally redundant residues on three bacterial outer membrane proteins with mRMR.** Selection via mutual information alone yields informative, but redundant, pairs (magenta). Selection via mRMR (blue) removes this redundancy. These residues are better distributed across the structures of all three bacterial proteins than the top-ranking MI pairs or ones selected by spectroscopists according to current practice in the field (green). a–c, d–e, and f–h show residues selected for FhuA, OprG, and Opa_{60}, respectively.

We used the mRMR algorithm on these data to select sets of pairwise distances that optimally report on undersampled regions of phase space (Fig 2.2).

In the case of FhuA, spectroscopists selected label pairs near the N-terminal domain, which is conformationally heterogeneous and regulates transport, and the periplasmic side of the beta-barrel using a standard triangulation strategy (Fig 2.2c).[58] Selection via mRMR identifies similar residues (Fig. 2.2b), with the addition of one pair spanning just the N-terminal domain. Label pairs not corrected for redundancy also specifically identify distances between the N-terminal domain and one side of the barrel as most informative (Fig 2.2a). DEER analysis independently identified this side of the barrel as interacting with the N-terminal domain. These two findings on FhuA, a relatively well-understood transport protein, show that the mRMR method can select label pairs that reflect best spectroscopic understanding and yield insight into conformational heterogeneity.

Our method provides even greater potential benefit when less is known about transport mechanism, as in the case of OprG, and may help test claims of loop involvement in OprG transport. Both the mechanism and the substrates for OprG transport are unclear: OprG may transport small, hydrophobic compounds via a lateral gating mechanism or small amino acids via the barrel channel; OprG crystal structures support the former hypothesis,[59] while recent NMR and mutational studies suggest the latter.[53] Non-transporting mutants studied via NMR have generally more ordered loops, and one loop has especially restricted motion, suggesting it may be critical to transport. Interestingly, this loop participates in all five informative OprG residue-residue pairs (Fig 2.2d) and in three of the five top-scoring mRMR pairs (Fig 2.2e). Thus, mRMR analysis yields label pairs that reflect existing mechanistic hypotheses and, most importantly, identify experiments to test these hypotheses.

As a robust test of mRMR-based label selection, we prospectively tested its ability to select DEER experiments and refine the conformational ensemble of Opa_{60}, the most challenging protein in our evaluation set. DEER data were acquired using label pairs selected via both mRMR and traditional structure-based selection, and we assessed the relative utility of each method in refining the ensemble. Opa_{60}'s long, flexible loops are both critical for function[30,31] and challenging for previous DEER pair selection methods. In contrast to FhuA or OprG, no structural or functional data provide strong guidance on which residues are responsible for function, in this case receptor engagement. Prior hybrid NMR-MD refinement of the apo conformational ensemble did not provide sufficient insight into the binding mechanism. Normal-mode approaches developed by Zheng and Brooks have been applied to identify informative, non-redundant label sets for DEER that differentiate pairs of structures when such structural data exist,[9,10,60] but this is not the case for Opa_{60}. Furthermore, normal-mode calculations from an

Opa_{60} elastic network model do not correlate with flexibility measured via NMR relaxation timescales (Fig A.1). Thus, spectroscopists must choose from more than 5,000 possible inter-loop pairs. We show below that mRMR selection method radically improves structural refinement compared to standard spectroscopic practice for systems that were previously intractable.

We prospectively tested mRMR pair selection by refining the Opa_{60} conformational ensemble[61] using two independently identified label sets: one selected using the mRMR algorithm and the other independently chosen by spectroscopists. The top five top-scoring mRMR pairs span multiple combinations of inter-loop distances (Fig 2.2g), and the top ten pairs capture all possible combinations of the loops (Table A.1). By contrast, the top ten pairs identified using maximum relevancy alone span a single loop-loop pair. Although the maximum-relevancy pairs define the most variable loop, they lose important information about the other loop (Fig 2.2f). The spectroscopist-selected pairs are primarily short barrel-loop distances because the length of the loops permits distances too long to be measured via DEER, so spectroscopic best practice is to select a more conservative set of pairs. However, this aside, the chance of manually selecting a loop-loop pair within the top 25% of those identified via mRMR is only 7%, showing a strong advantage for the systematic selection methods developed here.

Because Opa_{60} is so conformationally flexible, approximating the label-label distance distributions as C_α -C_α distributions introduces little error relative to the backbone motions of the protein. However, label flexibility becomes increasingly important to label selection as protein flexibility decreases, and explicit labels may be added as follows. First, unrestrained simulations of the wild-type protein may be used to calculate initial mRMR estimates. Explicit labels are introduced for each top-ranked residue-residue pair, and one additional simulation is performed per pair. The mRMR scores are recalculated for each simulation to determine the effects of label side-chain conformation on the final mRMR rankings. A “forward model" can be used for the spectroscopic observable, such as the predicted DEER spectrum,[10] using the explicit-label simulations.

To assess the quality of mRMR-guided versus structure-guided refinement, we estimated the Opa_{60} conformational ensemble using DEER data on pairs selected via each approach. We then compared the resulting ensembles using two independent metrics which we developed to quantitatively evaluate “quality of refinement." As a first metric, we measured how well each refined ensemble predicts DEER data held back from refinement as a test set. Refinement using mRMR-selected label pairs yielded significantly better agreement with the test DEER data: seven of eight test distributions are better captured by the mRMR-guided ensemble than the structure-guided ensemble (Fig 2.3).

We also analyzed the dimensionality of the conformational ensembles obtained from refinement using structured-guided versus mRMR-guided DEER data. Given sufficient

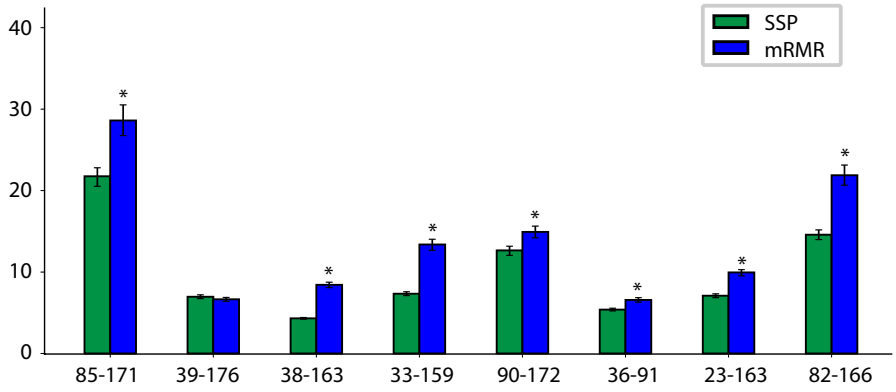

FIGURE 2.3: **mRMR-guided refinement predicts test DEER distributions better than structure-guided refinement.** Quality of refinement was evaluated by ability to predict additional 8 residue–residue pairs measured using DEER. Conformational ensembles refined using mRMR-selected pairs predict these DEER distributions significantly better than ones refined using spectroscopist-selected pairs (SSP) in 7 of 8 cases, quantified as inverse Jensen-Shannon divergence. Error bars show 90% CI from 1000 bootstrap samples; * denotes $p < 0.01$ via two-tailed t-tests.

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 (see A.1.1).

Because residue-residue distances yield an overcomplete basis set, we lumped together highly related distance variables at different thresholds of relatedness (ϵ) and calculated the number of independent distance variables required to describe the ensemble at each ϵ. At every ϵ tested, refinement with mRMR-selected DEER data yielded a conformational ensemble of lower dimensionality than with spectroscopist-selected DEER data (Fig. 2.4a). This indicates that DEER data from mRMR-selected pairs refine the conformational ensemble more efficiently than data from pairs selected according to current state of the art.

mRMR pair-selection also produces strikingly more informative structural results than spectroscopist-guided selection. We determined the major loop conformations in each ensemble by clustering loop-loop contact maps. After one iteration, mRMR-guided refinement yields four clusters, all of which show one loop protruding laterally and two loops closely interacting (Fig 2.4c). In contrast to mRMR-guided refinement, refinement using spectroscopist-selected pairs yields a larger number of structural clusters with poorly resolved loop conformations (Fig 2.4b) that also poorly predict additional DEER measurements (Fig 2.3). The loop conformations resolved by mRMR-guided refinement further yield a structural hypothesis for receptor recognition whereby either the two contacting loops or the one splayed loop is primarily responsible for receptor binding.

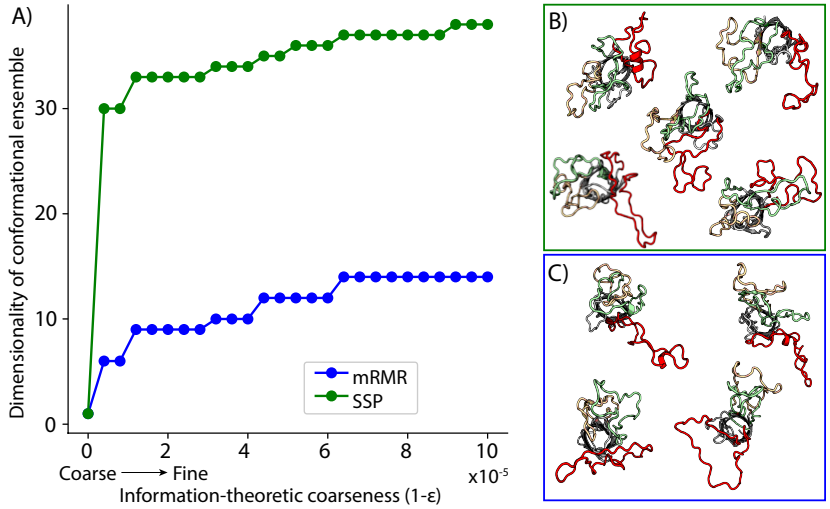

FIGURE 2.4: **mRMR-guided refinement produces ensembles of lower dimensionality than structure-guided refinement.** a) The dimensionality of the conformational ensemble (the number of independent distance variables), is plotted at each information theoretic resolution ϵ. Ensembles refined using mRMR-selected pairs are of lower dimensionality than those refined using SSPs by 20–25. b) Structures identified by cluster analysis of inter-loop contacts are also shown for each ensemble. mRMR refinement yields conformations in which a single loop extends from the base of the barrel while the two remaining loops interact. SSP refinement yields conformations with no well-defined loop–loop interaction patterns.

These tests demonstrate that mRMR provides a robust approach to spectroscopic label selection, particularly for flexible proteins where structural estimates are more challenging and the difference in data quality between optimally selected labels and poorly selected labels is greater. When strong mechanistic hypotheses have guided prior DEER experiments, mRMR yields label pairs that would test these hypotheses. For proteins such as Opa_{60} where mechanistic understanding is insufficient to guide experiment selection, we show via prospective testing that mRMR selection outperforms unaided spectroscopists. Therefore, we believe that mRMR will be of general use in guiding spectroscopic experiment selection for DEER and for other label-based methods such as smFRET and paramagnetic resonance enhancement. The method can also be extended to differentiate mechanistic hypotheses rather than conformations. For systems like OprG where two mechanistic hypotheses exist, mRMR could be used to identify which spectroscopic variables optimally distinguish conformational features specific to one mechanism or the other. Conformational flexibility and heterogeneity are some of the most challenging and exciting frontiers in understanding protein structure, and mRMR will increase the ability of these experimental methods to efficiently refine such conformational ensembles.

2.2 Preliminary refinement of the Opa_{60}-CEACAM ensemble

The contents of this section are published as part of a research article in:

Refinement of Highly Flexible Protein Structures using Simulation-Guided Spectroscopy. Jennifer M. Hays, Marissa K. Kieber, Jason Z. Li, Ji In Han, Kelley W. Moremen, Linda Columbus, Peter M. Kasson. *bioRxiv* January 1, 2018, 319335.

2.2.1 Review of the Opa_{60}-CEACAM interaction

Opacity associated (Opa) proteins bind to human carcinoembryonic antigenrelated cellular adhesion molecules (CEACAMs) triggering cellular uptake and mediating cellular invasion.[29] Opa_{60} is a canonical eight-stranded β-barrel integral membrane protein with four extracellular loops (Fig A.7) that are dynamic on the nanosecond time scale and predominantly disordered.[30] The long ligand-binding loops have high sequence diversity[28] and are hypothesized to be flexible in both the unbound state and during CEACAM engagement, which aids in immune evasion.[30] Despite high mobility and high sequence diversity among Opa variants, the loops bind with high affinity to CEACAM1.[31] 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 further refined

the conformational ensemble of Opa_{60} through additional DEER experiments directed by MD simulations. The ensemble estimate reveals previously uncharacterized loop-loop interaction modes and potential sites for CEACAM engagement.

2.2.2 Analysis of Opa_{60} loop-loop interactions

We obtained additional 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 ensemble simulations. These pairs span multiple biologically significant regions of the extracellular loops, namely the hypervariable and semivariable regions (Fig A.7), denoted HV1, HV2, and SV. The refined conformational ensemble yields several sets of specific loop-loop interactions that potential sites for CEACAM engagement.

To analyze these interactions systematically, we identified the most abundant loop conformations in the structural ensemble by clustering PCA-transformed loop-loop contact maps (Fig A.8). Ten well-separated clusters were formed in the projected subspace. Strikingly, in the centroids of three of the four the most populated clusters, HV1 interacts with HV2 or SV. 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 supports these interaction patterns (Fig A.8).

2.2.3 Opa_{60}conformations recognized by CEACAM.

HV1/HV2 chimeric Opa_{60} proteins have previously shown that specific HV1/HV2 combinations are required for CEACAM engagement, leading to a model in which HV1 and HV2 together directly engage receptors.[62] Our results on Opa_{60} 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.[62] These new hypotheses generated by the experimentally-derived conformational ensemble can now be further tested with carefully designed binding experiments.

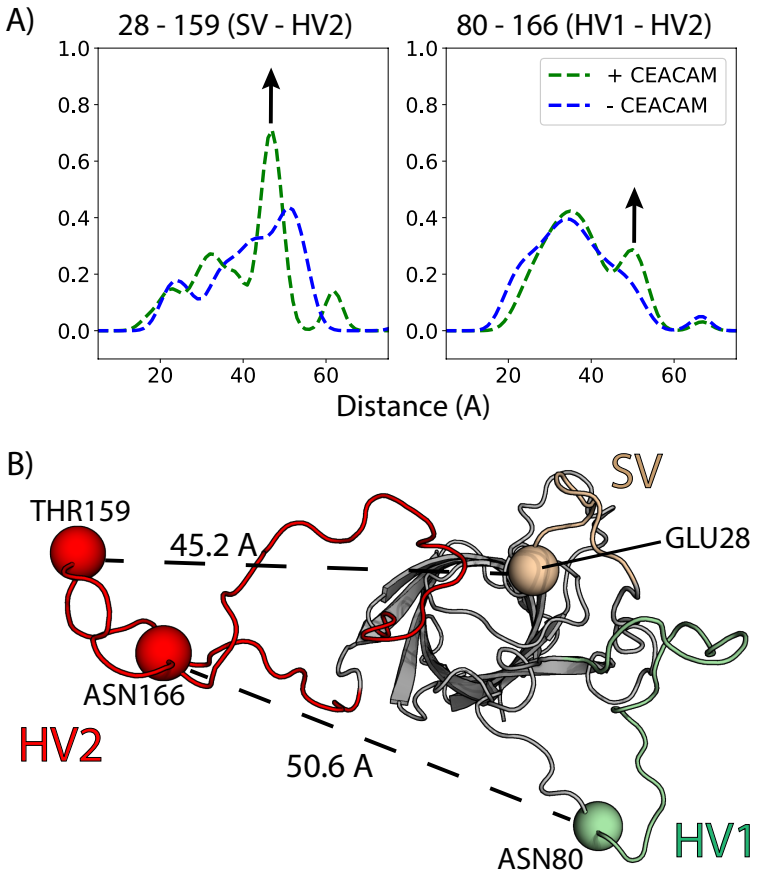

FIGURE 2.5: **Binding of CEACAM selects for HV2-extended Opa conformations.** DEER measurements of Opa-Opa residue pair distances show substantial shifts in the distance distributions upon binding to CEACAM (a), suggestive of conformational selection. Using these data to interpret the apo conformational ensemble, we find that HV2-extended conformations 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 Fig A.2.

To determine whether CEACAM binds the SV-extended Opa conformation or the HV2-extended Opa conformation, we performed additional DEER experiments on labeled Opa_{60} proteins with and without bound CEACAM1. The resulting data show substantial shifts in Opa_{60} loop-loop distance distributions upon CEACAM1 binding, consistent with conformational selection: a subset of the distances present in the *apo* protein increase, while others decrease (Fig 2.5a). 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 2.5B). 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).

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 2.4B); 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 2.4C). Since these conformational ensembles poorly predict the additional DEER distance measurements (Fig 2.3), they are incorrect and would have required further DEER measurements to yield similar hypotheses for the determinants of Opa-CEACAM binding.

2.3 The need for improved methods of incorporation: pitfalls of restrained-ensemble simulations

Although the modified version of restrained-ensemble simulations[23,61] used in Sections 2.1 and 2.2 successfully incorporated the Opa_{60} DEER-derived distributions, this method

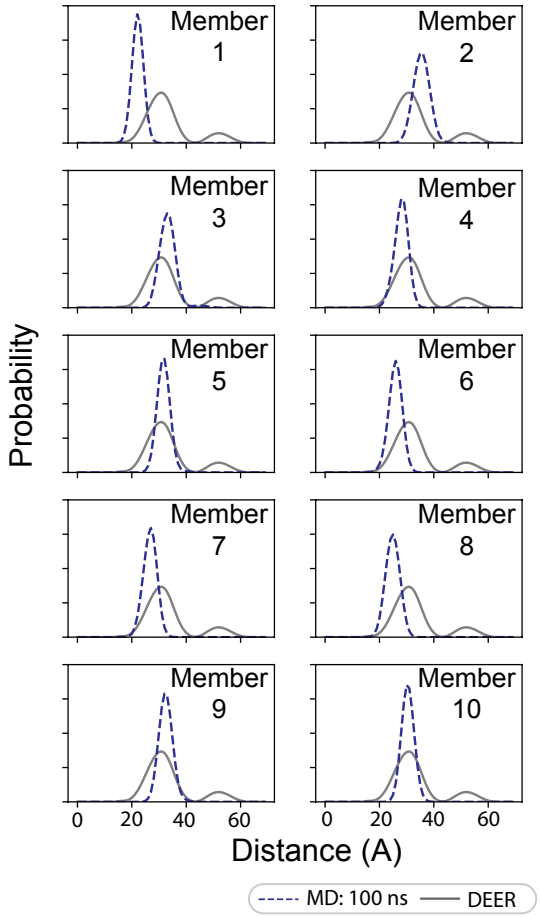

FIGURE 2.6: **Restrained-ensemble simulations damp exchange between well-separated probability modes.** None of the ensemble members from a 100 ns (1 µs aggregate) restrained-ensemble simulation sample the long distance mode of the DEER-derived distribution shown in gray. Indeed, each ensemble member samples only a narrow, unimodal Gaussian (dashed blue lines).

exhibits pathologies when distributions have well-separated probability modes (Fig 2.6). The following chapter develops a methodology for incorporation of distributional experimental data that is specifically intended to capture well-separated backbone conformational change.

Chapter 3

Incorporation of distributional data into high-resolution estimates of heterogeneous ensembles

3.1 Bias-resampling ensemble refinement (BRER): theory and applications

The contents of this section are published as a research article in:

Hybrid Refinement of Heterogeneous Conformational Ensembles Using Spectroscopic Data. Jennifer M. Hays, David S. Cafiso, and Peter M. Kasson. *The Journal of Physical Chemistry Letters* 2019 10 (12), 3410-3414.

Heterogeneous conformational ensembles play important roles in in a wide variety of cellular processes and diseases, including flexible recognition events during infection and in signal transduction pathways.[1–3,5,6,63] The major challenges in structural refinement of these flexible systems are twofold. Often, experimental data yield ensemble-average quantities and/or the experimental data are sparse.[9,13–15,64] Both of these difficulties lead to an ill-posed inverse problem: in both cases, an ensemble of structures, which is degenerate in the experimental quantity of interest, must be enumerated with very little other information.[65–67] One way to avoid this inverse problem is to use a forward model, such as a molecular dynamics (MD) force field, and directly integrate the experimental data into the model such that the final ensemble reproduces the experimental data. A great deal of work has been done to develop methods for integrating ensemble-average quantities into forward models, and this work has been quite successful.[16,19–21,68–74] However, a robust strategy for integrating sparse, distributional data has remained elusive.

Here we describe a hybrid maximum-entropy–stochastic-resampling approach for biasing molecular simulation ensembles toward experimental distributions rather than ensemble averages. We apply this method to double electron–electron resonance (DEER)

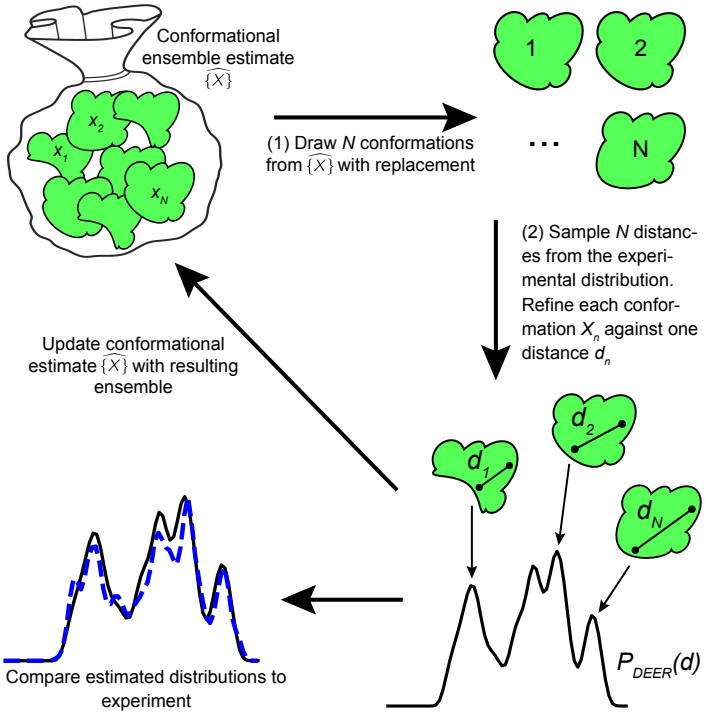

FIGURE 3.1: **Estimation of a conformational ensemble $\{X\}$ by stochastic resampling of experimental data.** An iterative update framework for bias resampling ensemble refinement from an initial estimate is schematized as follows: (1) a set of N conformations is drawn from the conformational ensemble estimate $\widehat{\{X\}}$, and (2) each conformation is refined against a single target which is stochastically resampled from an experimental distribution. At each iteration, the estimated distribution calculated from $\widehat{\{X\}}$ is compared against the experimental distribution. If the distribution $P_{\{X\}}(d)$ is significantly different from $P_{\text{experimental}}(d)$, the conformational estimate $\widehat{\{X\}}$ is updated with the refined structures and the refinement procedure is repeated.

data, but the method is extremely general and may be used for nearly any experimental method yielding distributional data. The method exhibits no instabilities in regions of zero probability and can sample important backbone conformational change. We describe how to incorporate a single distribution first; we then discuss a simple generalization to multiple distributions.

Our hybrid approach, which we call bias-resampling ensemble refinement (BRER), uses an iterative refinement scheme to update an estimate of the conformational ensemble $\widehat{\{X\}}$ based on a DEER distribution $P_{\text{DEER}}(d)$ (Fig 3.1). The simplest formulation of BRER is described here, while a more complex variant is given in Appendix B.1.1. During refinement, each conformation $x \in \widehat{\{X\}}_{i-1}$ is updated using a biased MD simulation such that the updated estimate $\widehat{\{X\}}_{1\ldots i}$ better reproduces $P_{\text{DEER}}(d)$. Thus, over

the course of multiple rounds of refinement, the conformational estimate $\widehat{\{X\}}$ should yield a distribution $P_{\widehat{\{X\}}}(d)$ that converges on $P_{\mathrm{DEER}}(d)$. The initial estimate $\widehat{\{X\}}_0$ may be obtained using experimental data (an NMR ensemble, a single-crystal structure) or an experimentally informed model.

The stochastic-resampling approach is implemented as follows: because any distribution $P_{\mathrm{DEER}}(d)$ may be represented as a sum of Gaussians, let the experimental distribution be a linear combination of M Gaussians with centers located at d_m:

$$P_{DEER}(D) = \sum_{m=1}^{M} \frac{p_i}{\sqrt{2\pi\sigma^2}} e^{-(d-d_m)^2/2\sigma^2} \tag{3.1}$$

with weighting factor p_i. During the i^{th} round of refinement, we randomly sample a set of N structures from the conformational estimate of the previous round, $\widehat{\{X\}}_{i-1}$. Each structure is assigned one target distance d_n via a probability-weighted draw from the set of experimental distances d_m. A maximum-entropy biasing potential is then applied to each ensemble member, driving the member toward its target distance.[17,75,76] We allow the ensemble to relax at the target distribution (which has been resampled from $P_{\mathrm{DEER}}(d)$) for some time t, at which point the resampling procedure is repeated. Because this approach is equivalent to performing Monte Carlo with an acceptance probability of one, the simulation ensemble distribution should converge on the experimental distribution after sufficient repetitions of the resampling procedure.

The two components of this method are demonstrated separately in Figure 3.2. In Figure 3.2A, we show how stochastic resampling converges on a complex target distribution demonstrated using a Gaussian stub in place of the biased MD. Figure 3.2B shows the results of an MD simulation biased to a single target. The biasing potential successfully drives the simulation distance to the target distance without disrupting secondary structural elements experimentally known to be preserved.[37] Details of the biased MD, implemented using the gmxapi package,[77] are provided in Appendix B.1.2.

This method is trivially extensible to multiple distributions because resampling can be performed on a joint distribution. If information on the correlation structure of the distributions is unavailable, they are assumed to be independent, and draws are performed on the convolution of the distributions. Our approach is especially powerful because information about correlation structure can be recovered from the ensemble; because the coupling constants are first trained using a maximum-entropy formalism, we can measure the work needed to drive the ensemble to its target distances. This quantity reports on the correlation between the particular distributional modes that have been sampled (see Chapter 4). However, in this Chapter we focus on the effectiveness of the fundamental method.

We have used the BRER methodology to refine the conformational ensemble of the soluble domain of syntaxin1-a using previously published DEER data.[37] We find that

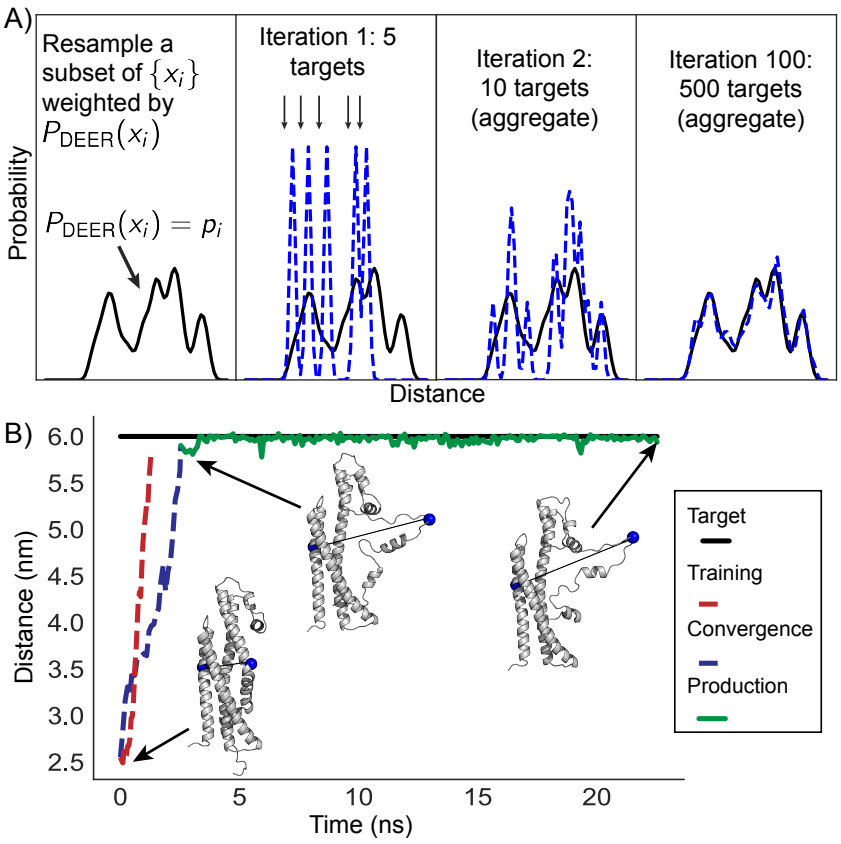

FIGURE 3.2: **Bias-resampling ensemble refinement has two components: stochastic resampling and a maximum-entropy biasing potential.** Iterative stochastic resampling of the target distribution yields an excellent approximation after 500 targets have been drawn (A). Here, the more complex MD engine has been replaced with a Gaussian stub such that each ensemble member samples a simple Gaussian distribution around its target. An example of the maximum-entropy biasing potential for a single target is shown in panel B. First, a maximum-entropy coupling constant is trained, which ensures that a minimally perturbative quantity of energy is introduced into the system. The simulation is then restarted using the pretrained coupling constant and converges to its target without excessively disrupting the secondary structure of the biased conformation.

BRER substantially outperforms current state-of-the-art methods at reproducing the experimental distributions and identifies previously unknown structural substates. These substates suggest that the open state of syntaxin may be more conformationally diverse than previously thought and thus have direct implications for formation of the soluble N-ethylmaleimide-sensitive factor attachment receptor (SNARE) complex.

We obtained three experimentally derived distributions from residue–residue pairs 52/210, 105/216, and 196/228 and used these distributions to refine the conformational ensemble of soluble syntaxin via three different methods: BRER, EBMetaD,[78] and restrained-ensemble MD.[61] The BRER-derived distributions reproduce the experimental distributions significantly better than either EBMetaD or restrained-ensemble (Fig 3.3A), quantified by Jensen–Shannon divergence (Fig 3.3B). BRER performs particularly well at reproducing the 52/210 distribution; the well-separated bimodal peaks in this distribution are important because they directly report on syntaxin's open–closed equilibrium.[37,79–81] Because EBMetaD and restrained-ensemble MD suffer from numerical instabilities in regions of zero probability, these methods fail to accurately reproduce this distribution. EBMetaD samples the open and closed states but with incorrect relative probabilities, and restrained-ensemble simulations simply fail to sample the open state. The 105/216 and 196/228 distributions pose a less challenging problem for EBMetaD and restrained-ensemble MD because neither distribution has very well-separated modes, yet they are still better reproduced by BRER. Thus, BRER is a top-performing, general method for refining conformational ensembles using DEER data.

Refinement of the syntaxin conformational ensemble with these pairs yields a previously unobserved family of structures that are partially open. It is generally thought that syntaxin must be in an open state to participate in formation of the SNARE complex and thus perform its critical role in neuronal exocytosis.[34,80–83] No experimental structures of the apo syntaxin open state exist, but it has been hypothesized that the open state is characterized by complete dissociation of the H3 domain from the Habc domain and an unwinding of the linker region between Hc and H3 (Fig 3.4A).[35,38,39] The BRER-refined ensemble identifies additional structures in which the H3 domain is only partially dissociated from Habc and the linker region retains is secondary structure (Fig 3.4B), with tight contacts remaining between residues 146–156 and 187–198. These BRER-refined structures are in close agreement with the DEER-derived distributions, as are structures in which H3 completely dissociates (Fig 3.4C). These results suggest a testable hypothesis: the syntaxin conformational ensemble is more diverse than was previously thought, and both the partially open and fully open states contribute to the ensemble. In this scenario, formation of the SNARE complex could result from a further conformational selection process. Additional DEER experiments informed by the BRER-refined ensemble could elucidate whether the open-state ensemble is indeed conformationally diverse and whether a conformational selection process takes place to

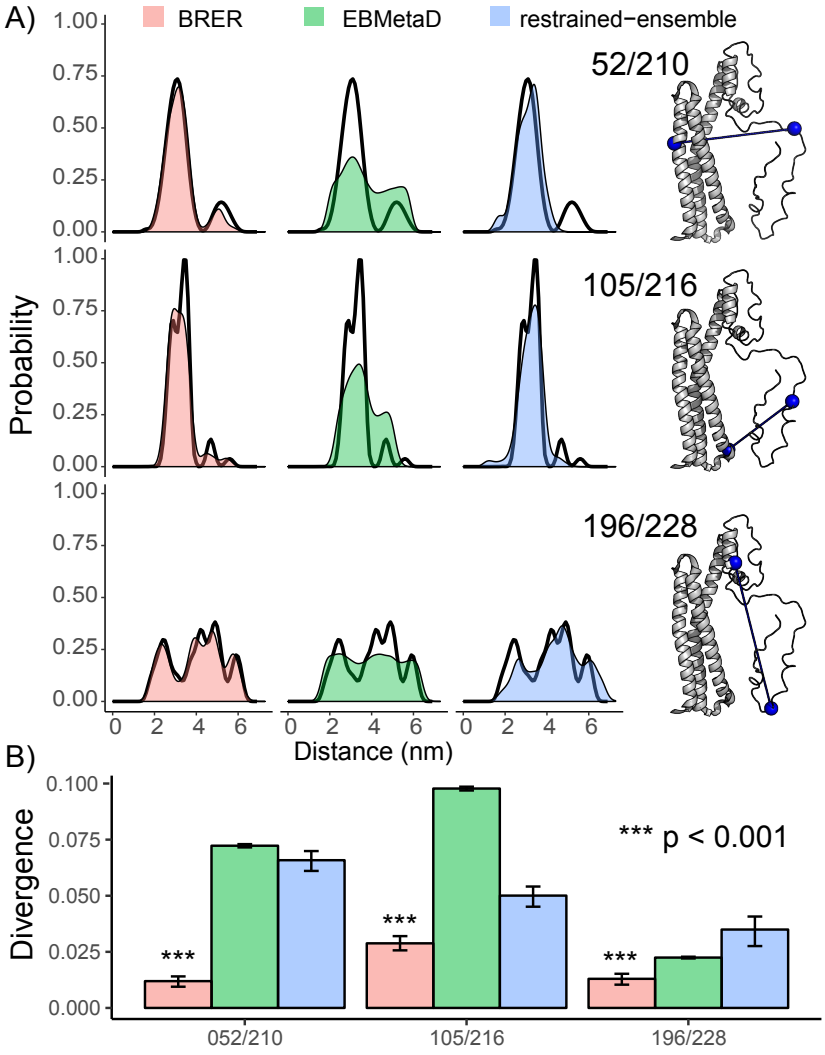

FIGURE 3.3: **BRER outperforms current state-of-the-art methods for refining conformational ensembles against DEER distributions.** Three sets of ensembles were refined against the three experimental distributions (shown in black in panel A). Distributions calculated from the BRER, EBMetaD, and restrained-ensemble conformational estimates are shown in color. BRER both qualitatively (A) and quantitatively (B) outperforms these other state-of-the-art methods for all three distributions. Agreement with the experimental distributions is quantified as Jensen–Shannon divergence (B).

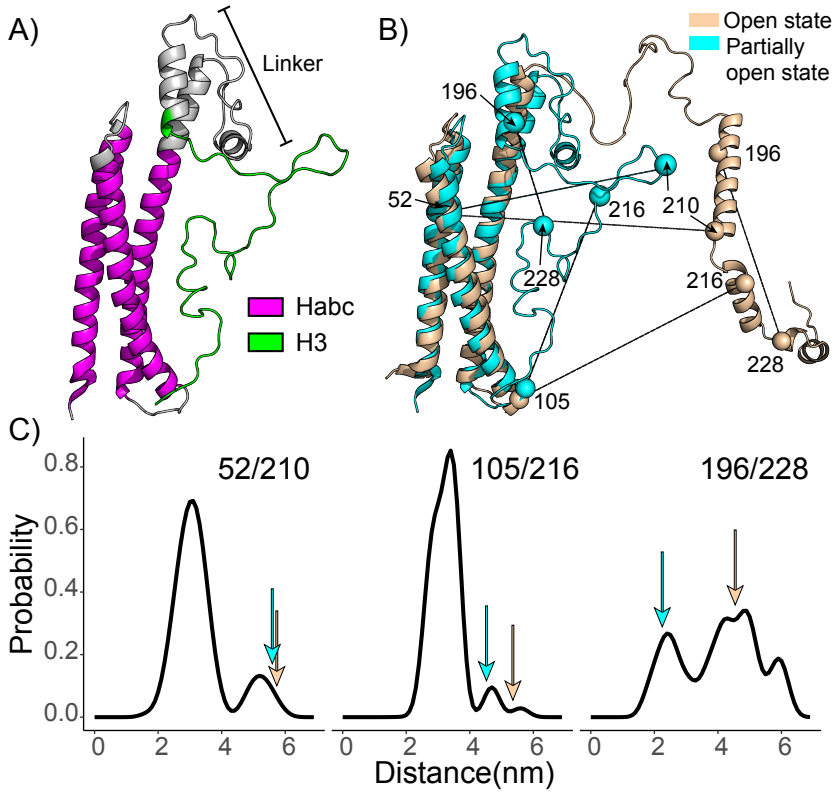

FIGURE 3.4: **BRER-guided refinement yields previously unresolved conformational substates.** Rendered in panel A is a new partially open conformation of syntaxin with key regions labeled. This conformation retains its compact structure, and part of the H3 domain remains in contact with Habc. Rendered in panel B is an overlay of this partially open state with a prior model for the open state of syntaxin, which is characterized by unwinding of the linker region and complete dissociation of H3 from Habc (B). Plotted in panel C are residue–residue distance distributions with values from the open and partially open states indicated by arrows. Both sets of distances agree with the DEER data, suggesting that the syntaxin open state may be more conformationally diverse than previously thought.

form the final SNARE bundle.

Preliminary refinement of the syntaxin conformational ensemble illustrates the importance of heterogeneous ensembles and of having refinement methods that treat them rigorously. Because bias-resampling ensemble refinement is explicitly designed to refine highly heterogeneous conformational ensembles, it significantly out-performs current refinement methods in estimating such ensembles from distributional data. Furthermore, bias-resampling ensemble refinement could be combined with other methods (such as metaynamics,[84] Rosetta,[50,85,86] or other non-MD sampling) to improve their treatment of heterogeneous data. Applied to syntaxin, where the DEER data reveal substantial heterogeneity, our new method uncovers previously unreported heterogeneity in the syntaxin open state. These new conformations may play an important role in mechanisms of SNARE complex assembly.

3.2 Further application of BRER methodology: solving the correlation structure of separately-measured distributions

When no information is available on the correlation structure of separately-measured DEER distributions, the methods developed in both the preceding and remaining chapters become extremely useful. The mRMR experiment-selection method of Chapter 2 is explicitly designed to select minimally-redundant (i.e., minimally correlated) measurements. This selection procedure thus mitigates the error introduced when independence is assumed. In the following chapter, a rigorous approach is developed which *explicitly calculates* the joint distribution of separate measurements. When combined, these methods should provide a powerful way to estimate conformational ensembles from sparse, separately-acquired experimental measurements.

Chapter 4

Solving the correlation structure of separately-measured distributions

Flexible proteins play a critical role in a wide variety of cellular processes, including flexible recognition events during infection and in signal transduction pathways, and this flexibility is essential to biological function.[1–7] Experimental methods that have traditionally been used to study the structural ensembles of biological systems, like X-ray crystallography and nuclear magnetic resonance (NMR) spectroscopy, tend to reduce the ensemble to just a few low energy states in order to achieve high-resolution structures. As awareness has increased of the fundamental role structural heterogeneity plays in biological function, new experimental methods have been developed to report directly on full ensembles. Methods such as double electron-electron resonance (DEER) spectroscopy and single molecule Förster resonance energy transfer (smFRET) provide distance distributions between labeled amino acids, and thus yield quantitative information on conformational populations in a sample. However, these methods come with an important set of challenges, described below.

Label-based experiments that yield distributional data are severely restricted in the number of labels that can be measured simultaneously, leading to two major limitations: since each distribution requires a separate, time-consuming experiment, the data tend to be sparse over atomic coordinates, and the data do not provide information on the joint distribution. Recent efforts have ameliorated the former limitation by optimizing label placement to ensure maximally informative measurements,[8,9,12] but little progress has been made in handling the latter. Here we present a general method for inferring joint probability distributions from separately-acquired measurements. The method not only estimates the correlation structure of the experimental distributions, but also provides a direct way to infer the conformational ensemble of interest.

We first lay out the theoretical underpinnings of the approach, then apply the method to a model of an alternating-access transporter. In the case of the alternating transporter, where the joint distribution is known, we find that our method accurately reproduces

the joint distribution and correctly estimates the true conformational ensemble. Although we have chosen to demonstrate the approach using a toy model, the method is general enough to be applied to any system for which distributional data can be obtained.

Let us denote a set of separately measured probability distribution functions $\{p(O_i)\}$, where O_i is a random variable representing the observable of interest. In this convention, particular values of O_i are denoted o_i. In the applications presented later, each $p(O_i)$ is a single DEER distribution and O_i is the distance variable of the i^{th} pair of atoms. We wish to estimate not only the joint probability distribution $p(O_1, O_2, \ldots, O_N)$, but the conformational ensemble $\{X\}$ that optimally reproduces the joint distribution. This inference problem can be stated in terms of conditional probabilities: what is the probability of an ensemble $\{X\}$ given a set of observed distances, i.e., what is $p(\{X\}|O_1, O_2, \ldots, O_N)$? The joint probability distribution is proportional to the free energy difference of the desired ensemble from some (arbitrary) reference ensemble:

$$p(\{X\}|O_1, O_2, \ldots, O_N) \propto e^{-\beta \Delta G(\{X\}|O_1, O_2, \ldots, O_N)} \tag{4.1}$$

If each random variable O_i can take on values $\{o_i\}$ with probability $p(o_i)$, then the probability of observing a particular conformation given a specific set of distances is trivially:

$$p(x|o_1, o_2, ..., o_N) \propto e^{-\beta \Delta G(x|o_1, o_2, ..., o_N)} \tag{4.2}$$

The challenge then lies in determining the free energy landscape ΔG as a function of the experimental observables. In some cases, it may be possibly to calculate this free energy analytically or via thermodynamic integration, but in general, it is prohibitively expensive to directly calculate the equilibrium free energy because of the large number of degrees of freedom and the slow relaxation timescales involved. Instead, the most robust and general method for calculating this free energy is via the Jarzynski equality: $e^{-\beta \Delta G} = \langle e^{-\beta W} \rangle$.[87] In the remainder of this section, we detail how to leverage the Jarzynski equality and the experimental data to estimate the free energy landscape.

We previously developed a methodology, bias-resampling ensemble refinement (BRER), to incorporate distributional data into molecular dynamics (MD) simulation.[26] The original method assumes that all $\{p(O_i)\}$ are independent, but a simple extension of this formalism enables estimation of the joint distribution. The original BRER method is an iterative approach as follows:

1. randomly sample a conformation x from the current ensemble estimate $\widehat{\{X\}}$
2. select a set of observables $\{O_1 = o_1, ..., O_i = o_i, ..., O_N = o_N\}$ via probability-weighted draws from the experimental distributions $\{p(O_i)\}$.
3. run a biased MD simulation to constrain the conformation x such that all $O_i = o_i$

4. update the estimate $\widehat{\{X\}}$ with the final conformation x

The method is trivially parallelized by drawing multiple conformations $\{x\}$ in a single iteration and applying the constraints to each x.

To estimate the free energy of a set of conformations $\{x\}$ given a set of observables $\{o_i\}$, we can leverage the data from the biased MD runs of step (3). Because we use a simple linear biasing potential, it is trivial to calculate the work done on the ensemble to enforce the constraints. We can thus apply this simple linear bias to drive the system out of equilibrium, then use the nonequilibrium work to estimate equilibrium free energy via Jarzynski's equality. Specifically:

$$e^{-\beta\Delta G(x|o_1,o_2,...,o_N)} = \left\langle e^{-\beta W\left(x^{(eq)}\to x|o_1,o_2,...,o_N\right)}\right\rangle_{x^{(eq)}\in\{X^{(eq)}\}} \tag{4.3}$$

where $\{X^{(eq)}\}$ is an equilibrium ensemble.

The general method for calculating both the joint distribution and the conformational ensemble from simulation, which we call Ensemble Estimation from Separate Measurements (EESM), can be summarized as follows:

1. Draw a set of conformations $\{x^{(eq)}\}$ from an equilibrium ensemble.
2. Select a set of specific observable values $\{o_i\}$ via stochastic draws from each $p\left(O_i\right)$.
3. Apply a linear biasing potential such that $O_i = o_i$ for all $x^{(eq)}$.
4. Calculate the work done in (3) and, consequently, the probabilities $p\left(x|o_1, o_2, ..., o_N\right)$.
5. Repeat 1-4 until the distribution $p\left(\{X\}|O_1, O_2, \ldots, O_N\right)$ has been estimated.

This method is demonstrated below for a simplified alternating-access transporter.

Alternating-access transporters are a class of membrane proteins that transport their substrates by switching between outward-facing and inward-facing conformations.[88–93] In order to test EESM, we studied a toy model of a "flexible" alternating transporter (Fig 4.1A). The transporter consists of two rigid rods connected at their midpoints by a spring with constant α. The rods rotate about their midpoints subject to two constraints: they mirror each other's rotation (the "channel" of the transporter is a symmetry axis) and the angle of rotation θ is constrained to a range $[\theta_{\text{min}}, \theta_{\text{max}}]$. For a given channel width x, all permitted values of θ have equal energy, while those outside the permitted range have infinite energy.

We can imagine performing three separate experiments on the transporter to try to estimate its conformational ensemble: one that measures the width of the channel midpoint ℓ, one that measures the distribution of the "inward-facing" mouth of the channel (D_1 of Fig 4.1A), and one that measures the "outward-facing" mouth of the channel (D_2

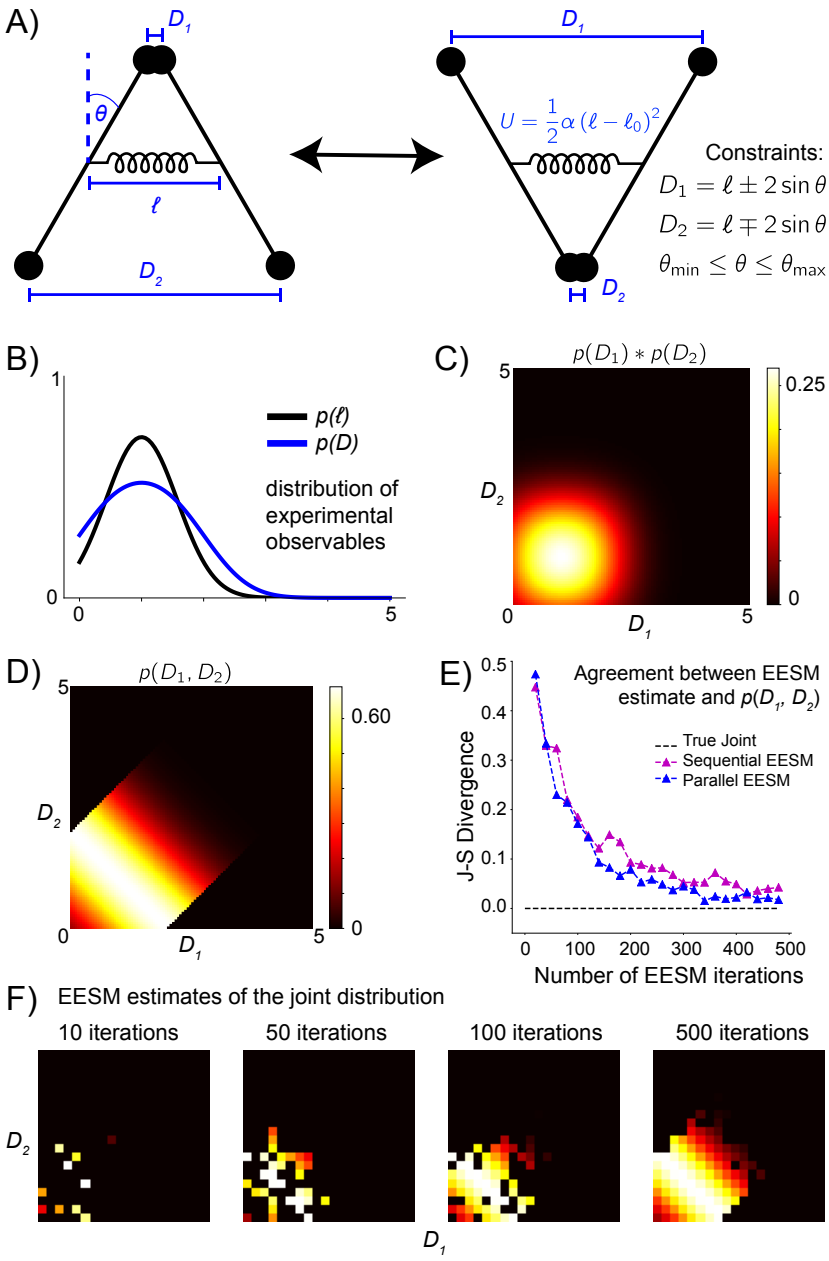

FIGURE 4.1: **EESM accurately reproduces the joint probability distribution for a toy alternating-access transporter.** A toy model of a "flexible" alternating-access transporter is schematized in (A). Experimental measurements of D_1, D_2, and ℓ would yield the distributions shown in (B). Because of the transporter's symmetry, the distributions $p(D_1)$ and $p(D_2)$ are identical and are plotted as $p(D) = p(D_1) = p(D_2)$. Assuming that the two distance distributions D_1 and D_2 are uncorrelated would yield the joint distribution in (C) as opposed to the true joint distribution in (D). Using EESM, we build a stochastic estimate of the true joint distribution $P(D_1, D_2)$ over 500 iterations of sampling performed sequentially and in parallel. Agreement between the estimate and true distribution is quantified as Jensen-Shannon divergence in (E); examples of the estimates over multiple iteration numbers are shown in (F).

of Fig 4.1A). The results of these “experiments" are shown in Fig 4.1B. Without any additional information, we would assume that the separately measured variables D_1 and D_1 are independent and we would estimate the joint probability distribution as shown in Fig 4.1. However, because of the constraints imposed on the channel, the true joint distribution is dramatically different (Fig 4.1D).

In order to estimate the true distribution from only the experimental observables, we performed 500 aggregate iterations of EESM. As the number of iterations increases, the estimate of the joint distribution approaches the true distribution. This is quantified as Kullback-Liebler divergence in Fig 4.1E and illustrated as plots of the joint distribution in Fig 4.1F. This simple but powerful example demonstrates that the method can indeed recover the correlation structure of separately-measured distributions.

We have thus developed a method, ensemble estimation from separate measurements, that can be used to infer the joint distribution of separately-acquired measurements and the conformational ensemble which optimally reproduces that distribution. The method was tested on a simplified model of an alternating-access transporter. We found that EESM converged to the correct distribution within relatively few iterations (Fig 4.1E), confirming that EESM can be used to calculate the correlation structure of separately-measured distributions. EESM is particularly designed to estimate the conformational ensembles of systems where it is impossible to obtain a ground truth ensemble and joint distribution, as is the case for the SNARE protein syntaxin-1a. The application of EESM to syntaxin-1a is presented in Chapter 5.

Label-based measurements that provide pair-wise distributions are an incredibly useful source of experimental data on heterogeneous ensembles. However, the utility of these measurements has been limited because each label pair must be introduced and measured separately. EESM can greatly improve our ability to leverage these separate experiments to refine complex, flexible conformational ensembles by successfully inferring their correlation structure.

Methods

Analysis of the toy alternating-access transporter

To validate the EESM approach, we performed 500 iterations of the method on the toy model shown in Fig 4.1 and compared the estimated joint distribution with the true distribution. The true distribution was calculated analytically as described below in *Analytical methods*. Details of EESM simulations follow in *EESM estimation of the joint distribution*.

Parameter	Value
α	$3\,k_BT$
x_0	1
Length of each rod	2
θ_{min}	$-\pi/6$
θ_{max}	$\pi/6$

TABLE 4.1: **Parameter choices for toy model.** All distance values are unitless. Angles are given in radians.

Analytical Methods

We analytically calculated the probability distributions shown in Fig 4.1B-D using the parameters in Table B.1 and the following equations. The probability of observing a particular channel width is given by

$$p(\ell) \propto e^{-\frac{3}{2}(\ell-1)^2} \tag{4.4}$$

and the probability of observing a particular distance d at either mouth of the channel given a particular channel width ℓ is:

$$p(d|\ell) \propto \begin{cases} e^{-\frac{3}{2}(\ell-1)^2} & \ell - 1 \leq d \leq \ell + 1 \\ 0 & \text{otherwise} \end{cases} \tag{4.5}$$

To calculate $p(d)$, we integrate over the channel widths. The joint probability distribution shown in Fig 4.1C is then the convolution of Equation 4.5 with itself.

The probability of observing a pair of distances (d_1, d_2) given a channel width ℓ is:

$$p(d_1, d_2|\ell) = p(d_1|d_2, \ell)p(d_2|\ell)p(\ell) \tag{4.6}$$

The symmetry constraint determines the probability $p(d_1|d_2, \ell)$:

$$p(d_1|d_2, \ell) \propto \begin{cases} 1 & d_1 = d_2 + 2\sin\theta \\ 0 & \text{otherwise} \end{cases} \tag{4.7}$$

From combining equations 4.4-4.7 and marginalizing over the channel widths, we can calculate the joint probability distribution in Fig 4.1D.

EESM estimation of the joint distribution

To estimate the toy model's joint distribution, we iteratively:

1. Selected a set of distances (d_1, d_2) by sampling from the distribution shown in Fig 4.1C.

2. Determined the conformation c associated with (d_1, d_2) using the constraints.

3. Calculated the work required to drive the conformation of the previous iteration to the conformation of (2).

When the EESM iterations are performed sequentially, the work is a function of the previous iteration's initial state and the final state:

$$W(x_i \to x_f) = \frac{1}{2}\alpha \left(\ell_f - \ell_i\right)^2$$

If the iterations are performed entirely in parallel and the initial conformation is drawn from equilibrium, then the work is simply

$$W(x_i \to x_f) = \frac{1}{2}\alpha \left(\ell_f - \ell_0\right)^2$$

Once we obtained a set of work estimates from this resampling procedure, we calculated the free energy as a function of the distance variables D_1, D_2 using the Jarzynski equality. To obtain the free energy $\Delta G(\{x\}|D_1 = d_1, D_2 = d_2)$, we average the exponential work values over all trials where x_f has $D_1 = d_1, D_2 = d_2$:

$$e^{-\beta \Delta G(\{x\}|d_1,d_2)} = \left\langle e^{-\beta W(x_i \to x_f|d_1,d_2)} \right\rangle$$

After approximately 500 iterations, both the sequential and parallel procedures converge on the analytically determined joint distribution (Fig 4.1F).

Chapter 5

Conclusions and future directions

5.1 Review: iterative refinement of flexible systems

5.1.1 Selection of experiments

We have demonstrated a method for selecting label locations to obtain distance distributions that optimally refine the conformational ensembles of flexible proteins (Chapter 2). 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 tested on three flexible bacterial outer membrane proteins, then prospectively validated for refinement of Opa_{60} Neisserial virulence-associated protein. We are actively using this method to refine bimolecular complexes in flexible molecular recognition such as occurs in the binding of CEACAM receptors by Opa_{60}.

In the case of Opa_{60} engagement of CEACAM1, we identified a set of loop conformations that account for the *apo* conformational ensemble. Further DEER measurements indicated a conformational selection event upon CEACAM1 binding, and we have shown that HV2-extended Opa conformations are the only ones consistent with the CEACAM1-engaged complex (Fig 2.5). Previous mutational data showed that specific HV1/HV2 loop sequence combinations were required for CEACAM1 engagement.[62] 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_{60} conformations recognized by CEACAM1 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 described in Section 2.1. 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 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 (Fig 2.3, A.5). Finally, incorporation of the distributions of mRMR pairs leads to improved refinement of the conformational ensemble by reducing the effective dimensionality of the ensemble (Fig 2.4, A.6). Ultimately, the same information can be obtained about an ensemble with significantly fewer spectroscopic measurements.

5.1.2 Incorporation of distributional data into estimates of conformational ensembles

We have demonstrated a method, bias-resampling ensemble refinement (BRER), for integrating heterogeneous, distributional data to refine the conformational ensembles of flexible ensembles (Chapter 3). The method was tested using DEER-derived distributions, but is sufficiently general to be used for any non-parametric data. We tested the method on the SNARE protein syntaxin-1a and compared it to two other state-of-the-art methods for data integration, EBMetaD[78] and restrained-ensemble MD.[61] BRER substantially outperformed both alternate methods at reproducing the experimental distributions (Fig 3.3).

BRER also enabled identification of previously unresolved set of open-state conformations that could elucidate the nature of SNARE assembly and function. The H3 SNARE-binding motif of syntaxin is known to be well-structured when in complex with the other SNARE proteins,[34,80–83] yet little is known about the unbound open state ensemble.[35,38,39] Our data suggest that the open state is actually quite heterogeneous and that the H3 domain is unstructured. This open-state heterogeneity suggests that SNARE complex formation may be the result of a conformational selection event: during assembly, conformations which retain a helical H3 domain are preferentially selected to form the final SNARE.

We are currently testing for the presence of these new open-state conformations using the mRMR algorithm developed in Chapter 2. We selected a set of pairs based on their ability to distinguish between the canoncial open state and the novel open state (Fig 3.4B). Measurements of these new pairs should reveal whether this new open state is present in the solution ensemble.

High-resolution refinement of the syntaxin conformational ensemble will likely require more sophisticated treatment of the correlation structure of the DEER distributions. Thus, in addition to measuring new mRMR-selected residue-residue pairs to confirm the presence of this new open state, we also developed a method to improve the estimate of the conformational ensemble by inferring the joint distribution of separate measurements.

5.1.3 Inferring joint distributions from separately-acquired measurements

We have demonstrated a method, ensemble estimation from separate measurements (EESM), that can be used to infer the joint distribution of separately-aquired measurements and improve the conformational ensemble estimate using that distribution (Chapter 4). The method was tested on a simplified model of an alternating-access transporter and converges to the correct joint distribution within relatively few iterations (Fig 4.1E). Here, the conformational ensemble and the joint distribution are known a priori, so it is possible to rigorously compare the EESM estimate with ground truth. In cases where no ensemble estimates exist, such as with syntaxin, we can evaluate the method based on two criteria: its convergence behavior and its ability to predict DEER measurements not used for refinement. This is discussed in detail in Section 5.2.

5.1.4 Summary

Together, the methods developed in Chapters 2-4 can be used to refine the conformational ensembles of biological systems that have challenged even the most sophisticated refinement procedures. We have tested the methods by refining the *apo* Opa_{60} ensemble and the soluble domain of syntaxin-1a, two systems which had otherwise been nearly impossible to characterize. In the future, the iterative approach shown in Figure 5.1, which leverages all three methods, may be used to refine the ensembles of significantly more complex systems: the Opa_{60}-CEACAM1 interaction and membrane-bound syntaxin-1a. Specific suggestions for refinement of these systems are described below.

5.2 Future directions

5.2.1 Refinement of the Opa_{60}-CEACAM ensemble

We have shown in Chapter 2 that the Opa_{60} likely engages CEACAM through a conformational selection process. Restrained-ensemble simulations performed with a set of five high-scoring mRMR pairs revealed a striking pattern of loop-loop interactions: in 60% of the resulting conformations, a single loop extended laterally from the base of the barrel while the two remaining loops formed multiple contacts. Further DEER experiments performed with and without CEACAM revealed that Opa engages its receptor

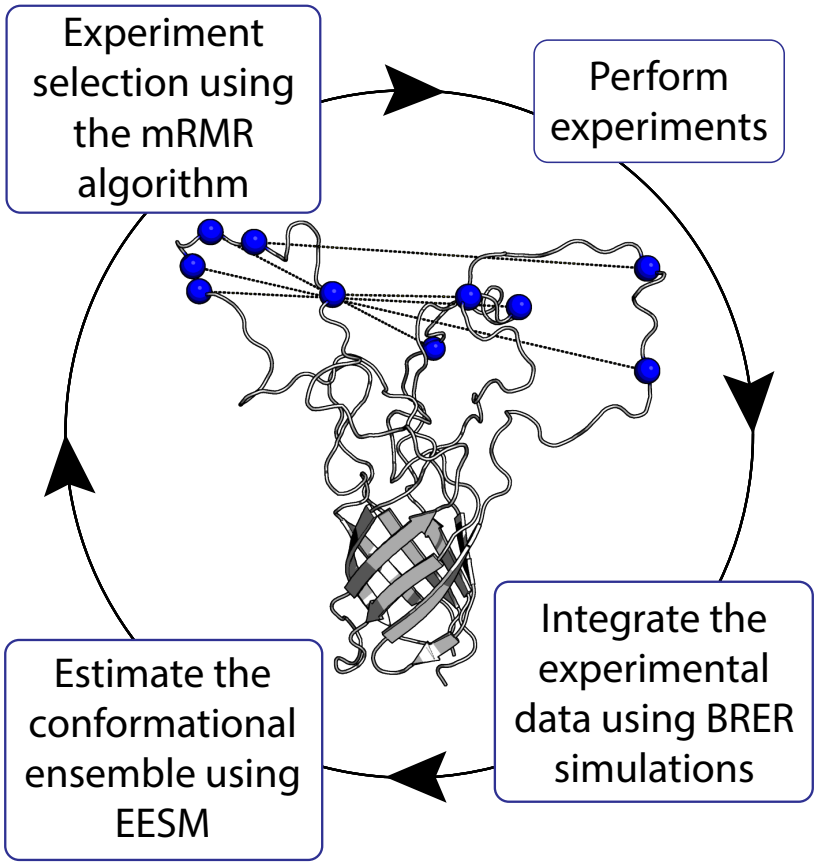

FIGURE 5.1: **Schema for refining heterogeneous conformational ensembles using mRMR, BRER, and EESM.** Data from MD simulations are used to select optimal experiments using the mRMR algorithm (Chapter 2). Data from these experiments are then incorporated into MD simulation using BRER (Chapter 3). EESM (Chapter 4) is used to infer that correlation structure of separate measurements and estimate the conformational ensemble.

through a subset of these particular conformations (Fig 2.5). However, it remains unknown whether Opa engages CEACAM through the single extended loop or through the interface formed by the contacting loops.

In order to further refine the conformational ensemble of Opa bound to CEACAM, we will perform additional DEER experiments using Opa_{60}-CEACAM1 pairs. These experiments will be selected via the mRMR algorithm on a set of unrestrained MD simulations of Opa_{60}:CEACAM1 using HV2-extended conformations as initial states. These pairs will be measured and incorporated into subsequent round of MD simulation using BRER simulations. Convergence of the ensemble can be estimated using cross-validation among DEER pairs not used in the refinement. Loop flexibility will be estimated using time-autocorrelation of the loop residues.

The proposed hybrid-refinement method and analysis will illuminate whether or not the Opa loops remain conformationally flexible upon binding to CEACAM. This would provide significant evidence in support of the hypothesis that Opa remains conformationally flexible in order to reduce the entropic penalty of binding (Fig 1.3).

5.2.2 Further refinement of the syntaxin-1a soluble domain and membrane interation

Although structures exist for the closed state of syntaxin,[81,94–96] few definitive structural data exist on the open state, and even fewer data provide conclusive insight into exactly how the conformational equilibrium of these states shifts in the presence of a membrane. Where data do exist, it has been particularly difficult to integrate the information since it has been acquired for different constructs of the protein. DEER data taken on just the soluble domain of syntaxin-1a indicate that the H3 domain is disordered in solution, yet NMR and FLIC data acquired in the presence of a membrane suggest that H3 is ordered near a membrane.[35,37,94] The BRER methodology developed in Chapters 3 and 4 is especially well-suited to integrate these data. We can use it to determine the syntaxin membrane interaction as follows.

We will first improve the conformational estimate of the syntaxin soluble domain using the method developed in Chapter 4. We can infer the correlation structure of the three separately-aquired DEER distributions which we used to obtain a prelimary estimate of the soluble ensemble (Chapter 3). We can reweight this ensemble estimate using the estimated joint distribution. Because no experimentally-validated joint distribution exists a priori, we will evaluate the EESM approach based on two criteria: its convergence behavior and its ability to predict DEER measurements not used for refinement. We anticipate that EESM will converge smoothly to a final estimate of the joint distribution that is distinct from the convolved distributions, as in the case of the

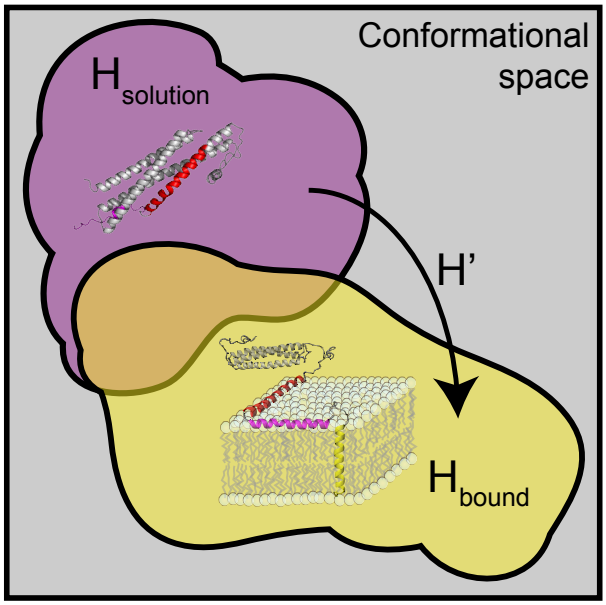

FIGURE 5.2: **Sequentially-trained BRER for determining the membrane-bound syntaxin-1a conformational ensemble.** Sequentially-trained BRER first learns the solution ensemble using DEER data, then the membrane ensemble using FLIC data, by minimally perturbing the solution ensemble with some perturbing Hamiltonian term H'.

alternating-access transporter. Most importantly, we expect that the EESM-refined ensemble will better approximate additional DEER distributions than an ensemble refined by assuming independent measurements.

Once we have obtained an improved estimate of the solution ensemble, we can leverage this estimate to elucidate the syntaxin membrane interaction. Because BRER obeys the maximum-entropy principle,[17,26,75] the solution ensemble incorporates experimental DEER data with minimal perturbation to the MD Hamiltonian. The FLIC measurements may then be incorporated via a subsequent round of BRER and EESM; these newly integrated measurements will minimally perturb the solution ensemble's Hamiltonian. With sufficiently many iterations, sequentially-trained BRER will yield an estimate of the membrane conformational ensemble that is "most compatible with," or minimally perturbed from, the solution ensemble (Fig 5.2). The quality of that refinement can then be determined by how well the estimated membrane ensemble predicts FLIC measurements not used for refinement.

Thus, sequentially-trained BRER and EESM leverage one ensemble to refine a distinct but related ensemble and will allow us to compare the solution and membrane-bound ensembles of syntaxin. A refined ensemble will provide insight not only into the general behavior of syntaxin-1a near and far from a membrane, but also how modulation of the conformational ensemble by the presence of a membrane may affect SNARE

complex formation and assembly.

Appendix A

Supplementary material for simulation-guided spectroscopy

A.1 mRMR theory and applications

The contents of this section are published as Supplementary Material to a research article in:

Refinement of Highly Flexible Protein Structures using Simulation-Guided Spectroscopy. Jennifer M. Hays, Marissa K. Kieber, Jason Z. Li, Ji In Han, Linda Columbus, and Peter M. Kasson. *Angewandte Chemie International* 2018 (130) 17356 –17360.

A.1.1 Methods

mRMR-based selection of optimal DEER measurements

For each bacterial protein, we selected residue-residue pairs using the mRMR algorithm on 2 µs MD ensemble simulations per protein. A C_α -C_α distance matrix was calculated using conformational snapshots at 500 ps intervals and these were histogrammed using 1 Å bins. Normalized histograms were used to calculate pair-configuration and pair-pair MI (Eqs 2.1 and 2.2) as follows:

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 residueresidue pairs. If the protein has n residues, there are $(n^2 - n)/2$ possible pairs, and we can define the following conformation variable:

$$\vec{C} = \begin{pmatrix} X_1 & X_2 & \cdots & X_i & \cdots & X_{(n^2-n)/2} \end{pmatrix}$$

where X_i is the distance between the i^{th}pair of residues. For mutual information calculations, these real-valued variables are then binned, such that each conformation variable

TABLE A.1: Ranking of top residue-residue pairs via mRMR and mutual information alone for Opa_{60}.

Rank	mRMR Pairs	mRMR score	MI Pairs	MI score
1	36 171	4.110195	36 171	4.110195
2	91 165	3.203021	37 164	4.109671
3	25 167	3.156535	36 161	4.098168
4	39 158	3.141428	35 171	4.095264
5	85 170	3.105476	36 164	4.094354
6	32 163	3.11782	37 171	4.091122
7	36 91	3.099664	34 169	4.089396
8	34 168	3.121006	37 168	4.089102
9	94 167	3.106743	36 172	4.087823
10	38 154	3.096239	37 165	4.087248
11	85 163	3.105664	37 162	4.087004
12	39 164	3.112969	38 164	4.086218
13	36 175	3.091163	37 166	4.085327
14	30 167	3.094117	36 165	4.08458
15	89 158	3.090069	36 170	4.084023
16	39 171	3.082419	35 166	4.081497
17	91 173	3.084754	37 169	4.080652
18	26 163	3.078114	36 168	4.079957
19	35 95	3.07494	37 161	4.078506
20	35 165	3.089347	36 162	4.078061

is represented as a vector of integers, with each integer being a bin number. We thus have a set of observed conformations $\vec{c}$.

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 conformation c and $P(x_i)$ and $P(c)$ are the marginal probability distribution functions of pair i and conformation c, respectively.

An ordered list of highest-ranking mRMR pairs was then generated using greedy mRMR selection (Table A.1).[51] Code implementing mRMR selection of residues for DEER experiments is available from: https://github.com/kassonlab/mRMR-DEER. The implementation also provides the ability to exclude user-defined residue-residue pairs, such as residues where spin label placement might disrupt function, but that feature was not needed here.

Setup and equilibration of MD simulations

FhuA Because the Ton box motif is highly mobile and thus poorly resolved with NMR and X-ray crystallography, no full-length *apo* structures of FhuA exist. We therefore used a previously published ensemble modeled using NIH-XPLOR to initialize our simulations.[58] This ensemble incorporated a set of MTSL spin-labels. The spin-labels were removed via a homology model with an incomplete *apo* structure (PDB ID 1BY3).[97] The final full-length *apo* structure was inserted into a membrane of 756 DLPC lipids using the Gromacs tool `g_membed`. In order to improve sampling of the heterogeneous Ton box motif, we ran an initial pulling simulation to extend the N-terminal domain into the periplasm. The simulation incorporated four harmonic, pairwise restraints between the C_β of the residue pairs 13-161, 13-228, 13-373, and 13-663. Each residue pair was pulled to a distance of approximately 5 nm over the course of 12 ns. This short simulation time is reasonable since this was intended only to generate initial states. Conformations were then sampled every ns to obtain 12 structures for subsequent unrestrained simulations. Finally, a brief 100 ps equilibration was run on each of the structures using the NPT conditions described in *Production MD simulations* below. The final ensemble consisted of two replicates of these 12 states for a total of 24 ensemble members.

OprG The 20 lowest energy structures previously identified (PDB ID 2N6L) were chosen as initial states. They were inserted into a DLPC membrane as follows: first, CHARMM-GUI was used to equilibrate a single OprG state obtained from the Orientations of Proteins in Membranes (OPM) database.[98] Then, each of the 20 low energy structures was aligned to the β-barrel of this single structure. Each system was solvated independently with approximately 40,000 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 using steepest-descent for 5000 steps or until the largest force was less than 1000 kJ mol nm^2. Finally, a brief 100 ps equilibration was run using the NPT conditions described in *Production MD Simulations* below. Of these initial 20 systems, only six fully relaxed in the membrane; many of the initial loop conformations extend downward into the plane of the membrane and thus are unlikely to be true conformational states.[99] The final ensemble consisted of four replicates of these six states for 24 total ensemble members.

Opa$_{60}$ The 20 lowest free-energy structures of Opa$_{60}$previously identified[30] (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 β-barrel of a single structure from the Fox simulations. The protein and membrane were energy-minimized using the steepestdescent 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 100 ps equilibration was run using the NPT conditions described in *Production MD Simulations* below.

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.

Production MD simulations

All production simulations were performed using a modified version of Gromacs 5.2 available at `https://github.com/kassonlab/reMDgromacs-5.2` and the CHARMM36[100,101] forcefield. Simulations were run under NPT conditions using the velocity-rescaling thermostat 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.[102] Covalent bonds were constrained using LINCS, and long-range electrostatics were treated using Particle Mesh Ewald.[103] For each protein, ensemble simulations were run until a total of 2 µs of data were collected.

Expression, purification, labeling, and refolding of Opa$_{60}$

The opa60 gene was sub-cloned into a pET28b vector (EMD chemicals, Gibbstown, NJ) containing N and C terminal His6 – 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 Co2+ 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 8 M 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 µM to 200 µM. The labeled proteins were rapidly diluted 20-fold into 20 mM TrisHCl, 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.[30,104] Folding efficiency was assessed with SDS-PAGE. Samples were dialyzed against 3×4L of 20 mM sodium phosphate, 150 mM NaCl for an hour each, removing any free spin. Opa proteins were concentrated to approximately 200 µM to 400 µM.

Double electron-electron 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. The spectrometer was connected to a 10 W solid-state amplifier (Bruker AmpQ). All samples were prepared to a final protein concentration between approximately 100 µM to 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 ν_1), and a π pump pulse (at a frequency ν_2) optimized at approximately 32 ns.[105] The pump frequency (ν_2) was set at the maximum of the nitroxide spectrum and the observed frequency (ν_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 h to 24 h, with a dipolar evolution time between 2 µs to 3 µs. Dipolar evolution data were processed using DEERAnalysis2016 software[106] using Tikhonov regularization to generate distance distributions. 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.

Restrained-ensemble biasing potentials

To compare the quality of mRMR-guided versus spectroscopist-guided refinement of Opa_{60}, two ensemble refinements were run. The first incorporated experimental DEER

distance distributions from high-ranking mRMR label pairs 31-166 and 88-162, while the second incorporated those from spectroscopist-selected label pairs 77-107 and 107-117. Restrained-ensemble biasing potentials previously developed by Roux were applied to match MD distance histograms to DEER-derived distance distributions (Fig A.3). Refinement was performed via restrained-ensemble simulation using a modified version of Gromacs 5.2 available at https://github.com/kassonlab/reMD-gromacs-5.2. 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 (SSP). 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 (Fig A.3). An initial spring constant K =10 kJ mol^{-1} 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^{-1} nm^{-2} in the mRMR-guided simulations in order to reverse the increase in J-S divergence observed from approximately 30 ns to 40 ns.

Information-theoretic clustering

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

$$I\left(X_1^{C_\alpha}, X_1^{C_\alpha}\right) = \sum_{\{x_1^{C_\alpha}\}} \sum_{\{x_2^{C_\alpha}\}} P\left(x_1^{C_\alpha}, x_2^{C_\alpha}\right) \log \frac{P\left(x_1^{C_\alpha}, x_2^{C_\alpha}\right)}{P\left(x_1^{C_\alpha}\right) P\left(x_2^{C_\alpha}\right)}$$

Because closely related sets of pairs (high $I\left(X_1^{C_\alpha}, X_1^{C_\alpha}\right)$) 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 C_α -C_α 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 C_α -C_α 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 C_α -C_α pairs using complete-linkage hierarchical clustering with an MI-based distance metric

$$D\left(X_1^{C_\alpha}, X_1^{C_\alpha}\right) = 1 - \frac{I\left(X_1^{C_\alpha}, X_1^{C_\alpha}\right)}{H\left(X_1^{C_\alpha}, X_1^{C_\alpha}\right)}$$

where $H\left(X_1^{C_\alpha}, X_1^{C_\alpha}\right)$ is the joint entropy of the pairwise C_α -C_α distance distributions:

$$H\left(X_1^{C_\alpha}, X_1^{C_\alpha}\right) = \sum_{\{x_1^{C_\alpha}\}} \sum_{\{x_2^{C_\alpha}\}} P\left(x_1^{C_\alpha}, x_2^{C_\alpha}\right) \log P\left(x_1^{C_\alpha}, x_2^{C_\alpha}\right)$$

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 C_α -C_α pairs are grouped together and considered redundant.

The information-theoretic resolution is reported in Fig 2.4 as $1 - \epsilon$, i.e., 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 four compact and well-separated clusters in the subspace formed by the first three principal components (Fig A.9). Similarly, for restrained ensemble simulations performed using SSP DEER data, all snapshots formed five well-separated clusters (Fig A.10). 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.

A.1.2 Additional Figures

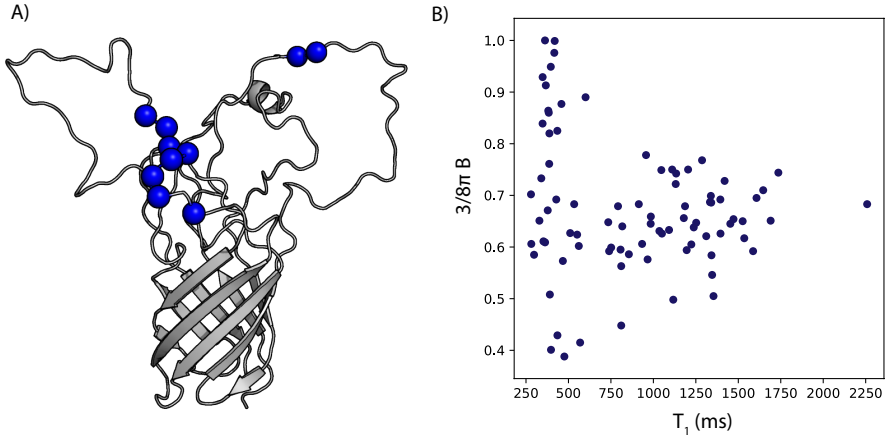

FIGURE A.1: **ENM-based scoring of flexibility correlates poorly with NMR data and identifies less informative loop regions.** Elastic network models provide a computationally efficient means of approximating some protein motions. To assess this approach for Opa loop prediction and DEER pair selection, a Gaussian Network Model was used to predict C_α 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 T1 decays2 (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}.

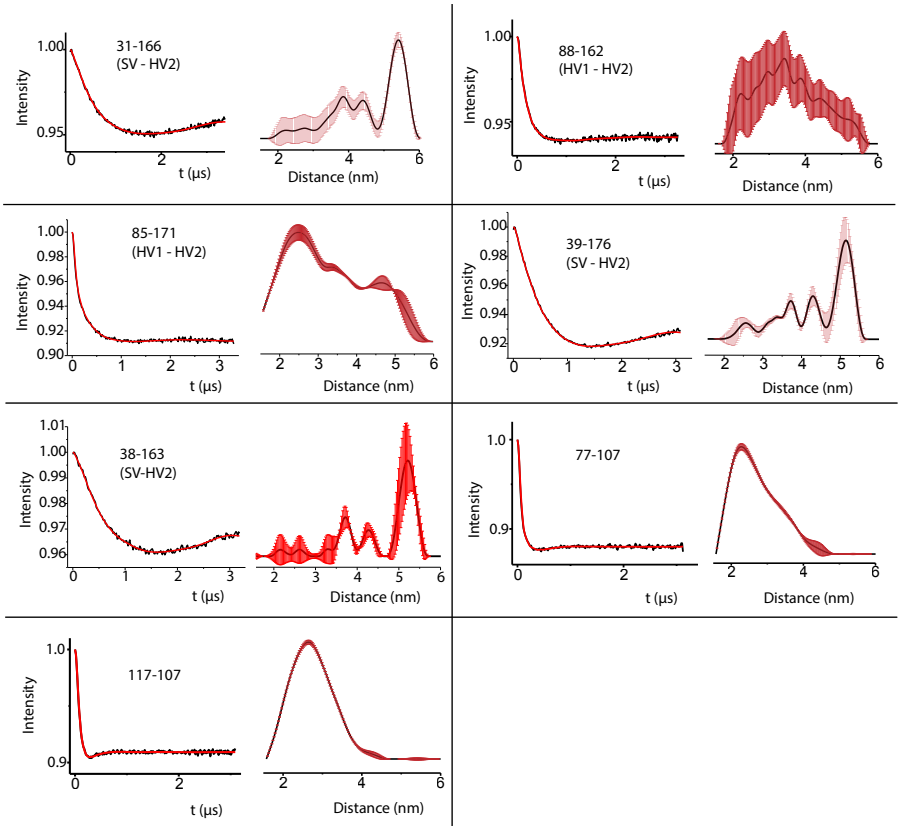

FIGURE A.2: **Measured spin-echo decays and fitted distributions**. Fits are superimposed in red on the decays. The red error bars in the distance distributions represent uncertainty due to the background subtraction form factor that produce fits within 15% RMSD of the best fit.

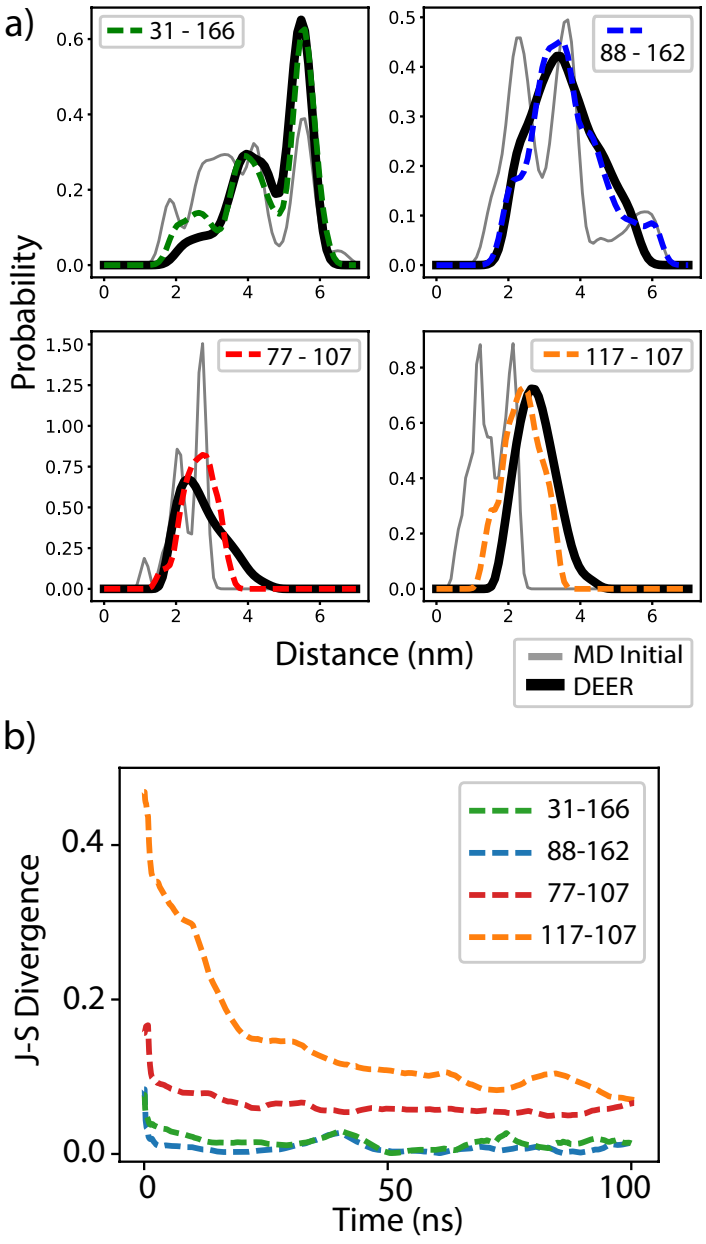

FIGURE A.3: **Restrained-ensemble simulations converge rapidly to experimental distributions.**. Convergence of restrained-ensemble simulations to DEERderived distributions over 100 ns is plotted in (a) for both the high-scoring mRMR pairs and spectroscopist selected pairs. Convergence of both ensembles is quantified in (b) using Jensen-Shannon divergence.

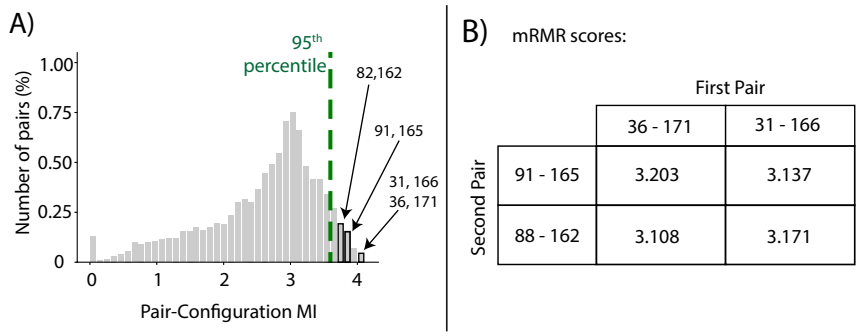

FIGURE A.4: **Top mRMR-predicted pairs and measured pairs have near-identical pair-configuration mutual information and mRMR values.** For operational reasons, the residue-residue pairs measured via DEER were slightly different than the top mRMR-predicted pairs. As shown in the histogram in a) and mRMR 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%.

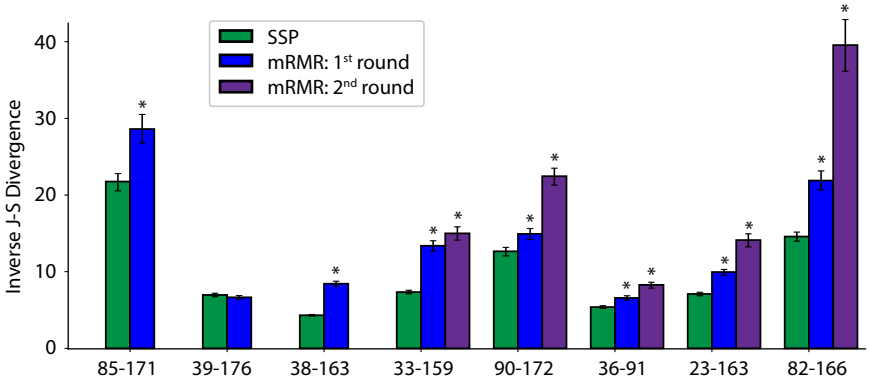

FIGURE A.5: **A second round of mRMR better refines the Opa_{60} conformational ensemble.** Conformational ensembles refined using mRMR-selected pairs predict these new DEER distributions significantly better than conformational ensembles refined using spectroscopist-selected pairs (SSP) in seven of 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 outperforms both 1st-round ensembles in predicting the five pairs not used for refinement. Error bars represent 90% confidence using 1000 bootstrap replicates.

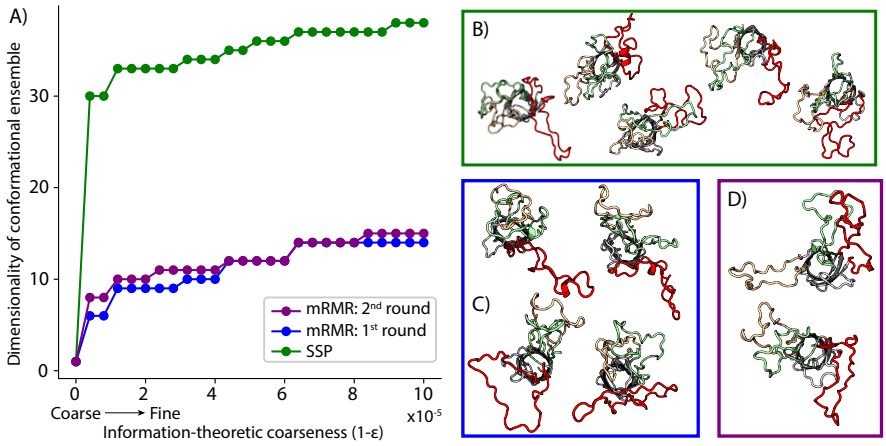

FIGURE A.6: **A second round of mRMR elucidates conformational heterogeneity of "two-and-one" loop configurations.** The same "two-and-one" interaction patterns observed in the first round of mRMR-guided refinement (C) predominate in a second round of refinement (D). The conformational heterogeneity of the two-and-one interaction pattern is better resolved in the second round as evidenced by the additional single-loop extension in (D) and the unchanged dimensionality in (A). Conformational clusters from SSP-guided refinement are shown in (B) for completeness.

A.2 Preliminary refinement of the Opa_{60}-CEACAM interaction

The contents of this section are published as Supplementary Material to a research article in:

Refinement of Highly Flexible Protein Structures using Simulation-Guided Spectroscopy. Jennifer M. Hays, Marissa K. Kieber, Jason Z. Li, Ji In Han, Kelley W. Moremen, Linda Columbus, Peter M. Kasson. *bioRxiv* January 1, 2018, 319335.

A.2.1 Methods

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.[107] 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, 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:

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

subject to the constraints

$$\begin{aligned} \alpha, \beta, \gamma &\geq 0 \\ \alpha + \beta + \gamma &= 1 \end{aligned}$$

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.

A.2.2 Additional figures

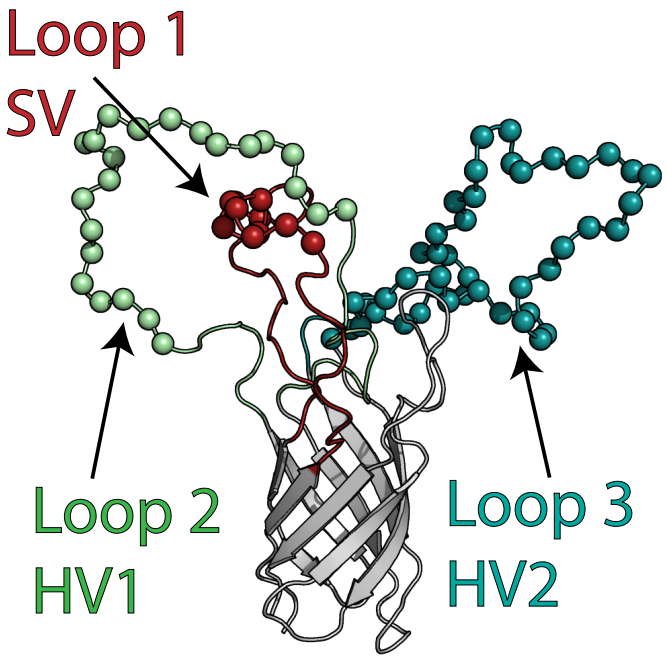

FIGURE A.7: **Loop nomenclature and variable regions of Opa_{60}.** Loops 1, 2, and 3 are shown in red, light green, and teal, respectively, on an Opa_{60}structure. C_α of the hyper-variable regions (HV1 and HV2) as well as the semivariable region (SV1) are rendered as spheres.

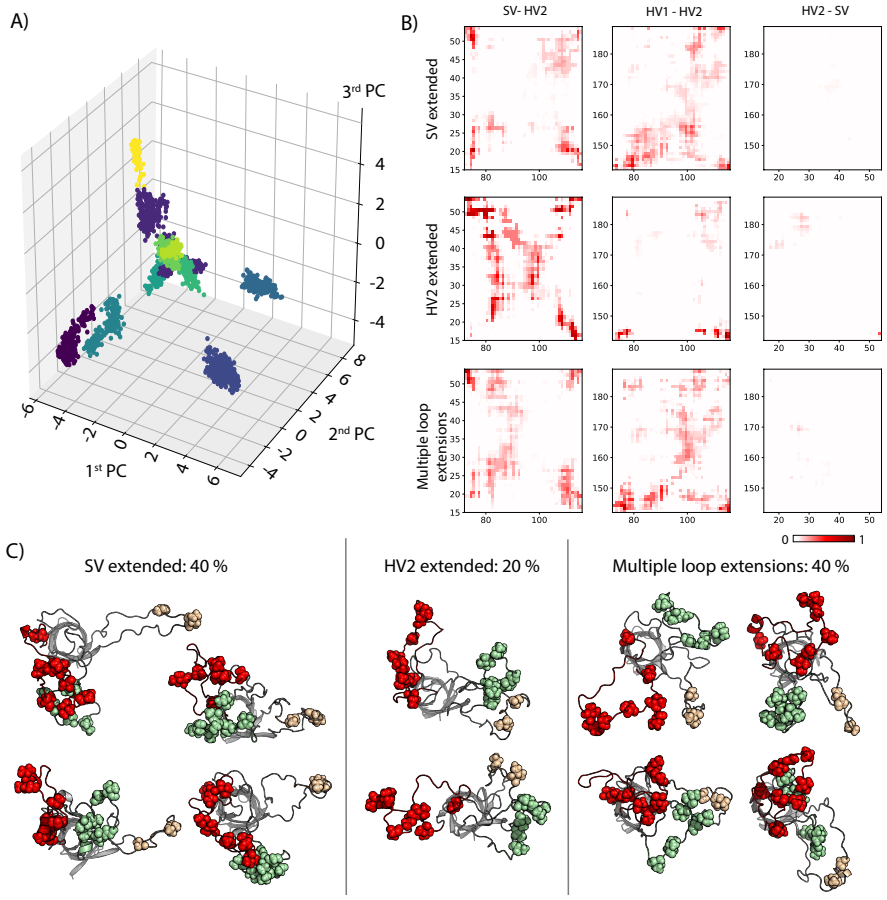

FIGURE A.8: **Loop-loop contact modes identified by second-round refinement using mRMR-selected DEER pairs.** Principal components analysis was performed on loop-loop contact matrices from restrained-ensemble simulations of the second-round mRMR-selected residue pairs. The first three of principal components 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.

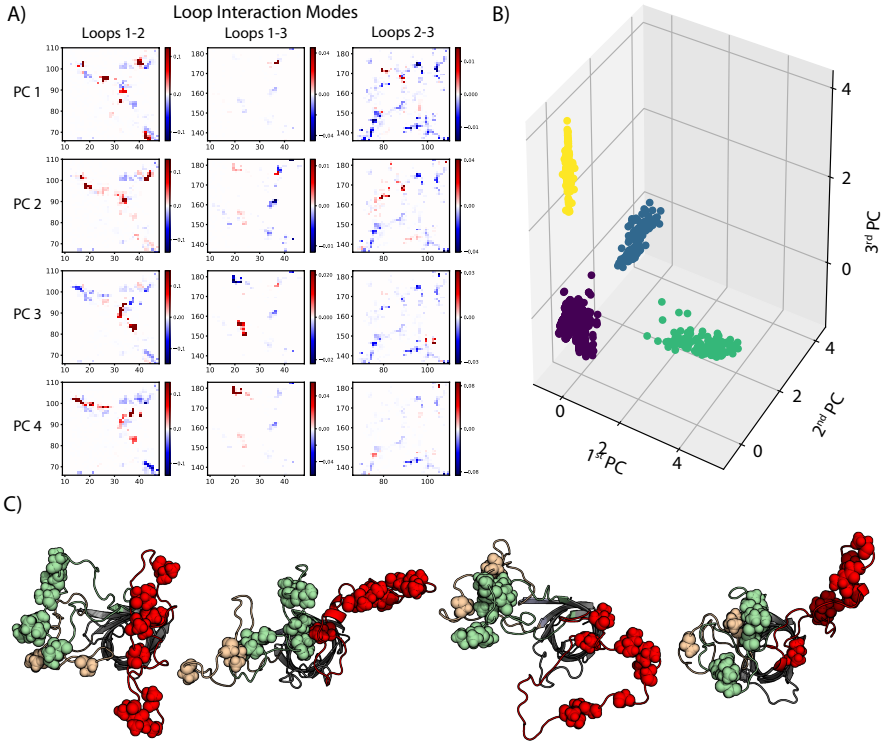

FIGURE A.9: **Loop-loop contact modes identified by first-round refinement using mRMR-selected DEER pairs.** Principal components analysis was performed on loop-loop contact matrices from restrained-ensemble simulations of mRMR-selected residue pairs. The first four principal components, which account 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.

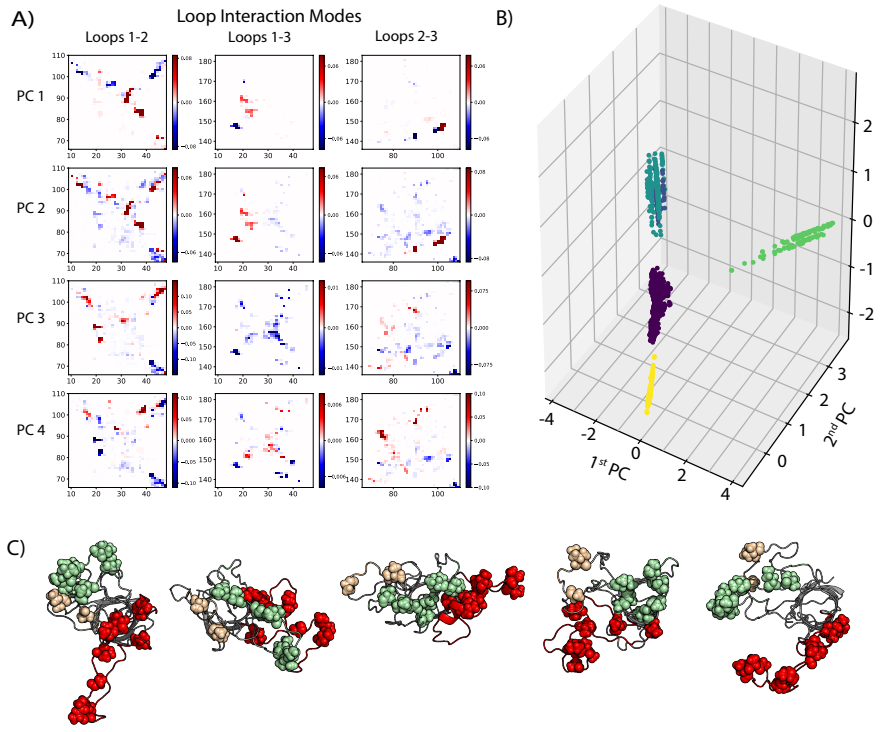

FIGURE A.10: **Loop-loop contact modes identified by first-round refinement of spectroscopist-selected DEER pairs.** Principal components analysis was performed on loop-loop contact matrices from restrained-ensemble simulations of spectroscopist-selected residue pairs. The first four principal components, 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.

Appendix B

Supplementary material for integrating distributional data on heterogeneous ensembles

B.1 Bias-resampling ensemble refinement (BRER): theory and applications

The contents of this section are published as Supplementary Material to a research article in:

Hybrid Refinement of Heterogeneous Conformational Ensembles Using Spectroscopic Data. Jennifer M. Hays, David S. Cafiso, and Peter M. Kasson. *The Journal of Physical Chemistry Letters* 2019 10 (12), 3410-3414.

B.1.1 Theory

A more complex formulation of bias-resampling ensemble refinement (BRER) may be used to perform more advanced sampling. Rather than draw each conformation x from the previous conformational estimate $\widehat{\{X\}}$, a history is maintained of k refinement rounds, so that the conformation x is drawn from the union of $\widehat{\{X\}}_{i-1}, \widehat{\{X\}}_{i-2}, \ldots, \widehat{\{X\}}_{i-k}$. The conformational estimate $\widehat{\{X\}}$ is then obtained as before: the conformations are updated using a biased MD simulation such that the updated estimate $\widehat{\{X\}}_{1\ldots i}$ will optimally reproduce $P_{\mathrm{DEER}}(d)$. Just as with the formulation provided in the main text, over the course of multiple rounds of refinement, the conformational estimate $\widehat{\{X\}}$ should yield a distribution $P_{\widehat{\{X\}}}(d)$ that converges on $P_{\mathrm{DEER}}(d)$.

B.1.2 Methods

Molecular dynamics simulations

Set up and equilibration of syntaxin-1a In order to best demonstrate the ability of our method to sample backbone conformational change and rare conformational states, we started all simulations of syntaxin-1a from its closed state. We obtained an initial structure of closed syntaxin by extracting the soluble domain from the crystal structure of syntaxin in complex with Munc-18 (PDB ID 3C98).[108] Simulations were run in Gromacs[100] using the CHARMM36[101] force field. The system was solvated with approximately 90,000 TIP3P water molecules and ions were added to obtain a system with 150 mM NaCl and no net charge. The system was energy minimized using the steepest-descent integrator for 5000 steps or until the largest force was less than 500 kJ mol^{-1} nm^{-2}, whichever came first. A brief 100 ps equilibration was run using NPT conditions using the velocity-rescaling thermostat[102] 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.[109] Covalent bonds were constrained using LINCS, and long-range electrostatics were treated using Particle Mesh Ewald.[103] For each set of ensemble simulations, we generated 50 identical replicas from the equilibrated structure and used these replicas as initial states for production runs.

Production simulations All production simulations were run under the same NPT conditions described above. DEER-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 all three distributions. Histograms were calculated using 1 Å bins. These distributions were then incorporated into MD simulation using each of three ensemble methods, detailed below. Production simulations were carried out using 50 ensemble members and 5 µs of simulation data were collected for each refinement method except EBMetaD. The reason for this exception is described in "EBMetaD simulations." Simulations were run using Gromacs[100] and the gmxapi Python API[77], which permits introduction of user-defined biasing potentials.

BRER simulations To sample the syntaxin conformational ensemble, we performed five iterations of BRER for each of 50 ensemble members. Each iteration is performed as follows: first, one target distance is chosen from each of the smoothed DEER distributions, then a linear biasing potential

$$U_{bias} = \sum_{n=1}^{N_{\text{distributions}}} \alpha_i \frac{d_{\text{MD}}^{(n)}}{d_{\text{target}}^{(n)}}$$

is applied to drive the simulation distance to the target.

Convergence to the target is achieved in two phases. During the training phase, the Hamiltonian coupling constants α are learned for each target using a modified version of the method described by White and Voth.[17] Each constant α is updated every 50 ps according to

$$\alpha_\tau = \alpha_{\tau-1} - \eta_\tau g_\tau$$

where η is the learning rate and g is the gradient:

$$g_\tau = -2\beta \left(\frac{\langle d_{\mathrm{MD}} \rangle_\tau}{d_{\mathrm{target}}} - 1 \right) \left(\langle d_{\mathrm{MD}}^2 \rangle - \langle d_{\mathrm{MD}} \rangle^2 \right),$$

$$\eta_\tau = \frac{A}{\sqrt{\sum_{i=1}^{\tau} g_i}}$$

At the end of the training phase, we select the maximum value of α to prevent underestimating α if the i^{th}degree of freedom converges much faster than the others. During the convergence phase, the simulation is restarted from the beginning of the iteration and a time independent potential (α fixed) is applied until the simulation converges to the target. The parameter A was chosen so as to achieve convergence between 1 ns to 5 ns ($A = 150\beta$). Once the simulation has converged to the target, a 20 ns production run is performed to relax the remaining degrees of freedom. The full procedure is then repeated, beginning with random resampling from the DEER distributions.

A python package to run BRER ensemble simulations is available at `https://github.com/jmhays/run_brer` and documentation can be found at `https://jmhays.github.io/run_brer`. A singularity container is also available at `https://singularity-hub.org/collections/1761`.

EBMetaD simulations EBMetaD simulations were implemented using the same modified version of gromacs and gmxapi[77] version as the BRER simulations. Because the EBMetaD potential is ill-defined in regions of zero probability, we add a small uniform prior to all experimental distributions: using the same notation as Marinelli and Faraldo-Gomez,[78] the modified EBMetaD potential is

$$V(\xi, t) = \sum_{t'=\tau,2\tau,\ldots}^{t} \frac{w \exp\left\{ -\left[\xi - \xi^f\left(X_{t'}\right)\right]^2 / 2\sigma^2 \right\}}{\exp\left\{S_\rho\right\}\left(\rho_{\exp}\left[\xi^f\left(X_{t'}\right)\right]\right) + \delta_{\mathrm{uniform}}}$$

where we have added the term $\delta_{\mathrm{uniform}}$. As $\delta_{\mathrm{uniform}}$increases, the simulations become more numerically stable, but the solution approaches standard metadynamics. Therefore, $\delta_{\mathrm{uniform}}$should be chosen carefully so as to maintain information about the DEER distributions but still produce stable simulations. We selected $\delta_{\mathrm{uniform}}$=0.1. Even with

this choice of δ_{uniform}, the method exhibited a high rate of stochastic failure: all 50 ensemble members failed in the range of 10 ns to 50 ns of simulation time. Because of this, we were only able to collect ∼3 µs of data.

This version of the EBMetaD method is available at https://github.com/jmhays/run_ebmetad. A singularity container is also available at https://www.singularity-hub.org/collections/1994.

Restrained-ensemble simulations Restrained-ensemble biasing potentials previously developed by Roux[23,61] were applied to match MD distance histograms to DEER-derived distance distributions. Refinement was performed via restrained-ensemble simulation using a modified version of Gromacs 5.2 available at https://github.com/kassonlab/restrained-ensemble. This method exhibits numerical instabilities when distributions are very tightly peaked, such as when an ensemble is started from copies of a single initial state. Thus, we initially used a very broad smoothing parameter, σ=10 Å for both the MD and DEER-derived distributions, which we modified to σ=1 Å once the ensemble had sampled enough of the distribution to be stable for small σ. Distance data were collected for all ensemble members for a period of 100 ps followed by an update of the biasing potential with a spring constant of K=100 kJ mol^{-1} nm^{-2}. Additionally, a boxcar averaging filter was applied so that the simulation distance distributions were calculated using the last 10 ns of data. These modifications were implemented in order to obtain sufficient sampling for generating the MD distance distributions as previously described in Hays et al.[12]

Calculation of final distributions and Jensen-Shannon divergence

For each ensemble, production simulations were sampled at 500 ps intervals and distances between the C_β of each residue-residue pair measured by DEER were calculated using MDAnalysis.[110] The distributions plotted in Fig. 3 of the main text were calculated using a Gaussian filter with smoothing parameter σ=2 Å and 1 Å bins for all three distributions. This was done for consistency with the experimental data, which was also smoothed with σ=2 Å and 1 Å bin width. J-S divergence was calculated using the smoothed experimental and simulation distributions.

Conformational ensemble analysis

We clustered the BRER-refined structures as follows: final, relaxed structures from each stochastic-resampling iteration were collected and the distances between the C_β of each residue-residue pair measured by DEER were calculated using MDAnalysis. Distances were calculated from C_α for glycine residues. We performed k-means clustering on these distance coordinates for a broad range of cluster numbers (2 – 50 clusters). We

selected the smallest cluster number (20) for which the average intra-cluster RMSD was substantially higher than the average inter-cluster RMSD (7 Å and 9 Å, respectively). Clusters were classified as "open" if the 52/210 distance was > 40 Å. The structure rendered in Fig 4 of the main text is the centroid of the most populated open cluster.

B.2 Summary of restrained-ensemble MD, EBMetaD, and BRER methods: properties of their biasing potentials

Method	U_{bias}	**Behavior when** $p_{DEER} = 0$	**Exchange between well-separated modes of** p_{DEER}	**Obeys maxEnt principle**
BRER	stochastic resampling with linear potential $\alpha \frac{d_{MD} - d_{target}}{d_{target}}$	stable	yes	yes
EBMetaD	$\sum_{t'=\tau,2\tau,\ldots}^{t} \frac{w \exp\left\{-\left[\xi - \xi^f(X_{t'})\right]^2 / 2\sigma^2\right\}}{\exp\left\{S_\rho\right\}\left(\rho_{\exp}\left[\xi^f(X_{t'})\right]\right)}$	unstable $U_{bias} \to \infty$	no	yes
restrained-ensemble	$\frac{1}{2}k\left(P_{\mathrm{DEER}}(d) - P_{\mathrm{MD}}(d)\right)^2$	stable	no	only in the limit of $k \to \infty$ or an infinite number of ensemble members

TABLE B.1: **Summary of the differences between BRER, EBMetaD, and restrained-ensemble**

Bibliography

(1) Boehr, D. D.; Nussinov, R.; Wright, P. E. *Nature Chemical Biology* **2009**, *5*, 789–796.

(2) Jimenez, R.; Salazar, G.; Yin, J.; Joo, T.; Romesberg, F. E. *Proceedings of the National Academy of Sciences* **2004**, *101*, 3803–3808.

(3) Motlagh, H. N.; Wrabl, J. O.; Li, J.; Hilser, V. J. *Nature* **2014**, *508*, 331–339.

(4) Willcox, B. E.; Gao, G. F.; Wyer, J. R.; Ladbury, J. E.; Bell, J. I.; Jakobsen, B. K.; van der Merwe, P. A. *Immunity* **1999**, *10*, 357–65.

(5) Norman, A. W.; Mizwicki, M. T.; Norman, D. P. G. *Nature Reviews Drug Discovery* **2004**, *3*, 27–41.

(6) Wright, P. E.; Dyson, H. J. *Nature Reviews Molecular Cell Biology* **2015**, *16*, 18–29.

(7) Wei, G.; Xi, W.; Nussinov, R.; Ma, B. *Chemical Reviews* **2016**, *116*, 6516–6551.

(8) Mittal, S.; Shukla, D. *The Journal of Physical Chemistry B* **2017**, *121*, 9761–9770.

(9) Jeschke, G. *Journal of Chemical Theory and Computation* **2012**, *8*, 3854–3863.

(10) Jeschke, G. *Protein Science* **2018**, *27*, 76–85.

(11) Krug, U.; Alexander, N.; Stein, R.; Keim, A.; Mchaourab, H.; Sträter, N.; Meiler, J. *Structure* **2016**, *24*, 43–56.

(12) Hays, J. M.; Kieber, M. K.; Li, J. Z.; Han, J. I.; Columbus, L.; Kasson, P. M. *Angewandte Chemie International Edition* **2018**, *57*, 17110–17114.

(13) Bernadó, P.; Mylonas, E.; Petoukhov, M. V.; Blackledge, M.; Svergun, D. I. *Journal of the American Chemical Society* **2007**, *129*, 5656–5664.

(14) Levin, E. J.; Kondrashov, D. A.; Wesenberg, G. E.; Phillips, G. N. *Structure* **2007**, *15*, 1040–1052.

(15) Bonvin, A. M.J. J.; Brü, A. T. *J. Mol. Biol* **1995**, *250*, 80–93.

(16) Bonomi, M.; Camilloni, C.; Cavalli, A.; Vendruscolo, M. *Science Advances* **2016**, *2*, e1501177.

(17) White, A. D.; Voth, G. A. *Journal of Chemical Theory and Computation* **2014**, *10*, 3023–3030.

(18) Hummer, G.; Köfinger, J. *The Journal of Chemical Physics* **2015**, *143*, 243150.

(19) Chen, J.; Chen, J.; Pinamonti, G.; Clementi, C. *Journal of Chemical Theory and Computation* **2018**, *14*, 3849–3858.

(20) Olsson, S.; Wu, H.; Paul, F.; Clementi, C.; Noé, F. *Proceedings of the National Academy of Sciences* **2017**, *114*, 8265–8270.

(21) Olsson, S.; Frellsen, J.; Boomsma, W.; Mardia, K. V.; Hamelryck, T. *PLoS ONE* **2013**, *8*, ed. by Fernandez-Fuentes, N., e79439.

(22) Piana, S. P.; Laio, A. **2007**, DOI: 10.1021/JP067873L.

(23) Islam, S. M. **2013**.

(24) Shen, R.; Han, W.; Fiorin, G.; Islam, S. M.; Schulten, K.; Roux, B. *PLOS Computational Biology* **2015**, *11*, ed. by Shehu, A., e1004368–e1004368.

(25) Kazmier, K.; Sharma, S.; Quick, M.; Islam, S. M.; Roux, B.; Weinstein, H.; Javitch, J. A.; Mchaourab, H. S. *Nature Structural & Molecular Biology* **2014**, *21*, 472–479.

(26) Hays, J. M.; Cafiso, D. S.; Kasson, P. M. *The Journal of Physical Chemistry Letters* **2019**, 3410–3414.

(27) Von Hagens, T.; Polyhach, Y.; Sajid, M.; Godt, A.; Jeschke, G. *Phys. Chem. Chem. Phys.* **2013**, *15*, 5854–5866.

(28) Sadarangani, M.; Pollard, A. J.; Gray-Owen, S. D. *FEMS Microbiology Reviews* **2011**, *35*, 498–514.

(29) McCaw, S. E.; Liao, E. H.; Gray-Owen, S. D. *Infection and immunity* **2004**, *72*, 2742–52.

(30) Fox, D. A.; Larsson, P.; Lo, R. H.; Kroncke, B. M.; Kasson, P. M.; Columbus, L. *Journal of the American Chemical Society* **2014**, *136*, 9938–9946.

(31) Martin, J. N.; Ball, L. M.; Solomon, T. L.; Dewald, A. H.; Criss, A. K.; Columbus, L. *Biochemistry* **2016**, *55*, 4286–4294.

(32) Südhof, T. C.; Rothman, J. E. *Science (New York, N.Y.)* **2009**, *323*, 474–7.

(33) Jahn, R.; Fasshauer, D. *Nature* **2012**, *490*, 201–207.

(34) Jahn, R.; Scheller, R. H. *Nature Reviews Molecular Cell Biology* **2006**, *7*, 631–643.

(35) Liang, B.; Kiessling, V.; Tamm, L. K. *Proceedings of the National Academy of Sciences* **2013**, *110*, 19384–19389.

(36) Liang, B.; Tamm, L. K. *Progress in Nuclear Magnetic Resonance Spectroscopy* **2018**, *105*, 41–53.

(37) Dawidowski, D.; Cafiso, D. *Biophysical Journal* **2013**, *104*, 1585–1594.

(38) Rizo, J.; Südhof, T. C. *Nature Reviews Neuroscience* **2002**, *3*, 641–653.

(39) Wang, S.; Choi, U. B.; Gong, J.; Yang, X.; Li, Y.; Wang, A. L.; Yang, X.; Brunger, A. T.; Ma, C. *The EMBO Journal* **2017**, *36*, 816–829.

(40) Sutton, R. B.; Fasshauer, D.; Jahn, R.; Brunger, A. T. *Nature* **1998**, *395*, 347–353.

(41) Antonin, W.; Fasshauer, D.; Becker, S.; Jahn, R.; Schneider, T. R. *Nature Structural Biology* **2002**, *9*, 107–111.

(42) Shaw, D. E.; Maragakis, P.; Lindorff-Larsen, K.; Piana, S.; Dror, R. O.; Eastwood, M. P.; Bank, J. A.; Jumper, J. M.; Salmon, J. K.; Shan, Y.; Wriggers, W. *Science* **2010**, *330*, 341–346.

(43) Jeschke, G. *Annual Review of Physical Chemistry* **2012**, *63*, 419–446.

(44) Jeschke, G.; Koch, A.; Jonas, U.; Godt, A. *Journal of Magnetic Resonance* **2002**, *155*, 72–82.

(45) Hubbell, W. L.; Cafiso, D. S.; Altenbach, C. *Nature Structural Biology* **2000**, *7*, 735–739.

(46) Todd, A. P.; Cong, J.; Levinthal, F.; Levinthal, C.; Hubell, W. L. *Proteins: Structure, Function, and Bioinformatics* **1989**, *6*, 294–305.

(47) Jeschke, G. *Proteins: Structure, Function, and Bioinformatics* **2016**, *84*, 544–560.

(48) Ward, R.; Zoltner, M.; Beer, L.; El Mkami, H.; Henderson, I.; Palmer, T.; Norman, D. *Structure* **2009**, *17*, 1187–1194.

(49) Rao, J. N.; Jao, C. C.; Hegde, B. G.; Langen, R.; Ulmer, T. S. *Journal of the American Chemical Society* **2010**, *132*, 8657–8668.

(50) Hirst, S. J.; Alexander, N.; Mchaourab, H. S.; Meiler, J. *Journal of Structural Biology* **2011**, *173*, 506–514.

(51) Peng, H.; Long, F.; Ding, C. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2005**, *27*, 1226–1238.

(52) Ding, C.; Peng, H. *Journal of Bioinformatics and Computational Biology* **2005**, *03*, 185–205.

(53) Kucharska, I.; Seelheim, P.; Edrington, T.; Liang, B.; Tamm, L. K. *Structure* **2015**, *23*, 2234–2245.

(54) Pautsch, A.; Schulz, G. E. *Nature Structural Biology* **1998**, *5*, 1013–1017.

(55) Arora, A.; Abildgaard, F.; Bushweller, J. H.; Tamm, L. K. *Nature Structural Biology* **2001**, *8*, 334–338.

(56) Moeck, G. S.; Coulton, J. W.; Postle, K. *Journal of Biological Chemistry* **1997**, *272*, 28391–28397.

(57) Hancock, R. E.; Brinkman, F. S. *Annual Reviews in Microbiology* **2002**, *56*, 17–38.

(58) Sarver, J. L.; Zhang, M.; Liu, L.; Nyenhuis, D.; Cafiso, D. S. *Biochemistry* **2018**, *57*, 1045–1053.

(59) Touw, D. S.; Patel, D. R.; Van Den Berg, B. *PloS one* **2010**, *5*, e15016.

(60) Zheng, W.; Brooks, B. R. *Biophysical Journal* **2005**, *88*, 3109–3117.

(61) Roux, B.; Islam, S. M. *The Journal of Physical Chemistry B* **2013**, *117*, 4733–4739.

(62) Virji, M.; Evans, D.; Hadfield, A.; Grunert, F.; Teixeira, A. M.; Watt, S. M. *Molecular microbiology* **1999**, *34*, 538–551.

(63) Mittag, T.; Kay, L. E.; Forman-Kay, J. D. *Journal of Molecular Recognition* **2010**, *23*, 105–116.

(64) Van den Bedem, H.; Fraser, J. S. *Nature Methods* **2015**, *12*, 307–318.

(65) Rieping, W. *Science* **2005**, *309*, 303–306.

(66) Bonomi, M.; Heller, G. T.; Camilloni, C.; Vendruscolo, M. *Current Opinion in Structural Biology* **2017**, *42*, 106–116.

(67) Bertini, I.; Giachetti, A.; Luchinat, C.; Parigi, G.; Petoukhov, M. V.; Pierattelli, R.; Ravera, E.; Svergun, D. I. *Journal of the American Chemical Society* **2010**, *132*, 13553–13558.

(68) Bonomi, M.; Camilloni, C.; Vendruscolo, M. *Scientific Reports* **2016**, *6*, 31232.

(69) Lindorff-Larsen, K.; Best, R. B.; DePristo, M. A.; Dobson, C. M.; Vendruscolo, M. *Nature* **2005**, *433*, 128–132.

(70) Habeck, M.; Rieping, W.; Nilges, M. *Proceedings of the National Academy of Sciences* **2006**, *103*, 1756–1761.

(71) Heo, L.; Feig, M. *Proceedings of the National Academy of Sciences* **2018**, *115*, 13276–13281.

(72) Buchete, N.-V.; Hummer, G. *The Journal of Physical Chemistry B* **2008**, *112*, 6057–6069.

(73) Noé, F.; Doose, S.; Daidone, I.; Löllmann, M.; Sauer, M.; Chodera, J. D.; Smith, J. C. *Proceedings of the National Academy of Sciences* **2011**, *108*, 4822.

(74) Sompornpisut, P.; Roux, B.; Perozo, E. *Biophysical Journal* **2008**, *95*, 5349–5361.

(75) Pitera, J. W.; Chodera, J. D. *Journal of Chemical Theory and Computation* **2012**, *8*, 3445–3451.

(76) Boomsma, W.; Ferkinghoff-Borg, J.; Lindorff-Larsen, K. *PLoS Computational Biology* **2014**, *10*, ed. by Levitt, M., e1003406.

(77) Irrgang, M. E.; Hays, J. M.; Kasson, P. M. *Bioinformatics* **2018**, *34*, ed. by Valencia, A., 3945–3947.

(78) Marinelli, F.; Faraldo-Gómez, J. *Biophysical Journal* **2015**, *108*, 2779–2782.

(79) Margittai, M.; Widengren, J.; Schweinberger, E.; Schroder, G. F.; Felekyan, S.; Haustein, E.; Konig, M.; Fasshauer, D.; Grubmuller, H.; Jahn, R.; Seidel, C. A. M. *Proceedings of the National Academy of Sciences* **2003**, *100*, 15516–15521.

(80) Carr, C. M. *Nature Structural Biology* **2001**, *8*, 186–188.

(81) Misura, K. M. S.; Scheller, R. H.; Weis, W. I. *Nature* **2000**, *404*, 355–362.

(82) Chen, Y. A.; Scheller, R. H. *Nature Reviews Molecular Cell Biology* **2001**, *2*, 98–106.

(83) Gerber, S. H.; Rah, J.-C.; Min, S.-W.; Liu, X.; de Wit, H.; Dulubova, I.; Meyer, A. C.; Rizo, J.; Arancillo, M.; Hammer, R. E.; Verhage, M.; Rosenmund, C.; Südhof, T. C. *Science* **2008**, *321*, 1507–1510.

(84) Laio, A.; Parrinello, M. *Proceedings of the National Academy of Sciences* **2002**, *99*, 12562–12566.

(85) Leaver-Fay, A. et al. In *Computer Methods, Part C*, Johnson, M. L., Brand, L., Eds.; Methods in Enzymology, Vol. 487; Academic Press: 2011, pp 545 –574.

(86) Chaudhury, S.; Lyskov, S.; Gray, J. J. *Bioinformatics* **2010**, *26*, 689–691.

(87) Jarzynski, C. *Physical Review Letters* **1997**, *78*, 2690–2693.

(88) Widdas, W. F. *The Journal of Physiology* **1952**, *118*, 23–39.

(89) Jardetzky, O. *Nature* **1966**, *211*, 969–970.

(90) Rees, D. C.; Johnson, E.; Lewinson, O. *Nature Reviews Molecular Cell Biology* **2009**, *10*, 218–227.

(91) Abramson, J.; Smirnova, I.; Kasho, V.; Verner, G.; Iwata, S.; Kaback, H. R. *FEBS Letters* **2003**.

(92) Ward, A.; Reyes, C. L.; Yu, J.; Roth, C. B.; Chang, G. *Proceedings of the National Academy of Sciences* **2007**, *104*, 19005.

(93) Wilkens, S. *F1000Prime Reports* **2015**, *7*, DOI: 10.12703/P7-14.

(94) Dawidowski, D.; Cafiso, D. *Structure* **2017**, *24*, 392–400.

(95) Chen, X.; Lu, J.; Dulubova, I.; Rizo, J. *Journal of biomolecular NMR* **2008**, *41*, 43–54.

(96) Lerman, J. C.; Robblee, J.; Fairman, R.; Hughson, F. M., DOI: 10.1021/bi0003994.

(97) Sali, A.; Blundell, T. L. *Journal of Molecular Biology* **1993**, *234*, 779–815.

(98) Lomize, M. A.; Pogozheva, I. D.; Joo, H.; Mosberg, H. I.; Lomize, A. L. *Nucleic Acids Research* **2012**, *40*, D370–376.

(99) Lee, J.; Patel, D. S.; Kucharska, I.; Tamm, L. K.; Im, W. *Biophysical Journal* **2017**, *112*, 346–355.

(100) Pronk, S.; Páll, 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. *Bioinformatics (Oxford, England)* **2013**, *29*, 845–854.

(101) Huang, J.; MacKerell, A. D. *Journal of computational chemistry* **2013**, *34*, 2135–2145.

(102) Bussi, G.; Donadio, D.; Parrinello, M. *The Journal of Chemical Physics* **2007**, *126*, 014101.

(103) Darden, T.; York, D. M.; Pedersen, L. G. In, 1993.

(104) Fox, D. A.; Columbus, L. *Protein Science* **2013**, *22*, 1133–1140.

(105) Pannier, M; Veit, S; Godt, A; Jeschke, G; Spiess, H. W *Journal of Magnetic Resonance* **2000**, *142*, 331–340.

(106) Jeschke, G.; Chechik, V.; Ionita, P.; Godt, A.; Zimmermann, H.; Banham, J.; Timmel, C. R.; Hilger, D.; Jung, H. *Applied Magnetic Resonance* **2006**, *30*, 473–498.

(107) Zhuo, Y.; Yang, J.-Y.; Moremen, K. W.; Prestegard, J. H. *The Journal of Biological Chemistry* **2016**, *291*, 20085–20095.

(108) Burkhardt, P.; Hattendorf, D. A.; Weis, W. I.; Fasshauer, D. *The EMBO journal* **2008**, *27*, 923–933.

(109) Parrinello, M.; Rahman, A. *The Journal of Chemical Physics* **1982**, *76*, 2662–2666.

(110) Michaud-Agrawal, N.; Denning, E. J.; Woolf, T. B.; Beckstein, O. *Journal of Computational Chemistry* **2011**, *32*, 2319–2327.