Inferring Structural Ensembles from Noisy Experiments

Kyle A. Beauchamp,[†] Rhiju Das[†,‡] and Vijay S. Pande[†*,¶]

Biophysics Program, Biochemistry Department, Stanford University, Stanford, CA, and Chemistry Department, Stanford University, Stanford, CA

E-mail: rhiju@stanford.edu,pande@stanford.edu

Abstract

Inferring conformation from experiment is the fundamental task of structural biology. Due to limited experimental resolution, structure determination often requires a combination of modelling and experiment. Current algorithms for structure determination, however, are limited to modelling a single conformation and provide only limited uncertainty information. Here we describe linear virtual biasing potential (LVBP), a method that combines simulation and experiment to infer solution ensembles. By using the machinery of Bayesian statistics, LVBP gives rigorous uncertainty estimates for structural features and equilibrium properties. We show that LVBP corrects significant forcefield bias in trialanine.

Introduction

Over the past forty years, structural biologists have solved "ground-state" structures of countless biological macromolecules.[1] Modern biology, however, presents many systems that do not fit the

*To whom correspondence should be addressed
[†]Biophysics Program
[‡]Biochemistry Department
[¶]Chemistry Department

single-structure paradigm–excited states of nucleic acids,[2] natively disordered proteins,[3] and protein folding intermediates[4] alike are poorly described by single conformation models. For such systems, conformational ensembles provide a rigorous framework for understanding structural and equilibrium properties.

Here we introduce a statistical approach to modelling solution ensembles of biological macromolecules. The algorithm, Linear Virtual Biasing Potential (LVBP), uses solution experiments to reweight a collection of atomistic models. In LVBP, Bayesian inference transforms experimental ambiguity into error bars on arbitrary structural features.

To validate LVBP, we investigate two model systems. We first use NMR data to infer conformational ensembles of trialanine.[5] We construct ensembles using five molecular dynamics forcefields with differing conformational propensities. LVBP corrects forcefield errors to provide self-consistent estimates of the α, β, and PP_{II} populations.

Theory: Linear Virtual Biasing Potential

Model Inputs

To model an ensemble using LVBP requires three components (Fig. 1). First, we need a set of conformations (x_j) sampled from the approximate equilibrium distribution of our system. In the present work, such conformations will be generated from molecular dynamics (MD) simulations. Second, we require a set of equilibrium experimental measurements (F_i) and their associated uncertainties (σ_i). Third, it is necessary to have a direct connection between simulation and experiment. This connection is achieved by predicting each experimental observable at each conformation: $f_i(x_j)$ is the predicted value of experiment i at conformation x_j.

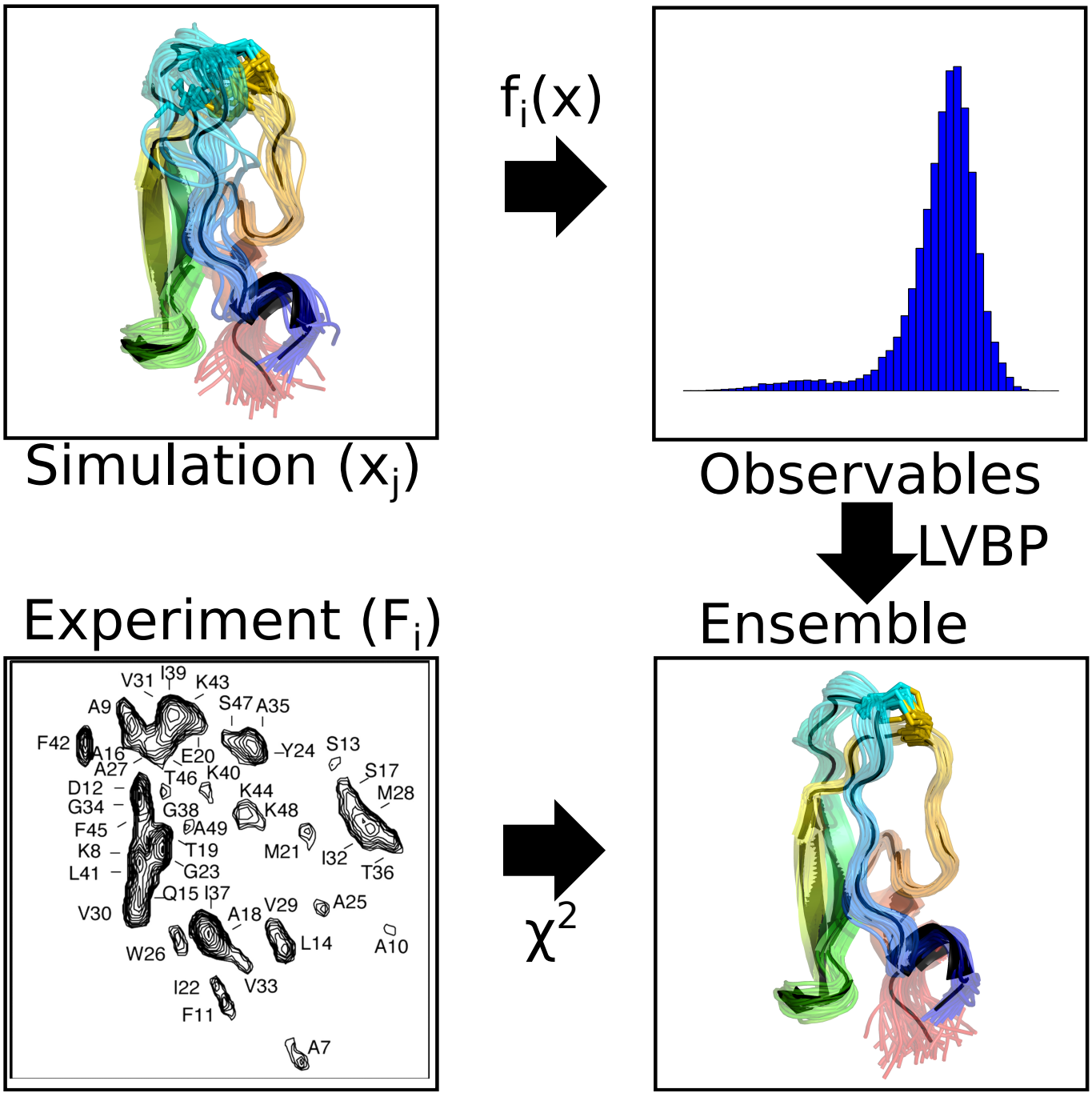


Figure 1: General scheme for LVBP modelling.

Reweighting

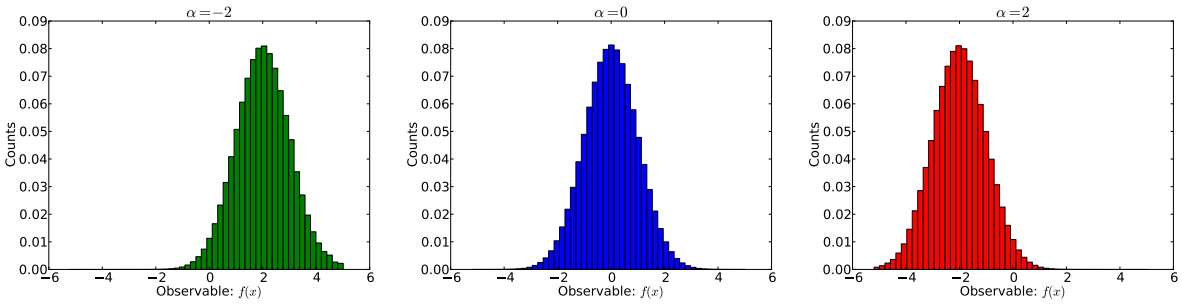
The next step in constructing an ensemble is to calculate the population of each conformation. For convenience, we work with the log populations (i.e. free energies). Inspired by a previous method for restraining simulations,[6] we reweight individual conformations by a biasing potential that is a linear combination of the predicted observables:

$$\Delta U(x;\alpha) = \sum_{i}^{n} \alpha_i f_i(x)$$

In $\Delta U(x;\alpha)$, the parameters α_i determine how strongly each experiment contributes to the biasing potential. Given the biasing potential, the populations of each conformation can be calculated using exponential averaging (see Appx. S1):

$$\pi_j(\alpha) = \frac{1}{\sum_k \exp[-\Delta U(x_k;\alpha)]} \exp[-\Delta U(x_j;\alpha)]$$

It is informative to consider the case of a single observable $f(x)$. Suppose the molecule of interest shows a normally distributed observable. If we let α = 0, then the biasing potential is 0 everywhere and our reweighted ensemble simply returns the results of the MD simulation (Fig. 2b). However, if we let $\alpha = -2$, conformations with large values of $f(x)$ are upweighted, while conformations with lower values of $f(x)$ are downweighted (Fig. 2a). Finally, if $\alpha = +2$, the ensemble shifts in the opposite direction (Fig. 2c).



(a) (b) (c)

Figure 2: Raw ($\alpha = 0$) and reweighted histograms of a one dimensional observable.

With the equilibrium populations, we can calculate the equilibrium expectations of an arbitrary observable $h(x)$:

$$\langle h(x)\rangle_\alpha = \sum_j h(x_j)\pi_j(\alpha)$$

In the above bracket notation, $\langle h(x)\rangle_\alpha$ is the ensemble average of $h(x)$ in an ensemble that is perturbed by a biasing potential $\Delta U(x;\alpha)$. At this point, the determination of α has not yet been discussed. The key idea, however, is that the α reweighted ensemble $\langle\rangle_\alpha$ should recapitulate the experimental measurements:

$$\langle f_i(x)\rangle_\alpha \approx F_i$$

A likelihood framework

We now derive a likelihood framework for determining the coefficients α used in the biasing potential. Given adequate sampling and self-consistent experiments, there should exist some *true* value of α whose ensemble matches the experimental data. However, experimental uncertainty (σ_i) prevents exact agreement between the measurements and the true ensemble. For the models in the current work, we model σ_i as the uncertainty associated with predicting chemical shifts and scalar couplings from structures; this dominant error is quantified by the RMS uncertainty estimated during the parameterization of chemical shift and scalar coupling models. We use an independent normal approximation (see Appx. S2) to model the agreement between the α ensemble and the measurements:

$$P(F_i|\alpha) \sim N(\langle f_i(x)\rangle_\alpha, \sigma_i^2)$$

Using Bayes Theorem, we can calculate the likelihood of α given the measurements:

$$P(\alpha|F_1,...,F_n) \propto P(F_1,...,F_n|\alpha)P(\alpha)$$

Now we let $LL(\alpha)$ denote the log likelihood of α and simplify, dropping terms that are independent of α:

$$LL(\alpha) = \log[P(\alpha|F_1,...,F_n)] = -\sum_i \frac{1}{2\sigma_i^2}(\langle f_i \rangle_\alpha - F_i)^2 + \log P(\alpha)$$

Note the simple form of the log likelihood. The first term measures the χ^2 agreement between the reweighted ensemble and measurements. The second term is the log of the prior distribution on α. Although other priors are possible (see Appxs. S3, S4), we recommend a maximum entropy prior, which penalizes ensembles as they deviate from the raw simulation results:

$$\log P(\alpha) = \lambda \sum_j \pi_j(\alpha) \log \pi_j(\alpha)$$

With large values of λ, the maximum entropy prior favors the raw simulation results (i.e. uniform conformational populations): $\pi_i \approx \frac{1}{n}$. The value of λ can be chosen via cross-validation or other methods (see Appx. S5).

Bayesian Modeling of Structural Ensembles

Because ensemble inference often presents many plausible solutions,[7,8] we avoid statistical methods that return a single solution (e.g. maximum likelihood) . We therefore use Markov chain Monte Carlo (MCMC), as implemented in PyMC,[9] to sample the distribution of structural ensembles consistent with experiment. The result is an ensemble of ensembles–a statistical ensemble of conformational ensembles. Averaging all MCMC samples provides posterior mean estimates of arbitrary structural features. Similarly, examining the MCMC variances provides statistical uncertainties. The shape of the posterior distribution indicates whether one or several models are consistent with the available data (Fig. S1). A Bayesian bootstrapping procedure[10] can also be used to model the statistical uncertainty of the MD simulations (see Appx. S6).

Results

Conformational Propensities of Trialanine

Short peptides provide crucial tests for evaluating and optimizing molecular dynamics forcefields.[5,11–14] Such peptides offer a window into the intrinsic conformational propensities of amino acids, free from the secondary structure bias found in statistical surveys of protein structures.[15] Here, we use LVBP to infer the conformational populations of trialanine from chemical shift and scalar coupling measurements.[5]

Trialanine was simulated (see Methods) in five different force fields; nine experimental measurements (six scalar couplings, three chemical shifts) probing the central ALA residue were used to construct an LVBP ensemble. The five force fields show considerable variation in their agreement with experiment (Fig. 3a). The amber96, amber99, charmm27, and oplsaa forcefields, for example, initially show significant deviation from the experimental measurements. Upon reweighting, however, all five forcefields agree with experiment.

Previous experimental studies have shown that short alanine peptides prefer the polyproline type helix (PPII) in solution.[5,14,16] Most molecular dynamics forcefields, however, are known to underpopulate the PPII state.[5,11–13] Our trialanine simulations recapitulate this known deficiency (Fig. 3 b-d; blue), with amber96 showing a strong β bias and amber99 showing a strong α bias. However, combining simulation and experiment leads to conformational ensembles that are robust to forcefield differences (Fig. 3 b-d, green; Fig. S2). Reweighting with limited experimental data is a powerful technique for correcting forcefield deficiencies.

Discussion

Structural (Ensemble?) Biology

Why model structural ensembles, rather than just structures? At least three compelling reasons favor ensembles. First, biological molecules are multi-state machines; they fold, unfold, bind lig-

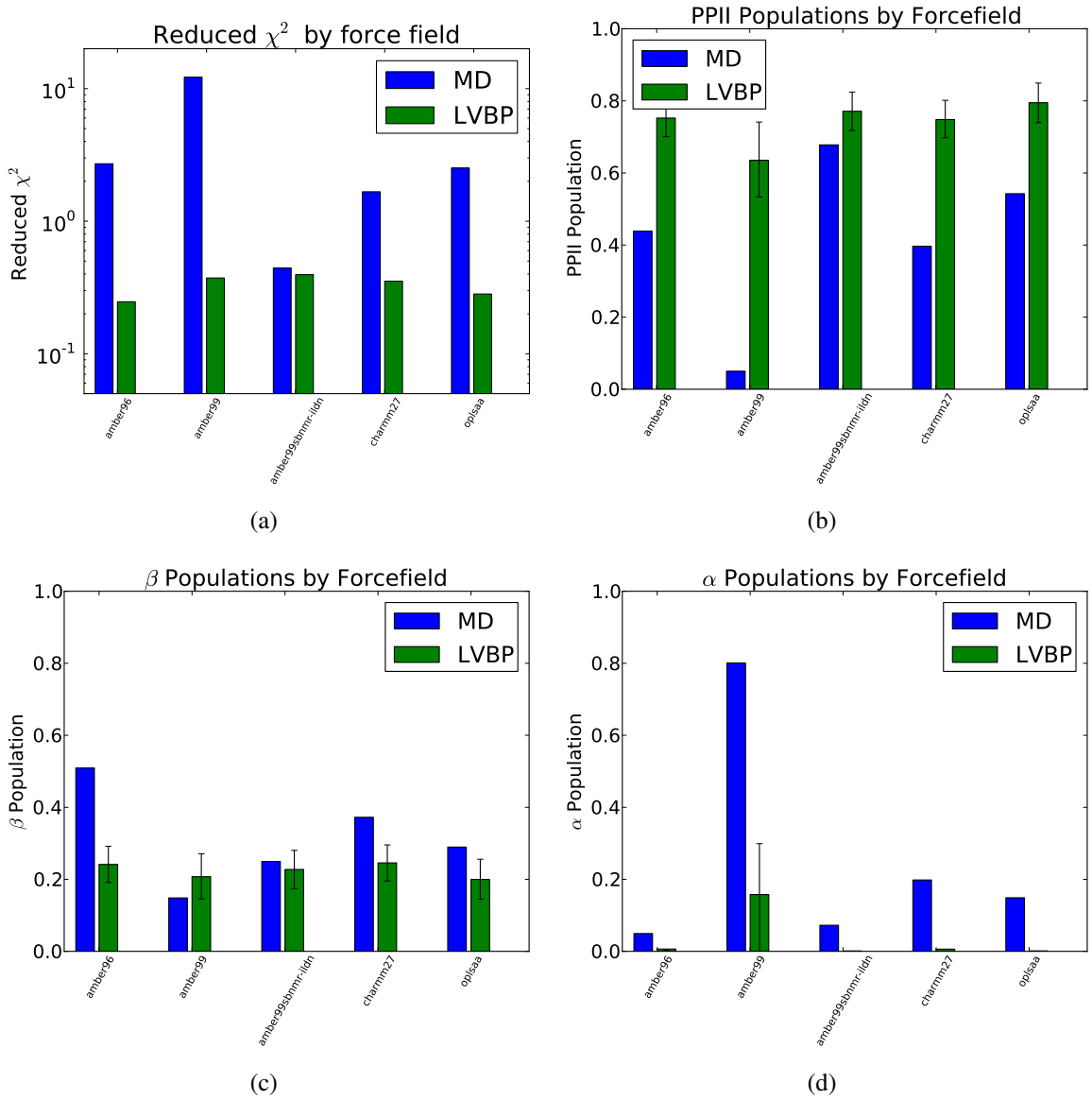


Figure 3: (a). The reduced χ^2 error for each raw and reweighted simulation. The LVBP reduced χ^2 is estimated as the mean reduced χ^2 over all MCMC samples. (b-d). Raw and reweighted conformational propensities for each force field. Note that amber99 shows larger uncertainties due to the extremely limited amount of PP_{II} sampling in the amber99 forcefield.

ands, aggregate, and change conformation. Biology is controlled by the relative populations of these states. Structural ensembles capture aspects of these phenomena by encoding equilibrium populations. A second argument for ensemble modelling is fidelity to experiment. Most solution experiments measure ensemble average equilibrium properties–chemical shifts, scalar couplings, NOEs, SAXS, and FRET are reasonably well-described as equilibrium properties. A truly quantitative connection to these measurements requires modelling the equilibrium ensemble. Finally, recent advances in atomistic simulation,[17–20] special-purpose hardware,[21] and distributed computing analysis[22,23] have enabled atomistic simulations to reach the millisecond timescale;[24–27] the computational cost of ensemble modeling is quickly becoming manageable.

One might argue that structural ensembles are unnecessary because many proteins occupy a single state under physiological conditions. For such proteins, it is probably safe to enforce single state behavior, as is done in current modelling approaches. However, we suggest that the number of states be *inferred*–not *assumed*.

How Accurate is the BPTI Ensemble?

It is critical to evaluate the accuracy limits of the current LVBP ensemble. Consider three different models for BPTI: the X-ray model, the raw MD simulation, and the LVBP ensemble. For the 53 trypsin-binding loop chemical shifts, these three models give reduced χ^2 values of 0.54, 1.6, and 0.67, respectively. According to this metric, we *reject* the raw simulation in favor of the other models. Comparing the X-ray and LVBP models is more subtle. According to the reduced χ^2 values, the X-ray model provides slightly better agreement with the chemical shift data. However, the difference (0.13) is small relative to the uncertainty in the reduced χ^2 (0.19). Regardless, both values lie below 1.0, suggesting that model comparison is limited by chemical shift prediction uncertainty. We also evaluated the three models using ten $^3J(H^N H^\alpha)$ couplings;[28] by this metric, the LVBP model performs best. Overall, we conclude that the LVBP ensemble provides a reasonable picture of BPTI in solution. Future improvements in forcefield accuracy and chemical shift prediction may help refine the current ensemble.

Implication for Trypsin Binding

The present LVBP ensemble suggests multiple conformational states in the trypsin binding loop of BPTI. These conformational states involve coupled motions involving the C14-C38 χ_1 torsions and the backbone torsions of K15, A16, and R17. Notably, these residues are critical determinants of specificity in binding trypsin or chymotrypsin;[29] peptide scission by trypsin occurs between residues K15 and A16.[30] Why is disorder present in the site of inhibitor specificity and peptide scission?

Conclusion

Bayesian inference of conformational ensembles allows the simultaneous characterization of structural and equilibrium properties. Limited experimental data can be used to supplement equilibrium simulations, leading to predictions that are robust to forcefield error and consistent with available measurements. Our equilibrium ensemble of BPTI suggests coupled motion in the trypsin specificity loop and the C14-C38 disulfide bridge. Falsifying or refining the present model offers a key challenge for understanding the conformational fluctuations in BPTI.

Acknowledgements

We thank John Chodera, TJ Lane, Frank Cochran, Pehr Harbury, Xuesong Shi, and Dan Herschlag for helpful discussions. We thank DE Shaw research for providing the millisecond simulation of BPTI.

Methods

Methods: Trialanine

Trialanine was simulated in the amber96,[31] amber99,[32] amber99sbnmr-ildn,[33] charmm27,[34,35] and oplsaa[36] force fields, as previously reported.[11] Simulations were performed using Gromacs 4.5[17] and run at constant temperature (300 K) and pressure (1.01 atm). Each simulation was at least 225 ns long. Conformations were stored every 5 ps. Chemical shifts (H, HA, CA, CB) for each frame were calculated using the average prediction of ShiftX2,[37] SPARTA+,[38] and PPM;[39] uncertainties were estimated as the mean uncertainty reported by SPARTA+. J couplings were calculated using the following Karplus relations: ${}^3J(H^N C')$,[40] ${}^3J(H^N H^\alpha)$,[41] ${}^2J(NC^\alpha)$,[5] ${}^3J(H^\alpha C')$,[40] ${}^1J(NC^\alpha)$,[5] ${}^3J(H^N C^\beta)$ citevogeli2007limits. J coupling uncertainties were approximated as the RMS errors reported when fitting the Karplus coefficients.

References

(1) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. *Nucleic Acids Res.* **2000**, *28*, 235–242.

(2) Dethoff, E. A.; Petzold, K.; Chugh, J.; Casiano-Negroni, A.; Al-Hashimi, H. M. *Nature* **2012**,

(3) Fink, A. L. *Current opinion in structural biology* **2005**, *15*, 35–41.

(4) Korzhnev, D. M.; Salvatella, X.; Vendruscolo, M.; Di Nardo, A. A.; Davidson, A. R.; Dobson, C. M.; Kay, L. E. *Nature* **2004**, *430*, 586–590.

(5) Graf, J.; Nguyen, P.; Stock, G.; Schwalbe, H. *J. Am. Chem. Soc.* **2007**, *129*, 1179–1189.

(6) Pitera, J.; Chodera, J. *Journal of Chemical Theory and Computation* **2012**,

(7) Fisher, C. K.; Huang, A.; Stultz, C. M. *Journal of the American Chemical Society* **2010**, *132*, 14919.

(8) Rieping, W.; Habeck, M.; Nilges, M. *Science* **2005**, *309*, 303–306.

(9) Patil, A.; Huard, D.; Fonnesbeck, C. J. *Journal of statistical software* **2010**, *35*, 1.

(10) Rubin, D. *The annals of statistics* **1981**, *9*, 130–134.

(11) Beauchamp, K.; Lin, Y.; Das, R.; Pande, V. *J. Chem. Theory Comput.* **2012**, *8*, 1409.

(12) Nerenberg, P.; Head-Gordon, T. *J. Chem. Theory Comput.* **2011**, *7*, 1220–1230.

(13) Best, R.; Buchete, N.; Hummer, G. *Biophys. J.* **2008**, *95*, L07–L09.

(14) Grdadolnik, J.; Mohacek-Grosev, V.; Baldwin, R.; Avbelj, F. *Proc. Natl. Acad. Sci. U. S. A.* **2011**, *108*, 1794.

(15) Jha, A.; Colubri, A.; Zaman, M.; Koide, S.; Sosnick, T.; Freed, K. *Biochemistry* **2005**, *44*, 9691–9702.

(16) Avbelj, F.; Grdadolnik, S.; Grdadolnik, J.; Baldwin, R. *Proc. Natl. Acad. Sci. U. S. A.* **2006**, *103*, 1272.

(17) Hess, B.; Kutzner, C.; Van Der Spoel, D.; Lindahl, E. *J. Chem. Theory Comput.* **2008**, *4*, 435–447.

(18) Pronk, S.; Pall, S.; Schulz, R.; Larsson, P.; Bjelkmar, P.; Apostolov, R.; Shirts, M. R.; Smith, J. C.; Kasson, P. M.; van der Spoel, D. *Bioinformatics* **2013**,

(19) Eastman, P.; Friedrichs, M. S.; Chodera, J. D.; Radmer, R. J.; Bruns, C. M.; Ku, J. P.; Beauchamp, K. A.; Lane, T. J.; Wang, L.-P.; Shukla, D.; Pande, V. S. *Journal of chemical theory and computation* **2012**, *9*, 461–469.

(20) Eastman, P.; Pande, V. *Comp. in Sci. Eng.* **2010**, *12*, 34–39.

(21) Shaw, D.; Deneroff, M.; Dror, R.; Kuskin, J.; Larson, R.; Salmon, J.; Young, C.; Batson, B.; Bowers, K.; Chao, J. *Commun. ACM* **2008**, *51*, 91–97.

(22) Senne, M.; Trendelkamp-Schroer, B.; Mey, A. S. J. S.; Schütte, C.; Noé, F. *J. Chem. Theory Comput.* **2012**,

(23) Beauchamp, K.; Bowman, G.; Lane, T.; Maibaum, L.; Haque, I.; Pande, V. *J. Chem. Theory Comput.* **2011**, *7*, 3412–3419.

(24) Voelz, V.; Bowman, G.; Beauchamp, K.; Pande, V. *J. Am. Chem. Soc.* **2010**, *132*, 1526–1528.

(25) Bowman, G. R.; Voelz, V. A.; Pande, V. S. *Journal of the American Chemical Society* **2011**, *133*, 664.

(26) 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.

(27) Lindorff-Larsen, K.; Piana, S.; Dror, R.; Shaw, D. *Science* **2011**, *334*, 517–520.

(28) Balasubramanian, S.; Nirmala, R.; Beveridge, D. L.; Bolton, P. H. *Journal of Magnetic Resonance, Series B* **1994**, *104*, 240–249.

(29) Scheidig, A. J.; Hynes, T. R.; Pelletier, L. A.; Wells, J. A.; Kossiakoff, A. A. *Protein science* **1997**, *6*, 1806–1824.

(30) Moh'd A, S.; Soares, A. S.; Hockla, A.; Radisky, E. S. *Journal of Biological Chemistry* **2008**, *283*, 4115–4123.

(31) Kollman, P. *Acc. Chem. Res.* **1996**, *29*, 461–469.

(32) Wang, J.; Cieplak, P.; Kollman, P. *J. Comput. Chem.* **2000**, *21*, 1049–1074.

(33) Li, D.; Bruschweiler, R. *Angew. Chem.* **2010**, *122*, 6930–6932.

(34) Mackerell Jr, A.; Feig, M.; Brooks III, C. *J. Comput. Chem.* **2004**, *25*, 1400–1415.

(35) Bjelkmar, P.; Larsson, P.; Cuendet, M.; Hess, B.; Lindahl, E. *J. Chem. Theory Comput.* **2010**, *6*, 459–466.

(36) Kaminski, G.; Friesner, R.; Tirado-Rives, J.; Jorgensen, W. *J. Phys. Chem. B* **2001**, *105*, 6474–6487.

(37) Han, B.; Liu, Y.; Ginzinger, S.; Wishart, D. *J. Biomol. NMR* **2011**, 1–15.

(38) Shen, Y.; Bax, A. *J. Biomol. NMR* **2010**, *48*, 13–22.

(39) Li, D.-W.; Brüschweiler, R. *Journal of Biomolecular NMR* **2012**, 1–9.

(40) Schmidt, J.; Blümel, M.; Löhr, F.; Rüterjans, H. *J. Biomol. NMR* **1999**, *14*, 1–12.

(41) Vögeli, B.; Ying, J.; Grishaev, A.; Bax, A. *Journal of the American Chemical Society* **2007**, *129*, 9377–9385.