JACS
JOURNAL OF THE AMERICAN CHEMICAL SOCIETY
Article
pubs.acs.org/JACS

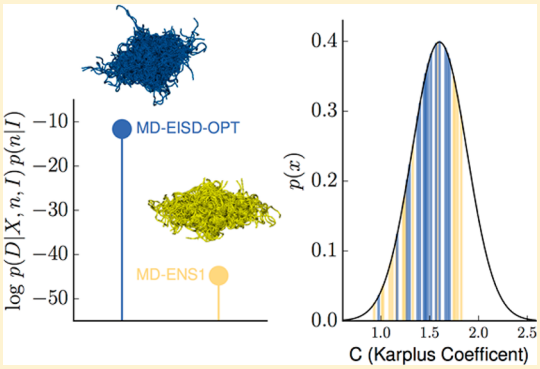
Experimental Inferential Structure Determination of Ensembles for Intrinsically Disordered Proteins

David H. Brookes[‡] and Teresa Head-Gordon*[,†,‡,§,∥]

[†]Department of Chemistry, [‡]Department of Bioengineering, [§]Department of Chemical and Biomolecular Engineering, [∥]Chemical Sciences Division, Lawrence Berkeley National Laboratory, University of California, Berkeley, California 94720, United States

Supporting Information

ABSTRACT: We develop a Bayesian approach to determine the most probable structural ensemble model from candidate structures for intrinsically disordered proteins (IDPs) that takes full advantage of NMR chemical shifts and J-coupling data, their known errors and variances, and the quality of the theoretical back-calculation from structure to experimental observables. Our approach differs from previous formulations in the optimization of experimental and back-calculation nuisance parameters that are treated as random variables with known distributions, as opposed to structural or ensemble weight optimization or use of a reference ensemble. The resulting experimental inferential structure determination (EISD) method is size extensive with O(N) scaling, with N = number of structures, that allows for the rapid ranking of large ensemble data comprising tens of thousands of conformations. We apply the EISD approach on singular folded proteins and a corresponding set of ∼25 000 misfolded states to illustrate the problems that can arise using Boltzmann weighted priors. We then apply the EISD method to rank IDP ensembles most consistent with the NMR data and show that the primary error for ranking or creating good IDP ensembles resides in the poor back-calculation from structure to simulated experimental observable. We show that a reduction by a factor of 3 in the uncertainty of the back-calculation error can improve the discrimination among qualitatively different IDP ensembles for the amyloid-beta peptide.



INTRODUCTION

X-ray and electron crystallography and microscopy have excelled at determining the structure of folded proteins and their complexes[1,2] since the atomic positions are over-determined by the available diffraction intensities from protein crystals. However, these methods are ill-suited for structure determination of intrinsically disordered proteins (IDPs),[3] since the primary characteristic of IDPs is that they are not singular well-folded structures but instead need to be characterized as diverse ensembles of conformational substates in solution.[4] While techniques such as nuclear magnetic resonance (NMR) are highly suitable for probing the solution structural ensemble of an IDP, the dynamical time scale for IDP motions results in highly averaged NMR observables that are typically unable to fully resolve the conformational substates. Therefore, building the connection between the experimental observables and the complete IDP structural ensemble depends critically on computational models.

Knowledge-based computational models are those that directly use experimental NMR, small-angle X-ray scattering (SAXS), and other biophysical information to derive the structural ensemble. Methods that use experimental constraints from NOE data, RDCs, J-couplings, and chemical shifts are the foundation of NMR structure determination of folded proteins and are embodied in software packages such as CANDID,[5] CYANA,[6] X-Plor-NIH,[7,8] SPARTA+,[9] and TALOS.[10] For the case of IDPs, knowledge-based approaches start with an extensive set of conformations derived from a variety of sources, such as MD,[11−13] or methods such as TraDES[14] and Flexible-Meccano,[15] which create ensembles of random statistical-coil conformers. The resulting "basis set" of structures is then culled for the subset of conformations that when back-calculated are in best agreement with experimental data, to create the IDP ensemble. Examples of such methodology are the energy-minima mapping and weighting method,[16,17] ASTEROIDS,[15] and the ENSEMBLE program which accommodates data from a very wide range of the aforementioned NMR sources, as well as hydrodynamic radii (R_h) and SAXS.[18]

By contrast, in recent work we have used *de novo* molecular dynamics sampling for amyloid-beta (Aβ) in which no experimental restraints are applied.[11,19] This *de novo* MD approach allows for the possibility of discovering new conformational ensembles and their time scales of interconversion that is important for correctly capturing NOESY data, and that may also be consistent with the experimental observable once validated through back-calculation. We have found that the unbiased MD calculation yielded qualitatively

Received: January 11, 2016
Published: March 11, 2016

© 2016 American Chemical Society

DOI: 10.1021/jacs.6b00351
J. Am. Chem. Soc. 2016, 138, 4530−4538

different structural ensembles than the TraDES or Flexible-Meccano approaches for Aβ,[12,20] in the sense that the *de novo* MD structures comprised a Boltzmann weighted combination of overall collapsed structures with heterogeneous populations of well-defined turns, α-helices, β-strands, and β-hairpins, as opposed to extended statistical coils with at most local secondary structure motifs.

The knowledge-based and *de novo* approaches are also complementary in that using experimental data directly can overcome challenges of insufficient sampling and force field inaccuracies, while using the power of unbiased sampling can compensate for gaps in experimental restraints in conformer generation and selection. Even so, while these approaches have been highly beneficial to the IDP field, limitations of the sampling and conformer selection methods are beginning to become apparent in either case. In particular, there may be a range of confidence in the IDP ensemble generated depending on how severe the problem is experimentally under-determined (unlike the folded protein case), whether the back-calculations from structure to experimental observable contain significant error, or whether the basis set of structures are actually representative and/or complete.

For example, while chemical shifts and scalar couplings can usually be experimentally measured with high accuracy, we require quantitative back-calculations of the NMR observables from structure to make the best use of that experimental data, in order to generate tighter spatial restraints to discriminate among alternative structural models. To illustrate this point, Figure 1 compares experimental chemical shifts measured for the IDP Aβ42 against chemical shift predictions using SHIFTX2[21] and SHIFTX[22] applied to the same IDP structural ensemble.

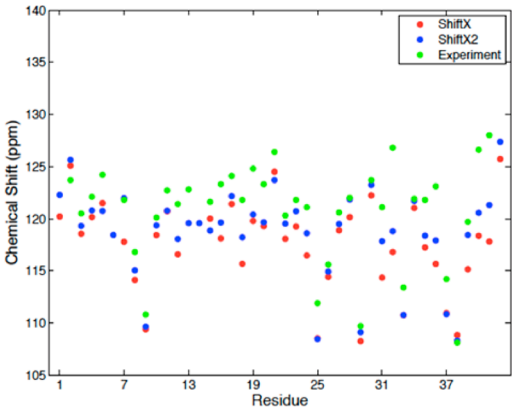


Figure 1. Experimental chemical shifts for Aβ42 (green) compared to back-calculated chemical shifts using SHIFTX (red) and SHIFTX2 (blue) on the same structural ensemble.

Although improvements realized by SHIFTX2 over SHIFTX were significant for folded proteins with the introduction of structural homology information, the level of difference between the calculators is relatively small for the Aβ42 example, since structural homology plays no role for IDPs. Therefore, while heuristic chemical shift calculators and parameter fits to the Karplus equation for J-couplings can be predicted with reasonable accuracy for folded proteins, their applicability to unstructured IDPs is currently problematic. Therefore, to most accurately represent our best knowledge about the IDP problem, one must be careful to extract as much information as possible from experiments, while accounting for any intrinsic measurement error or back-calculation uncertainties and adding as little information as possible in the form of heuristics and assumptions. Hence although the IDP problem is underdetermined for finding a unique solution, Bayesian optimization seems ideally suited for the IDP problem by narrowing the set of solutions to ones that are more relevant than others based on the highest or lowest probabilities.

The seminal work of Nilges and co-workers[23] used Bayesian inference to derive a probability distribution for the folded structure and its precision for well-defined macromolecules characterized using NMR. A number of groups have extended the inferential structure determination (ISD) method into the IDP structure determination domain, such as the Variational Bayesian Weighting (VBW) method,[24−26] Bayesian ensemble refinement;[27] maximum entropy approaches;[28−30] and other Bayesian formulations[31−33] that seek to define "the best" IDP ensemble given the data.

These important influences differ from the Bayesian model presented here in several ways. In particular, we define a set of "nuisance parameters" for each experimental data type that are associated with both the intrinsic experimental error, which for NMR data tends to be small, as well as for errors and any uncertainty in parameters used in the back-calculation from structure, which we illustrate using heuristic chemical shift calculators or Karplus equations for J-couplings. By modeling the nuisance parameters that represent uncertainty in experimental information as random variables whose distributions are known from the available literature, we then optimize over those distributions for each data point to arrive at an optimal combination of values sampled from these distributions. The resulting formulation of our posterior function will be shown to be both size extensive and to scale linearly with the number of structures N, as opposed to the $O(N^3)$ scaling and lack of size extensivity exhibited by other Bayesian methods.[24−26]

Our resulting experimental inferential structure determination (EISD) approach is tested and shown to be quite accurate when applied to three folded proteins using a uniform prior, and we provide cautionary evidence that Boltzmann priors can overwhelm the experimental information and degrade the quality of prediction of the native state for one of the folded proteins with a disordered section. We then extend the EISD method using a uniform prior to rank 7 qualitatively different IDP ensembles for Aβ42 by optimizing posterior distributions that are most consistent with chemical shift and J-coupling NMR data. We show that the problem of determining IDP ensembles is not strictly one of overcoming limited sampling, force field inadequacies, or uncertainty in experimental measurements, but that there are sorely needed improvements in the accuracy of the back-calculation from structure. Finally, we show that a reduction in back-calculation uncertainty by a factor of ∼3−5 could yield significant overall improvement in IDP structural ensemble determination.

■ THEORY

Rieping and co-workers presented a Bayesian framework for determining the most probable structure of a well-folded protein, illustrated using NOESY experimental data that was back-calculated under the isolated spin pair approximation.[23] Their ISD method attempted to find the most probable model from candidate structures from the posterior probability

DOI: 10.1021/jacs.6b00351
J. Am. Chem. Soc. 2016, 138, 4530−4538

distribution $p(X,\xi \mid D,I)$, which is decomposed using Bayes' Theorem:

$$p(X, \xi|D, I) \propto p(D|X, \xi, I)p(\xi|I)p(X|I) \quad (1)$$

where X is a structure, ξ is a set of so-called "nuisance" parameters which are uncertain values that cannot be determined directly from the data (such as the uncertainties in the experimental measurements or back-calculation equations), D is a set of experimental data, and I represents any prior information about the system. In their work, eq 1 models the conformational prior density $p(X|I)$ via Boltzmann weighting using an empirical energy function, the prior density of nuisance parameters $p(\xi|I)$ with Jeffrey's (uninformative) prior, $\pi(\xi)$, and finally assumes that all deviations from the experimental data fit a log-normal distribution to yield the following probabilistic model[23]

$$p(X, \xi|D, I) = \sigma^{-(M+1)}\pi(\xi) \exp\left(-\frac{1}{2\sigma^2}\sum_{i=0}^{M}\log^2\left[\frac{f(o_i, \xi)}{d_i}\right]\right) \exp\left(-\frac{E(X)}{k_bT}\right) \quad (2)$$

Equation 2 assumes we are given a set of M experimental data observations, $D = \{d_i\}_{i=1}^{M}$, a corresponding set of M simulated observables from a candidate structure $X \rightarrow \{o_i\}_{i=1}^{M}$, which are back-calculated using an approximate function f, and that all uncertainty and error is captured with a single variance σ parameter. We refer to the formulation in eq 2 as the original inferential structure determination (OISD) method throughout the rest of the paper.

While these assumptions proved robust for the folded class of protein, it requires significant reformulation if it is to be applied to the more underdetermined problem of IDPs, where using all known and reliable information well is crucial. For example, eq 2 uses an uninformative prior to represent the experimental nuisance parameters, even though often we know quite a lot about the distribution of these parameters. Additionally, the underlying assumptions about the experimental data effectively lumps all uncertainties into a single σ, which prevents us from using all the information we know about the separate distributions of different experimental data types and the variable quality with which we can back-calculate these observables from structure. For example, we can model the experimental error in C_α and H_α chemical shifts as a normal distribution with mean equal to 0 and variances equal to 0.1 and 0.01, respectively, and the corresponding error probability of 0.05 in the distributions would then be 0.1827 and 1.487×10^{-6}. So if the log-normal distribution in eq 2 is fit using many C_α shifts, than a large error in an H_α measurement might go unnoticed, although it is likely quite significant for discriminating for or against a candidate structure.

More recent Bayesian methods reformulate how we evaluate the $p(X,\xi \mid D,I)$ term to take better advantage of known and thus useful experimental information from the NMR method.[24−26,29,34] To model a more informative prior $p(\xi|I)$ for the nuisance parameters, one can decompose it into independent distributions for each experimental data, which allows for the modeling of uncertainties of individual data type more precisely instead of lumping it into one large variance σ. This essentially provides a higher resolution model that involves more refined experimental assumptions and is easily extensible as we gain more information about the system of interest, whether it is a folded protein or an IDP. Although Stultz and co-workers also model uncertainties associated with individual data types and their back-calculation, we use this information differently in the formulation of our posterior distribution by optimizing the experimentally related nuisance parameters to conform within the variance of their distributions and not weights on structures as they do in their Variational Bayesian Weighting method (VBW). This has important consequences for the scaling and size extensivity of the Bayesian model that we show in the Results section.

To construct the EISD model, we first assume independence of all (x_i,d_i) pairs and then take the log of eq 1 to yield

$$\log p(X, \xi|D, I) \propto \log p(X|I) + \sum_{i=1}^{M}\log[p(d_i|o_i, \xi_i, I)p(\xi_i|I)] \quad (3)$$

in which the structural prior distribution $p\ (X \mid I)$ is modeled either as a uniform prior or as a Boltzmann prior, which is explored in the Results section. For the experimental prior, we define a set of nuisance parameters defined as

$$p(\xi_i|I) = p(\xi_{(\mathrm{exp})_i})p(\xi_{(\mathrm{back})_i}) \quad (4)$$

where $p(\xi_{(\mathrm{exp})i})$ and $p(\xi_{(\mathrm{back})i})$ define a set of independent Gaussian distribution models for all experimental and back-calculation error for each data type and for each data point i, respectively. In this work, all of the experimental and back-calculation nuisance parameters are defined as Gaussian random variables whose distributions are taken from the literature and described in more detail in Methods and the Supporting Information.

These terms collectively consider all uncertainty in the experimental data and back-calculation, and we can therefore model the conditional distribution of data points given structural measurements and nuisance parameters as

$$p(d_i|o_i, \xi_i, I) = \begin{cases} 1 & \text{if } d_i + \xi_{(\mathrm{exp})_i} = f\left(o_i, \xi_{(\mathrm{back})_i}\right) \\ 0 & \text{otherwise} \end{cases} \quad (5)$$

This inference scheme is then easily extended to IDPs if we now assume that we are given N structures in the ensemble, $X = \{X^{(j)}\}_{j=1}^{N}$ each of which contain M structural measurements, $X^{(j)} \rightarrow \{o_i^{(j)}\}_{i=1}^{M}$ as well as the data and nuisance parameter terms of the particular experimental measurement. In addition, all that is known for a given NMR measurement on an IDP is that it corresponds to an average of that measurement over every structure in the ensemble. The only change to eq 5 that is required to make EISD suitable for IDPs is

$$p(d_i|o_i \in \{o_i^{(j)}\}_{j=1}^{N}, \xi_i, I) = \begin{cases} 1 & \text{if } d_i + \xi_{(\mathrm{exp})_i} = \left\langle f\left(o_i^{(j)}, \xi_{(\mathrm{back})_i}\right)\right\rangle_{j=1}^{N} \\ 0 & \text{otherwise} \end{cases} \quad (6)$$

where $\langle\ \rangle$ denotes an average over structures used to back-calculate experimental observables. For both folded proteins and IDPs, the posterior probability is determined by an optimization over the combination of nuisance parameters sampled from these distributions for each data point (eqs 5 and

DOI: 10.1021/jacs.6b00351
J. Am. Chem. Soc. 2016, 138, 4530−4538

6) to arrive at the model probability for structures or IDP ensembles.

METHODS

We created an ensemble of misfolded structures for three natively folded proteins, the 21 residue Trp-cage mini-protein (1L2Y),[35] a 135-residue retinol binding protein in its apo state (1JBH),[36] and a 71 amino acid 8.3KDA protein with unknown function from *methanobacterium thermoautotrophicum* (1GH9) which has a disordered section.[37] In each case, we used a reverse Metropolis algorithm to create a 25 000 member ensemble starting with the native PDB structure for a given protein, and at every iteration perturbing the current state by randomly sampling ϕ,ψ dihedral angles for a random residue from a Gaussian Mixture Model of dihedral angles trained with 10 000 PDB structures (thereby generating a Ramachandran plot). The result is a set of physically reasonable but misfolded structures with RMSD values from the native conformation ranging from 0.0 to 10.0 Å. For each structure we calculated $\log p(X,\xi \mid D,I)$ using both the OISD approach that lumps all uncertainty into a single variance, and the EISD formulation which treats experimental data types separately, as well as a physical energy using AMBER99 and an implicit solvent force field as implemented in MMTK.[38]

For the IDP ensembles, we use previously reported IDP data sets generated for the Aβ42 monomer: one random coil ensemble generated from TraDES;[14] one ensemble generated from a replica exchange simulation (*de novo* MD),[20] one statistical coil ensemble that incorporates bioinformatics knowledge about independent local secondary structure at each residue (Pred-SS),[20] and four ensembles generated by adding experimental restraints from NMR (RDCs, NOEs, scalar couplings, and chemical shifts) operating on the *de novo* MD and Pred-SS ensembles using ENSEMBLE (MD-ENS1, MD-ENS2, MD-ENS4, and Pred-SS-ENS).[18,39,40]

In this work we use chemical shifts and J-coupling data reported for 1L2Y,[35] 1JBH,[36] 1GH9,[37] and Aβ42.[11,20,41,42] We model $p(\xi_{(\mathrm{exp})_i})$ as Gaussian distributions centered at the reported NMR data value, and use the experimental uncertainty for each measured data point to define the variance, and the reader is referred to the original experimental studies for this nuisance data. To model the Gaussian distributions for $p(\xi_{(\mathrm{back})_i})$, there are differences in the treatment of back-calculations for scalar couplings and chemical shifts. For J-couplings we optimize over the three nuisance parameters A, B, and C of the back-calculation function $f(x)$, i.e., the Karplus equation (Table S1):

$$J = A\cos^2\phi + B\cos\phi + C \tag{7}$$

whose mean and variance are taken from Vuister and Bax.[43] For chemical shifts we use SHIFTX2[21] as the back-calculator, but it does not allow for direct optimization of parameters of f. The modeled Gaussian distributions in this case are given in Table S1 using the well-documented error distributions.[21]

We used the local optimization Powell algorithm in the SciPy package[44] to maximize the posterior probabilities by optimizing the complete (low-dimensional) set of nuisance parameters, $\{\xi\}$, for all available experimental chemical shifts and scalar coupling data and back-calculations, for both OISD and EISD. We found in practice that global optimizations were sometimes required to maximize the probability in the OISD model, whereas local optimizers were always sufficient for EISD. This lends an advantage to EISD since global optimization is much more computationally intensive than local optimization algorithms.

RESULTS

Our first test is to see how well OISD and EISD perform on predicting the native PDB structure of well-folded proteins with respect to ~25 000 other structures with larger RMSDs. Figure 2 shows the plot of optimized $\log p(X,\xi \mid D,I)$ vs RMSD using 115 measured chemical shifts (H_α, H_β, H_N, and all side chain hydrogens) for the 21 residue Trp-cage mini-protein.[36]

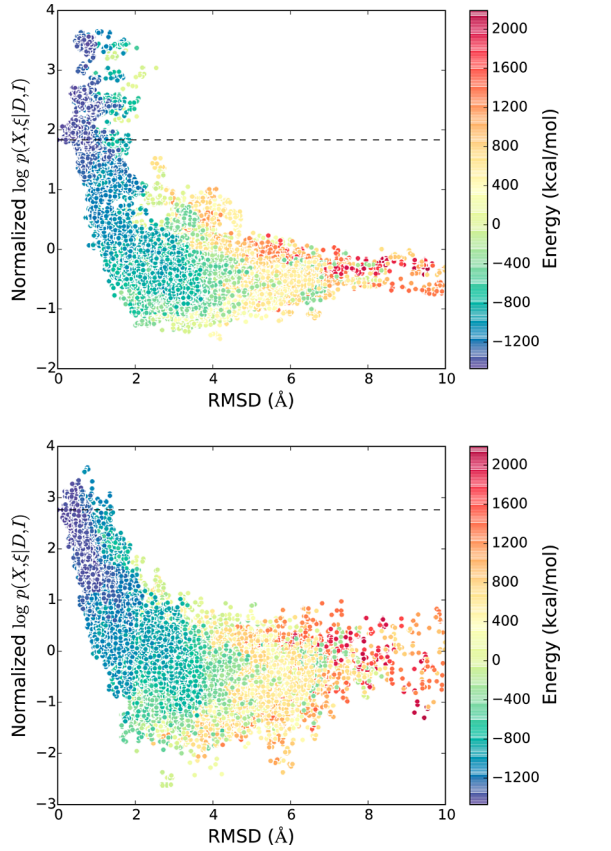


Figure 2. $\log p(X,\xi \mid D,I)$ vs RMSD for ~25 000 misfolded structures for 1L2Y using a uniform prior for (a) OISD and (b) EISD. Dotted black lines represent the fit-to-data probability of the native structure. All probabilities were normalized so the set had a mean of 0 and a variance of 1 (for easier comparison between schemes).

We first use an uninformative uniform prior to better ascertain the differences in how experimental information is handled. Qualitatively we can see that for both schemes the log probability has an overall negative correlation with RMSD, meaning that both methods can distinguish well between reasonable and unreasonable structures, even when energy is not considered via Boltzmann weighting as per eq 4.

However, we find that the EISD formulation assigns higher probabilities to lower RMSD structures compared to the OISD model. For OISD, the largest RMSD structure with a probability higher than the native state was 2.54 Å, whereas the structure with the highest overall probability had an RMSD of 1.02 Å. By contrast, for EISD the largest RMSD structure with a probability higher than the native state was 1.37 Å, whereas the structure with the highest overall probability had an RMSD of 0.76 Å.

The same conclusions apply when we perform the same test on the 1JBH protein using 855 chemical shifts (Figures S2 and S3). In this case, using the OISD method, the largest RMSD structure with a probability higher than the native state was 5.04 Å, while the highest overall probability had an RMSD of 2.41 Å. Using the EISD method, the largest RMSD structure with a probability higher than the native state was 0.84 Å, while the highest overall probability had an RMSD of 0.55 Å. Using

DOI: 10.1021/jacs.6b00351
J. Am. Chem. Soc. 2016, 138, 4530−4538

Boltzmann weighting as the prior improves these results so that the most probable structure is ~0.5 Å RMSD for both methods (Figure S1).

To put this result in perspective, the RMSD among ~90 different experimental X-ray structures of hen egg white lysozyme is 0.75 Å,[45] showing that we can obtain results with an uninformative uniform prior that are nearly as good as experimental uncertainty associated with X-ray crystallography but using NMR data. In fact the Boltzmann prior dominates the posterior distribution to overcome any difference in the experimental error handling between methods for the Trp-cage and retinol binding protein.

For the test on 1GH9 using 59 $^3J(H_N,H_\alpha)$ coupling constants, both posterior probability models were able to utilize the experimental data equally well (Figure 3) using a uniform prior. Here the results are more mixed for both methods since the largest RMSD structure with a probability higher than the native state was 2.02 and 2.36 Å, although the highest overall probability had an RMSD of 0.38 and 0.47 Å for EISD and OISD, respectively.

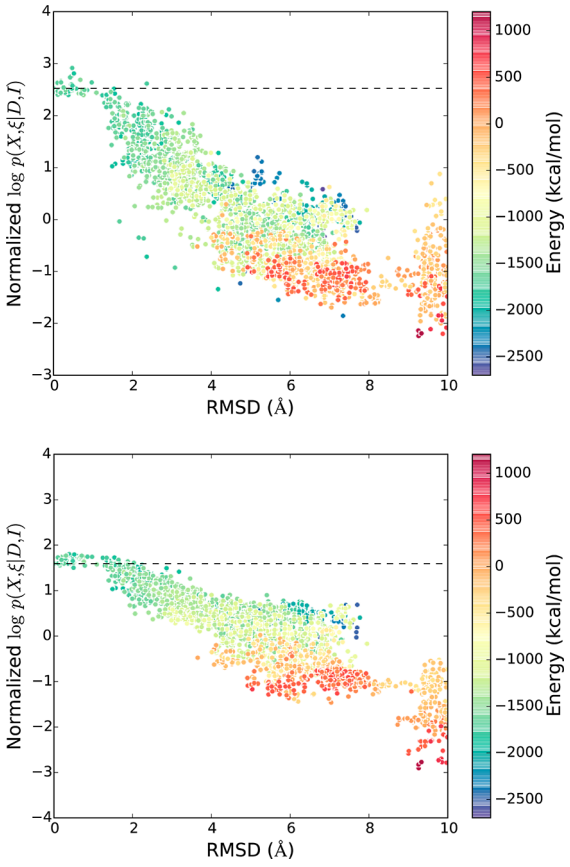


Figure 3. log $p(X,\xi \mid D,I)$ vs RMSD for ~25 000 misfolded structures for 1GH9 for (a) OISD and (b) EISD using a uniform prior. See Figure 2 caption for further details.

However, when the Boltzmann prior is applied, there are many structures in the 1GH9 test set with high-RMSD but with significantly lower energies based on the simple nonpolarizable protein force field and implicit solvent models we used here (Figure 4). 1GH9 is relevant to the IDP problem since it has a large disordered section, and we use the indiscriminate energy function to highlight the issue for IDPs, for which force fields may be suspect in general. This emphasizes the well-known problem with poorly chosen energy functions that are not able to discriminate the native state from misfolded structures.

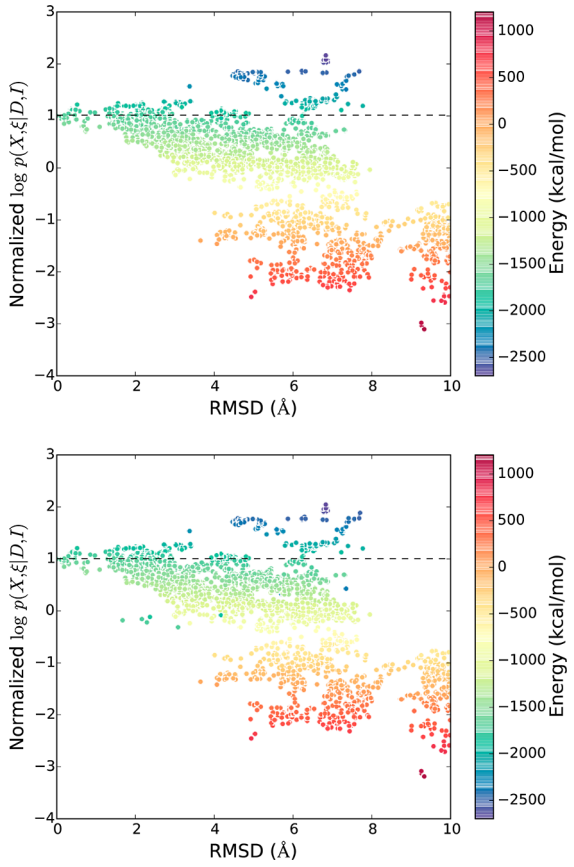


Figure 4. log $p(X,\xi \mid D,I)$ vs RMSD for ~25 000 structures for 1GH9 for (a) OISD and (b) EISD using a Boltzmann prior. See Figure 2 caption for further details.

This can in principle be remedied by using a different force field, such as the energy functions we have previously developed and which have undergone extensive validation for native state predictions for folded proteins.[46−48] Alternatively, in the Bayesian formulation of Hummer and Kofinger, this would be handled as an additional nuisance parameter that reflects low confidence in the reference ensemble they use for their prior.[27] However, for IDPs, we believe the best use of Boltzmann weighting, at present, is to use it to generate diverse conformational states via molecular dynamics or Monte Carlo sampling, and then to rank the resulting different ensembles using an uninformative uniform prior as we have shown above, i.e., using it directly for conformer selection as per previous studies would too often lead to unpredictable outcomes such as what we have demonstrated for 1GH9.

Next we turn to the application of the EISD method to a more directly relevant IDP case, namely, the amyloid-β IDP. This will illustrate multiple issues in regards to experimental error handling that both differs from the work of others and which yields insight into how to improve estimates of more probable IDP ensemble in the future. On the basis of the outcome on 1GH9, we use a uniform distribution for the structural prior for the IDP results below.

Our model for the experimental prior $p(\xi|I)$ for the individual NMR data types, including experimental measurement uncertainty and back-calculation error, is similar to that presented by Fisher et al.[24−26,34] However, our work diverges

DOI: 10.1021/jacs.6b00351
J. Am. Chem. Soc. 2016, 138, 4530−4538

from theirs since they do not treat the nuisance parameters as random variables to be optimized, but instead they optimize weights of structures keeping the nuisance parameters fixed. We refer the reader to eq 10 in ref 49, which is the posterior distribution equation that is minimized by VBW. To illustrate the implications of the difference between VBW and EISD on the choice of error handling, we implemented their method to perform several comparisons.

For the VBW method, the posterior distribution contains two sums over the number of structures and thus its computational cost scales $O(N^2)$; they further suggest that to optimize this equation, one employs a simulated annealing procedure with $100 \times N$ steps, making their full optimization procedure for an ensemble scale as $O(N^3)$. This clearly restricts the VBW approach to optimizations over very small ($N \sim 200$–300 structures) data sets (Figure 5a). By contrast, our EISD method scales as $O(N)$ as evident from eq 6.

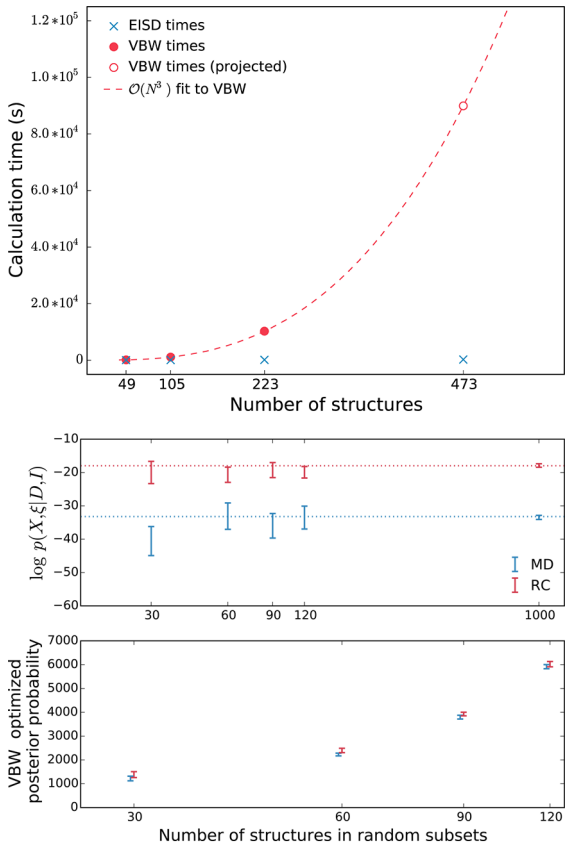


Figure 5. Scaling properties and size extensivity of the VBW vs EISD Bayesian models. (a) Computational scaling for VBW is $O(N^3)$ whereas the scaling for EISD is $O(N)$. We note that both models also scale with the number of experimental data points M. (b) The VBW posterior probability is not size extensive, whereas the EISD probability is size-extensive.

This more favorable scaling allows us to easily embed our EISD posterior probabilities into a Metropolis-Hastings Monte Carlo framework to optimize ensembles involving thousands of structures. Table 1 shows that the calculated probability of a ~1000 member ensemble derived from the *de novo* MD ensemble and then optimized for 5000 iterations of Monte Carlo sampling (MD-EISD-OPT). The new ensemble has significantly higher fit-to-data probabilities than the parent ensemble after relatively few iterations considering the combinatorics of the state space size of this search problem. This illustrates the strength of the $O(N)$ scaling of the EISD method for a calculation that would not be tractable under the VBW formulation.

Table 1. log $p(X,\xi|D,I)$ Probabilities Using Equation 6 for Seven Different IDP Ensembles for Aβ42[20a]

structural ensemble	J-coupling	chemical shift	both
MD-EISD-OPT	−20.443	−114.800	−133.296
MD-ENS4	−17.400	−116.588	−132.042
Random Coil	−18.471	−117.316	−133.841
MD-ENS2	−20.929	−116.013	−134.996
Pred-SS-ENS	−31.762	−116.414	−146.230
de novo MD	−33.221	−124.257	−155.532
MD-ENS1	−39.381	−121.202	−158.637
Pred-SS	−44.449	−120.344	−162.848

[a]See main text for their description.

Furthermore, since the nature of the VBW posterior distribution directly builds in a strong dependence on the sample size N, their results are not size extensive. Therefore, the VBW method relies heavily on the assumption that ~200–300 structures (what is tractable with their method) are representative of the IDP ensemble and that results will not change with respect to larger data sets. In order to test the impact of small data sets and lack of size extensivity, we performed a second test of the two methods for ranking two qualitatively different ensembles of the Aβ42 monomer. We randomly chose a "reservoir" of 5000 structures from the full *de novo* MD and RC ensembles (which have a total of about 42 000 and 83 000 structures, respectively) and then sampled random subsets of 30, 60, 90, 120 up to 1000 structures from this reservoir for each size.

Figure 5b shows the resulting optimized posterior probabilities using the VBW and EISD methods across these random data sets. It is evident that the VBW method shows significant overlap between the two ensembles given the small data sets used, indicating the sensitivity to incomplete data, and furthermore that the optimized probabilities change with the size of the ensemble N. By contrast the EISD method can resolve the differences between ensembles with much smaller data sets, and the EISD posterior probabilities are largely independent of system size beyond ~30 structures, since we always optimize over the same set of M nuisance parameters for any size N of discrete structures or structural ensembles.

Next we consider the aspect of IDP ensemble determination that is most problematic at present, i.e., back-calculation from structure, which we show competes with or even supersedes other issues such as the adequacy of force fields and conformational sampling. We have previously reported the generation of many different IDP ensembles for the Aβ42 monomer,[20] ranging in size from hundreds of structures to ~83 000 structures that we argue are *qualitatively* different. The qualitative differences among these IDP ensemble types would in fact lead to very different hypotheses about their biology and motivates the strong desire to differentiate between them. Thus, ranking of ensembles with well-separated probabilities using experimental information would significantly build confidence on the best hypothesis to pursue.

The first class of Aβ42 monomer ensemble comprises a structurally featureless random coil ensemble (RC) as well as random coils with statistical secondary structure motifs (Pred-SS); these are representative of structural ensembles that are

DOI: 10.1021/jacs.6b00351
J. Am. Chem. Soc. 2016, 138, 4530–4538

equivalent to an unfolded protein under very high denaturant conditions with a large radius of gyration. In addition we consider very heterogeneous but highly structured ensembles generated from a replica exchange simulation (*de novo* MD) that we would classify as an unfolded protein under very low denaturant conditions. In addition, we operate on the two classes structures using the ENSEMBLE method to create new ensembles that in principle agree with the available experimental data via restraints (MD-ENS1, MD-ENS2, MD-ENS4, and PRED-SS-ENS).

Table 1 tabulates the values of the optimized log $p(X,\xi \mid D,I)$ using eq 6, for seven qualitatively different ensembles for the Aβ42 monomer,[20] using 16 $^3J(H_N,H_\alpha)$ coupling constants and back-calculations from the Karplus equation and 194 hydrogen chemical shifts and using SHIFTX2 as the back-calculation from structure.

Figure 6 presents the results in more graphical form by showing how strongly the rankings depend on experimental data types. Figure 6a demonstrates that when only chemical shift data are used, the MD-ENS2, MD-ENS4, Pred-SS-ENS, and RC ensembles are within uncertainty of the sample size used for each case. When only J-couplings are considered, the rank order changes completely, and the relative rankings of ensembles are somewhat better differentiated as seen in Figure 6b. When we use both scalar couplings and chemical shifts together (Figure 6c), the relative rankings between ensembles are qualitatively unchanged from using J-couplings alone.

While it might suggest that J-coupling constants are a more discriminating measurement for determining IDP structure, in fact it is that the inherent errors of the heuristic chemical shift calculators are larger than uncertainties in the Karplus equations and add little to the discrimination among ensembles, as implied in Figure 1. Even so, the parameters of the Karplus equation do not escape scrutiny, since J-couplings alone or together with chemical shifts cannot differentiate between the extended RC ensemble and the collapsed and structured MD ensembles. Even the Metropolis scheme for optimizing new ensembles using $p(X,\xi|D,I)$ for MD-EISD-OPT are likely dominated by the problems with back-calculation errors (Table 1).

To more explicitly show the uncertainties that arise from back-calculation errors, we use Gaussian Kernel Density Estimation (KDE)[50] to approximate the probability distributions of back-calculated values corresponding to each experimental data point. Figure 7 shows the KDE result for the Aβ42 J-coupling data for Pred-SS and MD-ENS4, the lowest and highest probability ensembles for J-coupling, respectively, are shown in Figure 7.

We can see that for both ensembles, the mean of almost every distribution of back-calculations is within experimental error bars and is often nearly exactly the experimental mean value. In fact, we found that this is true for nearly every experimental measurement, including chemical shifts, in every tested ensemble. This demonstrates that most of the similarites in the EISD probability between ensembles is a result of the error and uncertainty in the back-calculation of experimental observables; in other words, optimizing the EISD model almost always favors lower-probability nuisance parameters over distributions whose means are outside experimental error bars. This suggests that improving the accuracy of the back-calculation from structure is the most crucial step that can be made toward the overall improvement of IDP ensemble determination.

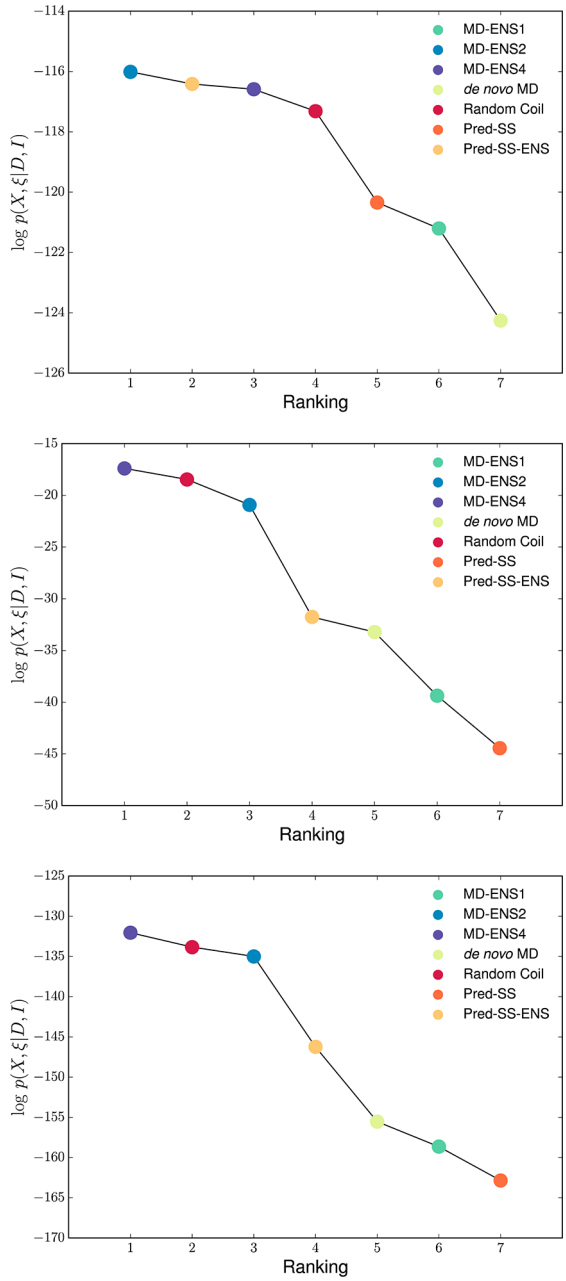


Figure 6. log $p(X,\xi \mid D,I)$ evaluated for X equal to the following qualitatively different ensembles for the Aβ42 monomer: random coil (RC), statistical secondary structure (Pred-SS), *de novo* MD, and ENSEMBLE optimized ensembles (MD-ENS1, MD-ENS2, MD-ENS4, and Pred-SS-ENS) using (a) J-coupling data only, (b) chemical shift data only, and (c) J-coupling and chemical shift data together.

Finally, we consider the improvement that would arise in IDP ensemble ranking if the variances used for the back-calculation from structure to experimental observable were smaller. We emphasize that this is highly artificial for the reason that SHIFTX2 is currently ill-suited to true chemical shift predictions for IDPs; the main utility of this test is to demonstrate what would happen if we had more *confidence* in the chemical shift prediction. Figure 8 illustrates the result if we artificially reduce the variances of the Gaussian distributions by a factor of 3, (improvements which in principle would be possible with highly accurate QM calculations). Now the rankings are becoming more differentiated among structural

DOI: 10.1021/jacs.6b00351
J. Am. Chem. Soc. 2016, 138, 4530−4538

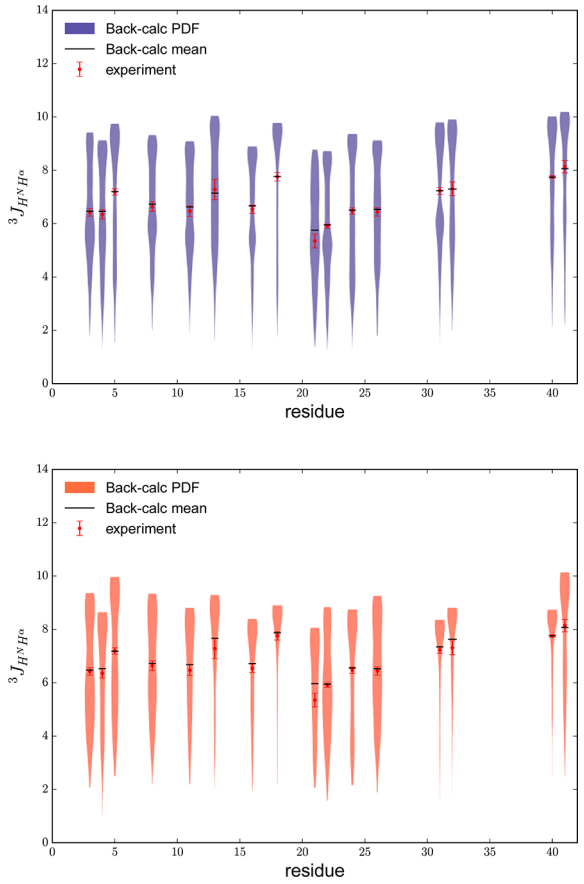


Figure 7. Gaussian Kernel Density Estimation of the probability distributions of back-calculated J-coupling constants from (a) MD-ENS4 and (b) Pred-SS. Wider areas represent higher probabilities.

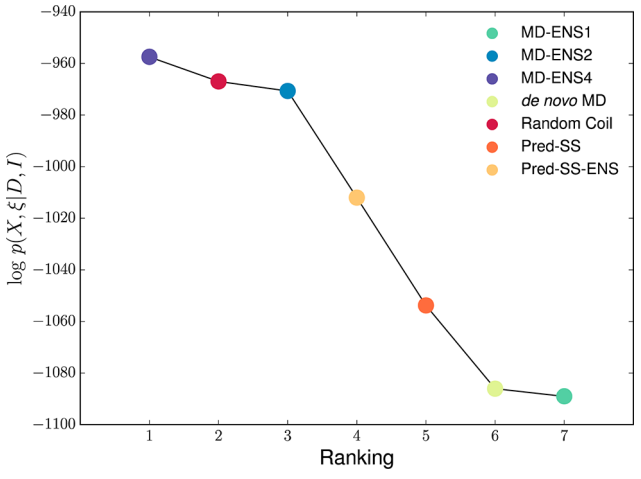


Figure 8. log $p(X,\xi \mid D,I)$ evaluated for X equal to seven different ensembles for the Aβ42 monomer shown in Figure 7 using both J-coupling and chemical shift data but artificially reducing back-calculation uncertainties by a factor of 3.

ensembles. If we had consistent and high quality back-calculations for other data types, such as SAXS, that would likely better differentiate between the MD and RC ensembles for amyloid-β. Combined with other data types and the development of better structural priors, EISD can tractably deliver on even better IDP rankings or structural ensemble refinement using Monte Carlo.

CONCLUSION

Our Bayesian approach differs from previous formulations in the optimization of experimental and back-calculation "nuisance" parameters that are treated as random variables with known Gaussian distributions. Our resulting EISD method is both size extensive with O(N) scaling that allows for the rapid evaluation across very large data sets. When we applied the EISD approach on singular folded proteins and a corresponding set of ∼25 000 misfolded states, we found that uninformative uniform priors performed nearly as well as Boltzmann weighting for two proteins. Furthermore, we showed the problems that can arise using Boltzmann weighted priors for a protein with a disordered segment, which directed us toward using an uninformative structural prior in the formulation of our EISD posterior probability for IDPs.

The EISD formulation presented here offers significant advantages over other existing Bayesian methods since it is size extensive, is able to clearly rank very different IDP ensembles, and the O(N) scaling allows the characterization of very large IDP ensembles of tens of thousands of structures and ease of Metropolis optimization to create new ensembles.

Finally, we showed that what is just as important as a greater range of experimental restraints, better force fields, or computational sampling to create candidate ensembles, is higher accuracy back-calculations from structure for important NMR data types such as chemical shifts and J-couplings. Since the error in NMR experimental measurements for these data types are relatively small, a factor of 3 improvement in the back-calculation error from structure could change this situation, allowing us to better discriminate among alternative structural ensembles and possibly extending the ability to refine for an IDP structural ensemble model given the experimental data. However, because the large number of degrees of freedom for the IDP is much larger than the number of experimental constraints, the underdetermined nature of the ensemble construction problem will continue to be a significant challenge in the future. In order to produce better IDP models, we must (1) produce better back-calculators from all types of experimental data, which reduces one source of degeneracy and (2) create a prior distribution that can accurately reflect the quality of an ensemble before experimental constraints are added in. EISD is sufficiently general to allow for both of these advances to be incorporated.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/jacs.6b00351.

Table of parameters of Gaussian distributions, log $p(X,\xi \mid D,I)$ vs RMSD plots, and optimized coefficients of the Karplus equation plots (PDF)

AUTHOR INFORMATION

Corresponding Author

*thg@berkeley.edu

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

We thank the National Science Foundation Grant CHE-1363320 for support of this work.

DOI: 10.1021/jacs.6b00351
J. Am. Chem. Soc. 2016, 138, 4530−4538

REFERENCES

(1) Kendrew, J. C.; Bodo, G.; Dintzis, H. M.; Parrish, R. G.; Wyckoff, H.; Phillips, D. C. *Nature* **1958**, *181*, 662.
(2) Henderson, R.; Baldwin, J. M.; Ceska, T. A.; Zemlin, F.; Beckmann, E.; Downing, K. H. *J. Mol. Biol.* **1990**, *213*, 899.
(3) Wright, P. E.; Dyson, H. J. *Curr. Opin. Struct. Biol.* **2009**, *19*, 31.
(4) Wright, P. E.; Dyson, H. J. *J. Mol. Biol.* **1999**, *293*, 321.
(5) Herrmann, T.; Güntert, P.; Wüthrich, K. *J. Mol. Biol.* **2002**, *319*, 209.
(6) López-Méndez, B.; Güntert, P. *J. Am. Chem. Soc.* **2006**, *128*, 13112.
(7) Schwieters, C. D.; Kuszewski, J. J.; Tjandra, N.; Clore, G. M. *J. Magn. Reson.* **2003**, *160*, 65.
(8) Schwieters, C. D.; Kuszewski, J. J.; Clore, G. M. *Prog. Nucl. Magn. Reson. Spectrosc.* **2006**, *48*, 47.
(9) Shen, Y.; Bax, A. *J. Biomol. NMR* **2010**, *48*, 13.
(10) Shen, Y.; Delaglio, F.; Cornilescu, G.; Bax, A. *J. Biomol. NMR* **2009**, *44*, 213.
(11) Ball, K. A.; Phillips, A. H.; Nerenberg, P. S.; Fawzi, N. L.; Wemmer, D. E.; Head-Gordon, T. L. *Biochemistry* **2011**, *50*, 7612.
(12) Ball, K. A.; Phillips, A. H.; Wemmer, D. E.; Head-Gordon, T. *Biophys. J.* **2013**, *104*, 2714.
(13) Vazin, T.; Ball, K. A.; Lu, H.; Park, H.; Ataeijannati, Y.; Head-Gordon, T.; Poo, M.; Schaffer, D. V. *Neurobiol. Dis.* **2014**, *62*, 62.
(14) Feldman, H. J.; Hogue, C. W. *Proteins: Struct., Funct., Genet.* **2000**, *39*, 112.
(15) Schneider, R.; Huang, J.-R.; Yao, M.; Communie, G.; Ozenne, V.; Mollica, L.; Salmon, L.; Jensen, M. R.; Blackledge, M. *Mol. BioSyst.* **2012**, *8*, 58.
(16) Huang, A.; Stultz, C. M. *PLoS Comput. Biol.* **2008**, *4*, e1000155.
(17) Yoon, M.; Venkatachalam, V.; Huang, A.; Choi, B.; Stultz, C.; Chou, J. *Protein Sci.* **2009**, *18*, 337.
(18) Krzeminski, M.; Marsh, J. A.; Neale, C.; Choy, W.-Y.; Forman-Kay, J. D. *Bioinformatics* **2013**, *29*, 398.
(19) Fawzi, N. L.; Phillips, A. H.; Ruscio, J. Z.; Doucleff, M.; Wemmer, D. E.; Head-Gordon, T. *J. Am. Chem. Soc.* **2008**, *130*, 6145.
(20) Ball, K. A.; Wemmer, D. E.; Head-Gordon, T. *J. Phys. Chem. B* **2014**, *118*, 6405.
(21) Han, B.; Liu, Y. F.; Ginzinger, S. W.; Wishart, D. S. *J. Biomol. NMR* **2011**, *50*, 43.
(22) Neal, S.; Nip, A. M.; Zhang, H.; Wishart, D. S. *J. Biomol. NMR* **2003**, *26*, 215.
(23) Rieping, W.; Habeck, M.; Nilges, M. *Science* **2005**, *309*, 303.
(24) Fisher, C.; Ullman, O.; Stultz, C. M. *Biophys. J.* **2013**, *104*, 1546.
(25) Fisher, C. K.; Stultz, C. M. *J. Am. Chem. Soc.* **2011**, *133*, 10022.
(26) Fisher, C. K.; Huang, A.; Stultz, C. M. *J. Am. Chem. Soc.* **2010**, *132*, 14919.
(27) Hummer, G.; Kofinger, J. *J. Chem. Phys.* **2015**, *143*, 243150.
(28) Cavalli, A.; Camilloni, C.; Vendruscolo, M. *J. Chem. Phys.* **2013**, *138*, 094112.
(29) Boomsma, W.; Ferkinghoff-Borg, J.; Lindorff-Larsen, K. *PLoS Comput. Biol.* **2014**, *10*, e1003406.
(30) Roux, B.; Weare, J. *J. Chem. Phys.* **2013**, *138*, 084107.
(31) Xiao, X.; Kallenbach, N.; Zhang, Y. *J. Chem. Theory Comput.* **2014**, *10*, 4152.
(32) Sethi, A.; Anunciado, D.; Tian, J.; Vu, D. M.; Gnanakaran, S. *Chem. Phys.* **2013**, *422*, 143.
(33) Antonov, L. D.; Olsson, S.; Boomsma, W.; Hamelryck, T. *Phys. Chem. Chem. Phys.* **2016**, *18*, 5832.
(34) Fisher, C. K.; Stultz, C. M. *Curr. Opin. Struct. Biol.* **2011**, *21*, 426.
(35) Scian, M.; Lin, J. C.; Le Trong, I.; Makhatadze, G. I.; Stenkamp, R. E.; Andersen, N. H. *Proc. Natl. Acad. Sci. U. S. A.* **2012**, *109*, 12521.
(36) Franzoni, L.; Lücke, C.; Pérez, C.; Cavazzini, D.; Rademacher, M.; Ludwig, C.; Spisni, A.; Rossi, G. L.; Rüterjans, H. *J. Biol. Chem.* **2002**, *277*, 21983.
(37) Christendat, D.; Yee, A.; Dharamsi, A.; Kluger, Y.; Savchenko, A.; Cort, J. R.; Booth, V.; Mackereth, C. D.; Saridakis, V.; Ekiel, I.; Kozlov, G.; Maxwell, K. L.; Wu, N.; McIntosh, L. P.; Gehring, K.; Kennedy, M. A.; Davidson, A. R.; Pai, E. F.; Gerstein, M.; Edwards, A. M.; Arrowsmith, C. H. *Nat. Struct. Biol.* **2000**, *7*, 903.
(38) Hinsen, K. *J. Comput. Chem.* **2000**, *21*, 79.
(39) Marsh, J. A.; Forman-Kay, J. D. *J. Mol. Biol.* **2009**, *391*, 359.
(40) Marsh, J. A.; Forman-Kay, J. D. *Proteins: Struct., Funct., Genet.* **2012**, *80*, 556.
(41) Hou, L.; Shao, H.; Zhang, Y.; Li, H.; Menon, N. K.; Neuhaus, E. B.; Brewer, J. M.; Byeon, I. J.; Ray, D. G.; Vitek, M. P.; Iwashita, T.; Makula, R. A.; Przybyla, A. B.; Zagorski, M. G. *J. Am. Chem. Soc.* **2004**, *126*, 1992.
(42) Sgourakis, N. G.; Yan, Y.; McCallum, S. A.; Wang, C.; Garcia, A. E. *J. Mol. Biol.* **2007**, *368*, 1448.
(43) Vuister, G. W.; Delaglio, F.; Bax, A. *J. Biomol. NMR* **1993**, *3*, 67.
(44) Oliphant, T. E. *Comput. Sci. Eng.* **2007**, *9*, 10.
(45) Kohn, J. E.; Afonine, P. V.; Ruscio, J. Z.; Adams, P. D.; Head-Gordon, T. *PLoS Comput. Biol.* **2010**, *6*, e1000911.
(46) Lin, M.; Head-Gordon, T. *J. Comput. Chem.* **2011**, *32*, 709.
(47) Lin, M. S.; Fawzi, N. L.; Head-Gordon, T. *Structure* **2007**, *15*, 727.
(48) Lin, M. S.; Head-Gordon, T. *J. Chem. Theory Comput.* **2008**, *4*, 515.
(49) Fisher, C. K.; Ullman, O.; Stultz, C. M. Efficient Construction of Disordered Protein Ensembles in a Bayesian Framework with Optimal Selection of Conformations. In *Proceedings of the Pacific Symposium*, Kohala Coast, HI, January 3−7, 2012; pp 82−93; DOI: 10.1142/9789814366496_0009
(50) Scott, D. W. *Multivariate Density Estimation: Theory, Practice, and Visualization*; John Wiley & Sons: New York, Chicester, 1992.

DOI: 10.1021/jacs.6b00351
J. Am. Chem. Soc. 2016, 138, 4530−4538