JCTC Journal of Chemical Theory and Computation

Article

pubs.acs.org/JCTC

Probabilistic Determination of Native State Ensembles of Proteins

Simon Olsson,*[†,§] Beat Rolf Vögeli,[‡] Andrea Cavalli,[§] Wouter Boomsma,[∥] Jesper Ferkinghoff-Borg,[⊥] Kresten Lindorff-Larsen,[∥] and Thomas Hamelryck*[†]

[†]Bioinformatics Centre, Department of Biology, Faculty of Science, University of Copenhagen, Copenhagen, Denmark

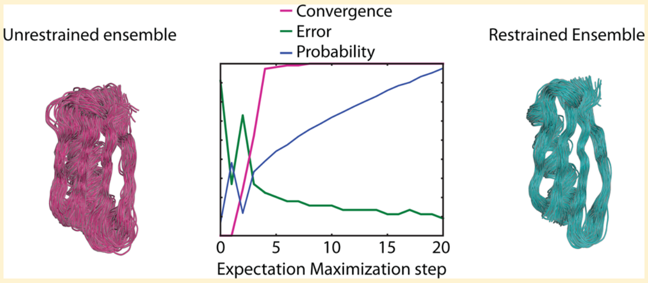
[‡]Laboratory of Physical Chemistry, Eidgenössische Technische Hochschule Zürich, 8093 Zürich, Switzerland

[§]Institute for Research in Biomedicine, CH-6500 Bellinzona, Switzerland

[∥]Structural Biology and NMR Laboratory, Department of Biology, Faculty of Science, University of Copenhagen, Copenhagen, Denmark

[⊥]Cellular Signal Integration Group, Center for Biological Sequence Analysis, Technical University of Denmark, Lyngby, Denmark

S *Supporting Information*



ABSTRACT: The motions of biological macromolecules are tightly coupled to their functions. However, while the study of fast motions has become increasingly feasible in recent years, the study of slower, biologically important motions remains difficult. Here, we present a method to construct native state ensembles of proteins by the combination of physical force fields and experimental data through modern statistical methodology. As an example, we use NMR residual dipolar couplings to determine a native state ensemble of the extensively studied third immunoglobulin binding domain of protein G (GB3). The ensemble accurately describes both local and nonlocal backbone fluctuations as judged by its reproduction of complementary experimental data. While it is difficult to assess precise time-scales of the observed motions, our results suggest that it is possible to construct realistic conformational ensembles of biomolecules very efficiently. The approach may allow for a dramatic reduction in the computational as well as experimental resources needed to obtain accurate conformational ensembles of biological macromolecules in a statistically sound manner.

1. INTRODUCTION

Under physiological conditions the motions of proteins span several spatial and temporal orders of magnitude. Fast and, in particular, slow molecular motions are essential to obtain, sustain, and regulate biological functions which include signal transduction[1] and catalysis.[2] The study of fast molecular motions may be routinely performed using experimental[3] or computational[4] means. Slower motions are, however, notoriously difficult to investigate,[5] requiring either special purpose supercomputers or restrained simulations using extensive data sets.[6,7] Experimental techniques such as NMR relaxation dispersion studies allow extraction of thermodynamic and kinetic information about the exchange between states,[8] but to obtain sufficient structural information to carry out characterization of the dynamics in atomic detail remains difficult.[9−11] In addition, these experiments are currently not widely applicable.[12]

Residual dipolar couplings (RDCs) provide geometric information averaged over a broad range of time scales (up to ms) and may be used to study the slow molecular motions of proteins.[13] Such application of RDCs have already been subject to extensive research.[6,14,15] Several reports describe fitting of discrete sets of conformations to large experimental data sets,[14,15] restraining molecular simulations[16] or selecting sets of protein conformations post hoc to describe structural variability in folded proteins,[17] intrinsically disordered proteins,[18] and flexible RNA molecules.[19] Common to these procedures is the assumption that motion reflected in the data is accurately represented by a weighted average of a discrete set of molecular conformations. It has been shown that, in the limit of an infinite number of conformers, this approach corresponds to a rigorous Maximum Entropy or minimally biasing strategy.[20−23] However, working with infinitely large sets is impractical, and thus, smaller sets are usually employed to good approximation.[22] Still, the procedure is subject to manual optimization of the number of conformations in a set, which introduces a superfluous free parameter. Furthermore, the computational demands of this approach increases linearly with the number of replicas.

In this work, we present a method to construct native state ensembles of biomolecules by combining experimental data with physical force fields in a statistically sound way. As an example, we study the native state ensemble of the third

Received: February 13, 2014
Published: June 17, 2014

immunoglobulin binding domain of protein G (GB3). For this, we curated a modestly sized data set of RDCs, which constitute a subset of previously reported RDCs.[24,25] These RDCs were obtained from conservative mutants of surface residues, which have minor effects on structure and dynamics.[26] The data for the different mutants give complementary views of the N$-$H^N and C$^\alpha-$H$^\alpha$ bonds in the backbone of GB3, which may be difficult to achieve by solely changing alignment media.[27] In principle, our method is applicable to data sets collected under fewer alignment conditions if RDCs between other nuclei are included. The RDCs were used as restraints in a native state simulation using the simple, highly efficient, empirical Profasi force field.[28] The restraints were employed using a strategy recently described by us, which imposes the least necessary bias on a force field that results in agreement with sparse, noisy experimental data.[29] Here, we devise a robust algorithm to iteratively refine this bias. We compare the accuracy of our ensemble to other structural models using an extensive set of complementary, experimental data. These other models include long simulations with state-of-the-art force fields, replica averaged molecular dynamics simulations (RAMD) and ensemble refinement methods. Apart from its computational efficiency, we find two key results support the method's merits. First, the resulting restrained ensemble accurately reproduces the complementary, experimental data to an extent that is comparable to state-of-the-art force fields. Second, we find that correlated motions in adjacent residues are not significantly distorted by our restraining procedure. The motions we find agree with those in previous reports.[6,30]

These results show that our method allows for the construction of native state ensembles of biomolecules by appropriately combining physical force fields with experimental data. The simulations are computationally inexpensive, dramatically reducing the computational and experimental resources necessary to obtain accurate descriptions of the native fluctuations in biomolecular systems.

2. RESULTS

Kullback–Leibler Optimal Restraining of Force Fields. Our aim is to use the empirical force field Profasi in a native state simulation restrained by experimental observations, which are subject to experimental noise, as well as averaging. The computational efficiency of Profasi was a key motivation for its use in this study. While Profasi has been successfully used both for protein folding[31] and aggregation studies,[32] the minimalistic nature of the force field sacrifices some physical details, and complete agreement with experimental data is therefore unlikely. The force field is thus ideal for testing methods, which use available experimental data to bias simulations to obtain more physically realistic models.

We here use previously reported RDC data of the N$-$H^N and C$^\alpha-$H$^\alpha$ bonds[25,26] to restrain the Profasi force field. While these data are highly informative about the relative angles of the interatomic bonds, the amount of information used in the restraining is modest compared to previous studies.[6,30] Consequently, the data alone is not sufficiently informative to provide an accurate description of the native state ensemble of structures. The data is therefore complemented with prior information from the Profasi force field, which provides a simplified view of the conformational behavior of a protein. We seek to restrain the force field using our data in a minimally biased way. This may be achieved by considering the restraining problem as equivalent to Kullback–Leibler optimization. Here, this corresponds to imposing a minimal bias on our prior distribution of the protein conformations (the Boltzmann ensemble of the Profasi force field) in order to make the average of the back-calculated representations match our data within the experimental uncertainty. We have previously described a general approach to perform this optimization.[29]

We start with the posterior distribution,

$$p(\mathbf{f}(\mathbf{x}),\ \mathbf{e},\ \mathbf{x}|\mathbf{d}) \propto p(\mathbf{d}|\mathbf{e},\ \sigma)\frac{\pi_{\mathbf{f}(\mathbf{x})}(\mathbf{f}(\mathbf{x})|\mathbf{e})}{\pi_R(\mathbf{f}(\mathbf{x}))}\pi_{\mathbf{x}}(\mathbf{x})\pi_{\mathbf{e}}(\mathbf{e}) \tag{1}$$

where $\mathbf{f}(\mathbf{x})$ denotes the back-calculated experimental data of a protein structure $\mathbf{x}$; $\mathbf{e}$ represents an *ensemble average* of this representation, and $\mathbf{d}$ represents the experimental data. Here we use a normal distribution, $\mathcal{N}(\cdot)$, for the likelihood $p(\mathbf{d}|\mathbf{e},\sigma)$[33] and a log–linear model $\mathcal{G}(\cdot)$[34] for the prior ratio $((\pi_{\mathbf{f}(\mathbf{x})}(\mathbf{f}(\mathbf{x})|\mathbf{e})/\pi_R(\mathbf{f}(\mathbf{x})))$, which corresponds to Maximum Entropy restraining.[29] The likelihood $\mathcal{N}(\mathbf{d}|\mathbf{e},\ \sigma)$ models the probability of the experimental data given its uncertainty σ and some ensemble average, $\mathbf{e}$. The prior $\mathcal{G}(\mathbf{f}(\mathbf{x})|\mathbf{e},\ \mathbf{B})$ models the distribution of protein conformations given the ensemble average and the scale matrix $\mathbf{B}$ in the data-space. The scale matrix $\mathbf{B}$ relates the units of the experimental data to the energies of Profasi and ensures that the average $\overline{\mathbf{f}(\mathbf{x})}$ matches $\mathbf{e}$. Finally, we chose a uniform prior for the ensemble average, $\pi(\mathbf{e}) \propto 1$. The hierarchical structure of this model was previously outlined.[29] However, the specific distributions have been altered to be in accord with the current experimental data. Further details of the model appear in the Materials and Methods section below.

We wish to obtain a point estimate of the unknown (latent) scale matrix $\mathbf{B}$ such that a given ensemble average, $\mathbf{e}$ yields a posterior expectation of $\mathbf{f}(\mathbf{x})$ according to eq 1, which is exactly $\mathbf{e}$. This corresponds to imposing the least necessary bias to the prior distribution of protein conformations to fulfill the experimental data. We here devise an expectation maximization (EM) algorithm to carry out this estimation (see Materials and Methods and Supporting Information). Additionally, we make use of a new method to account for the different experimental alignment conditions of the data sets (see Materials and Methods).

Expectation Maximization Algorithm Yields Minimally Biased Native State Ensembles. To investigate the native state ensemble of GB3 we use eight sets of previously reported residual dipolar couplings (RDCs)[24] to restrain the Profasi force field.[28] The data was acquired in seven different alignment conditions, resulting from mutants K19AD47K, K19ED40N, K19AT11K, K19EK4A, and two mutants that include a C- and N-terminal His-tag, K19EK4A-C-His6, and K19EK4A-N-His6 along with the wildtype of GB3. Seven of the data sets reported on the N$-$H^N bond whereas one reported on the C$^\alpha-$H$^\alpha$ bond. These conditions have been shown to accurately map the five independent parameters associated with molecular alignment.[26] Consequently, using appropriate structural prior information these data sets should accurately describe the fluctuations of the N$-$H^N bond and, to a lesser extent, the C$^\alpha-$H$^\alpha$ bond.

We used an EM algorithm[35] to refine iteratively the restrained native state ensembles of GB3 (EM$_i$, for i = 0$\cdots$20). We find that increased agreement with the training data (declining Q-factor, Q) correlates with the increase in expected posterior probability (declining expected negative log posterior

Table 1. Reproduction of Experimental Data in Various Structural Models of GB3[a]

	EM$_0$[b]	EM$_8$[c]	2LUM[d]	2OED[e]	MD1[f]	MD2[g]	RAMD[h]
eNOE$_{-6}$ (ρ)	0.66	0.73	0.85	0.79	0.8	0.85	0.7
eNOE$_{-3}$ (ρ)	0.69	0.77	0.89	0.87	0.81	0.86	0.74
$R_{\mathrm{HN}_i/\mathrm{HN}_{i+1}}$ (ρ)	0.63	0.94	0.9	0.93	0.78	0.92	0.91
$R_{\mathrm{HN}_i/\mathrm{H}^\alpha\mathrm{C}^\alpha_i}$ (ρ)	0.90	0.97	0.98	0.97	0.93	0.96	0.94
$R_{\mathrm{HN}_i/\mathrm{H}^\alpha\mathrm{C}^\alpha_{i-1}}$ (ρ)	0.94	0.98	0.96	0.99	0.92	0.97	0.97
$R_{\mathrm{H}^\alpha\mathrm{C}^\alpha_i/\mathrm{H}^\alpha\mathrm{C}^\alpha_{i+1}}$ (ρ)	0.96	0.93	0.92	0.93	0.93	0.93	0.95
$^{h3}J_{\mathrm{NC}'}$ (rmsd, s^{-1})[i]	0.16	0.13	0.14	0.13	0.19	0.19	0.14
$^3J_{\mathrm{H}^N-\mathrm{H}^\alpha}$ (rmsd, s^{-1})	0.97	0.51	0.36	0.45	0.92	0.56	0.85
$^3J_{\mathrm{H}^N-\mathrm{C}^\beta}$ (rmsd, s^{-1})	0.51	0.29	0.30	0.28	0.41	0.26	0.39
$^3J_{\mathrm{H}^N-\mathrm{C}'}$ (rmsd, s^{-1})	0.66	0.36	0.30	0.27	0.53	0.38	0.39
⟨RMSD⟩ (Å)[j]	2.76	1.42	0.72		1.63	1.07	0.85
mean Q-factor[k]	0.41	0.12	0.24	0.13	0.33	0.2	0.14
$\rho(\mathbf{B}_k,\mathbf{B}_{k-1})$[l]		1.0					

[a] Pearson's correlation coefficient is denoted by ρ. [b] Ensemble from unrestrained Profasi force field at 300 K. [c] Ensemble from RDC restrained Profasi force field at 300 K. [d] Refined against eNOE$_{-6}$, $^3J_{\mathrm{H}^N-\mathrm{H}^\alpha}$, $^3J_{\mathrm{H}^N-\mathrm{C}^\beta}$, and $^3J_{\mathrm{H}^N-\mathrm{C}'}$. [e] RDC refined X-ray crystal structure. [f] 10 μs simulation of GB3 using CHARMM22*.[38] [g] 10 μs simulation of GB3 using Amber ff99SB*-ILDN force field.[38] [h] RAMD simulation, restrained using the same data as EM$_8$. Back-calculated values are averages of four 1.1 ns simulations carried out in Almost (see Supporting Information, for simulation details).[40] [i] Ref 45. [j] Average pairwise C$^\alpha$ root-mean square deviation. [k] Average quality factor of training data used to restrain EM$_8$. [l] Correlation of $\mathbf{B}_k$ and $\mathbf{B}_{k-1}$.

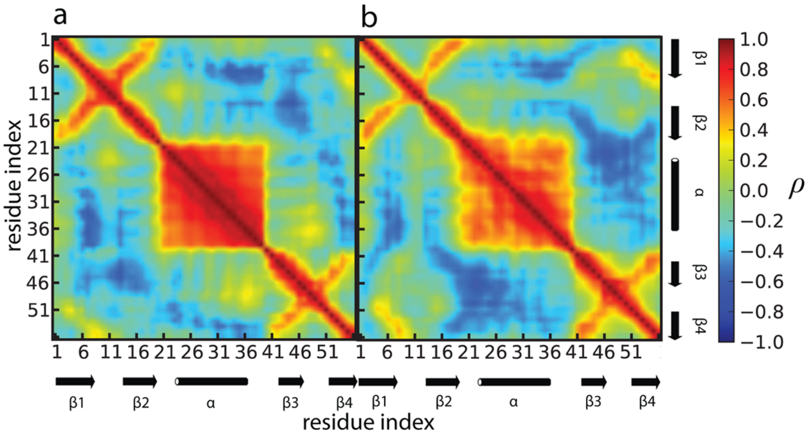


Figure 1. Heatmaps of pairwise correlations, ρ, of fluctuations of C$^\alpha$ atomic positions along the backbone of GB3. (a) unrestrained Profasi (EM$_0$) and (b) restrained Profasi (EM$_8$). Correlation matrices were calculated using THESEUS.[36]

probability, ENLP) of the model and agreement with complementary data not used in the restraining (see Table 1 and Supporting Information). We assess convergence of the estimation after eight iterations, as no significant change in the quality of the ensembles nor significant changes in the latent scale matrix $\mathbf{B}_k$ are observed. The final restrained ensemble (EM$_8$) is stable near the native state of GB3, with considerable flexibility. In contrast, the unrestrained ensemble (EM$_0$) partially unfolds during simulation, see Table 1. Importantly, the restraining of the force field using the RDCs does not significantly perturb the local pairwise positional fluctuations of the C$^\alpha$ atoms, as illustrated by a conservation of near diagonal elements in the correlation matrices calculated using THESEUS[36] (Figure 1). The entire procedure was repeated assuming absence of experimental uncertainty, yielding essentially indistinguishable results (Supporting Information Tables S1 and S2). Furthermore, we found strong agreement between estimated matrices, $\mathbf{B}_8$, resulting from two independent models and estimation runs, with a correlation coefficient of 0.90, suggesting the method is robust and reproducible.

Realistic Local and Nonlocal Motions. While it is promising that the restrained ensemble, EM$_8$, displays considerable flexibility and is in good agreement with the training data (Q = 12%), it does not entail physical realism. We test the quality of EM$_8$ by cross-validation with the wealth of experimental data available for GB3, providing local and nonlocal geometrical information. We compared the reproduction of the experimental data to a number of previously published models of GB3:2LUM, an ensemble refined using exact nuclear Overhauser enhancements (eNOEs), scalar couplings, RDCs, and chemical shifts;[30] 2OED, a RDC refined X-ray structure,[37] and 10 μs molecular dynamics (MD) simulations of GB3 performed with the CHARMM22* (MD1) and Amber ff99SB*-ILDN (MD2) force fields.[38] In addition, we included a comparison of replica-averaged molecular dynamics simulation (RAMD)[39] using the Almost simulation suite[40] restrained using the same data and computational resources as EM$_8$, in order to have a definitive benchmark (see Supporting Information for simulation details). Generally, we use scale sensitive quality measures, such as root-mean-square deviation (rms), for comparison when no scale ambiguity of the data is present—in other cases the Pearson's correlation coefficient is used to assess agreement. In addition, we include previously unpublished cross-correlated dipolar relaxation rate (CCR) data, which are particularly well suited for probing correlated motions (see Supporting Information, Materials and Methods, and Tables S3−6), providing a unique, complementary view of the backbone dynamics on a local length-scale.

As shown in Table 1, EM_8 is in excellent agreement with scalar couplings[41] and CCR data (see Supporting Information, Materials and Methods).[42] The unrestrained EM_0 ensemble is, in general, of considerably worse quality. In comparison, the RAMD ensemble similarly shows considerably better agreement with the data than the unrestrained EM_0. However, compared to the restrained EM_8 ensemble the agreement with the complementary data of the RAMD ensemble is slightly worse. The agreement of EM_8 with the data is comparable to that of previously published models of GB3:2LUM, 2OED and the 10 μs MD simulations. In particular, it is striking that EM_8 shows slightly better quantitative $^3J_{H^N-C^\beta}$ agreement when compared to the 2LUM ensemble, where these data were used in the refinement process. The cross-correlated dipolar relaxation rates $R_{HN_i-HN_{i+1}}$ are particularly sensitive to an accurate representation of the N−H^N bond.[42,43] We see a dramatic improvement in the correlation with these data when comparing the restrained to the unrestrained ensemble. This suggests that the RDCs used to restrain the Profasi force fields indeed accurately describe the fluctuations of the N−H^N bonds. This is further supported by the good correlation of back-calculated N−H order parameters of the EM_8 ensemble to previously reported RDC derived order-parameters,[24,44] see Figure 2. Again, the agreement of the RAMD (ρ = −0.04) ensemble is worse compared to EM_8 (ρ = 0.7).

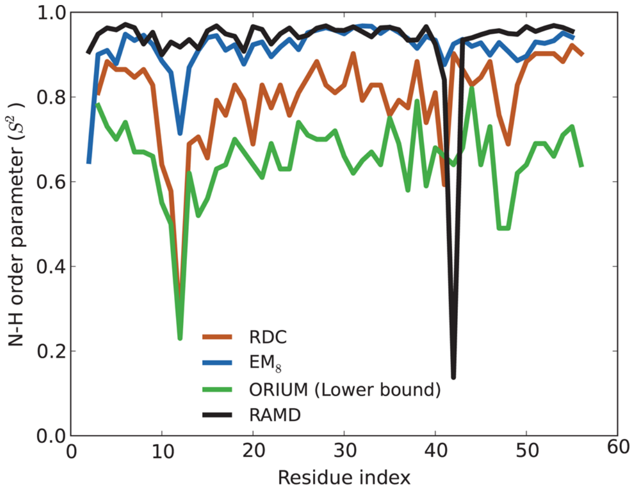


Figure 2. S^2 N−H order parameters (blue) back-calculated from the EM_8 ensemble and previously determined using residual dipolar couplings (red).[24] Order parameters corresponding to the lower-bound (green).[44]

The nonlocal motions were evaluated by cross validation with stereospecifically assigned exact nuclear Overhauser enhancements (eNOEs)[30] and hydrogen bond scalar couplings (HBCs).[45] As with the scalar couplings above, the 2LUM ensemble was refined against the eNOE data. However, here the 2LUM ensemble generally showed the best correlation with these data, closely followed by the simulation with the ff99SB*-ILDN force field (MD2). This is the case both when the conformational dynamics are assumed to be much faster ($eNOE_{-3}$) or slower ($eNOE_{-6}$) than the molecular tumbling.[46] The RDC refined structure, 2OED, and the CHARMM22* simulation (MD1) both show slightly worse agreement while the restrained (EM_8) and unrestrained (EM_0) Profasi ensembles are approximately 10% and 20% worse when compared to 2LUM.

The agreement with the HBCs is quantitatively best in the restrained Profasi ensemble EM_8 of the models presented in Table 1. This is surprising as the ensemble was restrained using local information, which is not expected to be closely related to the information represented in the HBCs. Indeed, this is interesting because the HBCs report long-range correlated motions[47] averaged on the same time scale as RDCs.[48]

Collectively, these results suggest that the local and nonlocal dynamics in EM_8 are of good quality. They may even suggest that the presented method generates ensembles which are compatible with ensembles from molecular dynamics simulations in state-of-the-art force fields on time-scales that are not routinely accessible. However, as Monte Carlo simulations do no provide dynamic information, we wanted to evaluate the agreement for simulations in the same force fields but simulated for one and 2 orders of magnitude shorter. In general, we observe a slightly worse, or on par, agreement with experimental data in the shorter simulations of 100 ns and 1 μs, respectively, see Supporting Information Table S3. One notable exception is the superior agreement of the 100 ns and 1 μs simulations in the ff99SB*-ILDN force field with the HBCs compared to all other ensembles. Consequently, as to be expected from Monte Carlo simulations, we are unable to make precise assessments of which time scales the fluctuations in EM_8 are most compatible.

3. DISCUSSION

Concerted Motion in the Native State Ensemble of GB3. Approximately 90% of the 56 residues of GB3 are in β-sheets or α-helices: an α-helix is wrapped by a mixed antiparallel/parallel four-stranded β-sheet.[49] Consequently, GB3 is considered a very stable and rigid protein, undergoing minor, albeit significant, structural transitions when binding to its biological partner, Fab (see Figure 3).[50] Still, fluctuations of varying magnitude and character throughout this fold have been reported in previous studies. These motions range from

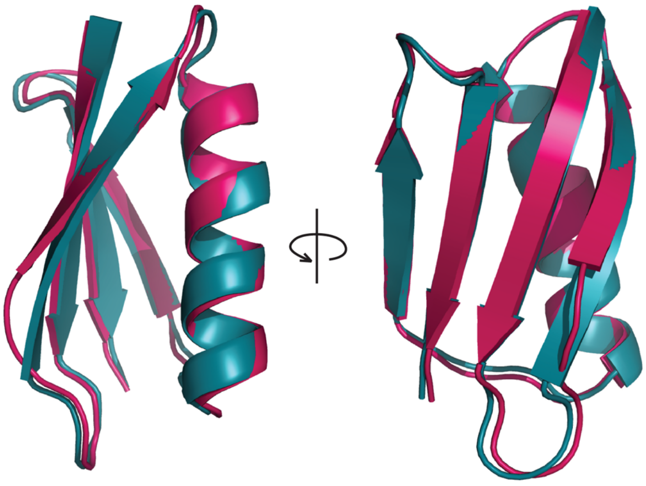


Figure 3. Superpositioning of the free (teal) and Fab-bound (red) forms of GB3.[50]

fast, isotropic bond librations over local crank-shaft moves characterized by anticorrelated changes torsion-angles in adjacent residues[51,52] to long-range correlated motions of the β-sheet across hydrogen bonds.[6,30] In particular the N-terminal sites of the α-helix and the β2-strand have been reported to undergo the most significant fluctuations on subnanosecond time scales. Curiously, the same sites are involved in binding to Fab. In EM_8, we see a similar behavior (Figure 1b), featuring locally correlated motions within the secondary structure elements. In particular, we observe strongly correlated motions within the α-helix and the β-sheet, respectively. While these motions are of relatively subtle amplitude, their correlated nature is strong, likely influenced by the hydrogen bond potential term of the Profasi force field, as the correlated nature of these motions is more subtle in other force fields and ensemble models (see Supporting Information Figure S1).

We observe a number of changes in the long-range correlated motions in the restrained EM_8 ensemble, compared to the unrestrained EM_0 simulation. First, a reinforcement of an inter α-helix and β-sheet anticorrelated motion. We also find a reduction of intra α-helix correlated motions. These changes are apparent in the pairwise correlation heatmap (Figure 1b). The first mode of motion is specifically seen in the off-diagonal areas corresponding to correlations between β-strands 3 and 4 and the α-helix, while the changes in intrahelix motions are observed specifically in the off-diagonal regions corresponding to the motion. Neither of these α-helixes are observed to the same extent in the other ensembles 2LUM, and the CHARMM22*, ff99SB*-ILDN, and RAMD simulations (Supporting Information Figure S1) and, consequently, may reasonably be attributed to the Profasi force field, and not simply the topology of GB3. However, in the first case, it appears that GB3 has to undergo a similar subtle conformational change going from the free-form to the bound form in the GB3:Fab-complex (see Figure 3). Meanwhile, the damping of the intrahelix motions appears to be in better agreement with the previously reported ensembles and the RAMD ensemble. Neither of the ensembles presented here, or those previously described, provide direct evidence for any binding mechanism of GB3 and Fab. However, the currently available evidence supports that GB3 may have an inherent structural plasticity which facilitates the transition between the bound and unbound form in solution.

Propagation of Local Experimental Restraints, into Subtle Translational Restraints. Our results show that the local experimental restraints significantly improve the overall quality of the resulting ensemble. This is similar to what we have previously observed in the context of structure determination when using prior information on the local structure in proteins.[53] Importantly, we do not observe severe distortions in pairwise atomic correlations by imposing the local restraints (Figure 1), which may be interpreted as resulting from the minimally biased nature of the restrained ensemble. We find that measures of both short- and long-range concerted motions are in better agreement with the restrained ensemble as compared to the unrestrained ensemble. This suggests that the angular information provided by the RDC restraints propagate into subtle translational information when combined with the prior information embodied in a physical force field.

In summary, we use the Profasi force field restrained by a set of experimental RDCs subject to averaging as well as experimental noise, to generate an ensemble of the protein GB3. Through cross-validation, we find that the ensemble is in good agreement with complementary experimental data. We identify correlated motions similar to what has been previously reported. In addition, we find an anticorrelated motion between the two major secondary structure assemblies of the protein—compared to the other ensembles discussed, we find this motion to be particularly pronounced in our restrained ensemble. This motion is compatible with the conformational change of GB3 when binding to its biological partner, Fab. While we cannot to directly infer the binding mechanism, we present an intriguing new mode of anticorrelated motions within α-helices and β-sheet assemblies, which are difficult to measure directly in experiments.

These results suggest that it may be generally possible to obtain realistic descriptions of a protein's conformational variability through minimal biasing of simple force fields, without the need of using replica averaged simulations. The quality of the experimental data as well as the force field, of course, remains the determining factor for the realism of these models—however, the results presented here suggest that we may obtain agreement with experiment comparable to state-of-the-art force fields using this approach. Consequently, the approach presented herein may provide an alternative not only to replica averaged simulations but also to simulations based on elaborate force fields. In addition, we find that assuming absence of experimental uncertainty, neither alters our conclusions nor compromises the quality of the inferred ensembles. However, this may be caused by the high quality of the RDC data used here. In the current implementation, we cannot make assessments about the time-scales we consider in the ensembles. Still, the results we obtain here are useful to understand the native state fluctuations of biological macromolecules. We envisage a combination of the presented statistical methodology with molecular dynamics simulations would allow us to discuss the relationship between averaged experimental observations and molecular kinetics. Furthermore, employing enhanced sampling techniques, such as generalized ensembles[54,55] or well-tempered meta-dynamics,[56] in combination with this technique may allow us to probe slow, large-scale motions in biomolecules. The method was implemented in the freely available open source software package Phaistos (www.phaistos.org).[57] An implementation will be made available through the public repository in the near future.

4. MATERIALS AND METHODS

Stochastic Model of Alignment. Here, we introduce a new approach to model the five independent tensor components, **s**, of folded proteins in restrained simulations. We make use of the fact that an ensemble average, **e**, is available during the simulation. One may equate **e** to the experimental data in an ordinary maximum likelihood tensor estimator akin to what was presented earlier.[33] We start with the linear system of equations, $\mathbf{A}\mathbf{s} = \mathbf{e}$ where $\mathbf{A}$ is a $N \times 5$ matrix of N bond vector projections of $\mathbf{x}$ onto the molecular frame, or alignment tensor, **s**. The variable **s** is a 5×1 vector of the alignment tensor components,[58] and **e** is some sampled ensemble average based upon the experimental data and their noise. The system of equations may readily be solved for **s** to obtain the maximum likelihood estimate $\tilde{\mathbf{s}}$. From this, we may obtain the 'back-calculated experimental data', **f**, of the conformational microstate given by the matrix **A** as $\mathbf{f} = \mathbf{A}\tilde{\mathbf{s}}$. This approach has a number of immediate advantages. First, it is independent upon the rotational reference frame. Second, the model is not limited to harmonic variations of the alignment tensor. This means that any anisotropic alignment is readily taken into account. An extreme example of anisotropic alignment is when multiple conformers exchange on a time-scale that is slow compared to the alignment event and consequently, align independently.[59] Finally, this approach avoids a computationally expensive simulation of the alignment tensor.[16,60]

Posterior Distribution. The full posterior distribution is given by

$$p(\mathbf{f}(\mathbf{x}), \mathbf{e}, \mathbf{x}|\mathbf{d}, \mathbf{B}_i) \propto \mathcal{N}(\mathbf{d}|\mathbf{e}, \sigma)\mathcal{G}(\mathbf{f}(\mathbf{x})|\mathbf{e}, \mathbf{B}_i)\exp(-\beta E_{\text{prof}}(\mathbf{x})) \tag{2}$$

where $\mathbf{f}(\mathbf{x})$ is calculated from a conformational microstate $\mathbf{x}$ using the alignment model described above, E_{prof} is the Profasi force field energy described in Supporting Information and β is the inverse temperature, $1/kT$, with Boltzmann constant, k. The experimental data and its uncertainty are represented by $\mathbf{d}$ and $\boldsymbol{\sigma}$, respectively. $\mathbf{e}$ is a variable which represents the average of the instantaneous representations $(\mathbf{f}(\mathbf{x}))$. $\mathcal{G}(\cdot)$ is a log−linear model described in further detail below. The parameter $\mathbf{B}_i$ is estimated iteratively using the EM algorithm described below.

Log−Linear Model. The prior ratio is modeled by a log−linear model $\mathcal{G}(\mathbf{f}(\mathbf{x})|\mathbf{e}, \mathbf{B}_i)$ with a linear link function, $l(\mathbf{B}_i, \mathbf{e}) = \mathbf{B}_i\mathbf{e}$.[34] The log−linear model allows us to define a probability density over the domain of real numbers, $\mathbb{R}$, relevant to residual dipolar couplings that only affect the first moment (the average). The link function allows us to introduce the diagonal $N \times N$-matrix $\mathbf{B}_i$, whose estimation will ensure an appropriate prior in the space of $\mathbf{f}(\mathbf{x})$. We have

$$\mathcal{G}(\mathbf{f}(\mathbf{x})|\mathbf{e}, \mathbf{B}_i) = \exp[c + \mathbf{f}(\mathbf{x})l(\mathbf{B}_i, \mathbf{e})] = \frac{\exp(\mathbf{f}(\mathbf{x})^T\mathbf{B}_i\mathbf{e})}{\mathcal{Z}} \tag{3}$$

where we define the constant $\mathcal{Z} = \exp(-c)$, which is the normalization constant of $\exp(\mathbf{f}(\mathbf{x})^T\mathbf{B}_i\mathbf{e})$. Both $\mathbf{f}(\mathbf{x})$ and $\mathbf{e}$ are $N \times 1$ vectors. Estimation of $\mathbf{B}_i$ proceeds in an Empirical Bayes fashion using an EM algorithm discussed next.

EM Algorithm. The aim of the expectation maximization algorithm is to estimate the scale matrix, $\mathbf{B}_k$, such that the optimal posterior probability distribution (eq 1) is obtained. This is achieved by minimizing the expected negative log-posterior (ENLP).[35]

- Initiate $\mathbf{B}_0$ to some initial guess; here we use the zero-matrix. This corresponds to an unrestrained simulation where the marginal posterior probability distribution of $\mathbf{x}$ and $\mathbf{f}(\mathbf{x})$ coincides with the prior distribution, $\pi_{\mathbf{x}}(\mathbf{x})$—here, the Boltzmann distribution of the Profasi force field at 300 K.
- **E-step** N samples $\mathcal{S} = \{\mathbf{f}(\mathbf{x}), \mathbf{e}, \mathbf{x}\}$ were obtained from the posterior distribution $p(\mathbf{f}(\mathbf{x}),\mathbf{e},\mathbf{x}|\mathbf{d},\mathbf{B}_\mathrm{k})$ by MCMC. Here, we use the Metropolis−Hastings algorithm.[61] The posterior expectations of $\mathbf{e}$ and $\mathbf{f}$ are estimated as the sample means.
- **M-step** This step yields a new scale matrix $\mathbf{B}_{\mathrm{k+1}}$, through importance sampling.[62] With the N samples $\mathcal{S} = \{\mathbf{f}, \mathbf{e}, \mathbf{x}\}$ from the posterior with $\mathbf{B}_\mathrm{k}$, and all other variables being equal the expectations $\bar{\mathbf{e}}$ and $\bar{\mathbf{f}}$ for a given $\mathbf{B}_{\mathrm{k+1}}$ may be approximated by,

$$\bar{\mathbf{e}}(\mathcal{S}, \mathbf{D}) \approx \sum_{i=1}^{N} \mathbf{e}_\mathrm{i}\frac{\exp(\mathbf{f}(\mathbf{x}_\mathrm{i})^T\mathbf{D}\mathbf{e}_\mathrm{i})}{\sum_{j=1}^{N}\exp(\mathbf{f}(\mathbf{x}_\mathrm{j})^T\mathbf{D}\mathbf{e}_\mathrm{j})} \approx \frac{1}{N}\sum_{i=1}^{N}\mathbf{e}_\mathrm{i} \tag{4}$$

and

$$\bar{\mathbf{f}}(\mathcal{S}, \mathbf{D}) \approx \sum_{i=1}^{N} \mathbf{f}_i\frac{\exp(\mathbf{f}(\mathbf{x}_\mathrm{i})^T\mathbf{D}\mathbf{e}_\mathrm{i})}{\sum_{j=1}^{N}\exp(\mathbf{f}(\mathbf{x}_\mathrm{j})^T\mathbf{D}\mathbf{e}_\mathrm{j})} \tag{5}$$

where $\mathbf{D} = \mathbf{B}_{\mathrm{k+1}} - \mathbf{B}_\mathrm{k}$ (see Supporting Information for further details). Here, the sample mean of $\mathbf{e}$ is used as it is insensitive to changes in $\mathbf{B}$.

We may now obtain a $\mathbf{D}$ as

$$\arg\min_{\mathbf{D}}\|\bar{\mathbf{e}}(\mathcal{S}, \mathbf{D}) - \bar{\mathbf{f}}(\mathcal{S}, \mathbf{D})\|^2$$

The minimization is carried out using a simple stochastic gradient descent heuristic:

1. Propose new $\mathbf{B}'$ as a diagonal matrix with elements $\mathcal{N}(0, \varepsilon)\cdot[\bar{\mathbf{e}}(\mathcal{S}, \mathbf{D}) - \bar{\mathbf{f}}(\mathcal{S}, \mathbf{D})]$, where ε is some small value.
2. Evaluate new expectations according to equations above. Accept or reject new value according to the Metropolis criterion.[63]

The iterative procedure was repeated for a total of 21 steps. Convergence was reached at the eighth iteration, as judged by inspecting RDC Q-factors, ENLP and correlation between $\mathbf{B}_k$ and $\mathbf{B}_{k-1}$ (Supporting Information Table S1). Although the ENLP continues to change beyond the eighth iteration, the magnitude of the changes in the changes and the stagnation of the other parameters suggests that sufficient convergence was reached for practical purposes. This phenomenon of slow final convergence is common in expectation maximization algorithms. Each E-step ran for 4×10^7 MCMC steps and the M step is run for 1.5×10^4 minimization steps. The MCMC simulation was carried out in the Phaistos framework[57] and covers the space of $\{\mathbf{x},\mathbf{e},\mathbf{f}(\mathbf{x})\}$. We seeded the simulation with a refined native state structure (PDB: 2OED) to ensure we sampled a relevant region of the phase-space with significant probability in the expected posterior distribution. Snapshots of $\{\mathbf{x},\mathbf{e},\mathbf{f}(\mathbf{x})\}$ were stored every 5×10^3 steps and used in the following M step. The sampling of the all-atom conformational micro states, $\mathbf{x}$, involves local[64] and nonlocal Monte Carlo moves, all of which fulfill detailed balance. A full EM step takes around 8 h on a quad-core desktop computer.

Assuming No Experimental Uncertainty. We repeated the estimation of $\mathbf{B}_k$ assuming experimental uncertainty to be zero, independently, using the protocol described above. We likewise ran the estimation for 21 steps. Results are presented in Supporting Information Tabel S2.

Final Ensembles. The final ensembles EM$_i$ with $i = \{0,1,2,...,20\}$ were constructed by uniformly subsampling 5×10^2 conformations of the 8×10^3 snapshots saved during the E step to ease subsequent analyses. No significant differences were found between different subsampled ensembles.

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

Experimental cross-correlated dipolar relaxation rate data, acquisition details, equations used to back-calculate NMR parameters, additional correlation heat-maps, and tables. This material is available free of charge via the Internet at http://pubs.acs.org/.

■ AUTHOR INFORMATION

Corresponding Authors

*Email: solsson@binf.ku.dk.

*Email: thamelry@binf.ku.dk.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work was funded by the Danish Council for Independent Research (FTP, 09-066546) to S.O., the Swiss National Science Foundation (Grant 140214) to B.V., the Villum Foundation (VKR023445) to W.B., and a Hallas−Møller stipend from the Novo Nordisk Foundation to K.L.L. We thank Lubomir D. Antonov for editorial assistance. Finally, we thank our colleagues at the Bioinformatics centre at the University of Copenhagen for valuable discussions.

■ REFERENCES

(1) Gardino, A. K.; Villali, J.; Kivenson, A.; Lei, M.; Liu, C. F.; Steindel, P.; Eisenmesser, E. Z.; Labeikovsky, W.; Wolf-Watz, M.; Clarkson, M. W.; Kern, D. *Cell* **2009**, *139*, 1109−1118.

(2) Henzler-Wildman, K. A.; Lei, M.; Thai, V.; Kerns, S. J.; Karplus, M.; Kern, D. *Nature* **2007**, *450*, 913−916.

(3) Palmer, A. G., 3rd *Annu. Rev. Biophys. Biomol. Struct.* **2001**, *30*, 129−155.

(4) Karplus, M.; Kuriyan, J. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 6679−6685.

(5) Freddolino, P. L.; Liu, F.; Gruebele, M.; Schulten, K. *Biophys. J.* **2008**, *94*, L75−L77.

(6) Bouvignies, G.; Bernadó, P.; Meier, S.; Cho, K.; Grzesiek, S.; Brüschweiler, R.; Blackledge, M. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 13885−13890.

(7) 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.

(8) Mulder, F. A.; Mittermaier, A.; Hon, B.; Dahlquist, F. W.; Kay, L. E. *Nat. Struct. Biol.* **2001**, *8*, 932−935.

(9) Vallurupalli, P.; Hansen, D. F.; Kay, L. E. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105*, 11766−11771.

(10) Korzhnev, D. M.; Religa, T. L.; Banachewicz, W.; Fersht, A. R.; Kay, L. E. *Science* **2010**, *329*, 1312−1316.

(11) Neudecker, P.; Robustelli, P.; Cavalli, A.; Walsh, P.; Lundström, P.; Zarrine-Afsar, A.; Sharpe, S.; Vendruscolo, M.; Kay, L. E. *Science* **2012**, *336*, 362−366.

(12) Hansen, D. F.; Vallurupalli, P.; Kay, L. E. *J. Biomol. NMR* **2008**, *41*, 113−120.

(13) Tjandra, N.; Bax, A. *Science* **1997**, *278*, 1111−1114.

(14) Lange, O. F.; Lakomek, N.-A.; Farès, C.; Schröder, G. F.; Walter, K. F. A.; Becker, S.; Meiler, J.; Grubmüller, H.; Griesinger, C.; de Groot, B. L. *Science* **2008**, *320*, 1471−1475.

(15) Fenwick, R. B.; Esteban-Marín, S.; Richter, B.; Lee, D.; Walter, K. F. A.; Milovanovic, D.; Becker, S.; Lakomek, N. A.; Griesinger, C.; Salvatella, X. *J. Am. Chem. Soc.* **2011**, *133*, 10336−10339.

(16) Montalvao, R. W.; De Simone, A.; Vendruscolo, M. *J. Biomol. NMR* **2012**, *53*, 281−292.

(17) Guerry, P.; Salmon, L.; Mollica, L.; Ortega Roldan, J.-L.; Markwick, P.; van Nuland, N. A. J.; McCammon, J. A.; Blackledge, M. *Angew. Chem., Int. Ed. Engl.* **2013**, *52*, 3181−3185.

(18) Salmon, L.; Nodet, G.; Ozenne, V.; Yin, G.; Jensen, M. R.; Zweckstetter, M.; Blackledge, M. *J. Am. Chem. Soc.* **2010**, *132*, 8407−8418.

(19) Salmon, L.; Bascom, G.; Andricioaei, I.; Al-Hashimi, H. M. *J. Am. Chem. Soc.* **2013**, *135*, 5457−5466.

(20) Pitera, J. W.; Chodera, J. D. *J. Chem. Theory Comput.* **2012**, *8*, 3445−3451.

(21) Roux, B.; Weare, J. *J. Chem. Phys.* **2013**, *138*, 084107.

(22) Cavalli, A.; Camilloni, C.; Vendruscolo, M. *J. Chem. Phys.* **2013**, *138*, 094112.

(23) Boomsma, W.; Ferkinghoff-Borg, J.; Lindorff-Larsen, K. *PLoS Comput. Biol.* **2014**, *10*, e1003406.

(24) Yao, L.; Vögeli, B.; Ying, J.; Bax, A. *J. Am. Chem. Soc.* **2008**, *130*, 16518−16520.

(25) Yao, L.; Vögeli, B.; Torchia, D. A.; Bax, A. *J. Phys. Chem. B* **2008**, *112*, 6045−6056.

(26) Yao, L.; Bax, A. *J. Am. Chem. Soc.* **2007**, *129*, 11326−11327.

(27) Higman, V. A.; Boyd, J.; Smith, L. J.; Redfield, C. *J. Biomol. NMR* **2011**, *49*, 53−60.

(28) Irbäck, A.; Mitternacht, S.; Mohanty, S. *PMC Biophys.* **2009**, *2*, 2.

(29) Olsson, S.; Frellsen, J.; Boomsma, W.; Mardia, K. V.; Hamelryck, T. *PLoS One* **2013**, *8*, e79439.

(30) Vögeli, B.; Kazemi, S.; Güntert, P.; Riek, R. *Nat. Struct. Mol. Biol.* **2012**, *19*, 1053−1057.

(31) Mohanty, S.; Hansmann, U. H. E. *Biophys. J.* **2006**, *91*, 3573−3578.

(32) Favrin, G.; Irbäck, A.; Mohanty, S. *Biophys. J.* **2004**, *87*, 3657−3664.

(33) Habeck, M.; Nilges, M.; Rieping, W. *J. Biomol. NMR* **2008**, *40*, 135−144.

(34) McCullagh, P.; Nelder, J. A. *Generalized Linear Models*, 2nd ed.; Chapman & Hall: London, 1989.

(35) Dempster, A. P.; Laird, N. M.; Rubin, D. B. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **1977**, *39*, 1−38.

(36) Theobald, D. L.; Wuttke, D. S. *Bioinformatics* **2006**, *22*, 2171−2172.

(37) Ulmer, T. S.; Ramirez, B. E.; Delaglio, F.; Bax, A. *J. Am. Chem. Soc.* **2003**, *125*, 9179−9191.

(38) Lindorff-Larsen, K.; Maragakis, P.; Piana, S.; Eastwood, M. P.; Dror, R. O.; Shaw, D. E. *PLoS One* **2012**, *7*, e32131.

(39) Showalter, S. A.; Brüschweiler, R. *J. Am. Chem. Soc.* **2007**, *129*, 4158−4159.

(40) Fu, B.; Sahakyan, A. B.; Camilloni, C.; Tartaglia, G. G.; Paci, E.; Caflisch, A.; Vendruscolo, M.; Cavalli, A. *J. Comput. Chem.* **2014**, *35*, 1101−1105.

(41) Vögeli, B.; Ying, J.; Grishaev, A.; Bax, A. *J. Am. Chem. Soc.* **2007**, *129*, 9377−9385.

(42) Reif, B.; Hennig, M.; Griesinger, C. *Science* **1997**, *276*, 1230−1233.

(43) Pelupessy, P.; Ravindranathan, S.; Bodenhausen, G. *J. Biomol. NMR* **2003**, *25*, 265−280.

(44) Sabo, T. M.; Smith, C. A.; Ban, D.; Mazur, A.; Lee, D.; Griesinger, C. *J. Biomol. NMR* **2014**, *58*, 287−301.

(45) Cornilescu, G.; Ramirez, B. E.; Frank, M. K.; Clore, G. M.; Gronenborn, A. M.; Bax, A. *J. Am. Chem. Soc.* **1999**, *121*, 6275−6279.

(46) Tropp, J. *J. Chem. Phys.* **1980**, *72*, 6035−6043.

(47) Barfield, M. *J. Am. Chem. Soc.* **2002**, *124*, 4158−4168.

(48) Cordier, F.; Grzesiek, S. *Biochemistry* **2004**, *43*, 11295−11301.

(49) Gronenborn, A. M.; Filpula, D. R.; Essig, N. Z.; Achari, A.; Whitlow, M.; Wingfield, P. T.; Clore, G. M. *Science* **1991**, *253*, 657−661.

(50) Derrick, J. P.; Wigley, D. B. *J. Mol. Biol.* **1994**, *243*, 906−918.

(51) Clore, G. M.; Schwieters, C. D. *Biochemistry* **2004**, *43*, 10678−10691.

(52) Clore, G. M.; Schwieters, C. D. *J. Mol. Biol.* **2006**, *355*, 879−886.

(53) Olsson, S.; Boomsma, W.; Frellsen, J.; Bottaro, S.; Harder, T.; Ferkinghoff-Borg, J.; Hamelryck, T. *J. Magn. Reson.* **2011**, *213*, 182−186.

(54) Swendsen, R. H.; Wang, J.-S. *Phys. Rev. Lett.* **1986**, *57*, 2607−2609.

(55) Tian, P.; Jónsson, S. Æ.; Ferkinghoff-Borg, J.; Krivov, S. V.; Lindorff-Larsen, K.; Irbäck, A.; Boomsma, W. *J. Chem. Theory Comput.* **2014**, *10*, 543−553.

(56) Barducci, A.; Bussi, G.; Parrinello, M. *Phys. Rev. Lett.* **2008**, *100*, 020603.

(57) Boomsma, W.; Frellsen, J.; Harder, T.; Bottaro, S.; Johansson, K.; Tian, P.; Stovgaard, K.; Andreetta, C.; Olsson, S.; Valentin, J.; Antonov, L.; Christiansen, A.; Borg, M.; Jensen, J.; Lindorff-Larsen, K.; Ferkinghoff-Borg, J.; Hamelryck, T. *J. Comput. Chem.* **2013**, *34*, 1697−1705.

(58) Losonczi, J. A.; Andrec, M.; Fischer, M. W.; Prestegard, J. H. *J. Magn. Reson.* **1999**, *138*, 334−342.
(59) Meier, S.; Blackledge, M.; Grzesiek, S. *J. Chem. Phys.* **2008**, *128*, 052204.
(60) Zweckstetter, M.; Hummer, G.; Bax, A. *Biophys. J.* **2004**, *86*, 3444−3460.
(61) Hastings, W. K. *Biometrika* **1970**, *57*, 97−109.
(62) Bishop, C. M. *Pattern Recognition and Machine Learning*; Springer: New York, 2006.
(63) Metropolis, N.; Rosenbluth, A. W.; Rosenbluth, M. N.; Teller, A. H.; Teller, E. *J. Chem. Phys.* **1953**, *21*, 1087−1092.
(64) Bottaro, S.; Boomsma, W.; E. Johansson, K.; Andreetta, C.; Hamelryck, T.; Ferkinghoff-Borg, J. *J. Chem. Theory Comput.* **2012**, *8*, 695−702.